Development and application of gene mapping approaches, towards an integrated physical, radiation-hybrid and comparative map of dog chromosome five

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Rachael Thomas B.Sc. (Hons.) Bath
Department of Biochemistry
University of Leicester

April 2000

Statement of originality

This thesis, submitted for the degree of Doctor of Philosophy, entitled "Development and application of gene mapping approaches, towards an integrated physical, radiation-hybrid and comparative map of dog chromosome five", is based on work conducted by the author in the Department of Biochemistry at the University of Leicester, and in the Centre for Preventive Medicine at the Animal Health Trust, mainly during the period between October 1996 and December 1999.

All the work recorded in this thesis is original unless otherwise acknowledged in the text or by references.

None of the work has been submitted for another degree in this or any other University.

Signed..... Date.....

Acknowledgements

I would like to thank Dr. Matthew Breen, Dr. N Holmes and Dr. M. Binns for their enthusiasm, encouragement and support during the course of this work, and throughout the past three years. I am grateful also to Dr. Jeff Sampson for his interest and input into the work described in chapter three, and to Dr Anne Willis for playing a role following relocation of the laboratory. I would like to acknowledge the Animal Health Trust, the BBSRC, the Guide Dogs for the Blind Association, and the Colin Wells Memorial Trust for financial support.

I am grateful to Dr. Panos Deloukas for enabling me to perform radiation-hybrid mapping data within his laboratory, and to the members of his group for their patience and advice. I gratefully acknowledge Cordelia Langford for provision of chromosome paint probe material, and for her interest and advice.

Thanks to Dr. Lucy Coles for friendship and understanding- I promise I will visit more often now! Thanks to all the members of the Genetics section at the Animal Health Trust, past and present, especially Ed Ryder, Kerry Benke and Emily McCann, and also Steve Shaw, for their sense of humour and continuous support, and for feeding me chocolate and coffee at regular intervals. Thanks also to Matthew and Tessa Breen for their 'pastoral care!'. Helen Spriggs has been a constant source of encouragement and friendship over the last few years, and I am grateful to Helen and her family for 'adopting' me.

Finally, I wish to acknowledge the love and understanding of my family, present and past, who have never let me down, and who have been so patient and supportive in helping me to reach my goals. This is for you!!

Publications

Thomas, R., Breen, M., Langford, C.F. and Binns, M.M. (1999). Zoo-FISH analysis of dog chromosome 5: Identification of conserved synteny with human and cat chromosomes. *Cytogenetics and Cell Genetics* **87** 4-10.

Breen, M., Thomas, R., Binns, M.M., Carter, N.P. and Langford, C.F. (1999). Reciprocal chromosome painting reveals detailed regions of conserved synteny between the karyotypes of the domestic dog (*Canis familiaris*) and human. *Genomics* **61** 145-155.

Breen, M., Langford, C.F., Carter, N.P., Holmes, N.G., Dickens, H.F., Thomas, R., Suter, N., Ryder, E.J., Pope, M. and Binns, M.M. (1999). FISH mapping and identification of canine chromosomes. *Journal of Heredity* **90** 27-30.

Dickens, H.F., Holmes, N.G., Ryder, E.J., Breen, M., Thomas, R., Suter, N., Sampson, J., Langford, C.F., Ross, M., Carter, N.P. and Binns, M.M. (1999). Use of cosmid-derived and chromosome-specific canine microsatellites. *Journal of Heredity* **90** 52-54.

Thomas, R., Holmes, N.G., Fischer, P.E., Dickens, H.F., Breen, M., Sampson, J. and Binns, M.M. (1997). Eight canine microsatellites. *Animal Genetics* **28** 152-153.

Fischer, P.E., Holmes, N.G., Dickens, H.F., Thomas, R., Binns, M.M. and Nacheva, E.P. (1996). The application of FISH techniques for physical mapping in the dog (*Canis familiaris*). *Mammalian Genome* **7** 37-41.

Contents

Chapter two:

Materials and Methods

Figures

Page numbers refer to the text page immediately preceding the corresponding figure.

Chapter five

Chapter six

Appendices

Tables

Page numbers refer to the text page immediately preceding the corresponding table.

Development and application of gene mapping approaches, towards an integrated physical, radiation-hybrid and comparative map of dog chromosome five

Rachael Thomas

Efforts to generate a genome map for the domestic dog (*Canis familiaris,* CFA) have to date largely focussed on the isolation and characterisation of anonymous markers. This study was aimed at the development and application of techniques by which gene markers might be added to the map and integrated with data both from other approaches and from the genomes of other species. This is essential for the continuing use of the dog as a model system for the study of inherited traits.

The generation of expressed sequence tags (partial cDNA sequences) was investigated as a route for achieving this aim. Seventeen of 76 cDNA clones analysed shared significant nucleotide similarity with previously annotated gene sequences from other species, forming a panel of resources for the extension of the existing dog gene map. Two gene markers were assigned to a specific dog chromosome by fluorescence *in situ* hybridisation.

In a contrasting approach, reciprocal chromosome painting analysis was used to identify evolutionarily conserved chromosome segments (ECCS) between dog chromosome five (CFA 5) and human chromosomes 1, 11, 16 and 17. The boundaries between ECCS, and their relative orientation in the two genomes, were investigated by the assignment to CFA 5 of 12 genes selected from loci located at proximal and distal extremes of the corresponding human ECCS. Radiation hybrid mapping was performed for eleven gene markers and ten anonymous markers from CFA 5. One marker of each type was also analysed by meiotic linkage analysis. Direct cDNA selection, by isolation of chromosome-enriched cDNA populations, was proposed and investigated as a technique for further enhancement of the dog gene map.

An integrated physical, radiation-hybrid and comparative map of CFA 5 is presented, and the relative merits of the mapping techniques investigated are discussed with reference to future prospects for the dog gene map.

Abbreviations

°C	degrees Celsius
λ	lambda
-ve	negative
+ve	positive
μ	micro-
μg	microgram
μl	microlitre(s)
μM	micromolar
θ	recombination fraction
A	adenosine
A_x	absorbance (at 'x' nanometers)
APS	ammonium persulphate
BAC	bacterial artificial chromosome(s)
BLAST	basic local alignment search tool
bp	basepair(s)
Bq	becquerel(s)
BSA	bovine serum albumin
BTA	*Bos taurus*
C	cytosine
CAR	ancestral carnivore karyotype
CATS	comparative anchor tagged sequence(s)
CCD	charge coupled device
cDNA	complementary DNA
CFA	*Canis familiaris*
CGH	comparative genomic hybridisation
Ci	curie(s)
cm	centimetres
cM	centiMorgan(s)
cps	counts per second
cR	centiRay(s)
CSNB	congenital stationary night blindness
DAPI	4', 6-diamidino-2-phenylindole
dATP	deoxyadenosine triphosphate
dbEST	database of expressed sequence tags
dCTP	deoxycytosine triphosphate
dGTP	deoxyguanosine triphosphate
dist.	distal
DNA	deoxyribonucleic acid
DNase	deoxyribonuclease
dNTP	deoxynucleoside triphosphate
DOP-PCR	degenerate oligonucleotide-primed polymerase chain reaction
dTTP	deoxythymidine triphosphate
dUTP	deoxyuridine triphosphate
ECA	*Equus caballus*
ECCS	evolutionarily conserved chromosome segment(s)
E. coli	*Escherichia coli*
EDTA	disodium ethylenediamine tetraacetic acid
EST	expressed sequence tag(s)
FCA	*Felis catus*
FCS	foetal calf serum
FISH	fluorescence *in situ* hybridisation
FITC	fluorescein isothiocyanate
g	gram
G	guanosine
GB4	Genebridge 4
gDNA	genomic DNA
HAT	hypoxanthine, aminopterin & thymidine
HCl	hydrochloric acid
HGND	Human Gene Nomenclature Database
HPRT	hypoxanthine phosphoribosyltransferase
HSA	*Homo sapiens*
kb	kilobase(s)
I.M.A.G.E.	Integrated Molecular Analysis of Genomes and their Expression
l	litre(s)
LB	Luria Bertani broth
LINE	long interspersed nucleotide element(s)
ln	natural logarithm
m	milli-
M	molar
MAU	*Mesocricetus auratus*
Mb	megabase(s)
mCi	milliCurie(s)
MBq	megabequerel(s)
MHC	major histocompatability complex
min	minute(s)

MMcC	microdissection-mediated cDNA capture
mmol	millimole(s)
MMU	*Mus musculus*
ml	millilitre(s)
MPC	magnetic particle concentrator
MQ	'milliQ' water
mRNA	messenger RNA
n/a	not applicable
NCBI	National Center for Biotechnology Information
nd	not determined
ng	nanogram(s)
NIH	National Institutes of Health
no.	number
nr	non-redundant
OD_x	optical density (at 'x' nanometers)
OMIM	Online Mendelian Inheritence in Man
ORF	open reading frame
ORI	origin of replication
PCR	polymerase chain reaction
pers. comm.	personal communication
pfu	plaque forming unit(s)
PNACL	protein and nucleic acids chemistry laboratory
poly(A)	polyadenylate
prox.	proximal
RH	radiation hybrid(s)
RNA	ribonucleic acid
RNase	ribonuclease
rpm	revolutions per minute
SCEUS	smallest conserved evolutionary unit segment
SCH	somatic cell hybrid(s)
SDS	sodium dodecyl sulphate
SINE	short interspersed nucleotide element(s)
SNP	single nucleotide polymorphism(s)
SSC	*Sus scrofa*/trisodium citrate and sodium chloride
STS	sequence-tagged site(s)
T	thymidine
TAE	tris-acetate-EDTA buffer
Taq	*Thermus aquaticus*
TBE	tris-borate EDTA
TBq	terabequerel(s)
TD	touchdown
TEMED	N, N, N', N'-tetramethylenediamine
TIGR	The Institute of Genome Research
TK	thymidine kinase
Tm	melting temperature
TNG	"The Next Generation"
TOAST	traced orthologous amplified sequence tag(s)
u	unit(s)
UDB	Unified Database
UK	United Kingdom
UM-STS	universal mammalian sequence-tagged site(s)
US	United States of America
UTR	untranslated region
UV	ultraviolet
v/v	volume:volume ratio
w/v	weight:volume ratio
WG-RH	whole genome radiation hybrid
YAC	yeast artificial chromosome

Chapter one:

Introduction

Introduction

1.1 Introduction

The domestic dog (*Canis familiaris*, CFA) is becoming increasingly popular as a model species for the study of inherited traits. Each distinct breed exhibits physical and behavioural traits that form the characteristics of that breed, and in order for these inherited traits to be propagated, breeding programs frequently revolve around a small number of individuals that demonstrate the desired features. One of the consequences of this limited gene pool is that inherited disorders, particularly simple autosomal recessive traits, are widespread in purebred dogs. Carriers of defective, recessive alleles may be asymptomatic, but when two such individuals are crossed, the faulty allele is transmitted to following generations, in a pattern that is determined by the inheritance characteristics of that trait. Inevitably, a proportion of affected offspring will be generated, with obvious impact on both the animal and its owner, and in turn further propagating the mutation within the breed. In excess of 370 genetic disorders are estimated to exist amongst over 400 dog breeds (Patterson, 1993; Patterson, in press, cited in Ostrander *et al.*, 2000; J. Sampson, pers. comm.). The dog has been described as the species for which the most naturally occurring genetic disorders have been identified, with the exception of the human (Ostrander *et al.*, 2000). Approximately 58% of inherited disorders of the dog have been recognised as sharing similar clinical features with human disorders, and of these, at least 20% have been shown to result from abnormalities of the homologous gene product in both species (Ostrander *et al.*, 2000). The various breeds collectively demonstrate an immense degree of natural polymorphism yet have maintained the ability to interbreed, and are therefore invaluable in investigating the genetic basis of both physical and behavioural characteristics. Amongst these are relatively simple features including size, coat colour and morphology, and more complex traits such as aggression and hunting instincts. The physiology of these domestic animals is more suited to gross comparison with that of the human than are the smaller rodent models, in which complex traits such as behaviour can less easily be analysed. In turn, the gestation period and litter size of the dog, compared to larger mammals, does not preclude its value in the generation of appropriate pedigrees for meiotic linkage analysis of segregating traits. The detailed pedigree records required for registration of purebred dogs with the Kennel Club can potentially enable the family history of each animal to be traced, whilst veterinary

health care records provide a similarly valuable source of information for the study of genetic disease.

In order to fully exploit the dog as a model system, it is necessary to develop a panel of resources with which genetic analysis may be performed. In particular, this requires the generation of a map of the dog genome that can be used for orientation in more detailed studies of the loci involved in inherited traits. Inevitably, one of the major aims of the dog genome mapping community is the development of techniques for the identification of loci implicated in inherited disease. Diagnostic tests to identify asymptomatic carriers of defective alleles will enable dog-breeders to make informed decisions in the planning of breeding programs that will not cause detriment to the overall quality of the breed but will enable a degree of control over the transmission of disease-causing alleles. In turn, detailed study of inherited traits in the dog will increase our knowledge of homologous traits in other species, in which direct study may not be as easily accomplished. Progress in determining the organisation of the dog genome, and that of other domestic animals and species of economic and experimental value, has been enhanced by the transfer of data, resources and technologies that have developed from the human genome mapping project. Consequently, a review of these issues is required in order to place the current efforts towards a dog genome map in context.

1.2 The human genome mapping project

1.2.1 Techniques for human genome mapping

December 1999 represents a landmark stage in the generation of a map of the human genome, with the first comprehensive report of the nucleotide sequence of a single chromosome (Dunham *et al.*, 1999). The implications of this report are widespread, perhaps most importantly proving that the ultimate aim of obtaining the entire human genome sequence in a rapid and efficient manner is a viable proposition. The data also provide a resource to initiate and extend detailed analysis of general parameters of the genome, such as total gene number and base composition, and the relationship between coding and non-coding regions. The identification of genes involved in the development of inherited traits is likely to be expedited, and other polymorphisms associated with genetic diversity between populations will be more easily identified. Comparative analysis of the genomes of different species will be enhanced by the first detailed description of genes within a single chromosome, for which corresponding regions have already been identified in several other organisms. By a range of comparative methods,

predictions of gene locations in less well-studied species will be advanced by this and further such reports. In turn, our understanding will increase of the modes of chromosome evolution that have occurred since divergence of modern species from their common ancestors.

Complete genome sequences exist for several species to date, including the eukaryotic yeast *Saccharomyces cerevisiae* (Goffeau *et al.*, 1996) and the pathogenic bacterium *Helicobacter pylori* (Tomb *et al.*, 1997). Amongst the most recent to be completed is that of the fruitfly *Drosophila melanogaster*, produced by Celera, a US genomics company (Adams *et al.*, 2000). It is proposed by Celera that a sequence map of the human genome will be available by 2002 (Venter *et al.*, 1998), having produced the *Drosophila* map as a test of the approaches to be used for this far greater task. This commercial venture runs simultaneously with a similar effort by the academic community, which has in turn announced that its attempt will yield an initial draft of the genome by spring 2000. This will account for 90% of the genome, and will be followed by the complete sequence by 2002 to 2003 (Little, 1999). It is perhaps to be expected that some members of the academic mapping community have expressed concern over the increasing commercial interest in genomics. However, it cannot be disputed that the financial and technological advantages held by certain commercial ventures have increased the rate at which new techniques are arising, enabling the mapping process to run at increasingly high efficiency with previously unattainable aims now coming into reach.

1.2.2 History and approaches

Prior to the initiation of the human genome mapping project in 1991, data regarding the structure, function and localisation of genes was limited to a small number of loci, for which detailed analysis had been performed as a result of a specific scientific interest in the gene or its product. The human genome mapping project was initiated to enable a more efficient and comprehensive approach to obtaining information concerning the structure, and ultimately the nucleotide sequence, of the entire genome, which could be used for orientation in subsequent detailed studies. The resulting genetic and physical maps would in turn lead to an efficient large-scale approach for the isolation of genes responsible for traits of interest. The project would also aim to develop resources, technologies and communication networks by which the resulting data could be analysed and stored by the scientific community.

In recent years the emphasis on genome mapping has moved towards the study of coding sequence, which constitutes only a small proportion of the genome of most eukaryotic organisms. It is these regions that are of the most direct significance to the downstream application of most mapping projects, whether they relate to the identification of disease genes or other specific loci of interest, to comparative analysis of genomes of different species, or to evolutionary studies, amongst others. It was therefore proposed that efforts should first be focussed on the localisation and characterisation of human gene sequences in preference to carrying out full-scale sequencing of the entire genome, since the latter would result in coding sequence data becoming lost amongst the overwhelming quantity of anonymous sequence data.

1.2.3 The gene content of the human genome

There is considerable variation in the total number of genes predicted to exist within the human genome. Estimations have traditionally been derived from the number of protein coding sequences from which a polyadenylated messenger RNA (mRNA) is produced. This has been used to predict that the human genome contains between 50,000 and 100,000 genes in total. In a review of methods for estimating gene number, Fields *et al.* (1994) provide a figure of 70,000 by generating data on gene density within highly studied chromosome regions, and then extrapolating this information to the whole genome. Other estimates, using methods such as RNA reassociation kinetics, direct genomic sequencing and comparisons with data from other species, fall as low as only 20,000 genes (reviewed in Fields *et al.*, 1994), whilst the authors themselves cite 60,000 to 70,000 as a more accurate estimate. The majority of sources propose that a total of between 50,000 and 80,000 genes exist in the human genome. Much of the variation reflects uncertainty over the average size of a human gene, and in particular its non-coding component. It is also apparent that some areas of the genome are considerably more gene-rich than others, and so estimations derived from studies of specific regions are likely to result in some variation depending on the relative gene content of the area of interest. For example, data on gene density from the recent sequence map of human chromosome 22 (*Homo sapiens*, HSA) (Dunham *et al.*, 1999) have been extrapolated to suggest that the human genome contains a minimum of 61,000 genes. This assumes that the gene distribution of this chromosome reflects that of the remainder of the genome. HSA 22 is in fact considered to be relatively gene-rich (Deloukas *et al.*, 1998 and others) and so the total gene estimate of the genome was reduced to approximately 45,000 by

Dunham *et al.* (1999). In contrast, a value closer to 140,000 has been proposed by the US genomics organisation Incyte through analysis of their own gene sequence database (http://www.incyte.com/news/1999/genes.html). The derivation of an accurate measure of the number of genes present within the human genome will therefore be dependant on the continuing generation of genome structure data.

1.2.4 Techniques for the mapping of specific genes

There are numerous approaches by which a gene of interest may be identified, and the method selected in each case is largely dependent on the resources available for study. In a positional cloning approach, the location of a gene may be established on the basis of a visible chromosome aberration that is consistently associated with a disease phenotype, or more typically as a result of genetic linkage analysis. In both cases, the locus is identified by virtue of its chromosomal location rather than by its function. Further detailed mapping analysis of the chromosome region implicated may then lead to identification of coding regions within the genomic sequence at this site, one of which may subsequently be shown to represent the locus responsible for the trait under study. Gross chromosome aberrations are typically limited to cancers and developmental abnormalities. In both, deviation from the normal karyotype generally involves changes in ploidy or a loss, gain or translocation of part of one or more chromosomes that enables the abnormality to be detected cytogenetically. Genetic linkage analysis is reliant on the existence of pedigrees in which the phenotype of interest is known to segregate. This method also requires that a polymorphic marker exists in sufficiently close proximity to the locus involved to enable linkage to be identified.

In contrast to the positional cloning approach, the identification of a specific gene may involve the identification of one or more candidate genes that may be responsible for the trait of interest. This would be conducted by searching for the causative gene on the basis of its function and by comparison with the phenotype of the trait in question, in the absence of knowledge of its chromosomal location. The gene may be classed as a candidate on the basis of having previously been demonstrated to function either temporally or spatially in a manner that correlates with the phenotype of interest Alternatively, a candidate may arise by comparison with a similar phenotype in another species for which the locus responsible has been identified. This gene, or its product, can then act as a probe for the isolation of the orthologous gene from any other species. This requires that the gene is sufficiently conserved between species that it performs a

comparable function in each. There must also be sufficient sequence identity to enable a clone from one species to be used successfully as a probe for another.

A third option combines both the above approaches to result in a 'positional candidate gene' method. Here, the locus responsible for a given trait is localised to a chromosome sub-region using one or more of the methods described above. Knowledge of the function of a gene that has previously been localised to that chromosome region then provides a strong candidate for the causative locus, which can be further investigated. Thus information regarding the location and the function of the gene are combined to enable its isolation.

In order for the positional candidate gene approach to be useful, the physical location of coding sequences within the genome must have previously been established. Information concerning the function of the gene product must also be available. Within the last decade, efforts to map and sequence the human genome, and in particular to characterise its gene content, have dramatically increased the availability of such information. Once a trait has been localised to a sub-chromosomal region, a simple database search for loci previously mapped to that site may identify one or more candidate genes whose function correlates with the trait under study.

1.2.5 Strategies for chromosomal localisation of DNA sequences

a) *In situ* hybridisation

The localisation of DNA probes to chromosomes may be achieved using variants of a well-established chromosomal *in situ* hybridisation approach. Typically, the sequence of interest is labelled with a reporter molecule, and allowed to hybridise with the complementary sequence on immobilised chromosome preparations. Non-specific hybridisation is reduced by a series of stringency washes, and the sites of probe hybridisation are then detected by one of a number of methods. Traditionally, radiolabelled probes have been detected by autoradiography, but the long exposure periods required, and the resulting high levels of background signal, mean that this has largely been superseded by fluorescence *in situ* hybridisation (FISH) techniques. Here, the probe is detected by the application of reagents that bind specifically to the reporter, and to which a fluorescent tag is attached. Single locus probes appear as fluorescent sites on pairs of sister chromatids.

Probe labelling is typically performed in a nick-translation reaction, incorporating a modified nucleotide to which the reporter molecule, most commonly biotin or

digoxigenin, is attached. Probe DNA is digested into fragments of less than 1kb with a DNA-specific endonuclease. DNA polymerase I 5' to 3' exonuclease activity is used to excise short regions of nucleotides from the digested fragments, which are then repaired by the same enzyme. The inclusion of a modified nucleotide into the reaction mixture results in the random incorporation of the reporter molecule into the probe during the repair process. Alternatively, modified nucleotides may be incorporated by the polymerase chain reaction (PCR), or a range of other methods. The labelled probe is then resuspended in a suitable hybridisation buffer. Since the probe may contain regions of repetitive sequence, which could create undesirable background hybridisation signals, an excess of unlabelled genomic DNA is typically allowed to pre-anneal to the labelled molecule prior to hybridisation. This saturates repetitive sequences and thus suppresses non-specific binding, such that true signal can be identified without interference from background noise. Following hybridisation of the probe to chromosome material, the presence of the reporter can be detected by incubation of the slide in a solution containing a reporter-binding molecule, to which a fluorescent tag is attached. For example, where biotin is used as the reporter, fluorochrome-conjugated avidin, or the related molecule streptavidin, is used. Both have a high affinity for biotin. Alternatively, antibodies raised to the reporter can be conjugated to a fluorochrome, as is typically the case with digoxigenin. Following stringency washing to remove excess detection reagents, discrete signals at the sites of probe hybridisation are visualised by fluorescence microscopy. The duration of hybridisation is largely probe dependent; however, the sensitivity of FISH analysis is such that data can be generated within 24 hours of labelling the probe, providing a major advantage over radiolabelling techniques. A further advantage is the availability of an increasingly wide range of reporter molecules, and fluorochromes with various excitation and emission characteristics. The use of a reporter is becoming increasingly negated by the development of nucleotides modified by the direct attachment of a fluorochrome, eliminating the need for immunological detection of the probe. Multicolour FISH analysis using combinations of fluorochromes enables a large number of distinct probes to be identified and analysed in a single experiment (for example, Schrock *et al.*, 1996; Speicher *et al.*, 1996).

Most commonly, FISH analysis is performed using metaphase chromosomes as the genomic target. Cultured cells, typically derived from the stimulation of peripheral blood lymphocytes, are arrested at metaphase. This is achieved using reagents such as colcemid, which act by disrupting the formation of the mitotic spindle apparatus required

for cytokinesis. Cells are swollen by incubation in a hypotonic salt solution, pelleted, and resuspended in a fixative solution. Fixed cell suspensions are dropped onto microscope slides from a height that results in the release of the chromosome complement of swollen cells on impact to generate a metaphase chromosome spread. The limit of resolution of FISH analysis using metaphase chromosomes has been estimated as three megabases (3MB) (reviewed in Trask, 1991), although various greater and lesser figures have been reported. More sophisticated methods have increased the potential resolution of FISH analysis to less than 1MB, and as little as tens of kilobases (reviewed in Trask, 1991). Such methods include the analysis of chromosomes in their elongated state at interphase, rather than when highly condensed at metaphase. FISH analysis of interphase nuclei can be highly useful in the ordering of closely linked markers where the relative distance between them is too small to allow their resolution on metaphase chromosomes. Further increase in resolution is achieved using artificially elongated chromosome material, generated by the mechanical extension of chromosomes, or chromatin fibres supported on a microscope slide, in a range of related 'fibre-FISH' techniques (for example, Heiskanen *et al.*, 1994). Such methods can allow resolution of probes separated by less than 10kb (reviewed in Heiskanen *et al.*, 1996).

The chromosome to which a probe hybridises is commonly identified by its unique banding pattern produced by the application of one of many DNA-binding dyes. The variation in internal structure of each chromosome is reflected by differential stain uptake, resulting in a banding pattern characteristic for each chromosome type. The precise physical location of a probe can then be determined in FISH analysis by simultaneous visualisation of the fluorescent signal with banded chromosomes. Staining may be performed either before or after the hybridisation of the probe, and largely depends on the stain used. Traditionally, the most common is the GTG-banding procedure, in which chromosomally-associated proteins are first partially digested with a protease such as trypsin. Chromosomes are then incubated in the presence of the DNA-binding dye Giemsa. Under these conditions, regions that take up the dye are largely AT-rich, and are described as G-bands, or G-positive regions. AT-binding dyes such as Hoechst 33258 or DAPI (4', 6-diamidino-2-phenylindole) create patterns similar to those generated by GTG-banding. Conversely, the application of GC-binding dyes such as Chromomycin A3 results in the production of R-bands, which are essentially the reverse of the G-banded pattern. DAPI-banding can be performed more rapidly than the more traditional GTG-banding method, and also enables simultaneous visualisation of banding

patterns with the fluorescent signals produced in FISH analysis. Consequently, DAPI-banding is increasingly becoming the method of choice, although both this and GTG-banding are widely used in cytogenetic analysis.

The correlation between banding pattern and base composition has been used to identify regions of high gene-density within different genomes. It has been suggested that AT-rich G-bands are relatively gene-poor, and that most coding regions exist within the GC-rich R-bands (reviewed in Craig and Bickmore, 1994). Chromosomes demonstrating extensive regions of R-banding are therefore often regarded as gene-rich. In a review of the status of the human gene map, Antonarakis (1994) described how the mapping of genes to human chromosomes has identified HSA 17 and 19 as containing a higher than average gene density, whilst HSA 13 and 18 are relatively gene-poor. This was found to correlate generally with the relative nucleotide composition of these chromosomes, determined by dual-laser flow cytometry. Gene-rich chromosomes, including HSA 17 and 19, exhibit a high GC-content, whilst HSA 13 and 18 are more AT-rich. Gene content also varies within a given chromosome. This has been demonstrated by Mouchiroud *et al.* (1991) in a survey of the location of previously mapped genes, and by several other similar studies. A detailed physical map of the most GC-rich (and therefore potentially most gene-rich) regions of human chromosomes has recently been described by Saccone *et al.* (1999), although the status of the human gene map is only now beginning to allow the validity of this hypothesis to be investigated in detail.

b) Meiotic linkage analysis

Meiotic linkage analysis is a well-established technique for producing framework genome maps that can be used for orientation in more focussed studies. The main prerequisite is the availability of polymorphic sequences for which inheritance of the various alleles can be traced through generations of reference family individuals. The most common markers used are microsatellites, sequence motifs consisting of a pattern of one to six nucleotides tandemly repeated many times in succession, and flanked by unique sequence. The presence of tandem repeats can result in strand slippage during DNA replication, leading to variation in the number of repeat units between individuals, each representing a different allele. This polymorphism can be utilised to follow the inheritance of the different alleles at meiosis through generations of a family. The frequency of meiotic recombination between two markers is defined by the term 'recombination fraction', or θ. This value can be used to estimate the genetic distance

between two loci, using mapping functions such as that described by Haldane (Haldane, 1919, cited in Cox *et al.*, 1990), in which distance (D) = $-\frac{1}{2}\ln(1-2\theta)$. A maximum value of $\theta = 0.5$ indicates independent segregation, such that the two loci exist on different chromosomes, or lie far apart on the same chromosome. As θ approaches zero, this implies that meiotic recombination between the two loci is infrequent, indicating that they lie in in close proximity to each other, and are described as genetically linked. Genetic map distance is measured in centimorgans (cM), where a distance of 1cM between two loci indicates a 1% rate of recombination between them. The logarithm (base 10) of the likelihood ratio for linkage between two loci is defined as the 'lod score'. A lod score of three or greater, equivalent to odds of 1000:1 against co-inheritance being due to chance alone, is typically regarded as significant evidence that the loci are genetically linked.

Groups of linked microsatellites can in turn be mapped to a chromosomal location by a range of techniques, which are discussed later. Ultimately, consistent co-inheritance of a given phenotype with a particular microsatellite allele suggests that that marker is situated close to the causal gene, narrowing down the search for the locus responsible for the trait. Microsatellite markers have been widely described (for example, Tautz and Renz, 1984; Tautz, 1989; Weber and May, 1989) and have played a major role in the generation of maps of many eukaryotic genomes, exemplified by studies of the human genome such as those of Weissenbach *et al.* (1992), Gyapay *et al.* (1994) and Dib *et al.* (1996). Due to the requirement for sequence conservation at coding (type I) loci, meiotic linkage analysis is of relatively limited value in direct gene mapping, unless polymorphism has been identified in untranslated regions of the gene, or in the case of certain trinucleotide repeats. Most of the polymorphic markers used for meiotic linkage analysis therefore represent anonymous (type II) loci. A small number of microsatellites have been found within gene sequences of various species, including the dystrophin gene of certain primates, rodents and carnivores (Zeiss *et al.*, 1998). Locus-specific microsatellite primers have been applied with relative success between closely related species, such as the Bovidae (Moore *et al.*, 1991), the Equidae (Breen *et al.*, 1994) and the Canidae (Fredholm and Winterø, 1995). However, they have relatively little to offer to comparative genome mapping, since the nature and location of repeats are not highly conserved between divergent species (Stallings *et al.*, 1991).

c) Somatic cell hybrid (SCH) analysis

Somatic cell hybrid (SCH) panels are formed by the artificially-induced fusion of cells from two different species to generate heterokaryons, cells containing two nuclei, one originating from each of the two species involved. For the purpose of clarity, the principle of SCH analysis will be described using the fusion of human and rodent cells as an example, although the technique has been applied to a range of other species. The species under study is commonly referred to as the donor, whilst that which provides the background cell line is known as the recipient, and is typically of rodent origin. As the cultured cells enter mitosis, the nuclear envelope dissociates and thereafter chromosomes from both species co-exist in a single nucleus. In subsequent rounds of cell division, donor chromosomes are lost at random. After several rounds of replication, stability is restored, resulting in the production of a cell line that contains a subset of the chromosome complement of the donor. Although loss of chromosomes from the donor is essentially random, selection procedures can be used to control these events as required. Perhaps the most common selection method involves the use of rodent cells deficient in the thymidine kinase (TK) enzyme, encoded by the TK gene. After the fusion process, hybrids are cultured in medium containing hypoxanthine, aminopterin and thymidine. This 'HAT' medium prevents the survival of TK-deficient cells. Since the TK gene resides on HSA 17, only cells containing this chromosome are able to survive. Thus rodent cells, and hybrid cells from which HSA 17 has been lost, will not propagate, leaving an SCH line that specifically retains this chromosome plus a combination of others. HPRT-deficient cell lines (lacking hypoxanthine phosphoribosyltransferase, encoded by a gene on HSA X) are also unable to survive in the presence of HAT medium. This can be used as an alternative selectable marker to the TK gene. Following fusion and exposure to HAT medium, the only viable cells are hybrids containing the donor chromosome that carries the corresponding selectable marker. Each cell line will also contain a random assortment of other donor chromosomes. The donor chromosome composition of each cell line can be determined by FISH analysis. Metaphase chromosome spreads are generated from the hybrid cell line, and are then screened with a probe that is specific for DNA derived from the donor. This is often achieved using *Alu*-PCR products from the donor species as the probe (Nelson *et al.*, 1989). A single oligonucleotide primer, specific for one end of the *Alu* repeat sequence (described in section 1.3.6), is used in a standard amplification reaction using the PCR. This repeat unit is present at an extremely high copy number within the human genome, and so two such

sequences are frequently found in close proximity. Since the repeats may also lie in opposing orientations, they can be used as sites from which PCR amplification of the intervening sequence can be primed. The intervening sequence is likely to be either unique or to be present a small number of times within the genome, such that the resulting *Alu*-PCR product represents a complex mixture of DNA derived from that species. The *Alu*-PCR product is then labelled and used in standard FISH analysis using metaphase chromosomes from each cell hybrid as the template. The *Alu* repeat is not present in the rodent genome, and therefore no cross-hybridisation is possible in FISH analysis. Sites of probe hybridisation therefore represent regions of donor chromosome present within the rodent recipient cell (Nelson *et al.*, 1989). Alternatively, chromosome-specific single locus probes or paint probes can be used to identify donor chromosome material, provided that cross-hybridisation with the recipient genome does not occur. Once each cell line has been characterised in this manner, combinations of cell lines, each with a different chromosome content, can be used to create a panel of hybrids for physical mapping.

DNA isolated from each cell line in a panel is used as the template in the PCR to determine the chromosome on which a given marker lies. Primers specific for that marker are used in a series of amplification reactions, each with template DNA from a different cell line. After gel electrophoresis, each cell line is scored for the presence or absence of an amplification product. This pattern is then compared to the known human chromosome complement of each cell line. The chromosome on which the marker lies must therefore be present in all cell lines for which a PCR product was obtained, and absent from those from which amplification does not occur. By this method, a marker can be assigned to a single chromosome from the species of interest.

A similar technique involves the generation of SCH lines that contain only a single chromosome from the species of interest. Generation of such monochromosomal hybrids was first described by Fournier and Ruddle (1977). Cells are arrested in mitosis by a mitotic spindle inhibitor. Chromosomes become partitioned into a series of micronuclei, which frequently contain only one chromosome. Micronuclei can be isolated from the cell, and then fused with a recipient cell line in the same manner that somatic cell hybrids are created. The resulting cell lines are therefore typically specific (or at least highly enriched) for a single donor chromosome, enabling easier assignment of markers to a specific chromosome by a simple PCR assay.

d) Radiation hybrid (RH) analysis

A natural progression from SCH mapping was the development of methods by which assignments could be made to a subchromosomal level. The advent of radiation hybrid mapping analysis satisfied this need, and was developed by Cox *et al.* (1990). The methodology is similar to that of the SCH mapping except that prior to the fusion event, donor chromosomes are exposed to a measured, lethal dose of irradiation, typically from an X-ray source. This procedure, first described by Goss and Harris (1975), causes each donor chromosome to be broken into a series of random fragments, which themselves are not able to propagate. Fragments are rescued by fusion with a recipient cell line, typically of rodent origin. A subsequent selection stage, comparable to that used in the generation of SCH panels, ensures that only hybrid cells survive. Donor fragments may be propagated subject to their stable integration into a chromosome from the recipient species. Each cell line generated in this manner carries a different combination of fragments from the donor chromosome, inserted non-contiguously in a random order into recipient cell chromosomes. The donor chromosome content of each cell line can be investigated in a similar manner to that described previously for SCH panels. The entire group of cell lines constitutes the radiation hybrid panel. It is common for the panel to consist of a number of cell lines that, in addition to suitable control samples (typically donor and recipient genomic DNA), can be analysed in a single, 96-well plate format. Analysis of a marker by radiation hybrid mapping is essentially the same as that for SCH mapping. A PCR assay is performed by the application of marker-specific primers to DNA isolated from each cell line within the panel. Each cell line is scored for the presence or absence of an amplification product, indicating the subset of cell lines within the panel that have retained the fragment of chromosome on which the marker lies.

Initially, RH mapping was performed using monochromosomal somatic cell hybrids as the source of donor cells. This technique was initially described by Cox *et al.* (1990) for the production of a relatively high-resolution map of a 20MB region of HSA 21, consisting of 14 markers. The use of a monochromosomal hybrid as the starting material requires an entire panel of radiation hybrids for the mapping of any given donor chromosome. Mapping an entire genome by this method would therefore require the analysis of an impracticably large number of cell lines. More recently, the irradiation of intact donor cells containing a full complement of donor chromosomes, and subsequent fusion with a rodent recipient, has enabled the procedure to be applied to the mapping of an entire genome with a single panel of radiation hybrids. This whole genome radiation

hybrid (WG-RH) method was first described by Walter *et al.* (1994). Irradiation-induced breakage of a human fibroblast cell line, and subsequent fusion with a rodent recipient, was used to generate a panel of 44 hybrid cell lines that was used to generate a preliminary map of HSA 14. Chromosome spreads were made for selected hybrid cell lines, and *in situ* hybridisation analysis was used to identify the human component of each. This demonstrated the presence of between five and eleven human chromosome fragments in each cell line examined. Their organisation fell into two categories. In some cell lines, the human-derived DNA fragments had become integrated into rodent chromosomes. In others, fragments derived from the donor existed as distinct 'new' chromosomes, which were assumed to be the result of the association of donor fragments containing telomeric sequence with others carrying centromeres. The panel was tested with randomly selected markers that indicated an average retention frequency of 20.9% for any given fragment. The retention frequency refers to the number of cell lines analysed that contain the fragment of interest, and is typically expressed as a percentage. Subsequently, 40 HSA 14 markers were mapped using the panel, demonstrating an overall average retention frequency of 23.7%.

Cox (1992) showed that the retention frequency of individual chromosome fragments within an RH panel was relatively constant, regardless of their location within the chromosome, although some variation in the retention frequency of a small number of markers was observed. Others have demonstrated an increased retention frequency for markers originating near the centromere of the donor chromosome (for example, Benham *et al.*, 1989; Ceccherini *et al.*, 1992; Gorski *et al.*, 1992). This effect has been proposed to reflect the increased stability of fragments that retain a centromere-like structure (Walter *et al.*, 1994). A similar effect at the telomere has also been reported (for example, Ceccherini *et al.*, 1992). Donor fragments may be maintained as separate structures in hybrid cells, rather than being incorporated into recipient chromosomes. Therefore, fragments containing centromeric or telomeric regions are likely to demonstrate a higher than average retention frequency since these structures are necessary for propagation of donor chromosome fragments that are not incorporated into recipient chromosomes (discussed in Jones, 1996). It has been suggested that the effects on retention of centromeric and telomeric markers may be a species-specific issue (McCarthy and Soderland, 1998), although the limited number of comprehensive RH maps for non-human chromosomes makes this hypothesis difficult to assess at present. Retention of

fragments at a higher than average level also applies at the site of the selectable marker used in library construction, such as the TK locus (for example, Foster *et al.,* 1996).

Unlike the standard SCH mapping technique, radiation hybrid mapping requires complex statistical analysis. The basis of such analyses relates to the fact that two loci that exist in close proximity on a given donor chromosome are less likely to be separated onto different fragments after irradiation than two loci that lie far apart on the same chromosome, or that lie on different chromosomes. Thus, the pattern of retention of two markers that are close together will be significantly more similar than that of two markers which lie sufficiently far apart that a radiation-induced breakage separates them onto different fragments. However, since each cell line may contain multiple chromosome fragments, positive amplification of two markers within a single cell line can result from either of two possibilities. The markers may be closely linked on the donor chromosome such that the fragment on which they reside remains intact within that cell line. Alternatively, there may be retention of two separate fragments in the cell line, with one marker localised on each, such that the markers are not closely linked. Therefore, co-amplification of two markers in a given cell line does not necessarily imply that they are located in close proximity. Instead, a statistical analysis of the frequency of breakage between two markers is performed using DNA extracted from all cell lines within the RH panel.

Radiation hybrid mapping analysis is in many ways comparable to the procedure used for meiotic linkage analysis of marker alleles through a pedigree. Rather than separation of markers occurring through meiotic recombination, in RH mapping separation occurs as a result of chromosome fragmentation through X-irradiation. The major assumptions of the statistical analysis are that the frequency of breakage between markers, and their individual frequency of retention, are independent of each other, and also that the retention of any chromosome fragment is independent of any other fragment. As with the measurement of the recombination frequency in meiotic linkage analysis, the frequency of breakage between two markers in the RH mapping approach is typically defined as θ, but in the latter technique, the value of θ varies from 0 to 1. Where $\theta = 0$, this indicates that the markers are always retained on a single fragment, that is, their retention patterns are identical. Conversely, a value of $\theta = 1$ is found where markers are sufficiently far apart that they are apparently never retained within the same chromosome fragment, and are classed as unlinked. The distance between markers is described by the formula

$D = -\ln(1-\theta)$, which is highly comparable to the mapping function used in meiotic linkage analysis, described in section 1.2.5b (Haldane, 1919, cited in Cox *et al.*, 1990). The likelihood of two markers being separated onto different fragments by X irradiation is dependent on the dose of radiation given, which is measured in rads. This value in turn influences the estimated distance between two markers, which is measured in centiRays (cR), similar to the centiMorgans (cM) used in meiotic linkage mapping. Computer-based methods have been developed for the pair-wise analysis of multiple markers, again with strong similarities to those used in meiotic linkage analysis. Typically, a pairwise analysis is performed first, in which retention patterns of each marker are compared against those for all other markers to identify those which are linked. Data for each linked pair are then compared to result in a proposed likely order for all markers along the chromosome of interest. A given order is typically considered significant where it is calculated as at least 1000 times more likely than the next most likely order, reflecting a lod score of three.

Amongst the most popular software for analysis of RH data is the 'RHMAP' package devised by Boehnke *et al.* (1991) and further developed by Lange *et al.* (1995). Typically, the 'RH2PT' component of the program is used to identify groups of linked markers by pairwise analysis of retention patterns. Each marker within a given linkage group must be linked to at least one other member at a level of significance that can be user-defined. Preliminary ordering of markers within linkage groups can then be investigated using the 'RHMINBRK' program. This calculates the minimum number of chromosome breakages that must have occurred in order to generate the relative retention patterns between groups of linked markers. It is assumed that the closer two markers lie, the fewer chromosome breakages will have occurred between them. RHMINBRK therefore identifies the locus order for which the smallest number of obligate chromosome breaks is required. Typically, several possible orders are reported, the most likely of which can be further investigated using the third component of RHMAP, 'RHMAXLIK'. This provides an opportunity to perform analysis using one of several models of retention, of which the most simple assumes equal retention of all markers, whilst others allow for increased retention at centromeres and telomeres. An estimation of distance between linked markers can also be derived using RHMAXLIK (Boehnke *et al.*, 1991).

WG-RH offers a number of advantages that make it an attractive method for mapping studies. One of the major advantages of all the above methods over meiotic linkage

mapping is the ability to map non-polymorphic markers. This therefore widens the spectrum of markers that can be analysed. In addition, the dose of irradiation used in construction of the WG-RH panel can be varied to reflect the level of resolution required for a given mapping project, by determining the frequency of chromosome breakage and in turn the size of the average donor chromosome fragment. SCH and WG-RH methods also negate the need for suitable pedigrees in which to study inheritance patterns, which can often be a limiting factor in the success of meiotic mapping studies.

Several limitations also exist. The main requirement common to SCH and RH analysis is the ability of marker primers to selectively amplify by the PCR from the donor without co-amplification of the recipient background. Where co-amplification does occur, accurate scoring may not be possible, unless the rodent product can be unequivocally distinguished from that of the donor. In addition, the PCR assay performed for a given marker across a panel of hybrids must be carried out with a high degree of accuracy. Erroneous results arising from cross-contamination or errors in plating out reagents will clearly affect the analysis. To limit these effects, assays are typically carried out in duplicate, and retention patterns compared to ensure agreement between each. Where patterns differ, that hybrid would be scored as 'ambiguous' and is excluded from subsequent analysis.

1.3 Expressed sequence tags as tools for genome mapping

1.3.1 Generation of expressed sequence tags for large-scale gene mapping

Whilst positional cloning and candidate gene approaches are vital for the isolation and characterisation of genes involved in specific traits of interest, they are not an efficient manner by which to study gene distribution on a genome-wide basis. Consequently there has been a need to develop methods by which sequencing and mapping projects can be focussed on coding regions. The study of mRNA represents an obvious technique by which coding sequences can be analysed in isolation from the remainder of the genome. The labile nature of mRNA generally necessitates its conversion to the more stable complementary DNA (cDNA) prior to study. Particularly within the last decade, analysis of cDNA clones has been widely used for the large-scale characterisation and mapping of gene sequences. Short regions of single-pass sequence data have been obtained from many thousands of cloned cDNA sequences isolated from libraries made from an enormous range of tissues, from many different species. The sequence data can potentially be used to identify the nature of the coding sequence isolated, by comparison

with existing database sequences for previously annotated genes, or to study the location or frequency of expression of a given gene in different tissue sources. The data also provide a resource by which the coding sequence can be mapped to its chromosomal location. This approach therefore generates sequence tagged sites that are specific for transcribed regions of the genome, hence they have become known as expressed sequence tags, or ESTs. EST generation has resulted in a dramatic increase in the density of mapped genes within the human genome, and also those of many other species.

1.3.2 cDNA production and cloning

Access to a high quality source of cDNA is a prerequisite for successful EST generation. The initial stage of cDNA library generation involves the isolation of RNA from the required tissue. This must be performed by a method that limits exposure to and degradation by ribonucleases, present both in the cell and in the external environment. Typically, cells are lysed with a detergent to release RNA. Genomic DNA can be eliminated using DNA-specific endonucleases. Either entire cellular RNA can be isolated, or specifically that fraction present within the cytoplasm, which contains only fully processed transcripts. The proportion of this total RNA that contains a 3' polyA tail can be isolated using an oligo d(T) sequence immobilised on a solid surface. This, or total RNA, is then used to create a heteroduplex of RNA and single stranded complementary DNA (cDNA) by means of a reverse transcriptase enzyme. This may utilise an oligo d(T) sequence to prime cDNA synthesis, which anneals to the poly(A) tail at the 3' end of the transcript. Alternatively, a mixture of oligonucleotide sequences may be used, which prime the reaction from random sites within the transcript. The RNA is then degraded with an RNA endonuclease, and the cDNA is made double stranded with DNA polymerase. Linker sequences may be attached to the cDNA ends to create restriction sites to facilitate cloning of the product into a suitable vector. Various improvements have been made to this process to increase the ease and efficiency of cDNA library production; however, the fundamental principle involved, based predominantly on the method described by Gubler and Hoffman (1983), remains the same.

1.3.3 Temporal-spatial issues in cDNA library construction

The precise method of cDNA library construction frequently reflects the fact that gene expression, and by association, mRNA levels, vary both temporally and spatially for a

given organism. For example, the brain is often selected as the source of RNA for EST generation since it is a highly complex and active organ that is reliant on the transcription and translation of coding sequence to maintain its function as the control centre of the body. Up to 30,000 human genes are thought to be expressed in the brain (Sutcliffe, 1988), representing a considerable proportion of the total number of genes encoded by the human genome. As a consequence, studies on the hybridisation kinetics of RNA populations have shown that the level of RNA expression that occurs in the brain is consistently two to three-fold greater than that in such tissues as the liver and kidney (reviewed in Sutcliffe, 1988). The overall rate of transcription is increased at certain timepoints within the lifespan of an individual, such as during *in utero* development and puberty. The study of cDNA derived from early embryonic tissue may therefore enable the identification of genes involved in developmental defects that may not be present, or at least not as abundant, in adult tissue. Thus the isolation of RNA from a specific tissue and/or at a specific timepoint is a key consideration in the construction of a cDNA library for EST analysis.

1.3.4 Database similarity searches

Once cDNA has been cloned and partial sequence obtained, these data can be used to attempt to identify the nature of the transcript represented by the clone. Typically this involves using the cDNA sequence as a template for performing a search against data for previously annotated genes and other sequences that have been deposited into one of the many sequence databases available to the mapping community. Similarity to highly represented repetitive elements can be identified, or to gene sequences from other species. Where no match is found, the sequence may prove to represent a novel gene. Database similarity searches are therefore vital tools for the preliminary characterisation of sequence information prior to downstream study. The database searched largely depends on the nature of the query sequence. Those most commonly used are hosted by the National Center for Biotechnology Information (NCBI), including the GenBank database (http://www.ncbi.nlm.nih.gov/Genbank/index.html), and others designed for more refined searches.

BLAST (Basic Local Alignment Search Tool, Altschul *et al.*, 1990) is the most commonly used electronic resource for the comparison of a query sequence with existing database entries in order to assign an identity. BLAST uses a rapid database-searching facility to identify small regions of similarity between the query and those sequences

present within the database selected for comparison. The alignment is extended further in both directions depending on the criteria selected as representing sufficient similarity to still constitute a significant match between them. Matches generate a positive score, whilst mismatches receive a penalty. Extension of the alignment continues until the score falls below a specified value. The total score is then reported, representing the optimum score generated without further extension. This indicates the overall degree of similarity between the two sequences, a higher score indicating a greater degree of nucleotide match. The percentage nucleotide identity observed between the two sequences is reported, in addition to the distance over which the similarity exists. The probability that the similarity observed is due to chance alone is indicated by the 'P-value'. This value ranges from zero to one, with scores closer to zero representing increasingly significant matches between aligned sequences. The P-value is dependent on numerous factors, including the quantity of sequence data existing within the database at the time, and the overall size of the query sequence. The P-value that is classed as demonstrating significant similarity therefore varies widely between studies, with the maximum typically $P \leq 0.01$ (Adams *et al.*, 1991; Khan *et al.*, 1992, and others), although more commonly $P \leq 1x\ e^{-10}$ is used (for example, Hwang *et al.*, 1995; Frigerio *et al.*, 1995 and others), where $1x\ e^{-10}$ represents ten to the minus 10th power. Parameters can be user-defined and are typically chosen empirically, largely depending on the nature of the study. Subsequent improvements were made in the generation of an updated version of BLAST that is also significantly quicker to use without detriment to the sensitivity of the search and the consequent quality of the output (Altschul *et al.*, 1997). This is important when one considers the dramatic increase in the size of sequence databases with which such searches are performed. The main advantage of this version of BLAST is that it allows the generation of an artificial gap in an alignment between two sequences, which previously had reduced the reported significance of a match between two sequences to the extent that they could be overlooked. Such gaps may represent an insertion or deletion event in one of the sequences. This more recent version of BLAST ('GAPPED-BLAST') (Altschul *et al.*, 1997) requires the identification of two short regions of similarity between two sequences from which to extend an alignment, rather than one as was used in the original version. The rationale behind this was that sequences with a common functional and/or evolutionary origin are likely to share more than one of these short regions of highly similar nucleotide composition. Extension of the alignment occurs subject to the identification of further such regions either side of the initial match. The

overall time taken to perform a search is reduced by eliminating the need to perform the initial alignment extension from the short region of similarity shared by both sequences, since in the newer version two such regions are initially identified (reviewed in Goodman, 1997). The rationale behind 'GAPPED-BLAST' (Altschul *et al.*, 1997) has particular relevance in EST generation. It allows compensation for the presence of an intron in a genomic DNA sequence that is absent in the corresponding cDNA sequence with which comparison is being made. The identification of a cDNA sequence by comparison with existing genomic DNA sequences, from the same or other species, can therefore be established more accurately than with the earlier version of BLAST.

1.3.5 Initial reports of EST mapping data

EST mapping stemmed from a proposal by Wilcox *et al.* (1991) to generate sequence data from the 3' untranslated region (UTR) of human genes, using cloned cDNA as the template. Primers designed within this 3' UTR, which were specific for that locus, could be used to map the marker by techniques such as SCH or RH mapping. Wilcox *et al.* (1991) cite three reasons for their focus on 3' UTR sequences. Firstly, such regions are known to be generally devoid of introns. Consequently, both cDNA sequence and the gene sequence proper are effectively identical within this region. The primers generated from the cDNA sequence data can be used to treat the locus as a standard sequence tagged site, since there is minimal risk of mismatch where intron sequence is present within the gene but absent in the corresponding cDNA. The size of amplification products derived from the application of such primers to both templates would therefore also be the same. Secondly, the sequence at the 3' UTR is typically less well conserved compared to the remainder of the average gene. This factor can be used to enable distinction between members of a gene family whose nucleotide sequence may be highly similar within the coding region, but which are more likely to differ within the 3' UTR. Thirdly, 3' UTRs are reported to demonstrate significant sequence divergence between human and rodent homologues of the same gene. This is essential if sequence data from this region is to be used to design PCR primers specific for the former, for use in SCH or RH mapping analysis. As an example, sequence comparisons of 15 randomly selected human and rodent cDNAs demonstrated approximately 88.5% identity within the protein-coding region, but only 71.4% identity within the 3' UTR (Wilcox *et al.*, 1991). PCR primers were designed within the 3' UTR of seven human genes without

reference to the corresponding rodent sequence. Each of these markers was successfully mapped using a SCH panel, without interference from the rodent background.

Initially, gene assignments were limited to the whole-chromosome level. Adams *et al.* (1991) reported the production of 609 human ESTs derived from randomly selected cDNA clones. Amongst these, a minimum of 337 ESTs were proposed to represent novel human genes, including 48 sharing significant similarity with genes from other species. Forty six markers were successfully mapped to chromosomes by somatic cell hybrid analysis. Unlike the study of Wilcox *et al.* (1991), sequence data in this project were derived not from the 3' UTR of oligo(dT)-primed clones, but from randomly-primed, partial cDNAs. That is, the focus lay in the analysis of the 'internal' protein coding region itself. This was proposed to allow improved database similarity searching, since alignments could also be made with existing protein sequence entries, from which 3' UTR sequence has been eliminated during mRNA processing. This project was designed also as a pilot study to assess the efficiency and quality of various parameters of the EST production process, such as the composition of a typical commercially produced cDNA library, and the accuracy of a routine sequencing analysis run. Thirty two percent of the 609 sequences analysed matched previously annotated human sequences, including repetitive elements, mitochondrial genes, ribosomal RNA genes, and nuclear genes. Eight percent of ESTs demonstrated significant similarity to existing entries from other species, whilst 38% did not share significant similarity to any existing entry. The remaining 22% either contained no insert or consisted solely of polyadenylate tracts.

This study was extended to the generation of more than 2000 additional human brain ESTs the following year (Adams *et al.*, 1992a) and almost 3500 a year later (Adams *et al.*, 1993a). This was closely followed by the report of more than 1500 3' and 5' ESTs generated from a custom-made human infant brain cDNA library (Adams *et al.*, 1993b). This study importantly incorporated several features intended to improve the efficiency of EST generation. These included directional cloning of the cDNA library, size selection of clones prior to sequencing, and prescreening to identify mitochondrial and ribosomal transcripts, as well as polyadenylate tracts and repetitive elements. This reflected the general finding that many commercial libraries are limited in their use for EST generation by the presence of excessive numbers of repetitive sequences that are not themselves of direct value in the mapping of unique coding sequences.

1.3.6 Repetitive elements and cDNA libraries

Many repetitive elements are distributed throughout the vertebrate genome, and estimates have shown that in excess of 35% of the human genome consists of repetitive sequence (reviewed in Smit, 1996). Two repeat elements have been particularly widely documented, described as the short and long interspersed nuclear elements (SINEs and LINEs) on the basis of the size of the repeat unit. Both SINEs and LINEs are classed as transposable elements, since they are amongst those sequences able to migrate to different locations within the genome. In the case of SINEs and LINEs, this occurs via an RNA intermediate. The SINE class is exemplified by the *Alu* repeat element family, which comprises approximately 10% of the human genome, and which has counterparts in the genomes of several other mammals (reviewed in Smit, 1996). Although rare in protein coding regions, the *Alu* sequence is known to exist within the noncoding regions of many genes, mainly within introns, but also within the flanking untranslated regions (Adams *et al.*, 1995, and others). Consequently, they may be present in the primary transcript of a gene, and in turn may be represented within cDNA sequences. The 300bp *Alu* consensus sequence contains a promoter region which enables its transcription, and a poly(A)-like sequence. The precise method of transposition is not fully understood, but may involve reverse transcription mediated by a product of the LINE element (reviewed in Boeke, 1997). The LINE class includes the LINE-1 (L1) element, which represents in excess of 14% of the human genome, and is also found in other mammalian genomes (Smit, 1996). The full-length L1 consensus sequence is 6.4kb in length and contains two open reading frames (ORFs). ORF1 encodes a protein of unknown function, whilst the product of ORF2 shares similarity with some reverse transcriptases. L1 also contains a 5' promoter region and a 3' poly(A) tail, and in common with the *Alu* repeat closely resembles a typical transcript from a coding sequence. L1 is also sometimes found within the non-coding regions of genes, and may be present within the primary transcript, being subsequently eliminated in the formation of the mature messenger RNA. L1 is classed as a retrotransposon, since the presence of the reverse transcriptase homologue allows the RNA intermediate to be converted to cDNA within the cell, which may then become re-introduced into a different chromosomal location (reviewed in Fanning and Singer, 1987). As both L1 and *Alu* contain polyadenylate tracts, and can be associated with coding sequences, these repeat elements are frequently identified within cDNA libraries. Since both are widely dispersed throughout the human genome, their repeated isolation

from randomly selected cDNA clones can have a detrimental effect on the efficiency of EST generation.

1.3.7 Normalisation and subtraction of cDNA libraries

In addition to repetitive elements, the issue of sequencing redundancy of clones corresponding to highly abundant gene transcripts has been addressed in order to improve the efficiency of the EST generation method. Such clones may represent genes that are expressed at high levels in the tissue used in library generation, whether specific to that tissue, or generally in the form of housekeeping genes. Reassociation kinetics have demonstrated that transcripts from a typical cell fall into three classes relating to their level of frequency (Bishop *et al.*, 1974). The most highly abundant class consists of less than 20 different mRNAs that together constitute up to 20% of the total mass of mRNA in that cell. The second, intermediate class comprises between 1000 and 2000 different mRNA types, representing between 40% and 45% of the total mRNA mass. Up to 20,000 different transcripts are represented in the third, complex class, and also form up to 45% of the final mRNA mass of the cell (Bishop *et al.*, 1974). Redundancy thus approaches 60% once most transcripts from within the first two classes have been isolated. To circumvent this, strategies have been implemented that enable distinction between highly abundant transcripts and those present at lower levels. Wilcox *et al.* (1991) and Khan *et al.* (1992) have shown that sequencing redundancy can be reduced by prescreening a partial cDNA library with a total cDNA probe, and selecting for further study only those clones for which no probe hybridisation was detected. This was shown to reduce the overall level of redundancy sevenfold compared to an unscreened library (Khan *et al.*, 1992), limiting the repeated isolation of highly abundant transcripts. In addition, prescreening was found to reduce the proportion of ribosomal and mitochondrial sequences present. More sophisticated 'normalisation' methods have been widely used and developed for this application by Soares *et al.* (1994), amongst others. These include the co-hybridisation of total cDNA and genomic DNA, and subsequent selection of bound cDNA. Individual cDNA clones are therefore adjusted to a level proportional to the abundance of the corresponding gene from genomic DNA. Approaches applied by Ko (1990) and Patanjali *et al.* (1991) relate to the reassociation kinetics of cDNA, in which rarer transcripts are found to reassociate more slowly than more abundant sequences. As reassociation proceeds, the remaining single-stranded cDNA becomes increasingly normalised as highly represented transcripts pair, and is

subsequently isolated for cloning. In a similar procedure described by Soares *et al.* (1994), it was estimated that the relative levels of all transcripts present were brought within one order of magnitude. Libraries constructed by this and similar methods (reviewed in de Fatima Bonaldo *et al.*, 1996) have been widely used for the generation of ESTs by the I.M.A.G.E. consortium (Lennon *et al.*, 1996). The I.M.A.G.E. (Integrated Molecular Analysis of Genomes and their Expression) consortium comprises human genome researchers from groups in the USA and France whose focus lies in cDNA analysis, and who have jointly made a major contribution to the development of human ESTs.

As more human genes have become identified and characterised by the EST approach, the repeated isolation of transcripts common to many tissues has also become a limiting factor in the efficiency of this method. This has been reduced to a great extent by the development of subtraction methods by which sequences common to transcript populations isolated from two different tissues can be excluded prior to library construction (for example, Diatchenko *et al.*, 1996). cDNA from the tissue of interest (the 'tester' population) is hybridised to an excess of cDNA or mRNA from another source (the 'driver'). Those 'tester' transcripts that remain unbound, due to the absence of a counterpart within the 'driver' population, are isolated and cloned. The subtraction method therefore results in a resource enriched for sequences specific to the tissue of interest, enabling a focussed analysis of such transcripts whilst limiting duplication of existing sequence data.

1.3.8 Chromosomal assignment of ESTs

As well as the need to improve the quality of cDNA resources for EST generation, the issue of mapping techniques has also been addressed. The value of each EST can be markedly increased if the corresponding gene can be assigned to a chromosomal location. This is essential for the successful application of positional candidate gene methods, and for the continued identification of regions of high gene density within a genome.

a) Meiotic linkage analysis

Although efforts have been made to eliminate repetitive elements from cDNA libraries, some have exploited their presence for the assignment of sequences to chromosomes. The possibility of generating ESTs containing microsatellites has been discussed and investigated by Khan *et al.* (1992) and others. These would simultaneously provide a

linked genetic marker for a specific gene, enabling coding loci to be directly incorporated into the extensive linkage map comprising anonymous markers. Khan *et al.* (1992) demonstrated that the screening of a human brain cDNA library with a radiolabelled dinucleotide repeat probe identified approximately 6% unique repeat-positive clones after suppression of highly-abundant repetitive sequences. Two clones were selected and mapped both physically and genetically to the same chromosome in both cases, each having been shown to be sufficiently polymorphic to permit genotyping analysis. Similarly, Polymeropoulos *et al.* (1993) reported that approximately 1.5% of clones within a human retinal cDNA library demonstrated hybridisation with a radiolabelled dinucleotide repeat probe, suggesting the presence of a microsatellite repeat. Six polymorphic cDNAs were successfully mapped by genetic linkage analysis.

The possibility of extending this type of analysis to the investigation of polymorphic trinucleotide repeats within coding sequences was also briefly discussed by Polymeropoulos *et al.* (1993), extending previous discussions by Levitt (1991) on the exploitation of 3'UTR polymorphisms for gene mapping. Expansion of exonic trinucleotide repeats is implicated in several human inherited disorders, and so the identification of such repeats within cDNAs generates a useful resource for the identification and study of the loci involved (reviewed in Riggins *et al.*, 1997).

b) Cell hybrid analysis

Meiotic linkage analysis of polymorphic ESTs is clearly limited since only a subset of coding sequences are associated with simple sequence repeats. An alternative involves the use of hybrid cell lines. This was the approach used by Wilcox *et al.* (1991) in their pilot study that initiated large-scale EST generation. Much of the early work on assignment of ESTs stemmed from the reports of Polymeropoulos *et al.* (1992) and Polymeropoulos *et al.* (1993). In the latter, 3' UTR sequences were used to design PCR primers for mapping with a somatic cell hybrid panel. The assignment of more than 300 randomly selected ESTs to human chromosomes demonstrated their non-random distribution amongst those chromosomes with the highest GC content. These data, in combination with a review of previously assigned genes, demonstrated that HSA 13 contains far fewer genes that would be predicted by its length. HSA 19 was proposed to contain approximately the same number of genes as HSA 2, which is almost four times larger than HSA 19 (Polymeropoulos *et al.*, 1993).

Further mapping was performed by Durkin *et al.* (1994), who mapped more than 60 human brain cDNAs by somatic cell hybrid (SCH) analysis. The study was aimed at increasing the efficiency of the method by developing automated, fluorescent-based scoring of multiplexed markers. This enabled several markers to be analysed in a single amplification reaction, using combinations of fluorescently tagged PCR primers designed from several different cDNA clones.

The development of WG-RH mapping provided an improved method by which ESTs could be assigned to chromosomes. Whereas the somatic cell hybrid panels used in previous studies contained fragments of only a single chromosome, the advent of WG-RH mapping allowed the sub-chromosomal assignment of markers from any chromosome with the use of a single panel of hybrids. Hayes *et al.* (1996) discussed the application of both SCH and WG-RH techniques for the mapping of 45 ESTs, and the potential for integration of these data with existing genetic linkage markers present throughout the genome. Hudson *et al.* (1995) integrated more than 3,000 human ESTs with over 15,000 further markers using a combination of genetic linkage and WG-RH mapping. This represented the first large-scale approach to the incorporation of EST sequences within a framework of anonymous STS markers, towards a transcript map covering the entire human genome.

c) *In situ* hybridisation

The assignment of EST sequences by direct *in situ* hybridisation methods has also been investigated, although to a lesser extent than the SCH and RH techniques. In an attempt to integrate EST data with existing physical and genetic mapping information, Korenburg *et al.* (1995) mapped 41 cDNA clones to human chromosomes by FISH analysis, where clones were an average of less than 2kb in length. Assignments were considered successful where signal was observed at a single site on at least one chromosome in a minimum of 10% of 200 metaphase spreads. 'Preliminary' and 'possible' assignments were classed as those where signal was present in 5% - 10% and 2% - 5% of cells studied, respectively. Mapping of cDNAs required the analysis of a considerably greater number of metaphase spreads than is required for the assignment of typical genomic probes. The time required to achieve this would be an important consideration if this method were to be applied routinely. The studies of Korenburg *et al.* (1995) and others indicate that the frequency of observed signals is also far lower than is found for large insert genomic DNA probes such as yeast- and bacterial artificial chromosomes (YACs

and BACs) and cosmids, for which conventional FISH analysis is highly suited. The success of FISH analysis for cDNA mapping is also likely to depend on the sensitivity of the equipment available. Although the cytogenetic data presented by Korenburg *et al.* (1995) would appear conclusive, the assignment of only one of the 14 clones mapped can be correlated with that subsequently reported within the Human Gene Nomenclature Database (http://www.gene.ucl.ac.uk/nomenclature/). Of the remaining sequences, five were mapped to different chromosomes by Korenburg *et al.* (1995) compared to the subsequent database entry, and no database information was available for the remaining eight loci. In contrast, complete correlation was observed for 14 loci mapped by Korenburg *et al.* (1995) for which others had previously made an assignment. The difficulties associated with the assignment of probes which are members of gene families is also highlighted by Korenburg *et al.* (1995), although it is suggested that accurate assignments are possible even where greater than 75% of nucleotides are identical between family members. The authors suggest the ideal study would combine FISH analysis with other physical mapping methodologies to increase the overall level of accuracy. More recently, Jay *et al.* (1997) attempted to map 55 cDNA clones, with inserts of greater than 1kb, of which 36 could be unambiguously assigned to a single human chromosome location. Of those that failed, ten were found to contain repetitive elements that resulted in a high level of background hybridisation that was likely to have masked any true signal, whilst four clones did not generate any signal at all. It must be considered that the size of the cDNA clone insert is only part of the issue. The extent of the corresponding chromosomal gene sequence, complete with its intron component, will also be a factor in determining the likely success of the probe for FISH analysis. For example, a partial cDNA sequence of 1kb may correspond directly to a single exon of a gene, such that both probe and chromosomal target are of an equal size. Alternatively, if the 1kb cDNA spans two exons, the chromosomal target could be considerably larger, and will depend on the size of the intervening intron. This will influence the length of chromosome over which any potential signal will extend; the larger the intron, the greater the extent. However, the presence of a large intron in the gene sequence that is absent in the cDNA probe will adversely affect the binding capacity of probe and target, such that sufficient annealing of the two may not occur for hybridisation to be detectable. Therefore *in situ* hybridisation analysis of cDNA clones is largely a clone-specific issue.

1.3.9 Development of a database for EST sequences

The rapidly increasing volume of partial cDNA sequences generated and deposited into GenBank and similar resources led to the construction of a new database specifically for EST data (Boguski *et al.*, 1993). This in part reflected recognition of the significant difference between ESTs and the more 'traditional' genomic DNA and protein data that had previously formed the vast majority of the content of sequence databases such as GenBank. EST sequences, by their nature, typically represent data from only single pass analysis and thus the inherent error rate is higher than data for more highly characterised sequences. Typically, the minimum criteria for submission to nucleotide databases require that an EST contain less than 3% unknown bases over at least 150bp of sequence data, following the guidelines proposed by Adams *et al.* (1991). The average EST entry is therefore short, and typically is less well annotated than sequence data from previously characterised gene or protein sequences. The rate of growth of GenBank had been dramatically altered by the advent of EST analysis. For example, it was estimated that between 1991 and 1993 over 14,000 new ESTs had been deposited, compared to 19,000 human genomic DNA sequences over the previous decade in total (Boguski *et al.*, 1993). GenBank had come to represent a mixture of genomic DNA and cDNA sequence, both partial and complete, processed and unprocessed. Thus it was decided to create 'dbEST', a database specifically for these single pass partial cDNA sequences (Boguski *et al.*, 1993). This would aid the continuing evaluation of the quality and content of the entries, and facilitate the task of repeated database similarity screening to attempt to derive an identity for each EST. The EST database would be kept constantly evolving in order to take account of all new sequencing data being performed. Genomic DNA, cDNA and RNA sequence derived by more 'traditional' studies would remain within the main GenBank database.

1.3.10 Large-scale and commercial EST mapping efforts

EST generation became a proprietary interest with the launch of a large scale sequencing project at The Institute for Genomic Research, TIGR, (and formerly at the National Institute of Health, NIH). Data began to be diverted into privately administered databases. The high throughput nature of EST generation, with its need for considerable financial and technological support, was felt by some to be a task ideally suited to an industrial environment. In time, the opposing views of academic and industrial researchers became apparent. The former largely believed that free and equal

contribution of all data was required to achieve the goal of the human genome mapping community, whilst the latter maintained their right to exclude their data from the public domain in order to protect their own commercial interests.

However, by the beginning of 1995, 1500 human cDNA sequences were being deposited daily into the public sector databases by a collaboration between several groups in the USA (Boguski and Schuler, 1995; Williamson *et al.* 1995, cited in Aaronson *et al.* 1996). This consisted predominantly of members of the Washington University Sequencing Centre, the National Centre for Biotechnology Information and of the I.M.A.G.E consortium (Lennon *et al.*, 1996), in association with Merck & Co. In a single report by Hillier *et al.* (1996), over 280,000 human ESTs were described, representing 22 libraries generated from 12 tissues. The subsequent mapping of markers derived from this readily available source of sequence data advanced the project to a higher level. The goal of the 'WashU-Merck' project was set at the production of up to 400,000 ESTs by the end of March 1996, to add to the 65,000 such sequences previously entered into the public domain by other groups. These were to represent a combination of both 3' and 5' sequence data from 200,000 oligo(dT)-primed, directionally cloned cDNAs from a variety of human tissues

Within the same year, Adams *et al.* (1995) published the results from the sequencing of over 50 million nucleotides of cDNA sequence from 37 different tissue types, releasing much of the previously inaccessible data into the public domain. Contiguous cDNA sequences were combined to produce a set of 'tentative human consensus' sequences, each of which was proposed to represent a single gene. In combination with analysis of a further 30 million nucleotides of data existing in dbEST, this 'Human Gene Anatomy Project' was said to provide an overview of gene expression throughout the human body at various stages of development of both normal and abnormal tissue.

EST sequences were screened against existing gene sequences and mRNAs within GenBank, such that entries could be compacted into groups of sequences likely to represent the same gene. These were described as the 'UniGene' set, and were used to characterise newly created EST sequences by virtue of sequence similarity. In the absence of similarity, additional ESTs were classified as potentially novel genes, and classed as the 'UniEST' set, of which 14457 were available by mid-1995 (reviewed in Boguski and Schuler, 1995). By 1996, almost half a million entries had been deposited into dbEST and the UniGene set had risen to approximately 50,000 (Schuler *et al.*,

1996), of which only 9% could be assigned a gene identity whilst the rest remained anonymous.

1.3.11 Production of an EST map of the human genome

The first comprehensive human gene map was presented in 1996 (Schuler *et al.*, 1996), describing a collaborative effort by groups from North America, Europe and Japan. Over 16,000 genes were assigned to chromosomes using radiation hybrid mapping and a YAC library, integrated with over 1000 polymorphic genetic linkage markers. The integration of these map types was a major step towards the possibility of using the positional candidate gene approach for identification of the locus responsible for a given trait. Since all the 1000 polymorphic markers in this report had been chromosomally assigned, ESTs mapped in the vicinity of the linked marker would represent potential candidates for the causative gene, depending on whether their known function correlated with the phenotype itself. Thirty thousand cDNA clone assignments were subsequently added to the human transcript map by the radiation hybrid mapping of human ESTs, a figure that was predicted to include half of all protein coding sequences within the human genome (Deloukas *et al.*, 1998).

1.3.12 The wider applications of EST generation

EST generation has moved from being predominantly focussed on human brain transcripts to begin to encompass other tissues, including skeletal muscle (Muraro *et al.*, 1997), prostate (Nelson *et al.*, 1998) and cochlea (Skvorak *et al.*, 1999). Each represents an enriched resource for the identification of loci involved in tissue-specific disorders. Others have focussed on specific timepoints such as foetal development, using complete foetal tissue representing all organs of the body (Jay *et al.*, 1997), or specific organs such as foetal heart (Hwang *et al.*, 1995). An alternative approach has been the analysis of abnormal tissue in order to assess differences with normal tissue that may implicate genes in the aetiology of the disorder. For example, Frigerio *et al.* (1995) constructed a cDNA library derived from colorectal cancer tissue, and subsequent sequencing analysis led to the identification of a set of novel genes whose expression patterns appeared to differ compared to normal tissue. These sequences therefore formed a source of markers for genes potentially involved in tumourigenesis.

1.3.13 The future use of EST data

At present, dbEST contains 3,828,199 entries from more than 200 different species (release 033100, March 31st 2000, http://www.ncbi.nlm.nih.gov/dbEST/ dbEST_summary.html), of which 75% are derived from human or mouse. A total of 1,513,537 human UniGene clusters have been created (March 28th 2000, http://www.ncbi.nlm.nih.gov/UniGene/Hs.stats.shtml). With the increasing number of EST sequences available, the downstream applications of these markers, and the future directions to be taken in genome analysis have been widely discussed. The current emphasis on structural analysis, and the need to identify and chromosomally assign all human genes, is likely to give way in the near future to more functional analysis. The wealth of data generated by the EST approach is likely to be applied to more detailed studies of the role of specific genes and their products, and of their patterns of expression in the human body. This proposition, and the need to create a publicly accessible database resource for the storage of such information, has been discussed by Strachan *et al.* (1997). Existing techniques, including *in situ* expression studies in embryos and experimental animal systems, and measuring expression variation in RNA/protein blotting studies, are likely to continue to play a large part. However, traditionally these have concentrated on only a small number of gene products at one time. New methods will therefore be needed to cope with the increased workload generated by EST projects. One method that is becoming increasingly popular is expression profiling using microarrays of cDNA clones from a source tissue that can be probed with fluorescently-labelled cDNA derived from another tissue (for example, DeRisi *et al.*, 1996; Strachan *et al.*, 1997). Differences in expression between the two tissues will result in varied levels of hybridisation between probe and target, which can be assessed by relative quantitation of fluorescence. The production, screening and scoring of cDNA arrays are all highly suited to automation, and potentially allow the rapid monitoring of expression levels of thousands of genes at one time. Abnormal expression levels of a particular gene may then lead to the identification of the cause of a particular phenotype. This has obvious implications from diseases such as neoplasms in which changes in gene expression are common (for example, DeRisi *et al.*, 1996).

The identification of single nucleotide polymorphisms (SNPs), sites at which two alternative nucleotides exist at an appreciable level within a population, is increasing in interest. SNPs are thought to exist once in every kilobase of sequence (reviewed in Little, 1999), and can be easily detected by sequence analysis, or more recently, using

microarrays ('DNA chips') fixed on microscope slides (Wang *et al.*, 1998). The consequent genetic diversity created between individuals is thought likely to be responsible for many genetic disorders and other inherited traits, for which an SNP can be used as a marker to analyse the genetic status of individuals at a given locus. However, SNPs are, as would be expected, relatively rare within coding sequences (Wang *et al.*, 1998), such that their isolation from ESTs is less common than from random genomic sequence. The data produced by the various EST projects will no doubt be used for a multitude of such diverse studies in the near future. One such program that has developed from the EST approach relates to the analysis of transcripts potentially involved in human cancer. The Cancer Genome Anatomy Project (CGAP) project, launched by the National Cancer Institute at the NCBI, is making use of the EST database to select likely candidates for further study (http://www/ncbi.nlm.nih.gov/ncicgap). EST generation has been applied to numerous other species. These include species of scientific interest such as mouse (for example, Stewart, G.J. *et al.*, 1997), the nematode *Caenorhabditis elegans* (McCombie *et al.*, 1992; Waterston *et al.*, 1992), and animals of economic value, such as the pig (Winterø *et al.*, 1996; Fridolfsson *et al.*, 1997; Jorgensen *et al.*, 1997) and cattle (for example, Ma *et al.*, 1998). Large scale EST generation for the mouse and other mammals is likely to increase, being one of the aims of the 'Mammalian Gene Collection' project recently launched by several US genomics organisations, chiefly the National Cancer Institute (Strausberg *et al.*, 1999). This will enable comparison of the location of transcribed sequences in the genomes of diverse species, which in turn will be vital to increase our understanding of genome evolution and speciation. The characterisation and mapping of ESTs therefore also has much to offer in the field of comparative genome mapping.

1.4 Comparative genomics

1.4.1 The concept of conserved synteny

Since the divergence of mammalian lineages from their common ancestor, the ancestral karyotype has been subject to breakages, fusions and rearrangements in the generation of the chromosome complements seen in extant species. Despite millions of years of divergent evolution, it is still possible to detect and compare regions corresponding to segments of ancestral karyotypes, in the chromosomes of different species. Often, breakages of ancestral chromosomes have resulted in the separation of formerly linked genes onto separate chromosomes during the period of evolution. Similarly, previously

unlinked loci have become associated through chromosome fusion. In many cases genes have remained tightly associated such that they co-exist on a single chromosome in many modern species. Loci that co-exist on a single chromosome are described as syntenic, regardless of the physical distance between them. The existence of combinations of genes that are syntenic in many divergent species provide evidence to suggest that these regions correspond to remnants of an ancestral karyotype. These regions of conserved synteny have been described by a number of different terms, but herein are referred to as 'evolutionarily conserved chromosome segments', or ECCS. Gene content, and often also gene order, is highly conserved within the corresponding ECCS of different species. As a result, the identification of corresponding ECCS within the genomes of different species provides invaluable information for predicting the likely chromosome location of any specific gene sequence in one species where the ECCS containing that sequence has been identified in another. Consequently, as well as increasing our understanding of genome structure and evolution, the identification of conserved synteny has had an enormous impact on the ability to extrapolate mapping information obtained from one genome to that of another species. This can be of particular significance for those species for which our knowledge is limited, whether by technical, financial, practical or ethical constraints.

1.4.2 Chromosomal rearrangements during evolution

It is known that minimal interchromosomal rearrangements involving the ancestral X chromosome have occurred, since the vast majority of X-encoded loci remain syntenic in most species studied to date (Ohno, 1973, and others). Gene order of X-encoded loci is frequently conserved between species, although differences in relative locus order has been observed between species such as human and mouse (Francke and Taggart, 1980). Banding analysis of X chromosomes of different species has also revealed distinct differences, proposed to represent intrachromosomal rearrangements. For example, in the case of cattle and goat (Piumi *et al.*, 1998), this was proven by the assignment of gene markers to the X chromosome of both species, showing disruption of gene order, but conservation of synteny.

In contrast, it has been shown for autosomes that tightly linked genes have tended to remain syntenic in many different genomes, whereas loosely linked genes in one species are more likely to have been separated onto two or more different chromosomes in other species during evolution (Lalley *et al.*, 1978). This finding reflects the important

difference between two terms commonly used in comparative mapping studies. The first, conserved synteny, merely reflects the fact that a given set of loci exist on the same chromosome. Conserved linkage occurs when loci exist in sufficient proximity that they are not separated either by chromosome evolution or meiotic recombination, and gene order is thus more frequently retained. Conserved linkage of loci can occur at the level of an entire chromosome, as is the case for the X chromosome. For autosomes, this is more likely to occur at the level of chromosome arms, or in the form of sub-chromosomal blocks (ECCS), particularly with increasing evolutionary divergence between the species for which such comparison is made.

1.4.3 Identification of ECCS by comparative banding and gene mapping

Various techniques can be used to identify ECCS in the genomes of two or more species. Comparison of the banded chromosomes of different species at high resolution has been used to identify homologous regions within chromosomes of the Equidae (Ryder *et al.*, 1978) and the Bovidae (Hediger *et al.*, 1991; Gallagher and Womack, 1992), amongst others. This approach has also been used by Sawyer (1991) to compare and align G-banded chromosomes of seven species from different orders, namely primates, carnivores and rodents. Observations widely corresponded with known conserved linkage data for specific chromosome regions, and rearrangements including inversions could be determined. Comparative banding analysis is somewhat laborious and depends on the ability to produce high quality extended chromosome preparations for all species under study. It is also more applicable to highly related species, since the increased karyotypic rearrangements typically observed between the chromosomes of more diverse species are likely to mask the identification of less well conserved banding patterns.

Identification of ECCS can be achieved by mapping individual coding sequences directly to chromosomes, using methods such as the isolation and physical mapping of a genomic or cDNA clone for that gene, or genetic linkage or radiation hybrid analysis. It then becomes possible to identify sets of markers that are consistently syntenic in two or more genomes, which may represent segments of a common ancestral karyotype. Amongst the earliest comparisons were between the two genomes for which genome maps have historically been the most comprehensive, namely those of human and the mouse. Nadeau and Taylor (1984) compared the chromosomal locations of more than 80 homologous genes that had been mapped in both species at this time, identifying 13 putative segments of conserved synteny between their genomes. A total of at least 46

such regions were assumed to exist between their genomes, by the inclusion of additional data for genes with less well-refined mapping data. However, this approach is clearly labour intensive when a genome-wide analysis of multiple species is required, and is reliant on the localisation of a given gene in all species studied. Species for which gene maps are at an early stage are thus not directly suitable for such analysis.

1.4.4 Resources for comparative mapping of type I markers

The increasing interest in the ability to compare the physical location of a given gene between species has given rise to a series of resources and methods designed to expedite gene mapping on a large scale. Such approaches have made use of the increasing number of non-human and non-rodent entries deposited in publicly available sequence databases. Amongst the most comprehensive has been the selection of a set of 321 genes mapped evenly throughout the human and mouse genomes (O'Brien *et al.*, 1993). The markers were proposed to act as a panel of anchored reference loci for the generation of comparative gene maps between mammalian species. O'Brien *et al.* (1993) also introduced the concept of the 'smallest conserved evolutionary unit segment' (SCEUS), regions in which groups of homologous genes were linked without interruption in all species studied. Genes representing each SCEUS were included within the panel of anchored reference loci, with an average spacing of 5 to 10cM across the whole human genome. Preliminary comparative mapping generated a physical location for each anchored reference locus in human, mouse, cattle and cat, demonstrating the existence of many putative regions of conserved synteny. Where possible, PCR primers were designed for each reference anchor locus. These could be used to allow amplification of short regions of coding sequence from diverse species, that could be used to derive a chromosomal assignment for that locus (Lyons *et al.*, 1997). A total of 410 primer pairs were designed to these and other gene sequences, creating a panel of 'comparative anchor tagged sequences' (CATS). These were created by aligning two or more previously annotated exon sequences from a range of vertebrate species, and selecting the most highly conserved regions for primer design. In most cases, primer pairs were designed to amplify both exonic and intronic regions of a given gene, the former allowing gene identification and the latter providing the potential for polymorphism analysis. Of the 410 primer pairs designed, 318 pairs (78%) were optimised for the PCR using genomic DNA from the domestic cat. Sequence data were obtained for 71 cat PCR products, and of these, 59 (83%) demonstrated sequence similarity to the expected gene.

PCR analysis of the 318 optimised primer pairs was extended to 20 further species consisting of mammals, birds and fish, using the conditions established for the cat. The rate of successful amplification of a single product from the 318 loci varied from 52% (Père David deer) to 35% (vole). It was predicted that up to 75% of the total of 410 CATS could be optimised for any eutherian mammal, with a lower rate of success expected for less evolutionarily related species. All but 12 genes for which a CATS primer pair is available have now been ordered in the mouse by various linkage mapping techniques (Chen *et al.*, 1999, and others). In contrast, most of the human homologues were mapped by FISH analysis and as such, relative ordering of genes has to date been at a somewhat lower overall resolution. Until recently, only 195 CATS loci had been mapped to human chromosomes in various studies by radiation hybrid analysis, since the development of the original CATS primer set (summarised in Chen *et al.*, 1999). To further the quality of the high resolution human gene map, 120 more gene markers have since been more finely mapped by radiation hybrid analysis by Chen *et al.* (1999) bringing to 314 the total number of genes represented by CATS mapped in this manner. In all but two cases, cytogenetic and radiation hybrid mapping positions of gene markers were in complete agreement. Comparisons between human and mouse CATS locations led to the identification of 83 conserved chromosome segments between the two genomes. The CATS primer set has therefore provided a useful resource for increasing the number of genes that have been mapped in more than one species, and has made a significant contribution to comparative genome analysis.

The application of PCR primers designed to human EST sequences for the amplification of the corresponding region of pig genomic DNA has been investigated as a method for the rapid generation of comparative gene mapping data (Lahbib-Mansais *et al.*, 1999). Of 344 human ESTs selected in this 'Zoo-PCR' approach, primers from 186 (54%) generated single products from pig genomic DNA. In turn, 60 (56%) of 107 ESTs enabled somatic cell hybrid mapping of the porcine homologue, although reproducible amplification was observed for a greater percentage of the total tested. Approximately 10% of 10,000 primer pairs tested permitted amplification from at least one non-human species, although significant variation in reproducibility of some amplification products was reported. Although this level of success might appear small, elimination of the need to perform cDNA sequencing analysis and primer design might partly compensate, providing that free access to human primer pairs is possible.

1.4.5 Generation of chromosome paints as tools for genome mapping

A considerably more rapid approach for the identification of ECCS between genomes involves a derivation of the standard FISH technique. Whilst standard FISH analysis can be performed for the routine chromosomal assignment of single locus probes, similar techniques can also be used for the analysis of entire chromosomes, or extensive chromosome sub-regions, using a 'chromosome painting' approach. The single locus probe is replaced by one that contains a collection of DNA fragments derived from the entire chromosome of interest. Initially, chromosome-specific libraries were used as the probe (for example, Cremer *et al.*, 1988; Pinkel *et al.*, 1988). Chromosome paint probes are now more typically generated by the sorting of pools of individual chromosomes by bivariate flow cytometry. This technique makes use of the fact that chromosomes vary with respect to both size and base composition. Chromosomes from cultured cells are arrested at metaphase and freed into suspension, and then stained with a combination of DNA-binding dyes, each of which has an affinity for regions of different base composition. A common combination is the AT-binding dye Hoechst 33258, and Chromomycin A3, which binds preferentially to GC-rich sequences. The difference in overall base composition, and the relative amount of DNA contained in a given chromosome (reflecting its size), in turn influences the relative amount of each of the dyes bound. This generates a specific binding profile for each chromosome. Stained chromosomes are passed singly through two fine laser beams, specific for each dye used, and their fluorescence emission profile resulting from this excitation is recorded. These data can be plotted to show the relative binding of both dyes by each chromosome to generate a 'bivariate' flow karyogram, in which each peak corresponds to a different chromosome from the karyotype of that species. Individual peaks can be selected such that all chromosomes exhibiting the corresponding fluorescence pattern are diverted to a collection tube, resulting in a sample highly enriched for a single chromosome. A flow karyogram for the human chromosome complement is shown in figure 1.4a. Although most chromosomes are separated into clearly isolated peaks, some, in particular HSA 9, 10, 11 and 12, cannot be resolved as reliably as others. This reflects the fact that these chromosomes are of a more similar size and base composition to each other compared to all other human chromosomes. Consequently, it may be necessary to use flow-sorted material derived from all four chromosomes to represent all members of this group simultaneously in a given study.

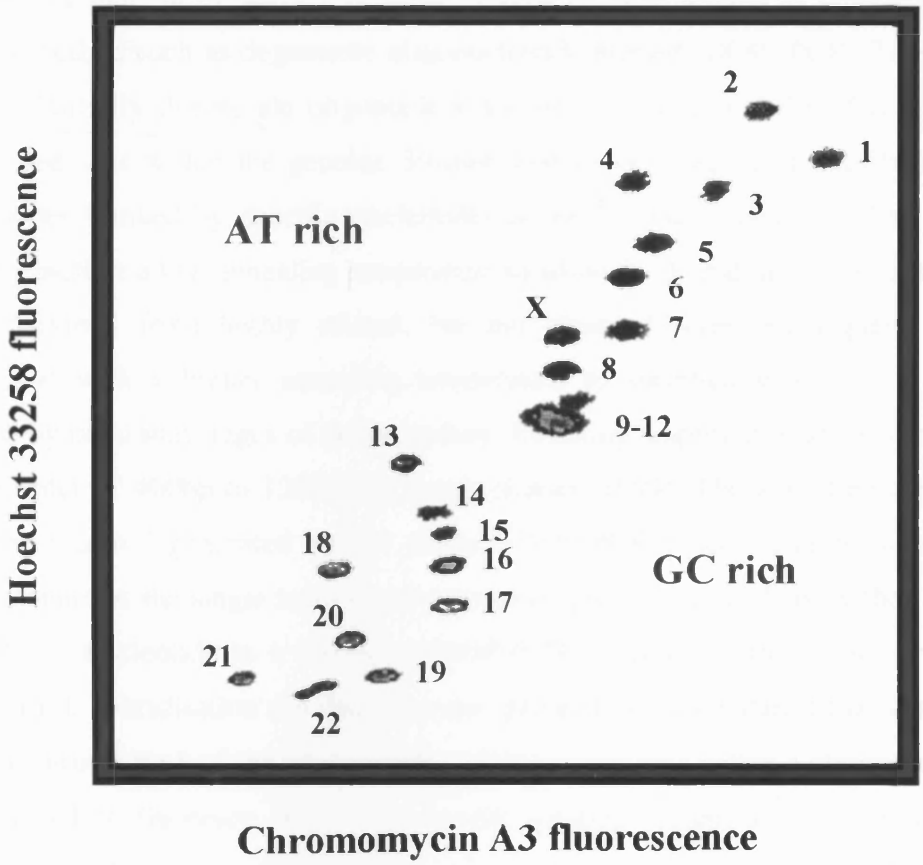


Figure 1.4a:

The human bivariate flow karyotype (adapted from Langford *et al.*, 1996). The human chromosome to which each flow peak corresponds is indicated. This clearly demonstrates the limited resolution of peaks corresponding to HSA 9-12, which lie near the centre of the flow karyotype. The remaining peaks are well isolated from their nearest neighbours.

The minute amount of starting material derived by flow sorting is typically amplified using a method such as degenerate oligonucleotide-primed (DOP) PCR (Telenius *et al.*, 1992a). Partially degenerate oligonucleotides are used to prime the PCR from evenly distributed sites within the genome. Primers consist of a region of usually six random nucleotides flanked by specific nucleotides at the 5' and 3' ends. Cycling conditions initially include a low annealing temperature to allow the 3' end of each primer to anneal to and extend from highly related, but not identical, sites. Subsequent cycles are performed with a higher annealing temperature to specifically amplify the material created by the initial stages of the procedure. Resulting amplified fragments are typically of the order of 400bp to 1200bp in length (Carter, 1994). The amplified chromosome-enriched material generated by this primary DOP-PCR process can be labelled in the same manner as the single locus probes discussed previously, such as by the inclusion of a modified nucleotide in a secondary DOP-PCR, to generate the chromosome 'paint' probe. Both hybridisation and detection also proceed as for standard FISH analysis, such that the entire length of the corresponding chromosome pair within a chromosome spread is observed to fluoresce. This chromosome painting technique has a wide range of applications. Where two chromosomes have a similar banding pattern, ambiguity can be overcome by the application of paint probes for these two chromosomes, where each is labelled and detected with a different reporter-fluorochrome combination. The two chromosome pairs can easily be distinguished by their differential fluorescence profiles. Single locus probes and chromosome paints can be combined in FISH analysis, so that the localisation of a probe to a chromosome can be achieved with confidence where ambiguity would exist in the absence of the co-hybridisation of a paint probe. Chromosome painting can aid the identification of certain forms of karyotype rearrangement. Grossly abnormal breakage, fusion, loss or gain of chromosomes can be detected by deviation from the expected two pairs of fluorescent entities within a metaphase spread. Similarly, flow cytometry of cells with an abnormal karyotype can be used to isolate aberrant chromosomes, and to produce a corresponding paint probe. Hybridisation of this probe to normal metaphases can then be used to determine the origin of the abnormal chromosome, in a procedure termed 'reverse chromosome painting' (Telenius *et al.*, 1992b). Flow-sorted material can be cloned to generate libraries enriched for sequences derived from a single chromosome (for example, VanDevanter *et al.*, 1994).

Flow cytometry has been applied to many species other than human, such that a panel of flow-sorted material that is both chromosome-enriched and species-specific can be generated. More recently, microdissection-mediated selection of single chromosomes fixed to a microscope slide has provided an alternative method for the production of chromosome-enriched material (for example, Christian *et al.*, 1999). This method has the advantage of enabling the isolation and collection of sub-chromosomal regions or individual aberrant chromosomes, for more focussed studies such as library generation (for example, Chaudhary *et al.*, 1998a) or chromosome breakpoint analysis (for example, Weimer *et al.*, 1999). The technique also offers an opportunity to overcome the resolution limitations of bivariate flow cytometry, such as that commonly encountered for HSA 9, 10, 11 and 12.

1.4.6 The use of chromosome paint probes in cDNA selection

Chromosome paint probes have relatively recently become tools for the selection of cDNA sequences that are highly enriched for transcripts from the corresponding chromosome. The basic principle, originating from work by Lovett *et al.* (1991) and Parimoo *et al.* (1991) involves filter hybridisation, or more typically solution hybridisation, of denatured 'whole-genome' cDNA with a biotinylated genomic DNA template derived from a single chromosome region or genomic clone. A proportion of the single-stranded cDNA sequences will anneal to their complementary genomic sequence if the latter is represented within the genomic probe. The cDNA-genomic DNA complex can be isolated from the remaining cDNA population using magnetic beads to which streptavidin molecules are covalently attached. The magnetic beads are added to the cDNA selection reaction, and allowed to hybridise to the biotinylated genomic DNA template. When placed in contact with the appropriate magnetic apparatus, the bead/genomic DNA/cDNA complex is concentrated at the bottom of the reaction tube, enabling the liquid phase to be removed and discarded. The latter contains free cDNA molecules for which the complementary sequence was not represented within the genomic probe. The bead complex is subjected to a series of stringency washes to ensure removal of all cDNA sequences that are not bound to a biotinylated genomic DNA fragment. What remains therefore predominantly comprises transcripts derived from the chromosome region of interest, that can be amplified by the PCR, and thus isolated from the bead complex. The selected cDNA sequences can be cloned and used as a source of type I markers highly enriched for transcripts derived from a single chromosome region,

in contrast to the random generation of EST sequences from the entire genome. The specificity of the procedure can be assessed by hybridisation of the selected transcript population to metaphase chromosomes in standard FISH analysis, for which highly specific hybridisation signals on the chromosome region of interest would be expected. If required, the amplified, selected cDNA can be cycled through a second round of selection, to increase the specificity of the resulting transcript population for those derived from the desired chromosome site. It has been estimated that each round of selection enriches the starting cDNA source by at least two to three orders of magnitude for transcripts from the chromosome region of interest (Korn *et al.*, 1992), although this value is dependent on the precise reaction components and conditions used. In particular, the level of enrichment decreases as the physical size of the genomic target increases (Lovett, 1994a).

The cDNA selection procedure has been applied in a range of approaches, predominantly differing in the nature of the biotinylated target DNA. The technique has most commonly been applied to the hybridisation of cDNA populations to biotinylated genomic clones from specific subchromosomal regions (for example, Korn *et al.*, 1992; Del Mastro *et al.*, 1995). Biotinylated chromosome segments isolated by microdissection have also been used for capturing transcripts within this region (for example, Morgan *et al.*, 1992; Su *et al.*, 1994). This is particularly valuable as an approach towards the isolation of candidate genes for a given trait, where a subchromosomal localisation for the causative locus is known. Others have isolated transcripts derived from the chromosome of interest by hybridising whole-genome cDNA populations to metaphase chromosome spreads, and then isolating the chromosome template and associated transcripts by microdissection (for example, Hozier *et al.*, 1994; Gracia *et al.*, 1997). The use of a chromosome paint probe as the genomic template for cDNA selection has been less widely exploited (for example, Chen-Liu *et al.*, 1995; Rouquier *et al.*, 1995), in part due to its greater sequence complexity compared to cloned DNA or chromosome subregions.

In order for the cDNA selection process to be effective, it is necessary to ensure that highly repetitive sequences within the genomic template are suppressed prior to the addition of the cDNA population. The common occurrence of highly repetitive sequences within cDNAs is often also counteracted by repeat suppression. This can be achieved in the same manner as that used in standard FISH analysis, in which the denatured genomic DNA is allowed to reanneal in the presence of competitor DNA, most typically the highly repetitive C_0t1 fraction of the genome. The relative concentration of the cDNA,

genomic DNA and competitor varies widely in previous descriptions of the technique, and is largely dependent on the nature and complexity of the genomic template.

1.4.7 Comparative chromosome painting analysis

Amongst the most valuable applications of the chromosome painting technique is the identification of evolutionarily conserved regions between the genomes of different species. Chromosome paint probes from one species can be hybridised to metaphase chromosomes of another. The pattern of hybridisation observed on the latter identifies ECCS in this karyotype that are the evolutionary equivalents of the chromosome of the former species from which the paint probe was derived.

Early applications of this 'heterologous chromosome painting' technique focussed on the human karyotype, and its comparison with various primates (for example, Wienberg *et al.*, 1990; Stanyon *et al.*, 1995), in order to increase understanding of their close phylogenetic relationships. In one of the first such studies, a considerable degree of rearrangement was observed between human and gibbon chromosomes by the application of all human paint probes to various gibbon species. However, human, orangutan, gorilla and chimpanzee karyotypes were shown to have maintained an extremely high level of conservation (Jauch *et al.*, 1992). Subsequently, more detailed studies have been performed to establish the precise nature of the chromosome rearrangements that have occurred in the evolution of primate karyotypes, including human and gibbon (Koehler *et al.*, 1995) and human and chimpanzee (Nickerson and Nelson, 1998).

Heterologous chromosome painting analysis has now been widely extended to the karyotypes of a broad range of species, more recently being extended to include species of economic and domestic importance. Frequently this has involved direct comparison with the human genome, allowing the transfer of information from this intensively studied species towards a better understanding of those for which genome mapping is at an early stage. This in turn has increased understanding of the wider issues involved in karyotype evolution, as well as providing highly valuable information for those in search of specific loci in one species for which the human orthologue has already been mapped. Comparison of the location of the human ECCS containing this locus with the corresponding segment in a less comprehensively studied genome immediately provides a valid starting point for locating the gene of interest in the latter, narrowing down the search to one or a small number of subchromosomal regions.

Heterologous chromosome painting analysis between non-primates stemmed largely from the work of Scherthan *et al.* (1994) in the application of a technique that has become widely known as 'Zoo-FISH'. One mouse and four human chromosome paint probes were applied to metaphases of a range of distantly related mammalian orders, including primates, rodents, ungulates and whales, which diverged from a common ancestor between 50 and 85 million years ago. The cross-species chromosome painting technique that had previously been applied successfully between primates was optimised to allow comparison of more diverse species without significant reduction in sensitivity or resolution. The three most important factors identified were the quality of the chromosome spreads used, the concentration of the probe applied and the duration of the hybridisation period. Comparison of two relatively distantly related genomes was assumed to involve a reduced overall level of probe-target homology in a given hybridisation experiment. However, it was also assumed that even this small proportion of total homology between the paint probe and the chromosome target would be sufficient to generate detectable and measurable hybridisation signal. This assumption was demonstrated by the successful comparative painting analysis between these four mammalian orders, simply by increasing both the amount of paint probe used (by as much as 150-fold), and the duration of hybridisation to high quality chromosome spreads (by up to five-fold). In all cases, the X-chromosome probe hybridised solely to the X chromosome of each species (with the exception of regions of constitutive heterochromatin), correlating with existing data regarding the extensive conservation of X-encoded loci during evolution. ECCS corresponding to HSA 1, 16 and 17 were also identified in the mouse, muntjac and fin whale, in all cases confirming existing data for regions of conserved synteny identified by comparative gene mapping. It was estimated that conserved segments of as little as 7MB in size could be detected by this Zoo-FISH method (Scherthan *et al.*, 1994).

1.4.8 Reciprocal chromosome painting analysis

Data derived from Zoo-FISH analysis can be enhanced under circumstances in which chromosome paint probes are available for both species of interest. This enables comparative chromosome painting analysis to be performed in a reciprocal manner. A bi-directional approach is valuable in confirming data obtained from the application of chromosome paints from one species to chromosomes of another. This is particularly vital where a degree of ambiguity exists in the interpretation of unidirectional analysis.

Ambiguity is a particular problem when Zoo-FISH is performed between highly diverged species where signal quality is low, or where ECCS between karyotypes approach the minimum size for detection by this method. Where data obtained from both aspects of the reciprocal procedure do not correlate, this may indicate a discrepancy that warrants further investigation. Most importantly, however, reciprocal chromosome painting allows the location of an ECCS to be refined to a sub-chromosomal location in both species. This is demonstrated in figure 1.4b, a schematic representation of both uni- and bi-directional Zoo-FISH analysis.

1.4.9 Comparative mapping analysis amongst vertebrate species

The development of techniques such as those outlined above has markedly increased our understanding of comparative genomic structure, particularly amongst the mammals. Amongst the earliest comparisons made were between the human and mouse genomes, since these represent the species for which most is known in terms of genome structure and composition. Nadeau (1989) furthered previous observations (Nadeau and Taylor, 1984) in a report of the chromosome assignment of over 240 genes mapped in both mouse and human. From the observed length of conserved chromosome segments, it was inferred that approximately 138 separate rearrangements have occurred since the divergence of mouse and human autosomes from their common ancestor. This gives some support to those who question the use of the mouse as a model for human disease and other traits, and as a resource for the identification of the causative loci, since their genome structure is clearly highly diverged. For example, the comparison of gene locations between HSA 22 and the mouse genome has shown that this single human chromosome shares an evolutionary origin with segments of no less than seven mouse chromosomes (reviewed in Dunham *et al.*, 1999).

A further karyotype to be studied in detail is that of the domestic cat (*Felis catus*, FCA), which has been suggested to represent an improved animal model for human studies. Initial reports identified 17 cat linkage groups by the somatic cell hybrid mapping of gene markers for the cat whose human homologue had been chromosomally assigned (O'Brien and Nash, 1982), with only three exceptions where linkage did not hold. Twelve of the linkage groups could be assigned to a cat chromosome, demonstrating a high degree of linkage conservation between the two genomes, in contrast to the frequent disruption of linkage in the mouse genome. A significant advancement of the cat-human comparative map was made by the application of unidirectional (Rettenberger *et al.*,

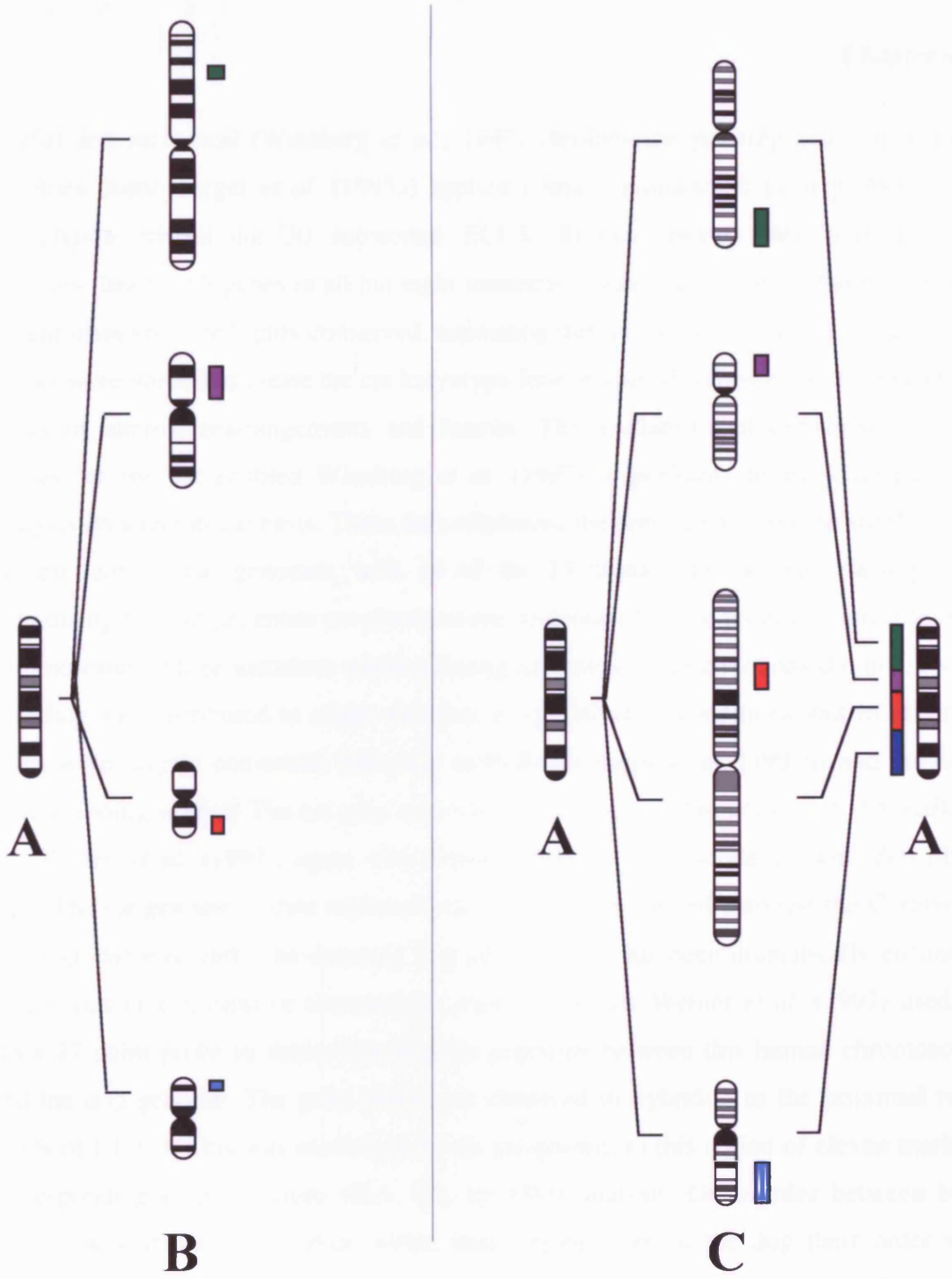


Figure 1.4b:

Schematic representation of unidirectional and reciprocal chromosome painting analysis. The location of each ECCS is represented by a vertical bar against the corresponding chromosome. The image on the left demonstrates that unidirectional painting analysis enables ECCS corresponding to a chromosome from one species (A) to be assigned at the subchromosomal level in the second species (B), with no information regarding where the corresponding chromosome regions exist in species A. Reciprocal chromosome painting analysis, represented on the right, enables subchromosomal localisation of ECCS in both species A and C.

1995a) and reciprocal (Wienberg *et al.*, 1997) chromosome painting analysis to their genomes. Rettenberger *et al.* (1995a) applied human chromosome paint probes to cat metaphases, identifying 30 autosomal ECCS. Results corresponded with existing mapping data for 68 genes in all but eight instances. Again, data implied that human and cat chromosomes are highly conserved, estimating that a total of only seven chromosome breaks were needed to create the cat karyotype from human chromosomes, followed by a series of internal rearrangements and fusions. The availability of chromosome paint probes for the cat enabled Wienberg *et al.* (1997) to perform chromosome painting analysis on a reciprocal basis. These data reinforced the previously observed similarity of the cat and human genomes, with 14 of the 23 human chromosome paint probes hybridising to a single, entire cat chromosome, and nine of 19 cat probes to single human chromosomes. Three instances where painting analysis in both directions did not wholly correlate were attributed to slight variation in hybridisation conditions, but in all other cases were largely consistent with data from Rettenberger *et al.* (1995a) and previous gene mapping studies. The cat gene map was further increased to consist of 105 markers by O'Brien *et al.* (1997), again with almost complete correspondence with Zoo-FISH data. The cat genome to date represents the most widely studied amongst the Carnivora, although more recently the domestic dog genome map has been dramatically enhanced by a series of comparative chromosome painting studies. Werner *et al.* (1997) used an HSA 17 paint probe to detect homologous segments between this human chromosome and the dog genome. The paint probe was observed to hybridise to the proximal two-thirds of CFA 9. This was confirmed by the assignment to this region of eleven markers corresponding to genes from HSA 17q, by FISH analysis. Gene order between both species was shown to correlate within these regions, but in the dog their order was inverted with respect to the centromere. The HSA 17 paint probe also hybridised to a mid-region of CFA 5, which was confirmed by the mapping of two further markers from HSA 17p. Selected markers were also assigned by meiotic linkage analysis, using microsatellite markers isolated from cosmid clones containing the relevant gene. These data represent the first detection of homologous segments between the dog and human genomes by chromosome painting analysis. The present study (Thomas *et al.*, 1999) describes the initial application of reciprocal chromosome painting between dog and human, whilst genome-wide reciprocal Zoo-FISH analysis between these genomes was first performed by Breen *et al.* (1999a). The latter study identified a minimum of 68 ECCS between their karyotypes, considerably higher than for other members of the

Carnivora. Twenty one of the 39 dog autosomes were painted exclusively by a single human chromosome paint, with the remainder painted by two to four human paint probes. The reciprocal painting analysis in turn demonstrated that only four human chromosomes were painted by a single dog chromosome paint, with seventeen further human chromosomes painted by two to five dog paint probes. A total of seven dog chromosome paint probes were shown to hybridise to human chromosome one. These data were used to assign 41 of 57 previously reported radiation hybrid linkage groups (Priat *et al.*, 1998) to a total of 23 dog chromosomes by Breen *et al.* (1999a). More recently, Zoo-FISH analysis between dog, human and red fox has been described by Yang *et al.* (1999). These data are discussed further in chapters four and seven.

Rettenberger *et al.* (1995b) have performed unidirectional analysis of human paint probes on pig chromosomes, identifying a total of 47 ECCS. A significantly greater degree of conservation was shown between the human and pig karyotypes than those of human and mouse. The same number of ECCS was observed by the study of Frönicke *et al.* (1996), with detailed integration of novel and existing genetic and physical mapping data for the pig. Variation in the intensity of hybridisation patterns seen for each human paint probe were also noted. One error in chromosome orientation in the study of Rettenberger *et al.* (1995b) was resolved, and several instances of disparity in the order and extent of certain ECCS were noted. The inclusion of gene mapping information therefore gives valuable supporting evidence to confirm and refine chromosome painting data. Additional refinement was made by Goureau *et al.* (1996) in the first reciprocal chromosome painting analysis of the human and pig karyotypes. This study identified subchromosomal regions in which the hybridisation patterns of two paint probes appeared to overlap. Co-hybridisation of both paint probes revealed alternation of both probes over the region of overlap. Further additions to the pig comparative map have included somatic cell hybrid mapping of porcine ESTs (Fridolfsson *et al.*, 1997; Jorgensen *et al.*, 1997; Winterø *et al.*, 1998) and detailed analysis of sub-chromosomal regions by radiation hybrid mapping (Robic *et al.*, 1999). FISH mapping (Van Poucke *et al.*, 1999) and somatic cell hybrid mapping (Sun *et al.*, 1999a) of gene markers from a known human chromosome location have generated an example of conserved linkage but not gene order between corresponding ECCS in two species, here involving HSA 3 and SSC 13. This implies a degree of internal rearrangement in one or both ECCS. Similar analysis of HSA 13 and SSC 11 shows an example of conserved gene order within ECCS (Sun *et al.*, 1999b).

The cattle genome has also been widely studied from a comparative mapping angle. Unidirectional genome-wide chromosome painting analysis has been reported by Hayes (1995), Solinas-Toldo *et al.* (1995) and Chowdhary *et al.* (1996), with the application of human paint probes to cattle chromosomes demonstrating between 46 and 50 ECCS. In the study of Solinas-Toldo *et al.* (1995), 15 of 29 cattle (*Bos taurus*, BTA) autosomes were entirely painted by a single human paint probe, indicative of a high level of conservation between their genomes. Chowdhary *et al.* (1996) compared their chromosome painting results with homologies identified by existing gene mapping data. Overall, 95% concordance was observed between these data sets. Observations of the high level of genome conservation between human and cattle led to an hypothesis that such correlation cannot solely be due to homology of genes, but must also extend to similarity between non-coding DNA segments (Chowdhary *et al.*, 1996). Data from the three chromosome painting studies were highly comparable, with variations attributed to the different sources of paint probes used, and to the use of different hybridisation conditions. Somatic cell hybrid and radiation hybrid mapping analysis have also been used in the generation of mapped gene markers for comparative analysis between the human and cattle genomes. Much of the focus has been on the entire conservation between BTA 19 and HSA 17 (Yang and Womack, 1995; Yang *et al.*, 1998) detected by the Zoo-FISH studies outlined above. Again, high resolution methods of hybrid analysis demonstrated the ability to detect internal gene order differences within two regions shown to be highly conserved at the cytogenetic level. The lack of corresponding high-resolution assignments for some human genes was noted to be a limiting factor in the ability to directly compare gene order between species.

Significant similarity between the cattle and sheep karyotypes, which diverged some 20 million years ago, has been determined by comparison of their chromosome banding patterns by Hediger *et al.* (1991), supplemented by the comparison of a small number of gene assignments. Subsequently, further similarity was noted by Chowdhary *et al.* (1996), detected by the extension of their studies to the application of selected human paint probes to sheep chromosomes, for which significant correlation was found with data from human-cattle Zoo-FISH analysis. Iannuzzi *et al.* (1999) performed extensive chromosome painting analysis of human probes on sheep metaphases, with the identification of 48 ECCS throughout the sheep genome, correlating with existing gene mapping data. It has therefore been proposed that the closely related cattle and sheep genomes may be useful models for the extrapolation of data between highly related

species. The bovine order may be a suitable candidate in particular, since the ability to directly interrelate their chromosome banding patterns (Hediger *et al.* (1991), facilitates the development of theories regarding the evolutionary rearrangements that have occurred since the divergence of their karyotypes.

A total of 43 ECCS have been identified between the horse and human genomes, which diverged 70-80 million years ago, by Raudsepp *et al.* (1996). Five horse chromosome regions demonstrated no apparent hybridisation of human paint probes, although this was attributed to features of the paint probes used. A high degree of conservation was observed between the horse and human genomes, with the majority of equine chromosomes (21 of 31 autosomes) being painted by a single human chromosome. It has been estimated from these data that only 20 chromosome breaks (plus a number of fusions) are necessary to recreate the horse karyotype from human chromosomes. A complete human-rabbit comparative map has recently been produced by reciprocal chromosome painting analysis (Korstanje *et al.*, 1999). Hybridisation of rabbit paint probes to human chromosomes demonstrated the presence of 42 ECCS, whilst the reciprocal experiment identified 38 human ECCS.

Chromosome painting analysis has largely been restricted to pairs of mammalian species that diverged less than approximately 100 million years ago, and most commonly to species sharing a more recent common ancestor. More recently, chromosome painting studies have been extended to less evolutionarily-related genomes including human and avian species, which diverged up to 350 million years ago (for example, Chowdhary and Raudsepp, 2000). Other non-mammalian species are joining the growing list of species for which comparative gene mapping analysis has been applied, including the Japanese pufferfish (*Fugu rubripes*) (Gilley and Fried, 1999). Human and pufferfish are separated by approximately 900 million years of evolution, significantly greater than that for most pairs of species for which comparative mapping has been performed. Despite their distant evolutionary relationship, the pufferfish has become an important experimental organism due to its compact genome. Although believed to contain a comparable gene content to that of human and other mammals, the pufferfish genome is less than one seventh of the size (Brenner *et al.*, 1993). The difference in size has been attributed to reduction in repetitive sequence and intron size, and to a high gene density. Such factors are exploited in the use of the pufferfish as a model for the human genome since its compact structure increases the efficiency by which some genome studies, such as identification of candidate disease genes, can be performed. However, this is dependent on homology

between the human and pufferfish genome being of a sufficient level that the results of such studies can be validly extrapolated to the human genome. Instances of extensive gene order differences between pufferfish and human were identified in the study of Brenner *et al.* (1993), amongst others. Caution is therefore advised in studies in which pufferfish genome structure is applied to human. However, since gene order differences have also been reported between several highly related species, the implications of these findings may not yet be fully apparent.

1.4.10 Exceptions to comparative chromosome painting data

Chromosome painting by its nature can only provide comparative mapping information at a gross level. A relatively small number of exceptions to Zoo-FISH data have been reported, such as that by Aleyasin and Barendse (1997). In this study, the mapping of a gene from HSA 22 to BTA 22 did not correlate with existing data for regions of conserved synteny between these genomes. This finding might suggest the existence of a previously uncharacterised ECCS existing between HSA 22 and BTA 22 that is less extensive than the 7Mb maximum resolution suggested for Zoo-FISH analysis (Scherthan *et al.*, 1994). Consequently, the gross data provided by Zoo-FISH must be supplemented by mapping of unique coding sequences to different genomes in order to generate high-resolution comparative maps. Instances of conserved linkage in the absence of conserved gene order have been more widely reported. One example involves the mouse homologue of the region implicated in Huntington's Disease in man, caused by a trinucleotide repeat expansion (The Huntington's Disease Collaborative Research Group, 1993) at HSA 4p16.3 The corresponding segment lies on mouse (*Mus musculus*, MMU) chromosome 5. Linkage analysis of six markers to both chromosomes demonstrated the presence of a chromosome rearrangement within this region of conserved synteny that disrupted the relative order of these loci. A similar situation has been observed for conserved synteny, but disrupted gene order, between HSA 4 and SSC 8, which has been attributed to an internal inversion event in one of the ECCS (Ellegren *et al.*, 1993). Similarly, gene order between HSA 18 and BTA 24 has been shown to be highly rearranged, but with conservation of synteny (Larsen *et al.*, 1999). It is highly likely that further detailed comparative study of the gene composition of ECCS between species will reveal further such examples.

A combined analysis of comparative gene mapping data for ruminants and humans has shown that the level of chromosome rearrangement between karyotypes is far greater

than that demonstrated by Zoo-FISH analysis (Schibler *et al.*, 1998). This has been noted even for chromosomes of one species that are solely and entirely painted by a single chromosome paint probe from another. Intrachromosomal rearrangements are therefore proposed to be as common as those between different chromosomes, though the former are less easily detected. This emphasises the need for high-resolution gene mapping information in attempting to accurately establish the level of homology between genomes.

1.4.11 Advances in heterologous chromosome painting analysis

Reciprocal chromosome painting analysis enables sub-chromosomal localisation of regions of conserved synteny between two genomes, and is a significant advantage over unidirectional approaches. However, this relies on the availability of chromosome paint probes for both species of interest. In many cases, such resources are not readily available, and so alternative approaches must be sought. Refinement of ECCS can be achieved by the application of human (or other) paint probes representing single chromosome arms rather than entire chromosomes. This approach has been demonstrated by Chaudhary *et al.* (1998b) in the application of five arm-specific probes for human chromosomes to pig and horse metaphases. The recent development of resources by which paint probes for all 24 human chromosome types can be analysed in a single hybridisation experiment (Schrock *et al.*, 1996; Speicher *et al.*, 1996), also has implications in Zoo-FISH. These techniques have arisen through the availability of an increasingly wide range of reporter molecules, and fluorochromes with various excitation and emission characteristics. The spectral karyotyping approach of Schrock *et al.* (1996) has been used to hybridise all human chromosome paints simultaneously to chromosomes of the gibbon karyotype. The results of this study were reported to demonstrate complete correlation with existing data for human-gibbon Zoo-FISH analysis. This type of approach is likely to enable more rapid identification of ECCS between species than is currently possible by the application of individual paint probes. However, the number of suitable reporter molecules and fluorochromes available remain a major limitation to the widespread application of the method to other species, particularly those with a high chromosome number.

Comparative genomic hybridisation (CGH) has been widely used as a tool for identifying gross differences between two genomes from members of the same species. The method, developed by Kallioniemi *et al.* (1992), involves the differential labelling of genomic

DNA from two sources of interest, and simultaneous hybridisation of both probes to a standard metaphase template following a preannealing period. The most common use of the method has been in the identification of gross chromosome aberrations in tumour cells. Tumour DNA (the test probe) is labelled with one fluorochrome, and combined with normal genomic DNA (the reference probe) labelled with a different fluorochrome. Both are hybridised to normal metaphases, and analysed by fluorescence microscopy, under conditions highly analogous to standard FISH analysis. Relative fluorescence intensities are calculated along the length of each chromosome, and deviations from the expected 1:1 ratio represent chromosome imbalances generated by regions of loss or gain. Regions demonstrating fluorescence corresponding only to the reference probe indicate deletions in the tumour cell. Conversely, an excess of probe corresponding to a particular chromosome region, caused by amplification in the tumour cells, will result in the overall fluorescence becoming skewed towards that of the test probe. CGH can also be used in a cross-species manner, and has recently been applied to the comparative analysis of the human, pygmy chimpanzee and gorilla genomes (Toder *et al.*, 1998), identifying several amplified or species-specific sites, in all cases within heterochromatic regions. The same study also described the application of a 'representational difference analysis' (RDA) method, by which sequences present in one genome but absent in another can be isolated. The principle, first described by Lisitsyn *et al.* (1993), involves a subtractive hybridisation approach similar to that used for generating cDNA libraries enriched for sequences that are specific to a given tissue. Three sequences present only in the gorilla genome and absent in human DNA were isolated by RDA (Toder *et al.*, 1998). Each was shown to represent a form of repetitive element, typically associated with telomeres. Thus RDA is proposed as a useful general approach to the isolation of species-specific repetitive sequences, which are likely to be at least in part responsible for differences in heterochromatin content identified by CGH. It is inevitable that additional advances such as these will be made in the near future that will enable a better understanding of the precise relationships between the genomes of different species.

1.4.12 Chromosome evolution and the proposed ancestral karyotype

Varying opinions exist as to the nature of the ancestral vertebrate karyotype, and also the nature of the rearrangements that have occurred to result in the chromosomes of modern species. The 'fusion hypothesis' implicates an ancestral karyotype with a high number of acrocentric chromosomes, which has progressively been reduced by a series of fusions to

result in a smaller number of metacentric chromosomes in modern species. An opposing view, that ancestral species originally demonstrated a karyotype with a low number of metacentric chromosomes, has been termed the 'fission hypothesis'. It implies that centromeric fission events have occurred to split chromosomes into a greater number of separate entities during subsequent evolution. A third potential explanation is a combination of both fusion and fission from a karyotype of moderate chromosome number, combined with other more chromosome-specific rearrangements (reviewed in O'Brien and Nash, 1982; Qumsiyeh, 1994, and Rettenberger *et al.*, 1995a). This issue is unlikely to be resolved until considerably more detailed information is available from both comparative genome analysis and phylogenetic studies.

Ohno (1984) supported the view that the ancestral vertebrate karyotype was represented by a 2n = 48 structure. This would comprise 24 separate linkage groups, each representing an acrocentric chromosome, with the original X and Y chromosomes initially being highly homologous. This suggestion was in part the result of the observation that many organisms display such a karyotype arrangement, or a closely related structure. An early survey of genome sizes and karyotype structure in eutherian organisms demonstrated the most common diploid chromosome numbers to be 46 and 48 (Matthey, 1968, cited in Ohno, 1984), supporting the 2n = 48 hypothesis of the karyotype of a common mammalian ancestor. Similarly, the common occurrence of between 40 and 50 ECCS between the karyotypes of many species studied thus far has been noted (Schibler *et al.*, 1998), whether human or other chromosome paints are used as the probe.

It has frequently been noted that both human and cat karyotypes resemble a model of a putative ancestral mammalian karyotype (Rettenberger *et al.*, 1995a). This observation derived from comparison with their data from a similar study of the pig, with 47 ECCS shared with the human karyotype (Rettenberger *et al.*, 1995b) and Zoo-FISH data for the cat and human genomes (Rettenberger *et al.*, 1995a). It was noted that the pig, cat and human karyotypes are highly comparable in overall morphology, although homology of banded chromosomes could only be seen between cat and human. It was suggested that species with solely acrocentric autosomes, including cattle and also the domestic dog, underwent a further species-specific karyotype rearrangement after the major divergence into separate mammalian orders. Human-pig comparative data demonstrate the existence of 47 ECCS (Rettenberger *et al.*, 1995b), compared to 51 for human-gibbon (Jauch *et al.*, 1992). Thus it has been noted that the number of karyotypic rearrangements existing between the two genomes is more a reflection of the similarity of their mode of evolution

rather than being suggestive of a close evolutionary distance between them (Rettenberger *et al.*, 1995b).

This may also apply to the rodents, since their genomes appear to have been subject to a far greater degree of rearrangements than are found in most mammals (reviewed in Graves, 1996). One might assume that the close evolutionary relationship between mouse and rat, species which diverged only 10-20 million years ago, would result in a high degree of conservation between their genomes. However, reciprocal chromosome painting analysis has shown their chromosomes to be highly rearranged (Stanyon *et al.*, 1999), indicating that their relative karyotype evolution has occurred at a high rate. A minimum of 14 translocations appear to have occurred between the rat and mouse genomes within the 10-20 million year period of their evolution (Stanyon *et al.*, 1999). In contrast, a single translocation appears to have occurred in the 5-6 million years of evolution between humans and chimpanzees (Jauch *et al.*, 1992). Thus the results of the study by Stanyon *et al.* (1999) estimated that the rate of chromosome translocation that has occurred between the rat and mouse genomes appears to be up to ten times greater than that between those of other species, such as human and cat, and human and chimpanzee. The precise timing and nature of the various stages in evolution of the rodent karyotype is therefore less easily predicted.

The location of corresponding ECCS in different genomes has given some insight into the possible composition of some ancestral chromosomes. Comparison of heterologous chromosome painting data for eight species, representing four mammalian orders, has led to the identification of certain consistent patterns in ECCS distribution between their genomes (Chowdhary *et al.*, 1998). For example, regions corresponding to HSA 13, 17, 20 and X appear to be regularly conserved as a single chromosome segment. These human chromosomes therefore frequently correspond to an entire chromosome, or chromosome arm, in the karyotypes of other mammals. Others, including HSA 2, 4, 5, 6, 9 and 11, constitute large chromosomal regions of other species, rather than entire chromosomes or arms. A third pattern was identified, in which regions corresponding to two different human chromosomes are commonly found adjacent (or in close proximity) to each other in other karyotypes. These two segments are likely to represent the composition of an ancestral chromosome that has since become dissociated by chromosome fission in some species, including human. HSA 12 and 22 are amongst such examples. Chromosome regions shown to share an evolutionary origin with these human chromosomes remain associated in many species, including the cat (Rettenberger *et al.*,

1995a; Wienberg *et al.*, 1997), domestic dog (Breen *et al.*, 1999a; Yang *et al.*, 1999), dolphin (Bielec *et al.*, 1998), cattle (Hayes, 1995; Solinas-Toldo *et al.*, 1995; Chowdhary *et al.*, 1996), pig (Rettenberger *et al.*, 1995b; Frönicke *et al.*, 1996; Goureau *et al.*, 1996), and horse (Raudseppe *et al.*, 1996). Thus HSA 12 and 22 are likely to have arisen by the reciprocal translocation of two segments from this ancestral chromosome to become two separate entities. Comparisons such as these, particularly with the addition of data from more species, will be of great value in establishing the precise modes of chromosome evolution between species, and the karyotype structure of their common ancestors.

1.4.13 Genome evolution within the Carnivora

A model of an ancestral carnivore karyotype ('CAR'), based on banding homologies between six species from five different families, has been proposed (Dutrillaux and Couturier, 1983). This consists of a small number of bi-armed chromosomes, and is thought to most likely represent the karyotype structure of the common ancestor of the Carnivora, whose origins can be traced back some 40-60 million years (Wayne, 1993). The structure of the CAR closely resembles the modern domestic cat karyotype (Nash and O'Brien, 1982; Nash and O'Brien, 1987), which in turn is highly similar to that of the American Mink (*Mustela vison*) (Hameister *et al.*, 1997), and the Harbour Seal (*Phoca vitulina*) (Frönicke *et al.*, 1997). More recently, their comparison has led to the development of a CAR based on Zoo-FISH data (Z-CAR) (Frönicke *et al.*, 1997).

However, this level of similarity does not extend to all other members of the order Carnivora. Canids, in addition to Ursids, are amongst the main exceptions, with very limited regions of similarity between their banded chromosomes and the CAR (Wayne *et al.*, 1987a). An extreme degree of variation exists between the karyotypes of members of this family, and for this reason, the Canidae were not included amongst the species studied in the development of the CAR by Dutrillaux and Couturier (1983). It has been estimated that the Canidae diverged from other Carnivora families as long as 60 million years ago, early in the evolution of the Carnivora from other orders (Wayne, 1993). Amongst the Canidae, the highest chromosome numbers, 2n = 78, are found in the domestic dog and the Grey wolf (*Canis lupus*), in which all autosomes are acrocentric. In contrast, the red fox karyotype (*Vulpes vulpes*), 2n = 36 (plus a small number of B chromosomes), consists almost solely of metacentric chromosomes, and the remaining members of the family exhibit a relatively wide range of chromosome number and morphology. Comparison of banded chromosomes of the Canidae has enabled their

probable evolutionary history to be recreated (Wayne *et al.*, 1987a; Wayne *et al.*, 1987b), suggesting the primitive ancestral karyotype represented a small (approximately 2n = 40), mainly metacentric chromosome set. It is proposed that subsequent chromosome fission would have then occurred to generate the high chromosome numbers seen in species like the Grey wolf and the domestic dog, accompanied by further rearrangements specific to one or more species. Individual domestic dog breeds are likely to have developed by selective breeding within the last few hundred years (Wayne and Ostrander, 1999), all of which can potentially interbreed.

The G-banded karyotypes of the Grey wolf and the domestic dog appear to be entirely homologous (Wurster-Hill and Centerwall, 1982; Wayne *et al.*, 1987b). It is thought most likely that the domestic dog originated from a wolf progenitor, and recent estimates place the timing of this event as in excess of 100,000 years ago (Morrell, 1997). This estimation was based on comparative sequence analysis of mitochondrial DNA isolated from wolves and other wild canids from a range of geographical locations, and from more than 60 domestic dog breeds. Fewer nucleotide differences were found between the wolf and the domestic dog within this mitochondrial sequence than between the domestic dog and other wild canids. The development of the domestic dog from a wolf ancestor is reported to have resulted from two separate domestication events some 135,000 years ago, although the identity of the wolf progenitor is not clear (Vilà *et al.*, 1997).

1.5 Towards a genome map for the domestic dog

1.5.1 The dog as a model for human disease

The rationale behind the generation of a genome map for the domestic dog is strongly directed towards its value as a model system for the study of heritable traits, and in particular, towards a greater understanding of the aetiology of inherited disease. Canine homologues of several human disorders have been described (reviewed in Patterson *et al.*, 1982; O'Brien, 1986; Ostrander *et al.*, 2000). Amongst the first to be studied at the gene level was canine X-linked Severe Combined Immunodeficiency Disease (X-SCID) (Henthorn *et al.*, 1994), a disease that confers clinical, pathological and immunological symptoms highly similar to the human form of the disease. Human X-SCID had previously been shown to be caused by mutations in the γ-chain of the interleukin-2 receptor gene, mapped to HSA Xq13 (Noguchi *et al.*, 1993). Meiotic linkage analysis has been used to map the causal gene to the proximal region of chromosome Xq in the dog. Further analysis showed the presence of deletions in the protein-coding region of the

same gene in affected dogs, which prevented the production of a functional gene product (Henthorn *et al.*, 1994). This provided evidence that canine X-SCID is a true homologue of the human disease at the gene level. Thus the canine system may be a useful model for investigation of the precise mechanism of the disease in both species, and potentially for assessment of future methods of treatment. Recently, the gene mutation responsible for the sleep disorder narcolepsy has been identified in the Doberman pinscher (Lin *et al.*, 1999). The physiological effects of the autosomal recessive canine disorder have been shown to closely resemble those of human narcolepsy (for example, Nishino and Mignot, 1997). The mutation was localised to a region of conserved synteny with HSA 6 by meiotic linkage mapping. Only a single previously annotated gene, HCRT2, had been located in this human chromosome region, with no previously established relationship to narcolepsy. Detailed analysis demonstrated disruption of HCRT2 in affected dogs, by the insertion of a SINE, providing a potential candidate gene that is under investigation with respect to human narcolepsy (Lin *et al.*, 1999). A dog model for human Duchenne muscular dystrophy, caused by a deletion in Xp21 in both species, has also recently been described (Schatzberg *et al.*, 1999), whilst studies on dog models of human cancers, heart diseases and epilepsy are in progress (M. Binns, pers. comm.).

A number of genetic diseases appear to be specific to one or a small number of dog breeds. The identification of a microsatellite linked to disease phenotype has led to the development of a diagnostic test for copper toxicosis (CT) in Bedlington terriers (Yuzbasiyan-Gurkan *et al.*, 1997). CT results in a toxic accumulation of copper in the liver, and shares similarity with a human disorder, Wilson's disease. It is intended that this linked marker be used as the basis for a diagnostic test for CT in the Bedlington terrier in which diagnosis is difficult prior to one year of age, and carriers may only be identified when affected offspring are produced. This has particular significance in such late-onset disorders, for which breeding from carrier individuals may have occurred prior to development of symptoms. Pre-symptomatic testing therefore can provide breeders with the necessary information for developing future breeding programs that will limit the propagation of the disease throughout the breed. More recently, a linked marker for CT, identified by Yuzbasiyan-Gurkan *et al.* (1997) has been mapped by van de Sluis *et al.* (1999) to CFA 10, a region sharing conserved synteny with HSA 2p16-p13. However, no suitable candidate gene has been identified in this location. It was also demonstrated that the dog homologue of the human Wilson's disease gene maps to CFA 22, and thus

cannot be the cause of CT in the dog. The identity of the causative gene defect therefore remains to be established.

The dog is considered to develop cancer at twice the rate of humans, and the presentation of the disease in both species has been shown to share many common features (reviewed in Ostrander *et al.*, 2000). Certain breeds appear to be more susceptible to cancers, and conversely, a subset of tumours have been shown to be largely limited to particular breeds (MacEwen, 1990, M. Brearley, pers. comm.). The identification of genetic differences between breeds may therefore generate potential explanations for this observation that could also have implications for human neoplasia. The development of cancer is likely to be a complex process involving multiple loci, and as such the study of its genetic basis is likely to be significantly more complex than that required for single-locus traits. The production of an integrated genetic and physical map of the dog genome will be a major advantage in the analysis of such complex traits, and has already been successfully used to study a number of single-gene defects.

1.5.2 The domestic dog karyotype

The karyotype of *Canis familiaris* has often been described as one of the most challenging amongst the animal kingdom. This in part reflects the high diploid chromosome number (2n = 78), and the similar morphology and banding patterns of many of the autosomes, all of which are acrocentric. The first attempt to describe the dog karyotype by banding analysis was made by Selden *et al.* (1975) using the GTG-banding technique. This was followed by a number of studies using various banding techniques (Manolache *et al.*, 1976; Howard-Peebles and Pryor, 1980, Mayr *et al.*, 1983; Poulson *et al.*, 1990; Stone *et al.*, 1991; Graphodatsky *et al.*, 1995, and others). The difficulty in reliable identification of the smaller autosomes was a common issue, and the inevitable discrepancies in nomenclature between these reports led to difficulties in the development of a standard karyotype to which the dog genome mapping community could refer.

The issue of an internationally accepted dog karyotype was revisited by Reimann *et al.* (1996). Centromere location was confirmed by FISH analysis using a dog-specific alpha-satellite probe previously described by Fanning (1989), enabling the identification of errors in the orientation of three autosomes within the karyotypes of Selden *et al.* (1975) and Graphodatsky *et al.* (1995). Significantly, the 460-band resolution GTG-banded karyotype proposed by Reimann *et al.* (1996) was derived from a single metaphase

spread, such that the size of all chromosomes could be internally compared. However, the problem of identification of the smallest autosomes was reiterated, with the suggestion that higher resolution banding analysis or the use of chromosome-specific probes might alleviate the difficulties.

These points were also discussed in the first report from the Committee for the Standardisation of the Dog Karyotype (Switonski *et al.*, 1996). It was decided by the Committee to combine the use of standard banding techniques with molecular cytogenetic methods, including chromosome painting analysis and the mapping of single locus probes, towards the production of an internationally accepted dog karyotype. This was to retain the most widely accepted nomenclature of Reimann *et al.* (1996), following Selden *et al.* (1975), where possible. The first report from the Committee described the initial phase of this process, an investigation of the number of chromosome pairs that could be definitively identified solely with the use of G-banding, which would allow the banding patterns of these chromosomes to be finalised. This was achieved for the largest 21 autosomes and both sex chromosomes, for which G-banded ideograms were presented, comprising a total of 235 bands. The recent development of chromosome-specific paint probes (Langford *et al.*, 1996) and single locus probes (Fischer *et al.*, 1996) for the dog was proposed to be a focal point in the next stage in the standardisation of the remaining 17 autosomes.

This approach was recently applied by the Committee for the Standardisation of the Dog Karyotype, using the chromosome paint probes developed by Langford *et al.* (1996). Paint probes representing the 17 remaining unclassified autosomes were provided to members of the committee who applied them to dog metaphases allowing banding patterns to be compared between laboratories. This resulted in the successful description of these remaining chromosome pairs (Breen *et al.*, 1998). This study was extended by Breen *et al.* (1999b), in which 30 metaphases spreads were sequentially probed with chromosome-specific paint probes for each of the 38 autosomes and two sex chromosomes. Subsequent FISH analysis allowed each chromosome within a single spread to be identified unequivocally. Although eight paint probes are derived from material from two chromosome pairs, their banding patterns were described as being sufficiently different to allow them to be distinguished from each other with complete confidence. The first, complete DAPI-banded karyotype and ideogram for the dog were presented, reflecting the increasing use of this dye in conjunction with FISH analysis, with a resolution level of 460 bands. Again, the most commonly used nomenclature

originally described by Reimann *et al.* (1996), following Selden *et al.* (1975), was followed for the purposes of continuity. However, the authors note that this nomenclature does not reflect the relative size of each chromosome, which was previously estimated by Langford *et al.* (1996) through comparison of the human and dog bivariate flow karyotypes.

1.5.3 Linkage maps of the dog genome

The current dog genome map consists predominantly of microsatellite markers, several hundred of which have been described (for example, Holmes *et al.*, 1993; Ostrander *et al.*, 1993; Holmes *et al.*, 1995; Ostrander *et al.*, 1995). A small number of dog genes have been found to be associated with microsatellite repeats, including factor IX (Gu and Ray, 1997) and the androgen receptor gene (Shibuya *et al.*, 1993). Until recently, anonymous microsatellite markers have been the main resources used for the development of a dog genome map. The DogMap consortium was initiated in 1993 for the main purpose of generating a low-resolution physical and linkage map of the dog genome, towards the recognition of the loci responsible for inherited disease and other genetic traits. The first results, from the genotyping of 94 polymorphic markers on a common reference family, were published in 1997 (Lingaas *et al.*, 1997). The reference family consists of 129 purebred individuals, of which 42 are German shepherd dogs and the remainder beagles. Forty three of the 94 markers, mainly anonymous microsatellites, formed a total of 16 linkage groups. Two groups could be assigned to a specific chromosome since they contained a marker that had been assigned by FISH analysis, and a third to an unidentified small autosome. A further 20 markers, generating seven linkage groups, were added by Ryder *et al.* (1999).

A separate dog linkage mapping effort has been described by Mellersh *et al*, (1997). One hundred and fifty further microsatellite markers were typed on a reference family consisting of 17 three-generation pedigrees of five breeds, comprising a total of 212 individuals. One hundred and thirty nine markers formed 30 linkage groups, but were not assigned to chromosomes. Neff *et al.* (1999) extended the map by the addition of 126 microsatellite markers, resulting in 39 linkage groups. The current linkage map consists of 341 microsatellite markers, generating 37 autosomal linkage groups, of which 15 were assigned to a specific dog chromosome, with one chromosome being represented by two separate linkage groups. This map has recently been integrated with a WG-RH map of

the dog genome (Priat *et al.*, 1998) by Mellersh *et al.* (2000), which is discussed in chapter five.

1.5.4 Resources for dog gene mapping

Resources have been developed by which the number of mapped dog genes can be increased. These largely resemble the CATS primer set (Lyons *et al.*, 1997), for which the success rate for the dog was reportedly high, with single PCR products from 45% of the 318 primer pairs. Venta *et al.* (1996) aligned sequence data from two or more species to create 86 pairs of PCR primers for application to mammalian gene mapping. These primer pairs were described as 'gene-specific universal mammalian sequence-tagged sites', or 'UM-STSs'. In each case, the intron/exon boundary was known in at least one of the species used in the alignment analysis, and primers were typically designed such that the resulting PCR product contained a region of intronic sequence. Products amplified from the dog were sequenced to ensure similarity to the expected gene. A small number of UM-STS were also used to isolate corresponding cosmid clones, which were assigned to dog chromosomes by FISH analysis (Guevara-Fujita *et al.*, 1996). Eleven UM-STSs were also applied to other mammalian species, demonstrating an overall success rate of 84%, under the same conditions used for amplification of the dog product. A similar approach was used by Jiang *et al.* (1998) in the development of 225 primer pairs, corresponding to 146 different genes. Gene sequences from two or more different vertebrate species were aligned, of which at least one was a genomic sequence. Primers were then designed to regions of high nucleotide conservation, flanking at least 100bp of coding sequence and in some cases, also some intronic sequence. The primers were used to attempt amplification from human, pig, dog and hamster, and the resulting products were described as 'traced orthologous amplified sequence tags', or 'TOASTs'. One hundred and twenty six (56%) of the 225 primer pairs successfully amplified a single product from dog genomic DNA, which were sequenced to determine amplification of the expected sequence. Interestingly, where amplification within a single exon was performed, the size of the resulting fragment appeared identical for both dog and human products in 97.7% of cases. This figure fell to 90.7% for dog and hamster comparisons. Where intronic regions were present in the amplified product, 61.9% of human and dog products were of a comparable size, compared to 36.4% of dog and hamster products. This naturally has implications for the use of the markers in radiation hybrid mapping. The size difference may negate the need to exclude from analysis any marker for which

rodent amplification occurs when the corresponding PCR primers are applied to DNA from a hybrid cell line, since it may be possible to distinguish between the amplification products from the two species.

This factor was further investigated in a follow-up study where 139 TOASTs were amongst 463 markers assessed for their suitability in radiation hybrid mapping of the dog genome by Priat *et al.* (1999). In a further 27 cases, where the TOAST primer pair amplified a product of comparable size in both hamster and dog, a new primer pair was designed from the characterised dog sequence. These markers were described as 'canine TOASTs'. The remaining markers included 221 primer pairs designed to 204 of the then 250 previously characterised dog genes deposited in sequence databases. It was clearly shown that knowledge of the intron/exon structure of the sequence increased the overall success rate of such a strategy. Where such information was not known, there was a risk of designing primers either spanning an intron/exon boundary, or of including a large intronic region which prevented efficient amplification of the product. Three hundred human EST markers were also used to attempt to amplify the orthologous dog gene, of which 76 (25%) were successful. The identity of each was confirmed by partial sequencing and a subsequent database similarity search.

Markers were tested for their ability to produce a single PCR product from dog genomic DNA in the absence of amplification from the hamster background of the radiation hybrid panel. One hundred and ninety (86%) of the 221 primer pairs designed from dog genes were successful, compared to 48 of the 139 TOASTs (34.5%) and 37 of the 76 human ESTs (49%). However, subsequent analysis across the RH panel encountered difficulties with the analysis of some markers, such that a total of 208 of the 463 markers (45%) were successfully mapped on a dog radiation hybrid panel.

1.5.5 Hybrid cell-line panels and maps

A dog-rodent somatic cell hybrid panel has been constructed by Langston *et al.* (1997). FISH analysis, using total dog genomic DNA as the probe, was used to establish the number of dog chromosomes present in each of the hybrid cell-lines generated. Hybrids containing highly fragmented dog chromosomes were eliminated. PCR analysis of microsatellite markers was performed for the remaining 43 cell lines, in order to estimate the genome coverage of the panel. The inability to assign 12% of markers tested to a linkage group indicated that the panel did not demonstrate complete coverage of dog

chromosomes. Whilst the panel may prove useful for analysis of certain chromosomes or marker sets, it has not been widely used to date.

Two dog WG-RH panels exist, the first of which was developed by McCarthy and colleagues at the University of Cambridge Genetics Department, UK. Female dog fibroblasts were irradiated with 3000 rads of X-rays prior to fusion with a thymidine kinase-deficient hamster cell line. The resulting 'T72' panel of 92 hybrids was estimated to have an average retention rate of 24.5% of the dog genome. Twenty six markers, previously assigned to at least five different dog chromosomes, were mapped in duplicate following construction of the panel, focussing on the X chromosome (McCarthy *et al.*, unpublished). The T72 panel is now commercially available through Research Genetics (Huntsville, Alabama).

The $RHDF_{5000}$ radiation hybrid panel developed by Vignaux *et al.* (1999a) originally comprised 126 cell lines created by the fragmentation of male dog fibroblasts with 5000 rads of gamma rays, and subsequent fusion with a thymidine kinase-deficient hamster cell line. Using this panel, a preliminary whole genome radiation hybrid map has been produced for the dog, comprising 218 type I and 182 type II markers (Priat *et al.*, 1998). These were mapped to a total of 57 distinct linkage groups with a lod score of six or greater, each containing between two and twenty one markers. Fourteen linkage groups were assigned to nine different dog chromosomes on the basis of previous studies. Fifty three markers were not linked to any group. The average retention frequency for all 400 markers analysed was calculated as 21%, and the map was estimated to provide 80% coverage of the dog genome. The mapping of 121 markers previously analysed by meiotic linkage enabled a degree of integration between this and previously developed maps. Additionally, regions of conserved synteny were proposed through mapping of non-polymorphic coding sequences and comparison of the location of the orthologue in other species, predominantly human and mouse. More recently, efforts have been made to reduce the number of clones within the hybrid panel, the aim being to increase the efficiency of typing without significant loss of resolution. The panel has now been reduced to 111 clones by elimination of those clones containing less than 6% of the dog genome, which was found to have little effect on the mapping power of the resource (Vignaux *et al.*, 1999a). The reduced panel has been made available to the dog genome mapping community.

1.5.6 Applications of the dog BAC library

The generation of a dog BAC library has proved to be an enormous advantage for the mapping of single locus probes to specific chromosomes. The RPCI81 dog BAC library was constructed by Li *et al.* (1999) from male Doberman pinscher genomic DNA, cloned into the *EcoR*I site of the pBACe3.6 vector, based on the original BAC vector of Shizuya *et al.* (1992). Almost 160,000 recombinant clones were generated, each containing an average of 155kb of insert DNA representing an overall eightfold coverage of the genome. Clones have been gridded at high density onto nine nylon hybridisation filters by the UK Human Genome Mapping Resource Centre (Hinxton Hall, Cambridge, UK), in a manner that aids interpretation of hybridisation signals. Groups of eight clones are gridded in a four by four array, in which each clone is represented twice. The duplication facilitates the identification of false positive signals since a clone may only be considered a candidate for further study in the presence of hybridisation signals corresponding to both duplicates. The relative position of the pair of duplicates within each grid of 16 is unique to that clone. A combination of this information and the co-ordinates of the four-by-four grid provides each clone with a unique identification code. The BAC library has proved invaluable for the isolation and physical mapping of clones isolated by screening of the library with smaller probes that cannot themselves be mapped by FISH analysis.

1.5.7 Cytogenetic resources and results for the dog

Fischer *et al.* (1996) reported the first use of FISH analysis for physical mapping in the dog. Twelve anonymous, microsatellite-containing cosmid clones were assigned, and a further cosmid hybridised to the centromere of 18 autosomes. Early assignment of dog genes focussed on the X chromosome, for which loci have been mapped by Puck *et al.* (1993) and Deschenes *et al.* (1994) in the study of an X-linked disease of the dog. The first autosomal genes to be assigned to dog chromosomes by FISH analysis were reported by Guevara-Fujita *et al.* (1996). Six loci were assigned, using the G-banded karyotype nomenclature proposed by Stone *et al.* (1991), in addition to a further locus from the X chromosome. Four more genes, two autosomal and two X-encoded, were subsequently mapped using FISH analysis by Dutra *et al.* (1996).

A complete set of chromosome paints for the dog have been developed by Langford *et al.* (1996). A total of 32 peaks were identified by bivariate flow cytometry of peripheral blood lymphocyte cultures. Between 500 and 1000 sorted chromosomes from each flow peak were amplified and labelled by DOP-PCR (Telenius *et al.*, 1992a), and hybridised

to dog metaphases. Probes from eight peaks each hybridised to two chromosome pairs that could not be fully resolved by flow cytometry. All other probes painted single chromosome pairs, representing the remaining 22 autosomes and both sex chromosomes. A further peak, specific for a single chromosome pair, was identified in one individual. This enabled the production of a single paint probe for this chromosome, which co-sorted with another autosome in all other individuals studied. Probes corresponding to CFA 1-21, for which standardised nomenclature had been described (Switonski *et al.*, 1996), were assigned to the bivariate flow karyotype. A mixture of pooled dog and human chromosomes was sorted by the same method. The relative position of peaks in the resulting flow karyotype was used to estimate the physical size of the corresponding dog chromosome, by comparison with existing data for human chromosomes.

1.6 Project aims

A large panel of resources is available for the study of the dog genome, and in particular for the development of a map to enable identification of the loci responsible for inherited traits of interest. This applies in particular to disease genes, and those associated with the valued traits of a given breed. The overall aim of this project was the development and application of methods by which the gene map of the dog could be improved. Initially, this was to involve an investigation into the potential for achieving this goal by generating random ESTs for the dog genome. This would extend methods developed for the human genome to establish the nature of the dog EST, and to then assess means by which these sequences could be added to the dog genome map. The suitability of the dog cDNA library used for this purpose would be assessed in terms of the quality of EST data generated, compared to that from previous studies in other species. The value of the EST approach for the dog would then be evaluated, and if considered appropriate would be extended to larger scale isolation, characterisation and mapping of dog ESTs. Comparison of the chromosomal location of orthologous genes in the dog and human genomes would enable putative regions of conserved synteny to be identified. This would enable the dog to be included in the growing list of species for which comparative mapping data are available, and would also provide resources for future study of loci of specific interest to the dog genome mapping community.

During this period, relocation of the laboratory to the Animal Health Trust provided access to resources that allowed a contrasting approach to be investigated. The availability of a complete set of both dog and human chromosome paint probes enabled

the investigation of the reciprocal Zoo-FISH technique for the identification of regions of conserved synteny between their genomes. This was to be investigated by selecting a single dog chromosome for which experimental conditions could be optimised. Reciprocal chromosome painting analysis would be used to identify ECCS between this dog chromosome and the corresponding human chromosomes. The boundaries of each ECCS would then be further investigated through the chromosomal localisation of dog genes selected from within the corresponding human ECCS. Existing publications describing PCR primers for amplification of type I markers would be used to select loci from the appropriate human chromosome region, for which a dog PCR product could be generated. The availability of a dog BAC library at this time would expedite the mapping of the orthologous dog gene, since the PCR product could be used as a probe to isolate a large insert BAC clone that could be readily assigned by metaphase FISH analysis. Similarly, a whole genome radiation hybrid panel for the dog would be used to investigate the potential for generating a preliminary RH map of the selected dog chromosome. Gene markers mapped to the dog chromosome would be assessed for their suitability as RH markers, and mapping data would be integrated with those derived from similar mapping of a panel of anonymous cosmid clones previously assigned to the same chromosome. Where possible, the relative order of genes would be compared between the dog chromosome and the corresponding human ECCS. This approach would result in the generation of mapping information for a dog chromosome by the development and application of methods that could be equally extended to the entire genome.

The overall contribution to the dog genome map made by each of the techniques developed and applied would be evaluated and integrated with existing data. Improvements to these techniques, and potential areas of future interest arising from this study, would then be identified and discussed.

Chapter two:

Materials and methods

Materials and Methods

The composition of common reagents is given in appendix A, and PCR primer sequences are listed in Appendix B. Centrifugation was performed at maximum speed (16000*g*) at room temperature in an MSE microcentrifuge, unless otherwise stated. Chromosomal locations of human gene loci were derived from the Human Gene Nomenclature Database (HGND, http://www.gene.ucl.ac.uk/nomenclature/) unless otherwise stated.

2.1 General methods

2.1.1 Preparation of genomic DNA

Ten millilitres of blood from normal healthy individuals were collected into 15ml polypropylene tubes (Sterilin) containing a final concentration of 2mg/ml EDTA. Genomic DNA was extracted using a Nucleon DNA Extraction and Purification kit (Scotlab). Four volumes of reagent A were added to each sample, and the contents were rotary mixed for four minutes at room temperature. The contents were pelleted by centrifugation at 1300*g* for five minutes in an MSE Centaur centrifuge. The supernatant was discarded and the pellet was resuspended in 2ml of reagent B. The contents were transferred to a 15ml polypropylene tube (Sterilin) and 0.5ml 5M sodium perchlorate were added, mixing by inverting. Two millilitres of chloroform were added, and mixed by inverting. The Nucleon resin was resuspended fully and 300µl were added to the sample preparation, followed by a centrifugation step at 1300*g* for three minutes in an MSE Centaur centrifuge. The upper layer was transferred to a second 15ml tube. Two volumes of ice cold absolute ethanol (AnalaR grade, BDH) were added, and the contents were mixed by inverting until the DNA was observed to precipitate. Where possible, the precipitate was removed by spooling on a sterile glass Pasteur pipette, the end of which had been sealed by heating in a flame to create a loop. The DNA was briefly immersed in a 1.5ml Eppendorf tube filled with 70% (v/v) ethanol, and then transferred to one millilitre of MQ water (appendix A), or a smaller volume if a low yield was expected. If the DNA could not be collected in this manner, the contents were pelleted by centrifugation at 5000*g* for five minutes. Two millilitres of 70% (v/v) ethanol were added, mixed by inverting, and the DNA pelleted as before. The supernatant was removed and the pellet was allowed to air-dry prior to resuspension in a suitable volume of MQ water in a 1.5ml Eppendorf tube.

The concentration of the genomic DNA sample was estimated using a Pharmacia Ultrospec 2000 spectrophotometer with a quartz cuvette (Sigma), assuming that an A_{260} value of 1.0 corresponds to 50μg/ml of double-stranded DNA, using MQ water as the solvent. DNA concentration was also determined on some occasions by agarose gel electrophoresis (section 2.1.5) and comparison with control samples of known concentration (Lambda DNA Quantitation standards, Gibco BRL). DNA preparations were diluted to 25ng/μl for use in the PCR, and were stored at -20°C.

2.1.2 Sonication of genomic DNA

Prior to sonication, high molecular weight DNA preparations were diluted to 10mg/ml in a 15ml polypropylene tube. The sample was placed on ice and subjected to eight repetitions consisting of 15 second pulses (50% power output level) using an Ultrasonics W225 sonicator. The size range of the resulting fragments was assessed by agarose gel electrophoresis (section 2.1.5), compared to a 1kb DNA ladder (Gibco BRL). The sonication process was repeated as required until the range of fragment sizes fell between approximately 0.2kb and 1.5kb.

2.1.3 DNA preparation from buccal swabs

Buccal swabs were stored at -20°C prior to isolation of genomic DNA. Two hundred microlitres of 50mM NaOH were placed in a 0.5ml polypropylene tube. The swab was inserted into the tube, ensuring that the entire tip was submerged. After agitation of the swab to encourage release of the cells into the liquid phase, the tube and swab were incubated at 97°C for 10 minutes in a heat block. The swab was then discarded, and the contents of the tube were neutralised by the addition of 70μl of 1M Tris HCl pH8. Samples were stored at -20°C prior to use. Since this simple method yields relatively poor quality DNA, quantification of the preparation was not performed, and samples were used directly in further analysis with no additional treatment.

2.1.4 PCR reaction composition and cycling conditions

Reactions described in chapter three were performed in 25μl volumes in 0.5ml polypropylene tubes, consisting of 2.25μl LEI-PCR buffer (appendix A), 50ng each primer as required (appendix B), 1U *Taq* polymerase (ICN) or 'Red-Hot' *Taq* polymerase (Advanced Biotechnologies), 1μl of template DNA as required and MQ

water up to 25µl. Reactions were overlaid with 30µl of light mineral oil (Sigma), and subjected to an initial denaturation period of 5 minutes at 94°C. This was followed by 30 cycles comprising 60 seconds at 94°C, 60 seconds at the required annealing temperature, and 60 seconds at 72°C, with a final incubation at 4°C, in a Perkin Elmer DNA thermal cycler.

Unless otherwise stated, reactions in subsequent chapters were performed in 10µl volumes in 0.2ml tubes (Robbins Scientific) using a PTC-225 thermal cycler (MJ Research), in the absence of mineral oil. Optimisation of PCR conditions was performed by variation of annealing temperature and elongation time, and by the use of reaction buffers of varying $MgCl_2$ concentration (Perkin Elmer 10x PCR buffer II supplemented with 1mM to 4mM $MgCl_2$ as required, see appendix A). Reactions consisted of 25ng of each primer, 2nmol of each dNTP (Pharmacia), 1U AmpliTaq Gold (Perkin Elmer), 2µl of the appropriate reaction buffer and 25ng of dog genomic DNA (or other template as appropriate). Unless otherwise stated, cycling commenced with a 25 minute step at 95°C in order to activate the enzyme, followed by 30 cycles of 60 seconds at 94°, 60 seconds at the optimised annealing temperature as appropriate, and 60 seconds at 72°C, completing with a final extension period of five minutes at 72°C.

2.1.5 Agarose gel electrophoresis

PCR products were analysed by electrophoresis at 100 volts on 2% (w/v) agarose gels (Gibco BRL) in 1x TAE running buffer, using 6x Orange G (Sigma) gel loading buffer (appendix A). High molecular weight DNA preparations were analysed on 1% (w/v) agarose gels. Gels were stained subsequent to electrophoresis for 10 minutes in 0.5µg/ml ethidium bromide (Gibco BRL) diluted in running buffer. Stained gels were visualised on an ultraviolet transilluminator (Ultraviolet Products), and were photographed using a video camera and video copy processor (Mitsubishi).

2.1.6 Purification of amplified DNA

Subsequent to the PCR, amplification products were purified using a Qiagen PCR Purification Kit. Samples were transferred to 1.5ml Eppendorf tubes, and five volumes of buffer PB were added. The contents were mixed by vortexing and collected by pulse-spinning in a microcentrifuge. A QIAquick column was placed in the 2ml tube provided, and the sample was loaded onto the column matrix. The column was spun at maximum

speed in a microcentrifuge for one minute, and the flow-through discarded. The DNA was washed by the addition of 0.75ml buffer PE to the column, followed by a further one minute spin. The flow through was discarded and the remaining wash buffer was removed by a third centrifugation step of one minute.

The column was transferred to a sterile 1.5ml Eppendorf tube and 30µl buffer EB (10mM Tris HCl pH8.5), prewarmed to 65°C, were pipetted onto the centre of the column matrix. After a one minute incubation at room temperature, the column was spun at maximum speed for one minute to elute the DNA. Purified products were stored at -20°C. One microlitre of the sample was analysed by electrophoresis on a 2% (w/v) agarose gel. The concentration of the resulting DNA solution was estimated by comparison with Lambda DNA quantitation standards (Gibco BRL).

Purification of PCR products excised from agarose gels was performed using the Qiagen Gel Extraction kit following the above protocol with the following exceptions. The excised gel slice was weighed and 300µl of buffer QX1 were added for each 100mg of agarose. The preparation was incubated at 50°C for ten minutes, with occasional inversion. One hundred microlitres of isopropanol were added for each 100mg of agarose. This sample was loaded onto a QIAquick spin column, which was then processed as described above.

2.1.7 Cycle sequencing

For reactions described in chapter three, approximately 100ng of insert DNA were added to 8µl of dRhodamine dye terminator mix (Perkin Elmer) with 10ng of the required primer (appendix B), in a total volume of 20µl. Reactions were overlaid with 30µl of light mineral oil and subjected to the PCR (25 cycles of 96°C x 30 seconds, 50°C x 15 seconds and 60°C x four minutes) in a Perkin Elmer DNA Thermal Cycler. The product was removed from beneath the oil and was purified by ethanol precipitation. Each sample was transferred to a 0.5ml Eppendorf tube, and 50µl of absolute ethanol and 2µl of 3M sodium acetate pH5.2 were added. This was incubated on ice for 10 minutes, followed by centrifugation at maximum speed for 20 minutes. The supernatant was removed and 250µl of 80% (v/v) ethanol were added. The tube was agitated and the contents were pelleted by a further 15 minute period of centrifugation. The supernatant was removed with a drawn-out glass Pasteur pipette and the pellet was air-dried and stored at -20°C until required. Sequencing analysis was performed according to the manufacturer's

instructions, on 4% (w/v) polyacrylamide gels using an ABI 377 automated sequencer by the PNACL sequencing facility, University of Leicester.

For reactions described in subsequent chapters, DNA sequencing was performed as described above except that all volumes and concentrations were halved, and cycle sequencing was performed in a PTC-225 thermal cycler (MJ Research) in the absence of mineral oil. Sequencing gels were prepared by dissolving 18g of urea (Amresco) in 5.3ml of 40% 19:1 bis:acrylamide (Amresco) made up to a total volume of 50ml with MQ water. After ensuring complete dissolution, 0.5g of Amberlite beads (Perkin Elmer) were added and the solution was mixed for five minutes using a magnetic stirrer. Five millilitres of 10x TBE were passed through a 0.2μm filter (Nalgene) under a vacuum, and the acrylamide solution was then filtered in the same manner into the same collection vessel. Application of the vacuum was continued for five minutes to eliminate air bubbles from the solution, which was then transferred to a fresh 50ml polypropylene tube. Polymerisation was initiated by the addition of 250μl of 10% (w/v) ammonium persulphate (Amresco) and 35μl TEMED (Sigma), and the gel was cast following the manufacturers instructions, using a 36-well shark's tooth comb (Perkin Elmer). The gel was allowed to fully polymerise for two hours before use, and was then prewarmed to 51°C in 1x TBE running buffer. Sequencing analysis was performed following the manufacturer's recommendations. Immediately prior to electrophoresis, 4μl of acrylamide gel loading buffer (appendix A) were added to each sample and mixed well to ensure complete resuspension of the DNA pellet. Samples were denatured for 2 minutes at 95°C, maintained on ice and 2μl were loaded onto the 4% polyacrylamide gel. Electrophoresis was carried out in 1x TBE running buffer at for seven hours at 51°C using the 'Seq. Run 36E-1200' run module (1680volts, 50milliamps, 150watts). Analysis of data was performed using the Sequence Analysis 2.1.1 package (Perkin Elmer) in accordance with manufacturer's instructions.

2.1.8 Editing of sequence data

The FACTURA 2.0 sequence analysis program (Perkin Elmer) was used to edit data prior to further analysis. Sample files were imported into the program and analysed using default settings. All sample files were edited to remove data containing 5% or more ambiguous bases over a twenty base window. Where vector primers were used in the preparation of the sequencing sample, FACTURA was used to identify and remove any

vector and primer sequence from the data obtained. Edited data were saved as text files in FASTA format for use in database similarity searches.

2.1.9 Database similarity searches

Sequence data were searched against existing database entries using BLAST version 2 ('Gapped BLAST') hosted by NCBI (http://www.ncbi.nlm.nih.gov/blast/blast.cgi) with default parameters. Data from chapter three were searched against both the non-redundant (nr) and EST (dbEST) databases, whilst sequences described in subsequent chapters were searched only against the non-redundant database. The non-redundant database contains all GenBank entries (except ESTs) plus those from the DataBank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL). dbEST contains EST sequences from the same three databases. In all cases, sequence text files were input in FASTA format, and the number of output results and alignments were limited to ten.

2.1.10 Primer preparation

Unless otherwise stated, PCR primers were synthesised by GenSet or Pharmacia. Stocks were stored at -20°C and were diluted to 50ng/µl in MQ water for use in the PCR.

2.1.11 Genotyping analysis of polymorphic markers

PCR amplification was performed as described previously in section 2.1.4 except that 0.05µl of the Rhodamine-110 fluorescent dUTP (3nmol, Perkin Elmer) was added to each 10µl reaction containing 25ng of the appropriate genomic template DNA (appendices D and E). Unincorporated fluorescent nucleotides were removed by the addition of 10µl of 1:1 phenol:chloroform mix (Perkin Elmer) to the reaction, followed by vigorous mixing and centrifugation for four minutes at 4000rpm in an Eppendorf 5804 centrifuge.

Genotyping gels were prepared by combining 4ml of Long Ranger acrylamide solution (Flowgen), 4ml of 10x TBE and 14.4g of urea (Amresco), in a total volume of 40ml with MQ water. After ensuring complete dissolution of urea, 28µl of TEMED (Sigma) and 200µl of 10% (w/v) ammonium persulphate (Amresco) were added, and the gel was cast following manufacturer's recommendations, using a 34-well square-tooth comb. The gel

was allowed to fully polymerise for two hours prior to use, and was then pre-warmed to 51°C in 1x TBE running buffer.

One microlitre of each fluorescently-labelled PCR product was added to 3µl of acrylamide gel loading buffer (appendix A) and 0.5µL ROX350 size standard (Applied Biosystems). Samples were denatured at 95°C for 2 minutes, and maintained on ice. Two microlitres of each sample were loaded onto the 4.25% polyacrylamide gel, and electrophoresis was performed in 1x TBE running buffer for two hours according to the manufacturer's instructions, using run module 'GS Run 36C-2400' (3000volts, 60milliamps, 200watts). Data were analysed using Genescan version 2.0.2 (Perkin Elmer) and Genotyper version 1.1.1 (Perkin Elmer) in accordance with manufacturer's recommendations. Linkage analysis was performed using the 'TWOPOINT", 'BUILD' and 'FLIPS' options of CRI-MAP version 2.4 (Lander and Green, 1987) under default parameters, except PUK_LIKE_TOL and PK_LIKE_TOL (which relate to the tolerance limits for discarding potential locus orders) were reduced from 3.0 to 2.0. Existing meiotic linkage data for CFA 5 markers were provided by C. Mellersh (Fred Hutchinson Cancer Research Center, Seattle) or were generated internally (N. Holmes, pers. comm.) TWOPOINT was used to determine pairwise linkage between markers. Existing locus order data were entered into the BUILD option and novel markers were then added to identify their most likely position within the existing map. The FLIPS option was used to swap the order of loci in pairs to identify any potentially improved orders.

2.2 Fluorescence *in situ* hybridisation (FISH) analysis

2.2.1 Lymphocyte culture for slide preparation

Healthy dogs with an apparently normal karyotype were bled from the jugular vein and 10ml of blood were collected in RPMI-1640 medium (Sigma) in 15ml polypropylene tubes in the presence of a final concentration of 15U/µl sodium heparin (C.P. Pharmaceuticals). Cultures were prepared in 125ml Ehrlenmeyer culture flasks (Corning), consisting of RPMI-1640 medium (Sigma) supplemented with 20% foetal calf serum (Gibco BRL), 2% (v/v) of 200mM L-glutamine (Sigma), and 1% (v/v) penicillin/streptomycin (10,000u/ml penicillin and 10mg/ml streptomycin in 0.9% (w/v) NaCl, Sigma), to which 2ml of fresh whole blood were added and gently mixed. To each culture, pokeweed mitogen (Gibco BRL) and phytohaemagglutanin (Murex) were added

at a final concentration of 1% (v/v) each. Cultures were incubated at 38°C, 5% CO_2 for 72 hours, with occasional gentle mixing.

Colcemid (Karyomax, Gibco BRL) was added to a final concentration of 50ng/ml one hour prior to harvesting in order to arrest cells at metaphase. Cultures were divided into two equal aliquots in 15ml polypropylene tubes (Sterilin), and spun at 200*g* for 10 minutes in an MSE Centaur centrifuge. The supernatant was carefully aspirated and the cell pellet was resuspended in fresh hypotonic solution (75mM KCl) maintained at 37°C to cause swelling of the cells. This solution was added dropwise up to a volume of approximately 5ml with agitation, and then added more rapidly to a total of 10ml. The resulting cell suspension was incubated at 37°C for 20 minutes, prior to centrifugation at 200*g* for ten minutes. All but approximately 1ml of the supernatant was carefully aspirated, in which the cell pellet was resuspended.

A fresh solution of cell fixative was prepared (3:1 methanol:glacial acetic acid, BDH) and cooled on ice. This was added dropwise to the resuspended cells with constant agitation until approximately 4ml had been added. A further 10ml of fixative were then added more rapidly, with agitation to prevent clumping of cells. The sample was incubated on ice for 20 minutes, followed by a further 10 minute period of centrifugation at 200*g*. The supernatant was aspirated from above the cell pellet, which was resuspended in the remaining small volume of fixative. This was repeated at least five times until the cell pellet appeared clean. The sample tube was filled with fixative and stored at -20°C prior to use as required.

2.2.2 Preparation of slides of metaphase spreads

Previously fixed cell preparations were removed from storage at -20°C and cells were pelleted by centrifugation at 200*g* for 10 minutes. All but approximately 0.5ml of the supernatant was removed, in which the cell pellet was resuspended, and fresh fixative was added dropwise to a volume of 10ml. Cells were pelleted by a second 10 minute period of centrifugation at 200*g*, and all but 0.5ml of supernatant was aspirated. The cell pellet was resuspended in sufficient fresh fixative such that the cell suspension gave a slightly opaque appearance.

Glass microscope slides (BDH) were prepared by soaking overnight in 2M HCl, followed by a thorough rinse in MQ water, and were dried with lens tissue (Whatman). Ten microlitres of chromosome suspension were dropped onto each slide from a height

of at least 50cm, in an atmosphere of 50% to 60% relative humidity. The fixative was allowed to evaporate and chromosomes were examined by light microscopy. Further fixative was added to the cell suspension if a reduction in cell density was required, and then the desired number of slides was prepared, with two regions of cells dropped on each. Slides were maintained at room temperature in darkness until required, and were typically used four to ten days after preparation.

2.2.3 Labelling of DNA with biotin and digoxigenin by nick translation

Typically, 1μg of DNA was labelled on each occasion in a 25μl final volume, and the reaction mixture was maintained on ice at all times. The volume of DNA preparation required to yield 1μg was estimated and made up to 18.5μl with MQ water. To this was added 2.5μl of 10x nick translation buffer (appendix A) and 2.5μl of a dNTP mix specific for biotin or digoxigenin labelling (appendix A) as required. The enzyme component of the reaction was added last, comprising 0.5μl of DNA polymerase I (10units/μl, Gibco BRL) and 1μl of DNAse I working stock (0.0075units/μl, Gibco BRL, appendix A). The reaction was incubated at 14.5°C for 60 minutes, and stopped by the addition of 2.5μl of 0.5M EDTA pH8.

Digestion of the probe was assessed by electrophoresis of 10% of the reaction volume through a 2% (w/v) agarose gel in comparison with 1μg of a 100bp ladder (Gibco BRL) to ensure that the DNA fragments had been digested to approximately 0.2kb to 0.6kb.

Unincorporated nucleotides were removed by ethanol precipitation, by the addition of 2.5μl of 3.5M NH_4Ac and 70μl of ice-cold absolute ethanol. The sample was incubated at -70°C for 90 minutes or at -20°C overnight, followed by centrifugation at maximum speed in a benchtop microcentrifuge for 15 minutes at room temperature. The supernatant was removed and the pellet washed by the addition of 250μl of 80% (v/v) ethanol, followed by agitation and a further 10 minute period of centrifugation. The supernatant was removed with a drawn-out glass Pasteur pipette and the pellet was air-dried in a laminar flow hood until it became clear and glassy in appearance. Probes were typically resuspended at approximately 50ng/μl in MQ water. The concentration of each precipitated probe was estimated by comparison with serial dilutions of a known concentration of sonicated salmon sperm DNA in MQ water (Gibco BRL).

2.2.4 Preparation of probes for FISH analysis

For single locus probes, typically one biotin-labelled and one digoxigenin-labelled probe were combined in a single FISH reaction. Approximately 50ng of each single locus probe, or 150ng of paint probe, were added to 10μg of sonicated dog genomic DNA and 7.5μl of deionised formamide in a 1.5ml Eppendorf tube. This was mixed by trituration and incubated at room temperature for approximately one hour to aid complete resuspension of the DNA. Prior to FISH analysis, 7.5μl of 2x hybridisation buffer were added (appendix A) and mixed by trituration. For hybridisation to human and cat chromosomes, 1μg of human C_0t-1 DNA (Gibco BRL) or 10μg of sonicated cat genomic DNA were used respectively as competitor.

2.2.5 Slide preparation

Slides were dehydrated by passing through an alcohol series consisting of 2x 70%, 2x 90% and 1x 100% (v/v) ethanol (AnalaR grade, BDH), for three minutes at each stage, and allowed to air-dry in a laminar flow hood. Slides were prewarmed for five minutes on a hotplate maintained at 50°C, prior to immersion in 70% formamide, 2xSSC preheated to 65°C in a waterbath. After precisely two minutes, slides were immediately transferred to 70% (v/v) ethanol on ice. Slides were rehydrated through the ethanol series as described above, omitting the first 70% (v/v) ethanol step, and air-dried. Slides and coverslips were prewarmed on a hotplate maintained at 38°C to 40°C.

2.2.6 Denaturation of probe and hybridisation

Probes were denatured at 70°C for 10 minutes and transferred to 37°C for a 30 minute reannealing period. The probe was placed onto a prewarmed 22 x 22mm coverslip (BDH), the slide was inverted and lowered onto the coverslip, and then returned to the platform. A seal was made around each coverslip using Cow Gum rubber solution (Cow Proofings Ltd.). Slides were incubated overnight (or as required depending on the nature of the probe and chromosomes used) at 37°C in a humidified hybridisation chamber.

2.2.7 Post-hybridisation wash and blocking

All washing steps were performed in Coplin jars at 42°C. The rubber solution and coverslip were removed by soaking in 2x SSC at 42°C, and slides were then incubated in 50% (v/v) formamide, 2x SSC at 42°C for three periods of three minutes, with gentle

agitation by inversion. This was repeated for a further three periods of three minutes in 2x SSC at 42°C. Blocking of slides was performed by a 30 minute incubation at 42°C in 4x SSC/0.05% Triton-x100 containing 5% v/v non-fat dried milk powder, filtered through 3MM paper (Whatman).

2.2.8 Signal detection

Immunological detection reagents (Sigma and Vector Laboratories) were prepared and maintained in darkness. Reagents were prepared as follows in blocking solution (4xSSC/0.05% Triton-x100 containing 5% (w/v) non-fat dried milk powder):

Layer one: 4µg/ml Texas red avidin;

Layer two: 4µg/ml biotinylated anti-avidin, 4µg/ml fluorescein isothiocyanate (FITC) anti-digoxin conjugate (raised in mouse);

Layer three: 4µg/ml Texas red avidin, 4µg/ml FITC anti-mouse (raised in goat).

Reagents were mixed thoroughly and incubated at room temperature for 10 minutes prior to centrifugation at maximum speed for 10 minutes. One hundred and twenty microlitres of layer one detection reagent were applied to a 22 x 50mm strip of Parafilm (American National) and the slide was inverted and lowered onto the solution, ensuring complete distribution over the entire slide. Hybridisation was continued at 37°C in the humidified chamber for 30 minutes followed by three washes of three minutes each in blocking solution at 42°C. This process was repeated for layers two and three. The final wash was then followed by a rinse in 2xSSC at room temperature and slides were then allowed to air dry.

2.2.9 Counterstaining

Counterstaining of chromosomes was performed by the immersion of slides for four minutes in a solution of 4-6-diamidino-2-phenylindole dihydrochloric hydrate (DAPI, Sigma, at 80ng/ml in 2xSSC) in a Coplin jar. This was followed by a brief rinse in 2xSSC, after which the slide was air-dried. A single drop of Vectashield (Vector Laboratories) mounting medium was placed onto a clean 22 x 50mm coverslip (BDH). The slide was inverted onto the coverslip, and excess moisture was blotted away with firm, vertical pressure between layers of 3MM paper (Whatman). Slides were stored at 4°C in darkness.

2.2.10 Analysis

Slides were allowed to warm to room temperature prior to examination. Analysis was performed using a fluorescence microscope (Axioskop, Zeiss) equipped with a 50 watt mercury bulb, and separate FITC, Texas Red and DAPI filters (Chroma Technologies). Typically, a minimum of 20 metaphase spreads were examined for each probe reaction analysed. Suitable metaphases were selected by examination under a 20x objective (Zeiss), and further examined under a 100x oil immersion objective (Zeiss). Images were captured using a cooled CCD camera (KAF1400, Photometrics), driven by a Macintosh Quadra 800 computer with dedicated software (Smart Capture, Vysis), with a binning of two. Chromosomes were identified by enhanced linear filtration of the DAPI-banded component of the captured image. Dog and cat chromosomes were described using the nomenclature proposed by Breen *et al.* (1999b) and Cho *et al.* (1997a) respectively.

2.3 Development and characterisation of expressed sequence tags for the dog genome

2.3.1 Titration of the library

The dog cDNA library was provided by Dr. M. Binns (Animal Health Trust) and was generated using two rounds of oligo(dT) selection, with subsequent cloning of double stranded cDNA into the *Eco*RI site of bacteriophage lambda (λ) gt10. Ten microlitres of the glycerol stock of the *Eschericia coli* C600 host strain (Stratagene) were inoculated onto LB agar supplemented with 0.2% maltose and 10mM $MgSO_4$, and incubated at 37°C for 24 hours. A single colony was inoculated into 10ml of LB broth containing 0.2% (v/v) maltose and 10mM $MgSO_4$ and incubated at 37°C, shaking at 250rpm overnight. Sterile glycerol was added to a final concentration of 50%, mixed thoroughly, and stored at -70°C.

The titre of the cDNA library was estimated by plating out serial dilutions of the stock. An overnight culture of the host was prepared by inoculation of 4ml of LB broth containing 0.2% (v/v) maltose and 10mM $MgSO_4$, with a single colony of *E. coli* C600, followed by incubation at 37°C for 16 hours. The cell density of the culture was estimated using a Pharmacia Ultrospec 2000 spectrophotometer, and incubation was continued until OD_{600}=1.0 was reached. The culture was pelleted by centrifugation at 1000*g* for 10 minutes in an MSE Centaur centrifuge, and was gently resuspended in 4ml of sterile 10mM $MgSO_4$.

Two hundred microlitres of this culture were added to serial dilutions of the library stock (described in chapter three), each made up to a total volume of 200µl with SM buffer. The samples were briefly mixed and incubated at 37°C for 20 minutes. Each was added separately to 4ml of LB top agarose maintained at 48°C, inverted to mix and poured onto LB agar in 88mm plates. Two control plates were also prepared. Two hundred microlitres of C600 were plated in the absence of phage, to confirm satisfactory growth of the bacterial host. The second control plate contained only LB agarose without phage or host, confirming absence of contaminating microbial growth. After allowing to set, the plates were incubated at 37°C for approximately 12 hours. The number of plaques obtained on each plate was used to calculate the titre of the cDNA library.

2.3.2 Removal and storage of randomly selected plaques

Isolated plaques were picked at random from plates prepared as described in section 2.3.1, using a library dilution required to give approximately 50 plaques per plate. Sterile glass Pasteur pipettes were used to remove an agarose core containing the plaque, which was placed into 100µl of sterile MQ water, gently agitated and incubated for 30 minutes at 37°C. The water was removed to a fresh tube and stored at -20°C for use in the PCR. Two hundred microlitres of SM buffer were added to the core, and stored at 4°C.

2.3.3 Amplification and purification of insert DNA

One microlitre of phage DNA in MQ water (prepared in section 2.3.2) was used directly in the PCR. Amplification was performed using the λgt10 vector primers (appendix B). Five microlitres of each product were analysed by agarose gel electrophoresis and the insert size estimated by comparison with 1µg of a 1kb DNA ladder (Gibco BRL). For selected clones, the remaining 20µl reaction was purified using a Qiagen PCR purification kit, and the concentration was estimated by comparison with Lambda DNA Quantitation standards (Gibco BRL).

2.3.4 Cycle sequencing of insert DNA and database similarity searching

Insert DNA was subjected to cycle sequencing analysis, and data were subsequently edited, as described in section 2.1.8. Edited data were searched against the EST and nr databases using the BLAST search tool, as described in section 2.1.9.

2.3.5 Multiple sequence alignments

Alignment of sequences was performed using the multiple alignment program ClustalW version 1.7 (Thompson *et al.*, 1994), hosted at the Baylor College of Medicine (http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html).

Sequence data were submitted in FASTA format, and alignments were performed using default parameters.

2.3.6 Long range PCR

Amplification of PCR products greater than 2kb from genomic DNA was performed using an Expand Long Template PCR system (Boehringer). Two separate reaction mixes were prepared on ice. The first consisted of 350μM of each dNTP, 100ng of each PCR primer, and 500ng of dog genomic DNA, in a total volume of 25μl. The second mix contained 5μl of the supplied buffer I and 0.75μl of the supplied enzyme mix, also in a total volume of 25μl. Immediately prior to cycling, the two mixes were combined, overlaid with 30μl light mineral oil (Sigma), and placed in a Perkin Elmer DNA thermal cycler preheated to 93°C. Amplification conditions consisted of an initial denaturation step of two minutes at 93°C, followed by 10 cycles of 10 seconds at 93°C, 30 seconds at the optimised annealing temperature, and six minutes at 68°C. This was followed by 20 further cycles in which the elongation step was increased by 20 seconds each cycle. Cycling terminated with a seven minute incubation at 68°C, and reaction mixes were then maintained at 4°C until required. Amplification products were analysed by electrophoresis through a 1% (w/v) agarose gel, and were purified as described in section 2.1.6.

2.4 Identification of conserved synteny between chromosome five of the domestic dog and the human genome

2.4.1 Radiolabelling of probes

Probes for type I markers were radiolabelled by PCR incorporation of α^{32}P-dATP. The PCR was performed as described in section 2.1.4 under the optimised conditions, except that the reaction volume was increased to 50μl, and 1.1MBq of α^{32}P-dATP (Redivue, specific activity 22TBq/mmol, 370MBq/ml, 10mCi/ml, Amersham) were added.

Following the PCR, unincorporated nucleotides were removed by passing the product through a G50 Sephadex column (Pharmacia) following the manufacturer's instructions,

using TNE as the equilibration and elution buffer. The storage buffer was discarded, the column was rinsed with TNE, and then filled with further TNE, which was allowed to flow through. The probe was added, followed by the addition of 400µl of TNE, and the column was allowed to empty. A 1.5ml Eppendorf tube was placed under the column, into which the probe was eluted by adding 400µl TNE to the column. The column was discarded and the eluted probe was monitored with a Series 900 Mini Monitor (Mini Instruments Ltd.) to confirm incorporation of the radiolabelled nucleotide. A duplicate amplification reaction was included as a positive control, in which the radiolabelled nucleotide was replaced by MQ water. One microlitre of this product was spotted onto Hybond N+ nylon membrane (Amersham) and fixed using a UV crosslinker with default settings. A further 5µl of PCR product were analysed by electrophoresis through an agarose gel to ensure that the PCR had been successful.

2.4.2 Preparation and prehybridisation of BAC library filters

The dog BAC library (RPCI81) (Li *et al.*, 1999) was gridded and supplied by the Human Genome Mapping Project Resource Centre (HGMP-RC, Hinxton, Cambridge, UK). BAC library filters were prepared by the HGMP-RC on Hybond N nylon filters (Amersham) and were stored at 4°C prior to use. In preparation for hybridisation, each filter was briefly immersed in a small volume of 2xSSC, and excess liquid was drained away. Filters were laid between layers of nylon mesh (Amersham) to a maximum of four filters. The filters were rolled, avoiding air bubbles, and inserted into a hybridisation bottle (Amersham) with the Hybond N+ containing the positive control. To each bottle was added 30ml of hybridisation buffer (Church and Gilbert, 1984). The sealed bottle was placed in a hybridisation oven (Hybaid) at 60°C for a minimum of one hour with constant rotation at approximately six revolutions per minute.

2.4.3 Addition of labelled probe and hybridisation

The labelled probe was denatured by incubation at 95°C for five minutes, and transferred to ice. The prehybridisation solution was removed from the bottle and the denatured probe added to 30ml of fresh hybridisation buffer prewarmed to 60°C. The solution was briefly mixed and then added to the hybridisation bottle, which was returned to 60°C. Hybridisation was continued for 16-20 hours at 60°C, with constant rotation.

2.4.4 Posthybridisation washes

The radiolabelled probe was removed and if required, was retained and stored at -20°C for future use. Fifty millilitres of 2x SSC (prewarmed to 60°C) were added to the hybridisation bottle and the sealed bottle was inverted several times before the buffer was discarded. This was replaced by a further 50ml of prewarmed 2x SSC, and the bottle was returned to 60°C for 15 minutes with constant rotation. The buffer was then discarded and a further 50ml were added for the same period. This was replaced with 50ml of prewarmed 1x SSC, 0.1% (w/v) SDS, and washing was continued for two periods of 15 minutes each. Filters were removed and monitored, and washing was continued until the counts on a given filter fell to below 30 counts per second, using 0.5x SSC, 0.1% (w/v) SDS when a higher washing stringency was required.

The nylon mesh was removed and filters were briefly blotted between sheets of 3MM paper (Whatman). Each filter was securely enclosed in Saran wrap and placed in a 24cm x 30cm autoradiography cassette (GRI Molecular Biology). A sheet of BioMax FS film (Kodak) and an intensifying screen (GRI Molecular Biology) were added and cassettes were placed at -70°C for 24 hours.

Films were developed using an Agfa Curix 60 developer, and the presence of hybridisation to the positive control was confirmed. Clones corresponding to positive hybridisation signals were identified from their grid position on the nylon filter. In some cases it was not possible to determine without question the identity of a clone in this manner, and so multiple clones from the area of the signal were selected, amongst which the required clone would be represented.

2.4.5 Verification of BAC clones

Dog BAC clones were supplied by the HGMP-RC as agar slants (LB agar containing 20μg/ml chloramphenicol) in individual 7ml bijoux tubes. In order to provide rapid preliminary verification of the identity of BAC clones, a small quantity of culture was removed from the agar slant under aseptic conditions, and resuspended in 20μl MQ water. One microlitre of bacterial suspension was used as the template in a standard 10μl PCR, under the cycling conditions appropriate for the marker under study. A duplicate reaction was performed using dog genomic DNA as a positive control. The size of the products of each reaction were compared by subsequent agarose gel electrophoresis and those clones for which an appropriate band was observed were selected for further study.

2.4.6 Isolation of BAC DNA (C. Langford, pers. comm.)

Selected clones were streaked onto LB agar plates containing 20μg/ml chloramphenicol (Sigma) using sterile toothpicks. Plates were inverted and incubated at 37°C overnight. A single colony was inoculated into 10ml of LB broth containing 20μg/ml chloramphenicol (Sigma). The culture was incubated overnight at 37°C shaking at 250rpm. The culture was transferred aseptically to a 15ml polypropylene tube and the cells were pelleted by centrifugation for 10 minutes at 3400*g* in an MSE Centaur centrifuge. The supernatant was discarded and the tube inverted for approximately two minutes to allow any remaining medium to drain away. Two hundred microlitres of 200μl GET buffer (appendix A) were added to the cell pellet and the contents mixed gently. After a 10 minute incubation period at room temperature, the cell pellet was gently resuspended in the GET buffer by gentle trituration with a wide-bore disposable Pasteur pipette, and transferred to a 1.5ml Eppendorf tube. A fresh solution of 0.2M NaOH/1% (w/v) SDS was prepared, and 400μl were added to the cell suspension. The sample was incubated on ice for five minutes, followed by the addition of 300μl of 3M sodium acetate pH5.2 and a further incubation period of 10 minutes on ice. The mixture was spun for five minutes at maximum speed in a benchtop microcentrifuge and the supernatant transferred to a fresh 1.5ml Eppendorf tube. This was respun under the same conditions and the supernatant again transferred to a fresh tube. If the supernatant was not clear, it was incubated on ice for a further 30 minutes, and cell debris pelleted by a further five minute spin at maximum speed. Six hundred microlitres of isopropanol were added to the clear supernatant, mixed by inversion and then incubated at -70°C for 30 minutes or at -20°C overnight.

The sample was spun at maximum speed for five minutes, the supernatant removed and the tube inverted to allow any excess to drain away. Two hundred microlitres of 0.3M sodium acetate pH7 were added to the pellet and incubated on ice for 30 minutes. The pellet was resuspended and 200μl of 1:1 phenol:chloroform were added. The mixture was agitated using a vortex mixer and spun at maximum speed for three minutes. The upper, aqueous layer was removed to a fresh tube, avoiding any carryover of the organic layer. A further 150μl of 0.3M sodium acetate pH7 were added to the original tube and the sample was mixed and spun as before. The remaining aqueous layer was added to that previously transferred to the fresh tube, again avoiding the lower organic layer. An equal volume of isopropanol was added and the sample was incubated at -70°C for 60

minutes. This was spun at maximum speed for 10 minutes and the supernatant discarded. The resulting pellet was washed by the addition of 500µl of ice-cold 70% (v/v) ethanol followed by inversion of the tube, and a further five minute spin. The supernatant was discarded and the pellet air-dried at room temperature for approximately 10 minutes. The pellet was resuspended in 50µl MQ water and 1µl of a 10mg/ml solution of RNase A (Sigma) in MQ water was added, followed by a 10 minute period of incubation at 55°C. The concentration of each BAC DNA preparation was estimated by comparison with bacteriophage Lambda DNA quantification standards (Gibco BRL).

2.5 Towards a radiation hybrid map of dog chromosome five

2.5.1 Preparation of cosmid DNA

A dog genomic DNA library (Stratagene) in the pWE15 cosmid vector was maintained at -70°C in 50% (v/v) glycerol. Glycerol stocks of cosmid clones previously assigned to CFA 5 (M. Breen, pers. comm.) were revived by the addition of 10µl glycerol stock to 1ml of LB broth containing 50µg/ml ampicillin (AmpTabs, Stratagene) followed by 16 to 20 hours of incubation at 37°C, shaking at 250rpm. A 10ml volume of LB broth containing 50µg/ml of kanamycin was inoculated with 100µl of the overnight culture, and incubated for a further 16 to 20 hours. DNA was prepared from each 10ml culture using a Wizard SV Miniprep kit (Promega). The 10ml culture was pelleted by centrifugation at 3400*g* for 10 minutes in an MSE Centaur centrifuge. The supernatant was discarded and the pellet fully resuspended in 250µl of Cell Resuspension Solution. After transferring the solution to a sterile 1.5ml Eppendorf tube using a sterile disposable Pasteur pipette, 250µl of Cell Lysis Solution were added. The contents were mixed by inverting until clearing of the suspension was observed. Ten microlitres of Alkaline Protease Solution were added, and the contents again mixed by inverting. The lysate was spun at maximum speed in a microcentrifuge for 10 minutes and the cleared lysate removed to a fresh tube, avoiding the cell debris. To ensure complete removal of the debris, the lysate was spun for a further five minutes, and the contents removed to a second 1.5ml Eppendorf tube, ensuring that no cell debris was transferred. The lysate was added to a Wizard Plus Minipreps Spin Column inserted into a 2ml collection tube. The column was spun at maximum speed in a microcentrifuge for one minute, and the flowthrough discarded. The column was washed by the addition of 750µl of Column Wash Solution, and a further one minute spin was performed. The flowthrough was

discarded, a further 250µl of Column Wash Solution added, and a two minute spin was performed. The column was transferred to a sterile 1.5ml Eppendorf tube and 50µl of MQ water prewarmed to 65°C were added to the centre of the column matrix. After a one minute incubation at room temperature, the DNA was eluted by spinning the column for one minute. One microlitre of the sample was analysed by electrophoresis through a 2% agarose gel. The concentration of the resulting DNA solution was estimated by comparison with bacteriophage Lambda DNA quantification standards (Gibco BRL).The DNA sample was stored at -20°C.

2.5.2 End sequencing of cosmid DNA

Sequencing of cosmid DNA was performed as previously described in section 2.1.7 except that the Big Dye terminator mix (Perkin Elmer) was used in place of dRhodamine, and approximately 500ng of cosmid DNA were used in the reaction. Each reaction mix contained 5ng of $T3^{pWE15}$ or $T7^{pWE15}$ primer as required (appendix B). The products of cycle sequencing were purified by ethanol precipitation as described in section 2.1.7 except that the 10 minute incubation prior to the first spin was carried out at room temperature.

Sequence data were analysed as described in sections 2.1.7 to 2.1.9. PCR primers were designed using Primer version 0.5 (Whitehead Institute) under default parameters with the following exceptions; the optimum primer length was set at 22 nucleotides (minimum: 20, maximum: 25), the optimum melting temperature at 60°C (minimum: 57°C, maximum: 63°C), and the minimum G+C content at 40%, (maximum: 80%).

2.5.3 Optimisation of conditions for RH mapping analysis

Radiation hybrid mapping analysis was carried out in the laboratory of Dr. P. Deloukas (Sanger Centre, Hinxton, Cambridge, UK). Selected gene (type I) and anonymous (type II) markers were assessed for their potential use in RH analysis by an initial PCR optimisation process. Primer pairs for each marker (appendix B) were subjected to the PCR using dog genomic DNA, hamster genomic DNA and DNA from a randomly selected RH cell line as the template. Primer pairs were diluted to 200ng/µl in MQ water, and 3µl of each primer were added to 94µl of PCR dilution buffer (appendix A). For each primer pair, three reagent master mixes were prepared, differing only with respect to the template DNA included. The mix consisted of 7.2µl of 34.6% sterile-filtered sucrose in

MQ water, 2µl of Cox buffer pH8.8 (appendix A), 0.187µl of a fresh 1/10 dilution of β-mercaptoethanol (BioRad) in PCR dilution buffer (appendix A), 1.25nmol of each dNTP (Pharmacia), 25ng of DNA template and 1U of Taq polymerase (Perkin Elmer). Ten microlitres of the reagent mix containing dog genomic DNA were added to one well of each of three microtitre plates (Thermowell model M, Costar). Similarly, 10µl of the remaining two reagent mixes were added to each of two further wells. Ten microlitres of the primer mix were then added to each well, resulting in three 20µl reactions, differing with respect to genomic template, duplicated in each of three separate microtitre plates.

Rubber sealing mats (Hybaid) were treated twice in a UV crosslinker (Stratagene) using the 'auto crosslink' default settings, and were then used to seal each microtitre plate. The PCR was performed using a PTC-225 thermal cycler (MJ Research). Each plate was subjected to the PCR under one of three cycling conditions, differing only with respect to annealing temperature. Programs were based on a common 'Touchdown' PCR format (Don *et al.*, 1991), in which the annealing temperature of the first cycle was decreased by a defined increment every cycle for nine further cycles, followed by a further 30 cycles at the final annealing temperature reached. Specifically, the first cycle of each PCR program consisted of an initial denaturation stage of 94°C for five minutes. This was followed by 30 seconds at 93°C, 50 seconds at 5°C above the final annealing temperature to be reached, and 50 seconds at 72°C. This was repeated for nine further cycles, the annealing temperature being reduced by 0.5°C each time. An additional 30 cycles were then performed, each with the annealing temperature constant at that reached in the tenth cycle. Initial optimisation was performed with annealing temperatures of 50°C, 55°C and 60°C.

Subsequent to the PCR, reactions were analysed by electrophoresis through 2.5% (w/v) agarose gels (Gibco BRL) in 1x TBE running buffer. Whilst molten, 120µg of ethidium bromide (Gibco BRL) were added to 300ml of agarose, and the gel cast with six tiers each of 50 wells. Ten microlitres of each amplification product were loaded into separate wells, and 1µg of a 100bp DNA ladder (Gibco BRL) was added to one well of each tier. Electrophoresis was performed at 180 volts in 1x TBE for approximately 40 minutes. PCR products were visualised using an ultraviolet transilluminator (Ultraviolet Products) and results were recorded by Polaroid photography.

The results of the PCR were assessed to establish the optimum conditions for subsequent analysis of the marker across the entire radiation hybrid panel. Ideal conditions were

classed as those which gave strong amplification of the product from dog genomic DNA and no detectable amplification from hamster genomic DNA. In cases where this was not achieved at any of the three annealing temperatures used, further optimisation was performed. A lower annealing temperature was used when amplification of the dog product was weak, and a higher temperature when amplification from hamster genomic DNA was observed. Markers for which ideal cycling conditions could not be established were excluded from further analysis.

2.5.4 Marker analysis on the entire radiation hybrid panel

A single reagent master mix was prepared for each marker to be tested, containing sufficient mix for 110 individual reactions. Each reaction required 7.2µl of 34.6% sterile-filtered sucrose in MQ water, 2µl of Cox buffer pH8.8, 0.187µl of a fresh 1/10 dilution of β-mercaptoethanol (BioRad) in PCR dilution buffer, 1.25nmol of each dNTP (Pharmacia), 50ng of each primer and 1U of *Taq* polymerase, as described in section 2.5.3.

The dog radiation hybrid panel (L. McCarthy, unpublished, see section 1.5.5) was supplied in 2ml Cryovials (Nunc) at a concentration of 50ng/µl in MQ water. This was diluted to 15ng/µl in dilution buffer (appendix A) in a deep well microtitre plate (Beckman). Hybrid DNA was mixed and dispensed using an automated microdispenser (Robbins). This was prepared by an initial acid cleaning process, in which 250µl volumes of 0.25M hydrochloric acid were drawn up and then expelled three times. This process was then repeated three times with fresh changes of MQ water. The deep well plate containing the diluted hybrid DNA was placed under the microdispenser and the contents mixed using the 'wash' setting. Multiples of 10µl of DNA were then drawn up depending on the number of markers to be tested. A PCR microtitre plate (Thermowell model M, Costar) was placed under the microdispenser, and a single aliquot of 10µl of the entire panel was dispensed, ensuring that the plate was in the correct orientation and that DNA had been dispensed into all wells. This was repeated for all plates to be used, after which the microdispenser was washed with a further six changes of MQ water.

Ten microlitres of the PCR master mix were added to each well of the microtitre plate containing hybrid DNA The PCR was then performed under the previously optimised conditions as described above. Ten microlitres of each reaction were analysed by agarose gel electrophoresis as described in section 2.5.3. Each lane was scored for the presence or

absence of a PCR product of the expected size. Presence of such a band was denoted '+', absence as '-' and where there was uncertainty, this was represented by '?'. The PCR was then repeated and analysed under identical conditions and scored in the same manner. Patterns of presence and absence of PCR product were compared between both replicates in order to derive a consensus.

2.5.5 Analysis of data using RHMAP

RH mapping data were analysed using the RHMAP version 2.01 analysis package (Boehnke *et al.*, 1991). In all cases, the assumptions of Cox *et al.* (1990) were followed, namely a) that radiation-induced breakage occurs at random, b) that retention of any given chromosome fragments occurs independently of any other fragment, and c) that retention probabilities for all fragments are equal.

Retention data were recorded in Microsoft Excel '97, and then exported as space-delimited text into the required RHMAP program (see sections 2.5.6 to 2.5.8), which were run locally under MS DOS. Locus names were abbreviated to symbols of no more than four characters, in accordance with the requirements of the RHMAP analysis package, and were analysed in alphabetical and numerical order. Input commands were inserted into each text document as required, following the author's instructions. Retention status was recorded using the symbols described above, such that '+' indicates locus typed and present in a given hybrid, '-' indicates locus typed and absent, and '?' indicates locus not typed (unresolved ambiguity).

2.5.6 Twopoint analysis of RH data using RH2PT

Pairs of linked markers were identified using the RH2PT analysis program (Boehnke *et al.*, 1991). Data files containing retention patterns for all markers across the RH panel were analysed under default parameters. The retention frequency for each marker was recorded, and pairs of markers linked with a lod score of three or greater were identified. Linkage groups were identified by the analysis program as sets of markers within which each marker was linked to at least one other with a lod score of four or greater. Marker order within individual linkage groups was thereafter analysed separately using RHMINBRK and RHMAXLIK analysis programs, and these data were compared to chromosome assignments generated by FISH analysis.

2.5.7 Analysis of RH data using RHMINBRK

Data analysed using RH2PT were adapted for analysis under RHMINBRK (Boehnke *et al.*, 1991) by the amendment of input commands following the author's instructions. Data were analysed using the 'stepwise' option, in which locus orders are evaluated in terms of the minimum number of obligate breaks between loci required to generate that marker order. Partially typed hybrids were included in the analysis, and no forced marker orders were used. All other parameters were retained at the default setting as recommended.

2.5.8 Analysis of data using RHMAXLIK

All linkage groups were analysed using RHMAXLIK (Boehnke *et al.*, 1991) under the equal retention model, which assumes that each chromosome fragment has an equal probability of being retained in a given hybrid. Centromeric and telomeric linkage groups were also analysed using the centromeric retention model where appropriate. The stepwise ordering strategy was used, in which possible locus orders are compiled one at a time. No forced marker orders were used, and partially typed hybrids were included in the analysis.

2.6 Direct cDNA selection towards isolation of gene markers for dog chromosome five

2.6.1 Titration of the dog brain cDNA library

A partial cDNA library derived from dog cerebellum tissue was obtained from Dr. D. Sargan (Cambridge University, UK). The construction of the library has been described previously (Lin and Sargan, 1997) in which the poly(A)+ fraction of total cerebellar RNA was isolated using oligo(dT)-chromatography. Following reverse transcription, resulting cDNA was cloned non-directionally into the *Eco*R1 site of the λZAPII vector (Short *et al.*, 1988) (Stratagene), and transformed into the XL1Blue MRF' host strain (Stratagene). The library had been amplified once prior to receipt.

Titration of the library was essentially performed as described in section 2.3.1, using the XL1 Blue MRF' (Stratagene) host strain in place of C600. The host was plated onto LB agar containing 12.5μg/ml tetracycline, and after approximately 20 hours of incubation at 37°C, a single colony was inoculated into 10ml of LB medium containing 0.2% (w/v) maltose and 10mM $MgSO_4$. The cell density of the culture was estimated using a

Pharmacia Ultrospec 2000 spectrophotometer, and incubation was continued until an OD_{600}=1.0 was reached. The culture was pelleted by centrifugation at 1000*g* for 10 minutes in an MSE Centaur centrifuge, and was gently resuspended in 10ml of sterile 10mM $MgSO_4$.

Ten-fold serial dilutions of the cDNA library glycerol stock were prepared and were made up to a total volume of 200µl with SM buffer. Each dilution was combined with 200µl of the XL1 Blue MRF' host strain, and incubated at 37°C for 20 minutes. Samples were added to 4ml of LB top agarose maintained at 48°C, mixed by inversion, and poured onto LB agar plates. Two control plates were also prepared. Two hundred microlitres of XL1 Blue MRF' were inoculated into LB top agarose and plated in the absence of phage, to confirm satisfactory growth of the bacterial host. The second control plate contained only LB agarose without phage or host, confirming absence of contaminating microbial growth. After allowing to set, plates were incubated at 37°C for approximately 12 hours. The number of plaques obtained on each plate was recorded and used to estimate the library titre.

2.6.2 Phagemid excision

Excision of the pBluescript II SK^- phagemid vector from the λZAPII phage was carried out using the Stratagene ExAssist kit. This method is summarised in figure 2.6a. Splinters of the supplied glycerol stocks of the host strains XL1Blue MRF' and SOLR (Stratagene) were removed with sterile toothpicks and streaked onto LB agar plates containing 12.5µg/ml tetracycline or 50µg/ml kanamycin respectively. The glycerol stocks were stored at -70°C, and the plates were incubated at 37°C for 16 hours. A single colony of XL1 Blue MRF' was inoculated into 20ml of LB medium containing 0.2% (v/v) maltose and 10mM $MgSO_4$, in a sterile 250ml conical flask. A single colony of SOLR was added to 20ml of LB containing no supplements, and both cultures were incubated for 16 hours at 30°C, 250rpm. The lower incubation temperature was intended to slow the growth rate, thus preventing excessive cell density. The cell density of each culture was estimated using a Pharmacia Ultrospec 2000 spectrophotometer, and if necessary, they were returned to the incubator for further growth until OD_{600}=1.0 was reached. Cultures were transferred to sterile 50ml polypropylene tubes (Sterilin) and the cells were pelleted by centrifugation at 1000g for 10 minutes in an MSE Centaur

Figure 2.6a: (opposite)

Outline of the *in vivo* phagemid excision procedure. The experimental method is summarised on the left of the flowchart, and an explanation of the methodology is given on the right.

Lambda ZAPII library. The lambda vector contains initiation (I) and termination (T) sequences in the f1 ORI.

Co-infect **XL1 Blue MRF'**

ExAssist helper phage. This provides the filamentous phage gene II protein required for excision. Replication of the helper phage is possible in the XL1 Blue host, but not in SOLR.

Adsorb at 37°C 15 mins

(A)
Add liquid medium and incubate at 37°C for 2-3 hours with gentle agitation.

(B)
Incubate at 65-70°C for 20 minutes. Spin down and decant the supernatant.

(C)
Combine a range of volumes of the supernatant with the SOLR host. Incubate at 37°C for 15 minutes, and plate onto LBampicillin. Incubate overnight at 37 °C.

The titre is calculated by:

$$\frac{\text{(no. of cfu's x dilution factor x1000)}}{\text{volume plated (in } \mu\text{l)}}$$

The excision efficiency represents the ratio of the number of cfu's rescued compared to the number of pfu's entered into the excision process.

(A)
The helper phage proteins recognise the I sequence, nick one of the two strands, and initiate DNA synthesis. Single-stranded DNA synthesis continues in the 3' direction from the site of the nick, through the cloned insert, until the T sequence is reached. The single stranded DNA is then circularised by a gene II product from the helper phage, recreating a functional f1 ORI. This is packaged as a filamentous phage by the helper phage proteins, then secreted from the cell.
The T sequence also contains signals for packaging the excised phagemid DNA, and for secreting phage particles from the host cells. The phagemid contains all the DNA originally contained between the I and T sequences in the lambda vector. This includes the ampicillin resistance gene, and the T3 and T7 sites. The f1-M13 ORI allows the phagemid to be replicated, packaged into phage particles,and secreted from the bacterial host.

(B)
Heat treatment inactivates lambda particles and XL1-Blue cells. Only the phagemid and helper phage remain intact.

(C)
The SOLR host does not suppress the mutations in the helper phage, which can therefore no longer be propagated. SOLR is also resistant to infection by the lambda bacteriophage, so only the excised phagemid DNA can be propagated.
SOLR is infected with the phagemid/helper phage mixture, and plated out to recover the phagemid. The helper phage genome is lost as the library is amplified.

centrifuge. The cell pellets were gently resuspended in 10ml of a sterile solution of 10mM $MgSO_4$, at a cell density of OD_{600}=1.0, corresponding to $8x10^8$ cells/ml.
The XL1 Blue MRF' host and lambda phage were combined in a sterile 50ml conical flask, such that the ratio of host cells to phage was 10:1 (see chapter six), and the solution was swirled gently to mix. ExAssist helper phage particles were added at an equal ratio to the number of host cells added. The solution was gently mixed and incubated at 37°C for 20 minutes in a waterbath to allow the adsorption of phage particles to the host cells. A 20ml volume of sterile LB broth was added and the culture incubated at 37°C, 250rpm. After three hours, the culture was transferred to a sterile 50ml tube and heated at 70°C for 20 minutes, followed by centrifugation at 1000*g* for 10 minutes at room temperature. The resulting supernatant containing the excised phagemids was removed to a second sterile 50ml tube, and the pelleted cell debris was discarded

2.6.3 Titration of the excised phagemids

A series of dilutions of the excised phagemid were prepared and made up to 100µl with LB medium in a 1.5ml Eppendorf tube. Pairs of duplicate dilutions were prepared in each case for comparison. Each dilution was combined with 200µl of the SOLR host strain culture, and mixed gently. The contents were incubated at 37°C for 20 minutes, and 150µl were pipetted onto an LB agar plate containing 50µg/ml ampicillin. A sterile glass spreader was used to disperse the solution evenly across each plate, and was sterilised between uses by flaming with ethanol. After allowing the solution to be absorbed into the agar, plates were inverted and incubated at 37°C for approximately 12 hours. Where the dilution was such that single colonies were obtained on a plate, these were counted and an average calculated for each pair of duplicates. These values were used to estimate the titre of the excised phagemids, based on the dilution factor involved.

2.6.4 Phagemid DNA preparation

A culture of SOLR was grown to OD_{600}=1.0, pelleted by centrifugation and resuspended in 10mM $MgSO_4$. The phagemid culture generated in section 2.6.3 was added to 25ml of SOLR culture. After a 20 minute incubation at 37°C, the contents were added to a two litre sterile conical flask containing 480ml of LB medium supplemented with 50µg/ml ampicillin. The flask was incubated at 37°C, shaking at 250rpm, for three hours, and was then transferred to a water bath at 70°C for 20 minutes.

Phagemid DNA was extracted from the culture using a Qiagen Plasmid Mega kit. The culture was transferred to sterile 250ml tubes (Sorvall) and was pelleted by centrifugation at 6000*g* for 15 minutes at 4°C in a Sorvall RC5B ultracentrifuge. The medium was discarded and the pellet was resuspended in 50ml of buffer P1, using a sterile disposable Pasteur pipette. A further 50ml of buffer P2 were added, and the suspension was mixed by inversion. A 15 minute incubation period at room temperature was followed by the addition of 50ml of chilled buffer P3, and the suspension was immediately mixed by inversion, resulting in the formation of a cloudy precipitate of cell debris.

The cell lysate was poured into a Qiafilter cartridge attached to the neck of a sterile 500ml glass bottle (Duran), and incubated at room temperature for 10 minutes, during which time the precipitated material formed a layer on the surface of the lysate. The filter was attached to a vacuum source and the lysate allowed to flow through and was collected. After releasing the vacuum, 50ml of buffer FWB were added and the precipitate was mixed gently with a sterile Pasteur pipette. The vacuum was again applied, the liquid was allowed to flow through, and was added to that previously collected.

A Qiagen tip-2500 was prepared by the addition of 35ml of buffer QBT, which was allowed to pass by gravity flow. The filtered lysate material was then applied to the column, and the flow allowed to proceed to completion. The column was washed by the addition of two aliquots of 100ml of buffer QC. The column was then transferred to a sterile collection vessel, and the DNA was eluted by the application of 35ml of buffer QF.

The collected DNA was precipitated by adding 24.5ml of isopropanol at room temperature, and the solution was mixed and pelleted by centrifugation at 15000*g*, 4°C for 30 minutes in a Sorvall RC5B ultracentrifuge. The supernatant was removed, avoiding the glassy pellet which formed at the bottom and sides of the centrifuge tube. The pellet was washed by adding 7ml of 70% (v/v) ethanol at room temperature, and spun as before for 10 minutes. After removal of the supernatant, the pellet was allowed to air-dry at room temperature for 15 minutes, and was then dissolved in 1ml of MQ water. The sample was incubated at room temperature overnight to facilitate resuspension of the cell pellet. The resulting solution was then transferred to a 1.5ml Eppendorf tube and was stored at -20°C prior to use. The concentration of the sample was estimated by comparison with known concentrations of sonicated salmon sperm DNA (Gibco BRL)

electrophoresed through a 2% (w/v) agarose gel with 1μg of a 1kb DNA size standard (Gibco BRL).

2.6.5 Preparation of cDNA template

The phagemid DNA preparation was used in the PCR to generate insert DNA for use in the subsequent cDNA selection process. Dilutions of phagemid DNA were prepared in MQ water and the PCR performed using $T3^{SK}$ and $T7^{SK}$ primers, and an annealing temperature of 60°C. Duplicate reactions were carried out with elongation times of one, two and four minutes in order to establish whether this led to an increase in the average size of the resulting product. The optimum concentration of phagemid DNA was established by comparison of the yield and size range of the product as observed by electrophoresis through a 2% (w/v) agarose gel.

A large scale PCR was then prepared under the optimised conditions, consisting of 30 duplicates of a 10μl reaction. Five microlitres of the resulting pooled product were examined by electrophoresis to ensure that the yield and average product size reflected that produced in the previous optimisation stage. The remainder of the reaction was divided into three equal volumes and purified using a PCR purification kit (Qiagen), eluting each sample in 30μl of buffer EB prewarmed to 65°C. The elutions were pooled, and one microlitre of the product was examined by agarose gel electrophoresis, and was quantified by comparison with known quantities of sonicated salmon sperm DNA.

2.6.6 Removal of free biotin from chromosome paint probes

Unincorporated biotin-16-dUTP were removed from chromosome paint probes using a Qiagen PCR purification kit, as described in section 2.1.6. Samples were eluted in 30μl of MQ water prewarmed to 65°C, and the size range and concentration of the eluted products was estimated by electrophoresis through a 2% agarose gel, and comparison with a 1kilobase DNA ladder (Gibco BRL) and known quantities of sonicated salmon sperm DNA (Gibco BRL). The concentration of purified paint probes was adjusted to 100ng/μl with additional MQ water.

2.6.7 Preparation of the cDNA selection reaction

Conditions for cDNA selection are described fully in chapter six, and the methodology is outlined in figure 2.6b. Briefly, 1μg of the cDNA produced in section 2.6.5 was

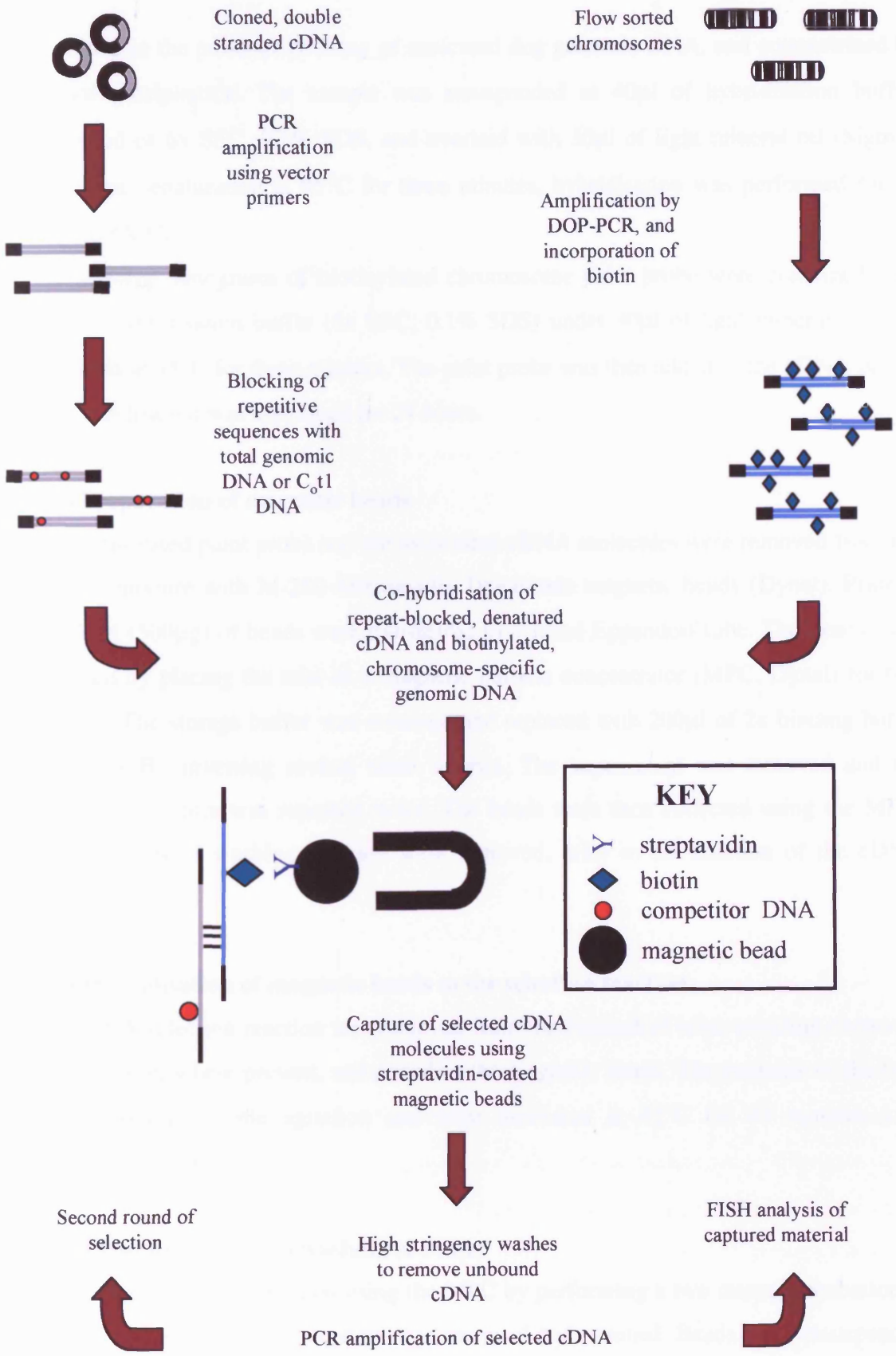


Figure 2.6b:

Direct cDNA selection

combined in the presence of 50μg of sonicated dog genomic DNA, and concentrated by ethanol precipitation. The sample was resuspended in 40μl of hybridisation buffer, composed of 6x SSC, 0.1% SDS, and overlaid with 30μl of light mineral oil (Sigma). After heat denaturation at 95°C for three minutes, hybridisation was performed for 16 hours at 65°C.

One hundred nanograms of biotinylated chromosome paint probe were combined with 10μl of hybridisation buffer (6x SSC, 0.1% SDS) under 30μl of light mineral oil, and denatured at 95°C for three minutes. The paint probe was then added to the cDNA probe, and hybridisation was continued for 24 hours.

2.6.8 Preparation of magnetic beads

The biotinylated paint probe and the associated cDNA molecules were removed from the reaction mixture with M-280 Streptavidin Dynabeads magnetic beads (Dynal). Prior to use, 50μl (500μg) of beads were transferred to a 1.5ml Eppendorf tube. The beads were collected by placing the tube in a magnetic particle concentrator (MPC, Dynal) for two minutes. The storage buffer was removed and replaced with 200μl of 2x binding buffer (appendix B), inverting several times to mix. The supernatant was removed and the washing procedure was repeated twice. The beads were then collected using the MPC, and all traces of washing solution were removed, prior to the addition of the cDNA selection sample.

2.6.9 Hybridisation of magnetic beads to the selection reaction

The cDNA selection reaction was removed from the Eppendorf tube, avoiding carryover of mineral oil where present, and added to the magnetic beads. The contents of the tube were mixed by gentle agitation and were incubated at 42°C for 45 minutes with occasional mixing.

2.6.10 Post hybridisation washing of beads

Magnetic beads were captured using the MPC by performing a two minute incubation at room temperature. The supernatant was removed and retained. Beads were resuspended in 200μl of wash buffer (0.1x SSC, 0.1% SDS) at room temperature, and mixed well by inversion. The beads were collected and the supernatant removed and retained, and a further two washes performed in the same manner. This was repeated for three periods of

15 minutes at 65°C, mixing by occasional inversion, and capturing beads with the MPC immersed in a 65°C water bath. Beads were finally washed twice with binding buffer at room temperature in order to remove traces of SDS, and were resuspended in 50µl of MQ water .

2.6.11 Amplification of captured material from magnetic beads

Amplification of captured cDNA was investigated under standard conditions (60°C annealing temperature, four minute elongation) as previously described except that 0.5µl or 1µl of resuspended beads was used as the template. Products were analysed by agarose gel electrophoresis, and the yield generated from reactions containing the two different template concentrations were compared. The quantity of template that produced satisfactory amplification was then used in each of 30 identical 10µl reactions. PCR products were purified using a Qiagen PCR purification kit as previously described, and quantified by comparison with serial dilutions of a known concentration of sonicated salmon sperm DNA (Gibco BRL).

2.6.12 FISH analysis of selected cDNA

Results of the selection procedure were assessed by FISH analysis of selected cDNA. Probes were labelled by direct incorporation of fluorochrome-conjugated dUTP. All manipulations involving directly labelled DNA were performed in reduced light. DNA was labelled as described in section 2.2.3 with the following amendments. All reactions were performed in 25µl volumes consisting of 500ng of template cDNA, 1.25µl of 0.2mM Spectrum Red dUTP (Vysis), 2.5µl of 0.1mM dTTP (Pharmacia), 5µl of each of dATP, dCTP and dGTP (Pharmacia) at 0.1mM each, 2.5µl of 10x NT buffer (appendix A), 0.5µl of DNA polymerase I (10units/µl, Gibco BRL) and 1µl of DNAse I working stock (0.0075units/µl, appendix A, Gibco BRL). Nick translation reactions were performed at 14.5°C for 90 minutes. Ethanol precipitation and analysis of reaction products by agarose gel electrophoresis were performed as described in section 2.4.3.

FISH analysis was performed as described in section 2.4, except that no immunological detection was required. Subsequent to the final 2x SSC wash, slides were air dried, DAPI-stained, mounted and visualised as previously described.

Chapter three:

Development and characterisation of expressed sequence tags for the dog genome

Development and characterisation of expressed sequence tags for the dog genome

Abstract

A dog liver cDNA library was used in a pilot study for the generation of expressed sequence tags, towards the addition of gene markers to the largely gene-poor genome map of the dog.

Following titration of the lambda library, insert DNA was amplified from randomly selected clones by the PCR. Sequence data were obtained for a total of 76 cDNA clones. Data were searched against the non-redundant GenBank and EST sequence databases, in order to identify similarity to previously characterised sequences. cDNA clones were assigned to one of eight categories on the basis of results obtained from each sequence database similarity search. Sequence data derived from both ends of the insert cDNA of thirteen clones demonstrated highly significant similarity to previously annotated gene sequences from other species present within one or both databases searched. Three further clones shared significant sequence similarity to existing gene sequences at one end of the insert cDNA only, and a further two clones demonstrated similarity to different genes at each end of the insert. No significant database match was found for 30 clones. The remaining cDNA clones were categorised as containing repetitive sequence motifs, or as sharing sequence similarity with anonymous human genomic DNA, mitochondrial DNA or microbial sequences.

Four of the 13 clones for which sequence similarity to previously annotated gene sequences was identified were further pursued. Unique PCR primers were designed from the cDNA sequence data, successfully enabling amplification of the corresponding sequence from dog genomic DNA in three of the four instances. Comparisons with human gene sequences enabled the location of dog intron/exon boundaries to be predicted. This was used in turn to predict the size of the product amplified from a dog genomic DNA template using the locus-specific primers. FISH analysis was used to assign two of the corresponding gene loci to a dog chromosome.

3.1 Selection and characterisation of random cDNA clones

3.1.1 Library titration

The titre of the dog liver cDNA library had previously been estimated as 10^{10} plaque forming units (pfu)/ml (M. Binns, pers. comm.), prior to storage at -70°C for a period of several months. Serial dilutions of the library stock were used to infect host cells, and these cultures were plated out to enable estimation of the titre after storage, as described in section 2.3.1. Assuming a stock titre of 10^{10} pfu/ml, tenfold serial dilutions were prepared, ranging from a predicted concentration of 10^9pfu/ml to 10^2pfu/ml. One hundred µl of each dilution were plated as described in section 2.3.1. The resulting density of plaques on three plates was such that the total number of pfu present could be accurately counted. One hundred µl of the 10^4pfu/ml dilution generated 555 plaques, and the same volume of the 10^3pfu/ml and 10^2pfu/ml dilutions generated 54 and seven plaques respectively. On the basis of the previous estimate of the stock library titre, these values were used to calculate the ratio of the observed:expected number of pfu obtained from the three library dilutions as 0.555, 0.54 and 0.7 respectively. The mean ratio of these data therefore implies that the library titre at the time of this analysis was 0.6x the expected value, or approximately 6.0 x 10^9 pfu/ml. These data are summarised in table 3.1a.

3.1.2 Amplification and sequencing of insert DNA

Phage clones were plated at a concentration resulting in approximately 50 plaques per 88mm plate. This ensured that clones could be isolated without a high incidence of mixed plaques being selected as a result of their close proximity. Two hundred plaques were selected at random. Insert DNA was amplified for each clone by the PCR using vector primers, and the size of the resulting product was estimated. Initially, all clones that amplified a single product were selected for further analysis. Subsequently, those with an insert size of one kilobase or more were preferentially selected, since this would increase the amount of sequence data available for downstream analysis. The average insert size of the 76 clones analysed was approximately 1400bp, with a range of 400bp to 3500bp. Sequencing analysis was performed from one end of each clone using the forward vector primer. After removal of low quality data, approximately 400 nucleotides of high quality sequence were generated for each clone. Reverse sequence data were obtained for those

Table 3.1a:

Preparation of dilutions of the phage cDNA library for titration. Tenfold serial dilutions of the library stock, with an estimated titre of 10^{10} pfu/ml, were prepared. One hundred µl of each dilution were plated out in order to re-evaluate the titre. Three dilutions generated countable numbers of pfu, which are indicated below. Column one shows the predicted concentration of the dilution used, on the basis of the previously estimated titre of 10^{10} pfu/ml. The predicted and actual number of plaque forming units obtained have been compared (columns two and three) to estimate the relationship between these values, and in turn to recalculate the total number of pfu/µl in the library.

Amount used	No. pfu predicted	No. pfu obtained	Observed/Expected
100µl x 10^{4}	1000	555	0.555
100µl x 10^{3}	100	54	0.54
100µl x 10^{2}	10	7	0.7
			average = 0.60

clones for which a gene identity could be assigned on the basis of database similarity search results.

3.1.3 Analysis of BLAST sequence similarity searches

Forward sequence data from 76 clones were searched against both the non-redundant (nr) and EST (dbEST) databases. In general, significant database matches were regarded as those generating at least 80% identity over a region of no less than 100bp, and with a P-value of 1 x e^{-10} or less. These guidelines follow the parameters used in many previous EST studies, which were discussed in chapter one. However, data for each cDNA clone generated within the present study were evaluated independently to enable investigation of similarities reported with lower levels of significance.

Results of database similarity searches are summarised below. Clones are grouped into eight categories reflecting the nature of the results obtained. The most significant database match reported for each clone is given in appendix C, and the chromosomal location of the human orthologue of the corresponding gene is taken from the Human Gene Nomenclature Database (HGND) (http://www.gene.ucl.ac.uk/nomenclature/).

- **CLASS A:**

Clones for which significant identity is shared with a previously annotated gene, with both forward and reverse sequence data.

This set of clones represents those most amenable to downstream analysis, since their identity is confirmed by database searching with sequence data from both ends of the clone. Thirteen clones (17.1%) satisfied these criteria. In each case, the extent of identity with the highest scoring match was in excess of 80%, typically spanning more than 200bp in total. The locus name, description and chromosome location of the human orthologues are given in table 3.1b.

LEI020

Forward sequence data from clone LEI020 demonstrated significant similarity to human aldehyde dehydrogenase II (ALDH2, also known as mitochondrial aldehyde dehydrogenase) in the non-redundant database. Eighty six percent identity exists with a human partial mRNA for ALDH2, over a region in excess of 400bp. Reverse sequence data were obtained from clone LEI020, and significant similarity was found between these data and human mRNA/cDNA sequence entries for ALDH2 in both the non-

Table 3.1b:

Proposed identity of class A clones. Sequence data from each end of insert cDNA from these thirteen clones demonstrate highly significant similarity to previously annotated gene sequences. The gene symbol and gene name with which similarity is observed is followed by the chromosomal location of that locus within the human genome, taken from the Human Genome Nomenclature Database. 'nd' indicates that this information could not be determined.

Clone	Symbol	Gene	HSA location
020	ALDH2	aldehyde dehydrogenase 2, mitochondrial	12q24.2
036	TAP1/2	transporter 1/2, ABC (ATP binding cassette)	6p21.3
042	HNRPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	7p15
053	FGB	fibrinogen, B beta polypeptide	4q28
072	PI	protease inhibitor 1 (anti-elastase), alpha-1-antitrypsin	14q32.1
077	ALB	albumin	4q11-q13
081	EEF1A1	eukaryotic translation elongation factor 1 alpha 1	6q14
096	EIF4A2	eukaryotic initiation factor 4AII	18p11.2
122	CYP2E	cytochrome P450 subfamily IIE (ethanol-inducible)	10q24.3-qter
134	ASL	argininosuccinate lyase	7pter-q22
143	PDZK1	PDZ domain containing-protein 1	1q21
151	SKB1	skb1 (*S. pombe*) homologue	nd
184	CYP2E	cytochrome P450 subfamily IIE (ethanol-inducible)	10q24.3-qter

redundant and EST databases. This implies that the anonymous cDNA sequence from dbEST with which the forward data share 86% identity also represents the ALDH2 gene. These data suggest that LEI020 shares approximately 85% identity with the human ALDH2 coding sequence over the region represented by the clone. ALDH2 maps to HSA 12q24.2.

LEI036

Both forward and reverse sequence data for LEI036 share similarity with rodent mRNA sequences for the ABC transporter gene (TAP) in the non-redundant database. In both cases, the region over which similarity was reported was only approximately 100bp, which falls near the minimum criteria required to be met for such data to be classed as significant in this study. The percentage of identical nucleotides across this region were 86% and 83% for forward and reverse sequence data respectively, which are also at the lower limit for being accepted as significant. Moreover, neither forward nor reverse sequence data identified any significant match with an entry from dbEST. This may simply reflect the absence of sequence data for this region of the TAP gene within the EST database. It can therefore be postulated that clone LEI036 represents part of the dog TAP gene, for which the human orthologue maps to HSA 6p21.3, although this must be classed as tentative in the absence of more definitive data.

LEI042

Data regarding clone LEI042 were amongst the most definitive identified in this study. Forward data demonstrated 83% identity over a region in excess of 400 nucleotides of the human sequence for the heterogeneous nuclear ribonucleoprotein (HNRPA2B1) gene, within the non-redundant database. Similar data were obtained by searching dbEST, in which 82% identity was found between LEI042 and a human cDNA clone that itself had been identified as representing HNRPA2B1. Comparable results were obtained by searching both databases with the reverse sequence data, with 90% and 91% identity observed with human HNRPA2B1 sequences in the non-redundant database and in dbEST respectively. Thus LEI042 can be confidently identified as representing part of the HNRPA2B1 gene, which has been assigned to HSA 7p15.

LEI053

As with LEI042, this clone could also be definitively identified on the basis of highly significant identity with existing sequences within both the non-redundant and EST databases. Both forward and reverse data shared approximately 88% nucleotide identity with part of the human gene encoding the beta subunit of fibrinogen (FGB), which has been assigned to HSA 4q28.

LEI072

A minimum of 83% identity over between 263bp and 474bp was found between sequence data from clone LEI072 and coding sequence entries representing the alpha-1-antitrypsin (PI) gene, in both the non-redundant and EST databases. Human PI has been assigned to HSA 14q32.1

LEI077

LEI077 generated the most significant sequence identity of all clones analysed since it represents a gene for which dog sequence data has recently been deposited within the non-redundant database. Forward and reverse sequence data share 96% and 99% identity respectively with the dog mRNA sequence for the serum albumin gene (ALB, accession no. AJ133489). Similarities to existing cDNAs in dbEST are also identified, although only one has previously been identified as representing the ALB gene itself. Human ALB maps to HSA 4q11-q13.

LEI081

LEI081 can be classed as representing part of the dog eukaryotic elongation factor 1 alpha 1 gene (EEF1A1), for which the human sequence has been assigned to HSA 6q14. Forward and reverse sequence data for LEI081 share 87% and 94% identity over 374bp and 476bp respectively with existing sequence entries for EEF1A1 within the non-redundant database. Forward and reverse sequence data for LEI081 also share 87% and 92% identity respectively with two anonymous human cDNAs within dbEST, which therefore can also be postulated to represent EEF1A1.

LEI096

Forward data demonstrate in excess of 95% identity to coding sequences for the eukaryotic initiation factor 4AII gene (EIF4A2) from HSA 18p11.2, in both databases

searched. This is also the case for the reverse sequence and an entry from dbEST. Thus LEI096 appears to represent part of the dog EIF4A2 gene. This also enables a tentative assignment to be given to a human cDNA from the non-redundant database, with which the reverse data share 95% identity over almost 500bp.

LEI122

Clone LEI122 represents part of the dog gene for cytochrome P450 subfamily IIE (CYP2E), also known as cytochrome P450-j, from HSA 10q24.3-qter. Forward and reverse LEI122 data respectively share 91% and 82% identity over 236bp and 476bp with existing entries for CYP2E within the non-redundant database. The same data for LEI122 respectively share 88% and 85% identity with cDNAs that have themselves been identified as representing CYP2E in dbEST. However, the degree of match with reverse data lies marginally outside the range that has been considered acceptable in this study, since the region of similarity extends only over 74bp. Despite this, the nature of the clone can be determined with confidence from the remaining three database similarity searches.

LEI134

Data from three of the four similarity searches performed with sequence data from clone LEI134 identify it as part of the dog arginosuccinate lyase gene (ASL), whose orthologue is located on HSA 7pter-q22. In each case, a minimum of 85% identity is observed with existing ASL sequence data. Forward sequence data generate BLAST search results spanning a region of less than 100bp (92bp and 87bp) in both databases searched, whilst reverse data share 88% identity over 209bp of human ASL in the non-redundant sequence database. In the fourth instance, comparison of the reverse data with the EST database demonstrated 87% identity with an anonymous human cDNA, over 199bp. This highly significant match indicates that the human cDNA clone also represents part of the ASL gene.

LEI143

As with LEI134, clone LEI143 can be identified by virtue of highly significant matches reported by three of the four searches performed, whilst the fourth enables a putative identification of an anonymous cDNA clone. LEI143 represents a region from the dog gene encoding the PDZ domain containing-protein 1 (PDZK1), whose orthologue is located on HSA 1q21. LEI143 reverse sequence data share approximately 90% identity

with the coding sequence of human PDZK1, over 182bp in both databases. Forward data are 89% identical to a 474bp region of the same gene sequence within the non-redundant database, and 88% identical to an anonymous human cDNA clone from dbEST over a total of 478bp. This human cDNA clone can therefore be proposed to represent part of the human PDZK1 gene.

LEI151

Respectively, forward and reverse data from LEI151 demonstrate 94% and 91% identity to 264bp and 436bp regions of two anonymous human cDNA sequences within dbEST. The nature of LEI151 can be better established by searching the non-redundant database with the same sequence information. Forward and reverse clone data share approximately 94% identity with the human SKB1 gene, a homologue of a gene previously identified in the yeast *Schizosaccharomyces pombe*. The extent of the region of similarity is 264bp for forward data from LEI151, and 472bp with reverse data. Human SKB1 has not been chromosomally assigned.

LEI184

This clone represents a similar case to that of LEI122, since both represent the same gene, cytochrome P450 subfamily IIE (CYP2E), from HSA 10q24.3-qter. The degree of nucleotide identity between forward and reverse clone sequence and non-redundant database entries is highly comparable, demonstrating between 81% and 85% identity with existing CYP2E sequences over a total of 248bp and 395bp respectively. The anonymous human cDNA clone with which LEI184 shares minor similarity (85% identity over 74bp) may therefore also represent part of the CYP2E gene, although further investigation of the human cDNA clone sequence would be required to confirm this proposition.

- **CLASS B:**

Clones for which significant identity is shared with a previously annotated gene, with either forward or reverse sequence data only.

These two clones, representing 2.6% of the 76 analysed, are not as amenable to further study as class A clones since the identity of class B clones cannot be confirmed using sequence data from both ends of the insert DNA. However, the fact that one end of the insert DNA can be assigned an identity may still prove valuable if this region of sequence

data is treated separately from the remainder of the clone. The locus name, description and human location for the corresponding genes are given in table 3.1c.

LEI136

Forward data from LEI136 share significant nucleotide similarity with part of the human coding sequence for the arginase gene (ARG1), whose human orthologue is located on HSA 6q23. Similarity searches against both the non-redundant and EST databases indicate that LEI136 shares approximately 88% identity with human ARG1 over regions in excess of 250bp. No matches were reported from searching the same databases with reverse sequence data from this clone.

LEI149

Forward sequence data from LEI149 show 95% identity over a 150bp region of a rat mRNA for serine/threonine kinase beta-PAK, from the non-redundant sequence database. These data also demonstrate 94% identity with a mouse cDNA for P21 activated kinase-3 from dbEST, again over a 150bp region. The HGND confirms that these matches represent the same gene, the standardised name of which is p21 (CDKN1A)-activated kinase 3 (PAK3). PAK3 is located on HSA Xq21.3-q24. Reverse data identified no significant matches in either database.

- **CLASS C:**

Clones for which significant identity is shared with a previously annotated gene with forward sequence data, but with a different gene with reverse data.

For these three clones, 3.9% of the total analysed, there is evidence to suggest that each contains sequence derived from two different transcripts. As with the previous set of clones, these may still be of value for mapping of the corresponding loci if the sequence data from each end of the insert DNA are considered as separate information. The locus name, description and human location for the corresponding genes are given in table 3.1d.

LEI079

LEI079 forward data demonstrate approximately 86% identity with existing sequence information for the human haptoglobin gene (HP) located on HSA 16q22.1, in both the non-redundant and EST databases. This similarity extends over 400bp of the clone.

Table 3.1c:

Proposed identity of class B clones. Sequence data from only one end of insert cDNA from these two clones demonstrate highly significant similarity to previously annotated gene sequences. The gene symbol and gene name with which similarity is observed is given, followed by the chromosomal location of that locus within the human genome, taken from the Human Genome Nomenclature Database.

Clone	Symbol	Gene	HSA location
136F	ARG1	arginase, liver 6q23	6q23
149F	PAK3	p21 (CDKN1A)-activated kinase 3	Xq21.3-q24

Table 3.1d:

Proposed identity of class C clones. Sequence data from insert cDNA from these three clones demonstrate highly significant similarity to previously annotated gene sequences. However, the two ends of the insert cDNA demonstrate similarity to unrelated gene sequences, and therefore appear to be chimaeric. The gene symbol and gene name with which similarity is observed is given, followed by the chromosomal location of that locus within the human genome, taken from the Human Genome Nomenclature Database.

Clone	Symbol	Gene	HSA location
079F	HP	haptoglobin alpha	16q22.1
079R	FMO3	flavin-containing monooxygenase	1q23-q25
130F	DDX5	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 5 (RNA helicase, 68kD)	17q21
130R	n/a	Human ribosomal DNA complete repeating unit	acrocentric p-arms
148F	PIK3R1	phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 1 (p85 alpha)	5q12-q13
148R	n/a	18S rRNA	acrocentric p-arms

However, the reverse data show similarity to the FMO3 gene from HSA 1q32-q35. This gene is termed flavin-containing monooxygenase, but the gene product is also referred to as dimethylamine monooxygenase, as seen in the dbEST match reported for this clone. The proportion of identity involved is 78% in both databases, over a region of 385bp in the non-redundant database, and 226bp in dbEST. Thus there is strong evidence to suggest that this clone contains sequence data from these two apparently unrelated genes, HP and FMO3.

LEI130

Forward data for clone LEI130 appear to contain part of the gene encoding the nuclear P68 protein (DDX5), which has been assigned to HSA 17q21. Identity is 95% over 468bp in the non-redundant database, and 96% over 451bp in dbEST. Thus both are highly significant matches. Reverse data, however, report 100% identity with human ribosomal DNA (accession no. U13369), over more than 400bp. The situation is further complicated by the fact that the reverse data identify significant identity with a pig cDNA within dbEST. The region of similarity extends again in excess of 400bp, with 98% identity. The most likely explanation would seem to be that this anonymous pig cDNA itself represents ribosomal DNA. This can be confirmed by obtaining sequence data for the pig cDNA (accession no. AJ241168.1) and using this to perform a non-redundant database similarity search. The result is a 98% match over 540bp with the human 18S ribosomal RNA sequence (accession no. X03205). Clone LEI130 can therefore be described as containing partial sequence for the DDX5 gene at one end of the insert, and ribosomal DNA at the other.

LEI148

Forward sequence data for this clone identify significant similarity to existing entries for the PI3-kinase-associated p85 protein (PIK3R1) from HSA 5q12-q13 in both databases searched. The region of similarity extends over almost 300bp in the non-redundant database, and nearly 200bp in dbEST. Reverse data, however, appear to represent ribosomal RNA sequence, since significant identity is also found in both databases with such entries. In the non-redundant database, 99% identity is found with 18S ribosomal RNA from the rabbit, whilst in dbEST the highest reported match is 99% identity with a rat cDNA which is itself similar to the mouse 45S pre rRNA gene.

- **CLASS D:**

Clones containing repetitive sequence

Nine clones, representing 11.8% of the total analysed, contain sequences that are known to be present in multiple copies within the genome.

LEI049, 071, 154

These clones appear to represent regions of repetitive DNA, and the highest reported matches in each case are with dog sequences. LEI049 is shown to contain sequence similar to the canine LINE-1 element, with 85% identity over 340bp, in the non-redundant database. A search of the EST database identifies an expressed sequence tag from the dog, with 87% identity over a less extensive region of 107bp. This suggests that this region of LEI049 contains part of a LINE-like repetitive element that is also present within the dog dbEST entry.

LEI071 identifies similarity with a dog microsatellite repeat (ZuBeCa2, accession no. AJ003059) within the non-redundant database, with 85% identity over only 68bp. This marker has been assigned to CFA 1q21 (Schläpfer *et al.*, 1998). The highest match reported in dbEST is with a dog mRNA sequence, with 81% identity over just 70bp. The level of significance of these matches is low, and would suggest that LEI071 also contains a region of repetitive DNA. Forward sequence data from LEI154 also appear to contain part of a LINE-1 repetitive element, with in excess of 80% identity over 284bp and 118bp with LINE-1 sequences in both the non-redundant and EST databases respectively.

LEI119, 162, 188, 199

The latter three clones demonstrate significant similarity to dog cDNA sequences within dbEST. Comparison of all four clones with entries from the non-redundant database also demonstrates their similarity to previously characterised dog gene sequences, namely part of the MHC beta chain, tyrosine aminotransferase, chymase and pancreatic colipase. The proportion of identical bases ranges from 85% to 96%, over regions of between 30bp and 142bp. It would be expected that matches with existing dog gene sequences would be significantly more extensive than the above data would suggest. The proportion of identical bases would be expected to near 100%, although allowing for sequencing error a more realistic estimate may be 95%. The region of identity would also be expected to be far more extensive.

Further investigation, described in section 3.2, demonstrates that the above four clones contain regions of repetitive DNA that are known to be associated with several genes, and that are classed as SINE elements. The fact that these genes contain part of this repeat element therefore results in identity being reported with a specific gene sequence when in fact the match is solely associated with the repeat region.

LEI104, 146

Forward and reverse sequence data for these clones demonstrate in excess of 90% identity over almost 500bp with ribosomal DNA, in the non-redundant database, enabling a confident assignment of their identity. The same data also identify anonymous cDNA sequences from pig and human cDNAs within dbEST, also with over 90% identity over more than 400bp. It is therefore likely that these cDNAs in fact represent ribosomal DNA. This is confirmed by performing a database similarity search using these cDNA sequences. The former demonstrates 98% identity over 540bp with human 18S ribosomal RNA, the latter 98% identity over 557bp with the human ribosomal DNA complete repeating unit (accession no. U13369).

- **CLASS E:**

Clones sharing similarity with microbial sequence

Eight of the 76 clones analysed (10.5%) demonstrated significant sequence similarity to database entries derived from microbial organisms, and are likely to represent contaminants of the cDNA library.

LEI001, 017, 052, 095, 105, 140, 156, 157

These clones showed highly significant similarity to sequence data from the yeast *Saccharomyces cerevisiae*. LEI001, 017, 052, 105 and 156 all show in excess of 90% identity with *S. cerevisae* sequence over a region of at least 459bp. In each case, this is so for further database matches reported. This supports the conclusion that these clones represent microbial sequence. Reverse sequence data obtained from LEI052 also confirms the nature of the insert DNA. No significant matches were found within dbEST for any of the five clones.

LEI095 and LEI157 also appear to represent microbial sequence, although the significance of the database matches is lower in each case compared to that for the five clones described above. The percentage of identity is reduced to 79% and 89%

respectively. The former in particular is at the lower limit of the criteria used to describe significant matches in this study. LEI095 forward data show 82% identity with an *S. cerevisae* cDNA clone in dbEST, over a total of 241bp. The region over which identity exists is extensive in all these cases. It therefore does appear that the sequences are of microbial origin although perhaps derived from a species other than *S. cerevisae*, which would explain the reduced overall percentage identity observed with data from this species.

LEI140 differs from the above in that significant similarity was found with a previously annotated yeast gene in a search of the non-redundant database with forward sequence data. This identified 91% identity in a region of more than 450bp, with the *S. cerevisae* serine hydroxymethyltransferase II gene. A less significant match was found with a cDNA from another yeast within dbEST, which may therefore represent the orthologous gene in this species.

- **CLASS F:**

Clones representing mitochondrial DNA

Three clones (3.9% of the total analysed) contain regions of sequence derived from the mitochondrial genome.

LEI055, 167, 183

These clones demonstrate significant matches to the dog mitochondrial genome sequence. Identity is between 90% and 99% in each case, and extends over between 325bp and 457bp. Sequence data for all three clones also share similarity with anonymous cDNA sequences within the EST database. This implies that these ESTs also contain regions of mitochondrial sequence.

- **CLASS G:**

Clones for which significant identity is shared with an existing database sequence, but which do not fit into the above categories.

Database search results for eight clones (representing 10.5% of the total analysed) demonstrated significant similarity to existing sequences, but cannot be assigned to any of the previously described categories.

LEI008, 069, 098, 113, 114

These clones demonstrate similarity to anonymous cloned human genomic DNA in the non-redundant database. The region over which similarity extends does not exceed 120bp in any case, and the percentage identity ranges from 86% to 93%. Only LEI098 identifies a match in dbEST, which in this case is with a human cDNA clone that contains sequence corresponding to the LINE-1 repetitive element. These LEI sequences therefore do not appear to correspond to any known coding sequence.

LEI083, 092

Both clones demonstrate significant identity with a human mRNA for 16G2, with 81% and 85% identity, respectively, over 262 bases. This sequence (accession no. AB017551) was recently added to the non-redundant database, representing a gene reported as specifically expressed in the human liver (T. Aihara, Y. Miyoshi, and Y. Nakamura, unpublished data). No additional information regarding the function or location of this sequence is available to date. Forward sequence data from both dog cDNA clones share similarity with anonymous cDNAs in dbEST, which may therefore also represent part of the 16G2 gene.

LEI094

Forward sequence data for this clone identified a match with an entry in the non-redundant database corresponding to the KIAA0370 gene (accession no. AB002368). Eighty eight percent identity is observed over 360bp of this database entry, which represents an uncharacterised gene described by Nagase *et al.* (1997). No other significant matches were found with sequences within this database. It maybe therefore be that this clone represents part of a rare transcript that has not been annotated by previous studies. The same forward sequence highlighted similarity to an anonymous human cDNA in dbEST, with 89% identity over 369bp. Reverse data did not generate any significant matches in either database.

- **CLASS H:**

Clones for which no significant similarity was found with any existing database entry

Thirty clones (39.5%) did not demonstrate any significant match with existing entries in either the non-redundant or EST database. These are listed in table 3.1e.

Table 3.1e:

Class H clones, for which no significant similarity was found with existing database entries.

No significant database match				
LEI009	LEI054	LEI103	LEI139	LEI179
LEI016	LEI074	LEI109	LEI155	LEI180
LEI022	LEI085	LEI121	LEI159	LEI181
LEI023	LEI088	LEI123	LEI161	LEI187
LEI025	LEI093	LEI131	LEI165	LEI189
LEI039	LEI100	LEI135	LEI177	LEI191

Table 3.1f:

Overview of categories of BLAST matches observed. The number of clones falling into each category is shown in column three.

Class	Description	Number	% of total
A	share significant identity with a previously annotated gene, with both forward and reverse sequence data	13	17.11
B	share significant identity with a previously annotated gene, with either forward or reverse sequence data only	2	2.63
C	share significant identity with a previously annotated gene with forward sequence data, but with a different gene with reverse data	3	3.95
D	contain repetitive sequence	9	11.84
E	share similarity with microbial sequence	8	10.53
F	represent mitochondrial sequence	3	3.95
G	share significant identity with an existing database sequence, but clones do not fit into the above categories	8	10.53
H	no significant similarity to any existing database entry	30	39.47
		76	Total = 100%

A summary of the data for each category is provided in table 3.1f, and the highest scoring BLAST match for each cDNA clone is listed in tables Ci and Cii in Appendix C. Sequence data for clones within classes A, B and C appear in the EMBL Nucleotide Sequence Database under accession numbers AJ402290 to AJ402324.

3.2 Further analysis of repetitive sequences

3.2.1 Clones containing putative SINE elements.

The dog SINE has previously been described by Minnick *et al.* (1992). The GenBank entry for this sequence (accession no. X57357.1) was used to search the non-redundant database to identify other entries containing part of this repeat element. The resulting output demonstrated that a region highly similar to the SINE is present in association with several gene sequences of the dog. Those reported within the 50 most significant matches are listed in table 3.2a. Amongst these are the four genes referred to in section 3.1 (class D clones) which highlighted the presence of the SINE element within clones LEI119, 162, 188, 199, namely the MHC beta chain, tyrosine aminotransferase, chymase and pancreatic colipase. No sequences were reported amongst the fifty most significant matches for any non-canid species, which reflects the specificity of this SINE for the dog. The proportion of bases identical between the LEI clone and the reported gene match ranged from 85% to 96%, over regions of between 30 and 142bp. Forward data for clone LEI162 appear to contain only a small region of SINE sequence since the database match extends only over a stretch of 30 bases within the initial 40 bases of the cDNA sequence. This region contains a (CT)-rich region, which is likely to represent part of the microsatellite-like sequence that is commonly associated with the 3' end of the dog SINE (Minnick *et al.*, 1992). The initial 40 bases of the LEI162 sequence were removed, and the database search was repeated. No significant match was reported, indicating that the moderately significant match reported previously was solely due to the presence of the (CT)-rich repeat motif both in the cDNA clone sequence and in association with the dog SINE.

Clones LEI119, 188 and 199 share a more extensive region of similarity to the dog SINE, over between 116 and 142 bases. The fact that the consensus SINE is 130bp in length therefore correlates with the range of these findings. The observed proportion of identical bases (85% to 96%) also compares well with the data summarised in table 3.2a, which range from 87% to 97% identity. It is therefore likely that these three dog cDNA clones contain the majority of the dog SINE sequence. This is demonstrated in figure 3.2a, a

Table 3.2a:

Dog genes associated with SINE elements

Locus name	Accession no.	Extent of match
pancreatic colipase	M63427	127/130 (97%)
retinal guanylate cyclase E	Y15484	124/130 (95%)
tyrosinase-related protein 2 gene	AF029683	124/130 (95%)
cGMP-gated channel alpha subunit	Y11309	123/130 (94%)
MHC class I DLA-53 gene	U55029	122/130 (93%)
retinoblastoma protein (RB1) gene	AF155737.1	121/130 (93%)
chymase gene	U89607	118/130 (90%)
microsatellite DNA, ZuBeCa2	AJ003059	117/130 (90%)
tyrosine aminotransferase gene	L47165	101/111 (90%)
sphingolipid Ca2+ release mediating protein	U33628	72/78 (92%)
CLN3 gene	AF033657	72/80 (90%)
MHC class II DLA DRB1 beta chain	U47339	69/76 (90%)
MHC DLA Class II DRB pseudogene	U46916	72/80 (90%)
rhodopsin	Y09004.1	60/66 (90%)
cytochrome P450 2C21 (CYP2C21)	AF049909	53/57 (92%)
PDEA gene	Y13283	70/80 (87%)

```
X57357.1     1 ---------------------------------------------------GGATCCCTGGG
188F       475 CACATACACAATATGTTAGGATAATAAGTGCTGTAAGAAACCCTAGGGGGGGATCCCTGGG
199F       301 --CTTTATCATTTCATTAAGA-AATATTTACTAT--------CTTGG--GCAGCCCCGGG-
119F       330 TAACTAGTGAAAAAATTCTAACAACATTGAATAA---------AAAGAGGGATGCCTGGG

X57357.1    12 TGGCGCA-CGGTTTGGCGCCTGCCTTTGGCCCAGGGAGCGATCCTGGAGACCCGGGATCG
188F       535 TGGCTCAGCAGTTTAGCGCCTGCCTTCGGCCCAGGGTGTGATCCTGGAGACCCAGGATAG
199F       347 TGGCGCAGCGGTTTANAGCC-GCCTGCARCCCAGGGCGTGATCCTGGAGACCCTGGATCA
119F       381 TGGCTTAGTGGTTGAGCGTCTGCCTTGGGCTCCGGGCATGATCCTGCATTCCTAGGATCG

X57357.1    71 AATCCCAC-ATCGGGCTCCCGGTGCATGGAGCCTGCTTCTCCCTCTGCCTGTGTCTCTGC
188F       595 AGTCCCAC-ATCGGGCTTCCTG--CATGGAGCCTGCTTCTTCCTCTGCCTGTGTCTCTAC
199F       406 AGTCCCAC-GTCAGGCTCCCTG--CATGGWGCCTGCTTCTCCCTCTGCCTGTGTCTCTGC
119F       441 AGTCCCACCATCAGGCTCCCTT--CAAGGAGGCTGCCTCTCCCTCT--------CTCCGG

X57357.1   130 G-------------------------------------------------------------
188F       652 CTCTCTCTCTCTCTCTATCTCTCATGAATAAATAAATAAAATCTTTAAAAAAAAGAAGAA
199F       463 NTCTCAATCTCTCTCTCTGTGKTCCAT----------------------------------
119F       491 ATT-------------------------------------------------------------
```

Figure 3.2a:

Alignment of the dog consensus SINE sequence (accession no. X57357.1) with data from clones LEI119, 188 and 199. Nucleotides that are identical between two or more sequences are highlighted in black, demonstrating that in excess of 80% identity is shared between the SINE motif and the SINE-like region of each of the three dog cDNAs.

multiple alignment of the consensus dog SINE sequence against data from LEI119, 188 and 199. Clone LEI199 data share similarity with the dog SINE in the same sense, whilst the reverse complement of data for LEI119 and 188 must first be derived in order to construct the alignment, since the SINE-like sequence is present on the complementary strand. Only a small number of deviations are seen between the cDNA clone data and the SINE consensus data, demonstrating the sequence conservation of the repeat motif. The degree of similarity between the 130bp SINE sequence and the corresponding region within the cDNA clone is 81.5%, 93% and 93.1% respectively for data from LEI119, 188 and 199. As with LEI162, the SINE motif was removed from the clone data, and the BLAST search was repeated. No significant matches were reported, again indicating that previous matches were solely due to the presence of the repeat sequence.

In section 3.1, LEI071 was found to share similarity with a microsatellite repeat (ZuBeCa2, accession no. AJ003059) within the non-redundant database, with 85% identity over only 68bp. Table 3.2a shows that the ZuBeCa2 sequence is also associated with a SINE element, sharing 90% identity over the 130bp of the repeat sequence. The region of repetitive DNA proposed to be present within the LEI071 sequence data can more accurately be described as part of the dog SINE, as a result of this database search. The nature of the dog SINE element is described in more detail in the discussion at the end of this chapter.

3.3 Further analysis of selected clones

3.3.1 Strategy for amplification of a partial genomic DNA sequence for dog genes

As discussed in chapter one, previous literature has indicated that the assignment of cDNA clones by standard FISH analysis is a relatively inefficient technique, and consequently an alternative approach was investigated. This involved the design of locus-specific PCR primers within the cDNA to enable amplification of the corresponding region of genomic DNA. Should this region contain intron sequence, the resulting product would be of a larger size. Careful selection of candidates for study would thus allow the conversion of an EST to a genomic PCR product of a size more amenable to FISH analysis.

At this time, dog genomic DNA sequence was not available for any of the genes for which a corresponding EST had been found in this study. It was therefore not possible to determine directly the exon structure of the gene. In the absence of this information, there would be a risk of primer design within a region of the dog cDNA sequence that spans

two separate exons. The application of such primers to genomic DNA would not be successful in this case, since the region would be interrupted by intron sequence. A potential approach for determining this information was by comparison of the dog cDNA sequence with existing database entries for the gene from other species. The assumption would be that intron-exon structure would be sufficiently conserved to enable this information to be extrapolated to the dog.

3.3.2 Selection of clones for further study

In order to investigate this assumption, four markers were chosen for further study, LEI042, 053, 077 and 122. These were suitable candidates since the complete sequences for the corresponding human genes were available within GenBank. For each locus, the dog cDNA sequence and the human gene sequence were compared using the ClustalW alignment program, as described in section 2.3.5. Those regions of the cDNA that share significant sequence similarity with the corresponding human exon were identified. This information was used to determine a suitable site for PCR primer design, such that the primers did not lie in any region of the cDNA sequence that spanned more than one of the exon regions defined by alignment with the human gene. Alignment data also indicated the proportion of the human gene sequence that was represented by the corresponding dog cDNA. This approach is demonstrated in detail using clone LEI042 as an example.

3.3.3 Application of the selected strategy using clone LEI042 as an example

A complete gene sequence for human HNRPA2B1 is present in GenBank (accession no. D28877), containing 11390 nucleotides of sequence data (Kozu *et al.*, 1995). The alignment of exons of this human gene sequence with the dog cDNA from clone LEI042 is shown in figure 3.3a. The reverse sequence data from clone LEI042 represent the sense strand. Consequently, it is necessary to derive the reverse complement of the forward sequence data in order to align this with the human gene sequence. Human exons are highlighted in red, demonstrating that these regions directly correspond to the dog cDNA sequence. The sole exception is exon two of the human gene, since no dog cDNA sequence aligns with this region. It therefore appears that sequence corresponding to exon two is not present within the dog coding sequence for this clone. Characterisation of the human HNRPA2B1 gene by Biamonti *et al.* (1994) and Kozu *et al.* (1995) has demonstrated that two splice variants of the gene naturally exist. These result in the

Figure 3.3a: (overleaf)

Alignment of clone LEI042 with the human HNRPA2B1 gene. The upper strand in each pair of aligned sequences represents i) reverse and ii) forward sequence data from LEI042. These data are aligned with exons from the complete genomic sequence for the human HNRPA2B1 gene. Exon data are shown in red, and flanking intron sequence in black. The site of primer design for PCR amplification of the dog genomic product are highlighted in grey.

i)

Human exon 1 (nucleotides 2427 - 2601)

```
CFA     ------------------------------------------------------------
HSA     AAAGCGGCGGCAGCGGCTCTAGCGGCAGTAGCAGCAGCGCCGGGTCCCGTGCGGAGGTGC
              2410      2420      2430      2440      2450      2460

CFA     ------------------------------------------------------------
HSA     TCCTCGCAGAGTTGTTTCTCGAGCAGCGGCAGTTCTCACTACAGCGCCAGGACGAGTCCG
              2470      2480      2490      2500      2510      2520

                                30        40        50        60
CFA     ----------------------CTCTCTCATCTCGCTCGGCTGCGGGAAATCGGGCTGAA
                              ::::::::::::::::::::::::::::::::::::::
HSA     GTTCGTGTTCGTCCGCGGAGATCTCTCTCATCTCGCTCGGCTGCGGGAAATCGGGCTGAA
              2530      2540      2550      2560      2570      2580

          70        80
CFA     GCGACTGAGTCCGCGATGGAG--------------------------------------
        :::::::::::::::::::::
HSA     GCGACTGAGTCCGCGATGGAGGTAACGGGTTTGAAATCAATGAGTTATTGAAAAGGGCAT
              2590      2600      2610      2620      2630      2640
```

Human exon 2 (nucleotides 5308 - 5343)

```
CFA     ------------------------------------------------------------
HSA     TTGATGTAAGTTTTTTTTCCCCCCAGAAAACTTTAGAAACTGTTCCTTTGGAGAGGAAA
              5290      5300      5310      5320      5330      5340

CFA     ------------------------------------------------------------
HSA     AAGGTACTCTGCCAGCAGGTCACCTCATATTTAAGAATTTAATTTCCTGCATACAAAGAG
              5350      5360      5370      5380      5390      5400
```

Human exon 3 (nucleotides 5442 - 5552)

```
                                                  90        100
CFA     ----------------------------------------AGAGAAAAGGAACAGTTCC
                                                ::::::::::::::::::
HSA     GAAAATGTAAATAAAAATTGAAATGGTATTTTCCTTTGCAGAGAGAAAAGGAACAGTTCC
              5410      5420      5430      5440      5450      5460

         110       120       130       140       150       160
CFA     GTAAACTCTTCATTGGTGGCTTGAGCTTTGAAACCACAGAAGAAAGTTTGAGGA-CTACT
        :::: ::::: :::::::::::: :::::::::::::::::::::::::::::: :::::
HSA     GTAAGCTCTTTATTGGTGGCTTAAGCTTTGAAACCACAGAAGAAAGTTTGAGGAACTACT
              5470      5480      5490      5500      5510      5520

           170       180       190       200
CFA     AC-AACCATGGGGGAAAAACTTACCGACTGTGTNGTT-----------------------
        :: ::: ::::::  ::: ::::: :::::::::.::
HSA     ACGAACAATGGGG--AAAGCTTACAGACTGTGTGGTATGTAAATTACTGAATTGTTACTG
              5530      5540      5550      5560      5570
```

Human exon 4 (nucleotides 5712 – 5858)

```
                                 210       220       230       240
CFA     ----------------ATGANAGATCCTGCCAGCNAAAGATCAAGAGGATTTGGTTTTGT
                         ::::. :::::::: :::.:::::::::::::::::::::::::
HSA     TTTGGTTTGAAAGGTAATGAGGGATCCTGCAAGCAAAAGATCAAGAGGATTTGGTTTTGT
      5700      5710      5720      5730      5740      5750

          250       260       270       280       290       300
CFA     GACCTTTTCATCCATGGCTGANGTTGATGCTGCTATGGCTGCAAGACCTCATTCAATTGA
         :: :::::::::::::::::.::::::::::: :::::::::::::::::::::::::::
HSA     AACTTTTTCATCCATGGCTGAGGTTGATGCTGCCATGGCTGCAAGACCTCATTCAATTGA
      5760      5770      5780      5790      5800      5810

          310       320       330       340
CFA     TGGGAGA-TGGTTGAGCCNAAACGTGCTGTTGCAAGANAAG-------------------
        ::::::: : ::::::::.::::::::::: ::::::.: :
HSA     TGGGAGAGTAGTTGAGCCAAAACGTGCTGTAGCAAGAGAGGTAAGCAAACAATGACTGTC
      5820      5830      5840      5850      5860      5870
```

Human exon 5 (nucleotides 6156 – 6366)

```
                                                  350       360
CFA     ---------------------------------------AATCTGGAAAACCNGGGGCTCA
                                               :::::::::::::.::::::::
HSA     TTCTGTTTGGATGACTTTTACCCCACCACTATTTTAGGAATCTGGAAAACCAGGGGCTCA
      6120      6130      6140      6150      6160      6170

        370       380       390       400       410       420
CFA     TGTTACTGTNAAAAAACTCTTTGTTGGGTGGAATTTAAGAAAGATTCTGAAGAAC--CCC
        ::: :::::.:: :: :: ::::::::  :::::: ::::: ::: :::: ::::  : :
HSA     TGTAACTGTGAAGAAGCTGTTTGTTGGC-GGAATTAAAGAA-GATACTGAGGAACATCAC
      6180      6190      6200      6210      6220      6230

                430       440
CFA     CT---------CTTTNAANATTAT-----------------------------------
        ::         ::::.: .: :::
HSA     CTTAGAGATTACTTTGAGGAATATGGAAAAATTGATACCATTGAGATAATTACTGATAGG
        6240      6250      6260      6270      6280      6290

CFA     ------------------------------------------------------------

HSA     CAGTCTGGAAAGAAAAGAGGCTTTGGCTTTGTTACTTTTGATGACCATGATCCTGTGGAT
        6300      6310      6320      6330      6340      6350
```

ii)

Human exon 5 (nucleotides 6156 – 6366)

```
                                                  840       850
CFA     ----------------------------------------TCTGGAAAACCAGGGGCTCAT
                                                :::::::::::::::::::::
HSA     TCTGTTTGGATGACTTTTACCCCACCACTATTTTAGGAATCTGGAAAACCAGGGGCTCAT
    6120      6130      6140      6150      6160      6170

          860       870       880       890       900       910
CFA     GTAACTGTGAAGAAGCTCTTTGTTGGTGGANTTAAAGAAGATACTGNGGAACACCATCTT
        :::::::::::::::::: :::::::: ::: :::::::::::::::: :::::: :: :::
HSA     GTAACTGTGAAGAAGCTGTTTGTTGGCGGAATTAAAGAAGATACTGAGGAACATCACCTT
    6180      6190      6200      6210      6220      6230

          920       930       940       950       960       970
CFA     AGAGATTATTTCGAGGAATATGGAAAGATTGNTACCNNTGAGATACNTACTNATAGNTAG
        :::::::: :: ::::::::::::::.:::: :::: .:::::::: .:::::.:::::. ::
HSA     AGAGATTACTTTGAGGAATATGGAAAAATTGATACCATTGAGATAATTACTGATAGGCAG
    6240      6250      6260      6270      6280      6290

          980       990      1000      1010      1020      1030
CFA     TTTGGAAAGAAAAGAGGCTTNGGTTTTGTTACCTTTGATGATCATGATCCTGTGGATAAG
        : ::::::::::::::::::.:: :::::::: :::::::: ::::::::::::::::::.
HSA     TCTGGAAAGAAAAGAGGCTTTGGCTTTGTTACTTTTGATGACCATGATCCTGTGGATAAA
    6300      6310      6320      6330      6340      6350

                                                                 1040
CFA     AT-------------------------------------------------------TGTGT
        ::                                                        :::::
HSA     ATCGTATGTAAGTGTCTAACCACAAATGTACTGTTTTTTTCCAGTGTATCAATTTTGTGT
    6360      6370      6380      6390      6400      6410
```

Human exon 6 (nucleotides 6514 - 6615)

```
                                                 1050      1060
CFA     ----------------------------------TGCAGAAATATCATACCATCAATGGT
                                          :::::::::: ::::::::::::::
HSA     TTAATTTAAAAGGAGTCTGAATTTTTCATTCCAGTGCAGAAATACCATACCATCAATGGT
    6480      6490      6500      6510      6520      6530

     1070      1080      1090      1100      1110      1120
CFA     CATAACGCAGAAGTAAGAAAGGCTTTGTNTAGACAAGAAATGCAGGAAGTCCAAAGTTCT
        ::::: :::::::::::::::::::::: :::::::::::::::::::: ::.::::::
HSA     CATAATGCAGAAGTAAGAAAGGCTTTGTCTAGACAAGAAATGCAGGAAGTTCAGAGTTCT
    6540      6550      6560      6570      6580      6590

     1130      1140
CFA     AGAAGTGGAAGAGGAGG-------------------------------------------
        ::.::::::::::::::
HSA     AGGAGTGGAAGAGGAGGTAATTTAATTCTGTTCTCTTTATTTTTGTTCATATATAAGGGC
    6600      6610      6620      6630      6640      6650
```

Human exon 7 (nucleotides 6688 - 6768)

```
                                        1150      1160      1170
CFA     ----------------------------CAACTTTGGTTTTGGAGATTCTCGTGGTGGT
                                    ::::::::: :::::.:::::.::::::::
HSA     TTGCTTCTAACTGGGGCATTTATTGTAGGCAACTTTGGCTTTGGGGATTCACGTGGTGGC
    6660      6670      6680      6690      6700      6710

        1180      1190      1200      1210
CFA     GGTGGAAATTTTGGACCAGGACCAGGAAGTAACTTT------------------------
        :::::::::::: :::::::::::::::::::::::
HSA     GGTGGAAATTTCGGACCAGGACCAGGAAGTAACTTTAGAGGAGGATCTGGTGAGTTTCAA
    6720      6730      6740      6750      6760      6770
```

Human exon 8 (nucleotides 7258 – 7320)

```
CFA     ------------------------------------------------------------

HSA     ATTAGCAATAATGCTGAACAGTTTATAGTAAACGTAATCTTTTTTTGGCCCCTAACAGAT
    7200      7210      7220      7230      7240      7250

CFA     ------------------------------------------------------------

HSA     GGATATGGCAGTGGACGTGGATTTGGGGATGGCTATAATGGGTATGGAGGAGGACCTGGA
    7260      7270      7280      7290      7300      7310
```

production of either the HNRPA2 or the HNRPB1 protein. The latter contains all 12 exons of the gene, whilst the former differs by the exclusion of exon two during processing of the mRNA transcript. The absence of sequence corresponding to exon two within the dog cDNA clone suggests that LEI042 represents part of the HNRPA2 variant. The locations of the primers designed to the dog cDNA sequence are highlighted in figure 3.3a. The intron/exon structure of the human gene sequence, and the alignment with the LEI042 cDNA sequence, is shown schematically in figure 3.3b. The sites at which PCR primers were designed are also indicated. The sense primer (designed from the LEI042 reverse sequence data) lies in the region corresponding to exon one of the HNRPA2B1 gene. The antisense primer (designed from the reverse complement of the LEI042 forward sequence data) lies in exon six. The primers span a region corresponding to nucleotides 2575 to 6590 of the human gene, a total of 4015 nucleotides. This information was extrapolated to predict that the product obtained from the application of these primers to dog genomic DNA would be approximately 4kb in length.

This assumption was tested by a standard PCR procedure, in which amplification was attempted from dog genomic DNA, and in a separate control reaction, from the cDNA template from which the primers were designed. The control reaction yielded a product of approximately 550bp, which correlated with the location of the primer sequence within the insert cDNA of clone LEI042. The corresponding product obtained from dog genomic DNA was approximately 4kb in length, but the yield of the product was markedly low compared to the control reaction. This was likely to reflect the increasing inefficiency of PCR amplification that is commonly observed for products in excess of approximately 3kb. Such limitations have been overcome to a great extent by the advent of long range PCR techniques (Barnes, 1994). These techniques utilise a combination of modified enzyme, buffer and amplification conditions to achieve amplification of products above this size with significantly greater efficiency than is possible by the standard PCR. Long range PCR typically uses a combination of two heat-stable enzymes, one of which enables efficient amplification of products in excess of 20kb, but with a relatively high rate of error. This is compensated for by the inclusion of a second, 'proofreading' enzyme that excises and repairs such errors and allows DNA extension to continue.

Long range PCR was performed using the primers designed from clone LEI042, using dog genomic DNA as the template. An annealing temperature of 60°C, with an elongation step of 4 minutes per cycle, generated a product of approximately 4kb. The

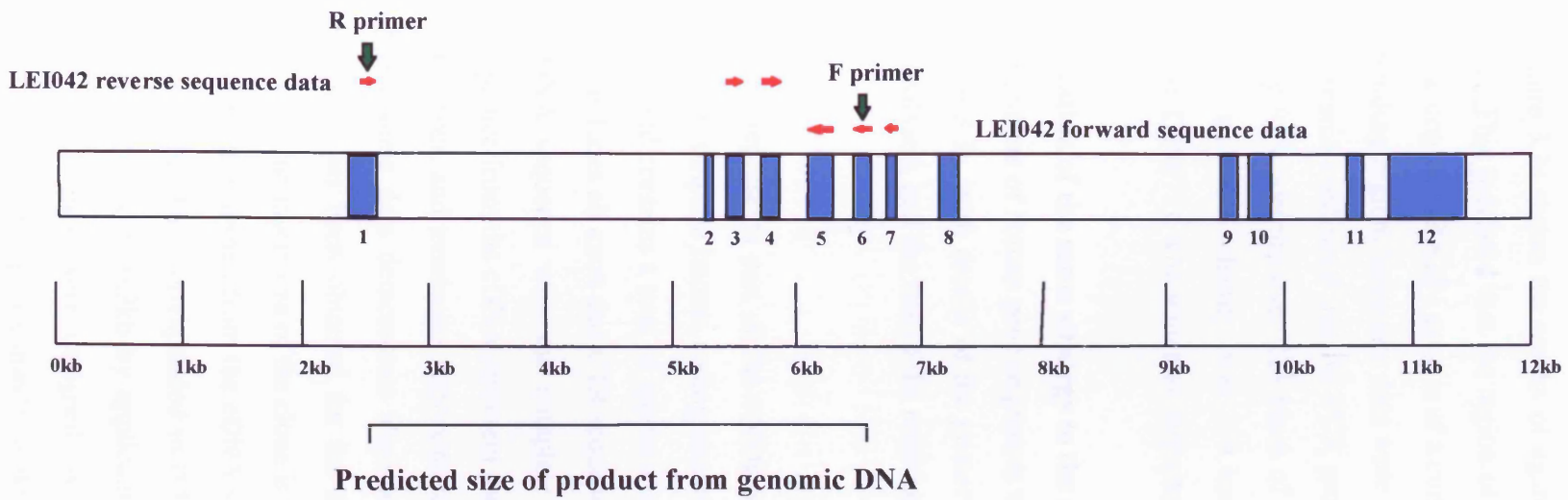


Figure 3.3b:

Schematic alignment of cDNA sequence from clone LEI042 and the human HNRPA2B1 gene. Exons within the human gene sequence are represented by blue blocks, and intervening introns are left blank. Below this is shown the approximate location of each intron and exon within the 12kb gene sequence, relative to the scale below marked in kilobase intervals. The relative location of clone LEI042 sequence is shown above this figure by red arrows, confirming that the dog cDNA sequence lies solely within human exon regions. The sites at which primers were designed for amplification from dog genomic DNA are indicated by vertical green arrows. These lie in exons one and six, resulting in the prediction of a 4kb product from dog genomic DNA.

yield of the product was comparable to that obtained from the cDNA control, thus long range PCR had successfully overcome the difficulties in amplifying this relatively large product. Figure 3.3c shows the results of agarose gel electrophoresis of the long range PCR product. This indicated that the region of non-coding sequence between exons one and six of the dog HNRPA2B1 gene is of a comparable size to that of the human gene in the corresponding region. Sequence data were obtained from the genomic PCR product. A database search confirmed that the PCR product represented part of the HNRPA2B1 gene, sharing 92% identity over 223 bases of the pig HNRPA2B1 gene (accession no. AF169290.1), and 92% identity over 219 bases of the human HNRPA2B1 sequence (accession no. D28877), with which an alignment was made with LEI042 data.

3.3.4 Application of the same strategy to the remaining cDNA clones selected

Accession numbers of human gene sequences with which alignments were performed are given in table 3.3a, with details of the primer sequences designed from the dog cDNA sequences analysed, and the size of the resulting PCR products.

LEI053

Clone LEI053 represents part of the coding sequence of the gene encoding fibrinogen beta chain. The complete human coding sequence for FGB exists in GenBank (accession no. M64983) and contains a total of approximately 8.9kb of sequence data (Chung *et al.*, 1991). This includes all exon and UTR sequence, and some intron data. Alignment of the LEI053 cDNA sequence with the complete human gene sequence showed that the forward sequence from the cDNA represent the sense strand. This extends from part way through exon four, and terminates approximately three-quarters of the way through exon seven. The reverse data demonstrate alignment over the same region, indicating that sequencing data has been obtained for the majority of the cDNA clone. This is not unexpected since the insert size of the clone is only 0.7kb.

PCR primers were designed from the cDNA sequence that spanned exons six and seven of the human gene. This corresponded to 0.52kb by comparison with the human gene sequence, as compared to 0.3kb by application of these primers to the LEI053 cDNA clone from which they were designed. Amplification using these internal primers generated a product of approximately 0.3kb using the cDNA clone as a template. Similarly, a product of approximately 0.55kb was obtained from a dog genomic DNA template (figure 3.3d). This indicated that the size of the intervening intron in both

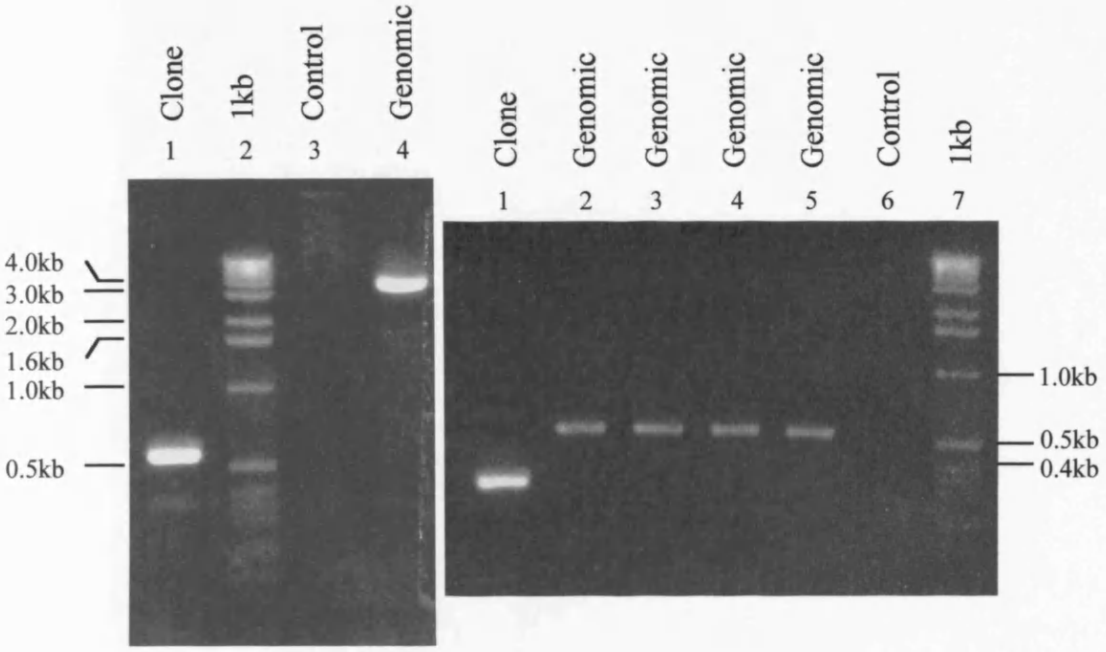


Figure 3.3c: (left)

Long range PCR analysis at the HRNPA2B1 locus. Amplification was performed using PCR primers designed from cDNA clone LEI042 sequence data, using both the clone (lane one) and dog genomic DNA (lane four) as the template. The product from the cDNA clone is approximately 0.55kb, whilst the same primers amplify a product of approximately 4.0kb (compared to the 1kb DNA ladder in lane two) demonstrating the presence of a region of intron sequence in the genomic DNA template. Lane three is the negative control, in which the PCR was performed with marker primers in the absence of a DNA template.

Figure 3.3d: (right)

PCR analysis at the FGB locus. Amplification was performed using LEI053 cDNA (lane one) and dog genomic DNA (lanes two to five) as the template. This demonstrates the presence of a small intronic region (approximately 0.2kb) within the genomic template flanked by the LEI053 primers. Lane six is the negative control, and lane seven is the 1kb DNA ladder.

Table 3.3a:

Primers designed for amplification of partial dog gene sequences. Primers were designed from sequence data obtained from four cDNA clones (column one). This was achieved by alignment of each cDNA with the corresponding human gene sequence (column two), the accession number of which is shown in column three. Primer sequences are given in columns four and five, and the size of the resulting product obtained from the cDNA clone is given in column six. The final two columns show both the expected and observed size of the product obtained from the same primers when used to amplify dog genomic DNA. The expected product size is derived by alignment of the dog cDNA sequence with the corresponding human complete gene sequence, and by assuming approximate correlation between the intron/exon composition of the gene in both species.

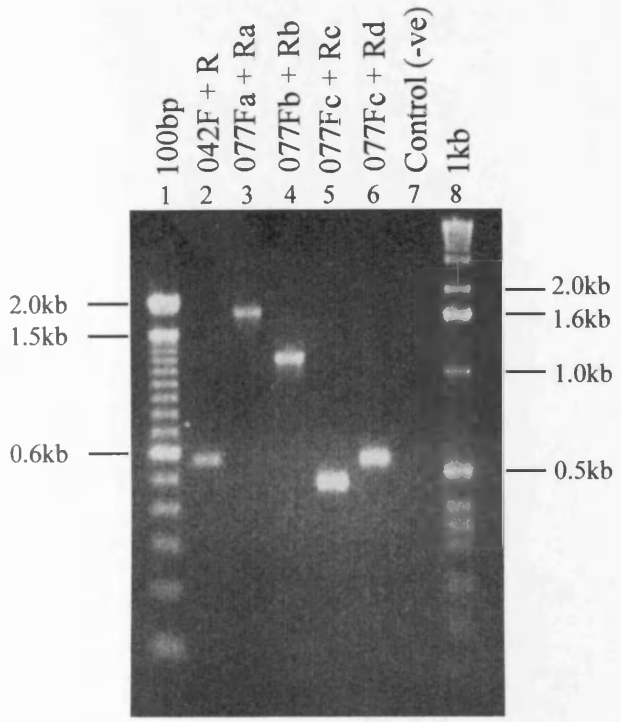
Clone	Gene	Accession no. of human gene	F Primer (5' to 3')	R Primer (5' to 3')	Product size from cDNA	Product size from gDNA	
						Expected	Actual
042	HNRPA2B1	D28877.1	ggacttcctgcatttcttctc	gctgaagcgactgagtccg	0.55kb	4.0kb	4.0kb
053	FGB	M64983.1	cagtggaggatggacagtga	cattgccagctgttccctta	0.3kb	0.52kb	0.55kb
077	ALB (a)	M12523.1	ccacaagcctttggcacaat	gctgcgcagcacttctctac	1.7kb	15.3kb	none
077	ALB (b)	M12523.1	gcctgtgcagctgaagagtc	caggacattctctctgattcag	1.2kb	11.0kb	10-11kb
077	ALB (c)	M12523.1	gagagattcaagtgtgccagc	ctctagtgtggcttcatattcc	0.5kb	4.8kb	4.2kb
077	ALB (d)	M12523.1	gagagattcaagtgtgccagc	ttctgaggctcatccacaagag	0.6kb	6.0kb	5.8kb
122	CYP2E	J02843.1	tcaggagcccgtatctcagg	gaatattctgcaggtggatatc	0.8kb	6.3kb	none

human and dog must be highly comparable. The genomic product was of a size unlikely to be amenable to chromosome assignment by FISH analysis, and was not pursued.

LEI077

Although sequence data are available for part of the dog ALB gene (accession no. AJ133489), these data were derived from an mRNA source, and do not contain any information on the intron/exon structure of the dog gene. Consequently, an alignment was performed between sequence obtained from clone LEI077 and the complete human ALB gene sequence (accession no. M12523). This entry (Minghetti *et al.*, 1986), is 19kb in length and provides complete information for the location and size of all 15 exons and the intervening introns. Forward sequence data from clone LEI077 represent the sense strand and extend from exon one into the beginning of exon four. Reverse sequence data begins at exon 13, and extends through to exon ten. The cDNA sequence thus spans nucleotides 1750 to 17080 of the human gene, a total of 15.33kb, more than 80% of the gene, and representing all but two of the 15 exons. Thus LEI077, with an insert of 1.9kb, can be considered as an almost full length cDNA clone for the ALB gene. Consequently, the design of PCR primers to enable amplification of part of the dog gene sequence was more difficult than for the other clones studied in this way. One pair of primers (077Fa and 077Ra) was designed within the forward and reverse sequence data for LEI077. Their location corresponded to a region spanning bases 1760 to 17040 of the human gene sequence, a total of 15.28kb from exon one to 13. These were used to generate further sequence data from within the cDNA clone from which new primers could be designed that would correspond to a product from genomic DNA that was within the capabilities of PCR amplification. These allowed the design of a second primer pair (077Fb and 077Rb) which spanned bases 4165 and 15105 (exons three to eleven) of the corresponding region of the human gene, a total of 10.94kb.

Both primer pairs were used in the PCR to ensure their ability to amplify from the cDNA template. The resulting products from pairs '077a' and '077b' were 1.7kb and 1.2kb respectively (figure 3.3e). In turn, long range PCR was attempted with both primer pairs using dog genomic DNA as a template. In the case of set '077a', although a product of approximately 1.7kb was amplified successfully from the cDNA positive control, no genomic product was generated. This was despite attempts to optimise conditions for amplification of large products by increasing the elongation time of the cycling program, and with the use of the range of buffers supplied with the enzyme. The positive control, a



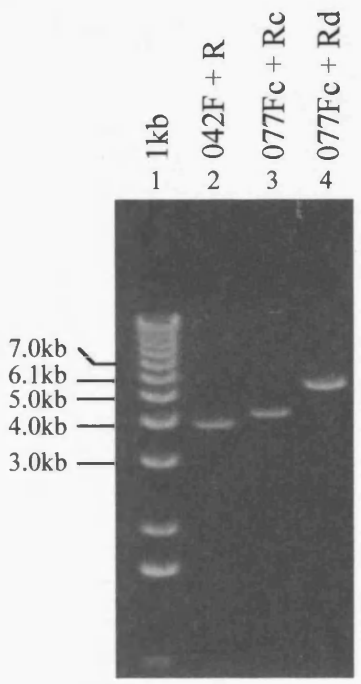


Figure 3.3e: (left)

PCR amplification from a cDNA template at the HNRPA2B1 and ALB loci The amplification product obtained from the application of the HNRPA2B1 primers to clone LEI042 cDNA are shown in lane two, following the 100bp DNA ladder. Four PCR primer combinations were developed from the LEI077 cDNA sequence, and were used to confirm their ability to amplify from the cDNA clone template. The resulting products are shown in lanes three to six, generating products ranging from 0.6kb to 1.7kb. The primer combinations used are shown above the corresponding lane. Lane seven contains the negative control reaction, followed by the 1kb DNA ladder in lane eight.

Figure 3.3f: (right)

Long range PCR amplification from a genomic DNA template at the HNRPA2B1 and ALB loci. The results of long range PCR using three primer pairs for two loci are shown. Lane one contains the 1kb DNA ladder, and lane two represents the product from 042F and 042R primers. Lanes three and four demonstrate the products obtained from genomic DNA using primers to the ALB locus, namely 077Fc and 077Rc (lane three), and 077Fc and 077Rd (lane four). Extended agarose gel electrophoresis enabled the size of the resulting products to be more accurately estimated as 4.0kb, 4.2kb and 5.8kb respectively, compared to the 1kb DNA ladder.

product of approximately 1.2kb, was also successfully generated from the cDNA clone with the use of primers 077b. In this case, a product of approximately 10kb to 11kb was produced from application of the same primers to dog genomic DNA, indicating that long range PCR had been successful. However, the product yield was very low. It therefore seemed that the corresponding size of the genomic product was at the limitations of the PCR amplification under these circumstances. Primers 077Fb and Rb were used to generate further sequence data from the internal region of the cDNA clone. This resulted in the design of three additional primers, 077Fc, 077Rc and 077Rd. The combination of 077Fc and 077Rc, and of 077Fc and 077Rd, corresponded to approximately 4.5kb and 6kb respectively of the human ALB sequence. Their application to cDNA successfully generated positive control products of 0.5kb and 0.6kb (figure 3.3e). In turn, long range PCR was attempted from dog genomic DNA. In both cases, this was found to be successful, with genomic products of approximately 4.2kb and 5.8kb produced with a high yield in each case (figure 3.3f). Therefore, as with LEI042, the relative intron size of both human and dog appeared to be highly comparable over the region of interest of the genes concerned.

LEI122

This clone corresponds to part of the cytochrome P450 IIE1 gene. A complete sequence for the human gene is available within the GenBank database (accession no. J02843) containing almost 15kb of sequence data, with intron and exon boundaries defined (Umeno *et al.*, 1988). Data from this clone extend between exon one and exon seven of the human gene sequence. The reverse clone data represent the sense strand, and span exons one to three, whilst the forward data align with exons seven, six and part of exon five. Primers were designed within exons one and six, spanning a total of 6.33kb of human gene sequence. Their application to the cDNA clone resulted in amplification of a product of 0.8kb. However, despite efforts to optimise conditions for long range PCR from dog genomic DNA, no product could be obtained.

3.4 Physical mapping of long range PCR products by FISH analysis

Long range PCR products were amplified using two primer pairs, namely 042F and R, and 077Fc and Rd, representing the HNRPA2B1 and ALB genes. The resulting products were approximately 4.0kb and 5.8kb respectively. One microgram of each product was labelled separately by incorporation of either biotin-16-dUTP or digoxigenin-11-dUTP,

by nick translation (section 2.2). The labelled probes were used in FISH analysis in the same manner as standard DNA probes with one exception. Due to their relatively small size in comparison with genomic clones, a greater quantity of each probe was hybridised to dog metaphases, such that approximately 250ng of probe were applied to the slide in each case. This quantity has routinely been found to be sufficient for FISH analysis of probes of 5kb to 10kb in size (M. Breen, pers. comm.). Each probe was competed with 10μg of sonicated dog DNA, and detection and analysis were performed with no further modifications.

Existing Zoo-FISH data from Breen *et al.* (1999a) and Yang *et al.* (1999) enabled the chromosomal location of both genes to be predicted prior to FISH analysis. HNRPA2B1 lies at HSA 7p15, within a region that is proposed to be evolutionarily related to the entire length of CFA 14q according to both Zoo-FISH reports. Dog HNRPA2B1 was therefore predicted to lie on CFA 14q. ALB lies within HSA 4q11-q13. HSA 4q13 lies at the extreme proximal end of an ECCS corresponding to CFA 3 according to Breen *et al.* (1999a), whilst no dog chromosome paint probe was observed to hybridise to HSA 4q11-q12. Yang *et al.* (1999) report the hybridisation of a paint probe representing CFA 15 within the region extending from HSA 4q12-q13mid, proximal to a region from HSA 4q13mid-q21mid that was painted by a CFA 13 paint probe. The chromosomal location of dog ALB could therefore be predicted with less certainty, but could be limited to CFA 3, 13 or 15 on the basis of existing Zoo-FISH data.

Both markers were confidently assigned on the basis of consistent and unique signal in all metaphases studied, representing a total of 15 cells for each probe. HNRPA2B1 was thus assigned to CFA 14q21.1 and ALB to CFA 13q22.2. Examples of the resulting images are given in figure 3.4a.

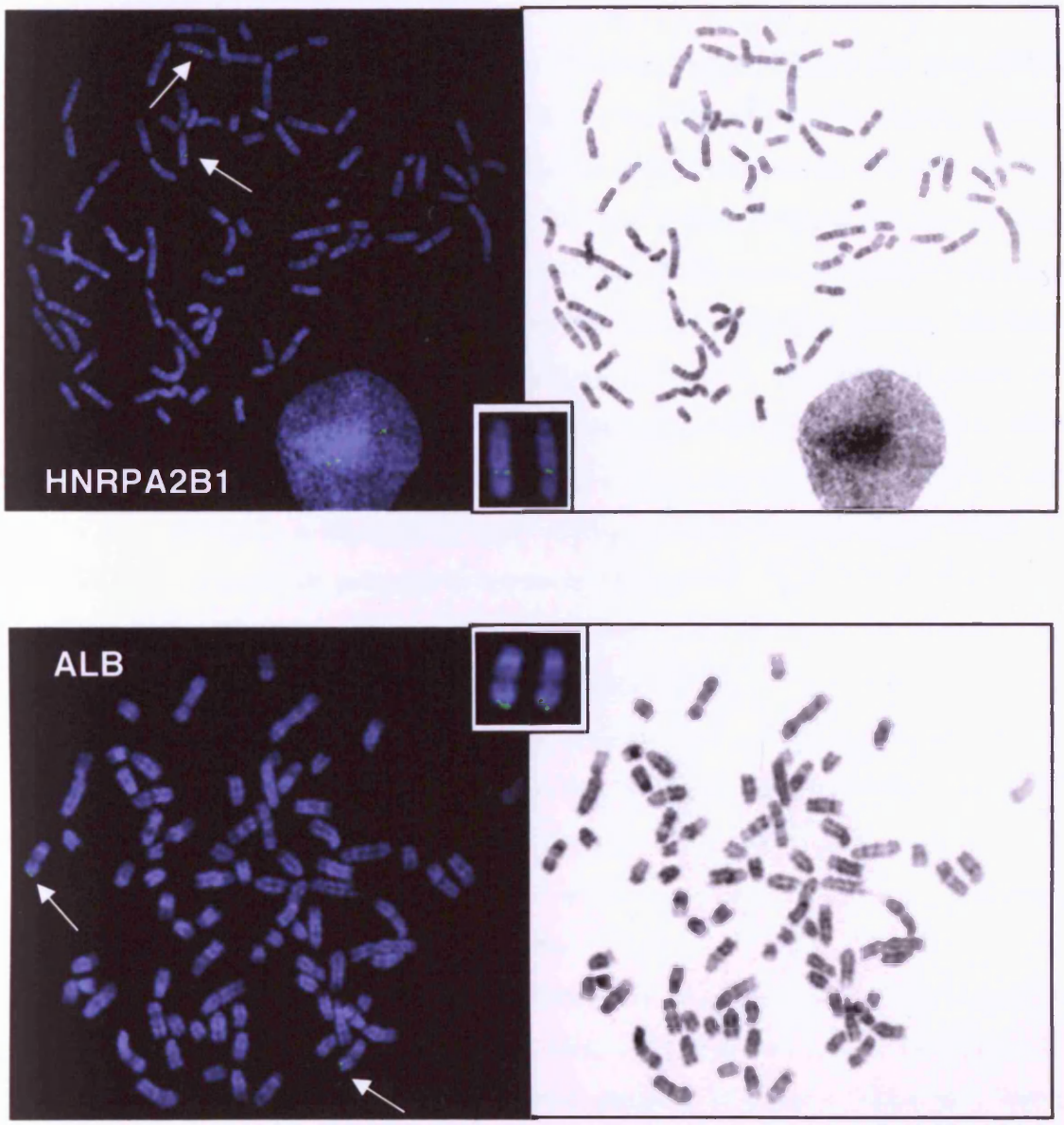


Figure 3.4a:

Chromosomal assignment of dog HNRPA2B1 and ALB by FISH analysis. Both probes demonstrated highly specific and consistent hybridisation to a single chromosomal location, indicated by the white arrows. HNRPA2B1 (upper image) was assigned to CFA 14q21.1, consistent with the location of the human orthologue on HSA 7p15. ALB lower image) was assigned to CFA 13q22.2, consistent with the location of the human orthologue on HSA 4q11-q13.

3.5 Discussion

Partial sequence data have been obtained for 76 clones isolated from a dog liver cDNA library, for the purpose of developing ESTs for the dog. Of the 76 clones, thirteen (17.1%) entirely satisfied the criteria set in this report for classification of a gene marker. Sequence data from both ends of the insert DNA of clones assigned to Class A were shown to share highly significant similarity with previously annotated genes from other species. The location of each of these gene loci within the human genome has been established via the Human Genome Nomenclature Database. The thirteen genes are spread throughout the human genome, with one marker on each of five autosomes (HSA 1, 10, 12, 14 and 18), and two markers on each of three further autosomes (HSA 4, 6 and 7). One gene (CYP2E) is represented twice amongst these markers, and the remaining gene (SKB1) has not been assigned to a human chromosome. Two cDNA clones have been assigned to Class B since they demonstrate significant similarity to existing gene sequences from HSA 6 and X, but only with sequence data from one end of the insert. This is likely to result from the fact that ESTs represent only partial gene sequences. As discussed in chapter one, sequencing analysis has frequently been performed only from the 3' end of poly(A)-selected transcripts. Consequently, many genes will only be represented in dbEST by data obtained from their 3' end rather than from the internal coding regions. Other projects have focussed on only 5' regions, while some have involved sequencing analysis of random internal coding sequences. These different approaches would explain a database match being found with one end of the insert, but not the other. Three cDNA clones have been assigned to Class C since they share significant similarity to different genes at each end of the insert DNA. Of the six sequences with which significant identity was found, two represent ribosomal sequences, which are located on the short arm of five human chromosome pairs. The four remaining gene sequences for which significant similarity was identified with Class C clones have been assigned to HSA 1, 5, 16 and 17. Thus partial sequences are available for 17 genes in total, located across 11 human autosomes and the X chromosome. Six of the genes identified in classes A to C are amongst those proposed by O'Brien *et al.* (1993) to represent 'anchored reference loci' for the generation of comparative maps. These genes are ALB, ALDH2, ASL, CYP2E, HP and PI. Their assignment will be of value in providing reference sites for comparison of the chromosomal location of type I markers in the genomes of different species. Furthermore, PI (alpha-1-antitrypsin) has previously been implicated in the development of chronic liver disease in the dog (Sevelius *et al.*,

1994). The isolation of a cDNA clone (LEI072) containing part of the dog gene may prove useful in investigating whether the disease has any hereditary basis. The EST sequences will form a valuable addition to existing resources for gene mapping in the dog, since sequence data for coding regions are limited for this species. As of March 31st 2000, 126 dog ESTs are present in dbEST, representing only 0.0033% of the total number of EST sequences contained within this database (http://www.ncbi.nlm.nih.gov/dbEST/ dbEST_summary.html). The submission of the 17 EST sequences generated within this study will increase this number to 143, representing a 13.5% increase in the number of publicly available dog ESTs.

Data from previous human EST projects using a liver cDNA library can be compared with genes for which corresponding dog ESTs were isolated in the present study. Amongst the most comprehensive liver cDNA library analyses performed and reported in detail was that of Okubo *et al.* (1992), who generated almost 1000 partial cDNA sequences from a human liver library. ESTs corresponding to seven genes were amongst these that have a counterpart in the present study, although in some cases the actual subunit with which sequence similarity was found differs, or is not specified. Okubo *et al.* (1992) reported the isolation of ESTs for elongation factor alpha, translation initiation factor eIF-2, aldehyde dehydrogenase, hnRNP coreprotein A1, serum albumin, fibrinogen alpha chain and alpha-1 antitrypsin, as well as various unspecified cytochromes. ESTs for related dog genes were identified in the present study by the isolation of clones containing sequence information from eukaryotic translation elongation factor 1 alpha 1 (LEI081), eukaryotic initiation factor 4AII (LEI096), aldehyde dehydrogenase II (LEI020), heterogeneous nuclear ribonucleoprotein A2/B1 (LEI042), albumin (LEI077), fibrinogen beta polypeptide (LEI053), alpha-1-antitrypsin (LEI072) and cytochrome p450 subfamily IIE (LEI122 and 184). A specific evaluation of the function of each gene corresponding to a dog EST lies outside the scope of the present study. However, the demonstration of similarity to predominantly enzymatic genes and those involved in various metabolic processes is consistent with the liver as a source of transcribed sequence, as well as with data obtained by Okubo *et al.* (1992).

The 17 dog ESTs described above have direct relevance to the reciprocal chromosome painting (Zoo-FISH) analysis between the human and dog genomes performed by Breen *et al.* (1999a) and Yang *et al.* (1999). As a result of these studies, the location of evolutionarily conserved chromosome segments (ECCS) between the human and dog genomes is now known. The human chromosome location of the ESTs developed in this

chapter can now be compared to the existing comparative map, enabling the prediction of where the corresponding dog gene is likely to exist. The data from tables 3.1b, c and d can therefore be extended to include these predictions, shown in tables 3.5a, b and c. As a minimum, each EST will be of value in confirming and refining the precise boundary of each ECCS. However, figure 3.5a demonstrates that some of the ESTs generated in this chapter will prove of additional value in determining the relative orientation of ECCS between the human and dog genomes. Three pairs of dog cDNA clones contain sequence data similar to genes that lie at opposing ends of ECCS on HSA 1 (LEI143 and 079), HSA 4 (LEI077 and 053), and HSA 6 (LEI036 and 081). The comparative map predicts that these clones will lie in corresponding ECCS on CFA 7, CFA 3 and CFA 12 respectively. The assignment of these clones would enable the gross relative orientation of the ECCS to be compared between the human and dog genomes, relative to the centromere of the chromosomes concerned. Similarly, LEI136, 042, 122 (and 184), and 072 share sequence similarity to genes that map to one extreme of an ECCS on HSA 6, 7, 10 and 14. These cDNA clones are likely to enable preliminary orientation of the corresponding ECCS on CFA 1, 14, 28 and 8 respectively.

Three further clones, LEI083, 092 and 094, demonstrated similarity to human genes that have not yet been fully characterised with respect to either their function or chromosomal location. Their relevance to gene mapping in the dog is somewhat limited until such characterisation is performed. The largest proportion of the clones studied, totalling 30 (39.5%), did not identify any significant similarity to existing database sequences. Their origin cannot be determined in the absence of further characterisation, but it is likely that some will represent novel genes for which no data are yet available from other species for comparison. Continued database searches would be required in order to identify future similar entries. Searches should be repeated at regular intervals to take account of the rapid rate of sequence generation and database submission. At a future time this might allow a definitive identity to be assigned to clones such as LEI083, 092 and 094, as well as those for which no similar sequences were identified by BLAST (Class H clones). The eight Class E clones (10.5%) appear to be of microbial origin, and represent contaminants of the library. The existence of contaminating sequences in cDNA libraries has been relatively well documented in previous studies, since the advent of large scale random cDNA sequencing has highlighted this issue. Prior to this time, when such libraries were typically used for screening and isolation of specific clones, the presence of contaminants would be likely to remain unnoticed, unless the probe used shared some

Table 3.5a:

Chromosomal location of Class A clones predicted from Zoo-FISH data. Data are identical to those given in table 3.2b, except that comparative chromosome painting data of Breen *et al.* (1999a) have been used to predict the dog chromosomal location of each gene for which significant identity was found by database similarity searching.

* The additional putative locations for the dog ALB gene (in brackets) were taken from Yang *et al.* (1999).

Clone	Symbol	Gene	HSA location	Predicted CFA location
020	ALDH2	aldehyde dehydrogenase 2, mitochondrial	12q24.2	26q12-q13
036	TAP1/2	transporter 1/2, ABC (ATP binding cassette)	6p21.3	12q
042	HNRPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1	7p15	14q
053	FGB	fibrinogen, B beta polypeptide	4q28	3q32-q35.2
072	PI	protease inhibitor 1 (anti-elastase), alpha-1-antitrypsin	14q32.1	8q
077	ALB	albumin	4q11-q13	3q32-q35.2 (or 13 or 15)*
081	EEF1A1	eukaryotic translation elongation factor 1 alpha 1	6q14	12q
096	EIF4A2	eukaryotic initiation factor 4AII	18p11.2	7q17dist-q24
122	CYP2E	cytochrome P450 subfamily IIE (ethanol-inducible)	10q24.3-qter	28q
134	ASL	argininosuccinate lyase	7pter-q22	14q
143	PDZK1	PDZ domain containing-protein 1	1q21	7q12-q17prox
151	SKB1	skb1 (*S. pombe*) homologue	nd	n/a
184	CYP2E	cytochrome P450 subfamily IIE (ethanol-inducible)	10q24.3-qter	28q

Table 3.5b:

Chromosomal location of Class B clones predicted from Zoo-FISH data. Data are identical to those given in table 3.2c, except that comparative chromosome painting data of Breen *et al.* (1999a) have been used to predict the dog chromosomal location of each gene for which significant identity was found by database similarity searching.

Clone	Symbol	Gene	HSA location	Predicted CFA location
136F	ARG1	arginase, liver 6q23	6q23	1q14-q25
149F	PAK3	p21 (CDKN1A)-activated kinase 3	Xq21.3-q24	X

Table 3.5c:

Chromosomal location of Class C clones predicted from Zoo-FISH data. Data are identical to those given in table 3.2d, except that comparative chromosome painting data of Breen *et al.* (1999a) have been used to predict the dog chromosomal location of each gene for which significant identity was found by database similarity searching.

Clone	Symbol	Gene	HSA location	Predicted CFA location
079F	HP	haptoglobin alpha	16q22.1	5q33-q36
079R	FMO3	flavin-containing monooxygenase	1q23-q25	nd
130F	DDX5	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 5 (RNA helicase, 68kD)	17q21	9q11.2-q25.1
130R	n/a	Human ribosomal DNA complete repeating unit	acrocentric p-arms	n/a
148F	PIK3R1	phosphatidylinositol 3-kinase, regulatory subunit, polypeptide 1 (p85 alpha)	5q12-q13	2q21dist-q24
148R	n/a	18S rRNA	acrocentric p-arms	n/a

Figure 3.5a:

Comparison of the dog-human comparative map with dog EST data. The human chromosome location of genes for which a dog EST was isolated is shown against the reciprocal chromosome painting data generated by Breen *et al.* (1999a). This demonstrates that a subset of the dog ESTs are likely to have particular value in refining the status of this comparative map. Class A markers are shown in bold type, class B markers are shown in normal type, and class C markers are italicised. The PAK3 marker has not been included in this figure since it maps to the X chromosome of both species. Human chromosome ideograms are taken from Francke (1994).

similarity to contaminating sequences. However, in an EST approach, where clones are selected at random for analysis, the chance of isolating contaminants is increased.

Contaminants can be classed into two groups, reflecting anomalies relating either to their species of origin, or to their structural nature. Group one contaminants represent sequences present within a library that have not been derived from the species under study. These typically represent sequences derived from microbial organisms, which become introduced into the library during at some stage of its construction. Adams *et al.* (1995) cited bacterial infection of the tissue used for RNA isolation as a potential source of a proportion of contaminating sequences. The presence of *Escherichia coli* sequence within a human myeloid cDNA library was described by Rosenberg (1993) after screening the library with a microbial gene sequence in order to isolate potential human orthologues. The sequences isolated from the library were shown to be identical to the *E. coli* gene, whilst being absent from total RNA from which the library was initially prepared. It was likely that the contaminants were introduced during the cloning stage of library construction. Of more concern was the identification of a large group of contaminating sequences present within the EMBL database. Savakis and Doelz (1993) studied almost 5000 cDNA sequences deposited within the database from a variety of EST projects and performed a similarity search against entries in the SWISS-PROT protein database. Of 2366 sequences deposited in one EST study on a human T lymphoblastoid library, 278 (11.7%) were found to be of prokaryotic origin, and were subsequently eliminated from the database. The identification of a contaminating sequence by virtue of a complete match to nucleotide sequence data from other species is only possible when that sequence has previously been documented and submitted to the relevant database. In this instance, further analysis indicated that up to 85% of cDNA sequences from one library that were deposited in GenBank by Généthon in October 1992 were of microbial origin (Anon, 1993). This resulted from the use of a computer algorithm developed for such a purpose, designed as an aid to the identification of non-human cDNAs within alleged human cDNA libraries. The origin of a given clone was assessed by performing an analysis of the frequency of occurrence of hexanucleotide combinations within the insert sequence (White *et al.*, 1993). This value was compared to that previously calculated for the known average occurrence of these hexamer combinations within a set of human control sequences. Sequences that deviated from the expected value could therefore be classed as potentially of non-human origin. Deviations are identifiable only between relatively unrelated species, and the overall difference in

hexamer composition will become less distinct the more related two species are. Thus a microbial contaminant in a human library can be detected, but the presence of contamination from another mammalian species cannot. Généthon responded by identifying the problematic library as one constructed by a particular commercial manufacturer, which had since been withdrawn from distribution due to previous reports of microbial contamination, predominantly of yeast origin. A second human cDNA library made by the same manufacturer was identified as a source of contaminants by Dean and Allikmets (1995), containing sequence from *S. cerevisae* and several bacterial species. An accompanying report described other microbial sequences found amongst clones from two further libraries, again originating from the same manufacturer. Although it is difficult to trace the origin of such sequences, the possibility of a microbial infection of the starting tissue source is an obvious starting point, as is the contamination of library construction reagents or apparatus. This is of particular concern in laboratories where libraries are routinely generated on a commercial level, since the possibility of cross-contamination is increased in an environment where resources for different species, or when different tissues from the same species, are used.

The second class of contaminating sequences represent those derived from the correct species, but which do not reflect the nature of a sequence required for EST generation. These predominantly represent partially spliced transcripts, and have been discussed by Bürglin and Barnes (1992) following the identification of cDNAs containing introns. Further investigations revealed the presence of several ESTs for which significant similarity existed to previously annotated cDNA sequence, but where the sequence alignment was interrupted by regions with little or no similarity. This suggested the presence of an intron in the EST that was not present within the previously annotated coding sequence with which comparison was made. In response, Adams *et al.* (1992b) discussed a potential explanation for such findings. Intron sequence could be present as a result of incomplete splicing of the transcript, which is highly likely if the library is constructed from total cell RNA rather than solely cytoplasmic transcripts, since only the latter will have undergone complete splicing and processing. Alternatively, intron sequence may reflect the presence of genomic DNA within the cDNA library. The possibility of genomic contamination should be limited if DNAse treatment of the RNA is performed prior to reverse transcription to generate cDNA. These factors may explain the origin of clones LEI008, 069, 098, 113 and 144 from Class G, which are similar to anonymous human genomic DNA clones. However, these human clones could

potentially contain uncharacterised coding sequences, and so the origin of the dog clones cannot be determined in the absence of more detailed analysis. The three mitochondrion-derived sequences in Class F are also likely to represent contamination of the library by apparently untranslated DNA.

Sequence data from Class D clones appear to be largely composed of repetitive elements. Interspersed repeat elements, including SINEs and LINEs, have been discussed in chapter one, with respect to their presence within primary gene transcripts. A family of highly similar canine SINEs has been described by Minnick *et al.* (1992). Each member consisted of a 130bp sequence flanked by a microsatellite repeat and a poly A tract of up to 65 nucleotides in length at the 3' end. Initially this sequence motif was found within untranslated regions of the dog interferon-ω and clotting factor IX genes. Comparable SINE elements have been isolated by Das *et al.* (1998), and further characterisation demonstrated the presence of this repeat motif every 5kb to 8kb within the dog genome. The SINE family was shown to share significant similarity with that previously described by Minnick *et al.* (1992). A database search performed in the present study, using the dog SINE element as the query, identified in excess of eight dog genes containing all or part of the consensus SINE repeat sequence. These are indicated in table 3.2a. The existence of a poly(A) tract in members of this repeat family is likely to explain their presence within a cDNA library generated from poly(A)+ mRNA, since there is likely to be association between the poly(A) of the SINE and the oligo(dT) probe. In turn, the dinucleotide repeat will in part have led to repeated isolation of sequences containing the SINE in previous studies, as a result of the screening of dog libraries for the isolation of potentially polymorphic genetic linkage markers. In a review of repetitive elements within the dog genome, Bentolila *et al.* (1999) identified a family of LINE sequences, many of which share highly significant identity with the human consensus L1 sequence. Again, the presence of a poly(A) tract associated with the repeat element is likely to explain its presence within an oligo(dT)-selected cDNA library.

Genes encoding ribosomal RNAs can also be considered as repetitive elements since they are highly abundant within the human genome. The loci encoding the human 28S, 18S and 5.8S ribosomal RNAs are found in clusters on the short arm of each of the five acrocentric chromosomes (HSA 13, 14, 15, 21 and 22), within the so-called nucleolar organiser regions (NORs). The transcript unit within the NOR is expressed as a 45S rRNA precursor that is then differentially cleaved to form the 28S, 18S and 5.8S ribosomal RNAs, via a series of intermediates. Ribosomal RNA constitutes an enormous

proportion of the total mass of RNA transcribed by an active cell. It is therefore not surprising that ribosomal transcripts are commonly isolated during the generation of ESTs, findings that have led to the development of methods to limit this redundancy. Clone LEI130 demonstrates a potential situation in which an incorrect gene identity may be assigned on the basis of a database similarity search which, on further inspection, is found to be solely related to the presence of a region of ribosomal sequence in both sequences.

Data from the three clones in Class C strongly suggest that they are chimaeric clones, representing different genes at each end of the insert. Two of these are highly similar to ribosomal sequences at one end of the insert. It would seem unlikely that this reflects anything other than the fact that ribosomal sequences may be present at relatively high frequency within a library, since there is no obvious reason why these specific gene sequences should become associated with ribosomal RNA genes. Similarly, there is no obvious explanation why the HP and FMO3 genes should become associated as was found in the case of LEI079. The existence of chimaeric clones is not uncommon in both genomic DNA and cDNA libraries, and under some circumstances may occur at a highly undesirable frequency. For example, Adams *et al.* (1992b) reported a high prevalence (30% to 40%) of chimaeric clones within one cDNA library used, resulting from the association of two unrelated transcripts prior to cloning. This library was subsequently eliminated from further studies. Wolfsberg and Landsman (1997) suggested that some apparently chimaeric cDNA clones may represent as yet undocumented biological events, although many are likely to have arisen as cloning artefacts during library preparation. Their third hypothesis was that some clones demonstrating similarity to different genes from opposing ends of the insert might simply be due to practical error, such as mislabelling of samples, or errors in lane tracking of sequencing gels. This possibility can be eliminated in the present study since comparable database search results were obtained when sequencing analysis was repeated for both extremes of the three apparently chimaeric clones.

The problem of identifying contaminating sequences is clearly a significant one, and was particularly so in the initial period of interest in EST generation. The fact that at this time ESTs could be deposited within GenBank directly, with no peer review or quality control assessment, meant that the potential for error was high. The responsibility of error checking prior to sequence submission therefore lay heavily with the individual researcher, though clearly for groups generating tens of thousands of ESTs this is not

easily achieved. The presence of some degree of contamination is perhaps inevitable, and more so is the potential for sequence inaccuracy at the nucleotide level, since EST data are by their nature typically generated in single pass sequencing procedures. The consequences of errors remaining unnoticed are potentially widespread, and therefore more comprehensive characterisation of ESTs is essential prior to embarking on long term analysis of a given clone. The stringent criteria used to describe a true database match in this project were intended to eliminate much of the potential for error. In the present study, the degree of similarity required for a significant match is greater than that used by many others. The length over which identity must be observed ensures that matches between short sequence motifs, that may not be a true reflection of the gross similarity between two sequences, can be identified. This may apply to cDNA clones such as those in Class D that contain SINE or LINE elements, motifs common to many gene sequences, for which database similarity results must be interpreted with caution. Finally, the requirement for matches to be generated from data taken from both ends of the cDNA clone enables the identification of chimaeric sequences, such as the Group C clones discussed in section 3.1.3. However, this is possible partly since this is a small-scale project for a species that has been little studied to date.

Preliminary attempts to physically map coding loci in the dog by FISH analysis using EST data has been applied to two clones, LEI042 and LEI077. The chromosomal location of the corresponding dog genes was predicted from Zoo-FISH data generated by Breen *et al.* (1999a) and Yang *et al.* (1999), prior to their assignment. The subsequent mapping of dog HNRPA2B1 to CFA 14q21.1, and of ALB to CFA13q22.2, demonstrates the powerful resource that data from these Zoo-FISH studies represent. Knowledge of dog-human synteny enabled the chromosomal location of HNRPA2B1 and ALB to be predicted to a single chromosome in the former case, and to three potential candidates in the latter. Thus the assignment of these loci in the dog has further confirmed and refined existing data for the location and boundaries of ECCS in the human and dog genomes. Physical mapping of dog HNRPA2B1 and ALB was possible by comparison of the dog cDNA sequence to the human complete gene sequence, and by the design of PCR primers to the former sequence to generate a genomic probe of sufficient size that FISH analysis was successful. Data regarding the size of human introns was used to predict the size of the corresponding dog PCR product. Although the average size of a vertebrate intron has been estimated as 622bp, it has been shown that intron sizes vary widely between species (Hawkins, 1988). Most lie within the range of 80bp to 100bp, but some

significantly longer introns have been reported, thus skewing the mean to this rather larger value. In contrast, the average exon was reported as only 137bp, and rarely greater than 800bp, although again a wide level of variation exists between different loci and different species (Hawkins, 1988). It is therefore extremely difficult to predict the size of a product generated by amplification across an intron from primers located within flanking exons. However, in two of the examples discussed in section 3.3 (LEI042 and LEI077), comparison with existing data for the corresponding human gene successfully enabled this prediction for the generation of an amplification product from dog genomic DNA. The absence of such a product in the third example (LEI122) may indicate that conservation of intron size between these two species does not always hold true. The recent report of the sequence of a single human chromosome (Dunham *et al.*, 1999) has provided a further resource for the estimation of average parameters for coding regions. This study demonstrated the extreme range that exists in these parameters. For example, the average gene was estimated as 19.2kb, but ranged from 1kb to 583kb. The total number of exons in a gene ranged from one to 54, with a mean of 5.4. This further indicates that prediction of the size of coding regions is likely to be extremely approximate. Until we have a far greater knowledge of the parameters of coding sequences within and between species, approaches such as that used in this chapter are amongst the only options for generation of genomic products by PCR amplification using data from partial cDNA sequences.

A further option is the direct physical mapping of cDNAs by FISH analysis, which has been discussed in chapter one. Chaudhary *et al.* (1997) mapped seven pig cDNA clones, ranging from 1.1kb to 1.3kb, to their subchromosomal location. The authors discussed the difficulties associated with FISH analysis of small probes, including the presence of background signal at non-specific chromosome sites. Assignments demonstrated concordance with the chromosomal location predicted from existing Zoo-FISH data. Attempts to optimise probe labelling procedures, hybridisation conditions and washing stringencies were reported to have had little effect on the ability to map five further pig cDNA clones. In a subsequent study, Thomsen *et al.* (1998) successfully mapped by FISH analysis 40% of 48 pig cDNA clones, with inserts of between 0.55kb and 2.7kb. Short inserts were shown to be a major limitation in the success of this approach, since only four of 18 clones with inserts of less than 1kb demonstrated detectable hybridisation signals. Since the average insert size of the clones analysed in this chapter was 1.4kb, it is likely that FISH analysis would not be widely successful for this purpose, particularly on

a routine basis. Fibre-FISH and similar techniques have been successfully used to map cDNA sequences to genomic clones hybridised onto DNA fibres (for example, Florijn *et al.*, 1996; Horelli-Kuitunen *et al.*, 1999). In the latter study, three ESTs of less than 1kb in size, and three cDNAs of less than 3kb were mapped, but required refinement of the technique to generate sufficient sensitivity to allow such small probes to be successfully hybridised and detected. Hybridisation to a metaphase chromosome template proved successful in only 15% of spreads, typically only on one chromatid. The use of extended fibres increased this frequency to between 25% and 45% for ESTs, and to 60% for cDNA clones, and was greater for larger probes. Korenburg *et al.* (1995) have shown that metaphase FISH has a success rate of 10% to 60% for cDNA probes of 1kb to 2kb, and that this rate is lower for smaller cDNAs. Assignments necessitated the analysis of 200 metaphase spreads for each probe. Clearly, direct assignment of cDNAs by FISH is a viable approach, but is reliant on specialised techniques and resources, and labour-intensive image analysis. It was not considered appropriate to pursue the method in this study, since the aim was specifically to investigate the value of EST generation for the dog. The recent chromosome painting data of Breen *et al.* (1999a) and Yang *et al.* (1999) are likely to confer an advantage in attempting FISH analysis of cDNA clones. The data arising from these studies enables the location of a dog gene to be predicted by reference to the human gene location. In turn, this would enable true hybridisation signal to be identified amongst background hybridisation. With knowledge of the expected gene location in the dog, concentrated optimisation of the method to improve the specificity of true hybridisation would then be possible. Similarly, the acquisition of a dog BAC library and WG-RH panel, and the ability to isolate and map genomic clones corresponding to dog ESTs has been demonstrated in the following chapter, which describes a more targeted approach to the mapping of coding sequences within the dog genome.

3.6 Summary

Partial sequence has been obtained for 76 clones selected at random from a dog liver cDNA library. Database similarity searches have identified 17 sequences with which a significant degree of nucleotide identity is shared with existing database entries. These correspond to gene loci previously mapped to 11 different human autosomes and the X chromosome, and will represent a useful contribution towards the mapping of coding sequences within the dog genome. Each has value in the confirmation and refinement of the boundary of ECCS between the human and dog genomes. Furthermore, information

regarding the relative orientation of ECCS between human and dog chromosomes can be provided by a subset of these ESTs. The ability to assign genes to dog chromosomes using EST sequence data has been investigated by the isolation and physical mapping by FISH analysis of genomic products corresponding to two gene loci.

Chapter four:

Identification of conserved synteny between dog chromosome five and the human genome

Identification of conserved synteny between chromosome five of the domestic dog and the human genome

Abstract

Reciprocal chromosome painting analysis between different species has been proven in many previous studies to be a useful tool in determining evolutionarily conserved segments between their genomes. Prior to the start of this project such an approach had not been applied to the study of the domestic dog. CFA 5 was selected as a candidate for the development and application of reciprocal chromosome painting analysis between dog and human. A chromosome paint probe specific for CFA 5 was applied to human metaphase spreads in order to establish regions of conserved synteny between this dog chromosome and the corresponding human chromosomes. Subsequent FISH analysis demonstrated that CFA 5 is composed of four consistent ECCS shared with regions of HSA 1, 11, 16 and 17.

In the reciprocal analysis, paint probes representing these four human chromosomes were applied separately to dog metaphases. This enabled the identification of the precise regions of CFA 5 that share conserved synteny with HSA 1, 11, 16 and 17. The CFA 5 paint probe was also applied to metaphase chromosomes of the domestic cat, identifying four regions of conserved synteny between their genomes. These data were compared with those obtained from the dog-human chromosome painting analyses, and with cat-human data from previously published studies.

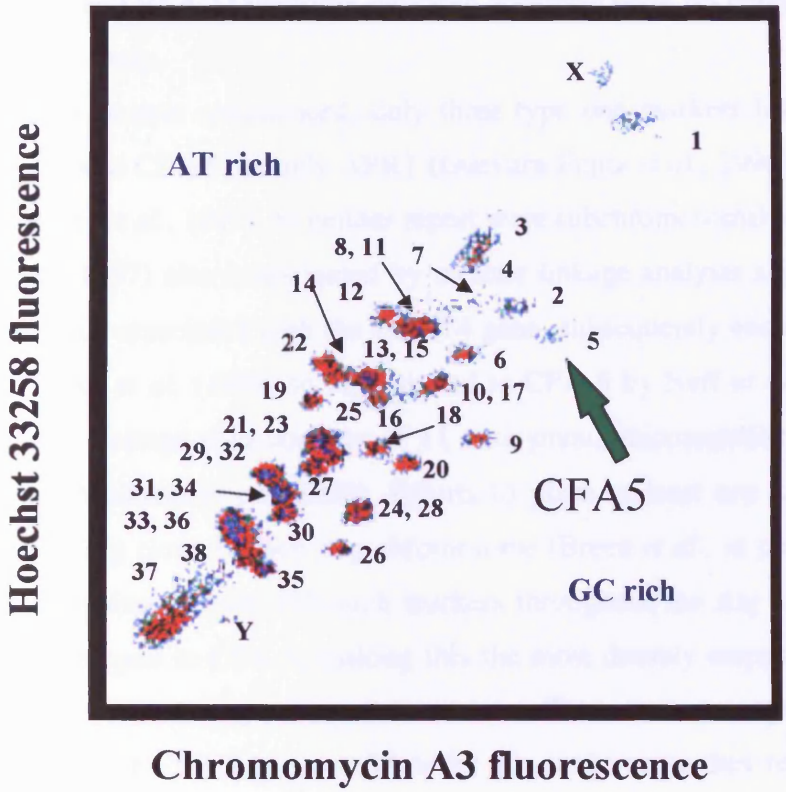
The boundaries of the conserved segments on CFA 5 were further investigated and refined by the isolation and physical mapping of dog BAC clones containing type one markers. Markers were selected for this analysis on the basis of their location within those regions of the human genome predicted to be evolutionarily related to CFA 5. The physical location of each type one marker within the dog genome was then compared with data resulting from reciprocal chromosome painting analysis, and the correlation between these data was discussed.

4.1 Selection of a chromosome for focussed study

In order to establish techniques for chromosome painting analysis and comparative gene mapping between human and dog, a single chromosome was selected to enable focussed study. The selection of a suitable candidate required consideration of a number of factors, predominantly relating to the availability and quality of mapping resources. Since dog-human reciprocal chromosome painting analysis had not previously been investigated, the selection of a dog chromosome for which cytogenetic resources were of a high quality was paramount. This would influence the quality of the hybridisation signals obtained from cross-species painting analysis.

Although the dog paint probes of Langford *et al.* (1996) have been shown to be highly enriched for the corresponding dog chromosome, the nature of the flow-karyotype indicates that some probes are likely to be more specific than others (figure 4.1a). The highest quality probes are likely to be those for which the flow-peak is well isolated from neighbouring peaks. CFA X and 1 are amongst the most isolated peaks, but since studies are in progress on the development of an integrated meiotic and radiation hybrid linkage map of anonymous markers for these chromosomes (H. Dickens, pers. comm.), CFA X and CFA 1 were not considered for this study. In addition, CFA X would not represent a suitable choice for comparative mapping analysis since the extreme level of conservation of X-linked sequences in a wide range of species is well established.

Other isolated flow peaks represent those for CFA 5, 9, 24 and 28, and 26. The mixed probe for CFA 24 and 28 was excluded since data from cross-species chromosome painting analysis would be complicated by the presence of material from two autosomes. CFA 9 has previously been studied by Werner *et al.* (1997) from a comparative angle, and therefore was not considered an appropriate candidate. Although CFA 26 remained a potential candidate, CFA 5 was selected as the most appropriate in terms of cytogenetic resources. Although Werner *et al.* (1997) included the mapping of two gene markers to CFA 5 in their preliminary analysis of HSA 17 counterparts in the dog, their main focus was on CFA 9. Thus the study of CFA 5 was considered unlikely to result in any duplication of effort. Werner *et al.* (1997) established that a paint probe for HSA 17 hybridises to the midregion of CFA 5, but no reciprocal analysis was performed. The authors mapped the genes GLUT4 and PMP22 within CFA 5qmid., the latter distal to the former, inferring inverted orientation compared to the human location of these genes on HSA 17p. The existing identification of a putative region of conserved synteny between



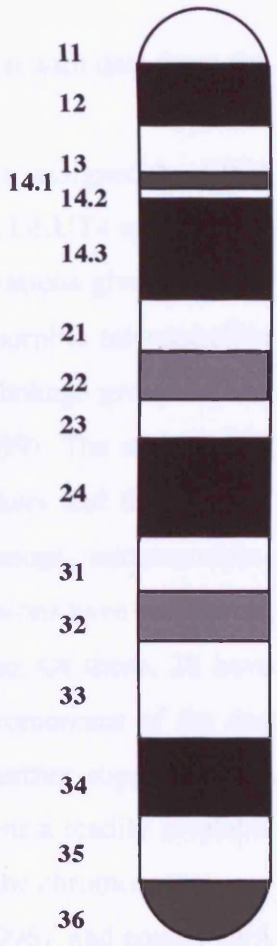


Figure 4.1a: (left)

The bivariate flow karyotype of the dog (adapted from Langford *et al.*, 1996). Chromomycin A3 fluorescence is plotted on the horizontal axis, and Hoechst 33258 fluorescence on the vertical axis. It can be clearly seen that the flow peak corresponding to CFA 5 is one of the most isolated, and lies towards the most GC-rich region of the flow karyotype. Assignments of flow peaks are shown according to the nomenclature of Breen *et al.* (1999b).

Figure 4.1b: (right)

DAPI-banded ideogram of CFA 5 (adapted from Breen *et al.*, 1999b).

CFA 5 and HSA 17 therefore provided a starting point for comparison with data from the present study.

As this project commenced, only three type one markers had been assigned by FISH analysis to CFA 5, namely APRT (Guevara-Fujita *et al.*, 1996), and GLUT4 and PMP22 (Werner *et al.*, 1997). In neither report were subchromosomal localisations given. Werner *et al.* (1997) also investigated by meiotic linkage analysis a polymorphic microsatellite sequence associated with the GLUT4 gene, subsequently enabling linkage group L19 of Mellersh *et al.* (1997) to be assigned to CFA 5 by Neff *et al.* (1999). The most recent CFA 5 linkage map consists of 11 anonymous microsatellite markers and the GLUT4 gene (Mellersh *et al.*, 2000). Efforts to place at least one anonymous, microsatellite-containing clone to each dog chromosome (Breen *et al.*, in preparation) have resulted in the mapping of over 350 such markers throughout the dog genome. Of these, 28 have been mapped to CFA 5, making this the most densely mapped chromosome of the dog karyotype with respect to such markers. These resources gave further support to the selection of CFA 5 as a candidate for this study, since they represent a readily available source of markers that could be used to create a framework map of the chromosome.

The size of CFA 5 has been estimated as 99Mb (Langford *et al.*, 1996), and equals CFA 2 as the fourth largest dog autosome. Both GTG-banding and DAPI-banding patterns are highly distinctive for CFA 5 such that it can be recognised unequivocally within a dog metaphase spread, unlike many of the smaller autosomes. The DAPI-banded CFA 5 ideogram proposed by Breen *et al.* (1999b) is shown in figure 4.1b. Banding analysis identifies an extensive GC-rich region at CFA 5q31-q33, and another smaller such region at CFA 5q21-q23. These can be considered indicative of a high gene density, and this is additionally supported by the position of the CFA 5 flow-peak, which is skewed towards the area occupied by those chromosomes that have a high affinity for the GC-binding dye Chromomycin A. CFA 5 is therefore likely to represent one of the most gene-rich chromosomes of the dog karyotype, and thus is appropriate for a comparative gene mapping study.

4.2 Reciprocal chromosome painting analysis of CFA 5

4.2.1 Assessment of the quality of the CFA 5 paint probe

The CFA 5 chromosome paint probe was first applied to dog metaphase chromosomes in order to demonstrate the quality and specificity of the resource. One hundred and fifty nanograms of biotinylated paint probe (Langford *et al.*, 1996) were combined with ten

micrograms of sonicated dog genomic DNA, and hybridisation was carried out at 37°C for 16 hours. Stringency washes were performed as described in section 2.2, and the biotinylated probe was detected with an FITC-avidin conjugate. Signal amplification was performed by the application of biotinylated anti-avidin followed by a further layer of FITC-avidin. A representative example of the resulting images captured is shown in figure 4.2a, in which distinct and even signal covered the entire length of CFA 5. No consistent signal was observed on any other chromosome.

4.2.2 Hybridisation of the CFA 5 paint probe to human metaphases

One hundred and fifty nanograms of the biotinylated CFA 5 paint probe were hybridised to four day old human metaphase spreads. The protocol used was that described in section 2.2, except that the hybridisation period was extended to seven days, following the recommendations of Scherthan *et al.* (1994) for heterologous chromosome painting analysis. Two separate reactions were performed, one with 10μg of sonicated dog genomic DNA as competitor, the other with 1μg of human C_0t1 DNA (in the presence of 9μg sonicated salmon sperm DNA as carrier). Comparison of the images resulting from these two reactions would indicate which competitor was most appropriate for cross-species chromosome painting analysis.

The hybridisation of the probe was detected using two layers of FITC-conjugated avidin. Analysis of 20 metaphase spreads from both hybridisation reactions demonstrated that human C_0t1 DNA was the preferred competitor for this form of hybridisation, since the ratio of signal to background was consistently higher than when sonicated dog genomic DNA was used. Subsequent analyses therefore were performed with competitor from the species corresponding to the chromosomal target used rather than that of the probe.

Thirty human metaphase spreads were studied in order to determine the hybridisation sites of the CFA 5 paint probe. This enabled an assessment of the degree to which hybridisation was consistent between different metaphase spreads. Hybridisation signals were observed to be generally more intense on shorter chromosome spreads. However, longer chromosomes permitted the subchromosomal location of the signal to be established at a higher resolution than was possible by analysis of shorter metaphases. Thus the presence of a range of early to mid metaphase chromosome spreads of varying length on each slide conferred a significant advantage in the analysis of hybridisation signals, an example of which is given in figure 4.2b.

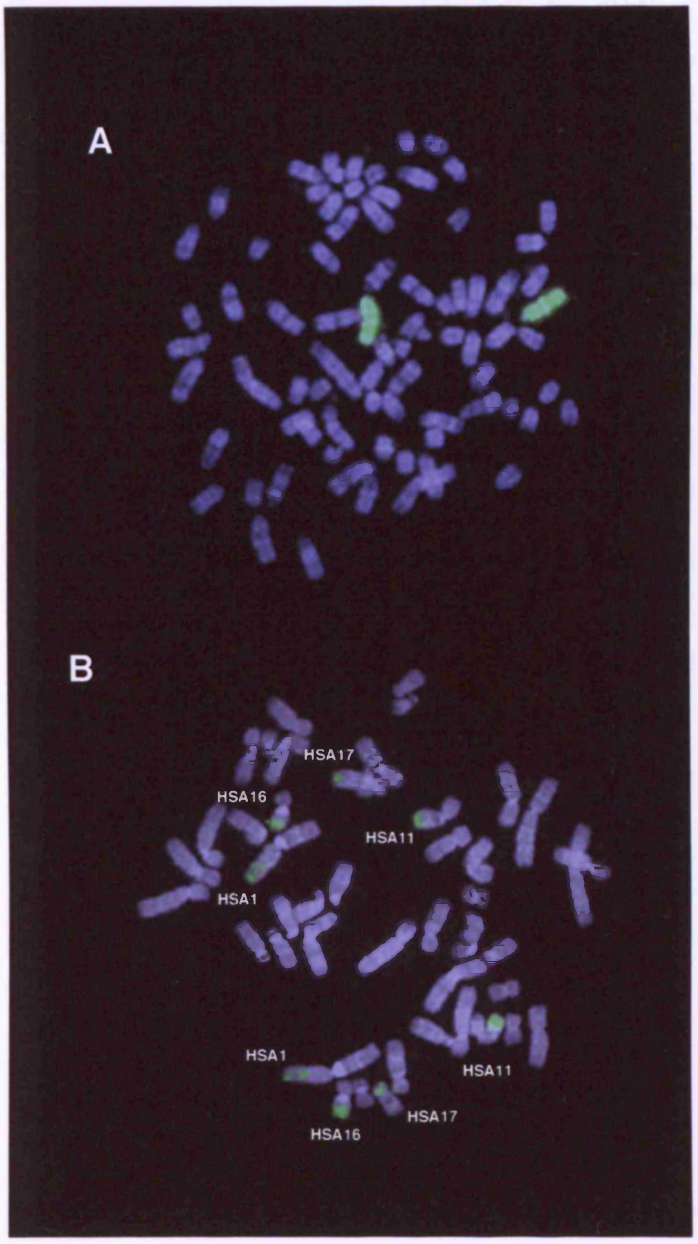


Figure 4.2a (above):

Hybridisation of the CFA 5 chromosome paint probe to a dog metaphase spread. The specificity of the paint probe for CFA 5 correlates with its isolated position in the bivariate flow karyotype of the dog.

Figure 4.2b (below):

Hybridisation of the CFA 5 paint probe to a human metaphase spread. Distinct regions of hybridisation signal can be seen on HSA 1p, 11q, 16q and 17p.

Consistent hybridisation of the CFA 5 paint probe was apparent on four human chromosomes. The most extensive region of probe hybridisation was observed on the distal tip of the long arm of HSA 11. Signal did not appear to extend entirely to the telomere in all metaphases studied, however it was more frequently present than absent in this region. In all cases, the signal extended at least to the distal extreme of band HSA 11q24, and typically extended into HSA 11q25prox. The proximal boundary of the region to which the CFA 5 paint hybridised was considerably more consistent between spreads, extending to, and including, HSA 11q23. The region of hybridisation of the probe to HSA 16 appeared to be of a similar size to that described for HSA 11, and was also located at the distal region of the long arm. Here, hybridisation extending to the telomere appeared to be more consistent, and extended proximally to include HSA 16q21. The CFA 5 paint probe showed consistent hybridisation to the short arm of HSA 17. Again, signal consistently extended to include the telomere, and proximally to include HSA 17p13 in its entirety. Here, however, the signal was frequently observed to extend further to include HSA 17p12, but again was present in most but not all metaphases studied.

A small region of hybridisation signal was present also on the midregion of the short arm of HSA 1. The small size of this region made this the most difficult to describe with reference to its subchromosomal location. Hybridisation was clearly present within bands HSA 1p32 and HSA 1p31, but did not appear to include the entire band 1p31. The region can thus be described as localised to HSA 1p32.3-p31.2. In a minority (approximately 10%) of metaphase spreads, hybridisation signal was observed also within the region of HSA 1p36-p35. In instances where this was apparent, the signal intensity was markedly lower than that for the previously described regions on HSA 1, 11, 16 and 17.

The application of the CFA 5 chromosome paint probe to human metaphases therefore led to the identification of four putative ECCS, on HSA 1p, 11q, 16q and 17p. The presence of a fifth such region, also on HSA 1p, was of less certainty.

4.2.3 Hybridisation of human chromosome paint probes to dog metaphases

One hundred and fifty nanograms of biotinylated paint probe corresponding to HSA 1, 16 and 17 were hybridised separately to dog metaphases aged for four days at room temperature. As a result of the observations described in section 4.2.2, 10μg of sonicated dog genomic DNA were used as the competitor in each case.

The similarity in size and G+C content of HSA 9-12 results in their close proximity within the human bivariate flow karyotype, such that it is frequently not possible for the corresponding flow peaks to be resolved from each other. As a consequence, in this study a single paint probe was available which contained material from all four of these human chromosomes. This was used to obtain data for the hybridisation of HSA 11 to dog metaphases. The concentration of this probe used was therefore increased fourfold to 600ng such that the quantity of material applied to dog metaphases which was derived from HSA 11 was comparable to that for the three other paint probes used. The quantity of competitor DNA was also increased fourfold. Hybridisation signals from the application of human paint probes to dog metaphases spreads were successfully detected after 48 hours of incubation at 37°C.

All four paint probes demonstrated hybridisation to CFA 5, as predicted from the data obtained in section 4.2.2. Examples of the images obtained are shown in figure 4.2c. Typically, between 25 and 40 human metaphase spreads were studied for each hybridisation reaction. Overall, hybridisation signals for the HSA 1, 16 and 17 probes were of a higher degree of specificity than those observed from the application of the dog paint probe to human chromosomes. The HSA 1 probe was shown to hybridise to the mid- to distal- region of CFA 5, and the analysis of metaphases chromosomes of varying length resulted in this region being localised to include and extend between CFA 5q24-q32. Similarly, the HSA 17 paint probe hybridised specifically to CFA 5q21-q23. The HSA 16 paint probe hybridised to the most distal region of the chromosome of interest, extending from CFA 5q33-q35. Again, signal within the most telomeric region at CFA 5q36 was commonly observed in most but not every metaphase spread analysed. Data obtained from the chromosome paint representing HSA 9-12 was considerably more difficult to analyse due to the increased complexity of the probe. However, the sole region of hybridisation to CFA 5 extended proximally from CFA 5q14. Again, signal was frequently observed within the most proximal band CFA 5q11, but this was not without exception. Thus the minimum region of hybridisation can be described as CFA 5q12-q14.

CFA 5 can thus be described in terms of four ECCS, corresponding to regions of HSA 11, 17, 1 and 16. This represents the order of the segments from the proximal to the distal tip of CFA 5. The segments corresponding to HSA 1, 11 and 16 are of relatively similar size, whilst that corresponding to HSA 17 is markedly smaller. The segments can be described to the level of a single band, and the localisation in each case does not indicate

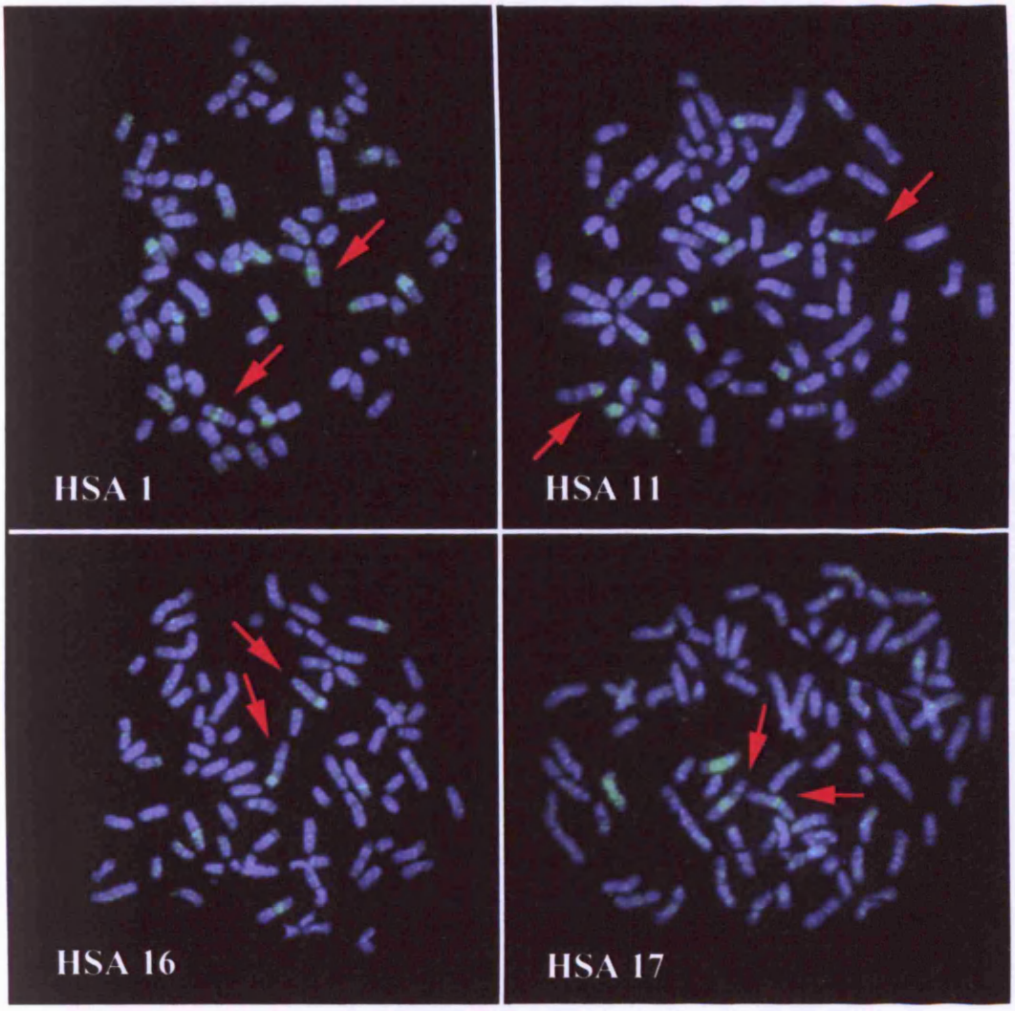


Figure 4.2c:

Hybridisation of paint probes corresponding to HSA 1, 11, 16 and 17 to a dog metaphase spread. This enables the chromosomal location of CFA 5 ECCS (indicated by red arrows), and their counterparts in the human genome, to be described to the subchromosomal level.

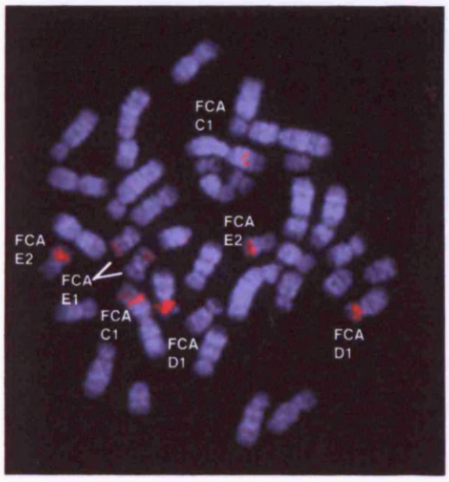


Figure 4.2d:

Hybridisation of the CFA 5 paint probe to a cat metaphase spread. Distinct regions of hybridisation signal can be seen on FCA C1p, D1p, E1p and E2q (using the nomenclature of Cho *et al.* (1997a).

that any overlap exists between them, neither were any gaps observed within CFA 5 to which no human paint probe hybridised. The reciprocal chromosome painting analysis appears to give complete coverage of CFA 5 and its counterparts in the human genome, with the exception of the extreme telomeric regions both of dog and human chromosomes.

4.2.4 Hybridisation of the CFA 5 paint probe to cat metaphases

One hundred and fifty nanograms of the biotinylated CFA 5 paint probe were combined with 10μg of sonicated cat genomic DNA, and hybridised to cat metaphase spreads aged for four days at room temperature prior to use. Hybridisation signals were successfully generated after 48 hours of incubation at 37°C. Chromosome assignment was based on the nomenclature described by Cho *et al.* (1997a). Twenty five metaphase spreads were analysed, resulting in the identification of four ECCS within the cat genome that correspond to CFA 5. Most striking was the hybridisation to FCA C1, since this was remarkably similar to that obtained previously for the hybridisation of the CFA 5 paint probe to HSA 1. An intense region of signal was observed at FCA C1p22, and in some cases was seen to extend as far as FCA C1p23-p21. This signal was extremely consistent between all spreads studied. However, additional, less intense signal was also seen at the most distal tip of the arm, at FCA C1p28-p27, although as with the situation regarding HSA 1pdist. this signal was noticeably less consistent. Distinct and consistent signal was also present at FCA D1p13-p12. Again, the region of hybridisation was frequently seen to extend further towards the telomere, to also include FCA D1p14.

The remaining two ECCS were located on two of the smaller cat chromosomes, namely FCA E1 and E2. The small region of signal on the former was localised to FCA E1p14-p13, most frequently extending fully to include the most telomeric band. In a small number of cases (<10%), the region was observed to extend further to include FCA E1p12. The fourth putative ECCS was localised to FCA E2q13-q14, including the most telomeric region of the chromosome in approximately 75% of metaphases studied.

The quality of hybridisation signals observed from the hybridisation of the CFA 5 paint to cat chromosomes was high, with very little background apparent. Figure 4.2d shows a typical example of the images obtained.

Data obtained in sections 4.2.2 to 4.2.4 were combined to produce figure 4.2e, a schematic representation of the regions of conserved synteny on CFA 5 and its

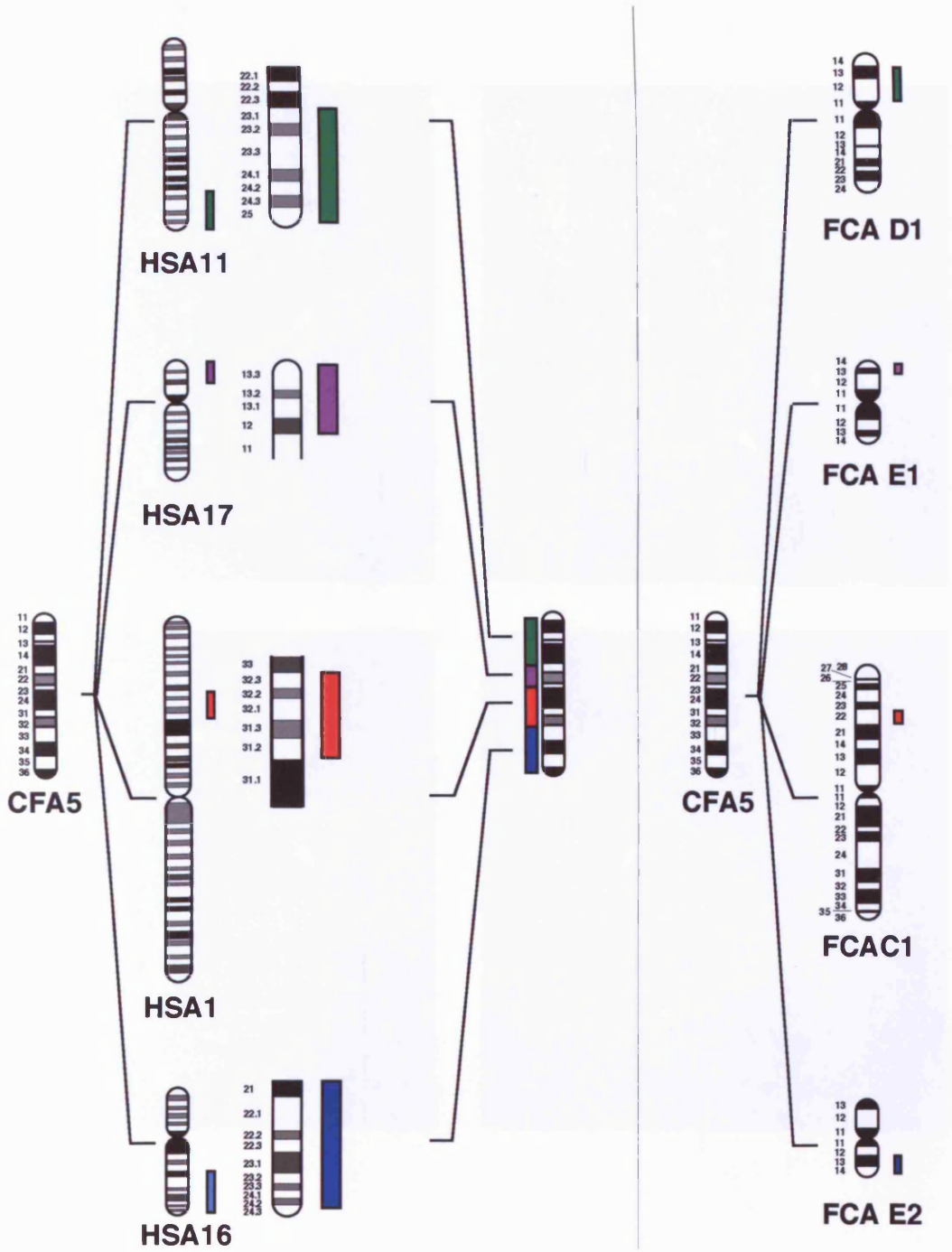


Figure 4.2e:

Schematic representation of ECCS on CFA 5 and their counterparts in the human and cat genomes. The CFA 5 paint probe was hybridised to human (left) and cat (right) metaphases, enabling the location of ECCS corresponding to CFA 5 to be identified in both the latter genomes. The reciprocal experiment using human chromosome paint probes enables corresponding ECCS in both the human and dog genomes to be described to the subchromosomal level (left). This information can be related to figure 1.4b, which compares the relative value of unidirectional and reciprocal Zoo-FISH analysis. Human, cat and dog ideograms were adapted from Francke (1994), Cho *et al.* (1997a), and Breen *et al.* (1999b) respectively.

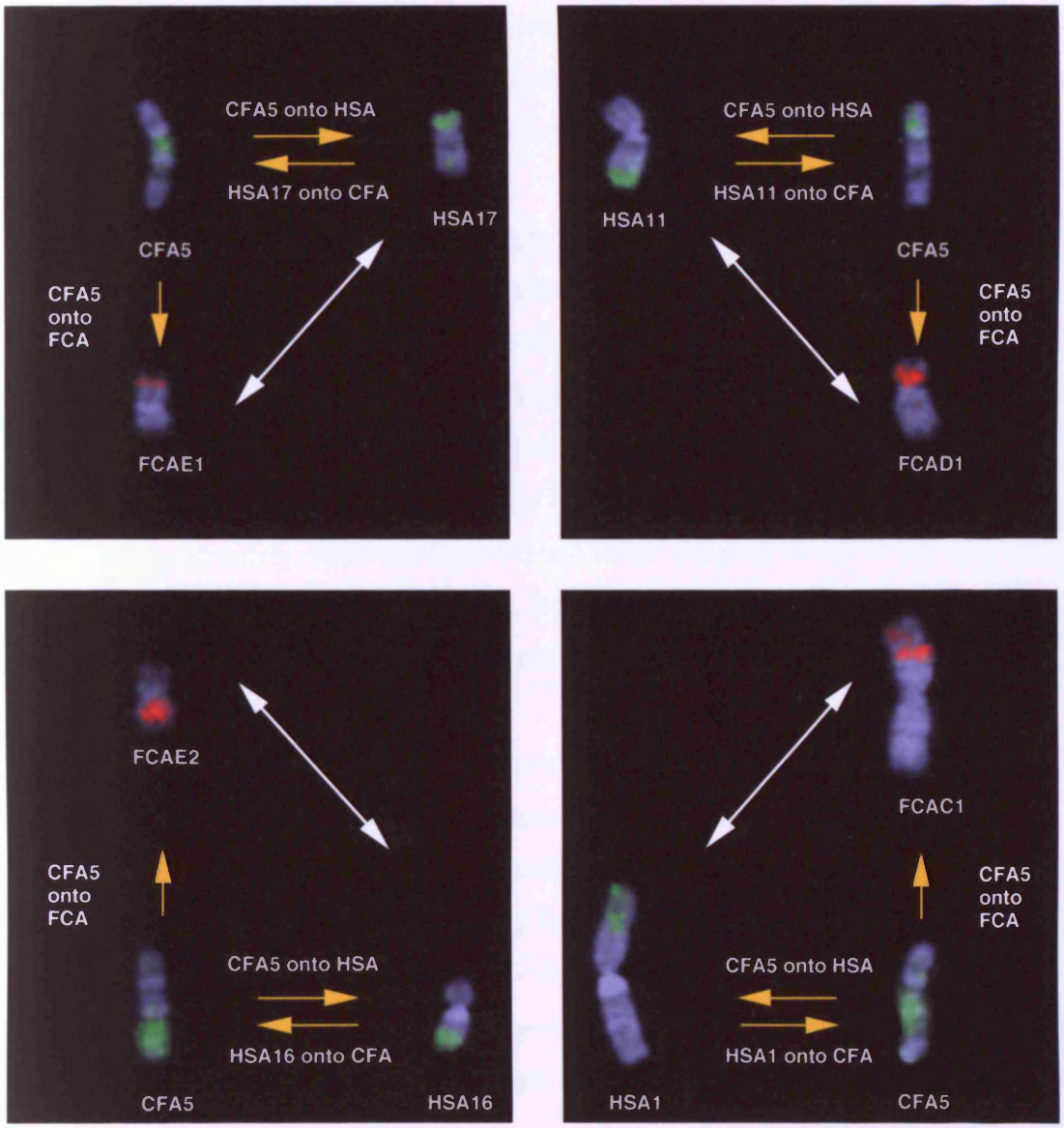


Figure 4.2f:

Comparison of corresponding human, dog and cat ECCS by reciprocal chromosome painting analysis. Unidirectional and reciprocal chromosome painting analysis enables corresponding ECCS on CFA 5 and on four chromosomes in both the human and cat genomes to be directly related, despite the absence of cat chromosome paint probes. Cat/human Zoo-FISH data (indicated by white arrows) were taken from Wienberg *et al.* (1997).

counterparts in the human and cat genomes. Figure 4.2f shows how this information can be extrapolated to enable a three-way comparison between the chromosomal location of dog, human and cat ECCS with respect to CFA 5.

4.3 Physical assignment of dog genes using comparative mapping resources

4.3.1 Selection of gene markers for confirmation of chromosome painting

Existing publications were studied in order to identify markers mapping to those human chromosome regions shown by this study to share conserved synteny with CFA 5. The nature of these sets of markers is described in detail in chapter one. Those which were suitable for analysis are listed in table 4.3a, which gives the locus name reported in the relevant publication. Both the standardised name assigned to the locus, and its human chromosome location, were obtained from the Human Gene Nomenclature Database (http://www.gene.ucl.ac.uk/nomenclature/). Markers are grouped by their physical location within the human genome.

Sixteen potentially informative markers were available amongst the CATS developed by Lyons *et al.* (1997), of the 318 such primer pairs described. These were spread relatively evenly throughout the four ECCS identified, and also within the putative fifth ECCS at the distal end of HSA 1p. Six of the 86 UM-STS markers (Venta *et al.*, 1996) lay within the ECCS of interest. Of these, three lay within the HSA 16q ECCS, and one marker corresponded to each of the HSA 11 and HSA 17 ECCS. The sixth marker was derived from a gene mapped to the putative fifth ECCS on HSA 1p. No UM-STS markers were available for the major ECCS on HSA 1p. Markers described by Priat *et al.* (1999) also generated a number of potentially useful loci amongst the 463 described. Eight dog gene markers were located throughout the five human chromosomal regions of interest. These were supplemented by a single TOAST marker and a human EST, both derived from HSA 1, and a canine TOAST marker from HSA 17. The nature of each subtype of marker is described in chapter one. Jiang *et al.* (1998) described six markers that were of potential use in this study, of the total of 225 reported. These markers were widely spread throughout the human chromosomal regions of interest. At least one such marker was available for each of the four ECCS and for the putative fifth such region, with two markers being available for the HSA 11q region.

Forty two potentially useful type one markers were identified within the publications described above. Within these were a number of duplications of the same locus, such that the 42 markers represented a total of 33 genes. The NPPA gene on HSA 1p is represented

Table 4.3a (opposite):

Type I markers available for refinement of ECCS. Columns one and two list the accepted locus name and human chromosome assignment for each of the markers investigated, taken from the Human Gene Nomenclature Database. The locus name and human chromosome assignment reported in the marker publication are given in columns three and four. The final column indicates the original source of the marker: (A) Lyons *et al.* (1997) (B) Venta *et al.* (1996) (C) Priat *et al.* (1999) (D) Jiang *et al.* (1998) (E) Aguirre *et al.* (1998) (F) chapter three of this study and (G) A. Curson, pers. comm. The markers described by Priat *et al.* (1999) are further subdivided to classify each as a 'TOAST' marker (C^{T}), dog gene (C^{D}), human EST (C^{E}) or 'Canine TOAST' (C^{CT}). These categories are further described in sections 1.4.4 and 1.5.4.

Locus name (HGND)	HSA location (HGND)	Marker name	HSA location cited	Source of marker
HSA1				
NPPA	1p36	ANP	1p36.3-p36.2	C^T
NPPA	1p36	NPPA	1p36.2	D
NPPA	1p36	NPPA	1p36	A
NPPA	1p36	PND	1p36	B
FGR	1p36.2-p36.1	FGR	1p36.2-p36.1	A
DDOST	1p36.1	DDOST	1p36.1-p35	C^D
CSF3R	1p35-34.3	CSF3R	1p35-34.3	A
DIO1	1p33-p32	DIO1	1p33-p32	C^D
JUN	1p32-p31	JUN	1p32-p31	A
VCAM1	1p32-p31	VCAM1	1p32-p31	C^D
VCAM1	1p32-p31	VCAM1	1p32-p31	C^E
NFIA	1p31.3-p31.2	NFIA	1p31.3-p31.2	D
ACADM	1p31	ACADM	1p31	A
RPE65	1p31	RPE65	1p31	A
RPE65	1p31	RPE65	n/a	E
HSA11				
MMP1	11q21-q22	MMP1	11q22-q23	D
DRD2	11q22-q23	DRD2	11q23.1	A
THY1	11q22.3-q23	THY1	11q22.3-q23	A
CD3E	11q23	CD3E	11q23	C^D
CD3D	11q23	CD3D	11q23.3	A
APOA1	11q23-q24	APOA1	11q23	D
APOC3	11q23-qter	APOC3	11q23-qter	B
ETS1	11q23.3	ETS1	11q23.3	A
HSA16				
GOT2	16q12-q22	GOT2	16q21-q22	B
HP	16q22.1	HP	n/a	F
TAT	16q22.1	TYRA	16q22.1-q22.3	C^D
ALDOA	16q22-q24	ALDOA	16q22-q24	C^D
CTRB1	16q23-q24.1	CTRB	16q22.3-q23.2	B
CTRB1	16q23-q24.1	CTRB1	16q23.1	A
APRT	16q24	APRT	16q24	B
APRT	16q24	APRT	16q24.2-qter	A
DPEP1	16q24	RDP	16q24.3	D
HSA17				
ASPA	17pter-p13	ASPA	17pter-p13	A
GP1BA	17pter-p12	gpIb	17p12	C^D
SLC2A4	17p13	SLC2A4	17p13.1	C^{CT}
GUCY2E	17p13.1	GUCY2E	n/a	G
MYH2	17p13.1	MYH2	17p13.1	A
POLR2A	17p13.1	POLR2A	17p13.1	A
TP53	17p13.1	TP53	17p13.1	A
TP53	17p13.1	TP53	17p13.1	C^D
TP53	17p13.1	TP53	17p13.1	B
UBB	17p12-p11.2	UBB	17p12-11.1	D

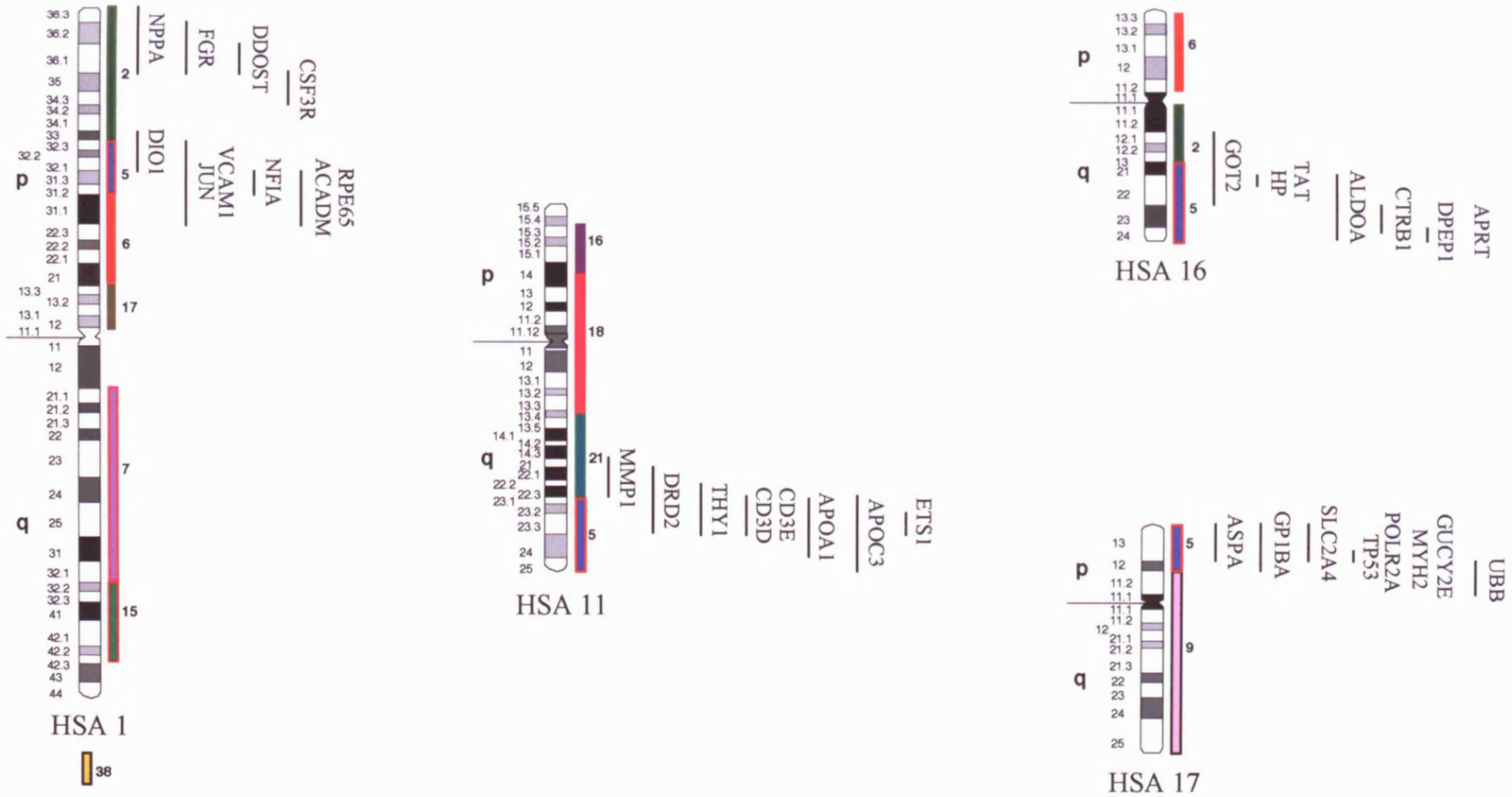


Figure 4.3a:

Location of type I markers on human chromosomes. Loci listed in table 4.3a were investigated as potential markers for the refinement of ECCS corresponding to CFA 5. The chromosomal locations of the orthologous human loci are shown above against the corresponding ideogram (Francke, 1994), and are taken from the Human Gene Nomenclature Database. The location of ECCS in both the dog and human genomes was derived from this study, and from Breen *et al.* (1999a).

by four markers, although both Priat *et al.* (1999) and Venta *et al.* (1996) use a different name for the locus (ANP and PND respectively). The TP53 gene on HSA 17p was represented by three markers. VCAM1 (HSA 1p), CTRB1 (HSA 16q) and APRT (HSA 16q) are each represented by two different markers. The location of all available markers is shown against the corresponding human ideogram in figure 4.3a, using the assignments taken from the Human Gene Nomenclature Database (http://www.gene.ucl.ac.uk/ nomenclature/). Where the assignment provided by the Human Gene Nomenclature Database differs from existing publications, both assignments are given in the text. Also shown are three markers derived from other sources (GUCY2E, HP and an alternative RPE65 marker), which are discussed later.

4.3.2 Initial assessment of the marker primers in the PCR

Preliminary PCR analysis was carried out in standard 10µl reaction volumes using the published annealing temperatures where available. For markers where this information was not provided, an annealing temperature of 55°C was used in the first instance. Where a single product was not obtained under these conditions, further optimisation of amplification conditions was performed as described in section 2.1.4. Markers for which conditions could be fully optimised are listed in table 4.3b. The size of the resulting PCR product is also shown and, where such information was available, that obtained/predicted by the authors is also given. Table 4.3c indicates the proportion of each marker type available from the publications of Venta *et al.* (1996), Lyons *et al.* (1997), Jiang *et al.* (1998) and Priat *et al.* (1999) that was successfully optimised.

4.3.3 Verification of the identity of PCR products

Amplification of part of the expected gene was verified by sequencing analysis of the PCR product and subsequent comparison of these data with information present within the GenBank database. Since most amplification products were less than 500bp in size, sequencing analysis was typically performed in a single direction only.

Table 4.3d lists the highest scoring database match for each of the markers analysed with the probability value (P-value) and number of identical bases out of the total size of the aligned region.

Table 4.3b:

Optimised PCR conditions for selected markers. Column one lists the markers using the locus name given in the Human Gene Nomenclature Database, and column two indicates the original source of those markers for which amplification conditions could be optimised (see legend for table 4.3a for details). These conditions are given in column three, with the annealing temperature (°C) followed by the reaction buffer used (mM $MgCl_2$). The final two columns list the PCR product size predicted by the authors, and that observed in this study, from application of marker primers to dog genomic DNA. Note that predicted product sizes for markers described by Lyons *et al.* (1997) (denoted 'A') were derived by aligning sequences from a variable number of different species, from which a consensus primer sequence and product size was inferred. This product size does not therefore relate specifically to the dog. The chromosomal location of the orthologous human gene locus was derived from the Human Gene Nomenclature Database, as shown in table 4.3a.

Marker name	HSA location	Source	Optimised conditions	Product size from CFA gDNA	
				Predicted (bp)	Observed (bp)
HSA 1					
PND	1p36	B	57/1.5	360	380
CSF3R	1p35-34.3	A	53/1.0	376/500	450
DIO1	1p33-p32	C^D	60/1.0	239	250
VCAM1	1p32-p31	C^D	55/1.5	304	300
NFIA	1p31.3-p31.2	D	60/1.25	319	300
RPE65	1p31	E	60/1.5	105/109	100
HSA 11					
MMP1	11q21-q22	D	60/1.5	390	400
THY1	11q22.3-q23	A	57/1.5	955/950	800
CD3E	11q23	C^D	55/1.5	178	170
HSA 16					
GOT2	16q12-q22	B	see text	1300	300 + 1600
HP	16q22.1	F	60/1.0	n/a	170
TAT	16q22.1	C^D	55/1.5	180	180
ALDOA	16q22-q24	C^D	55/1.5	177	180
APRT	16q24	B	57/1.5	1300	750
DPEP1	16q24	D	55/1.5	158	150
HSA 17					
GP1BA	17pter-p12	C^D	55/2.0	252	250
SLC2A4	17p13	C^{CT}	55/2.0	124	120
$SLC2A4^{RT}$	17p13	see text	60/1.5	n/a	150
GUCY2E	17p13.1	G	57/1.5	n/a	120
TP53	17p13.1	A	57/1.5	296/900	1200

Table 4.3c:

Success rate of different marker types. The categories of marker used in this study are listed in columns one and two. Column three lists the number of markers available within each category, which mapped to human chromosome regions for which a corresponding ECCS lies on CFA 5. Column four indicates the number of available markers for which PCR optimisation from dog genomic DNA was attempted, followed in column five by the number of markers for which optimisation was successful. This fraction is given in the final column, indicating the success rate of each marker category (with respect to the number of markers of the total tested for which robust amplification was achieved).

Marker type	Source	Number available	Number tested	Number optimised	Rate of success
CATS	A	16	15	3	3/15 (20%)
UM-STS	B	6	6	2	2/6 (33.3%)
TOAST	C^T	1	0	nd	n/a
Dog gene	C^D	8	6	6	6/6 (100%)
Human EST	C^E	1	0	nd	n/a
Canine TOAST	C^{CT}	1	1	1	1/1 (100%)
TOAST	D	6	5	3	3/5 (60%)
TOTAL		39	33	15	15/33 (45.5%)

Table 4.3d:

Database similarity search of PCR products amplified using type I markers. The locus name (column one) is followed by a description of the highest scoring sequence match identified by BLAST, the accession number of which is given in column three. The proportion of identical nucleotides, as a percentage of the total number of nucleotides within the aligned region, is given in column three, followed by the P-value reported for that alignment.

Marker	Highest scoring BLAST match	Accession no.	Identity	P-value
NPPA	Dog atrial natriuretic polypeptide mRNA	M12045	280/289 (96%)	e-140
CSF3R	no significant match (see text)	-	-	-
DIO1	*Canis familiaris* type I iodothyronine deiodinase (dio 1) mRNA	U11762	190/191 (99%)	e-102
VCAM1	*Canis familiaris* vascular cell adhesion molecule-1 mRNA	U32086	268/272 (98%)	e-143
NFIA	*Homo sapiens* chromosome 19, cosmid F17972, complete sequence	AC004660	232/249 (93%)	e-119
RPE65	*Canis familiaris* mRNA for retinal pigment epithelium abundant protein	Y16567	74/78 (94%)	2e-31
MMP1	Human collagenase and stromelysin genes	U78045	128/151 (84%)	1e-23
MMP1^{RT}	Human collagenase and stromelysin genes	U78045	147/172 (85%)	3e-30
THY1	*Canis familiaris* Thy-1 gene	AF103747.1	419/425 (98%)	0.0
CD3E	Dog CD3 epsilon subunit mRNA	M55410	136/140 (97%)	1e-68
GOT2	*B.taurus* mitochondrial aspartate aminotransferase mRNA	Z25466	106/119 (89%)	2e-30
HP	Chimpanzee haptoglobin (Hp) gene	M20760	112/119 (94%)	5e-43
TAT	*Canis familiaris* tyrosine aminotransferase gene	L47165	75/76 (98%)	7e-36
ALDOA	*Canis familiaris* aldolase A gene	U03565	115/117 (98%)	1e-55
APRT	Human adenine phosphoribosyltransferase (APRT) gene	M16446.1	160/187 (86%)	1e-17
DPEP1	Pig mRNA for dipeptidase	D13143	74/80 (92%)	2e-23
GP1BA	*Canis familiaris* glycoprotein Ib mRNA, complete cds	U19489	216/223 (96%)	e-108
SLC2A4	*Sus scrofa* mRNA for glucose transporter protein	X17058	69/77 (89%)	2e-19
SLC2A4^{RT}	*Canis familiaris* mRNA for partial Insulin-responsive glucose transporter, glut4 gene	AJ388533 .1	81/81 (100%)	3e-38
TP53	*Canis familiaris* tumor suppressor protein (p53) gene, partial cds	U51857.1	324/348 (93%)	e-156

4.3.4 Isolation of BAC clones using PCR primers for dog genes

A corresponding BAC was isolated for all markers for which the PCR produced a single product whose identity was confirmed by sequencing. A radiolabelled probe was generated for each gene marker, and was used to screen three filters (representing 18,432 clones per filter) selected at random from the dog BAC library generated by Li *et al.* (1999). Table 4.3e lists the markers for which a BAC was isolated, and the identity of the clones that were tested PCR positive with marker primers. In some instances, probes for more than one gene were pooled and used to screen the library for multiple genes simultaneously, to increase the efficiency of the procedure. On two occasions, one of the genes in the pool was not represented amongst those clones selected on the basis of their hybridisation signals. When this occurred, the probe was relabelled and used as the sole marker to screen a further three filters from the library. In such cases, the second screening process successfully identified at least one BAC clone for the gene of interest. At least one PCR-positive clone was examined by sequencing analysis to confirm whether the clone represented the expected locus. Database similarity search results from these analyses are shown in table 4.3f.

4.3.5 Mapping of gene markers to dog chromosomes by FISH analysis

DNA was prepared from each BAC as described in section 2.4.6. Typically a yield of approximately two to four micrograms of DNA were isolated from each 10ml culture. DNA was labelled with biotin-16-dUTP or digoxigenin-11-dUTP by nick translation, and a proportion of the product was analysed by electrophoresis to determine the size range and concentration of the resulting fragments. Labelled clones were hybridised for 16 hours to dog metaphases to determine their genomic location. Chromosomal assignments of these markers are listed in table 4.3g. In each case, hybridisation signals were consistently observed on both homologous chromosomes, enabling assignments to be made with a high degree of confidence. Where more than one BAC clone was analysed for a given gene, all were assigned to the same chromosomal location, as expected. Representative FISH analysis results for all markers mapped are shown in figure 4.3b. In addition, the successful mapping by FISH analysis of two dog BAC clones (THY1 and TAT) to cat metaphases is shown in figure 4.3b, for which the hybridisation period was extended to approximately 30 hours, under the conditions described in section 2.2.4. These assignments are discussed further in the legend for figure 4.3b.

Table 4.3e:

BAC clones isolated using type I markers. The markers used to screen the dog BAC library are listed in column one. This is followed by a list of clones that demonstrated positive hybridisation signals in the screening process, and which were subsequently shown to test positive by PCR amplification from bacterial suspensions using marker primers.

Marker	Clone address of BAC clones identified by type I marker screening					
ALDOA	197e17	285l15				
APRT	3f16	101e3	147h12	197d17		
CD3E	276o24	284p3				
CSF3R	76c17	82n11	105d11	105e21	115o8	143n18
DIO1	308j12					
GOT2^{L}	4d10					
GOT2^{S}	10l2	15d6	114g13			
GP1BA	218g20	264h2				
GUCY2E	381b19					
HP	318n14	319j10				
MMP1	38i15	142p16				
NPPA	90m7	167j20	209g20	214g17	352n7	357g20
DPEP1	6h5	7p5	95e6	96k10		
RPE65	253o18	260a4	266k5	337l8	345j10	
SLC2A4	197f8	225f12				
SLC2A4^{RT}	20j14	66g16	113h16			
TAT	202n18					
THY1	57o5	168e19	191a15			
TP53	278e24					
VCAM1	243j24	249e1	344n13	404h4		

Table 4.3f:

Database similarity search results from BAC clone templates. Sequence data derived from amplification of BAC clone DNA using marker primers were used in a BLAST database search to confirm the identity of the marker product. The highest scoring database match is followed by the proportion of identical bases over the alignment generated by BLAST, and the P-value reported for the alignment.

Marker	Highest scoring BLAST match	Accession no.	Identity	P-value
NPPA	Dog atrial natriuretic polypeptide mRNA, complete cds	M12045	294/296 (99%)	e-162
CSF3R	see text	-	-	-
DIO1	*Canis familiaris* type I iodothyronine deiodinase (dio 1) mRNA	U11762	190/196 (96%)	5e-99
VCAM1	*Canis familiaris* vascular cell adhesion molecule-1 mRNA, complete cds	U32086	267/271 (98%)	e-138
RPE65	*Canis familiaris* mRNA for retinal pigment epithelium abundant protein	Y16567	49/52 (94%)	1e-14
MMP1^{RT}	Human collagenase and stromelysin genes, complete cds	U78045	128/151 (84%)	2e-22
THY1	*Canis familiaris* Thy-1 gene, partial cds	AF103747.1	398/407 (97%)	0.0
CD3E	Dog CD3 epsilon subunit mRNA, complete cds	M55410	117/119 (98%)	7e-58
GOT2^{S}	*B.taurus* mitochondrial aspartate aminotransferase mRNA	Z25466	105/119 (88%)	4e-27
GOT2^{L}	*Homo sapiens* glutamic-oxaloacetic transaminase 2 (aspartate aminotransferase 2)	NM_002080	91/104 (87%)	3e-24
HP	Chimpanzee haptoglobin (Hpp) gene, exon 5	M20760	112/125 (89%)	1e-39
TAT	*Canis familiaris* tyrosine aminotransferase gene, partial cds	L47165	97/97 (100%)	3e-48
ALDOA	*Canis familiaris* aldolase A gene	U03565	134/141 (95%)	8e-44
APRT	Human adenine phosphoribosyltransferase (APRT) gene, complete cds	M16446.1	95/108 (87%)	2e-20
DPEP1	Pig mRNA for dipeptidase	D13143	115/134 (85%)	2e-27
GP1BA	*Canis familiaris* glycoprotein Ib mRNA, complete cds	U19489	127/151 (84%)	1e-41
SLC2A4	*Sus scrofa* mRNA for glucose transporter protein	X17058	69/76 (90%)	3e-19
SLC2A4^{RT}	*Canis familiaris* mRNA for partial Insulin-responsive glucose transporter, glut4 gene	AJ388533.1	83/85 (97%)	1e-39
GUCY2E^{RT}	*Canis familiaris* gene encoding retinal guanylate cyclase E	Y15484	123/124 (99%)	6e-65
TP53	*Canis familiaris* tumor suppressor protein (p53) gene	U51857.1	302/305 (99%)	1e-170

Table 4.3g:

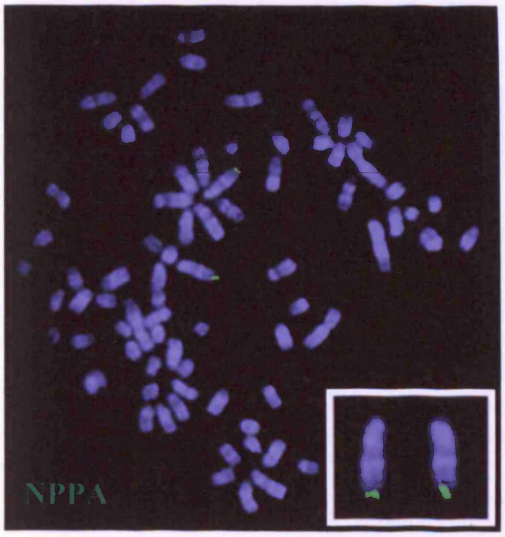
Dog chromosome assignment of type I markers. The locus name and human chromosome assignment of each locus is given in columns one and two. These data are followed by the predicted subchromosomal location of the dog orthologue, on the basis of the Zoo-FISH data generated in this study. For the two markers from HSA 1pdist. (NPPA and CSF3R), this prediction is shown in brackets to indicate that these are tentative predictions on the basis of a possible ECCS at this location that is shared with a region of CFA 5. The actual chromosomal localisation of each dog gene is then given in the final column. Human gene assignments were taken from the Human Gene Nomenclature Database.

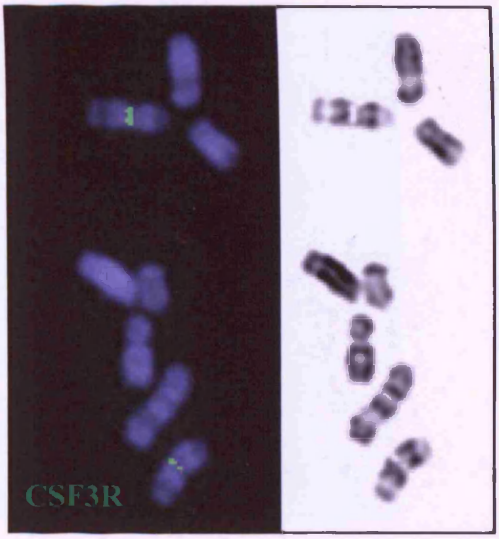
Locus name (HGND)	HSA Location	Predicted CFA location	Actual CFA location
NPPA	1p36	(CFA 5)	2q35 dist.
CSF3R	1p35-p34.3	(CFA 5)	9q21
DIO1	1p33-p32	5q24-q32	5q31 prox.
VCAM1	1p32-p31	5q24-q32	6q22
RPE65	1p31	5q24-q32	6q25
MMP1	11q21-q22	5q11-q14	5q14.3
THY1	11q22.3-q23	5q11-q14	5q13
CD3E	11q23	5q11-q14	5q13
GOT2	16q12-q22	5q33-q36	2q31-q32
HP	16q22.1	5q33-q36	5q34-q35prox.
TAT	16q22.1	5q33-q36	5q34-q35prox.
ALDOA	16q22-q24	5q33-q36	6q14
APRT	16q24	5q33-q36	5q33prox.
DPEP1	16q24	5q33-q36	5q33prox.
GP1BA	17pter-p12	5q21-q23	5q21
SLC2A4	17p13	5q21-q23	15q12
SLC2A4^{RT}	17p13	5q21-q23	5q21
GUCY2E	17p13.1	5q21-q23	5q21
TP53	17p13.1	5q21-q23	5q21

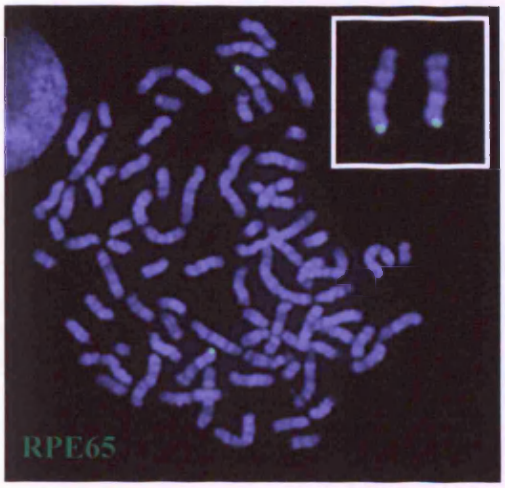
Figure 4.3b: (overleaf)

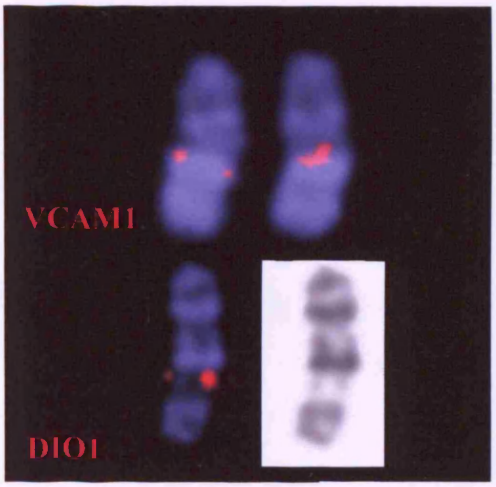
Assignment of gene markers to dog chromosomes by FISH analysis. Representative images are shown for each marker for which a corresponding BAC clone was isolated. The locus name is indicated in each case in the colour corresponding to the fluorescent signal observed. Where signals were of a small size, the individual chromosome pairs to which hybridisation were observed are enlarged in an inset figure. Markers are shown in order of their location in the human genome, with loci from HSA 1, 11, 16 and 17 shown in that order. Details of the isolation and chromosome assignment of each BAC clone are provided in section 4.3.5.

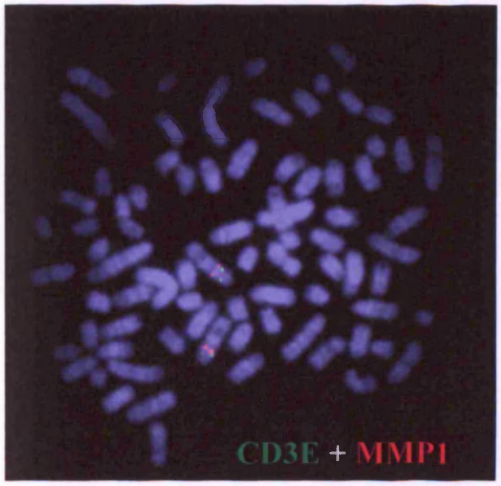
The final image shows the successful hybridisation of two dog BAC clones to cat metaphase chromosomes, demonstrating that cross-species FISH analysis is a potential method for generating comparative gene mapping data. THY1 and TAT mapped to FCA D1p and CFA E2q respectively, correlating with the locations of THY1 on HSA 11 and TAT on HSA 16. These assignments are therefore in complete agreement with the three-way comparison of human, dog and cat ECCS given in figure 4.2f.

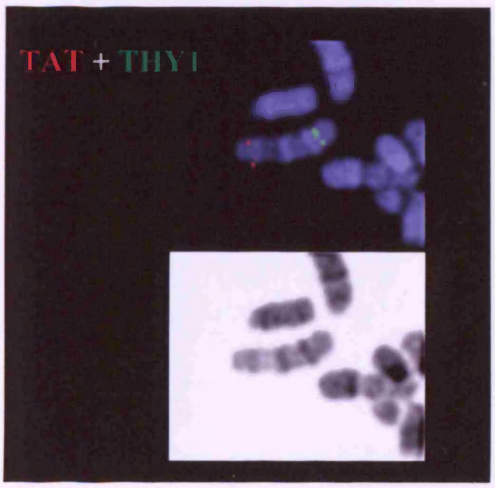

NPPA

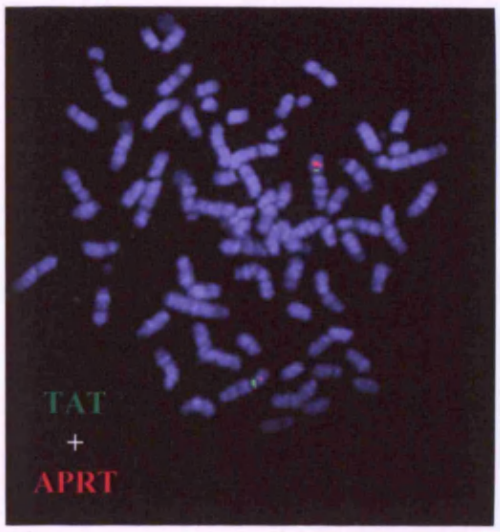

CSF3R

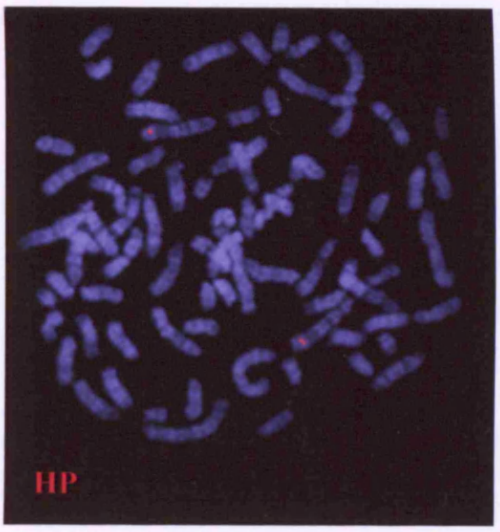

RPE65

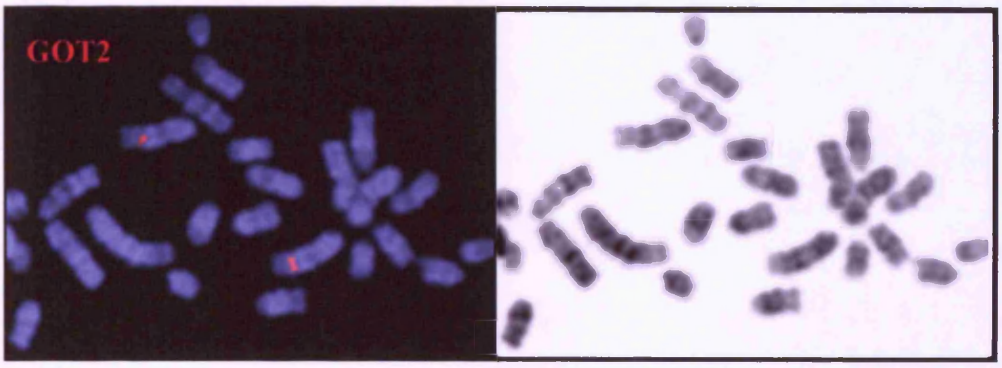

VCAM1
DIO1

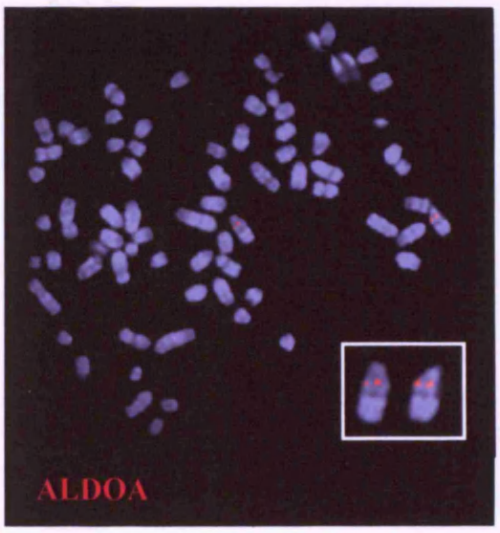

CD3E + MMP1

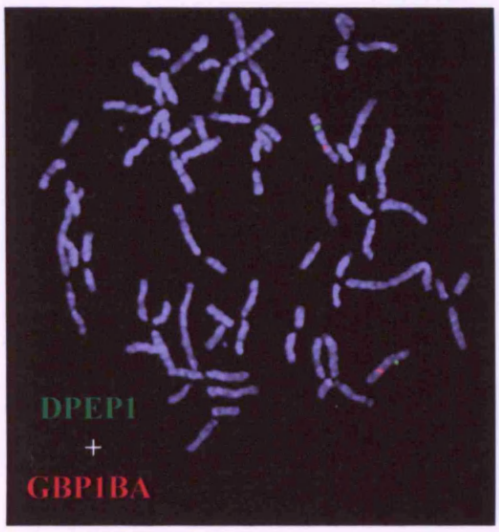

TAT + THY1

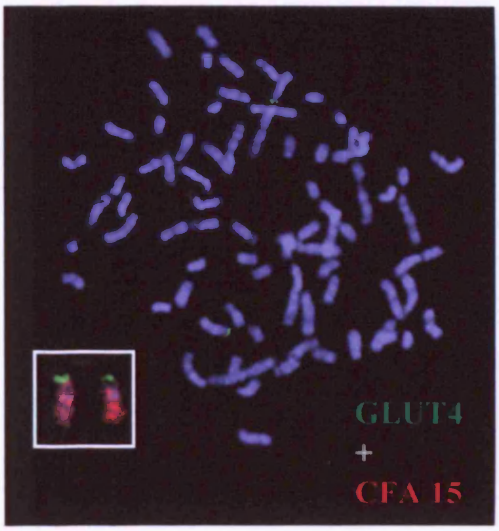

TAT
+
APRT

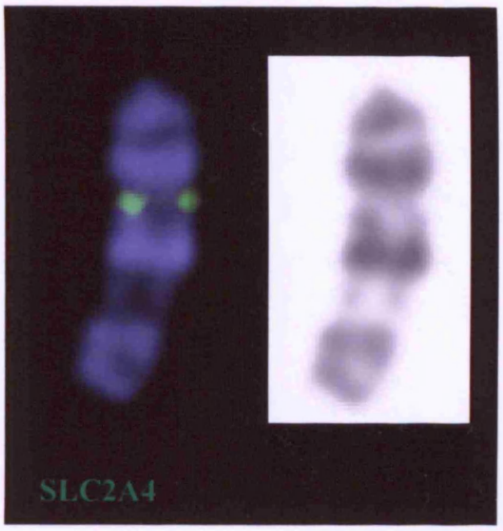

HP

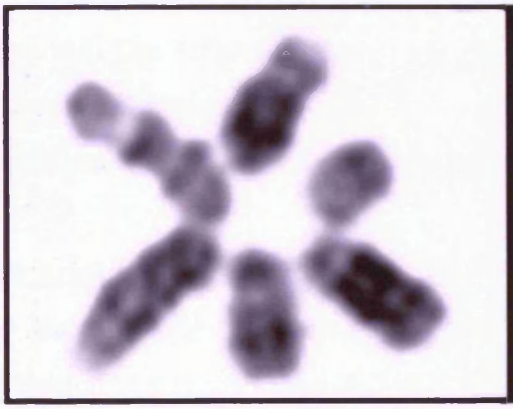

GOT2

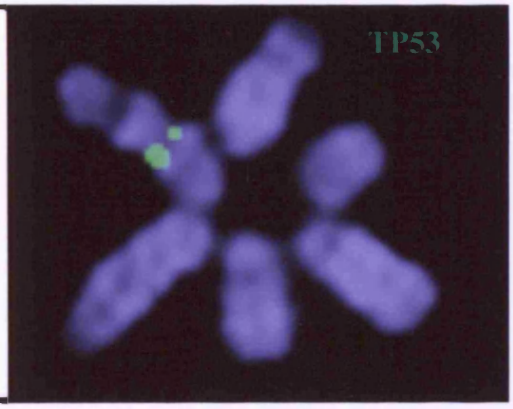

ALDOA

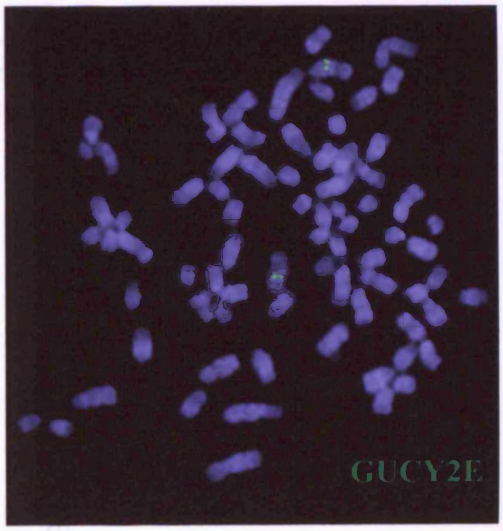

DPEP1
+
GBP1BA

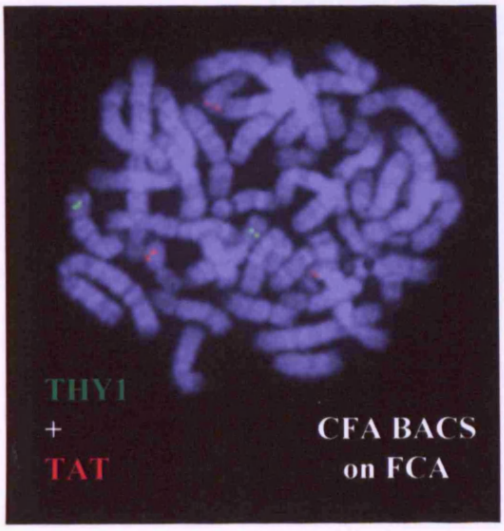

GLUT4
+
CFA 15


SLC2A4




TP53


GUCY2E


THY1
+
TAT
CFA BACS
on FCA

Data for the mapping of gene markers derived from each ECCS are summarised below, and reflect the approximate chronological order in which markers became available and were subsequently analysed. The identity of each BAC clone was verified by obtaining partial sequence data, and performing a database similarity search sequencing analysis prior to FISH analysis. Unless otherwise stated below, all were confirmed as representing the expected gene (see table 4.3f). Human gene assignments were taken from the Human Gene Nomenclature Database, unless otherwise stated.

- **HSA 1p (distal)**

Seven markers, representing four genes, were available to study this potential ECCS. Four of the seven markers represented the NPPA gene. The first of these four to be tested proved to be suitable for further analysis, and so the remaining three were not pursued. The CSF3R marker from the opposite end of the putative ECCS was also successfully optimised. The FGR marker consistently produced multiple products that could not be resolved, and was rejected. DDOST was not pursued as a result of data obtained from the mapping of the aforementioned markers.

Natriuretic peptide precursor A (NPPA)

The human NPPA gene is assigned to HSA 1p36 in the Human Gene Nomenclature Database, but a refined assignment to HSA 1p36.2 has been reported (Yang-Feng *et al.*, 1985). A marker for dog NPPA has been described by Venta *et al.* (1996), where the locus was described by one of its former names, PND. The predicted product size of 360bp was confirmed by PCR analysis that yielded a product from dog genomic DNA of approximately this size, and its identity was confirmed by sequencing analysis. The NPPA marker product was used to screen the BAC library and two clones were identified which were confirmed by the PCR. Both were used in FISH analysis to establish whether they mapped to dog CFA 5, which would provide evidence for the presence of a fifth ECCS. However, both BAC clones demonstrated hybridisation to CFA 2q35dist. This implies that HSA 1p36 shares conserved synteny with CFA 2 and therefore does not support the existence of a genuine fifth CFA 5 ECCS.

Colony stimulating factor 3 receptor (granulocyte) (CSF3R)

CSF3R has been assigned to HSA 1p35-p34.3 (Inazawa *et al.*, 1991) and to HSA1p34-p32 by Tweardy *et al.* (1992). This gene, in common with NPPA, lies where a putative

fifth region of conserved synteny between CFA 5 and the human genome was identified by reciprocal chromosome painting analysis. Lyons *et al.* (1997) developed a marker for CSF3R that was found to yield a product of approximately 450bp from dog genomic DNA. This product was sequenced and used to confirm the identity of the sequence using BLAST. However, neither forward nor reverse sequence data demonstrated significant similarity to existing database entries. In both cases, the highest ranked match was with anonymous human genomic DNA over regions of approximately 30 bases. Thus the identity of the marker product could not be confirmed in this manner. To further investigate the nature of the CSF3R marker, the product was used to screen the dog BAC library, from which six clones were shown to be positive using the marker primers. Sequencing analysis of these PCR products did not identity any significant similarity to existing database sequences. Three clones were selected at random for FISH analysis, and all mapped to CFA 9q21. This does not correlate with the HSA 1/CFA 5 ECCS identified by Zoo-FISH analysis. Furthermore, whole genome Zoo-FISH studies between human and dog (Breen *et al.*, 1999a; Yang *et al.*, 1999) have not identified detectable evolutionary similarity between HSA 1 and CFA 9, which might otherwise explain the assignment of the putative dog CSF3R locus.

Evidence against the presence of an extensive region of conserved synteny between CFA 5 and HSA 1pdist. was generated by the mapping of the NPPA gene from this human chromosome region to CFA 2. Data from whole genome reciprocal chromosome painting between dog and human had begun to provide further evidence to preclude this region of HSA 1p as sharing conserved synteny with CFA 5 (M. Breen, pers. comm.). This is discussed fully in the discussion at the end of this chapter.

- **HSA 1p32.3-p31.2**

Seven markers were available for analysis of the major region of conserved synteny shared by HSA 1 and CFA 5. Two markers represented the same gene, VCAM1, for which the first to be analysed (Priat *et al.*, 1999) was successfully optimised, and therefore the second was not pursued. DIO1 and NFIA were also successfully optimised. All three CATS markers, namely JUN, ACADM and RPE65 (Lyons *et al.*, 1997) were initially tested by the PCR, but none could be fully optimised to result in a single product. Such findings were also reported by the authors, and thus the markers were not pursued. However, a publication concerning the potential involvement of the RPE65

gene product in a canine visual disorder (Aguirre *et al.*, 1998) provided an alternative source of primers for this locus.

Deiodinase, iodothyronine type 1 (DIO1)

DIO1 maps to HSA 1p33-p32 (Jakobs *et al.*, 1997), spanning the distal boundary of the region of HSA 1 that corresponds to an ECCS on CFA 5. A marker for DIO1 has been described by Priat *et al.* (1999), with a predicted size of 239bp, which was confirmed by the PCR yielding a product of approximately 220bp. This was sequenced and confirmed as representing part of the DIO1 gene. A BAC clone was isolated and tested positive using the DIO1 marker, and was confirmed by sequencing analysis. This clone was used in FISH analysis to map the dog DIO1 gene to CFA 5q31prox.

Vascular cell adhesion molecule 1 (VCAM1)

The human VCAM1 gene has been mapped to HSA 1p32-p31 (Cybulsky *et al.*, 1991) and thus represents a marker on the proximal side of the ECCS on this chromosome that corresponds to CFA 5. This was used as an opportunity to locate a marker that delineated the proximal boundary of the ECCS. A VCAM1 marker for the dog has been described by Priat *et al.* (1999), with a predicted product size of 304bp. A product of approximately this size was obtained by the PCR using dog genomic DNA. A similarity search using sequence data from this product confirmed its identity. Four clones were isolated from the BAC library that tested positive on amplification with the VCAM1 marker primers, and were confirmed by sequencing analysis. This enabled the dog VCAM1 gene to be assigned to CFA 6q22. Human VCAM1 must therefore lie outside the region of HSA 1 that shares an evolutionary origin with CFA 5, and lies within an ECCS corresponding to CFA 6. This may also suggest that the assignment of the human gene can be refined to HSA 1p31.1, since a more distal location would have resulted in the mapping of the dog orthologue to CFA 5, according to the data generated in this study.

Nuclear factor I/A (NFIA)

NFIA is one of a family of four related genes that encode DNA binding proteins. The human chromosomal location of each has been established by *in situ* hybridisation, and NFIA has been mapped to HSA 1p31.3-p31.2 (Qian *et al.*, 1995). Jiang *et al.* (1998) have described a marker for the NFIA locus, which was used in the present study to amplify a product of approximately 300bp from dog genomic DNA. Sequence data were used to

establish the identity of the product. The most significant match was with an HSA 19 cosmid sequence (accession no. AC004660) with 232 identical bases over a total of 249 bases (93%). Of the ten most significant matches, seven are with the NFIX gene (HSA 19p13.3) from human or other species, and none are specifically with the NFIA gene. It is the 37th most significant match which first specifically details similarity with the NFIA gene, which in this case is from the mouse (accession no. AF111263), where the identity is reduced to 113 identical bases of a total of 137 (82%). The identity of the marker was thus uncertain, and it was not pursued.

Retinal pigment epithelium-specific protein (65kD) (RPE65)

The human RPE65 gene maps to HSA 1p31 (Hamel *et al.*, 1994), and thus lies within a cytogenetic band of which only part appears to represent an ECCS shared with CFA 5. As with VCAM1. the mapping of this locus in the dog could therefore be used to further confirm and refine the boundary of the ECCS. The RPE65 marker described by Lyons *et al.* (1997) could not be optimised for use on the dog. More recently, the RPE65 gene has been the subject of an investigation into its potential role in congenital stationary night blindness in the Swedish Briard (Aguirre *et al.*, 1998). This is caused by a 4bp deletion within the coding sequence of RPE65, and thus can be detected by standard genotyping analysis using PCR primers flanking the deletion site. The PCR primers discussed for this purpose were therefore also ideally suited for use in the present study. A product of 100bp was confirmed by the PCR, and by a database similarity search with sequence data from this product.

Screening of the BAC library led to the isolation of five BAC clones, which tested positive on amplification with the corresponding marker primers. FISH analysis resulted in the assignment of the dog RPE65 gene to CFA 6q25. This therefore indicates that as with VCAM1, the human RPE65 gene lies beyond the proximal boundary of the ECCS shared with CFA 5, and thus lies within the adjacent HSA 1 ECCS that corresponds to CFA 6. Similarly, it is proposed that the assignment of the human RPE65 gene may be refined to HSA 1p31.1, since the data generated in this study would infer that a more distal location would have resulted in the mapping of the dog homologue to CFA 5.

The RPE65 locus is of particular clinical significance to the dog mapping community, and was considered worthy of further study. Genotyping analysis at this locus was used to determine whether the mutation described by Aguirre *et al.* (1998) is present within the UK Briard population. Data obtained through the testing of 34 individuals led to the

identification of five carriers of the defective allele. It has therefore been demonstrated that the mutation is indeed present within the breed in this country, and at a rate significant enough to generate interest in a diagnostic test. Data are described in appendix D.

The assignment of a type I marker (DIO1) derived from HSA 1p33-p32 to CFA 5 confirmed the existence of a region of conserved synteny between these two chromosomes that had initially been identified by Zoo-FISH analysis. The proximal boundary of the conserved region was refined by the mapping of two further genes, VCAM1 and RPE65, to CFA 6.

- **HSA 11q23-q25**

Eight type I markers were available for the analysis of this ECCS, each representing a different locus. Three (THY1, MMP1 and CD3E) were successfully optimised for PCR analysis, all of which lay towards the proximal region of the ECCS on HSA 11. The remaining five markers (DRD2, CD3D, APOA1, APOC3 and ETS1) could not be optimised, of which only one (CD3D) had previously been reported to have been successful on dog genomic DNA (Lyons *et al.,* 1997).

Matrix metalloproteinase 1 (interstitial collagenase) (MMP1)

Human MMP1 has been assigned to HSA 11q21-q22 in the Human Gene Nomenclature Database, and to HSA 11q22-q23 in the NCBI OMIM database (Online Mendelian Inheritance in Man, http://www3.ncbi.nlm.nih.gov/htbin-post/Omim/). These assignments indicate that MMP1 lies close to the boundary between two ECCS on HSA 11. A more refined assignment has been provided, which maps the gene to HSA 11q22.3 (Pendás *et al.*, 1996), at the proximal boundary of the CFA 5/HSA 11 ECCS. The dog marker for MMP1 (Jiang *et al.*, 1998) produced a product of approximately 400bp from dog genomic DNA, corresponding with the predicted size of 390bp reported by the authors. Sequence data derived from this PCR product confirmed the identity of the marker. This also demonstrated the presence of a $(TA)_{25}$ microsatellite repeat sequence approximately 70 bases from the start of the forward sequence. The sequence data obtained beyond this point were of poor quality, most likely reflecting the fact that the individual from which the genomic DNA template was acquired was heterozygous at this locus, therefore producing a mixed sequence. Since the presence of a microsatellite in the

dog PCR product would be likely to create a high degree of non-specific background hybridisation on screening of the BAC library, a new primer was designed to avoid this repeat region. This was achieved using the reverse sequence data from which approximately 240 bases of high quality sequence were available prior to the repeat region beyond which quality deteriorated. This marker is hereafter referred to as $MMP1^{RT}$. The resulting PCR product was approximately 200bp in size, and a BLAST search confirmed its identity, with 85% identity over 172bp of the human MMP1 gene (accession no. U78045). Two BAC clones that tested positive with the new primer combination were isolated, and used in FISH analysis to map dog MMP1 to CFA 5q14.3. Genotyping analysis of the MMP1 gene marker was performed, enabling the locus to be integrated into the meiotic linkage map of CFA 5. Results are summarised and discussed in appendix E.

Thy-1 cell surface antigen (THY1)

The Human Gene Nomenclature Database places THY1 on HSA 11q22.3-q23, which spans the proximal boundary of the ECCS that corresponds to CFA 5. A more refined assignment to HSA 11q23.3 has been described (Tunnacliffe and McGuire, 1990). The THY1 marker described by Lyons *et al.* (1997) yielded a single product of approximately 800bp. The identity of the marker was confirmed by sequencing analysis of the PCR product, and by comparison of these data with existing sequences. Three BAC clones were isolated and tested positive with the THY1 marker primers, and the dog THY1 locus was thus assigned by FISH analysis to CFA 5q13. This would imply that human THY1 lies no further proximal than HSA 11q23, which represents the boundary of the HSA 11/CFA 5 ECCS.

CD3E antigen, epsilon polypeptide (TiT3 complex) (CD3E)

The T-cell antigen receptor consists of multiple subunits, of which both the delta (CD3D) and epsilon (CD3E) subunits have been mapped to HSA 11q23 (Gold *et al.*, 1987). This assignment of CD3E therefore places it as a proximal marker for the region of HSA 11 that shares an evolutionary origin with CFA 5. A marker for CD3E has been described by Priat *et al.* (1999) with a predicted product size of 178bp. Amplification of dog genomic DNA using the marker primers generated a PCR product of 180bp, whose identity was confirmed by a database similarity search. Two clones were isolated from the BAC library that tested positive with the marker primers, and their identity was confirmed by

sequencing analysis. Both were used in FISH analysis, and CD3E was thus assigned to CFA 5q13.

Three markers were therefore mapped to CFA 5 that correspond to genes located within the corresponding ECCS on HSA 11. No deviations from the expected mapping locations arose. Markers THY1 and CD3E could not be resolved by FISH analysis using metaphase chromosomes, but MMP1 clearly mapped distally to both these markers on CFA 5.

- **HSA 16q22-q24**

Eight type I markers were available for this human chromosome region, of which two loci (CTRB1 and APRT) were each represented by two primer pairs. Neither of the published CTRB1 primer pairs could be optimised for the PCR. The APRT marker developed by Venta *et al.* (1996) was successfully optimised and was used in preference to that of Lyons *et al.* (1997), which the authors had reported as not demonstrating amplification from dog genomic DNA. ALDOA, TAT and DPEP1 were also successfully optimised. GOT2 consistently generated two amplification products. A further marker for this region of HSA 16 was generated using data for the HP EST isolated in chapter three.

Glutamic-oxaloacetic transaminase 2 (mitochondrial) (GOT2)

The GOT2 gene maps to HSA 16q12-q22 (Jeremiah *et al.*, 1982), and was selected as a proximal marker for the evolutionarily conserved segment on this chromosome that corresponds to CFA 5. A marker for this gene has been described by Venta *et al.* (1996). Amplification from dog genomic DNA using the marker primers generated two products, one of approximately 1600bp, the other 300bp, the latter typically less intense. Attempts to optimise amplification conditions to yield a single product were not successful, and both bands were consistently observed. This finding had previously been observed by the authors (P. Venta, pers. comm.), who reported products of approximately 1500bp and 300bp. It was suggested that the latter represents a processed pseudogene, since the size of the product generated from application of the GOT2 marker primers to dog cDNA is predicted as 298bp (P. Venta, pers. comm.). Three such sequences have been described within the HGND, namely GOT2L1 (glutamic-oxaloacetic transaminase 2-like 1) located on HSA 1p33-p32, GOT2L2 on HSA 1q25-q31 and GOT2L3 on HSA 12p13.2-p13.1. It

is therefore likely that the two bands observed correspond to part of the dog GOT2 gene sequence and part of one of the corresponding pseudogene sequences that lie on HSA 1 and HSA 12.

The PCR was repeated with the GOT2 marker primers in order to generate a radiolabelled probe that consisted of both the 1600bp and 300bp products. This probe was used to screen the BAC library, from which eight positively hybridising clones were selected. All were tested with the GOT2 marker primers. Five clones generated a product of approximately 300bp, and were denoted $GOT2^S$. The remaining three clones, denoted $GOT2^L$, yielded a product of approximately 1600bp. It was therefore proposed that the latter three clones may represent the GOT2 gene, whilst the former five may correspond to one or more of the associated pseudogenes.

A single representative from the two GOT2 clone categories was assigned to a dog chromosome by FISH analysis. The $GOT2^L$ marker mapped to CFA 2q31-q32. This region corresponds to an ECCS at HSA 16q11.1-q13 (Breen *et al.*, 1999a; Yang *et al.*, 1999), proximal to the CFA 5/HSA 16 ECCS. It can therefore be proposed that GOT2 lies within the HSA 16/CFA 2 ECCS, and thus lies outside the region of HSA 16q that shares an evolutionary origin with CFA 5. The Zoo-FISH data generated in this study would therefore imply that the assignment of the human GOT2 gene, presently given as HSA 16q12-q22, may be refined to HSA 16q12-q13. This is possible as the mapping of GOT2 to CFA 2 implies that the human orthologue cannot lie distal to HSA 16q13, since this would be expected to result in its assignment to CFA 5. The $GOT2^S$ marker was assigned to CFA 27q. Although CFA 27 and CFA 30 share similar morphology, this assignment was confirmed by the co-hybridisation of the $GOT2^S$ marker with a chromosome paint probe for these dog chromosomes, in separate reactions. The entire length of CFA 27 corresponds to an ECCS at HSA 12pter-q14. The GOT2L3 pseudogene lies on HSA 12p13.2-p13.1, within the CFA 27/HSA 12 ECCS. It can therefore be proposed that this $GOT2^S$ marker contains part of the GOT2L3 pseudogene.

Haptoglobin (HP)

The human haptoglobin gene lies within HSA 16q22.1 according to the Human Gene Nomenclature Database, and therefore lies towards the proximal extreme of the HSA 16/CFA 5 ECCS. No marker for dog HP has previously been described. Hence, the isolation of a partial cDNA sequence for haptoglobin described in chapter three represented an opportunity to develop a novel marker for the dog gene. The forward

sequence data obtained from cDNA clone LEI079 had been shown to share 87% identity (466 of 533 bases) with the human mRNA for haptoglobin alpha 1S (accession no. X00637). This sequence was aligned with the human full length haptoglobin sequence (accession no. M69197), and showed that the majority of the partial sequence from LEI079 lay within the most 3' exon of the human haptoglobin gene. This enabled the design of a pair of primers which lay within this exon, avoiding the need to attempt amplification across an intron of unknown length.

The resulting product from dog genomic DNA was expected to be 175bp by the location of the primers within the sequence data previously obtained in chapter three. This was confirmed by the PCR from which a product of approximately 180bp was obtained, whose identify was confirmed by a database similarity search. Screening of the BAC library resulted in the identification of two clones which tested positive with the HP primers. Both were used in FISH analysis to map the dog HP gene to CFA 5q34-q35prox.

Tyrosine aminotransferase (TAT)

This marker has been assigned in separate studies to HSA 16q22-q24 (Barton *et al.*, 1986) and HSA 16q21 (Pol *et al.*, 1989). The standardised assignment of the gene, given within the HGND, is HSA 16q22.1, at the proximal extreme of the ECCS from this chromosome that is shared by CFA 5. TAT is therefore a useful marker for establishing the relative orientation of these segments on the two chromosomes. A marker for dog TAT (using the alternative locus name TYRA) has been described (Priat *et al.*, 1999). A product of 180bp was predicted, and was confirmed by PCR analysis, which yielded a product of this size.

The product was sequenced and confirmed as representing the expected gene. The BAC library was screened and a single clone was isolated, which tested positive with the marker primers. FISH analysis using this BAC clone enabled the dog TAT gene to be mapped to CFA 5q34-q35prox.

Aldolase A, fructose bisphosphate (ALDOA)

Human aldolase A lies on chromosome 16q22-q24 (Kukita *et al.*, 1987), an assignment which encompasses the majority of the ECCS shared with CFA 5. Priat *et al.* (1999) developed a dog marker for ALDOA, with a predicted size of 177bp. This was confirmed by the amplification of a product of approximately 180bp from dog genomic DNA using

the marker primers. The product was sequenced and its identity confirmed by a database similarity search. The probe was successfully used to isolate a single BAC clone that tested positive for amplification with the published marker primers, and whose identity was further confirmed by sequencing analysis and comparison with existing sequences. The BAC clone was assigned to CFA 6q14 by FISH analysis. This assignment was considered to indicate that the ALDOA marker represents an exception to the conservation of synteny between HSA 16 and CFA 5. This hypothesis is further investigated in the discussion that follows at the end of this chapter.

Adenine phosphoribosyltransferase (APRT)

The human APRT gene has been mapped to chromosome 16q24 (Fratini *et al.*, 1986), and thus lies in the distal extreme of the HSA16/CFA5 ECCS. Venta *et al.* (1996) described a marker for the dog APRT gene, which was predicted to amplify a product of 1300bp from dog genomic DNA. However, amplification using the APRT primers actually yielded a product of approximately 750bp in size. The product was sequenced and a BLAST search confirmed that the marker did represent part of the APRT gene. It was proposed that the size discrepancy is likely to reflect error in estimating the size of the dog PCR product reported by Venta *et al.* (1996) (P. Venta, pers. comm.).

Four BAC clones were isolated and tested positive for amplification with the APRT marker primers of Venta *et al.* (1996) and subsequently verified by sequencing analysis. Two were selected at random and used in FISH analysis, resulting in the assignment of the dog APRT gene to CFA 5q33prox.

Dipeptidase 1 (renal) (DPEP1)

The renal dipeptidase gene is assigned to HSA 16q24 within the HGND, and the assignment refined to HSA 16q24.3 by Austruy *et al.* (1993). A marker for DPEP1 has been described by Jiang *et al.* (1998) using the alternative locus name RDP. Application of the corresponding PCR primers to dog genomic DNA yielded a single product of approximately 150bp, which corresponded to the predicted size of 158bp. Sequencing analysis confirmed that this product represented the expected gene. Screening of the dog BAC library with the DPEP1 probe yielded four clones that tested positively on application of the marker primers. Two clones were used in FISH analysis, resulting in the mapping of the dog DPEP1 gene to CFA 5q33prox., near the proximal boundary of the ECCS that CFA 5 shares with HSA 16.

Four genes were assigned to the ECCS on CFA 5 that corresponds to HSA 16. FISH analysis demonstrated that TAT and HP map to the distal region of the ECCS on CFA 5, whilst DPEP1 and APRT map near the proximal boundary. Two further genes from the HSA 16 ECCS did not map to CFA 5, namely GOT2, which mapped to CFA 2q31-q32, and ALDOA, which mapped to CFA 6q14.

- **HSA 17pter-p12**

This ECCS produced the largest number of type I markers available for its study, with a total of nine gene markers described. Of these, three represent the TP53 gene. The TP53 marker (Lyons *et al.*, 1997), in addition to SLC2A4 and GP1BA, were successfully optimised. The remaining TP53 markers were thus not pursued. ASPA, MYH2, POLR2A and UBB all produced multiple products that could not be resolved to a single band by optimisation of conditions

Solute carrier family two (facilitated glucose transporter), member 4 (SLC2A4)

SLC2A4 (formerly known as GLUT4) is located on HSA 17p13 (Fan *et al.*, 1989), representing the fourth member of the solute carrier family number two, of which there are a total of five members plus a pseudogene. A product of approximately 120bp was produced from dog genomic DNA using the primers from the SLC2A4 markers of Priat *et al.* (1999). The product was sequenced and the data showed significant similarity to several existing sequences. The highest match was with the *Sus scrofa* mRNA for glucose transporter protein (accession no. X17058) with 69 identical bases over 77 in total, giving an overall 89% identity. However, this and the following two highest matches gave no information regarding the specific solute carrier family member involved. The fourth most similar sequence was the *Homo sapiens* solute carrier family 2 (facilitated glucose transporter), member 1 (SLC2A1) mRNA (accession no. K03195), with 52 of 55 identical bases, giving 94% identity overall. This suggested that the marker shares similarity with the SLC2A1 protein, and thus caused the real identity of the product to be questioned.

To further investigate the marker, the BAC library was screened with the PCR product derived from the SLC2A4 marker. Five clones were isolated and verified by the PCR. Sequencing of the product derived from applying the SLC2A4 primers to the BAC clones demonstrated comparable BLAST search results. Two BAC clones were used in FISH analysis, and mapped to a mid-sized chromosome that was identified as either CFA

15 or CFA 16. One BAC clone was co-hybridised in two separate experiments with paint probes representing these dog chromosomes. These experiments clearly showed that the SLC2A4 marker of Priat *et al.* (1999) does not lie on CFA 16, since the BAC clone and the paint probe mapped to different autosomes. The second paint probe represents both CFA 13 and 15, since these could not be resolved in the bivariate flow karyotype of the dog generated by Langford *et al.* (1996). Co-hybridisation with the putative SLC2A4 BAC clone showed that the BAC maps to one of the two chromosome pairs represented by the paint probe. It is readily possible to distinguish between CFA 13 and CFA 15 by DAPI-banding analysis, such that this experiment confirmed that the BAC clone isolated using the SLC2A4 marker of Priat *et al.* (1999) maps to CFA 15q12. Zoo-FISH data indicate that this assignment does not correlate with the location of the human orthologue on HSA 17p13, since CFA 15 and HSA 17 do not share an ECCS (Breen *et al.*, 1999a). Thus the identity of the marker product was in question.

A second primer pair was designed for the analysis of the dog SLC2A4 locus. A partial mRNA sequence for the dog gene has recently been deposited in GenBank (accession no. AJ388533). New primers were designed manually within this sequence at sites spanning approximately 130bp of the human SLC2A4 sequence, and this marker is hereafter denoted SLC2A4^{RT}. A product of approximately 120bp was produced with the dog template, implying that the primers lie within a single exon of the dog coding sequence. The resulting product was sequenced and a database search confirmed similarity to the dog SLC2A4 sequence (accession no. AJ388533). This product was radiolabelled and used to screen the dog BAC library. Three clones were demonstrated to be PCR positive for this marker, and two were used in FISH analysis. SLC2A4 was thus mapped to CFA 5q21.

Glycoprotein Ib (platelet), alpha polypeptide (GP1BA)

The human GP1BA gene has been assigned to HSA 17pter-p12 (Wenger *et al.*, 1989), thus the assignment spans the entire region of this chromosome, which has been shown to be evolutionarily related to CFA 5. A marker for dog GP1BA has been described by Priat *et al.* (1999), with a predicted product of 252bp from dog genomic DNA. This was confirmed by the PCR, which yielded a product of approximately 250bp, whose identity was confirmed by a database similarity search. Screening of the BAC library yielded two clones that tested positive with the marker primers, and both were used in FISH analysis, enabling the dog GP1BA gene to be assigned to CFA 5q21.

Tumour protein p53 (Li-Fraumeni syndrome) (TP53)

The TP53 tumour suppressor gene has been mapped to HSA 17p13.1 (Isobe *et al.*, 1986; McBride *et al.*, 1986; Miller *et al.*, 1986), within the midregion of the ECCS shared with CFA 5. Dog TP53 has previously been isolated in studies by Chu *et al.* (1998) and Veldhoen and Milner (1998), but no chromosomal assignment was made.

The marker for dog TP53 described by Lyons *et al.* (1997) yielded a product of approximately 1200bp. This was confirmed to correspond to the dog TP53 gene by sequencing analysis and a database similarity search. One BAC clone was isolated, which tested positive with the marker primers, and was used in FISH analysis, indicating that dog TP53 maps to CFA 5q21. This has also been confirmed by the mapping of an alternative TP53 clone to CFA 5q21 (M. Breen, pers. comm.).

Guanylate cyclase 2E (GUCY2E)

The human GUCY2E gene was mapped to HSA 17p13 by Yang *et al.* (1996) and subsequently given a map location of HSA 17p13.1 within the HGND. Both the complete mRNA (accession no. Y15483.1) and genomic DNA (accession no. Y15484.1) sequence have previously been determined for the dog gene (Veske *et al.*, 1996). This information had previously been used (A. Curson, pers. comm.) to design PCR primers to amplify a region of approximately 120bp of the dog GUCY2E gene. The product was sequenced to confirm its identity, and was subsequently used to isolate a BAC clone corresponding to the dog GUCY2E locus (A. Curson, pers. comm.) The gene was assigned by FISH analysis to CFA5q21.

All four gene markers analysed from the ECCS on HSA 17 were mapped to CFA 5q21, at the most proximal region of the CFA 5/HSA 17 ECCS. Their relative order could not be resolved by metaphase FISH analysis.

Chromosome assignments for all dog type I markers mapped in this chapter are shown in figure 4.3c against the relevant ideogram. Comparison is also made with the corresponding human ECCS on which the orthologous locus is located, enabling these data to be correlated with chromosome painting data from this and other studies.

Figure 4.3c: (opposite)

Comparison of chromosome assignments of human and dog genes with the location of corresponding ECCS. Chromosome assignments for all dog type I markers mapped in this chapter, and those of the human orthologue, are shown in against the relevant ideogram. The location of the corresponding ECCS in both genomes is also shown, and can be compared with gene marker assignments to determine the degree of correlation between the two sets of data. Anomalies, such as ALDOA, can therefore be readily identified and further investigated, as described in section 4.3.5. Human chromosome ideograms were adapted from Francke (1994), and dog ideograms from Breen *et al.* (1999b).

HSA 16
GOT2
ALDOA
HP
TAT
APRT
DPEP1
CFA 2
GOT2
CFA 5
HP APRT
TAT DPEP1
CFA 6
ALDOA
HSA 11
MMP1
THY1
CD3E
CFA 5
MMP1
THY1
CD3E
HSA 17
SLC2A4
GP1BA
GUCY2E
TP53
CFA 5
SLC2A4
GP1BA
GUCY2E
TP53
HSA 1
NPPA
DIO1
RPE65
VCAM1
38
CFA 2
NPPA
CFA 5
DIO1
CFA 6
VCAM1
RPE65

4.4 Discussion

Dog chromosome five was considered the most suitable candidate for the development of detailed reciprocal chromosome painting analysis in this species. This was in part due to the high quality cytogenetic resources available for this chromosome, a major consideration where cross-species FISH analysis is concerned. The isolated position of the peak corresponding to CFA 5 in the dog bivariate flow karyotype had already established this as a highly pure source of material derived from this chromosome. The application of the biotinylated CFA 5 paint to dog metaphases confirmed the probe as one that is highly specific for the expected chromosome at the cytogenetic level. Figure 4.2a demonstrates that background signal on other chromosomes is undetectable. The flow peak corresponding to CFA 2 is the nearest neighbour to the CFA 5 peak, and represents that most likely to contaminate the latter. However, no consistent signal was observed on CFA 2. Thus the CFA 5 paint probe may be considered pure at the cytogenetic level.

One consideration for cross-species chromosome painting is the nature of the competitor DNA to be used in suppressing repetitive elements located both within the chromosomal target and the paint probe itself. Human FISH analysis almost without exception utilises the highly repeat-enriched C_0t1 fraction of genomic DNA for this purpose. Approximately 50ng to 200ng of a single locus probe are typically competed with 1μg of human C_0t1 DNA. An additional quantity of sonicated salmon sperm DNA may be added to the probe mix to maintain a consistent DNA concentration and to act as a carrier to aid recovery of the sample should a precipitation step be necessary. For mapping single locus probes in the dog, and in the absence of a readily-available source of dog C_0t1 DNA, sonicated dog whole genomic DNA has been the choice to date. Typically this is used at a concentration ten-fold greater than that of C_0t1 DNA to compensate for the fact that the former contains a lower fraction of repetitive sequence. To establish which of these options was preferable for cross-species FISH analysis using the CFA 5 paint probe, parallel experiments were performed using equal quantities of probe but with different competitor. This indicated that the use of competitor corresponding to the chromosomal target rather than the probe generated more highly specific signal, with a lower level of background hybridisation. This has been reinforced by subsequent reciprocal chromosome painting analysis involving all human and dog chromosomes by Breen *et al.* (1999a). In the absence of a comparable source of dog C_0t1 DNA, it is not possible to determine whether this would have generated superior results. However, since the paint

probe corresponds in each case to a single chromosome, it would perhaps seem more logical for a greater need to suppress repeats within the more complex chromosomal target. It would be interesting to establish whether a combination of competitor from both species would provide even more specific hybridisation.

The hybridisation of the CFA 5 paint probe to human metaphase spreads demonstrated highly specific and consistent signals on four human chromosomes. The analysis of spreads of varying length from each reaction enabled these regions to be described to defined sub-chromosomal locations, which are indicated in figure 4.2b. Initially, five such regions were identified, one each on HSA 11q, 16q and 17p, and two on HSA 1p. Although there was a small degree of variation between the intensity of hybridisation observed between different chromosome spreads, it was noticed that the putative ECCS on the distal tip of HSA 1p was markedly less consistent than the other four. Further analysis generated data to indicate that this region of HSA 1p did not represent a genuine HSA 1/CFA 5 ECCS. The relative physical size of the confirmed ECCS can be estimated, using data extracted from the Unified Database for Human Genome Mapping (UDB, http:// bioinformatics.weizmann.ac.il/udb/) and the Genetic Location Database (http://cedar.genetics.soton.ac.uk/public_html/ldb.html). These data indicate that the human chromosome regions to which the CFA 5 paint probe hybridised are approximately 21.5Mb on HSA 1, 29.2 Mb on HSA 11, 33.1 Mb on HSA 16 and 16.6 Mb on HSA 17, a total of approximately 100.4 Mb. This compares well with the estimated size of CFA 5 as 99 Mb (Langford *et al.*, 1996).

This procedure therefore resulted in the identification of at least four regions of the human genome that can be considered to share an evolutionary origin with CFA 5. A unidirectional approach is in itself a major aid to the ability to compare their genome maps, enabling one to predict loci that lie on the latter by virtue of genes previously assigned within the former. It is not, however, possible to ascertain from these data where such loci are likely to lie on CFA 5, since they provide no information on which regions of this chromosome correspond to the identified human ECCS. However, the availability of paint probes from both species allows the reciprocal experiment to be performed, in which the corresponding human chromosome paints were hybridised to dog metaphases. A reciprocal approach significantly increases the information that can be generated by this technique, since it directly enables the sub-chromosomal location of corresponding ECCS in both species to be described. Thus the gene content of corresponding ECCS can be compared between species as separate entities rather than as part of a single

chromosome. In turn, prediction of a gene location mapped in one species may be refined to the corresponding ECCS rather than to the chromosome as a whole.

The hybridisation of the paint probes corresponding to HSA 1, 11, 16 and 17 to dog metaphases successfully resulted in the localisation of the four ECCS within the length of CFA 5. No significant overlap was detected, although a small degree of variation of the precise location of the boundaries between each ECCS was apparent across the cells studied. This is to be expected since variation in the degree of chromosome elongation within a single slide will have an effect on the appearance of the hybridisation signals. There is also likely to be some degree of variation in the actual hybridisation conditions across different regions of a given slide. The deviation of signals can be estimated as not in excess of 5 Mb in any case, which represents less than 5% of the total chromosome length. The absence of hybridisation signals for human and dog telomeric and centromeric regions is an observation that has been widely reported in previous Zoo-FISH studies, including the whole genome dog-human chromosome painting analysis of Breen *et al.* (1999a). This is widely thought likely to reflect the presence of repetitive elements in these regions that do not cross-hybridise between the species of interest.

The use of a chromosome paint pool (HSA 9-12) to obtain data for HSA 11 does not present uncertainty with respect to the FISH analysis of this probe on dog chromosomes, since the CFA 5 paint showed no detectable hybridisation to HSA 9, 10 or 12. It can therefore be concluded with confidence that CFA 5 does not share any gross homology with the additional three human chromosomes present in this pool. The only visible difference in the data obtained reflected the greater complexity of this probe, since the hybridisation sites of the probe were more numerous as a result of containing material from four chromosomes. The generation of a paint probe specific for HSA 11 could be achieved using a single flow-sorted or microdissected chromosome as the template for DOP-PCR. This approach was not required in this instance, since although a higher level of background signal was apparent on hybridisation of the mixed human paint probe to dog metaphases, the sites of hybridisation to CFA 5 could be clearly discerned with confidence.

Overall, the hybridisation of human chromosome paint probes to dog metaphases appeared to generate noticeably less background signal than was apparent in the reverse experiment. Individual peaks on the human bivariate flow karyotype are well isolated compared to those of the dog (figure 4.4a). As a consequence, it is not surprising that human paint probes are typically of very high purity, with little contamination from

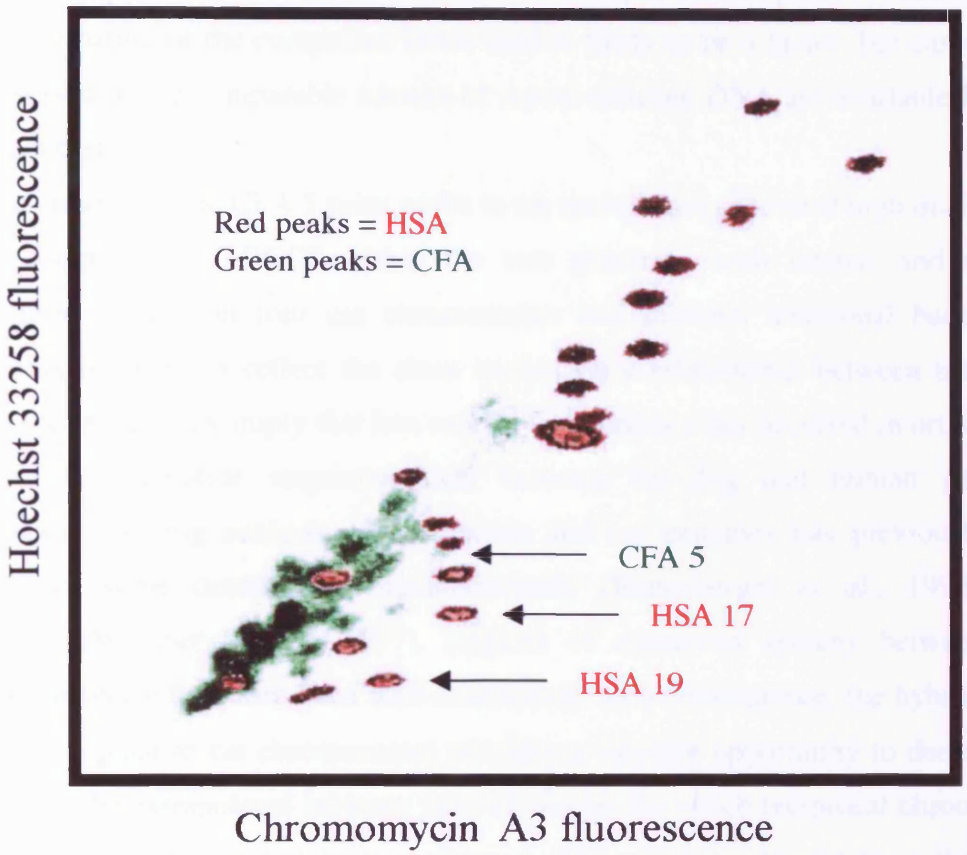


Figure 4.4a:

Comparison of the human and dog bivariate flow karyotypes (adapted from Langford *et al.*, 1996). Chromosomes from both species were subjected to bivariate flow sorting analysis in separate studies. The resulting flow peak profiles were overlaid, and are shown to scale. Peaks corresponding to human chromosomes are shown in red, and those representing dog chromosomes are shown in green. This shows that human chromosomes exhibit considerable variation in physical size and in their relative affinity for the DNA-binding dyes Hoechst 33258 and Chromomycin A3, compared to the dog karyotype. This results in a greater degree of spatial separation between human flow peaks than is possible for dog chromosomes. It is therefore likely that human chromosome paint probes generated from this flow-sorted material are of greater average purity than is possible from flow-sorted dog chromosomes. HSA 17 and 19, and CFA 5, are highlighted for the purpose of further discussion in chapter six.

adjacent flow peaks. Even though the CFA 5 paint probe appears to be highly specific at the cytogenetic level, it is likely that the degree of purity of human probes is considerably greater. The nature of the competitor DNA used is likely to be a factor, but can only be fairly assessed where comparable sources of repeat-enriched DNA are available for both species concerned.

The hybridisation of the CFA 5 paint probe to cat metaphases generated high quality data for the comparison of ECCS within the two genomes, with intense and specific hybridisation signals on four cat chromosomes and minimal additional background signal. This is likely to reflect the close evolutionary relationship between these two species, and in turn may imply that less extensive divergence has occurred in orthologous dog and cat nucleotide sequences than between the dog and human genomes. Chromosome painting analysis of the human and cat genomes has previously been described in some detail, both unidirectionally (Rettenberger *et al.*, 1995a) and reciprocally (Wienberg *et al.*, 1997). Regions of conserved synteny between their chromosomes have therefore been well established. As a consequence, the hybridisation of the CFA 5 paint to cat chromosomes provides a valuable opportunity to demonstrate how data can be extrapolated between pairs of species for which reciprocal chromosome painting analysis has not yet been performed. This can be achieved by a three-way comparison between the data obtained for other species, and is demonstrated in figure 4.2f. The CFA 5 paint probe hybridised to regions of HSA 1p, 11q, 16q and 17p. Similarly, hybridisation signals were observed on FCA C1p, D1p, E1p and E2q. Data from reciprocal chromosome painting analysis between human and cat showed that ECCS exist on HSA 1p and FCA C1p, HSA 11q and FCA D1p, HSA 16 and FCA E2q, and HSA 17p and FCA E1p (Rettenberger *et al.*, 1995a; Wienberg *et al.*, 1997). This enables four ECCS to be described across three different species, in the absence of reciprocal chromosome painting analysis between two of them (dog and cat). It should therefore be possible to suggest an approximate chromosomal location for an ECCS between the dog and any other species, providing that reciprocal chromosome painting analysis has been performed between this latter species and the human karyotype.

A physical map of the most gene-rich sites within the human genome has been produced (Saccone *et al.*, 1999). It is therefore possible to compare the ECCS that comprise CFA 5 with the corresponding regions of the human genome to determine whether CFA 5 is itself related to regions of high gene density. HSA 11q23.3, HSA 16q22.1, HSA 16q24.1 and HSA 16q24.3, and HSA 17p13.1 and HSA 17p11.2 are classed within the 9% most

GC-rich regions of the human genome. These all lie within ECCS corresponding to CFA 5. The correlation between gene density of homologous bands in the genomes of different species has not yet been widely established, but it would seem likely that some degree of extrapolation is possible. Thus CFA 5 does indeed appear to represent a chromosome with a higher than average gene content.

The mapping of type I markers to dog chromosomes was considered vital for refinement of the location of ECCS between the human and dog genome. Their assignment would also confer the potential to determine the relative orientation of the corresponding ECCS within the two genomes. The recent reports of gene markers for application to comparative mapping were considered the most logical and efficient resources with which this could be achieved. These reports (Venta *et al.*, 1996; Lyons *et al.*, 1997; Jiang *et al.*, 1998; Priat *et al.*, 1999) together describe 42 markers that lie within the five regions of HSA 1, 11, 16 and 17 that were identified as sharing a putative evolutionary origin with CFA 5. The overall spread of the markers within the human chromosomal regions concerned was highly comparable, with between seven and nine located within each, despite the varied relative sizes of the four ECCS. Taking account of duplications of loci, between six and eight unique gene markers were available for the mapping of the four confirmed ECCS. Simplistically, it would seem likely that the number of markers available for a given ECCS would reflect the physical size of the region of interest. However, it is also likely that chromosome regions that are of specific interest, and that have been extensively studied as a consequence, will contain a greater density of mapped markers. Sequence data are also likely to be more comprehensive. The ECCS on HSA 17p consists of only two chromosome bands, yet nine markers were available. However, of these, three represent the TP53 locus which has to date generated considerable interest due to its involvement in neoplasia. A total of six unique type I markers were therefore available for the study of the HSA 17/CFA 5 ECCS.

An evaluation of markers from each report gives an insight into the relative suitability of the different types for this application. Table 4.3c indicates the proportion of markers tested for which amplification conditions could be optimised. Overall, of 33 markers tested, 15 (45.5%) satisfied this criterion. It can be seen that the single canine TOAST and all six of the dog genes tested (Priat *et al.*, 1999) were successful. This is to be expected since the primers for both types of marker were designed directly from dog sequence information. Primers representing the remaining marker types were designed by alignment of available gene sequence data from multiple species, which did not

specifically include the dog. Thus their success for dog gene mapping relies on there being sufficiently high sequence conservation between the dog sequence and that from which the primers were designed that PCR primers can anneal and prime amplification. Three of the five TOAST markers (60%) (Jiang *et al.*, 1998) were successfully optimised, compared to two of the six UM-STS markers (33.3%) (Venta *et al.*, 1996) and only three of the 15 CATS primers (20%) (Lyons *et al.*, 1997). The numbers involved are too limited to be statistically valid, but the distinction between the value of markers designed specifically for the dog versus those designed to apparently conserved sequence data from various species is as clear as would be expected. The generation of more gene sequence data for the dog is therefore vital to increase the efficiency of mapping and characterisation of such loci.

Of the 15 markers successfully optimised for the PCR, 12 were confirmed to represent the expected gene by sequencing analysis. The three anomalous markers were CSF3R (Lyons *et al.*, 1997), SLC2A4 (Priat *et al.*, 1999) and NFIA (Jiang *et al.*, 1998). Sequence data obtained from the CSF3R marker product did not demonstrate significantly high similarity to any known gene. Six clones were isolated from the dog BAC library using the PCR product as a probe. Three clones selected at random were mapped to CFA 9q21. This does not correlate with the presence of an ECCS shared between CFA 5 and the distal tip of HSA 1p. This may simply reflect an exception to existing Zoo-FISH data, such as the presence of an HSA 1/CFA 9 ECCS that has not thus far been detected, perhaps due to its limited size. Breen *et al.* (1999a) determined that the proximal two thirds of CFA 9 (including CFA 9q21) corresponds to an ECCS consisting of the vast majority of HSA 17, namely HSA 17p11-qter. The remaining region of this human chromosome, HSA 17pter-p12, represents the ECCS corresponding to CFA 5 identified in this study. The mapping of the putative CSF3R clones to CFA 9q21 would infer that this locus would lie within the corresponding ECCS at HSA 17p11-qter. This suggests an alternative explanation, that the marker represents sequence associated with the colony stimulating factor 3 (CSF3) gene itself, and not the intended receptor protein, CSF3R. The CSF3 gene maps to HSA 17q11.2-q12, a region that has been well established to share conserved synteny with CFA 9 (Werner *et al.*, 1997; Breen *et al.*, 1999a; Yang *et al.*, 1999). This hypothesis would correlate with comparative data generated by each of these studies, and could be explained by error in the interpretation of a database search result and incorrect annotation of the marker. However this issue

may be resolved by more extensive sequencing analysis and subsequent comparison of these data with existing database sequences.

The human SLC2A4 gene maps to HSA 17p13. The two BAC clones isolated using the SLC2A4 primers of Priat *et al.* (1999) mapped to CFA 15q12. This is not consistent with the ECCS on HSA 17p that has been shown to correspond to CFA 5. Sequencing data provided evidence that the PCR product in fact represents the SLC2A1 gene from HSA 1p35-p31.3. This region spans ECCS corresponding to CFA 5 (HSA 1p32.2-p31.2) and CFA 2 (HSA 1pter-p32.3), as determined by the Zoo-FISH analysis of Breen *et al.* (1999a). The Zoo-FISH data of Yang *et al.* (1999) identity a further, smaller region of conserved synteny between HSA 1p34-p32 and CFA 15, which lies between the ECCS from CFA 5 and CFA 2 identified by Breen *et al.* (1999a) (figure 4.4b). The assignment of a BAC clone to CFA 15q21 therefore supports the hypothesis that the SLC2A4 primers of Priat *et al.* (1999) in fact amplify part of the SLC2A1 gene (HSA 1p35-p31.3). Subsequently, PCR primers were designed to an existing dog SLC2A4 sequence extracted from the GenBank database (accession no. AJ388533). Two BAC clones isolated with this product mapped to CFA 5q21, which correlate with Zoo-FISH data generated in this chapter. The anomaly concerning the SLC2A4 marker described by Priat *et al.* (1999) may have resulted from misinterpretation of the results of a BLAST search during the initial development of this marker. Where gene families are involved, this can easily occur since sequences deposited within the GenBank database are not always well annotated, and the specific identity of the subunit of a gene involved is often not made clear. Amplification of sequences not consistent with the expected identity, but sharing similarity to members of the same gene family, has been noted before (for example, Schibler *et al.*, 1998) A similar situation was identified with the NFIA marker product, which generated several matches with the NFIX gene. Although related, the nucleotide sequence of the four human members of this gene family (Qian *et al.*, 1995) has been shown to differ sufficiently to eliminate the possibility that the database similarity search is incorrect in identifying a discrepancy. For this reason, the marker was not pursued.

The identity of the remaining 12 markers was successfully confirmed by sequencing analysis and a subsequent database similarity search. In all cases, at least one corresponding BAC clone was isolated and mapped by FISH analysis. Initially the focus was placed on establishing whether the putative ECCS at the distal tip of HSA 1p did correspond to CFA 5. The NPPA locus (HSA 1p36) was mapped to CFA 2, which

therefore did not correlate with this hypothesis. Similarly, Mellersh *et al.* (2000) have recently reported the assignment of DDOST (HSA 1p36.1) to CFA 2. Although these data alone are clearly not sufficient to resolve the question, further evidence was provided at this time by the whole genome reciprocal chromosome painting analysis between dog and human (Breen *et al.*, 1999a). The hybridisation of the CFA 2 paint probe to human metaphases clearly exhibited hybridisation signal extending from HSA 1p36-p33. Similarly, the HSA 1 paint probe hybridised to CFA 2q35. The assignment of NPPA to CFA 2q35 thus correlates with these observations. The presence of hybridisation signal at the distal end of HSA 1p was found not only with the CFA 5 chromosome paint probe, but also with those for CFA 7 and CFA 15 (M. Breen, pers. comm.). Yang *et al.* (1999) identified a region of hybridisation of the CFA 5 paint probe to HSA 1p36.2-p36.1, but this was not confirmed by the assignment of type I markers. These data are further discussed in the following text. It may be that this region of HSA 1p contains some sequence motif that allows cross-hybridisation with these probes but which does not represent a true ECCS. This reinforces the value of the reciprocity of this approach, and of confirming data with the assignment of single locus probes in both genomes.

Only one marker (DIO1) corresponding to a gene from HSA 1 actually mapped to CFA 5. This in part reflects the relatively small size of this human chromosome region, for which until recently only the three CATS markers of Lyons *et al.* (1997) were available. None of these proved to be suitable for analysis due to non-specific amplification from dog genomic DNA, in common with the findings reported by the authors. The NFIA marker (Jiang *et al.*, 1998) was not pursued due to its ambiguous nature as determined by sequencing analysis. The mapping of VCAM1 and RPE65 to CFA 6 provided evidence that the HSA 1 region on which these loci lie does not form part of the ECCS that is shared with CFA 5. These markers therefore provide some indication as to where the proximal boundary of this ECCS lies. Similarly the mapping of NPPA (from HSA 1p36) to CFA 2 delineates the distal boundary, although subsequent whole genome chromosome painting analysis has demonstrated that the HSA 1/CFA 5 ECCS does not extend distally past HSA 1p33 (Breen *et al.*, 1999a).

The remaining three ECCS provided more potential gene markers for study and in turn offered greater opportunity for comparative analysis. Three markers from HSA 11q were mapped to CFA 5, which lay towards the proximal extreme of the ECCS in this region. The most proximal marker, MMP1 (assigned to HSA 11q21-q22 by the HGND) was

assigned to CFA 5q14.3, which is in the distal region of the corresponding ECCS. This assignment may imply that the ECCS on HSA 11 that corresponds to CFA 5 in fact extends proximally to include HSA 11q21-q22. Alternatively, the assignment of MMP1 to CFA 5 may indicate that the human MMP1 locus can be refined more distally to lie within the HSA 11/CFA 5 ECCS, which would imply that its location lies no further proximally than HSA 11q23. THY1 and CD3E, which lie distal to MMP1 on HSA 11q, appear to lie proximal to this marker on CFA 5, confirming the presence of the HSA 11/CFA 5 ECCS. These assignments also suggest that the ECCS lie in opposing orientations relative to each other with respect to the centromere of the chromosome concerned.

Four markers corresponding to genes on HSA 16 were mapped to CFA 5. APRT and DPEP1 are both located towards the distal end of HSA 16q, whilst TAT and HP lie at the proximal boundary of the ECCS. However, APRT and DPEP1 were found to lie proximal to TAT and HP on CFA 5. Again, this implies that the ECCS on HSA 11 lies in an opposing orientation to the corresponding region on CFA 5.

The GOT2 marker proved complicated to analyse due to the presence of two bands subsequent to the PCR, the smaller of which is thought to represent a pseudogene (P. Venta, pers. comm.). FISH analysis of a BAC clone, $GOT2^{L}$, which appeared to represent the true gene, mapped the locus to CFA 2q31-q32. This correlates with GOT2 being located beyond the proximal boundary of the HSA 16/CFA 5 ECCS, and within the adjacent ECCS shared by HSA 16 and CFA 2. Thus the boundaries involved are well defined in this instance. The existence of GOT2 pseudogenes was demonstrated by the assignment of an alternative BAC clone, $GOT2^{S}$, from which a smaller amplification product was generated using marker primers. The assignment of this BAC clone to CFA 27 could be correlated with the presence of one such pseudogene in the corresponding ECCS on HSA 12. Further sequencing analysis would be required to fully investigate the implication of this finding, however, this was considered to lie outside the scope of this study. The nature of pseudogenes is such that there is likely to be a significant degree of sequence similarity between such loci, although both $GOT2^{L}$ and $GOT2^{S}$ probes were observed to hybridise to different, unique sites within the dog genome. A decrease in the stringency of FISH hybridisation and washing conditions may reveal additional, weaker signals at the sites of the other members of this gene family.

A further marker for the HSA 16/ECCS also did not map to CFA 5. ALDOA is located on HSA 16q22-q24 and this assignment lies entirely within the region of HSA 16 ECCS

that corresponds to CFA 5. A BAC clone containing the dog ALDOA locus was mapped to CFA 6q14, and this assignment is not consistent with the HSA 16/CFA 5 ECCS. Neither is it consistent with the adjacent HSA 16/CFA 2 ECCS that explained the mapping data for GOT2. However, an ECCS corresponding to CFA 6 lies on HSA 16p, since Zoo-FISH analysis has shown homology between HSA 16p13-p11.2 and CFA 6q14-q22prox. (Breen *et al.*, 1999a). The mapping of ALDOA may reflect an internal rearrangement of the ECCS corresponding to HSA 16, such as a pericentric inversion that has resulted in a chromosome region corresponding to CFA 6 becoming adjacent to the CFA 5 ECCS on HSA 16. This is demonstrated in figure 4.3c. Alternatively, these data may identify an additional ECCS located on HSA 16 and CFA 6 of a size that falls below the resolution limits of Zoo-FISH analysis. It is only possible to speculate on the explanation of this observation at this early stage of comparative mapping between the dog and human. More comprehensive gene mapping in the dog will be necessary in order to understand fully this apparent exception to synteny conservation.

Four genes from HSA 17 were mapped to CFA 5. SLC2A4, TP53, GUCY2E and GP1BA lie within the ECCS from this human chromosome. Existing cytogenetic data are not of sufficient resolution to allow their relative order to be determined in either species. Consequently the relative orientations of the corresponding ECCS cannot be established with great certainty in the absence of more detailed mapping information. The relationship between HSA 17 and CFA 5 has previously been investigated by Werner *et al.* (1997), who mapped clones representing HSA 17 loci to dog chromosomes by FISH analysis. The karyotype of Selden *et al.* (1975) was used, for which the nomenclature corresponds directly with that used in this study. Of 11 clones studied from HSA 17q, all mapped to the proximal two thirds of CFA 9, whilst two clones (GLUT4 and PMP22) from HSA 17p both mapped to CFA 5. The mapping of PMP22 (HSA 17p12-p11.2) to CFA 5 may suggest that the region of conserved synteny with CFA 5 extends proximally to 17p11.2, whereas in this study the hybridisation of the CFA 5 paint probe did not appear to extend proximally beyond HSA 17p12. Alternatively, the human locus may lie specifically in HSA p12, in which hybridisation signals were clearly apparent in a large proportion of the spreads studied. Co-hybridisation of the two cosmid clones demonstrated that GLUT4 lies proximal to PMP22 on CFA 5 (Werner *et al.*, 1997). These data indicate that the corresponding ECCS on CFA 5 and HSA 17 lie in inverted orientations with respect to the centromere of each chromosome concerned. Application of an HSA 17 paint probe to dog metaphases also demonstrated hybridisation to the

proximal two thirds of CFA 9, and also to the region of CFA 5 to which GLUT4 and PMP22 were assigned. This correlates with the data obtained in this study, and also that for the whole genome human-dog reciprocal chromosome painting analysis (Breen *et al.*, 1999a).

Chromosome assignments for type I markers were compared with existing gene mapping information for the dog. Guevara-Fujita *et al.* (1996) described the assignment of six genes to dog autosomes by FISH analysis, including the APRT and TP53 loci discussed above. PCR products were generated using the markers of Venta *et al.* (1996), and were used as probes to screen a dog genomic cosmid library. Corresponding cosmid clones were mapped by FISH analysis and assignments were made on the basis of G-banding and comparison to the karyotype proposed by Stone *et al.* (1991). APRT and TP53 were thus assigned to CFA 3 and CFA 5 respectively by Guevara-Fujita *et al.* (1996). Clearly this does not correlate with the data from the present study in which both loci mapped to the same chromosome. From the nomenclature used by Stone *et al.* (1991), CFA 3 corresponds to CFA 5 of the karyotype used throughout this report (Breen *et al.*, 1999b). Thus the assignment of APRT by Guevara-Fujita *et al.* (1996) agrees with the data described in this chapter. The assignment of TP53 by Guevara-Fujita *et al.* (1996) to CFA 5, using the nomenclature of Stone *et al.* (1991), equates to CFA 7 of that used in this study (Breen *et al.*, 1999b). This therefore indicates a discrepancy between the mapping data for TP53, which has been further confounded by the absence of a universally accepted nomenclature for the dog karyotype. Guevara-Fujita *et al.* (1996) did not appear to have co-hybridised APRT and TP53, and so it is possible that this is simply a misassignment. However, the identity of their TP53 clone is unclear since the cytogenetic data of Guevara-Fujita *et al.* (1996) demonstrate the hybridisation of their putative TP53 marker to the extreme end of a dog autosome, for which no banded images were presented. This does not correlate with the localisation in the present study of the dog TP53 gene to the midregion of CFA 5, specifically to CFA 5q21 (using the nomenclature of Breen *et al.*, 1999b). This assignment was further confirmed by co-hybridisation with three other CFA 5 clones, and is supported by reciprocal chromosome painting data, since TP53 (located on HSA 17p13) maps to the region of CFA 5 to which the HSA 17 paint probe also hybridises. Furthermore, the cat TP53 gene has been assigned to E1p14-p13 (Cho *et al.*, 1997b), a region which corresponds to an ECCS on HSA 17p and CFA 5q, which are also the locations of the homologous gene in these species. This issue demonstrates the power of the reciprocal chromosome painting

technique. The ECCS to which a given type I marker is mapped can readily be compared to that in which the homologous gene lies in the human (or other) genome. Any anomalies, such as that discussed for TP53, are therefore more easily recognised than would be the case in the absence of such data.

Of the 19 loci mapped in this study, six appear in the most recent dog genome map (Mellersh *et al.* 2000), in which assignments by meiotic linkage and radiation hybrid analysis are given to the whole chromosome level. The assignments of three genes (ALDOA and VCAM1 on CFA 6, and TAT on CFA 5) correlate in both studies. TP53 and GP1BA, mapped in this chapter, remained unlinked to any other marker in the report of Mellersh *et al.* (2000), whilst NPPA was mapped by Mellersh *et al.* (2000) to an unassigned radiation hybrid linkage group. Of the 12 genes assigned to CFA 5 in this study, three (CD3E, SLC2A4 and DIO1) are represented in the CFA 5 map of Mellersh *et al.* (2000). DIO1 and CD3E are represented only in the RH map, whilst SLC2A4 is present only in the meiotic map. One further gene (TAT) is assigned by both methods to CFA 5 by Mellersh *et al.* (2000), but is not included in the CFA 5 map. The remaining eight genes are therefore additional to the current CFA 5 map of Mellersh *et al.* (2000). This study represents nine novel FISH assignments of CFA 5 genes, two refined assignments (APRT and SLC2A4, following the previous assignments by Guevara-Fujita *et al.,* 1996 and Werner *et al.*, 1997, respectively), and one corrected assignment (TP53).

These gene assignments may be correlated with existing information for comparative mapping between the dog and human genomes. The dog-human chromosome comparison of Yang *et al.* (1999) differs from that of Breen *et al.* (1999a) in a number of areas. Yang *et al.* (1999) used the red fox karyotype (*Vulpes vulpes*, 2n = 34) as an intermediate species for which chromosome assignment is less complicated than that for the domestic dog. Both species belong to the order Canidae, and are believed to have shared a common ancestor some 10 million years ago (Wayne, 1993). Metaphase chromosome spreads and a set of chromosome paint probes were generated for human and both canids from peripheral blood lymphocyte and fibroblast cultures. Additional dog chromosome preparations were derived from an established dog kidney cell line (MDCK), and these formed the majority of the material used for flow-sorting and chromosome paint production. Hybridisation conditions were highly comparable to those used by Breen *et al.* (1999a), except that no competitor DNA was included in the probe mixture. Since the MDCK cell line used was known to have a rearranged karyotype, each paint was hybridised to normal dog, fox, and MDCK metaphases, to identify the DNA

content of each. Three paint probes were eliminated from further study due to detection of rearrangements, whilst the remaining 35 were considered normal.

Domestic dog paint probes were hybridised to fox metaphases, typically hybridising to a single chromosome arm or region, or in four cases, to two separate segments. In the reciprocal experiment, fox paint probes were hybridised to dog metaphases, and typically painted entire arms of between two and four dog chromosomes. A complete comparative map of both fox and dog karyotypes was generated from these data. Human chromosome paint probes were then hybridised to both dog and fox metaphases, identifying 73 ECCS between human and canid karyotypes in each case. Subsequently, all dog paints were hybridised to human metaphases, detecting 90 ECCS. A comparative map of the karyotypes of all three species was then generated, using the fox chromosome complement to correlate the location of dog and human ECCS. Their proposition was that the red fox be used as the karyotype to which human markers are assigned for the identification of the corresponding dog ECCS, rather than using the domestic dog chromosome itself as a template. A probe from a known human chromosome location would thus be mapped to a fox chromosome, and the dog assignment would then be inferred from the known dog-fox comparative map. A number of dog markers were assigned to chromosomes by performing the PCR using marker primers and template DNA from each flow peak, and determining the peak for which a product was obtained (Sargan *et al.*, 1999). Instances where products were obtained using material obtained from more than one peak were attributed to minor cross-contamination of sorted chromosome material, presumably by that from adjacent flow peaks.

Data relating to CFA 5 are highly comparable between this study and that of Yang *et al.* (1999), and are compared in figure 4.4b. The most significant difference relates to the presence of a small region of homology to CFA 9 at HSA 17p13.3. The identification of this region is likely to reflect the difference in composition of the paint probes used in these studies. It may reflect a more complete representation of this CFA 9 within the corresponding paint probe described by Yang *et al.* (1999), compared to those used in this study and that of Breen *et al.* (1999a). Differences in the quality of hybridisation signal obtained with human chromosome paint probes produced by different manufacturers has been noted before, by Vezuli *et al.* (1997) amongst others. Here, apparent variation in the intensity and uniformity of hybridisation was attributed to the degree of repetitive sequence present in the probe, and its overall concentration. It is possible also that the difference with respect to the use of competitor DNA in probes for

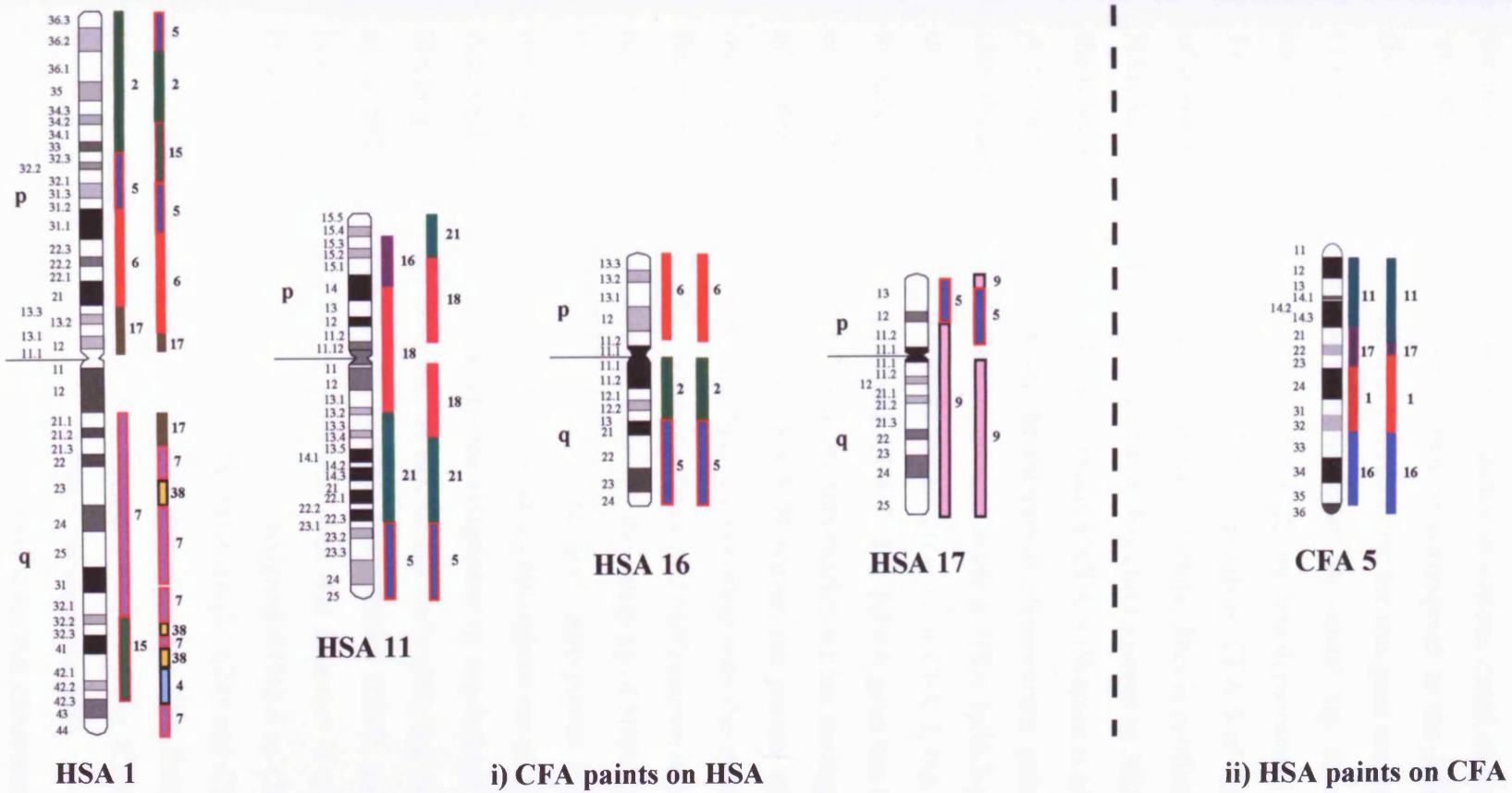


Figure 4.4b:

Comparison of chromosome painting data obtained by two different Zoo-FISH studies, with reference to CFA 5. To the right of each ideogram lies a series of blocks representing the corresponding ECCS on the chromosome of the second species. The chromosome number represented is shown to the right of each block. Data from this study combined with that of Breen *et al.* (1999a) are shown immediately to the right of each ideogram, followed by that from the study of Yang *et al.* (1999). Human ideograms are taken from Francke (1994), and the CFA 5 ideogram from Breen *et al*,. (1999b).

FISH analysis in the two studies may have had some effect. The presence of this additional ECCS at HSA 17p13.3 that correlates with CFA 9 is supported by data from the physical mapping of HSA 17 genes to various canid chromosomes by Park (1996). The MDCR locus has been shown to correspond to the LIS1 gene at HSA 17p13.3 (Reiner *et al.*, 1993). A human cosmid clone for this gene was mapped in the dog by Park (1996), and was reported to lie near the distal tip of CFA 23. However, the ideogrammatic representation of the mapping data demonstrate that the chromosome that Park (1996) described as CFA 23 corresponds to CFA 9 of the karyotype nomenclature of Breen *et al.* (1999b) used in the present study. This is confirmed by the mapping of the RARA gene from HSA 17 to the same dog chromosome as MDCR by Park (1996), since the RARA gene has since been assigned to CFA 9 (Werner *et al.*, 1997).

A further difference between the reciprocal chromosome painting data of Breen *et al.* (1999b) and Yang *et al.* (1999) concerns a site at HSA 1p36.3-p36.2. In the former study this region is reported to correspond to a region of CFA 2, but in the latter is considered to share sequence similarity with CFA 5. The NPPA gene has been mapped (Yang-Feng *et al.*, 1985) to HSA 1p36.2, placing this marker within the region described by Yang *et al.* (1999) as corresponding to CFA 5. However, the present study has demonstrated the mapping of NPPA to CFA 2q35dist., correlating with the chromosome painting data of Breen *et al.* (1999a). The previously discussed explanations regarding HSA 17p13.3 may equally apply to this situation, although the mapping of NPPA provides more support for the data reported by Breen *et al.*, 1999a. Such ambiguities will only be resolved by an increased density of mapped type I markers throughout the dog genome.

A further anomaly is provided by the assignment of dog linkage groups, generated by RH mapping analysis (Priat *et al.*, 1998), to the corresponding dog chromosome. Linkage group RH.49a contains DIO1, and RH.55a contains CD3E, and the assignment of both linkage groups to CFA 5 corresponds with that reported in chapter four of this study. However, Yang *et al.* (1999) have also assigned RH42.a to CFA 5. This linkage group contains two gene markers, ALDOA (HSA 16q22-q24) and CD19 (HSA 16p11.2). The assignment of RH42.a to CFA 5 was based on the fact that PCR primers for CD19 generated a product from flow-sorted material from CFA 5, whilst the human chromosome assignment of ALDOA correlates with its location on this dog chromosome. However, Yang *et al.* (1999) note that chromosome-painting data do not indicate any detectable homology between HSA 16p11 and CFA 5, and Zoo-FISH data would place CD19 on CFA 6. Chapter four describes the mapping of ALDOA to CFA

6q14, and would thus imply that CD19, linked to ALDOA by RH analysis, also lies on CFA 6. These anomalies involve assignments to a dog ECCS that does lie on HSA 16, but is not that predicted on the basis of the human gene location and the related Zoo-FISH data. Rather than any error in assignment, these data would imply an internal rearrangement of ECCS within HSA 16 relative to the corresponding dog chromosome regions, that cannot be detected at the resolution possible using chromosome painting analysis.

4.5 Summary

Reciprocal chromosome painting analysis has identified regions on HSA 1, 11, 16 and 17 that share homology with CFA 5. This has been further confirmed by hybridisation of the CFA 5 paint probe to cat metaphases, and comparison with existing cat-human synteny data. Nineteen gene markers, existing and novel, have been optimised for the PCR, and each has been mapped to the corresponding dog chromosome. This has resulted in the assignment of 12 genes to CFA 5 and refinement of the location of the ECCS boundaries. Anomalies regarding the identity of two gene markers have been identified and confirmed by sequencing analysis and by their chromosome assignment. A further five gene markers have been assigned to dog chromosomes other than CFA 5, providing evidence for other ECCS between the human and dog genomes, and aiding the refinement of those directly relating to CFA 5. The assignment of type I markers to CFA 5 is extended in the following chapter by the integration of their cytogenetic assignments with those generated by WG-RH mapping analysis.

Chapter five:

Towards a radiation hybrid map of dog chromosome five

Towards a radiation hybrid map of dog chromosome five

Abstract

Twelve type I markers from CFA 5, developed in chapter four, were assessed for their suitability in radiation hybrid mapping analysis, towards the production of a low resolution map of this chromosome. In addition, a panel of ten anonymous cosmid clones was selected from those previously mapped to CFA 5, and their chromosomal location was verified by FISH analysis. The ten clones were chosen by virtue of being relatively evenly spaced along the chromosome. End sequencing of the cosmid clones enabled the design of unique PCR primers such that each could be subjected to radiation hybrid mapping analysis, creating a framework map for localisation of additional markers.

Each marker was initially included in an optimisation procedure to establish suitable conditions for the PCR. It was also necessary to establish whether amplification occurred from the rodent recipient cell line used to create the panel, and if so, whether the dog product could be distinguished from the rodent product. For three type I markers, suitable amplification conditions could not be established, and new primer pairs were therefore designed.

Eleven type I and all ten type II markers were successfully optimised and assayed across the radiation hybrid panel. For each marker, the pattern of amplification was recorded from each cell-line comprising the panel. These data were used to identify pairs of linked markers, and to form larger linkage groups containing three or more markers. Where possible, the order of markers within linkage groups was investigated, and the most likely order was inferred from the data obtained. The observed gene order was compared with data obtained by FISH analysis, and where possible, comparisons were also made with data from existing maps of CFA 5.

5.1 Development of framework markers from anonymous CFA 5 cosmid clones

5.1.1 FISH analysis of a panel of ten cosmid markers

Ten evenly spaced cosmid markers were selected from a panel of 28 that had been assigned to CFA 5 in a previous study (M. Breen, pers. comm.). Their cytogenetic location was verified by standard FISH analysis. One microgram of each cosmid DNA was labelled with either biotin-16-dUTP or digoxigenin-11-dUTP. Fifty nanograms of one biotinylated and one digoxigenin-labelled cosmid were combined with 10μg of sonicated dog genomic DNA. Probe mixtures were hybridised overnight at 37°C to dog metaphase spreads. The physical assignment determined for each clone is shown against the CFA 5 ideogram in figure 5.1a., following the nomenclature of Breen *et al.* (1999b).

5.1.2 End sequencing of cosmid clones

Approximately 500ng of cosmid DNA were used as the template in a cycle sequencing reaction. The 'Big Dye' sequencing kit (Perkin Elmer) was used in preference to the dRhodamine dye set used for PCR products since the former is considered by the manufacturer as more suitable for obtaining data from large insert clones. Initially sequencing was performed using the $T3^{pWE15}$ primer only. Where sufficient quality data were not obtained, a second reaction was prepared using the $T7^{pWE15}$ primer.

Approximately 350 bases of high quality sequence data were obtained for each cosmid clone. Edited data were first searched for similarity to existing sequences using BLAST, and results are shown in table 5.1a. These data show that of the ten cosmids analysed, data for eight do not share any significant similarity with existing database sequences. End sequence from clone H166 shares 88% identity with an anonymous human genomic DNA sequence entry, over a total of 70 bases. Data for clone H045 yield the most interesting results from the database similarity search (figure 5.1c). The highest reported match is with an anonymous human cDNA (accession no. AL050223.1), with 96% identity over 78 bases, and a further region of 88% identity over 116 bases. The second most significant match reported is with part of the rat vesicle associated membrane protein (VAMP2) coding sequence (accession no. M24105). H045 shares 88% identity over a total of 93 bases with the rat VAMP2 coding sequence and a further region of 90% identity over 54 bases. In both alignments, a gap exists which indicates a region of H045 sequence that does not align with the human and rat database sequence. This may indicate untranslated sequence that is present in the dog genomic clone but not the coding sequences reported in the BLAST output. Human VAMP2 (also known as synaptobrevin

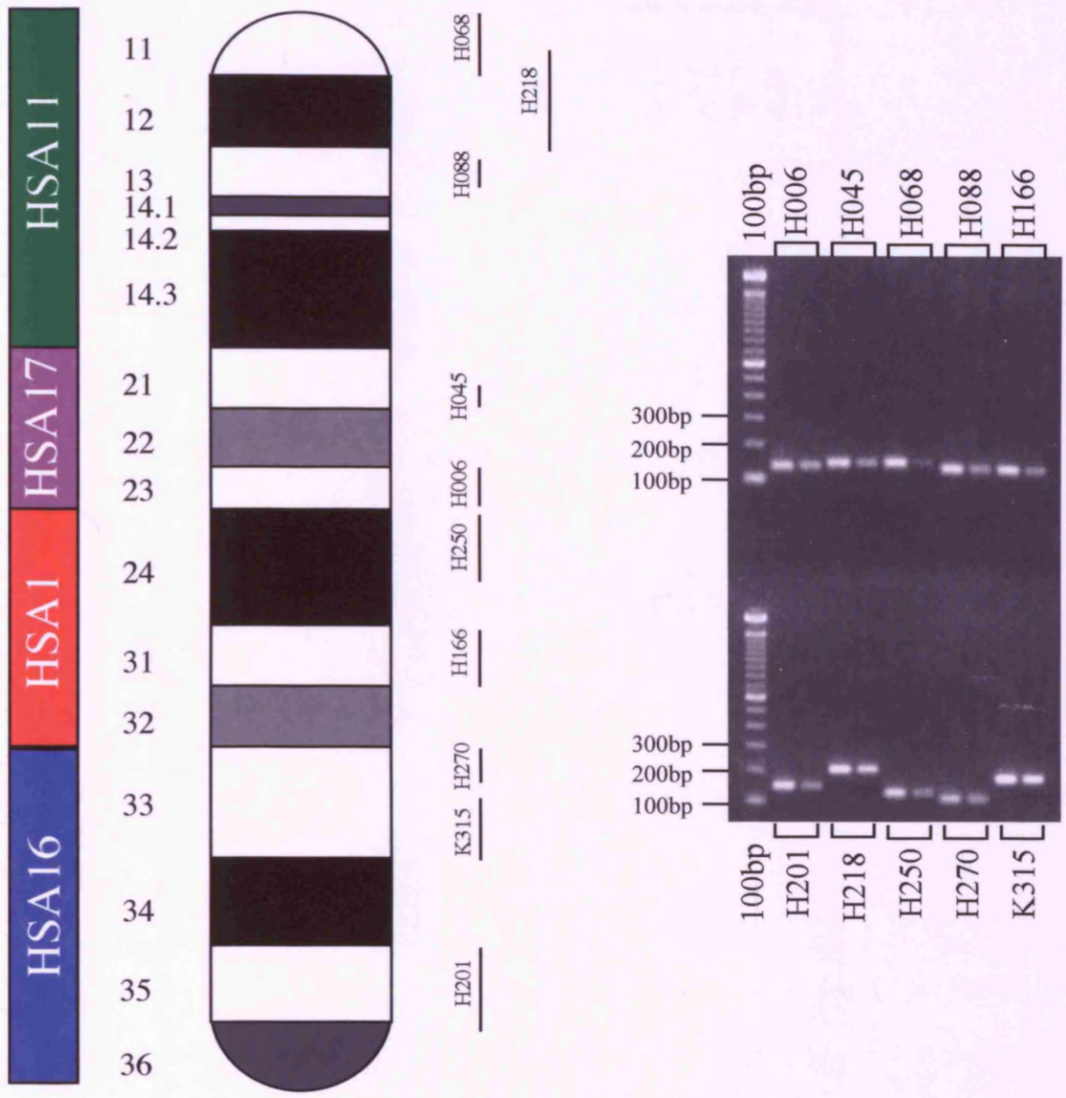


Figure 5.1a:

Ideogrammatical representation of the location of CFA 5 cosmid clones (above left). The bar to the left of the ideogram shows the location of the four ECCS on CFA 5 that correspond to regions of HSA 1, 11, 16 and 17, identified in chapter four.

Figure 5.1b:

Agarose gel analysis of type II marker products amplified from dog genomic DNA and from the corresponding cosmid clone (above right). The locus represented is shown above the relevant lane. Products derived from clone DNA are shown in the left lane of each pair, and that from dog genomic DNA is shown in the adjacent lane to the right. The size of each product was estimated relative to the 100bp DNA ladder.

Table 5.1a:

Database similarity search results for end sequence data derived from ten CFA 5 cosmid clones. The database sequence with which highest sequence identity was obtained is reported for each clone, followed by the proportion of identical nucleotides and the probability score, represented by the P-value.

Marker	Highest scoring BLAST match	Accession no.	Identity	P-value
H006	*Homo sapiens* unconventional myosin (MYO15) gene	AF051976	51/58 (87%)	$5e^{-6}$
H045	*Homo sapiens* mRNA; cDNA DKFZp586L1323	AL050223.1	75/78 (96%)	$4e^{-26}$
H068	*Aspergillus fumigatus* beta-1,3-glucanosyltransferase (BGT1) gene	AF038596.1	24/25 (96%)	0.08
H088	Human Chromosome 11 Cosmid cSRL97a6, complete sequence	U73649	42/47 (89%)	$9e^{-6}$
H166	*Homo sapiens* DNA sequence from PAC 163M9	AL021920	62/70 (88%)	$5e^{-12}$
H201	*Homo sapiens* clone NH0260M02, complete sequence	AC007038.3	21/21 (100%)	0.095
H218	*Caenorhabditis elegans* cosmid R04A9	U41550	20/22 (90%)	0.67
H250	Human DNA sequence from clone 550H1	AL035420.15	32/34 (94%)	$8e^{-5}$
H270	*Homo sapiens* DNA sequence from PAC 934G17	AL021155	20/20 (100%)	0.28
K315	*Homo sapiens* Chromosome 11q12.2 PAC clone pDJ519o13	AC004228	36/39 (92%)	$6e^{-7}$

Figure 5.1c: (opposite)

Database similarity search result for clone H045. Sequence data obtained using the $T3^{pWE15}$ primer were used to search the non-redundant sequence database using BLAST. Results demonstrate that clone H045 shares significant similarity to the rat VAMP2 gene sequence. The two most highly significant matches are shown, with the proportion of identical bases over the aligned region indicated in each case.

Query= **H045 T3** (392 letters)

Sequences producing significant alignments:	Score (bits)	E Value
emb|AL050223.1|HSM800520 Homo sapiens mRNA; cDNA DKFZp586L1...	**123**	**4e-26**
gb|M24105|RATVAMPB Rat vesicle associated membrane protein ...	**90**	**5e-16**

emb|AL050223.1|HSM800520 Homo sapiens mRNA; cDNA DKFZp586L1323 (from clone DKFZp586L1323)

Length = 1503

Score = 123 bits (62), Expect = 4e-26

Identities = 75/78 (96%), Gaps = 1/78 (1%)

Strand = Plus / Plus

```
Query: 99  ttgggaggccctgagactggccccacttggtcctaagaatcccaaggtcttctgggagct 158
           ||||||||| ||||||||||||||||||||||||||||||||||||||||||| |||||||
Sbjct: 309 ttgggaggctctgagactggccccacttggtcctaagaatcccaaggtctt-tgggagcg 367

Query: 159 tccagcatgttaattagc 176
           ||||||||||||||||||
Sbjct: 368 tccagcatgttaattagc 385
```

Score = 103 bits (52), Expect = 3e-20

Identities = 103/116 (88%), Gaps = 3/116 (2%)

Strand = Plus / Plus

```
Query: 243 aacccgtgtttctttttt-actgaggagttagtccccacg-gtccttgtatcagtctttc 300
           |||| ||||||||||||| |||||||||||||||||||   || |||||||||   ||||
Sbjct: 443 aacctgtgtttctttttttactgaggagttagtccccattagttcttgtatcacattttc 502

Query: 301 atttgcatgacattacttgcaggtggt-gggagcctgggcttttggggagccaggc 355
           ||||||| ||||||||| ||||||||| |||||||||||||||||||||| ||||||
Sbjct: 503 atttgcacgacattactcgcaggtggtggggagcctgggcttttggggaaccaggc 558
```

gb|M24105|RATVAMPB Rat vesicle associated membrane protein (VAMP-2) mRNA, complete cds. >gi|3013194|gb|I77040|I77040 Sequence 15 from patent US 5693476

Length = 2071

Score = 89.7 bits (45), Expect = 5e-16

Identities = 82/93 (88%), Gaps = 1/93 (1%)

Strand = Plus / Plus

```
Query: 99  ttgggaggccctgagactggccccacttggtcctaagaatcccaaggtcttctgggagct 158
           ||||||| | || ||||||| || |||||||||||||||||||||||  ||||||||||
Sbjct: 849 ttgggagcctctaagactggtcctacttggtcctaagaatcccaaggatttctgggagct 908

Query: 159 tccagcatgttaattagcat-tcatttacatac 190
           |||| |||||| |||||||| |||||| |||||
Sbjct: 909 tccaacatgttgattagcatatcatttgcatac 941
```

Score = 60.0 bits (30), Expect = 4e-07

Identities = 49/54 (90%), Gaps = 1/54 (1%)

Strand = Plus / Plus

```
Query: 296  ctttcatttgcatgacattacttgcaggtggtggg-agcctgggcttttgggga 348
            |||| ||||||| |||||||  ||||||||||||| |||||||||||||||||
Sbjct: 1063 cttttatttgcacgacattatctgcaggtggtggggagcctgggcttttgggga 1116
```

2, SYB2) has been assigned to HSA 17pter-p12 (Archer *et al.*, 1990), and has been implicated in familial infantile myasthenia gravis (Christodoulou *et al.*, 1997), an autosomal recessive neuromuscular disorder relating to disrupted function of acetylcholine receptors. A dog model of the disease has been described (Miller *et al.*, 1984), in which the occurrence of this lethal trait disease in the smooth fox terrier was discussed. Cosmid clone H045 has been mapped to CFA 5q21, within the region shown to be evolutionarily related to HSA 17p. These data indicate that clone H045 contains a region of sequence from the dog VAMP2 gene.

5.1.3 Primer design and PCR amplification

PCR primers were designed for each clone using criteria highly comparable to those used by Priat *et al.* (1998) in their RH mapping analysis, as described in section 2.5.2. PCR products were designed to be between 100bp and 200bp where possible. Duplicate amplification reactions were prepared for each primer pair, the former containing approximately 25ng of purified cosmid DNA, and the latter containing the same quantity of dog genomic DNA to ensure that amplification was successful from both templates. All were subjected to the PCR with an annealing temperature of 60°C, and denaturation, annealing and elongation steps of one minute each, following the procedure described in section 2.1.4.

In each case, PCR primers generated the same size product from both DNA templates following agarose gel electrophoresis and comparison with a 100bp DNA size standard, as shown in figure 5.1b. Primer sequences are listed in appendix B.

5.2 Radiation hybrid mapping of type I and type II markers

5.2.1 Optimisation of PCR conditions for RH analysis

All markers were initially assessed in an optimisation step to establish suitable conditions for the PCR. This procedure also demonstrated whether PCR primers permitted any amplification from the rodent background of the panel in addition to that from dog genomic DNA. The optimisation process is described in section 2.5.3, and examples of the resulting agarose gel analysis are shown in figure 5.2a. The cycling conditions selected for each marker are indicated in table 5.2a.

Three markers failed the optimisation step due to co-amplification of the rodent background that could not be eliminated or distinguished from the dog product. The SLC2A4^{RT} PCR primers designed in chapter four amplified a product of approximately

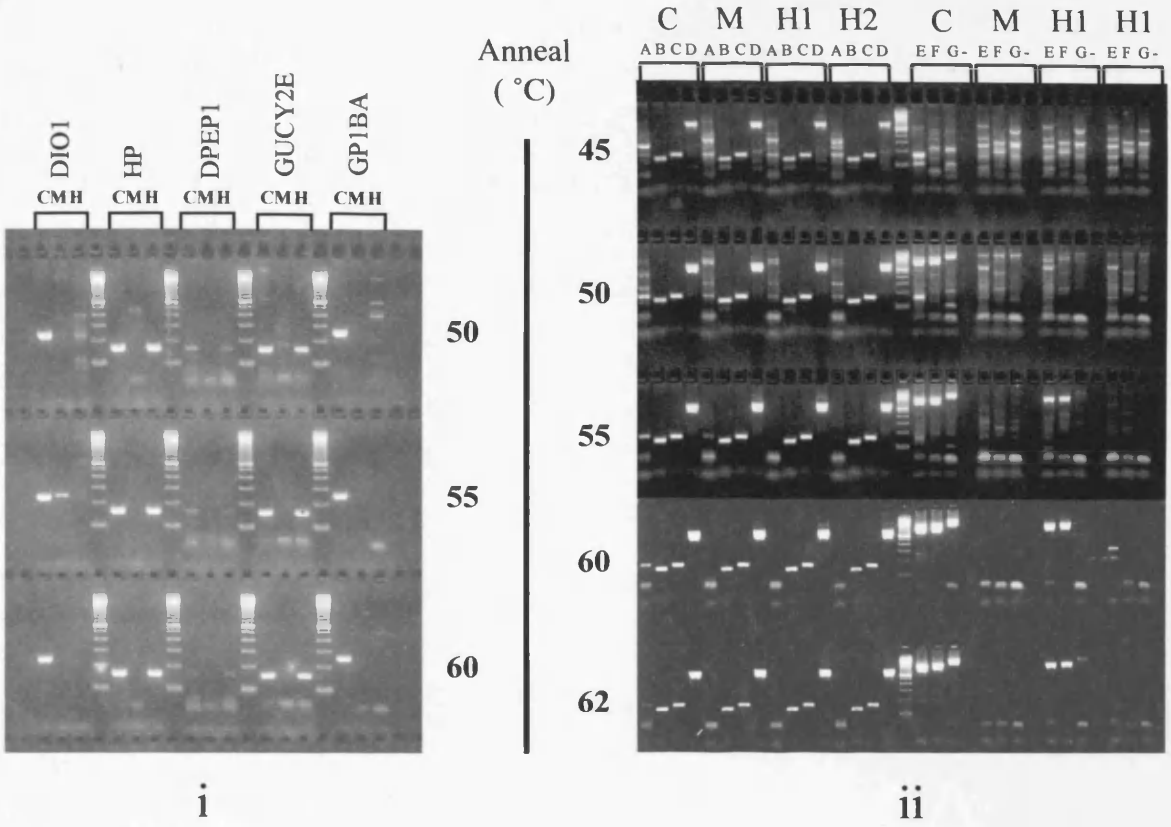


Figure 5.2a:

Optimisation of amplification conditions for selected markers. Amplification was performed using dog, hamster and radiation hybrid template DNA, under a range of annealing temperatures. The annealing temperature used is shown beside each tier of the agarose gel image. Figure 5.2ai (left) demonstrates the optimisation of five markers, for which the amplification products from dog (C), hamster (M) and randomly selected hybrid (H) DNA are shown. DIO1, HP, GUCY2E and GP1BA generated robust amplification from dog genomic DNA in the absence of any hamster template amplification. HP and GUCY2E both amplified from the hybrid DNA used in the optimisation procedure, whilst this hybrid was negative for DIO1 and GP1BA. DPEP1 could not be successfully optimised since amplification from the dog template was weak under all cycling conditions used.

Figure 5.2aii (right) shows the optimisation process for (A) DPEP1 and (B-G) five primer combinations for the SLC2A4 locus, at five different annealing temperatures: (B) F242 + R398, (C) F242 + R485, (D) F168 + R369, (E) F168 + R398 and (F) F168 + R485. Amplification for DPEP1 was improved but remained weak when subsequently tested across the entire RH panel. All SLC2A4 primer combinations generated robust amplification from dog genomic DNA, but the former three also amplified from the hamster background of the RH panel, and were eliminated from further analysis.

Table 5.2a:

Optimised PCR conditions for RH analysis of CFA 5 markers. The original publication from which primer sequences were obtained is shown in column two. This is followed by details of the optimised PCR conditions for each, indicated by the annealing temperature used (°C), and the duration of the elongation step in minutes (min). Column four shows the group to which each marker was assigned on the basis of the quality of results obtained by PCR analysis of the locus across the entire RH panel, and is described in section 5.2.2. The size of the PCR product obtained is shown in column five. Where suitable amplification conditions could not be optimised, this is indicated 'nd' (not determined).

Marker	Source	Optimised conditions	Group	Dog product size (bp)
TYPE I				
APRT	Venta *et al.* (1996)	50°C, 3 min	2	750
CD3E	Priat *et al.* (1999)	60°C, 1 min	4	170
DIO1	Priat *et al.* (1999)	64°C, 1 min	4	250
GUCY2E	A. Curson. pers. comm.	nd	1	120
GUCY2E^{RH}	this study	64°C, 1 min	4	150
SLC2A4	this study	nd	1	120
SLC2A4^{RH}	this study	62°C, 3 min	3	1100
GP1BA	Priat *et al.* (1999)	64°C, 1 min	4	250
HP	this study	64°C, 1 min	4	170
MMP1^{RT}	this study	60°C, 1 min	2	200
DPEP1	Jiang *et al.* (1998)	nd	1	150
TAT	Priat *et al.* (1999)	55°C, 1 min	4	180
THY1	Lyons et al. (1997)	nd	1	800
THY1^{RH}	this study	60°C, 1 min	2	320
TP53	Lyons et al. (1997)	60°C, 3 min	4	1200
TP53^{AHT}	M. Binns, pers. comm.	55°C, 1 min	3	350
TYPE II				
H006	this study	60°C, 1 min	4	120
H045	this study	60°C, 1 min	4	120
H068	this study	60°C, 1 min	4	140
H088	this study	60°C, 1 min	2	120
H166	this study	60°C, 1 min	3	120
H201	this study	62°C, 1 min	3	150
H218	this study	60°C, 1 min	4	200
H250	this study	60°C, 1 min	4	110
H270	this study	60°C, 1 min	4	110
K315	this study	60°C, 1 min	2	180

120bp from both dog and hamster DNA. Primers for THY1 (Lyons *et al.*, 1997) amplified a product of approximately 700bp from hamster genomic DNA, compared to the 800bp dog product. PCR primers for the GUCY2E marker described in chapter four also strongly amplified the rodent background of the RH panel, generating a product of approximately 120bp from both dog and hamster genomic DNA. In none of these cases could the rodent product be eliminated by increasing the annealing temperature used in the PCR. Since no alternative primers were available, additional pairs were designed for these three markers. The new primers are assigned the suffix '^{RH}' to identify these from the original set, and were used in all subsequent studies. These primers were successfully optimised for amplification from dog genomic DNA in the absence of rodent amplification, generating a product of the expected size.

The original primers for the GUCY2E marker (described in section 4.3.5e) amplified strongly from the rodent background of the RH panel. Partial sequence data were obtained for the hamster product, and the design of a new forward primer was investigated within a region where sequence mismatch between the dog and hamster product was apparent. The combination of this new primer and the existing reverse primer would be expected to yield a dog product of less than 100bp, at the lower limit of the preferred size for radiation hybrid mapping. Therefore, the complete gene sequence for dog GUCY2E (accession no. Y15484.1) was used to design a new reverse primer that lay within the same exon as the original product, using the PRIMER program. The new primers (GUCY2E^{RH}) were predicted to amplify a product of 160bp, which was confirmed by PCR analysis from dog genomic DNA.

The partial dog cDNA sequence for SLC2A4 (accession no. AJ388533) was imported into the PRIMER program in order to identify further suitable sites for primer design. The optimal sites reported are shown in figure 5.2b, which is taken from the PRIMER output. Each of these primers was used in a series of PCR amplifications, in several combinations, using dog and hamster genomic DNA as the template. Initially amplification was performed with an arbitrarily chosen annealing temperature of 57°C. This would enable a suitable pair to be selected that did not amplify from the recipient background of the RH panel, but could be used as a robust marker in the dog. The results of the amplification process are shown in figure 5.2c, and the primer combinations used are listed in table 5.2b. Figure 5.2c shows the co-amplification of both dog (CFA) and hamster (MAU, *Mesocricetus auratus*) genomic DNA generated with the primer pair (referred to as F^{242} and R^{369}) described in chapter four for the isolation of a dog BAC

```
                +10        +20        +30        +40
    ACCAGCATCT TCGAGACGGC AGGGGTCGGG CAGCCAGCCT ATGCGACCAT
  1 ---------- ---------- ---------- ---------- ----------
    TGGTCGTAGA AGCTCTGCCG TCCCCAGCCC GTCGGTCGGA TACGCTGGTA
               +10        +20        +30        +40
    AGGAGCTGGT GTGGTCAACA CAGTCTTCAC CTTGGTCTCG GTGTTCCTGG     F-052
 51 ---------- ---------- ---------- ---------- ----------
    TCCTCGACCA CACCAGTTGT GTCAGAAGTG GAACCAGAGC CACAAGGACC
               +10        +20        +30        +40
    TCGAGCGGGC TGGCCGTCGG ACACTCCATC TCCTGGGCCT GGCAGGAATG
101 ---------- ---------- ---------- ---------- ----------
    AGCTCGCCCG ACCGGCAGCC TGTGAGGTAG AGGACCCGGA CCGTCCTTAC
               +10        +20        +30        +40
    TGTGGCTGTG CCATCTTGAT GACCATAGCC CTGCTTCTGC TGGAGCGGCT     F-168
151 ---------- ---------- ---------- ---------- ----------
    ACACCGACAC GGTAGAACTA CTGGTATCGG GACGAAGACG ACCTCGCCGA
               +10        +20        +30        +40
    TCCAGCCATG AGCTACGTCT CCATCGTGGC CATCTTTGGC TTTGTGGCCT     F-242
201 ---------- ---------- ---------- ---------- ----------
    AGGTCGGTAC TCGATGCAGA GGTAGCACCG GTAGAAACCG AAACACCGGA     R-236
               +10        +20        +30        +40
    TCTTTGAGAT TGGCCCAGGC CCCATTCCCT GGTTCATCGT GGCCGAGCTG
251 ---------- ---------- ---------- ---------- ----------
    AGAAACTCTA ACCGGGTCCG GGGTAAGGGA CCAAGTAGCA CCGGCTCGAC
               +10        +20        +30        +40
    TTCAGCCAGG GCCCCCGCCC AGCCGCCATG GCCGTGGCTG GCTTCTGCAA
301 ---------- ---------- ---------- ---------- ----------
    AAGTCGGTCC CGGGGGCGGG TCGGCGGTAC CGGCACCGAC CGAAGACGTT     R-369
               +10        +20        +30        +40
    CTGGACAAGC AACTTCATCA TTGGCATGGG TTTCCAGTAT ATCGCGGANG     R-398
351 ---------- ---------- ---------- ---------- --------?-
    GACCTGTTCG TTGAAGTAGT AACCGTACCC AAAGGTCATA TAGCGCCTNC
               +10        +20        +30        +40
    CCATGGGGCC CTATGTCTTC CTTCTGTTCG CGGTTCTCCT GCTCGCCTTC
401 ---------- ---------- ---------- ---------- ----------
    GGTACCCCGG GATACAGAAG GAAGACAAGC GCCAAGAGGA CGAGCGGAAG
               +10        +20        +30        +40
    TTCATCTTCA CCTTCCTAAA AGTGCCTGAA ACCCGA
451 ---------- ---------- ---------- ------
    AAGTAGAAGT GGAAGGATTT TCACGGACTT TGGGCT                    R-485
```

Figure 5.2b:

Sites of primer design within the dog SLC2A4 partial cDNA sequence (accession no. AJ388533). The name of each primer is shown to the right of the alignment, the number indicating the position of the most 5' base in the primer. F denotes the 'forward primers' (red type), and R the 'reverse primers' (blue type).

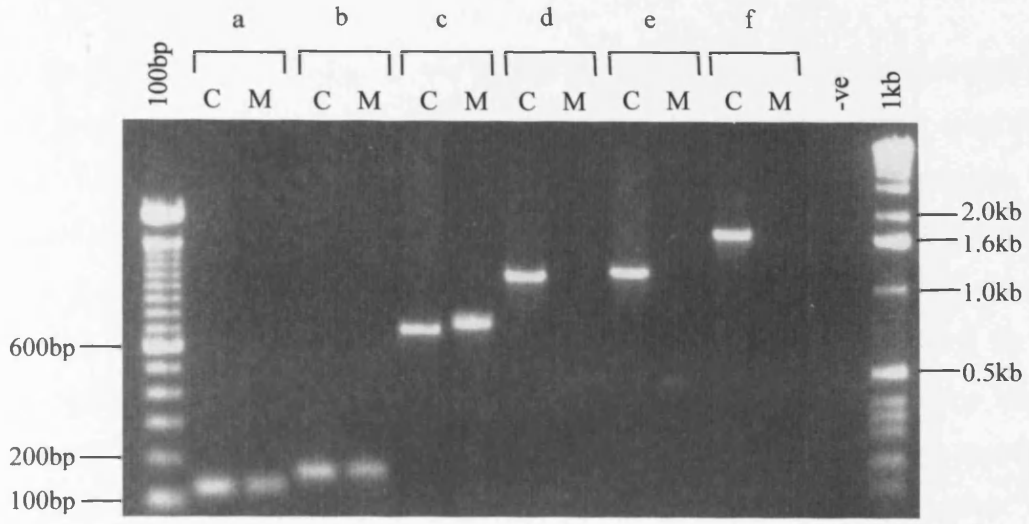


Figure 5.2c:

PCR amplification using different SLC2A4 primer combinations, on dog (C) and hamster (M) genomic DNA. The primer pairs used are as follows:

a) F242 + R369 b) F242 + R398 c) F242 + R485

d) F168 + R369 e) F168 + R398 f) F168 + R485

Table 5.2b:

Primer combinations used in amplification of the dog SLC2A4 gene. Primer sequences and locations are shown in figure 5.2b, from which the predicted distance between each primer pair is derived, shown in column three. Column four indicates whether amplification from dog (CFA) genomic DNA was successful, and whether amplification from hamster (MAU) genomic DNA was observed under the conditions used. The final column lists the size of the amplification product generated from dog genomic DNA. Differences between the values in columns three and five infer the presence of intron sequence between the two primers used. The primer pair used in RH analysis is highlighted in grey.

Primer one	Primer two	Distance from cDNA	Amplification status	Size of dog genomic product
F242	R369	127bp	CFA + / MAU +	120bp
F242	R398	156bp	CFA + / MAU +	150bp
F242	R485	243bp	CFA + / MAU +	700bp
F168	R369	201bp	CFA + / MAU -	1100bp
F168	R398	230bp	CFA + / MAU -	1200bp
F168	R485	317bp	CFA + / MAU -	1900bp

clone for SLC2A4. The design of two further reverse primers, used in combination with the original forward primer from chapter four, also demonstrated rodent amplification. Using a new forward primer in combination with the three new reverse primers did not generate any detectable rodent amplification. All combinations successfully amplified from dog genomic DNA. Table 5.2b shows that two of the amplified products were of a size comparable with that predicted from the cDNA sequence from which they were designed. The remaining four products were considerably larger, most likely indicating that the cDNA sequence spans at least one intron within the corresponding region of the dog gene sequence. This investigation indicated that the most appropriate marker combination to be used in RH analysis was F^{168} and R^{369} (hereafter referred to as SLC2A4 F^{RH} and R^{RH}), since this combination (highlighted in table 5.2b) generated the most robust and rapidly analysed PCR product from the dog, in the absence of amplification from hamster.

The THY1 marker (Lyons *et al.*, 1997) used in section 4.3.5c for the isolation of a BAC clone for this gene amplified strongly from the rodent background of the dog RH panel. This amplification could not be eliminated by optimisation of PCR conditions. The rodent product was approximately 700bp in size, marginally smaller than the dog product which measured 800bp. The THY1 primers (Lyons *et al.*, 1997) were used to amplify and obtain partial sequence data from the rodent product, and these data were aligned with the dog sequence for the same locus. Regions of mismatch were identified, at which two primers were designed that were more highly specific for the dog sequence. The resulting product was approximately 320bp in size, and could be amplified from radiation hybrid DNA without interference of a rodent PCR product. This locus is hereafter referred to as $THY1^{RH}$.

5.2.2 Application of markers to the radiation hybrid panel

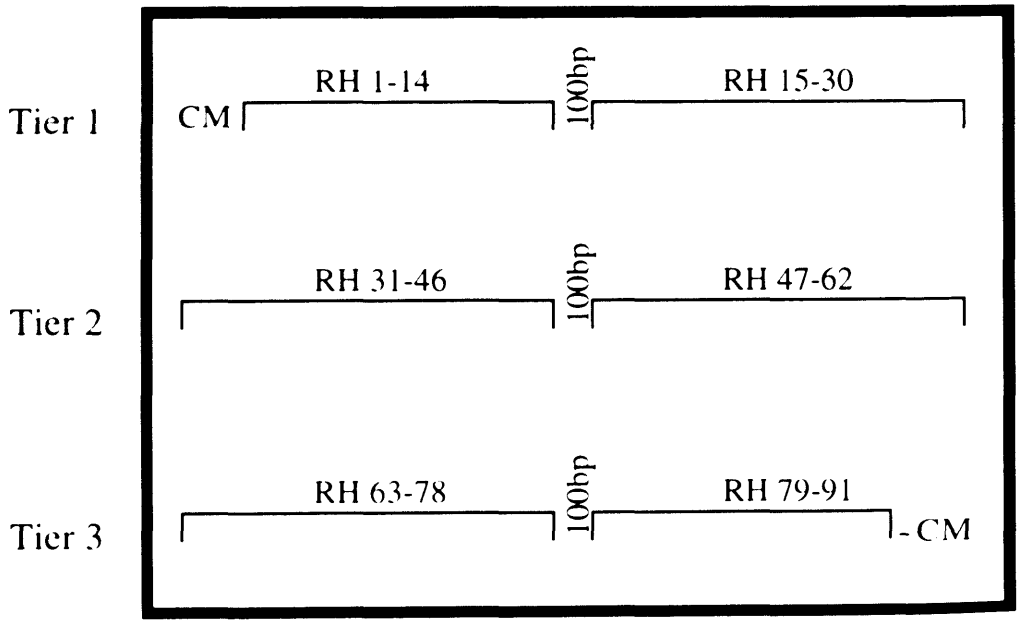
Each marker for which the optimisation step was successful was analysed across all ninety one hybrids within the dog WG-RH panel. Each marker was assigned to one of four classes (Groups 1 to 4) to reflect its potential for RH mapping, on the basis of observations during PCR optimisation and analysis across the RH panel. A summary of these observations is given below where appropriate. Examples of markers typed across the RH panel are shown in figure 5.2d.

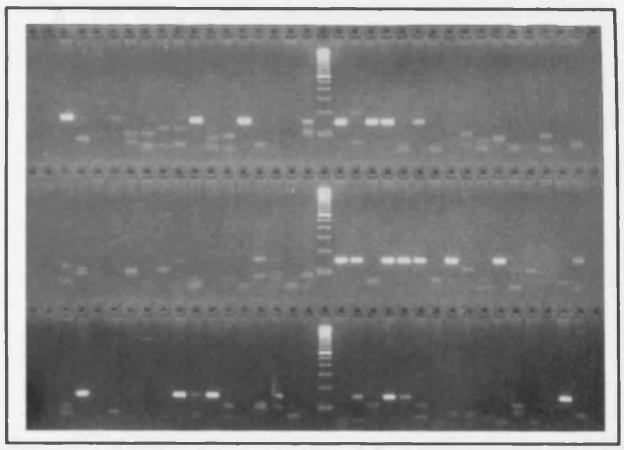
Figure 5.2d:

Analysis of selected markers across the dog RH panel (opposite). A schematic representation of the layout of the RH assay is shown below. The first two lanes of tier one in each case represent the control samples, in which amplification was performed using dog (C) and hamster (M) genomic DNA respectively. These controls are repeated in the last two lanes of the third tier, preceded by the negative control (-), in which no DNA template was included. Amplification products resulting from the application of marker primers to each hybrid constituting the RH panel are shown in numerical order (RH 1 to RH 91).

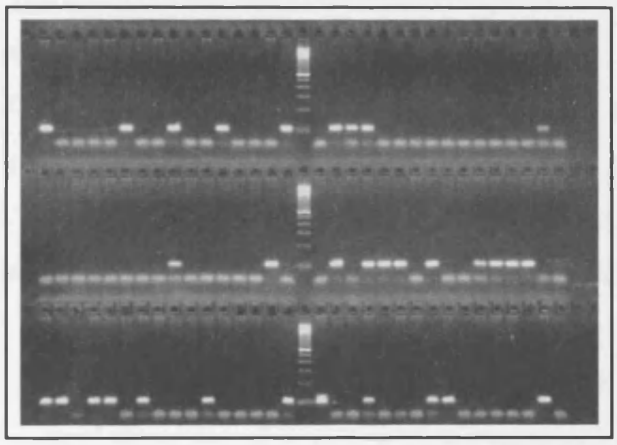
Image A (opposite) represents analysis of marker H201, for which spurious amplification from the rodent background can be observed. Consequently this marker was assigned to Group 2. Image B represents marker H270, which was the most robust of those analysed in this study, and was assigned to Group 4.

Images C and D represent analysis of the same locus (TP53) amplified using two different primer pairs. Those used in image C were taken from Lyons *et al.* (1997) whilst amplification observed in image D was generated using primers designed from an existing DNA sequence for dog TP53 (M. Binns, pers. comm). In the latter case, the dog amplification product is approximately 350bp, and can be clearly distinguished from the product generated from the rodent background template, which was of approximately 1kb in size.

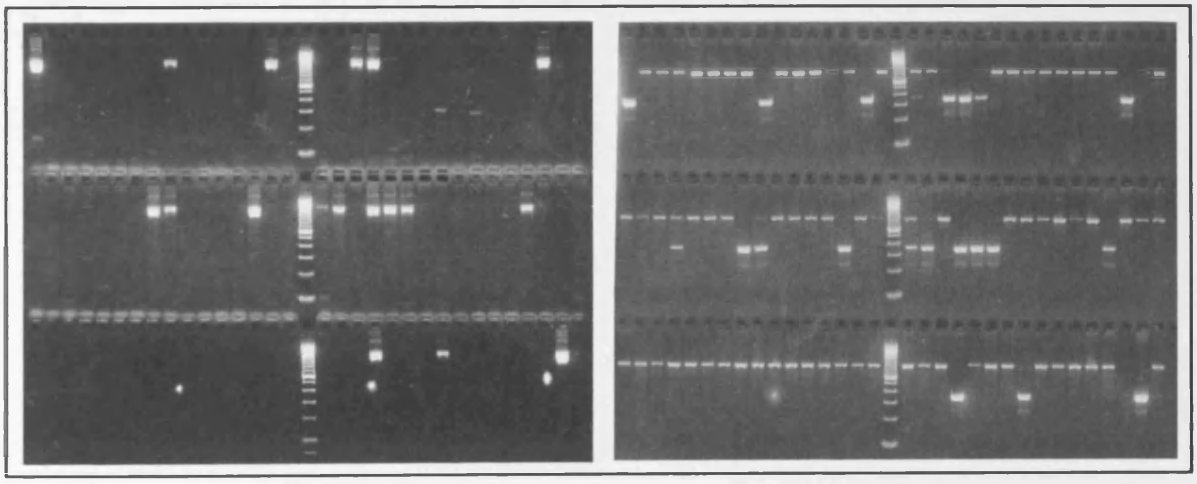




A



B



C D

- **<u>Group 1</u>: failed the optimisation step**

GUCY2E, SLC2A4, DPEP1, THY1

As described in section 5.2.1, the original marker primers for THY1, SLC2A4 and GUCY2E were rejected due to co-amplification of the rodent background that could not be eliminated or distinguished from the dog product. New primers were designed to enable these markers to be mapped by RH analysis, as described in section 5.2.1. DPEP1 gave consistently weak amplification from both dog genomic DNA and hybrid DNA, and could not be optimised. This marker was not pursued since time constraints prevented the design of additional PCR primers.

- **<u>Group 2:</u> optimisation possible with some scoring ambiguities**

APRT, MMP^{RT}, $THY1^{RH}$, H088, K315

These markers could be sufficiently optimised to allow continuation to the analysis stage. Three markers (APRT, THY1 and H088) generated moderate amplification from dog genomic DNA, and the remaining Group 2 markers demonstrated amplification of the recipient background of the RH panel that complicated scoring of data.

The product generated by the APRT marker primers was 750bp in size, and the elongation time for the PCR was therefore increased to three minutes. The product was consistently weakly amplified under all conditions investigated, but amplification from the hybrids was comparable to that from the dog genomic DNA positive control. The primers used to assay the MMP marker were those designed and described in section 4.3.5c for screening the dog BAC library. These primers (denoted $MMP1^{RT}$) were selected in preference to those of Jiang *et al.* (1998) since the former generated satisfactory results during optimisation of PCR conditions. Therefore it was deemed unnecessary to test the suitability of the alternative marker at this time.

- **<u>Group 3:</u> scoring straightforward with limited recipient amplification**

$SLC2A4^{RH}$, $TP53^{AHT}$, H166, H201

Optimisation of conditions was possible for these markers, and scoring of results from the analysis stage was rarely ambiguous. Markers were assigned to this category due to the presence of rodent background amplification that did not interfere with scoring. The $TP53^{AHT}$ marker is discussed amongst the Group 4 markers below.

- **<u>Group 4:</u> scoring straightforward with minimal or no recipient amplification**

CD3E, DIO1, GUCY2E^{RT}, GP1BA, HP, TAT, TP53, H006, H045, H068, H218, H250, H270

These represent the most robust markers, which could be optimised and fully analysed with minimal or no apparent amplification from the recipient background, and with strong amplification from dog DNA.

The TP53 locus was assayed using two primer pairs. The 1.2kb product generated by the TP53 marker primers of Lyons *et al.* (1997) used in chapter four was strongly amplified from a subset of hybrids, with no amplification from the recipient template, and this marker was thus assigned to Group 4. The TP53^{AHT} primers were designed in an earlier study from a partial DNA sequence for the dog TP53 gene (accession no. L27630) (M. Binns, pers. comm.). The primers amplify a product of approximately 350bp. This provided an opportunity to assess the RH data for analysis of the same locus but using a different STS. The TP53^{AHT} primers strongly amplified from the hamster background of the panel and the marker was therefore assigned to Group 3. Since the rodent product (approximately 1kb) was considerably larger than the dog product, both could be readily distinguished

5.2.3 Scoring of amplification patterns for each marker analysed by WG-RH

Data for each duplicate PCR assay were scored and used to identify the retention pattern for each marker. The scoring system followed the standard criteria for mapping analysis, and is summarised in table 5.2c. Presence of an amplification product from a given hybrid in both duplications of the PCR assay was scored '+'. Similarly, absence of product for a hybrid in both assays was scored '-'. A '+' score was assigned where amplification was strong in one assay but weak in the second. Weak amplification in both duplicate assays was also scored '+'. Where amplification was obtained from a given hybrid (whether strong or weak) but absent from the second duplicate, this hybrid was assigned the score '?' for that marker, indicating an ambiguity. This hybrid was not included in further analysis for the marker concerned. The number of ambiguous scores for each marker is indicated in table 5.2d.

Data for both duplicate PCR assays of each marker analysed across the entire RH panel are shown in appendix F.

Table 5.2c:

Summary of scoring criteria used in WG-RH analysis. Columns one and two show the combinations of scores possible for amplification of a given locus from a single hybrid, from the two duplicate assays (**I** and **II**). Column three shows the final score that would be assigned as a result of the amplification patterns obtained in the duplicate assays.

Assay score		Final score
Duplicate		
I	II	
+	+	+
-	-	-
+	-	?
+	weak	+
weak	weak	+
-	weak	?

Table 5.2d:

Number of ambiguous scores for each marker analysed. The proportion of the 91 hybrids within the RH panel number that were scored '?' are given as a percentage value for each marker. Markers are listed in alphabetical/numerical order. DPEP1 is marked 'nd' (not determined) since the marker could not be optimised for RH analysis and was not pursued.

Marker	Ambiguous scores	Marker	Ambiguous scores
APRT	3.3%	H006	0
CD3E	4.4%	H045	5.5%
DIO1	6.6%	H068	2.2%
DPEP1	nd	H088	7.7%
GP1BA	2.2%	H166	4.4%
GUCY2E	2.2%	H201	5.5%
HP	4.4%	H218	3.3%
MMP1	4.4%	H250	6.6%
SLC2A4	6.6%	H270	0
TAT	1.1%	K315	13.2%
THY1	7.7%		
TP53	0		

5.2.4 Calculation of the retention frequency of each locus

To derive the retention frequency for any locus, it is first necessary to establish the total number of hybrids analysed for that locus. Specifically, this represents the number of hybrids for which a score of either '+' or '-' was assigned, excluding those scored '?'. This can be expressed as the percentage of the total number of hybrids within the panel, which in this instance is 91. This is shown in table 5.2e. The retention frequency for a given marker represents the number of hybrids scored for which a product was obtained by the PCR assay. This indicates that the fragment of chromosome on which that marker lies is present within each of these hybrids. The retention frequency, which is typically expressed as a percentage of the number of hybrids scored for the marker concerned, is also shown in table 5.2e.

The total number of hybrids analysed for a given marker is in general a measure of the suitability of that locus as an RH marker. For example, H006, H270 and TP53 were scored in all 91 hybrids, and were also classed amongst the highest quality markers analysed, with absence of background amplification, and no ambiguous amplification patterns. In contrast, K315 could be analysed in only 86.8% of hybrids, due to the presence of weak amplification from the rodent background of the RH panel that resulted in a relatively high proportion of ambiguous scores (13.2%), requiring 12 hybrids to be excluded from further analysis. K315 was thus categorised into Group 2, amongst the lowest quality markers analysed. The percentage of the RH panel scored for the 21 markers analysed therefore ranged from 86.8% to 100%, with a mean of 95.7%.

The retention frequency of a given marker is determined by a number of factors, and is known to be influenced by the proximity of a marker to a centromere, which has a tendency towards an increase in retention, as discussed in chapter one. The order of markers within table 5.2e is therefore arranged to approximately reflect their physical location down CFA 5, beginning with those nearest the centromere. The highest retention frequencies of all markers analysed are for H068 and H218, which are located closest to the centromere. In both cases, retention approaches 50% (49.4% and 46.6% respectively), compared to the average of approximately 27.2%. Retention was not observed to be significantly increased near the telomere, since the most distal markers (TAT, HP and H201) demonstrate a retention frequency that is lower than the mean for all markers analysed. These data are shown graphically in figure 5.2e. The centromeric effect on retention frequency has been introduced in chapter one and is discussed further at the end of this chapter.

Table 5.2e:

Retention frequency and percentage of the RH panel scored for each marker. The percentage of the 91 hybrids in which each locus was retained is given as a percentage of the total number of hybrids analysed for the presence of the marker product. Markers are listed in approximate order of their location on CFA 5 predicted from existing FISH assignments and Zoo-FISH data, with the most proximal markers listed first.

Locus	% hybrids scored	% retention
H068	97.8	49.4
H218	96.7	46.6
H088	92.3	15.5
CD3E	95.6	31.0
THY1	92.3	31.0
MMP1	95.6	26.4
GUCY2E	97.8	19.1
TP53	100.0	19.8
SLC2A4	93.4	16.5
GP1BA	97.8	23.6
H045	94.5	23.3
H006	100.0	23.1
H250	93.4	24.7
DIO1	93.4	28.2
H166	95.6	26.4
H270	100.0	33.0
K315	86.8	39.2
APRT	96.7	23.9
TAT	98.9	23.3
HP	95.6	26.4
H201	94.5	19.8
TOTAL	**95.7**	**27.2**

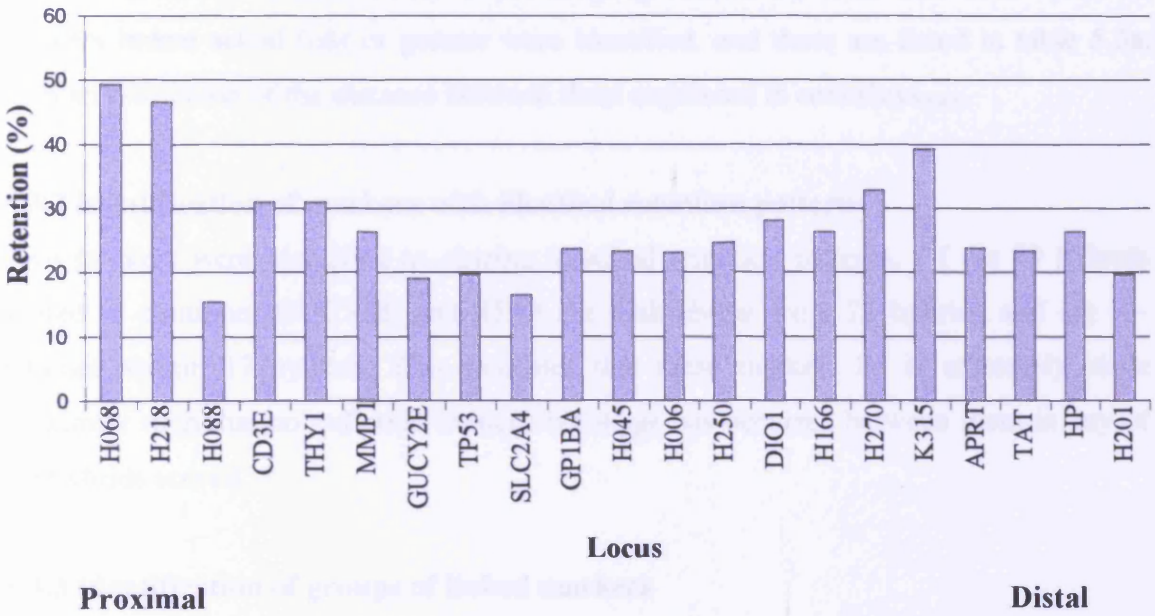


Figure 5.2e:

Graphical representation of the retention frequency of markers assayed on the RH panel. Markers are listed in order of their physical location on CFA 5, with the most centromeric markers shown nearest the origin. Retention of the chromosome fragment on which the marker lies is expressed as a percentage of the total number of hybrids assayed for the presence of that locus.

5.3 Twopoint analysis of radiation hybrid data

5.3.1 Pairwise linkage analysis

Pairs of linked markers were identified using the RH2PT component of the RHMAP program (Boehnke *et al.*, 1991), as described in section 2.5.6. A total of 28 pairs of markers linked at lod four or greater were identified, and these are listed in table 5.3a, with an estimation of the distance between them expressed in centiRays$_{3000}$.

5.3.2 Identification of markers with identical retention patterns

Two markers were identified as sharing identical retention patterns. Of the 89 hybrids scored in common, GUCY2E and TP53 are both absent from 72 hybrids and are co-retained within 17 hybrids. This indicates that these markers lie in extremely close proximity such that no radiation-induced breakage has occurred between them in any of the hybrids scored.

5.3.3 Identification of groups of linked markers

RH2PT uses lod four as the default minimum criterion for reporting groups of linked markers. The composition of each linkage group is described in table 5.3b. Six linkage groups were identified, in which each marker is linked to at least one other member of the group with a lod of four or greater. Two linkage groups comprise a single pair of markers, and two further groups contain three markers. The remaining groups contain five and six markers respectively. Therefore, each of the 21 markers analysed shows significant linkage to at least one other locus.

This information can be compared to the known physical location of each of the markers, as determined by FISH analysis (figure 5.3a). Linkage group one is comprised of two markers, H068 and H218, located at the proximal end of CFA 5, at CFA 5q11 and CFA 5q11dist.-q12 respectively. Absence of linkage to other markers is not unexpected from the observed physical distance between these and the neighbouring linkage group two, comprising CD3E, THY1 and H088, which lie in CFA 5q13. No detectable linkage was observed between H088 and either H068 or H218. It is therefore likely that the assignment of H218 can be refined to the proximal region of the assignment generated by FISH analysis, explaining the absence of linkage with the three markers in CFA 5q13.

The assignment of CD3E, THY1 and H088 to linkage group two is consistent with their physical distance from MMP1, the nearest distal neighbour, with which no significant linkage was found. MMP1 lies in linkage group three, with GUCY2E, GP1BA, SLC2A4,

Table 5.3a:

Identification of pairs of linked markers using RH2PT. Markers linked at lod four are listed in alphabetical order with respect to the first locus of the pair. The estimated distance between them follows in column three, expressed in centiRays. Lod scores demonstrating the significance of linkage are shown in column four.

Linked markers		Distance	Lod score
I	II	(cR)	
APRT	DIO1	56.2	5.02
APRT	H166	39.7	7.35
APRT	H270	17.6	13.21
APRT	K315	42.2	6.65
CD3E	THY1	25.8	10.17
CD3E	H088	40.2	6.43
DIO1	H166	37.3	7.76
DIO1	H270	54.7	5.59
GUCY2E	GP1BA	10.3	14.16
GUCY2E	MMP1	61.1	4.55
GUCY2E	SLC2A4	8.6	12.60
GUCY2E	TP53	0.0	18.85
GUCY2E	H045	6.9	15.12
GP1BA	SLC2A4	4.1	14.47
GP1BA	TP53	9.8	14.88
GP1BA	H045	9.5	14.86
HP	TAT	6.1	16.77
HP	H201	62.2	4.30
MMP1	TP53	65.4	4.24
SLC2A4	TP53	8.6	12.67
SLC2A4	H045	12.9	11.24
TAT	H201	61.4	4.49
TP53	H045	6.6	15.77
H006	H250	19.8	11.64
H068	H218	25.8	11.59
H166	H270	48.7	6.52
H166	K315	58.3	4.69
H270	K315	15.8	13.64

Table 5.3b:

Formation of groups of linked markers at lod four using RH2PT. Groups are listed in the random order generated within the analysis output of RH2PT, and markers are listed in alphabetical and numerical order within each linkage group.

Linkage group	Markers	Number of loci
1	H068, H218	2
2	CD3E, THY1, H088	3
3	GUCY2E, GP1BA, MMP1, SLC2A4, TP53, H045	6
4	H006, H250	2
5	APRT, DIO1, H166, H270, K315	5
6	HP, TAT, H201	3

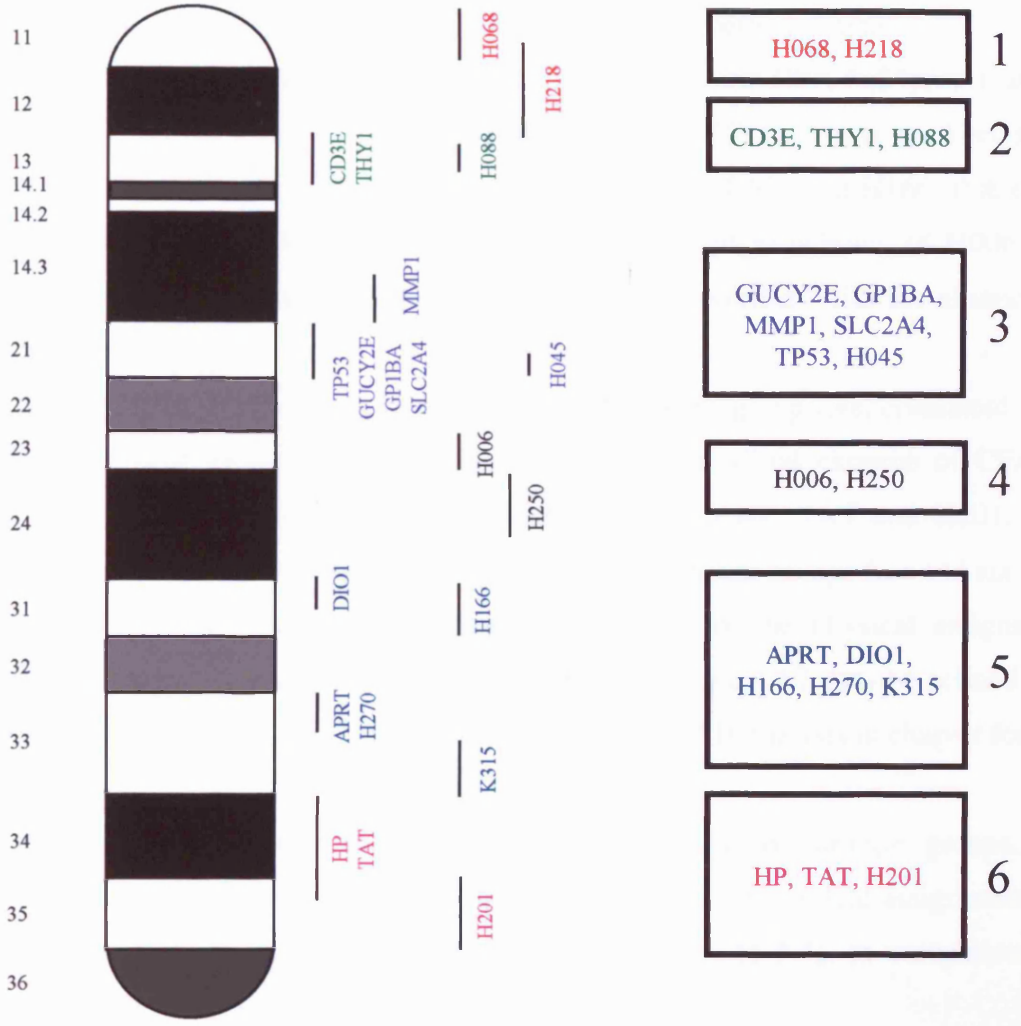


Figure 5.3a:

Comparison of FISH assignments with pairwise RH data for CFA 5 markers. Cytogenetic locations obtained from FISH analysis are shown on the left beside the CFA 5 ideogram. Data derived from RH analysis of these markers are shown on the right. The six linkage groups formed with a lod of four or greater are represented by a separate box, within which are listed the loci forming the linkage group, in alphabetical/numerical order. Locus names are colour-coded to aid interpretation, such that all loci assigned to the same linkage group are represented in the same colour.

TP53 and H045, consistent with the assignment of these markers by FISH analysis to CFA 5q21. The physical distance between these six markers and the nearest distal marker, H006, correlates with absence of significant linkage between them.

H006 and H250 lie in the mid region of CFA 5 (CFA 5q23 and CFA 5q24prox.), at the boundary between the HSA 17/CFA5 and HSA 1/CFA5 ECCS. These markers form linkage group four, and are flanked distally by two markers (DIO1 and H166) that lie in CFA 5q31. The physical distance between the nearest distal neighbours of H006 and H250, which can be approximated to in excess of 5 Mb, correlates with the absence of significant linkage between them.

APRT, DIO1, H166, H270 and K315 are assigned to linkage group five, consistent with the known physical location of these markers towards the distal extreme of CFA 5. Linkage group six lies distal to these markers, comprising HP, TAT and H201. The absence of significant linkage between the markers from linkage groups five and six may suggest that K315 lies within the most proximal region of the physical assignment proposed in section 5.1, and/or that the assignments of HP and TAT may be refined to a more distal location within the boundaries determined by FISH analysis in chapter four.

Twopoint linkage analysis of RH data therefore generated six linkage groups, the composition of which correlates with the previously generated physical assignments of these loci. These data are summarised schematically in figure 5.3a, in comparison to previously obtained cytogenetic data for each marker.

5.4 Preliminary analysis of marker order within linkage groups

5.4.1 Identification of gene order on human chromosomes

In order to compare the location of type I markers between the dog and human genomes, it was necessary to identify a detailed gene location, and potential gene order for the human loci. This information was obtained from two sources. Data derived from RH mapping of the relevant human genes was obtained from the 'GeneMap 98 database' (http://www.ncbi.nlm.nih.gov/genemap/). The physical location for each locus was identified, determined by mapping reports using the GB4 WG-RH panel (Gyapay *et al.*, 1996). Data from only three markers (DIO1, TAT and GP1BA) were available from the G3 panel (Stewart, E.A. *et al.*, 1997) and are therefore not included since they did not aid marker ordering. Data for the VAMP2 locus are also given, since this study has indicated that the anonymous dog cosmid clone H045 contains sequence associated with the dog

orthologue of this gene. The location is described in terms of the centiRay map distance of the marker from the top of the human chromosome. A proportion of markers list more than one such location, reflecting a degree of experimental error between different mapping projects. A cytogenetic assignment was obtained from 'The Unified Database' (UDB, http://bioinformatics.weizmann.ac.il/udb/), in which loci are described in terms of their megabase distance from the top of the chromosome, through the extrapolation and integration of existing RH, meiotic linkage and cytogenetic data. Since these data are continually being updated by further mapping work, they are considered in this context in terms of the relative order of markers, rather than being indicative of the physical distance between them.

The data extracted from these resources are shown in table 5.4a. Data for DIO1 are shown for completeness, since ordering does not apply in the case of this single marker from HSA 1. The three markers analysed in this study from HSA 11 can be ordered by the combination of RH data obtained from the GB4 panel, and their cytogenetic location from UDB. The data indicate that CD3E and THY1 lie in close proximity to each other, whilst MMP1 lies proximally, which correlates with the chromosomal assignments obtained from the HGND in chapter four. The four markers from HSA 16 comprise two pairs of closely located markers. HP and TAT are assigned to the same chromosomal location (HSA 16q22.1) in the HGND. These markers cannot be ordered using data from GeneMap '98 or UDB since there are no data common to both markers within either database. They are therefore listed in alphabetical order in table 5.4a. The second pair of markers, APRT and DPEP1, are also assigned to the same chromosomal location (HSA 16q24) in HGND, but RH data from the GB4 panel can be interpreted to indicate that APRT lies proximal to DPEP1. Cytogenetic assignments for TAT and APRT can in turn be used to confirm that HP and TAT lie proximal to APRT and DPEP1. This is supported by data generated by somatic cell hybrid mapping of HSA 16q (Callen *et al.*, 1992), which determined the locus order as HSA 16cen - GOT2 – ALDOA – HP/TAT – APRT – DPEP1 – qtel. The four gene markers from HSA 17 were shown to map within a small region of CFA 5 in chapter four, which correlates with their close physical proximity on this human chromosome. Data from RH mapping on the GB4 panel imply that GP1BA lies distal to VAMP2, TP53 and SLC2A4, which is supported by the cytogenetic assignments from UDB. The same data indicate that VAMP2, TP53 and SLC2A4 all lie in extremely close proximity. No RH or cytogenetic data were available for GUCY2E.

Table 5.4a:

Location and order of human orthologues of CFA 5 type I markers. Data were compiled from the GeneMap '98 and Unified Databases (see text). Where possible using these data, gene markers are listed in order of their inferred location, starting from the tip of the human chromosome. Where no data were available, this is indicated 'n/a'. Markers included in brackets are included for the purpose of further discussion (see text). The chromosome assignment of the dog PMP22 gene was taken from Werner *et al.* (1997).

Locus	Location from GB4 panel (cR3000)	Physical location from UDB (MB)	Chromosomal location from HGND	Dog chromosomal location
HSA1				
DIO1	152.12-152.63	51.42	1p33-p32	5q31 prox.
HSA11				
MMP1	351.61-352.42	112.989-113.717	11q21-q22	5q14.3
CD3E	380.67	122.833	11q22.3-q23	5q13
THY1	388.08	121.897-125.224	11q23	5q13
HSA16				
HP	427.18-434.60	n/a	16q22.1	5q34-q35prox.
TAT	n/a	82.413	16q22.1	5q34-q35prox.
APRT	489.83	94.515	16q24	5q33
DPEP1	501.78	n/a	16q24	5q33
HSA 17				
GP1BA	40.15	6.789-7.164	17pter-p12	5q21
(VAMP2)	53.53-53.73	9.047-9.080	17pter-p12	5q21
TP53	53.60-54.01	9.064-9.133	17p13.1	5q21
SLC2A4	53.8	9.097	17p13.1	5q21
GUCY2E	n/a	n/a	17p13.1	5q21
(PMP22)	70.82	n/a	17p12-p11.2	5qmid.

5.4.2 Assessment of CFA 5 RH data using RHMINBRK and RHMAXLIK

The potential to order markers within linkage groups was investigated using RHMINBRK and RHMAXLIK. The methodology used in these analysis programs was described in chapter one. RHMINBRK requires the identification of the marker order(s) requiring the smallest number of obligate chromosome breakages between adjacent loci. A break can be identified where a hybrid retains a given marker but not a second marker, such that they are not present on a single chromosome fragment in that hybrid. RHMAXLIK explores various breakage and retention models and estimates the marker order that is most likely under the model(s) tested.

Ordering of markers is performed by investigating the minimum number of obligate breaks that must have occurred between them in order to generate the retention patterns observed in the PCR assay. Consequently it is not possible to include in these analyses any markers that share an identical retention pattern. It was therefore necessary to remove one marker from the pair identified in section 5.3.2 for which identical retention was observed, namely TP53 and GUCY2E. The eliminated marker was selected at random since both can be considered as representing the same locus under both the RHMINBRK and RHMAXLIK analysis protocols. TP53 was thus excluded from further analysis.

The four linkage groups listed in table 5.3b that contain three or more markers were each analysed using RHMINBRK and RHMAXLIK. Data obtained from RHMINBRK and RHMAXLIK were compared to generate table 5.4b, which lists the highest ranked marker order for each linkage group. Data regarding the ordering of markers within linkage groups, and the relative distance between them, are presented schematically in figure 5.4a.

- **Linkage group two**

Linkage group two consists of three markers, CD3E, THY1 and H088. FISH analysis located all three loci to CFA 5q13. Analysis of RH data using RHMAXLIK identified two possible locus orders, of which the highest ranked was H008-CD3E-THY1, requiring a total of 18 obligate breaks across the entire RH panel. Seven influential hybrids are responsible for reporting this as more likely than the next best order, which requires 21 obligate breaks. RHMAXLIK analysis identifies the same most likely order, which is reported as 205.7 times more likely than the second ranked order, equivalent to a lod score of 2.31. Therefore there is significant evidence to indicate that the locus order

Table 5.4b:

Multipoint analysis of linkage groups. Each linkage group was analysed using both RHMINBRK and RHMAXLIK. The highest ranking locus order in each case is given below, with the most proximal marker listed first where this can be determined from FISH analysis. The lod score assigned to the highest ranking order under RHMAXLIK analysis is also given, demonstrating the calculated likelihood that this order is correct as opposed to the next most likely order. Data for linkage group six were analysed and compared under both the equal-retention and centromeric retention models, due to the location of the constituent markers near the telomere of CFA 5.

Linkage group	**Highest ranked marker order**	
	MINBRK	**MAXLIK**
2	H088 CD3E THY1	H088 CD3E THY1
Lod score		**2.31 (205.7 x)**

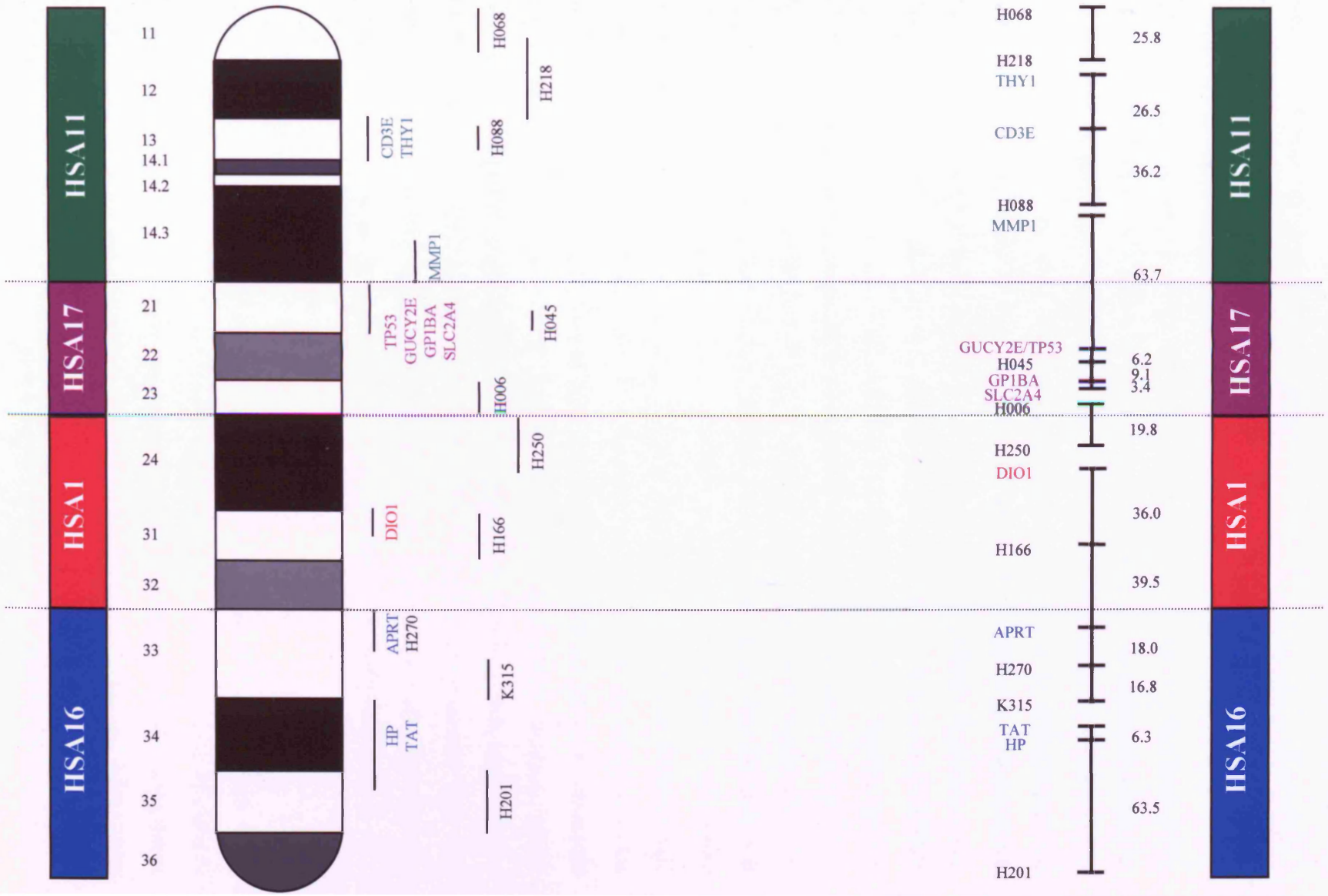
	MINBRK	**MAXLIK**
3	MMP1 TP53/GUCY2E H045 GP1BA SLC2A4	MMP1 TP53/GUCY2E H045 GP1BA SLC2A4
Lod score		**1.13 (13.3 x)**

	MINBRK	**MAXLIK**
5	DIO1 H166 APRT H270 K315	DIO1 H166 APRT H270 K315
Lod score		**1.53 (33.6 x)**

	MINBRK	**MAXLIK**	**Centromeric model**
6	HP TAT H201	TAT HP H201	TAT HP H201
Lod score		**0.36 (2.3 x)**	**0.08 (1.2 x)**

Figure 5.4a:

Multipoint analysis of RH data, and comparison with cytogenetic assignments of CFA 5 loci. The highest ranked locus order for markers within each linkage group was derived using RHMAXLIK, and is shown on the right side of the figure. Relative distances between adjacent loci generated by this analysis are shown in cR_{6000}. These data can be compared with those derived from the cytogenetic assignment of the same panel of CFA 5 loci, which are shown on the left against the CFA 5 ideogram. Type I marker names are colour coded to indicate the ECCS on which the human orthologue lies, and the physical extent of each ECCS is represented by the bars shown against both the RH and FISH data. The three boundaries between adjacent ECCS are highlighted by horizontal, broken lines, and demonstrate that each boundary is spanned by an RH linkage group.


HSA11
HSA17
HSA1
HSA16
11
12
13
14.1
14.2
14.3
21
22
23
24
31
32
33
34
35
36
CD3E
THY1
MMP1
TP53
GUCY2E
GP1BA
SLC2A4
DIO1
APRT
H270
HP
TAT
H068
H218
H088
H045
H006
H250
H166
K315
H201
H068
25.8
H218
THY1
26.5
CD3E
36.2
H088
MMP1
63.7
GUCY2E/TP53
H045
6.2
9.1
GP1BA
3.4
SLC2A4
H006
19.8
H250
DIO1
36.0
H166
39.5
APRT
18.0
H270
16.8
K315
TAT
HP
6.3
63.5
H201
HSA11
HSA17
HSA1
HSA16

within linkage group two is H008-CD3E-THY1. Since all three markers have been assigned to the same chromosomal location, the orientation of this linkage group on CFA 5 cannot be determined from multipoint analysis.

- **Linkage group three**

Linkage group three consists of six markers, GUCY2E, GP1BA, MMP1, SLC2A4, TP53 and H045. TP53 and GUCY2E share the same retention pattern, and therefore TP53 was not included in multipoint analysis, effectively reducing the number of distinct loci in this linkage group to five. Nine possible locus orders were reported by RHMINBRK, of which the highest ranked was MMP1-GUCY2E(+TP53)-H045-GP1BA-SLC2A4, requiring 21 obligate breaks in the RH panel. In all nine orders, MMP1 is placed at one boundary of the linkage group. This correlates with the chromosomal location of MMP1 to CFA 5q14.3, whilst the remaining markers in this linkage group lie distally in CFA 5q21. Thus linkage group three can be orientated with confidence. Twenty possible locus orders are reported by RHMAXLIK, of which the highest ranked is identical to that reported by RHMINBRK. This is described as 13.6 times more likely than the next best order, equivalent to a lod score of 1.13. MMP1 is a boundary marker in all twenty possible orders, correlating with cytogenetic data and those derived from RHMINBRK analysis. In common with RHMINBRK, the order of the remaining markers varies throughout the twenty possible orders reported. This is reflected by the small distances between these loci reported by RHMAXLIK, which does not exceed $10cR_{3000}$, whilst these loci are separated from MMP1 by approximately $63.7cR_{3000}$. This suggests that GP1BA, GUCY2E, SLC2A4, TP53 and H045 lie in such close proximity that they cannot be resolved by RH analysis under the conditions used in this study. Therefore the highest ranked order must be treated as tentative in the absence of additional data.

- **Linkage group five**

Linkage group five consists of five markers, APRT, DIO1, H166, H270 and K315. Only a single possible locus order is reported by RHMINBRK, namely DIO1-H166-APRT-H270-K315, requiring a total of 36 obligate breaks within the RH panel. The assignment of DIO1 to CFA 5q31, and of K315 to CFA 5q33dist., enables this linkage group to be orientated with confidence.

This same locus order is also reported as the most likely under RHMAXLIK analysis, 33.6 times more likely than the next ranked order, equivalent to a lod score of 1.53. In

common with linkage group four, the highest ranked locus order can be chromosomally orientated for linkage group five, but the internal order of markers can be predicted with less confidence. DIO1 and H166 (CFA 5q31) are consistently listed at one boundary of the locus order, whilst APRT and K315 (CFA 5q33prox. and CFA 5q33dist. respectively) are reported at the opposite boundary. H270 (CFA 5q33prox.) is consistently located centrally within these two pairs of markers in multipoint RH analysis. These data are consistent with the chromosomal location of these markers as determined by FISH analysis (figure 5.4a).

- **Linkage group six**

Linkage group six consists of three markers, HP, TAT and H201. HP and TAT have been assigned to CFA 5q34-q35prox. by FISH analysis, and H201 to CFA 5q35. The highest ranked marker order for this linkage group is HP-TAT-H201, requiring 28 obligate breaks within the RH panel.

Since these markers lie close to the CFA 5 telomere, RH data were analysed under RHMAXLIK using both the equal- and centromeric-retention models. In both cases, the same highest ranked locus order was reported, identical to that generated by RHMINBRK. This order was reported as 2.3 times and 1.2 times more likely (lod 0.36 and 0.78) than the next best order (TAT-HP-H201), under the equal- and centromeric-retention models respectively. No orders are suggested in which H201 lie between HP and TAT. The relative order of HP and TAT cannot be determined unequivocally by these data. However, these data, in combination with the cytogenetic assignments of the loci within this linkage group, suggest that H201 represents the most distal marker, and that HP and TAT are closely associated to each other at a more proximal location.

5.5 Discussion

This chapter described the radiation hybrid mapping of type I and type II markers previously assigned to CFA 5 by FISH analysis. Amongst the first considerations in generating a preliminary radiation hybrid map of CFA 5 regarded the selection of the RH panel to be used. Two panels are currently available for the dog (discussed in section 1.5.5). That of McCarthy and colleagues (unpublished) available through Research Genetics Inc. (Huntsville, AL) was generated using 3000 rads of X-irradiation, whilst that described by Vignaux *et al.* (1999a) was produced using 5000 rads of gamma rays. The quantity of irradiation used in the fragmentation procedure determines the average size of the resulting chromosome fragments, and in turn the overall mapping resolution possible using that panel. The use of the lower resolution panel of McCarthy and colleagues was considered most appropriate for use in this study. Few markers had previously been assigned to CFA 5, and the number of markers available for analysis in this study was relatively small. Since RH panels generated with a low radiation dose result in fewer chromosome breakages, the average size of a retained fragment must be greater than for higher resolution panels. Consequently, the conservation of linked marker pairs on a single chromosome fragment is more likely to be observed using a lower resolution panel. The dog RH panel of McCarthy and colleagues was generated using the same value of X-irradiation as that used in the production of the GB4 human RH panel (Gyapay *et al.*, 1996). GB4 is considered to contain human genomic DNA fragments of approximately 10 Mb on average (Schuler *et al.*, 1996), with the potential to order markers at a resolution level of 1 Mb (Stewart, E.A. *et al.*, 1997). Markers separated by between one and eight Mb have been successfully mapped using the GB4 panel (for example, Hudson *et al.*, 1995). It is likely that some of the markers analysed in this study are separated by physical distances significantly in excess of this figure, in particular the distance between CD3E/THY1/H088 and MMP1 appears to span approximately 10% of the proximal region of CFA 5, equating to approximately 10Mb. A less significant gap in marker coverage exists between H250 and DIO1/H166, in the midregion of CFA 5. Significantly, it has been shown that the panel described by Vignaux *et al.* (1999a) does not contain the TP53 locus (Mellersh *et al.*, 2000). This would imply that surrounding markers will also be absent from this panel, which would be a major limitation for the analysis of CFA 5 loci. It was therefore deemed more appropriate to use the RH panel generated by McCarthy and colleagues, in order to generate a preliminary but meaningful map of a single chromosome.

A total of 12 type I and 10 type II markers were assessed within the initial optimisation stage of RH analysis. Smaller PCR products were found to be more readily assayed, since their size could be more accurately determined from agarose gel electrophoresis, enabling distinction from rodent background amplification if necessary. Smaller products were also more rapidly analysed by electrophoresis, and also tended to amplify more robustly. The average proportion of the WG-RH panel analysed for the 21 markers assayed was 95.7%, and the mean retention was 27.2%. This compares well with previously obtained retention frequency data for this panel (24.5%, L. McCarthy, unpublished), as well as from other, similar, WG-RH panels. For example, the GB4 panel (Gyapay *et al.*, 1996) was shown to have an overall retention of 29%, and was generated using 3000rads of irradiation, as was the panel used in the present study. It has been shown that retention frequencies have a tendency to be higher at centromeres (for example, Benham *et al.*, 1989; Ceccherini *et al.*, 1992; Walter *et al.*, 1994, also see chapter one). This observation is supported in this study by the markedly higher than average retention of H088 and H218, both present in almost 50% of the 91 hybrids that form the dog RH panel, compared to an average of 27.2%. A similar effect at telomeric regions has also been observed, but is less well documented (for example, Ceccherini *et al.*, 1992). Telomeric retention was not shown to be significantly increased for the most distal markers (HP, TAT and H201) analysed in this study, since their retention was less than the mean for all markers assayed (table 5.2e and figure 5.2e.) FISH analysis of these three distal markers indicated that their distance from the telomere was greater than the distance of H088 and H218 from the centromere. It is therefore likely that HP, TAT and H201 lie sufficiently far from the telomere that their retention follows the equal retention model applied to the remaining markers assayed in this study. RHMAXLIK analysis of linkage group six, in which HP, TAT and H201 lie, supports this hypothesis, since the use of the equal- and centromeric-retention models had negligible effect on the results of this analysis.

Of the type I markers, a single marker (DPEP1) could not be optimised due to consistently weak amplification from hybrid DNA. It is possible that amplification of the DPEP1 locus under the PCR conditions used in chapter four would enable its analysis across the RH panel, and the design of new primers would be an alternative approach to enable the marker to be included in this study. However, it was necessary to eliminate DPEP1 from further study due to time limitations. Three further markers (GUCY2E, SLC2A4 and THY1) failed the optimisation step since both amplified strongly from the

rodent background of the RH panel. In each case, an additional primer pair was designed, using a combination of existing sequence data for dog and rodent, and that derived from the amplification products generated during the optimisation procedure. The new primer pair successfully enabled these markers to be assayed across the RH panel. The remaining markers, derived mainly from existing publications, were successfully assayed using the marker primers discussed in chapter four. The presence of the TP53 locus in the RH panel of McCarthy and colleagues has been confirmed in this study. The reason for the absence of this locus in the panel developed by Vignaux *et al.* (1999a) remains unclear. In this study, TP53 was assayed using two different marker primer pairs. The retention patterns generated by a single assay of the $TP53^{AHT}$ marker across the entire RH panel were in complete agreement with duplicate assay data for the TP53 marker of Lyons *et al.* (1997) described in group 4.

All 10 type II markers were successfully analysed using the RH panel. It is perhaps to be expected that type II markers would be more suited to RH analysis, since these anonymous sequences are less likely to be conserved between species. Thus the degree of amplification of the rodent background would be expected to be less than for markers containing coding sequences. Although none of the type II markers failed optimisation, they are spread throughout the remaining classifications (groups 2 to 4), indicating that amplification from the rodent recipient was also encountered using anonymous markers, and was not an issue limited to coding loci. Indeed, six of the ten type I markers fall within group 4, those shown to be most robust in RH analysis. Type I markers have the additional advantage that sequence data from a range of species are more likely to be available for previously annotated genes, which can be used to aid the selection of suitable regions for primer design. Therefore, in the context of this study, conservation of sequence data at coding regions has not proved to be a limiting factor in successful RH analysis.

Duplicate assays for each marker across the WG-RH panel demonstrated that a subset of hybrids generated weak amplification of marker products in both assays. This was the case for 16 of the 21 markers analysed. Amongst these markers, the number of hybrids for which weak amplification was consistent ranged from one (1.1% of the panel) to a maximum of four (4.4%). Variation in the intensity of the PCR product generated from each hybrid in a somatic cell hybrid panel from application of the same primer pair has previously been reported (Lahbib-Mansais *et al.*, 1999). This was attributed to variation in the relative quantity of the fragment of interest retained within each hybrid cell line

from which a DNA template was made in the generation of the panel. It is highly likely that the same situation applies to whole genome radiation hybrid panels such as that used in this study. Weak amplification could lead to assignment of false negative scores for a given marker on such a hybrid, such that consistency in the scoring of such data is paramount between markers. Weak amplification from a subset of hybrids was most noticeable for H068, THY1, and H088, for which four hybrids yielded consistently weak amplification. Under the scoring system used, any hybrid that generated weak amplification for a given marker in both duplicate assays was scored '+'. This follows the most commonly encountered scoring system, which assumes that consistently weak amplification is an artefactual phenomenon resulting from the cell-line culturing process. The instability of hybrid cell lines is considered to result in the loss of DNA fragments at random at a low level (L. McCarthy, pers. comm.). If this occurs in the early stages of the culturing process, continued replication of cells from which a fragment has been lost will result in a degree of clonal variation within the hybrid cell-line. With each passage, it is therefore possible that the distribution of a given chromosome fragment within a hybrid cell-line shifts as each distinct clone population demonstrates a different cycling time (Matise *et al.*, 1999). The relative concentration of the fragment of interest present in the template DNA prepared from the final cell population will therefore be lower than that within hybrids for which the fragment has been stably retained. The yield of the marker product for a locus present on that fragment is therefore likely to be less prolific from hybrids where the relative quantity of starting template is lower, resulting in consistently weak amplification. This is typically scored '+' if amplification is weak under both duplicate assays and under such circumstances it is particularly essential that scoring is performed in a consistent manner. Hybrids for which consistently weak amplification was observed for a given marker are highlighted in appendix F, which contains all the raw RH data generated in this study. Since markers are listed in descending order of their chromosomal location, it therefore becomes possible to identify hybrids that appear to demonstrate the clonal variation phenomenon. Where weak amplification is observed for a given hybrid when assayed for the presence of two or more closely located markers, this suggests that the chromosome fragment on which those markers lie has been lost from the cell-line during culturing. For example, hybrid 19 weakly amplified markers TP53, H045 and GP1BA, all of which lie in CFA 5q21. Hybrid 15 weakly amplified the latter two markers only. Similarly, hybrid 86 weakly amplified THY1 and CD3E, whilst hybrid 90 weakly amplified CD3E and H088, all of which lie in CFA 5q13. This

phenomenon may therefore also prove useful in establishing and refining the relative order of closely located markers. However, consistently weak amplification was more commonly limited to single loci, which may reflect the fact that the chromosome fragment lost in each case is smaller than the distance between adjacent markers. The identification of consistently weak amplification from a subset of hybrids within a WG-RH panel has not been widely reported in previous literature, but has been previously observed in panels generated for a range of different species (L. McCarthy, pers. comm.). A proportion of hybrids were scored '?' to indicate that retention patterns were not in agreement under duplicate assays. The number of hybrids for which ambiguity existed ranged from zero (H006, TP53 and H270) to a maximum of 12 (K315). Ambiguity resulted from the identification of hybrids for which an amplification product (regardless of intensity) was obtained in one of the duplicate assays, but not in the second. A hybrid was also scored as ambiguous if the presence of rodent recipient amplification masked the appearance of the donor product, such that it was not possible to unequivocally determine whether the observed band was of donor or recipient origin. Finally, ambiguity existed where the marker product was obscured due to technical reasons, such as an imperfection in the agarose gel, which also prevented unequivocal scoring for that hybrid under duplicate assays. As with the aforementioned issue, it is essential that the scoring system used is consistently applied in order to maintain the validity of the resulting data. This is particularly pertinent in studies such as this, where the aim is to produce a panel of markers for which RH mapping data can be integrated with existing data to generate a robust framework of loci that can be used for orientation in more extensive studies. A degree of caution is therefore appropriate at this stage, since error in interpretation of tentative data could have a significant effect on future placement of additional markers. Since in this study, all ambiguous hybrids were excluded from further analysis, they will have had no effect on the linkage data obtained. Their exclusion will however have limited the overall number of informative 'breakages' analysed between markers within the hybrid panel. This effect is highly analogous to the need for a large reference family pedigree for meiotic linkage analysis, in which each individual increases the number of informative meioses available to follow inheritance of marker alleles through subsequent generations. Where lod scores are at the limits of significance, this may have a considerable effect on the composition and relative ordering of the resulting linkage groups. In all cases, ambiguities could potentially be resolved by retyping individual hybrids.

The identification of a small number of ambiguous scores in this study was possible because all PCR assays were performed in duplicate. This follows the conventional approach to RH mapping analysis, since it is imperative that anomalies are identified prior to map construction. The nature of the PCR assay is such that a low level of discrepancy is inevitable. The large scale on which RH analysis is performed, where several hundred amplification reactions may be performed daily, implies that a small degree of practical error is likely to be incurred. Evaporation of individual samples during thermal cycling is a less significant issue since this is easily detected, particularly if a pigmented reagent is included in the reaction mixture as in this study. Total failure of the cycling procedure can also easily be detected by the absence of amplification from the positive control, whether this reflects error in the composition of the reaction mix, or a fault in the cycling apparatus. However, if the latter affects only part of a heating block in a thermal cycler, this effect may not be as easily determined, depending on whether the positive control is amongst those individual reactions affected. The inclusion of more than one positive control distributed across a set of PCR assays is a useful safeguard to enable cycling failure to be identified, although the capacity of the cycler and the overall size of the hybrid panel is a limiting factor. Errors are perhaps most likely to occur during the aliquoting of reaction reagents prior to amplification. Since all reagents except the template DNA are added as a master mix, this effect has generally been minimised in this study. The most likely source of error is in the addition of the template DNA itself. The scale of the RH analysis procedure is such that some degree of automation of the process is common. The use of robotic pipettors is a great advantage in terms of efficiency, but it is advisable that each microtitre plate is manually checked for the addition of template to each well prior to the addition of the master mix containing the remaining ingredients. The duplication of all PCR assays remains essential since this is by far the most satisfactory way to identify discrepancies. Ideally, hybrids for which anomalies are detected would be retyped individually, although this would significantly reduce the efficiency of the analysis procedure, which is one of the main attractions of RH analysis.

It is likely that the retyping of certain markers for which ambiguous results were obtained may aid the ordering of markers within linkage groups, although multipoint analysis of RH data correlated well with marker orders determined by FISH analysis. Figure 5.4a demonstrates that each of the three boundaries between adjacent ECCS on CFA 5 has been spanned by an RH linkage group. The density of these linkage groups could rapidly be increased by the progressive selection and mapping of novel type I markers for CFA 5

whose orthologue lies within the appropriate human chromosomal region delineated by Zoo-FISH and WG-RH mapping. The combination of FISH data and RH data therefore provides a logical method by which the junctions between ECCS may be further characterised.

RH multipoint analysis strongly indicates that the locus order within linkage group two is H008-CD3E-THY1. FISH analysis located all three loci to CFA 5q13, hence the chromosomal orientation of this linkage group cannot be determined from multipoint analysis. This correlates with the human RH data for CD3E and THY1, since they lie within an $8cR_{3000}$ region of HSA 11q when analysed using the GB4 panel (table 5.4a), and could not be resolved cytogenetically.

Linkage group three could be orientated since the confident location of MMP1 at one boundary of the linkage group correlated with its chromosomal location at CFA 5q14.3, whilst the remaining members of this group lie within CFA 5q21. Data from chapter four and from Werner *et al.* (1997) have indicated that the CFA 5/HSA 17 ECCS lie in opposing orientation with respect to the centromere of the respective chromosomes. This would imply that GP1BA would represent the most proximal marker mapped to CFA 5 from those within the CFA 5/HSA 17 ECCS, since GP1BA is the most distal marker from HSA 17p that was analysed in this study (table 5.4a). In turn, this would imply that GP1BA lies closer to MMP1 than the remaining markers from this linkage group. However, the highest ranked locus order reported by both RHMINBRK and RHMAXLIK places TP53 and GUCY2E adjacent to MMP1. This may indicate a gene order rearrangement within the CFA 5/HSA 17 ECCS. However, multipoint analysis demonstrated that ordering of GUCY2E, GP1BA, TP54, SLC2A4 and H045 could not be determined with a significant lod score. This may reflect the fact that the close proximity of these loci lies below the resolution power of the RH panel used in this study. Table 5.4a demonstrates that the human orthologues of GP1BA, VAMP2 (corresponding to CFA cosmid H045), TP53 and SLC2A4 lie within a $15cR_{3000}$ region of the HSA 17 RH map generated using the GB4 WG-RH panel. Of these, the latter three markers could not be ordered, since they lie within a region estimated as $53.5cR_{3000}$ to $54cR_{3000}$ from the HSA 17p telomere. No data were available for the human GUCY2E locus; however, the dog marker for this gene and for TP53 generated identical retention patterns in this study and were therefore required to be considered as representing the same locus in multipoint analysis. The GB4 panel was generated using 3000 rads of irradiation, equivalent to that used in the production of the dog WG-RH panel by McCarthy and colleagues

(unpublished). On the assumption that the dog and human orthologues of these genes lie within chromosomal regions of comparable size, the absence of unequivocal ordering data for markers within linkage group three correlates with data from the corresponding region of the human genome.

RHMINBRK reported only a single locus order for linkage group five, namely DIO1-H166-APRT-H270-K315, which is supported by the highest ranked order from RHMAXLIK. The assignment of DIO1 to CFA 5q31, and of K315 to CFA 5q33dist., enables this linkage group to be orientated with confidence, although internal markers can be ordered with less confidence. DIO1 and H166 (CFA 5q31) are consistently placed at one boundary of the locus order, whilst APRT and K315 (CFA 5q33prox. and CFA 5q33 dist. respectively) are reported at the opposite boundary. H270 (CFA 5q33prox.) is consistently located between at least one member of each these two pairs of markers. These data are consistent with the chromosomal location of these markers as determined by FISH analysis (figure 5.4a). Table 5.4a shows that both HP and TAT are closely located on HSA 16q22.1, whilst both APRT and DPEP1 lie approximately 12Mb distally at HSA 16q24. The order of these human loci is therefore HSA 16qcen. – HP/TAT – APRT/DPEP1 – HSA 16qtel. FISH analysis of the dog orthologues demonstrated that this relative order is reversed, such that the marker order is CFA 5qcen. – DIO1 – APRT/DPEP1 - HP/TAT – CFA 5qtel. This order is supported by the RH data generated in this chapter (with the exception that DPEP1 could not be included in this analysis). These data therefore correlate with the relative orientation of these type I markers on CFA 5 compared to the human orthologues, as discussed in chapter four. FISH analysis discussed in chapter four indicates that APRT and H270 lie in very close proximity within CFA 5q33prox., and were not resolved by standard metaphase FISH analysis. K315 was assigned to a marginally more distal location, at CFA 5q33dist. Although there is no evidence to suggest that the cytogenetic assignments of these markers are inaccurate, figure 5.4a demonstrates a region of overlap in the chromosome assignments of these markers, and therefore higher resolution FISH analysis would be required to establish an unequivocal locus order within this region. Elimination of ambiguous scores for these markers under RH analysis, increasing the number of informative hybrids that can be analysed, may also clarify this issue.

An identical marker order is reported under both RHMINBRK and RHMAXLIK analysis for linkage group six, namely HP-TAT-H201. As discussed earlier, loci within this linkage group did not demonstrate increased retention frequency despite their sub-

telomeric location. This correlates with the greater physical distance between these markers and the telomere, compared to the distance between H068 and H218 and the centromere, for which increased retention was observed. HP and TAT have been assigned to CFA 5q34-q35prox., and H201 to CFA 5q35, suggesting that the HP and TAT represent the proximal markers in this linkage group. The orientation of these two type I markers with respect to the telomere cannot be determined unequivocally, since both RH data and FISH analysis demonstrate that they lie in close proximity to each other.

Locus orders are generally considered significant where the highest ranking order is at least 100 times more likely than the next order, corresponding to a lod score of two. Under certain circumstances, particularly where the density of markers is high, a lod score of three is used (for example, Cox *et al.*, 1990; Walter *et al.*, 1994; Gyapay *et al.*, 1996) as is typically the case in multipoint analysis of linked markers in meiotic linkage analysis. Other reports use less stringent criteria, and it has been remarked that locus orders with odds of 10:1 (lod one) should not be discounted (McCarthy and Soderland, 1998). Markers located in close proximity, which by their nature are likely to be separated by only a small number of obligate breakages, may generate several possible orders, with highly similar likelihoods. Conversely, markers at each end of an ordered linkage group that are themselves separated by large physical distances are less easily ordered than the more closely located internal markers (McCarthy and Soderland, 1998). An increase in marker density is therefore required for ordering to be improved. In the mapping of 39 markers to a subregion of HSA 17q, Foster *et al.* (1996) reported orders that were 15 times more likely than the next possible order, although other linkage groups were ordered with significantly higher scores. The order of selected markers within a linkage group was determined by comparison of the likelihood of this order compared with that ranked next highest, for which the order of this subset of markers differed. This subset of HSA 17q markers could therefore be ordered independently of other markers within the linkage group. It was noted by Foster *et al.* (1996) that in several instances many possible orders existed that were almost equally as likely. As a consequence, ordering at lod three was determined as necessary for the construction of a framework map, around which the remaining markers could be placed with lower levels of significance. Foster *et al.* (1996) also report the identification of certain markers that demonstrated a retention frequency significantly lower than that for other markers,

suggesting that these markers may exist within chromosome regions that are somehow selected against during RH panel construction.

Although locus orders approaching or exceeding a value of lod two were identified in this study, the absence of highly significant data for ordering of all loci was not unexpected. This is in part due to the limited number of markers analysed, spanning a physical region of up to 100Mb. Markers were selected with care to ensure their even distribution where possible. However, their density is not consistent along the entire length of CFA 5, with potential gaps of in excess of 10Mb between adjacent markers. In contrast, one previous study in which cytogenetic and RH maps for a single chromosome were integrated involved the RH mapping of over 500 markers (McPherson *et al.*, 1997), with an average of 350kb between loci. Although human chromosome mapping projects are inevitably performed on a far greater scale than is currently possible for species such as the dog, it is clear that relatively few RH projects are performed at the whole, single chromosome level, other than those forming part of the international human genome mapping project itself. It is more common either for mapping to be performed by saturation of an entire genome with thousands of markers, for which prior chromosomal assignments are limited (such as Gyapay *et al.*, 1996; Stewart, E.A. *et al.*, 1997), or for more focussed studies on specific subchromosomal regions of interest (such as Gorski *et al.*, 1992; Foster *et al.*, 1996). The approach used in the present study must therefore be considered as a developmental process for both techniques and markers that will form the basis for a detailed map of CFA 5 as more markers become available. Similarly, until the density of mapped markers for the dog is increased, the estimation of physical distance from RH mapping data is not simply achieved. Retention frequencies are not consistent either inter- or intrachromosomally (for example, Jones, 1997). Although the degree of variation has been proposed to have a limited effect on the ordering of markers (Boehnke *et al.*, 1991), this factor reduces the ability to directly extrapolate RH data between different chromosomes and subchromosomal regions. Gyapay *et al.* (1996) showed that using the GB4 panel, the relationship between physical distance and centiRays ranged between 110kb/cR_{3000} for HSA 19 to 305kb/cR_{3000} for HSA 9. In comparison, for a panel created by 8000 rads of radiation, 1 cR_{8000} was estimated as corresponding to approximately 50kb (Cox *et al.*, 1990). It is therefore of limited value to attempt to derive an estimation of the physical distance between sets of markers within a genome in this manner, and as such RH data are more commonly used simply to compare relative rather than absolute distances. The fact that all markers mapped in this study of CFA 5 have

also been assigned by FISH analysis enables correlation between cytogenetic and RH maps to be performed with little difficulty. In addition, the knowledge of the structure of CFA 5, with respect to its counterparts in the human genome, will be of great value in the selection of additional markers for the improvement of the developing RH map of this dog chromosome. The remaining 18 cosmid clones that have been mapped to CFA 5 (M. Breen, pers. comm.) represent a valuable source of markers for expansion of the CFA 5 radiation hybrid map. Those analysed in this study were selected on the basis of their even distribution along the chromosome, since these were most likely to generate a set of useful framework markers that could be used for integration of additional loci, with the limited quantity of hybrid DNA available. Cosmid clones that have been cytogenetically assigned to chromosomal locations where marker density is poor represent those most likely to enable the joining of two or more smaller linkage groups into a single larger group, towards the development of a single linkage group spanning the entire chromosome. The isolation of the polymorphic repeat sequence from each cosmid clone, and their subsequent placement on the meiotic linkage map, will further the process of integration of data from these complementary approaches.

Yang and Womack (1998) have applied a similar approach to mapping the bovine genome, using WG-RH mapping analysis to obtain detailed information on gene order between ECCS on HSA 17 and BTA 19, identified by previous gene mapping and Zoo-FISH analyses. Previous studies had shown both chromosomes to correspond to a single ECCS (Yang and Womack, 1995, and others). To refine this information, Yang and Womack (1998) analysed 17 gene markers and 12 microsatellite-containing markers on a bovine WG-RH panel, to generate a radiation–hybrid map of BTA 19. The data were combined with those for seven previously mapped type I loci, and compared with mapping data for the human homologues of the gene markers studied, most of which had been assigned to HSA 17 by FISH analysis. The limited resolution of the FISH data was reported to have created some ambiguity in their comparative data, since assignments could not be made for human genes with as much regional information as for the cattle homologue. Consequently, assignments were also refined by mapping the human homologues using a human WG-RH panel. A comparison of gene order between these chromosomes, derived from this 'parallel radiation hybrid mapping' technique, demonstrated the existence of three putative evolutionary breakpoints that have caused changes in relative gene order, separating the two chromosomes into four comparable regions. Two such regions showed that genes lie in opposing orientations with respect to

the centromere of each chromosome, and internal gene order differences were also identified. A third region showed directly comparable orientation and gene order in both species, whilst the fourth region solely showed inverted orientation, with identical gene order. A similar approach has recently been described for the cat (Murphy *et al.*, 1999a). Previously determined Zoo-FISH data were exploited in the generation of a preliminary RH map of FCA B4 and D3, and comparison with the map location of loci within the corresponding ECCS on HSA 12 and 22. Many of the markers used were derived from the CATS primer set (Lyons *et al.*, 1997). The authors note the common occurrence of rodent co-amplification, which was overcome by the design of additional marker primers at regions of mismatch between rodent and cat sequence, which was also the approach used in this study for the generation of suitable marker primers for GUCY2E and THY1. Parallel radiation hybrid mapping has since been used to demonstrate the conservation of synteny of loci encoded on the sex chromosomes of human, mouse and cat, and in turn, to establish that gene order is also conserved between human and cat X chromosomes but is disrupted in the mouse (Murphy *et al.*, 1999b).

Radiation hybrid mapping is therefore becoming an increasingly popular technique for genome analysis in a range of diverse species. The development of two WG-RH panels for the dog has enabled this technique to be applied and integrated with data for this species from other mapping approaches. The first dog genome map generated by WG-RH mapping was described by Priat *et al.* (1998), comprising 218 type I and 182 type II markers. These were mapped to a total of 57 distinct linkage groups with a lod score of six or greater, each containing between two and twenty one markers. Fourteen linkage groups were assigned to nine different dog chromosomes on the basis of previous studies. Fifty three markers were not linked to any group. The radiation hybrid mapping analyses of Priat *et al.* (1998) were performed with only a single PCR assay across the whole panel for each marker, although forty markers were selected for duplicate testing and no major discrepancies were reported. Up to four ambiguous scores were identified in some cases over the 126 cell lines tested, although it was reported than none caused any significant difference between the mapping data calculated for the two duplicate sets of assays. These data have been incorporated into an integrated RH and meiotic linkage map of the dog genome (Mellersh *et al.*, 2000) comprising 600 RH markers, of which 523 were placed with an average distance of 24.3 cR_{5000} between markers. The map was compared with data for 341 markers assigned by meiotic linkage analysis, identifying 37 autosomal linkage groups, of which 14 could be assigned to the corresponding

chromosome (Werner *et al.*, 1999). A total of 724 markers were included within the integrated map, of which 217 were assigned by both methods. A total of 12 markers were placed on the meiotic linkage map of CFA 5, representing a single type I marker (GLUT4, also known as SLC2A4) and 11 type II markers. The accompanying RH map comprises 16 markers, of which three (DIO1, CD3E and GLUL) represent coding loci. Nine markers are included in both the meiotic linkage and RH maps, all of which are anonymous type II loci. The assignment of GLUT4, CD3E and DIO1 to CFA 5 correlates with data generated in this study, in which all three of these loci have been assigned to this chromosome by both RH and FISH analysis. The data obtained in the present study demonstrate that the meiotic linkage and RH maps of Mellersh *et al.* (2000) are shown in the opposing orientation to their location on CFA 5, such that the uppermost marker in both maps represents the most distal locus assigned to CFA 5. The inclusion of the fourth type I marker, GLUL, in the integrated map of Mellersh *et al.* (2000) does not correlate with data obtained in the present study. The human GLUL gene has been assigned to HSA 1q25 (Wang *et al.*, 1996) and subsequently to HSA 1q31 (Helou *et al.*, 1997). Neither assignment correlates with the chromosome painting data obtained in this study, or that of Yang *et al.* (1999), since no hybridisation of the CFA 5 chromosome paint probe was observed on HSA 1q. One potential explanation is the presence of an ECCS at this site that was below the resolution limits of Zoo-FISH analysis, and which may only become apparent with the assignment of single locus probes for which the chromosomal location of the human orthologue is known. The position of the GLUL marker on the RH map of Mellersh *et al.* (2000) places this locus proximal to CD3E, at the centromeric end of the CFA 5 RH map. An additional potential explanation can be proposed on the basis of the assignment of several GLUL-like human genes by Wang *et al.* (1996), of which one (GLULL3) was assigned by FISH analysis to HSA 11q24. This human chromosome region has been shown in this study to be evolutionarily related to the proximal region of CFA 5q. This was confirmed by the assignment of three markers, CD3E (HSA 11q22.3-q23), THY1 (HSA 11q23) and MMP1 (HSA 11q21-q22) to CFA 5q13 (CD3E and THY1) and CFA 5q14.3 (MMP1). The assignment of a locus from HSA 11q24 to a region proximal to CD3E therefore correlates with the data obtained in the present study, since it been established that the CFA 5/HSA 11 ECCS lie in opposing orientations with respect to the centromere of each chromosome. It may therefore be the case that the GLUL marker assignment by Mellersh *et al.* (2000) in fact represents the GLULL3 locus. This could result from misinterpretation of database similarity search results when the

marker product was first characterised, or incorrect annotation of a database sequence with which similarity was identified. There may be sufficient sequence similarity between the GLUL and GLULL3 loci that a locus other than that expected was amplified, or the assignments of the human loci may be inaccurate. However, since the anomaly involves only a single marker it is not possible at this time to establish the validity of this result. An increased density of mapped markers within this region will be required to establish whether an additional ECCS lies on CFA 5 that was not identified in the present study.

It is not yet clear whether either of the two dog RH panels will be used more extensively than the other. At a future time, it should be relatively straightforward to incorporate data from chapter five into maps generated using the alternative panel produced by Vignaux *et al.* (1999a). Similarly, data from the latter can be extrapolated for analysis using the Research Genetics RH panel that was originally produced by McCarthy and colleagues (unpublished) used in this study. It seems likely that both panels will continue to be used, dependent on the availability of each resource to the individual researcher, the nature and size of the chromosome region of interest, and the number of markers to be mapped. Parallel use of two such resources is common, exemplified by the use of three RH panels for human genome mapping, namely GB4 (Gyapay *et al.*, 1996), G3 (Stewart, E.A. *et al.*, 1997) and TNG (http://www-shgc.stanford.edu/Mapping/rh/). Maps generated by the individual panels can be integrated by assaying a subset of key markers mapped using one panel with one or both of the other panels. This approach could easily be applied to the transfer of mapping information generated with one dog RH panel to a map generated using the alternative resource. As the number of mapped markers for the dog genome increases, a further, higher resolution RH panel may be required, and again existing data may be extrapolated by the analysis of selected markers from existing maps on the new panel.

This study has described the mapping of 21 markers to CFA 5, eleven of which represent coding loci, making a significant contribution towards the development of an integrated RH and cytogenetic map of this chromosome. All previously reported gene markers from human ECCS corresponding to regions of CFA 5 have been mapped by both techniques, and novel markers have also been generated. It must therefore be a priority to develop an extensive panel of new type I markers that can be integrated into the CFA 5 map developed in this study. The following chapter describes the preliminary development and application of a method by which this may be achieved.

5.6 Summary

The basis of an integrated cytogenetic and RH map of CFA 5 has been formed, consisting of eleven type I and ten type II markers assigned by both methods. One marker mapped by FISH analysis could not be optimised for RH analysis, and new PCR primers were designed for three further loci for which rodent recipient amplification could not be eliminated. Markers have been formed into six linkage groups, comprising between two and six markers. All pairwise linkage data were in support of those generated by FISH analysis. Multipoint analysis enabled locus order to be predicted for the four linkage groups comprising three or more markers, which correlated with data generated by FISH analysis of these loci. RH data have confirmed the proposed orientations of corresponding ECCS within CFA 5 and the human genome discussed in chapter four, and comparisons between the relative order of gene loci within these regions have been made.

Chapter six:

Direct cDNA selection towards isolation of gene markers for dog chromosome five

Direct cDNA selection towards isolation of cDNA sequences from dog chromosome five

Abstract

Direct cDNA selection was performed using CFA 5 as the target for the isolation of a chromosome-enriched transcript population, towards the generation of novel gene markers for this dog chromosome. A whole-genome dog cDNA library generated from cerebellum tissue was initially examined by PCR, sequencing and FISH analysis to confirm its suitability for use in cDNA selection. A pool of cDNA clones was fluorescently labelled by the PCR and hybridised to dog metaphase chromosomes to investigate the genome coverage of the transcript population. This showed that the pattern of hybridisation correlated strongly with the location of GC-rich genome regions. A transcript population was amplified by the PCR and hybridised in solution to a biotinylated CFA 5 chromosome paint probe, in the presence of competitor DNA. Hybrids were isolated from the probe mixture by immobilisation of biotinylated paint probe fragments, and in turn, of associated cDNA, on streptavidin-coated magnetic beads. Following stringency washing of the bead-template complex, captured cDNA was amplified *en masse* by the PCR and fluorescently labelled for FISH analysis. The pattern of hybridisation of the labelled cDNA population to dog metaphase chromosomes was compared to that obtained using preselected cDNA from the original transcript population. Enrichment of the cDNA population for sequences derived from CFA 5 was distinctly apparent, since more intense hybridisation signal was observed on this chromosome, relative to other chromosomes, than was apparent for the preselected cDNA probe. In addition, hybridisation signal was consistently observed within a number of GC-rich chromosome regions, demonstrating suboptimal efficiency of the selection procedure. Explanations for these findings were considered, and possible improvements of the method were proposed on the basis of the preliminary data obtained.

6.1 Titration of the dog cDNA library

A partial cDNA library derived from dog cerebellum tissue was obtained from D. Sargan (Cambridge University, UK), and was used as the source of a dog transcript population in cDNA selection. The construction of the library has been described previously (Lin and Sargan, 1997). Briefly, the poly(A)+ fraction of total cerebellar RNA was isolated using oligo(dT)-chromatography. Following reverse transcription, resulting cDNA was cloned non-directionally into the *Eco*R1 site of the λZAPII vector (Short *et al.*, 1988), and transformed into the XL1Blue MRF' host strain. The library had been amplified once prior to receipt.

The titre of the cDNA library had previously been estimated at 10^{10} pfu/ml (D. Sargan, pers. comm.). In order to assess the status of the library subsequent to a period of storage, the titre of the stock aliquot provided was re-estimated. Ten-fold serial dilutions of the stock were prepared in SM buffer. From the previous estimation of the library titre, the expected number of pfu/μl was calculated for each dilution. A series of titration plates were prepared using the quantities of each dilution indicated in table 6.1a, which shows both the expected and observed number of pfu for each plate. These data were used to estimate a mean value of 0.022 for the ratio of the observed:expected number of pfu/μl. This implies that the titre at the time of the analysis was 0.022 times that predicted, or approximately 2.2 x 10^5pfu/μl (2.2 x 10^8 pfu/ml).

6.2 Preliminary analysis of random dog cDNA clones in the lambda vector

6.2.1 Assessment of insert size

Eighty lambda clones were selected at random from the plates used for the library titration described in section 6.1.2. The size of the insert DNA from each clone was investigated by PCR analysis as described in section 2.1.4, using vector primers ($T3^{SK}$ and $T7^{SK}$, appendix B) with an annealing temperature of 60°C, and a two minute elongation period. Data are summarised in table 6.2a. Of the 80 clones, 58 (72.5%) demonstrated the presence of a single product representing insert DNA of at least 0.2kb, satisfying the minimum criteria required for further analysis, as discussed in chapter three. The average size of the insert within these 58 clones was calculated as approximately 1.1kb, with a maximum of 2.8kb. A total of 31 clones (39%) contained inserts of the optimum size for further analysis (discussed in chapter three), generating single amplification products in excess of 1kb. PCR analysis of nineteen clones (23.75%)

Table 6.1a:

Quantity of serial dilutions plated for titration of the dog cDNA library. The number of pfu expected on the basis of the previous titration assay are given in column two. The observed number of plaque-forming units are given for each duplicate assay in columns three and four. The mean values for these data (column five) were used to re-evaluate the library titre, by comparison of the ratio of observed:expected pfu, calculated in column six.

Amount used	No. pfu expected	No. of pfu obtained		Mean obtained	Observed/Expected
		Plate 1	Plate 2		
50µl x 10^5 pfu/ml	5000	96	148	122	0.024
20µl x 10^5 pfu/ml	2000	48	57	53	0.027
100µl x 10^4 pfu/ml	1000	30	32	31	0.031
50µl x 10^4 pfu/ml	500	7	6	7	0.010
20µl x 10^4 pfu/ml	200	3	5	4	0.020
					average = 0.022

Table 6.2a:

Range of insert size of 80 randomly-selected cDNA clones. Insert size was estimated by comparison with a 1kb DNA ladder. Of the 80 clones analysed in this manner, 31 (39%) demonstrated optimal results for further analysis as discussed in chapter three, with single amplification products in excess of 1kb. These clones are highlighted in the table below. Percentage values are calculated to the nearest whole number.

PCR product size	Number of clones (%)
Single product < 0.2kb	19 (24%)
Single product > 0.2kb < 1.0kb	27 (34%)
Single product > 1.0kb < 2.0kb	**25 (31%)**
Single product > 2.0kb	**6 (8%)**
Multiple products	1 (1%)
No product	2 (3%)
	Total = 80 (100%)

Table 6.2b:

Results of database similarity searches with sequence data from 25 cDNA clones.

Result of similarity search	Number of clones (%)
Significant match with previously annotated gene	3 (12%)
Significant match with dog mitochondrial DNA	2 (8%)
Significant match with bacterial DNA	1 (4%)
No significant match	19 (76%)
	Total = 25 (100%)

indicated the presence of less than approximately 0.2kb of insert DNA, and were not considered suitable for further analysis. A single clone yielded three products, suggesting that more than one plaque had been isolated. The remaining two clones did not generate any detectable product from the PCR.

6.2.2 Sequence analysis of randomly selected phage clones

Phage DNA was used in the PCR with $T3^{SK}$ and $T7^{SK}$ primers under the conditions described in section 6.2.1, to generate insert DNA for sequencing analysis. Twenty five clones analysed in section 6.2.1 were selected for further analysis. These represented the clones containing the largest inserts, with an average insert size of 1.7kb, which would therefore be expected to yield the most sequence data for downstream analysis. Sequencing data were obtained using the $T3^{SK}$ primer, and where satisfactory data were not obtained, this was repeated using the $T7^{SK}$ primer. Edited data were searched for similarity with previously generated sequences using the BLAST search tool, as described in section 2.1.9. Data are summarised in table 6.2b.

Of the 25 clones, sequence data from three (12%) generated highly significant similarity scores with previously annotated genes. Clone R22 showed 93% identity over 366 nucleotides of the human phosphoglycerate kinase 1 (PGK1) gene sequence (HSA Xq13). Clone R32 showed 88% identity over 344bp of the sequence encoding the bovine selenoprotein P-like (SEPP1) protein (HSA 5q31). The third clone corresponding to a highly significant database match was clone R61, sharing 92% sequence identity over 363 nucleotides of the pig FOS gene (murine osteosarcoma viral oncogene homologue, HSA 14q24.3). Two clones (8%) shared in excess of 96% identity with the dog mitochondrial DNA sequence, namely R13 and R94, over a minimum of 200 nucleotides. Clone R38 represents a bacterial contaminant, sharing 98% identity over a region of approximately 390bp of *E. coli* genomic DNA. No other significant database matches were identified, under the criteria described in section 3.1.3.

6.2.3 Chromosomal assignment of the dog FOS gene

Since FOS represents an oncogene for which the dog orthologue has not previously been assigned to a chromosome, the characterisation of clone R61 was extended to the isolation of a dog BAC clone containing this sequence. A pairwise alignment of data from clone R61 and the complete sequence for the human FOS gene (accession no. K00650.1), shown in figure 6.2a demonstrates that the dog cDNA sequence lies solely

```
                2881       2895 2896       2910 2911       2925 2926       2940 2941       2955 2956       2970
1 R061          --------------- --------------- --------------- --------------- --GTGGAGCCCGTCA AGAGCATCGGCAGCA
2 K00650.1      CCCCGGAGTCTGAGG AGGCCTTCACCCTGC CTCTCCTCAATGACC CTGAGCCCAAGCCCT CAGTGGAACCTGTCA AGAGCATCAGCAGCA

                2971       2985 2986       3000 3001       3015 3016       3030 3031       3045 3046       3060
1 R061          TGGAGCTGAAGGCCG AGCCCTTTGATGACT TCCTGTTTCCAGCAT CATCCAGGCCCAGCG GCTCGGAGACCGCCC GCTCCGTGCCAGACA
2 K00650.1      TGGAGCTGAAGACCG AGCCCTTTGATGACT TCCTGTTCCCAGCAT CATCCAGGCCCAGTG GCTCTGAGACAGCCC GCTCCGTGCCAGACA

                3061       3075 3076       3090 3091       3105 3106       3120 3121       3135 3136       3150
1 R061          TGGACCTGTCTGGTT CCTTCTATGCAGCAG ACTGGGAGCCCCTGC ATGGTGGCTCCCTGG GGATGGGGCCCATGG CCACAGAGCCCGAGC
2 K00650.1      TGGACCTATCTGGGT CCTTCTATGCAGCAG ACTGGGAGCCTCTGC ACAGTGGCTCCCTGG GGATGGGGCCCATGG CCACAGAGCTGGAGC

                3151       3165 3166       3180 3181       3195 3196       3210 3211       3225 3226       3240
1 R061          CTCTGTGCACCCCCG TAGTCACCTGTACTC CTAGCTGCACTACCT ATACGTCTTCCTTCG TCTTCACCTACCCTG AGGCTGACTCCTTCC
2 K00650.1      CCCTGTGCACTCCGG TGGTCACCTGTACTC CCAGCTGCACTGCTT ACACGTCTTCCTTCG TCTTCACCTACCCCG AGGCTGACTCCTTCC

                3241       3255 3256       3270 3271       3285 3286       3300 3301       3315 3316       3330
1 R061          CCAGCTGTGCGGCC- --------------- --------------- --------------- --------------- ---------------
2 K00650.1      CCAGCTGTGCAGCTG CCCACCGCAAGGGCA GCAGCAGCAATGAGC CTTCCTCTGACTCGC TCAGCTCACCCACGC TGCTGGCCCTGTGAG
```

Figure 6.2a:

Alignment of sequence data from cDNA clone R61 and the human FOS gene (accession no. K00650.1). Nucleotides that are identical in both sequences are highlighted in red text, and demonstrate that clone R61 shares 92.3% identity with the 312 nucleotide region of the human FOS gene shown above.

within a single exon (exon four) of human FOS. Assuming that the gene structure of human and dog FOS is conserved, PCR primers could therefore be designed with confidence from clone R61 sequence data, with limited probability of spanning an intron-exon boundary. A pair of primers were designed (appendix B) that spanned 278bp of the dog cDNA sequence. A product of this size was observed when both R61 clone cDNA and dog genomic DNA were subjected to PCR amplification using these primers, with an annealing temperature of 60°C, and a one minute elongation step. The identity of the amplification product was confirmed by sequencing analysis, which demonstrated 91% identity with human FOS (accession no. K00650.1) over a region of 211bp.

The PCR product was radiolabelled and this probe was used to screen three filters from the dog BAC library, as described in section 2.4. A single positive hybridisation signal was identified, corresponding to BAC clone 145a3. This clone was confirmed as PCR positive with the dog FOS primers, generating a product of comparable size to that obtained from dog genomic DNA. BAC clone 145a3 was used in FISH analysis to assign the dog FOS gene to CFA 8q31. This assignment correlates with dog-human chromosome painting data (Breen *et al.*, 1999a; Yang *et al.*, 1999), since CFA 8 and HSA 14 are shown to correspond to a single ECCS in both species. This assignment also suggests that the ECCS lie in the same orientation in both genomes, since FOS is located within the distal half of both CFA 8 and HSA 14, at CFA 8q31 and HSA 14q24.3 respectively. A representative example of the FISH analysis data is shown in figure 6.2b. The corresponding ECCS on CFA 8 and HSA 14 are compared in figure 6.2c with respect to the chromosomal location of the FOS gene in both species.

6.3 Preparation of phagemid DNA for cDNA selection

6.3.1 Isolation of phagemid from the lambda vector

Since the dog cDNA library was to be used in downstream studies, and no data were available to demonstrate its sequence composition, the lambda library was converted to a phagemid library by *in vivo* excision, and further characterisation was performed. *In vivo* excision of the pBluescript phagemid vector from the lambda library was performed as described in section 2.6.2. The methodology is outlined in figure 2.6a. Assuming that the number of unique sequences represented in the cDNA library is approximately 100,000 (on the basis of the estimated number of gene sequences contained within a mammalian genome, section 1.2.3), phage representing a 100-fold excess of this number were treated in the excision procedure, in order to maintain the representation of the library. Based on

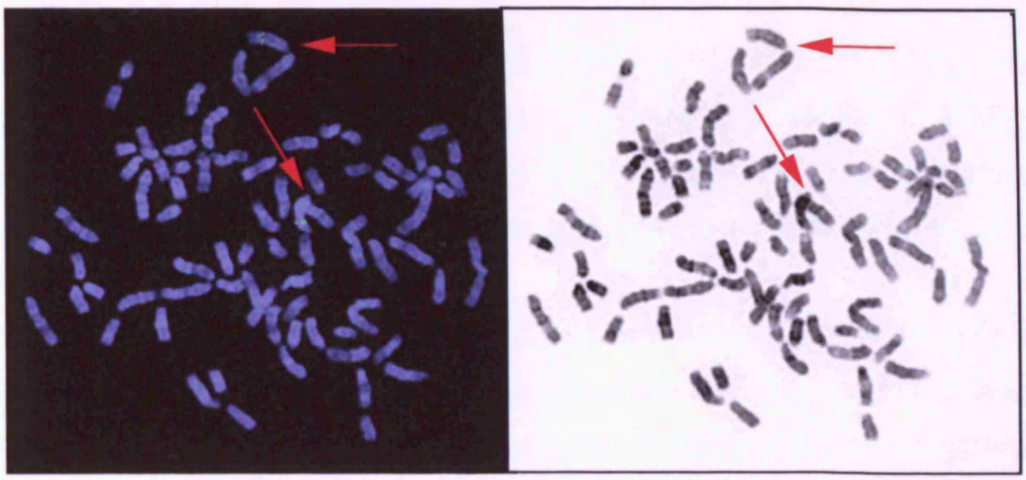


Figure 6.2b:

Assignment of the dog FOS gene to CFA 8q31 by FISH analysis. Hybridisation signal corresponding to the FOS gene is indicated by a red arrow in the image on the left, and also in the DAPI-banded image of the same metaphase spread to the right.

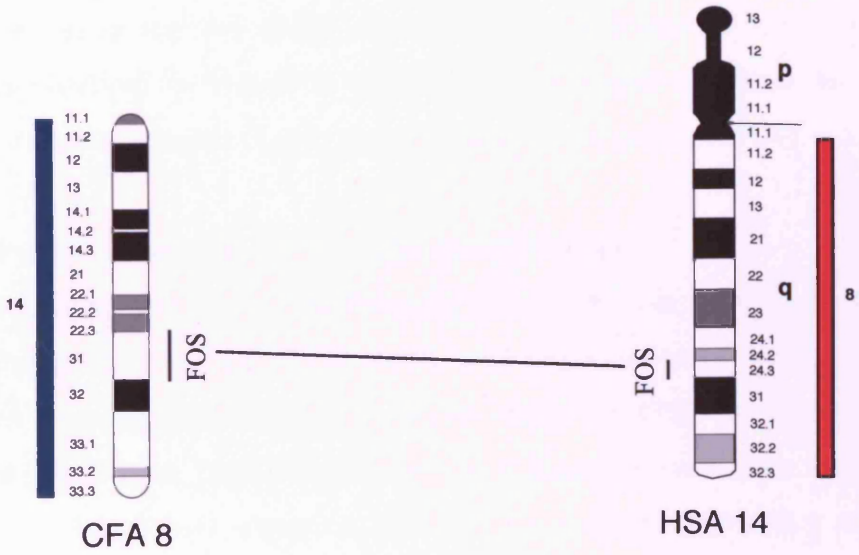


Figure 6.2c:

Comparison of the corresponding ECCS on CFA 8 and HSA 14, with respect to the chromosomal location of the FOS gene (Breen *et al.*, 1999a).

the library titre calculated in section 6.2.2, 460µl of a tenfold dilution of the library were used in phagemid excision, corresponding to approximately 10,000,000pfu. Following the manufacturer's recommendations, the phage were combined with an equivalent number of ExAssist helper phage particles and a ten-fold excess of the XL1 Blue MRF' host strain.

Serial dilutions (table 6.3a) of the resulting 20ml of excised phagemid preparation were combined with SOLR host cells in order to estimate the titre. The number of colonies obtained in this procedure indicated the concentration of the excised phagemids as approximately 120 colony forming units (cfu)/µl, corresponding to a total of 2,400,000 cfu isolated in total. Since approximately 10,000,000 pfu were entered into the excision process, this corresponds to approximately 24% excision efficiency.

Large-scale isolation of phagemid DNA was performed as described in section 2.6.4. The resulting preparation was analysed by electrophoresis as shown in figure 6.3a. The concentration of the sample was estimated as 1µg/µl by comparison with known quantities of sonicated salmon sperm DNA, indicating a total yield of approximately 1mg of phagemid DNA. The size of the product ranged from a minimum of between 2kb and 3kb, which represents the size of the non-recombinant pBluescript vector. A smear of products was observed to extend beyond the largest fragment within the 1kb DNA ladder, indicating the presence of amplification products in excess of 12kb.

6.3.2 Preliminary analysis of a subset of phagemid clones

Ten phagemid clones were selected from the plates used for titration of the excised phagemid preparation (section 6.3.1). Each was inoculated into 10ml of LB broth supplemented with 50µg/ml ampicillin, and incubated at 37°C for approximately 16 hours, shaking at 250rpm. Duplicate cultures were also prepared for the wild-type XL1 Blue MRF' and SOLR host strains, in order to demonstrate the inability of both host strains to propagate in the absence of phagemid, since the latter is required to confer ampicillin resistance. A further 10ml of LB-ampicillin were incubated under the same conditions to confirm the absence of contaminating growth from other sources. Both wild-type host strains were also plated onto LB-ampicillin agar, as were three of the ten clones selected at random from those for which a liquid culture was prepared.

After overnight incubation, no growth of either wild-type host strain was visible on solid medium, but all three clones taken from the phagemid titration plate produced prolific

Table 6.3a:

Phagemid dilutions used in library titration.

Quantity of phagemid added	Quantity plated	No. of cfu produced	Estimated cfu/µl of stock
100µl	50µl	Uncountable	~
50µl	25µl	Uncountable	~
10µl	5µl	550	110
1µl	0.5µl	67	134
1/10µl	1/20µl	6	120
1/50µl	1/25µl	1	~
			Mean = 120cfu/µl

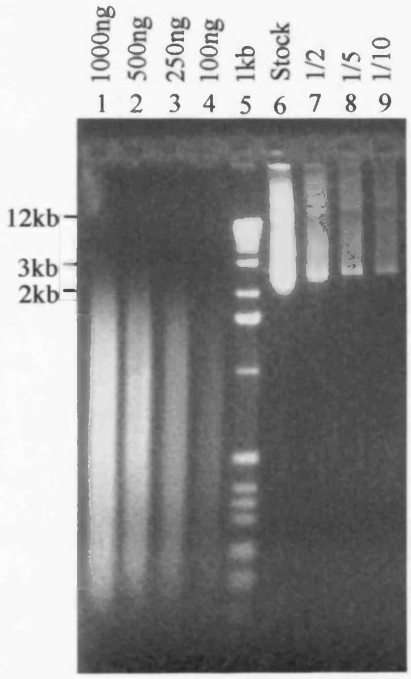


Figure 6.3a:

Large-scale isolation of phagemid DNA. Dilutions of the phagemid DNA preparation were analysed by agarose gel electrophoresis (lanes six to eight). The dilution factor is indicated above the relevant lane of the gel. The size range of the isolated phagemid DNA was estimated by agarose gel electrophoresis and comparison with a 1kb DNA ladder. The concentration of dilutions of the DNA preparation were also compared with known quantities of sonicated salmon sperm DNA, which are indicated above lanes one to four.

growth on the same medium. Growth of host strains was also absent in liquid medium, but was clearly visible in the liquid cultures of all ten test samples. No growth was apparent in any of the negative control samples that had not been inoculated.

Phagemid DNA was extracted from the liquid cultures of each of the ten test samples. One microlitre of the resulting DNA preparation was analysed by agarose gel electrophoresis and compared with size and quantitation standards as previously described. Approximately 25ng of phagemid DNA were used as the template in a standard 10µl PCR, using $T3^{SK}$ and $T7^{SK}$ primers, to assess insert size. Of the 10 clones, four generated PCR products of less than 0.2kb, and one did not yield product. The remaining five demonstrated the presence of insert DNA of a size that would be appropriate for further analysis, the range being from approximately 0.7kb to 2kb.

6.4 Amplification of insert DNA for use in cDNA selection

6.4.1 Establishing PCR conditions for cDNA amplification from phagemid DNA

Optimum template concentration and PCR cycling conditions were established by the preparation of a set of reactions varying in these criteria. Serial dilutions of phagemid DNA isolated in section 6.3.1 were prepared in MQ water, corresponding to 1/10, 1/20, 1/50, 1/100 and 1/250-fold dilutions of the stock. Five 10µl amplification reactions were prepared, each including 1µl of one of the template dilutions. These were subjected to standard cycling conditions with an annealing temperature of 60°C, and one minute denaturation, annealing and elongation steps, using $T3^{SK}$ and $T7^{SK}$ primers. A second identical set of reactions was amplified under the same conditions except that the elongation step of each cycle was increased from one minute to two minutes. This step was further increased to four minutes for a final set of five reactions.

Subsequent to the PCR, five microlitres of each reaction were analysed by electrophoresis through a 2% agarose gel, as shown in figure 6.4a. The products were compared to establish the effect of the varying template concentration on the product yield. A distinct increase in the average product size was observed when the elongation time was increased. The increase was more marked when the duration of the elongation step was increased from one minute to two minutes, compared to the difference between two and four minute elongation steps. High molecular weight DNA, representing excess template, was apparent when the 1/10 and 1/20 dilutions of the phagemid stock were

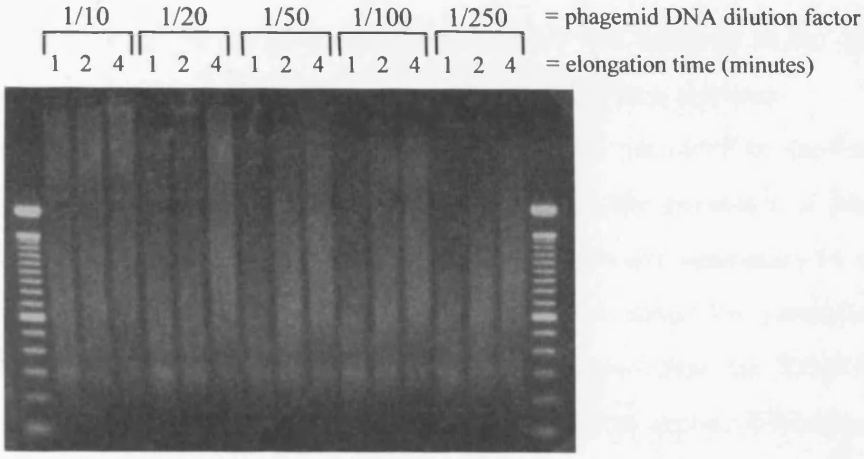


Figure 6.4a:

Comparison of the size range of insert cDNA amplified under different PCR conditions. Dilutions of the DNA preparation shown in figure 6.3a were used as the template in a series of amplification reactions in which both template concentration and elongation time were varied. PCR products were analysed by agarose gel electrophoresis and the results are shown above. The dilution factor of phagemid DNA used in the reaction is shown above the relevant lane of the gel. The duration of the elongation period used in the amplification reaction is also indicated (in minutes) above the relevant lane. The increase in the average size of the resulting PCR products can be seen clearly as the elongation time is increased.

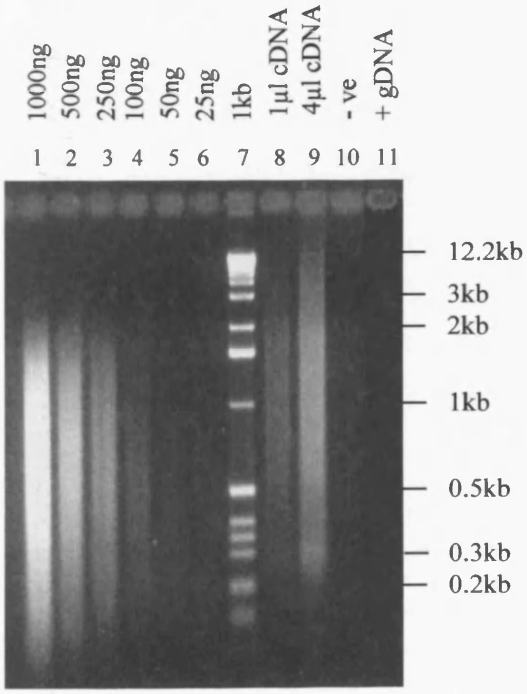


Figure 6.4b:

Generation of cDNA template by PCR amplification. Vector primers were used to amplify insert cDNA for use as the probe in cDNA selection. The sample was quantified by reference to known concentrations of sonicated salmon sperm DNA (lanes one to six), and the size range of the amplified cDNA was assessed relative to a 1kb DNA ladder (lane seven). Lanes eight and nine contain 1µl and 4 µl of amplified cDNA respectively. Lane ten contain the negative control (no template DNA) and lane eleven demonstrates that no amplification was generated in the presence of a total dog genomic DNA template.

used. A 1/50 dilution, representing approximately 20ng/µl, was selected as the optimum template quantity to use in the PCR, with an elongation step of four minutes.
Using the optimised conditions, further PCR analysis was prepared to confirm that amplification using $T3^{SK}$ and $T7^{SK}$ primers occurred only in the presence of phagemid DNA, and did not result from a genomic DNA template. This was necessary in order to confirm that amplified product generated from the cDNA selection procedure was derived from the selected cDNA, and not from the biotinylated genomic DNA used to capture the cDNA. A set of duplicate reactions was prepared as above, differing only in the template DNA used. Twenty five nanograms of dog genomic DNA were added to one reaction, and the same quantity of phagemid DNA to a second, acting as a positive control. A negative control was prepared using MQ water in place of DNA, from which absence of amplification was confirmed. No amplification was detected from dog genomic DNA, whilst that from phagemid DNA was comparable in size and yield to that observed in the previous amplification experiment. These controls were included in all subsequent cDNA amplification procedures.

6.4.2 Preparation of amplified insert DNA for cDNA selection

A large-scale amplification procedure was performed in order to generate sufficient material to perform cDNA selection, under the optimised conditions identified in section 6.4.1. A reagent mix was prepared containing quantities sufficient for a final volume of 500µl, and the mix was dispensed into 50 identical 10µl volumes. This was intended to avoid any effect of a larger reaction volume on the optimised amplification conditions. Two 10µl negative control reactions were prepared, one including 25ng of dog genomic DNA, and the other MQ water in place of template DNA. All reactions were subjected to the PCR with an annealing temperature of 60°C and a four minute elongation period. Amplified insert DNA reactions were pooled and purified as described in section 2.1.6. The size range and concentration of the resulting preparation was assessed by agarose gel electrophoresis, as shown in figure 6.4b. Amplified cDNA was quantified at 100ng/µl, with a size range of approximately 0.3kb to in excess of 12kb. The two negative control samples were also analysed on the same gel, confirming absence of product.

6.5 Preparation for cDNA selection

6.5.1 Removal of unincorporated biotin from the CFA 5 paint probe

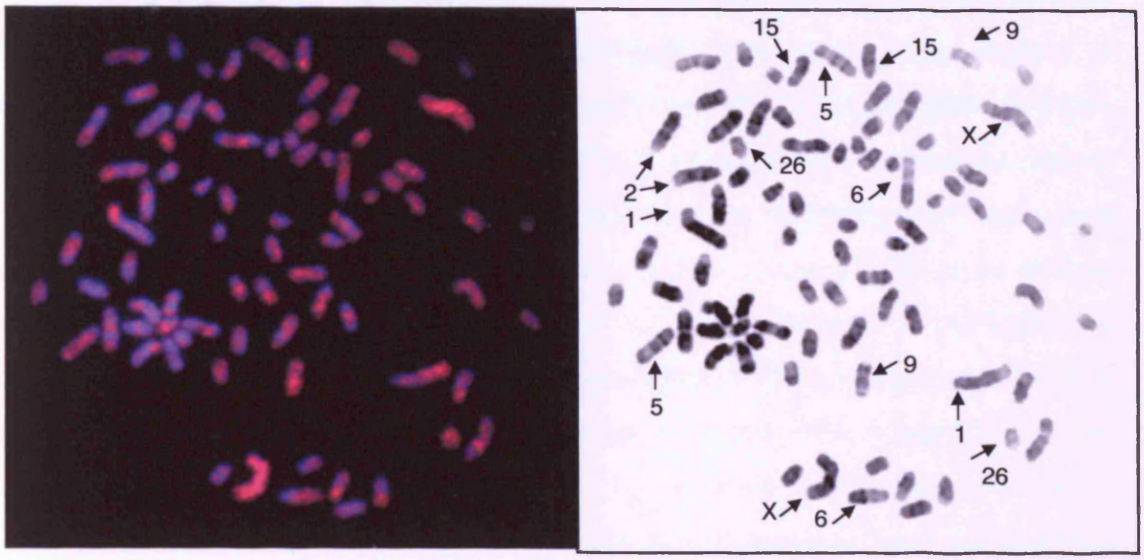
Unincorporated biotin residues were removed from the biotinylated CFA 5 paint probe. This step was essential since binding sites on streptavidin-coated beads may be blocked by free biotinylated nucleotides, reducing the efficiency of the capturing process. A 1µl aliquot of the biotinylated paint probe was analysed by gel electrophoresis, in comparison with a 1kb DNA ladder and known quantities of sonicated salmon sperm DNA. The probe fragments demonstrated a size range of approximately 0.3kb to 1.0kb, and the probe was quantified. A volume corresponding to a total of 3µg of probe was removed and unincorporated biotin was eliminated as described in section 2.6.7. The resulting products were eluted in 30µl of MQ water, and the concentration and fragment size range was established by comparison with known concentrations of sonicated salmon sperm DNA, and a 1kb DNA ladder, respectively. No detectable change in the range of fragment sizes was detected after the purification procedure, and the probe concentration was adjusted to 100ng/µl with additional MQ water.

6.5.2 FISH analysis of the CFA 5 paint probe

Prior to cDNA selection, the purified paint probe was used in FISH analysis to confirm successful and specific hybridisation to CFA 5. One hundred nanograms of paint probe were combined with 10µg of sonicated dog genomic DNA. The probe was used in standard FISH analysis on dog metaphases, as described in section 2.4, and hybridisation was performed for 16 hours at 37°C. FISH analysis confirmed intense and highly specific hybridisation of the paint probe to CFA 5, with no detectable background signal on any other chromosome.

6.5.3 FISH analysis of pre-selected cDNA

The suitability of the cDNA library for use in cDNA selection was also investigated by FISH analysis of a subset of the library prior to the selection procedure. Five hundred nanograms of the purified cDNA preparation generated in section 6.4.2 were labelled by incorporation of Spectrum-Red conjugated-dUTP by nick translation (section 2.6.13). The probe was combined with 50µg of sonicated salmon sperm DNA to aid recovery of the sample after precipitation, and hybridised to dog metaphases for 16 hours at 37°C. Examples of the resulting images are shown in figure 6.5a.



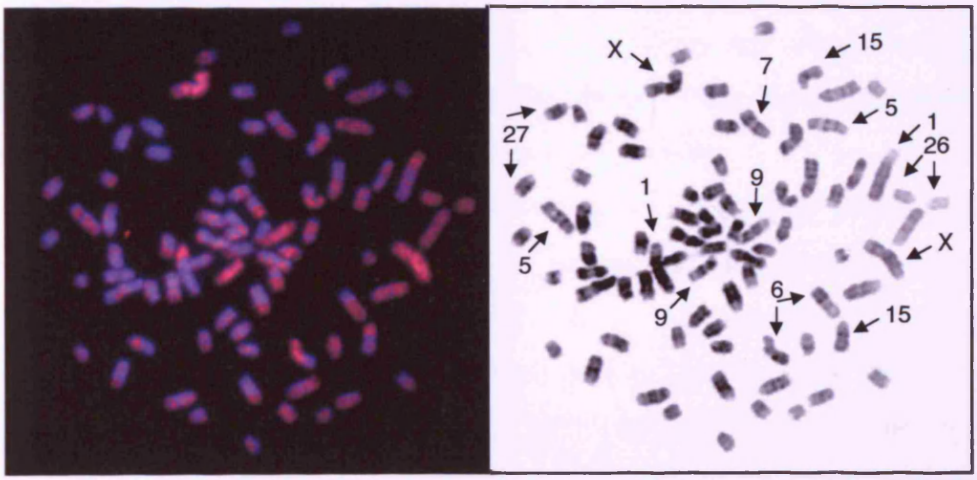


Figure 6.5a:

FISH analysis of preselected cDNA. The images on the left demonstrate the genome coverage of the initial cDNA population prior to any selection. Hybridisation signal can be seen on all chromosomes, but is clearly more intense within a subset of chromosomes, examples of which are indicated by arrows. These regions largely correspond to GC-rich sites, which appear pale in the DAPI-banded images to the right.

This analysis demonstrated that the library was not significantly skewed towards transcripts from any specific chromosome, although there were clearly regions of stronger signal density compared to others, examples of which are highlighted in figure 6.5a. Examination of the DAPI-banded image of each metaphase showed that the regions of strong signal intensity correlated well with the location of DAPI-negative bands, such that the cDNA signal itself generated a banding pattern. Amongst the most striking signals were identified at the distal tip of CFA 1, within the region CFA 1q34-q38, and also along the length of CFA 9 and CFA 26. This region of CFA 1 is amongst the most extensive DAPI-negative regions of the dog genome, whilst CFA 9 and CFA 26 are amongst the most DAPI-negative chromosomes, as confirmed by the position of their flow peaks towards the GC-rich region of the dog flow karyotype (figure 4.1a). Strong hybridisation signals were also observed on CFA X, predominantly within the centromeric region, which is not regarded as being highly GC-rich. This finding correlates with the common occurrence of signal on this chromosome when probed with a range of genomic targets, including both single locus probes and chromosome paint probes (M. Breen, pers. comm.). It is therefore possible that CFA X contains sequence motifs that commonly cross-hybridise with other genomic regions.

CFA 6 is a distinct example of a cDNA hybridisation pattern correlating with the DAPI-banded pattern of the chromosome. CFA 6 is strongly DAPI-negative within the proximal half of the chromosome, and DAPI-positive in the distal half. In contrast, hybridisation of the cDNA probe was considerably more intense in the proximal half of the chromosome, correlating with the common hypothesis that GC-rich regions are also largely gene-rich. It is important to note also that the probe did not show considerable hybridisation to CFA 5 in excess of that which would be predicted from the banding pattern of the chromosome. That is, the starting cDNA population is not highly enriched for sequences from CFA 5 compared to other GC-rich chromosomes, and probe signal is in fact more highly concentrated in other chromosome regions, predominantly within the distal region of CFA 1, and in CFA 9 and CFA 26.

6.6 Preliminary investigation of the cDNA selection method

6.6.1 Hybridisation and capture of cDNA using a genomic DNA probe

The cDNA selection method was investigated by the use of the CFA 5 paint probe as a target for selection of transcripts derived from this dog chromosome, under the

hybridisation conditions described by Rouquier *et al.* (1995), unless otherwise stated. These conditions are described in sections 2.6.7 to 2.6.10.

Captured cDNA sequences were amplified by the PCR under the conditions described in section 2.6.11. Standard 10µl amplification reactions were performed using 0.5µl of bead suspension from the selection reaction as the template. Five µl of the resulting products were analysed by agarose gel electrophoresis, in order to confirm the presence of amplified cDNA. The absence of product in the negative control samples was also confirmed, in which the bead suspension was replaced by MQ water or by 50ng of dog genomic DNA. The size of the amplified products ranged from approximately 0.3kb to in excess of 4.0kb. Large-scale preparation of amplified cDNA was performed in 30 duplicates of a 10µl reaction each containing 0.5µl of bead suspension. Products were analysed by agarose gel electrophoresis, confirming the size range of fragments as approximately 0.3kb to in excess of 4.0kb, and the concentration of amplified fragments was estimated by comparison to known concentrations of sonicated salmon sperm DNA.

6.6.2 FISH analysis of amplified products

Five hundred nanograms of amplified insert cDNA were labelled using Spectrum Red-dUTP, as described in section 2.6.13. The probe was combined with 50µg of sonicated salmon sperm DNA, which was intended to generate an accurate reflection of the sequence composition of the cDNA probe, in the absence of homologous suppression, in contrast to the study of Rouquier *et al.* (1995) (discussed in section 6.7). In this instance, the sonicated salmon sperm DNA was used to aid recovery of the sample after precipitation, and to maintain the total DNA concentration of the probe at approximately that used in standard FISH analysis elsewhere in this study. Hybridisation was performed at 37°C for 16 hours.

The results of the cDNA selection process were examined by hybridisation of selected, amplified cDNA sequences to dog metaphases. In each case, preselected cDNA was hybridised within the same procedure onto a separate slide to ensure that resulting images were comparable to those obtained previously. This ensured that any difference on the pattern of hybridisation of selected cDNA was a result of the selection procedure itself and was not related to variation in hybridisation conditions. Examples of the hybridisation of selected cDNAs to dog metaphases are shown in figure 6.6a.

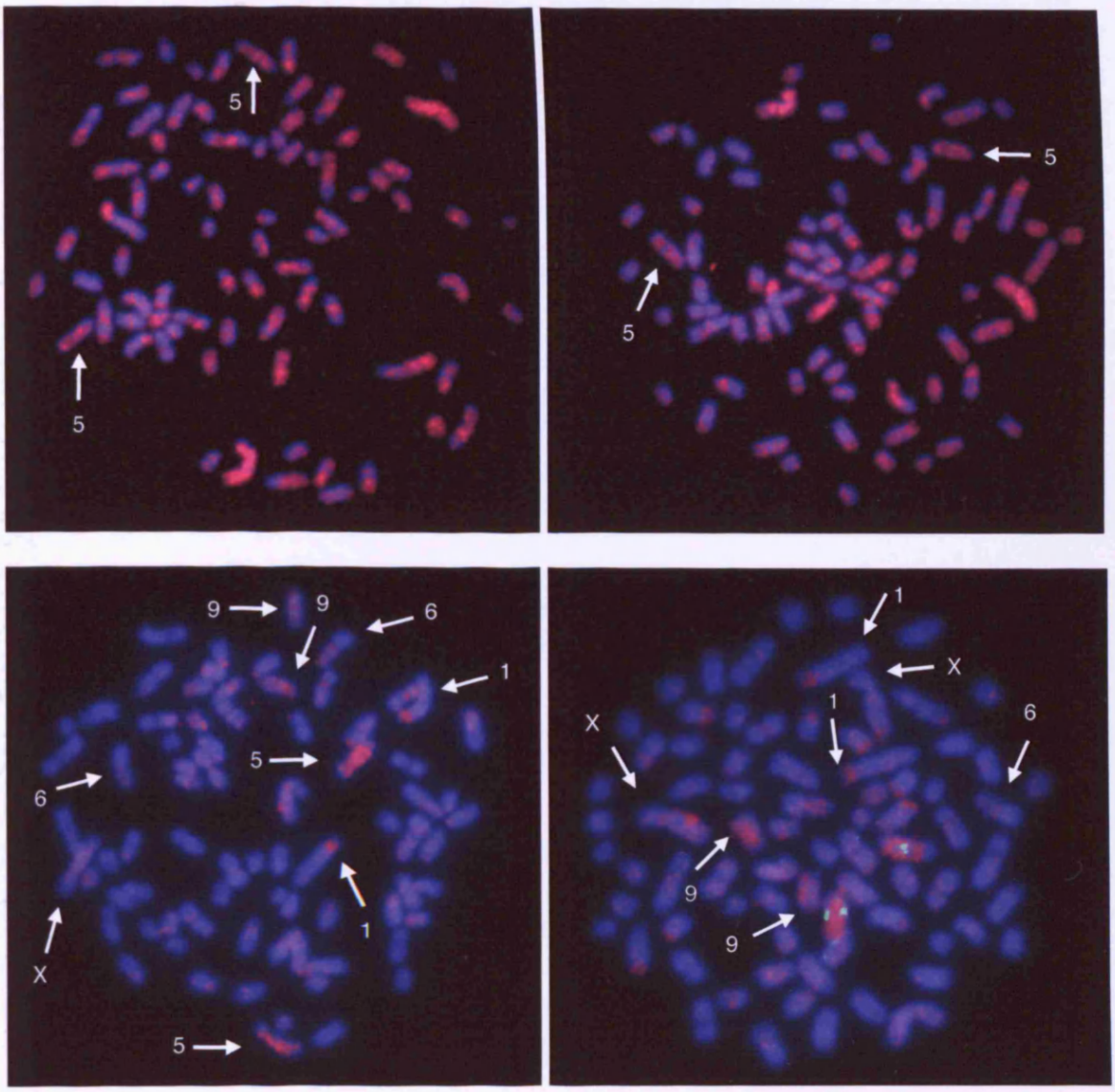


Figure 6.6a:

Comparison of pre- and post-selected cDNA populations by FISH analysis. The upper two images demonstrate the genome coverage of the preselected population, and are taken from figure 6.5a. The lower two images represent examples of transcripts recovered after cDNA selection using a biotinylated CFA 5 probe. Comparison of the two sets of images demonstrates that enrichment of the population for sequences derived from CFA 5 has been achieved by the selection procedure, since fluorescent signal is more concentrated on this chromosome in the lower images. Certain chromosome regions that were intensely painted in the preselected population can be identified in the postselection images, demonstrating that enrichment has not occurred with optimal efficiency. A subset of these sites are indicated for reference. CFA 5 is indicated in the lower right image by co-hybridisation of a labelled cosmid probe.

Section 6.5.3 discussed the fact that the starting cDNA population was not highly enriched for sequences from CFA 5, since probe hybridisation to this chromosome was no more intense than would be predicted from the position of CFA 5 in the dog flow karyotype. Preselected transcripts showed a higher affinity for other chromosomes and chromosome regions, particularly CFA 9 and CFA 26, and their pattern of hybridisation largely corresponded to DAPI-negative sites of the karyotype. It is apparent from comparison of these images with those presented for post-selected cDNAs that the selection procedure had reduced the complexity of the cDNA population, since the genome coverage of the probe was distinctly reduced and the DAPI-negative banding pattern was no longer apparent throughout the karyotype. Probe signal on CFA 5 was the most intense within each metaphase spread, and the signal consistently extended along the length of the chromosome, as was observed prior to selection. It is therefore apparent that CFA 5 transcripts had been selectively retained from the preselected cDNA population, whilst those from the rest of the karyotype had been reduced or eliminated. As demonstrated in figure 6.6a, a considerable proportion of the karyotype demonstrated no or minimal detectable probe hybridisation, despite the relative uniformity of probe hybridisation throughout each metaphase prior to selection. In addition to the intense signal on CFA 5, probe hybridisation was also apparent at a number of additional locations. These largely correlated with the regions of intense signal observed in preselected cDNA, namely at CFA 1qdist., CFA 9 and CFA 26. Signal on other chromosome regions that were strongly apparent in the preselected cDNA population were considerably reduced by the selection procedure, particularly the distal region of CFA 6, and also CFA X. The degree of hybridisation signal on all other chromosomes was markedly reduced, such that a greater proportion of the overall signal visible within each spread was concentrated on CFA 5. It was therefore concluded that cDNA selection had enriched the cDNA population for transcripts derived from CFA5, but had not occurred with complete efficiency.

6.7 Discussion

cDNA selection, using a chromosome paint probe as the genomic target, has been proposed as a method for generating novel type I markers for CFA 5. This technique was investigated following the previous report of a similar method by Rouquier *et al.* (1995) using human chromosome paint probes as the genomic targets. The aim therefore was to attempt to reproduce the data from this publication using dog cytogenetic resources, and if successful, to optimise the method for downstream sequencing analysis of selected cDNAs and subsequent marker development.

The study commenced with characterisation of the dog cDNA library that was to be used as the source of transcript material. Lovett (1994a) has commented on the importance of a high quality cDNA source in the optimisation of cDNA selection procedures. The presence of contaminants in many commercial libraries, and their typically skewed sequence representation, was cited amongst the potential disadvantages of their use over custom-made libraries. Evaluation of a subset of cDNA clones in this study indicated that the library appeared suitable for cDNA selection. The proportion of non-recombinant clones was estimated as only 3%, whilst 39% generated single amplification products in excess of 1kb. Sequencing analysis of a subset of 25 clones showed that the majority (76%) of clones did not demonstrate significant sequence similarity to existing database entries. Only a single bacterial contaminant was identified, whilst 12% of clones shared similarity with existing gene sequences, which led to the novel assignment of the FOS oncogene locus to CFA 8q31, correlating with the Zoo-FISH data of Breen *et al.* (1999a) and Yang *et al.* (1999).

The library was subsequently subjected to *in vivo* excision to convert the bacteriophage lambda library to a transcript population contained within the pBluescript phagemid cloning vector, facilitating downstream applications related to this and subsequent studies. The precise efficiency of the procedure is difficult to assess due to the considerable number of clones involved and the accompanying large error rate inherent in the estimation of the titre of both pre- and post-excision populations. However, the resulting estimation that 24% of phage clones entered into the excision process were recovered in the form of phagemid clones correlates well with the expected value of 10% efficiency stated by the manufacturers (Stratagene). Excised phagemid clones behaved as expected in both bacterial culturing and PCR analysis, and optimisation of amplification conditions demonstrated that the average size of amplified products could be greatly increased by extension of the elongation step to four minutes. This investigation was

considered appropriate in view of previous comments (Rouquier et al, 1995) regarding the limited average size (0.4kb to 0.5kb) of transcripts obtained following cDNA selection. Increase of the average size of transcripts entered into the cDNA selection process was thus intended in this study to generate a chromosome-enriched transcript population that would be of a size appropriate for sequencing analysis in the manner discussed in chapter three.

cDNA was amplified *en masse* under optimised PCR conditions and hybridised to dog chromosomes by FISH analysis. This enabled the genome coverage of the transcript population to be evaluated, prior to its use in cDNA selection. FISH analysis demonstrated relatively uniform coverage of the genome, with hybridisation signals consistently observed on all chromosomes, and signals were not observed to be excessively concentrated on any specific chromosome. Intense signal was apparent on certain chromosome regions, which were discussed in section 6.5.3. These hybridisation patterns correlate strongly with regions proposed to be highly gene-rich on the basis of DAPI-banding analysis. Therefore FISH analysis of the preselected cDNA population indicated that the library was suitable for use in subsequent cDNA selection.

FISH analysis of transcripts isolated by cDNA selection demonstrated that a degree of enrichment for CFA 5 transcripts had been achieved, although the efficiency of the procedure could not be matched with that apparent from the related study of Rouquier *et al.* (1995), despite repeated attempts. Whilst considerable probe signal was observed on CFA 5 following the selection procedure, intense signal was also apparent in other chromosome regions, predominantly the distal tip of CFA 1, and on CFA 9. These correspond to the sites at which most intense probe hybridisation was observed in the preselected transcript population. It can therefore be proposed that the selection procedure has enriched the cDNA population for CFA 5 transcripts, with a level of efficiency that reduced background hybridisation on chromosome regions other than those sites highly represented within the starting transcript source. Visual inspection of a range of metaphases from post-selection FISH analysis suggested that in excess of 60% of the probe signal was concentrated on CFA 5, compared to less than 5% to 10% in the preselection images. This itself would represent a significant improvement to the efficiency of identification of CFA 5 gene markers, but prior to cloning the selected transcripts for this purpose, further optimisation of the protocol would be appropriate.

Comparison of the FISH analysis data from the studies of Rouquier *et al.* (1995) and Chen-Liu *et al.* (1995) suggest that the cDNA selection procedure performed in the

present study has resulted in a procedure of efficiency intermediate between that of the two previous reports. Cytogenetic data of Chen-Liu *et al.* (1995) show that after one round of selection, the probe intensity on the chromosome of interest was greater than that on any other chromosome, but that signal was present on all chromosomes in a highly uniform manner. In comparison, data from the present study demonstrate a lesser degree of background hybridisation to other dog chromosomes, with the exception of the highly GC-rich sites on CFA 1 and CFA 9. The degree of enrichment was shown to be considerably improved by Chen-Liu *et al.* (1995) when a second round of selection was performed. Transcripts isolated from the first round were amplified by the PCR and hybridised with a second aliquot of biotinylated target genomic DNA, and the selection procedure was repeated under the same conditions as in the first round. The possibility of performing a second round of selection to increase the degree of enrichment of selected cDNAs corresponding to CFA 5 would be worthy of investigation, although it has been generally considered that the vast majority of enrichment occurs within the first round of the procedure (for example, Lovett, 1994b).

In investigating the cDNA selection approach for generation of chromosome-enriched cDNA sub-libraries, one of the most problematic issues was the wide variation in parameters used in previous reports of the method. This introduced a large number of variables in attempting to optimise reaction conditions that could not be fully investigated in the time available. As discussed in chapter one, cDNA selection has primarily been used as a method to isolate cDNA sequences using biotinylated single locus probes as the genomic template. The use of a chromosome paint probe as the target has been considerably less well documented. The approach followed in this chapter was based on the report of Rouquier *et al.* (1995), who used HSA 17 and HSA 19 as the genomic targets. Human chromosomes were flow-sorted and 500 copies of HSA 17 and HSA 19 were collected separately. These were amplified by DOP-PCR (Telenius *et al.*, 1992a) and biotinylated by nick translation. Human foetal brain cDNA was amplified by the PCR using vector primers, and 1μg was combined with 5μg of human C_0t1 DNA. The selection procedure was then performed as described above in section 6.5, using 100ng of the biotinylated human chromosome paint probe as the genomic target. Certain exceptions to this published method were required in this chapter, reflecting the resources available. The absence of a source of dog C_0t1 DNA required that this competitor be replaced by sonicated dog genomic DNA. This substitution is routinely made in FISH analysis for species where the C_0t1 component of the genome is not available in an

enriched form, and the replacement of this resource by a tenfold mass of sonicated genomic DNA is common (for example, as discussed in chapter four for both the dog and cat). In previous cDNA selection reports, competitor DNA has been used to suppress repetitive sequences in both the cDNA population (for example, Lovett *et al.*, 1991) and the genomic DNA target (for example, Parimoo *et al.*, 1991). Lovett (1994a) has commented on the fact that these strategies appear to work equally well. However, the use of sonicated total dog genomic DNA as the competitor in this chapter is likely to at least in part explain the decreased efficiency of the procedure, compared to that using human and mouse C_0t1 DNA by Rouquier *et al.* (1995) and Chen-Lui *et al.* (1995) respectively. The use of a whole genomic DNA competitor results in an increase in the sequence complexity and DNA concentration of the reaction mixture, which is highly likely to have an effect on hybridisation kinetics. Development of a source of dog C_0t1 DNA would therefore be a priority in further optimisation of this technique. Rouquier *et al.* (1995) do not describe denaturation of either the cDNA or the genomic DNA template prior to hybridisation. This step was included in the protocol investigated in this chapter since it would seem unlikely that efficient hybridisation would occur between the two DNA sources unless in single stranded form. It has since been confirmed that a comparable denaturation step was indeed performed in the original report of this method (S. Rouquier, pers. comm.) Thirdly, rather than elute bound cDNA molecules from the streptavidin-coated magnetic beads by alkaline denaturation, as was performed by Rouquier *et al.* (1995), beads were incorporated directly into a PCR amplification step, from which amplification was primed from cDNA sequences whilst they remained associated with the bead complex. This has been discussed as a preferable technique by the manufacturers of the magnetic beads since the efficiency of elution of bound sequences from the bead complex by alkaline denaturation is low (K. Dobson, Dynal, pers. comm.). The technique requires that the concentration of beads does not exceed 1μg/μl in the final amplification reaction volume, and so the use of 0.5μl of beads (5μg) in a 10μl volume in the protocol used in this chapter is within the recommended limit.

It is possible that the decreased efficiency of the selection procedure described in this chapter was related to the method of biotinylation of the chromosome-enriched genomic target. Chen-Liu *et al.* (1995) derived their genomic probe from ten microdissected copies of MMU 11, which were amplified by DOP-PCR and biotinylated. The method of biotinylation was not indicated. Rouquier *et al.* (1995) incorporated biotinylated nucleotides by nick translation of the secondary DOP-PCR product, using a commercial

labelling kit (BioNick, Gibco BRL). This kit is analogous to the method of nick translation described in chapter two of this study, and has been estimated to result in the incorporation of one biotin residue every twenty nucleotides (D. Nickson, Gibco BRL, pers. comm.). The biotinylated CFA 5 paint probe used in this chapter was generated by incorporation of biotinylated nucleotides within the secondary DOP-PCR itself (Langford *et al.*, 1996). This method has been suggested to result in the substitution of biotin residues at a similar rate to that of the nick translation procedure described above (N. Carter, pers. comm.). Both nick translation and PCR labelling methods have been used in previous studies of cDNA selection, although since the overwhelming majority of reports have used high molecular weight, single locus probes as the source of genomic DNA, nick translation methods have been more commonly used. The application of nick translation labelling in this instance would have to take into account the small size of the CFA 5 probe material, which in section 6.5.1 was shown to contain fragments ranging in size from 0.3kb to 1.0kb.

The nature of the hybridisation buffer used also varies greatly within previous cDNA selection reports, ranging from simple high-salt solutions to more complex buffers containing various blocking agents and those containing formamide, which allow the reaction to proceed at relatively low temperatures without reduction in specificity. The use of formamide in the hybridisation buffer is regarded as undesirable by the manufacturers of the magnetic beads, since its effects on the kinetics of the reaction, and in particular on the structure of the streptavidin molecule, have not been fully investigated (K. Dobson, Dynal, pers. comm.). For effective binding of the biotinylated target to streptavidin beads, a final concentration of 1M NaCl within the hybridisation reaction is required. The use of 6x SSC, 0.1% SDS as the hybridisation buffer by Rouquier *et al.* (1995) is therefore appropriate for this purpose, since this solution is 0.9M with respect to NaCl. Investigation of alternative hybridisation buffers might be considered appropriate for further optimisation of the selection method, but would require adjustment of the salt concentration of the final reaction mixture prior to exposure to the magnetic beads.

The duration of the hybridisation step in previously reported cDNA selection protocols is less variable, with most reports citing periods of 16-48 hours, comparable to the conditions used in many standard FISH analysis techniques. The most distinct exception is the six day hybridisation period used by Chen-Liu *et al.* (1995), which was intended to increase the potential for capturing low abundance transcripts. Extended hybridisation

periods were not investigated in this study, since the capture of rare transcripts was not an important issue. It is apparent from the resulting FISH data of Chen-Liu *et al.* (1995) that considerable levels of background hybridisation were observed on all chromosomes, with more intense signal present on that corresponding to the genomic target. It would seem likely that the elongated hybridisation period used is partly responsible for this observation. However, more detailed study would be required of the results obtained following hybridisation of varying periods in order to establish whether this is the case, and to determine whether capture of low abundance transcripts is enhanced as a consequence.

The ratio of cDNA to genomic DNA template used in the selection process described in this chapter was maintained at 1000ng:100ng, following the protocol of Rouquier *et al.* (1995), whereas Chen-Liu *et al.* (1995) used a ratio of 1000ng:10ng. The precise correlation between genomic target and transcript population is clearly dependent on the physical size of the chromosome region of interest. Rouquier *et al.* (1995) focussed on HSA 17 and HSA 19, which are generally accepted as approximately 89MB and 66MB respectively, whilst MMU 11, studied by Chen-Liu *et al.* (1995), is approximately 145MB in size (http://www.informatics.jax.org). The targets used in these previous reports are therefore grossly similar in physical size to CFA 5. In contrast, comparison with template ratios from cDNA selection protocols using single locus probes is of relatively little value since the complexity of the genomic target is reduced by several orders of magnitude compared to probes representing entire chromosomes. Optimisation of cDNA selection using a whole chromosome paint probe may, however, be aided by parallel investigation of the range of relevant experimental parameters using a single locus probe as the genomic target. The efficiency of this procedure is less easily determined since captured material may not be detectable unless the locus contains a considerable number of coding sequences and covers a physical distance that is within the resolution limits of metaphase FISH analysis. Similarly, the efficiency of selection has been shown to be inversely related to the size of the genomic target, such that selection procedures using single locus probes and genomic paint probes are highly dissimilar in many experimental respects. Using genomic targets of less than 1MB, 100,000-fold enrichment of transcripts has been achieved (Lovett, 1994b). Larger targets enable less efficient enrichment, for example, Lovett (1994a) used a 174Mb cosmid contig from HSA 5 (comprising approximately 25,000 clones) as the genomic target, demonstrating between 100 and 1000-fold enrichment for transcripts from this

chromosome. This suggests that cDNA selection using a CFA 5 chromosome paint probe as the genomic target is unlikely to enrich the transcript population by a value in excess of three to four orders of magnitude.

It is perhaps most likely that the difference in the level of efficiency of cDNA selection using the CFA 5 paint probe compared the report of Rouquier *et al.* (1995) is directly related to the relative purity of the genomic template. Comparisons between the human and dog flow karyotypes have been made previously in figure 4.4a, in which it was shown that individual peaks on the human bivariate flow karyotype are well isolated compared to those of the dog. Flow peaks corresponding to HSA 17 and 19, and to CFA 5, have been highlighted in figure 4.4a for comparison. This implies that the average molecular purity of a human chromosome paint probe is considerably greater, with less contamination from adjacent flow peaks, than can be achieved for the dog. The human chromosome paint probes used in this study have previously been estimated as greater than 95% pure (M. Ross, pers. comm.) whilst the CFA 9 paint probe, which occupied one of the most isolated positions within the dog bivariate flow karyotype, has been shown to be less than 70% pure (C. Andre, pers. comm.). It is therefore likely that the more recently developed, microdissection-based methods for cDNA selection, will increase the efficiency of this technique for the dog at the whole-chromosome level. Preparative *in situ* hybridisation (Prep-ISH) (Hozier *et al.* (1994) and microdissection-mediated cDNA capture (MMcC) (Gracia *et al.*, 1996; Gracia *et al.*, 1997) involve the hybridisation of linkered cDNA directly to metaphase chromosomes in a manner analogous to standard FISH analysis. Microdissection is then used to isolate one or more copies of the chromosome of interest, complete with associated cDNAs. The cDNA population can then be amplified by the PCR, in isolation from transcripts derived from other chromosomes. This method is particularly appropriate for species for which whole chromosome paint probes are not available, or are not of sufficient purity to enable cDNA selection by solution hybridisation to be performed efficiently. The technique may therefore prove valuable for the selection of cDNA populations derived from dog chromosomes, for which the bivariate flow karyotype indicates that molecular purity of paint probes is likely to be lower than for species such as the human. In addition, cDNA capture by microdissection will allow the individual to determine precisely the chromosome region for which corresponding cDNAs are to be selected, since the genomic target may comprise of a whole chromosome or any sub-chromosomal region as required.

Having isolated a population of chromosome-enriched cDNAs, it would be essential to perform detailed analysis to determine the precise degree of enrichment. In their cDNA selection procedures using an MMU 11 genomic template, Chen-Liu *et al.* (1995) used a panel of seven gene markers from this chromosome to assay for their presence in both the pre- and post-selected cDNA populations. Dot blots of cDNA were screened with radiolabelled gene products, demonstrating that all seven genes were present in both cDNA populations, and in the MMU 11 chromosome paint probe used as the genomic target. All seven genes were absent from negative control blots represented by chromosome-enriched genomic DNA probes from MMU 1 and MMU 2. The panel of gene markers confirmed to map to CFA 5 in chapter four of this study would be ideally suited for the incorporation of positive controls within the selection procedure, whilst those confirmed to map to other dog chromosomes could equally act as negative controls.

Cloning, sequencing and mapping of selected cDNAs remains the sole true test of the degree of chromosome enrichment, but has rarely been performed on a large-scale basis. Mancini *et al.* (1996) used a human-mouse somatic cell hybrid containing HSA 1 to select cDNAs from this chromosome. One hundred and ninety two clones were gridded, and 37 were radiolabelled and used to screen the panel of gridded clones to determine the copy number of selected sequences. This showed that 26 of the 37 clones analysed were uniquely represented amongst the panel, whilst the remaining 11 were present with a copy number of between two and four. The 37 cDNA clones were used to screen a human YAC library, from which positively hybridising YACs were identified for 19 cDNA clones, each of which was confirmed to map to HSA 1. Five further cDNA clones were mapped to HSA 1 by other means, whilst the remainder were considered to represent non-specific products of the selection procedure. Sequence data were obtained from all 37 cDNA clones, of which three were identical to existing, anonymous database entries, and the remainder detected no significant similarity to existing sequences. It was estimated that the procedure resulted in a five- to six-fold enrichment of the starting cDNA library for transcripts derived from HSA 1. Following the generation of a gridded panel of cDNA clones isolated by the cDNA selection procedure, it may be of value to screen this panel with probes representing common repeat motifs, particularly the SINE and LINE sequences discussed in chapters one and three. Although it would be anticipated that the selection procedure itself would confer a degree of normalisation and subtraction to the resulting cDNA population, the removal of positively hybridising

cDNA clones from further study may increase the efficiency of EST generation from selected transcripts. Their exclusion would limit sequencing redundancy and reduce the frequency of isolation of uninformative clones that do not enable a gene identity to be assigned, enabling concentrated effort on the identification of novel gene markers for the chromosome of interest.

In further optimisation of cDNA selection for application to the dog genome, it would seem that the most appropriate starting point would be the use of microdissection to generate a purer source of genomic DNA from the chromosome of interest. This might be carried out by the microdissection or flow-sorting of single copies of the chromosome and the amplification of this material by the PCR. It would be anticipated that this technique would result in a paint probe more highly specific for that chromosome than can be achieved by flow-sorting methods. Alternatively, the methods described by Gracia *et al.* (1996) and others may be more effective, in which the entire starting cDNA population is first hybridised to metaphase chromosomes prior to microdissection of the required chromosome region and associated cDNAs. Both techniques would be expected to increase the efficiency of cDNA selection, by enabling improved control over the generation of a chromosome-specific genomic template, or by enabling direct physical isolation of cDNAs bound to the chromosome of interest. Cloning and sequencing analysis would then be used to assess the degree of enrichment resulting from these procedures. This may be achieved by screening both the pre- and post-selected cDNA populations with a panel of gene markers from the chromosome of interest, and comparing their relative increase in frequency with a baseline level determined by similar analysis with gene markers from other genomic locations. Sequences that share significant nucleotide identity with existing database entries would then be used to design PCR primers for RH analysis, and to isolate a large-insert genomic clone suitable for metaphase FISH analysis. These novel gene markers would be used to increase the density of mapped type I markers within the dog genome, extending the existing, relatively limited, transcript map. Comparison of the chromosomal locations of each locus in the dog and human genomes would then be used to further enhance comparative mapping data between these species, refining the boundaries between adjacent ECCS and enabling assessment of relative internal gene order.

6.8 Summary

cDNA selection has been investigated as a technique for the generation of novel gene markers for CFA 5. Preliminary characterisation by PCR, sequencing and cytogenetic analysis demonstrated that the dog cDNA library to be used as the source of transcripts for cDNA selection appeared to be suitable for this purpose. FISH analysis of the starting pool of transcripts demonstrated hybridisation to dog chromosomes in a pattern consistent with proposed sites of high gene density. A probe composed of post-selection transcripts was shown to hybridise strongly and evenly to CFA 5, indicating that the method had resulted in a detectable increase in the proportion of sequences derived from CFA 5. Hybridisation signal was also observed at a subset of other chromosomal regions that correlate with GC-rich, and therefore putatively gene-rich sites, indicating that the selection process had not occurred with absolute efficiency. cDNA selection was therefore shown to be a technique of promise for the isolation and characterisation of novel gene markers for a single chromosome, and potential means for further optimisation of the methodology were proposed.

Chapter seven:

General discussion and future work

General discussion and future work

The primary aim of this study was the development and application of techniques for improving the status of the dog gene map. This was initiated in chapter three by the generation of expressed sequence tags from a dog cDNA library, which led to the identification of a panel of 17 ESTs that share significant nucleotide identity with previously annotated gene sequences from other species. This panel represents a valuable contribution to the resources available for mapping type I markers to the relatively gene-poor genome map of the dog, since at the present time none of these markers have been assigned to dog chromosomes outside this study. None appear in the most recent integrated RH and meiotic linkage map of the dog genome (Mellersh *et al.*, 2000). It will therefore be a priority to integrate these loci with the current genome map.

Perhaps more importantly, the study initiated in chapter three enabled the identification of key issues that would require consideration in the future application of this pilot study to a large scale EST project for the dog genome. As is apparent from existing literature, the quality of the starting cDNA resource is of paramount importance to the success of the approach. Data from chapter three demonstrate the existence of a considerable proportion of undesirable clones within the cDNA library used, which limited the efficiency of this study. Amongst these were clones containing repetitive elements, whose repeated isolation could be negated by prescreening the library with repeat probes prior to the selection of potentially unique clones for further sequencing analysis. Normalisation of the library is an option that should therefore be investigated, since this would reduce the level of sequencing redundancy that is inevitable in large scale EST development. The issues of chimerism of cDNA clones, and of the existence of contaminating microbial sequences, are less easily resolved, and are largely dependent on the manner in which library construction is performed. Directional cloning of insert DNA would enable selected analysis of 3' and 5' gene sequences, which would be an advantage should downstream RH mapping of clones be required. In the planning of a future dog EST project, it may therefore be appropriate to consider the generation of a custom-designed cDNA library, which would enable these issues to be more easily addressed and controlled.

The study discussed in chapter three was conducted in the form of a random cDNA clone selection procedure. Rather than focussing on a specific genomic region, the study was

performed in the same manner as for a large-scale, genome-wide EST project, without modification to the cDNA source or to existing methodologies for clone characterisation. This enabled pertinent issues regarding the techniques involved in random EST generation to be identified and considered, prior to embarking on a more comprehensive study of this form. Recent progress in dog genome mapping now enables each random locus assignment to attain relevance at the whole-genome level. For example, knowledge of the structural relationships between the human and dog genomes (Breen *et al.*, 1999a; Yang *et al.*, 1999) enables the chromosomal location of dog genes to be accurately predicted from the map position of the human orthologues. Thus, the mapping of each coding sequence to the dog genome not only has value in its own right, but assists in the confirmation and refinement of regions of conserved synteny between genomes that were previously identified by chromosome painting studies. Similarly, the development of a large insert genomic library for the dog (Li *et al.*, 1999), and of two WG-RH panels (L. McCarthy, unpublished; Vignaux *et al.*, 1999) will enable the chromosome assignment of EST sequences to be performed on a scale that is more comparable to the rate at which sequence information can be produced. Chapter three discussed issues relating to chromosomal localisation of ESTs. A review of existing literature indicated that FISH analysis of small insert cDNA clones is not efficient on a routine basis, whilst the generation of extended genomic probes using long range PCR analysis proved successful for this purpose, but was labour-intensive. However, the ability to integrate cDNA sequence data with maps generated by WG-RH mapping, particularly where aided by the advantages of a directionally cloned library, will represent a major improvement to the efficiency of this technique. The isolation of a corresponding BAC clone for each EST will enable this locus to be incorporated into the developing physical map of the dog genome, in turn providing the potential for isolation of a tightly linked, polymorphic marker. This approach would aid the continuing correlation of existing meiotic and RH linkage data with a specific chromosome assignment, whilst enhancing comparative mapping data between the dog and other species.

In contrast to the EST study, chapters four and five describe the development and application of a range of complementary techniques for the focussed mapping of a single chromosome. Amongst these, the heterologous chromosome-painting technique (Scherthan *et al.*, 1994) represents the greatest contribution towards the ability to extrapolate mapping information between species. In a series of comparative chromosome painting studies, it was possible to identify regions of HSA 1, 11, 16 and 17

that share an evolutionary relationship with CFA 5. This enabled extrapolation of existing data from the highly advanced human gene map to this dog chromosome. The validity of this extrapolation was investigated and confirmed by the assignment of gene markers derived from HSA 1, 11, 16 and 17 to their position within the dog karyotype. These assignments in turn enabled the boundaries between ECCS to be refined, and their relative orientations within the corresponding human and dog chromosomes to be proposed.

The acquisition of gene markers for the dog was largely dependent on existing resources for other species, since few dog gene sequences are publicly available. As a consequence, comparative gene mapping resources that have been designed to aid the assignment of coding sequences in a range of species were considered the most appropriate means by which to rapidly identify and chromosomally assign dog genes in the absence of dog-specific sequence data. This step proved to be the limiting factor in the efficiency of this study. Loci for which comparative mapping markers are available are relatively evenly distributed throughout the human genome, particularly the CATS primers developed by Lyons *et al.* (1997) from a panel of loci identified by O'Brien *et al.* (1993). However, the value of these resources is largely dependent on the ability of each primer pair to successfully amplify from a genomic template derived from the species of interest. The gene mapping resources described by Venta *et al.* (1996), Lyons *et al.* (1997), Jiang *et al.* (1998), Priat *et al.* (1999) and others have made extensive use of existing database entries, sequence alignment tools and computer-aided primer design in attempts to optimise the cross-species application of their resources. However, it is apparent from the data presented in chapter four that successful amplification from these gene primers is by no means a certainty. Although a relatively limited proportion of the available markers were assessed, it was possible to make a preliminary evaluation of the success rate for each of these sets of gene primers, as shown in table 4.3c. It will be interesting to see whether these values prove to be representative of the success achieved on a genome-wide mapping basis. This will also demonstrate the proportion of the total number of these markers that do not appear to represent the expected locus, such as CSF3R (Lyons *et al.*, 1997), and SLC2A4 (Priat *et al.*, 1999). In both cases, sequencing analysis of the amplified product indicated that a discrepancy existed between observed and expected results. However, physical mapping of the product to a dog chromosome, and evaluation of this information against dog-human comparative mapping information (Breen *et al.*, 1999a; Yang *et al.*, 1999) enabled an improved understanding of the nature of the

anomaly. This emphasises the need to verify the identity of each marker product by sequencing analysis prior to physical assignment, and further reinforces the value of the reciprocal Zoo-FISH technique for the dog and human genomes.

A further limitation relates to the quality of information for human gene assignments, for which a range in resolution was widely apparent in the selection of potential loci for further study, as shown in figure 4.3a. Existing cytogenetic assignments varied from the level of a specific chromosome subband (particularly in the case of HSA 17 markers) to a region that in the case of the GOT2 and ALDOA markers extended to include in excess of a third of the chromosome (HSA 16). Figure 4.3a also shows that in a number of cases, human gene assignments spanned the boundary between adjacent ECCS. The variation in the resolution of physical assignments is in part likely to reflect the relative scientific interest in specific chromosome regions, and had the effect of decreasing the efficiency of the immediate aim of selecting markers whose dog orthologue would map to CFA 5. However, the outcome was the identification of markers that map to dog chromosomes other than CFA 5, confirming and refining the boundaries of adjacent ECCS. This demonstrates that for 'gene-poor' species such as the dog, the assignment of any coding sequence is likely to be of value on a genome-wide level. The assignment of all gene sequences identified in this study will therefore be continued in order to maximise their potential contribution to existing maps of other dog chromosomes.

For future fine mapping of each CFA 5 ECCS, knowledge of the gene location and order in the appropriate human chromosome region will increase the efficiency by which loci can be selected that will map to CFA 5. Although gene order information may be available from existing RH maps of human chromosomes, an accurate and high-resolution physical assignment for at least a subset of human loci is initially required, in order to interpret this information with reference to Zoo-FISH data. The selection of loci immediately flanking markers confirmed to map to the dog ECCS of interest can then be used to extend mapping information in both directions, further refining the boundaries. Under circumstances such as those reported in this study, once the extreme ECCS boundaries have been determined, there is less need for physical assignments for human loci that are known from RH analysis to lie within these boundaries. Where existing human gene assignments are not of sufficiently high resolution, it would be advantageous to obtain a human clone for the locus of interest, refining existing mapping information prior to isolation of a dog orthologue, although routine access to resources such as these may be limited. The approach taken in performing comprehensive gene mapping within

an ECCS would therefore largely depend on the quality of the mapping data available for the region of interest. An alternative approach by which to increase the resolution of orthologous human gene assignments may result by investigating the ability to hybridise single locus dog gene probes to human chromosomes. Figure 4.3c demonstrates that dog BAC clones can be successfully hybridised to cat metaphases by FISH analysis, and optimisation of this method may enable this analysis to be performed onto human chromosomes.

Further development of the dog-human comparative map is dependent on an increase in the density of gene markers mapped to the dog genome. High resolution comparative gene mapping will be required in order to identify small regions of conserved synteny between species that are not apparent from gross chromosome painting analysis. For example, in chapter four, the mapping of the ALDOA marker (Priat *et al.*, 1999), located on HSA 16q22-q24, to CFA 6q14, identified a putative region of conserved synteny between these chromosomes that was not detected by chromosome painting analysis. All assignments generated by FISH analysis were derived from analysis of chromosomes from at least two unrelated individuals, eliminating the possibility that the anomaly surrounding ALDOA was a result of a unique chromosome rearrangement in one karyotype. This finding clearly warrants further investigation, initially by the physical mapping of the dog orthologue of genes flanking the human ALDOA locus, to determine the extent of this putative ECCS. High resolution mapping would ultimately enable identification of the precise site at which deviation from the Zoo-FISH data occurs. This will be of interest in exploring the nature of the relative chromosome rearrangements that have occurred in the evolution of the human and dog karyotypes from their common ancestor.

Eleven type I markers assigned to CFA 5 in this study were subjected to WG-RH mapping analysis, generating a preliminary map comprising six linkage groups spanning the chromosome. The cytogenetic and RH maps were anchored by the analysis of all markers by both methods, with the exception of the DPEP1 locus, which could not be optimised for RH analysis. In addition, ten anonymous cosmid clones from CFA 5 were assigned by both FISH and RH analysis. Since each clone is known to contain a microsatellite repeat sequence (N. Holmes, pers. comm.), genotyping analysis of each marker across the dog reference panel could be used to rapidly and efficiently link the RH and cytogenetic maps to the meiotic linkage map. Efforts towards this aim have been made by genotyping at the MMP1 and K315 loci (appendix E), which has enabled

preliminary integration of the RH and cytogenetic map of CFA 5 with the current meiotic linkage map.

The enhancement of the CFA 5 RH map by the conversion of the six separate linkage groups into a single group would rely on the availability of markers at the boundaries of the existing groups, where physical gaps exist between mapped markers. The acquisition of markers within these locations could be achieved by any of a number of techniques. Large numbers of supplementary type II markers could be rapidly isolated by the production of a CFA 5-enriched genomic library, which could be achieved by the cloning of flow-sorted or microdissected chromosome material followed by mapping of randomly selected, large-insert clones by FISH analysis. Prescreening of the library with a tandem nucleotide repeat probe could be used to simultaneously identify markers suitable for meiotic linkage mapping. Those that map close to the region of interest on CFA 5 could then be used in RH mapping via the generation of end-sequence data and the design of PCR primers, following the technique used in chapter five for the ten framework cosmid clones. A further 18 anonymous, microsatellite-containing cosmid clones have been mapped to CFA 5 (M. Breen, pers. comm.), however, the ten selected for RH analysis in chapter five represent those that combine to result in the most even and complete coverage of the chromosome possible in the time available.

As this project was focussed on the addition of gene markers to the dog genome map, it would be more appropriate to identify more coding sequences for extension of the study. All currently available cross-species primer resources for CFA 5 have been exhausted, therefore this option is not be available unless further examples are developed in the near future. Human genome mapping resources may be a worthy area of investigation in this respect. Of particular value may be the extensive panels of PCR primer sets designed to cloned human cDNA sequences, such as those shown by Lahbib-Mansais *et al.* (1999) to successfully amplify from non-human genomic templates in 10% of instances. However, it would seem preferable to focus on the production of additional gene sequence data specifically for the dog, which may be generated by the EST approach. Isolation and physical mapping of a corresponding dog BAC clone could then be used to rapidly increase the density of mapped type I markers. End-sequence data from each BAC clone could serve as a resource for addition of the marker to the WG-RH map, through the design of appropriate marker primers. The efficiency of this approach could be increased by the screening of the dog BAC library filters with pools of probes representing a number of different loci. Data in chapter four were obtained using pools of probes for up

to four different gene markers, and theoretically it should be possible to increase this number significantly. In practice the hybridisation protocol is likely to require adaptation to allow for the increased probe complexity and the consequent increase in background signal generation. Alternatively, the conversion of the BAC library to a panel of pooled BAC DNA populations that could be assayed by the PCR would increase the rate of clone identification, since it would eliminate the need for hybridisation screening. The development of such a resource has recently been announced as a future plan of the international dog genome mapping community (M. Breen, pers. comm.) Limitations on the physical mapping of more than two or three clones in a single hybridisation experiment will be lifted by the advances in FISH technology that have arisen within the last few years. Of particular value is availability of a range of fluorochromes for differential labelling and efficient detection of considerably larger numbers of probes simultaneously than is possible with three-colour systems such as that used in this study, in which one colour is required for counterstaining. The rate of locus assignment to the dog genome is therefore likely to demonstrate a dramatic increase in the near future.

Large-scale random EST generation specific for any given species has the potential to rapidly yield a considerable number of gene markers, but this is more appropriate on a genome-wide level than for the study of a specific chromosome. Chapter six discussed the assessment of a technique by which a cDNA library can be enriched for transcripts derived from a single chromosome, and issues to be addressed in the further optimisation of this technique were highlighted. cDNA selection may then represent a valuable method by which to enable the randomised EST approach discussed in chapter three to be performed in a focussed manner, by the generation and characterisation of chromosome-specific cDNA libraries. Technological advances now permit one individual to generate sequence data from several hundred templates in a single 24 hour period. Similarly, the introduction of mapping techniques such as WG-RH mapping and multicolour FISH analysis enable gene assignment through the characterisation of chromosome-specific EST libraries to be performed on an exceedingly efficient scale. Furthermore, as a result of the chromosome painting studies of Breen *et al.* (1999a) and Yang *et al.* (1999), the dog genome mapping community is now in a position to consider in detail the relative chromosome structure within corresponding ECCS in the dog and human genomes. This will be of paramount importance in the isolation of candidate genes for inherited traits where an approximate chromosomal location has been established by meiotic linkage studies, or by identification of chromosome aberrations,

amongst other techniques. The combination of cDNA selection with microdissection enables the generation of a genomic template that is more pure at the molecular level than is possible using a chromosome paint probe generated from flow-sorted material. Furthermore, this combined approach would confer the ability to generate a genomic probe for selection of transcripts from subchromosomal regions. cDNA selection using the corresponding genomic target would enable the isolation of gene transcripts potentially involved in the trait of interest, which is of particular value under circumstances where a suitable candidate cannot be inferred from human mapping data.

In theory, all of the techniques described in chapters four and five, and potentially chapter six, are applicable to any dog autosome, and indeed to those of any species for which the appropriate cytogenetic and molecular mapping resources are available. It has been shown by Breen *et al.* (1999a) that Zoo-FISH analysis is only moderately limited for those paint probes that contain material for more than one chromosome, and that all the dog paint probes generated by Langford *et al.* (1996) are of sufficiently high quality for conclusive Zoo-FISH data to be obtained. However, the decision to focus on a single dog chromosome has enabled different resources and techniques to be evaluated in a more comprehensive manner than would have been possible from a whole-genome analysis in the time available. Chromosome painting data for CFA 5 have since been confirmed to be amongst the highest quality for all members of the dog karyotype (M. Breen, pers. comm.), and identified regions of four human chromosomes that are evolutionary related to CFA 5. Existing mapping data for these human chromosome regions can be examined to identify additional genes and regions of particular interest to the dog genome mapping community that are also likely to be located on CFA 5, and that may be worthy of further studies. For example, HSA 11q23 lies within an ECCS that corresponds to CFA 5, and is associated with both balanced and unbalanced chromosomal aberrations in a considerable number of malignancies (reviewed in Mitelman *et al.*, 1997). In particular, HSA 11q23 has been widely studied as a fragile site frequently associated with haematological malignancies (for example, Yunis *et al.*, 1989; Rowley *et al.*, 1990). In the former study, cases of leukaemia and lymphoma demonstrated chromosome rearrangements involving a breakpoint at HSA 11q23 and fusion with another autosome, or the X chromosome. A similar fragile site exists within the corresponding ECCS on MMU 9 that contains the THY1 and CD3E genes (Antonucci *et al.*, 1984; Sanz *et al.*, 1986; Searle *et al.*, 1987, cited in Yunis *et al.*, 1989). The corresponding ECCS in the dog was located to CFA 5q12-q14 in chapter four, the

region to which both the CD3E and THY1 genes were subsequently mapped. Similarly, abnormalities of tumour cell karyotypes, most commonly those found in haematological malignancies, have also been described within HSA 16q21-q24 (reviewed in Mitelman *et al.*, 1997), a region that shares conserved synteny with CFA 5q33-q36. Deletions of regions of HSA 17p, demonstrated in this study as evolutionarily related to CFA 5q21-q23, have also been identified in a range of neoplasms (reviewed in Mitelman *et al.*, 1997). The HSA 17/CFA 5 ECCS is of particular clinical interest due to the assignment of the tumour suppresser gene TP53 to HSA 17p13 (Isobe *et al.*, 1986; McBride *et al.*, 1986; Miller *et al.*, 1986) and subsequently of the dog orthologue to CFA 5q21 in this study. CFA 5 must therefore be considered as a strong candidate for involvement in chromosome rearrangements associated with neoplasia. It would be interesting to determine whether the dog karyotype demonstrates structural abnormalities at the corresponding sites within CFA 5 compared to those aberrations seen in human haematological malignancies.

Data generated in this study indicate that CFA 5q33-q34 is the most likely location of the dog melanocortin 1 receptor gene (MC1R), since the human orthologue has been assigned to HSA 16q24.3 (Gantz *et al.*, 1994). This locus is of interest to the dog genome mapping community since the gene product is responsible for determining the pigmentation of the coat in this and other mammalian species. Variation in the gene sequence, and its relationship to coat colour in several domestic dog breeds, has recently been studied (Newton *et al.*, 2000), and is clearly of interest to dog breeders with respect to the control of this genetic trait. The isolation and physical mapping of a BAC clone containing the dog MC1R gene is currently in progress.

HSA 17p is also the location of several loci involved in human retinal disorders (reviewed in Joshi *et al.* (1997). These include an autosomal dominant form of retinitis pigmentosa, Leber's congenital amaurosis (an early onset, autosomal recessive visual disorder), and an autosomal dominant progressive cone degeneration, with a varied period of onset. In each case, the causative locus has been mapped to HSA 17p by meiotic linkage analysis, and for some the mutation involved has been established. Several retinal proteins are encoded on HSA 17p, including the retinal guanylate cyclase E (GUCY2E) mapped to CFA 5 in chapter four. Visual disorders, particularly those involving the retina, are of considerable interest of the dog mapping community. In the search for causative gene defects, concentrated mapping at CFA 5q21-q23, the region sharing an evolutionary origin with HSA 17p, may therefore be appropriate.

Conclusions

The primary aim of this study was the development and application of techniques for improving the status of the gene map for the dog. The generation of EST sequences from a dog cDNA library has led to the isolation of 17 partial gene sequences that correspond to loci mapped to 11 different human autosomes and the X chromosome. This panel of markers will represent a useful contribution towards the mapping of coding sequences within the dog genome, providing a 13.5% increase in the number of publicly available dog ESTs. The majority also have direct value in the confirmation and refinement of the boundaries of ECCS between the human and dog genomes, and in the determination of their relative orientation in both genomes. This study also identified issues that should be addressed prior to the commencement of a large-scale EST generation effort for the dog, and should therefore prove valuable in optimising the efficiency and accuracy of this technique in a future study.

Focussed comparative mapping has been performed between CFA 5 and its counterparts in the human genome, using a heterologous chromosome painting technique. These data have been extended and refined by the mapping of 12 gene markers to CFA 5. Five further markers have been assigned to other dog autosomes, confirming the presence of additional ECCS within the human and dog genomes, identifying one exception to conservation of synteny (ALDOA), and assigning a disease locus to its chromosomal location (RPE65). The assignment of 12 gene loci to CFA 5 places this chromosome second only to CFA 9 in terms of the density of RH-mapped type I markers, since 15 genes have been mapped by this technique to CFA 9 (Mellersh *et al.*, 2000). The present study represents the most comprehensive in terms of integrated cytogenetic and radiation hybrid mapping of a specific dog autosome. The assignment of ten type II cosmid clones to CFA 5 by FISH analysis and RH mapping provides a basis for integration of these data with meiotic linkage maps, via the isolation and genotyping analysis of a polymorphic repeat sequence from each cosmid clone. This approach has been demonstrated in this study for one anonymous marker (K315) and one gene marker (MMP1).

cDNA selection has been investigated as a technique by which to develop additional gene mapping resources for CFA 5. The procedure was demonstrated to result in a distinct degree of enrichment of a whole-genome cDNA library for transcripts derived from this chromosome. Further optimisation of this process is required in order for the technique to be efficiently applied for the required aim, and potential areas of future investigation have been discussed by which to achieve these aims.

This study has generated an integrated map of CFA 5, focussing on the correlation of these data with the more comprehensive human genome map by taking a comparative mapping approach. In addition, the issues raised and investigated by this study, in combination with the ongoing advancement of the dog genome map, will serve to aid the development of the dog as a valuable system for the study of inherited traits, that will undoubtedly continue to benefit both man and dog alike.

Appendices

Appendix A:
Reagent composition

Unless otherwise indicated, chemicals were obtained from Sigma, Gibco BRL or BDH, and were of analytical grade, with the exception of culture media. All solutions were prepared using double-distilled, deionised MQ water (Millipore). Reagents were sterilised by autoclaving for 15 minutes at 121°C, 15lbs/inch^2, or by passage through a 0.2μm syringe filter, where indicated below. All reagents were stored at room temperature unless otherwise stated.

- **CULTURE MEDIA**

All culture media were sterilised by autoclave, and were stored at room temperature.

LB medium

10g/l bactotryptone, 5g/l bactoyeast extract, 10g/l NaCl (plus 15g/l agar where appropriate)

LB medium containing maltose and $MgSO_4$

As above, with the addition of 2.4g/l $MgSO_4$, and with maltose (filter sterilised) added to a final concentration of 0.2% (w/v) subsequent to autoclaving

LB top agarose

As LB medium containing maltose and $MgSO_4$, with the addition of 15g/l low melting point agarose in place of agar

- **ANTIBIOTICS**

All antibiotic solutions were filter sterilised and stored at -20°C.

Chloramphenicol

20mg/ml working stock prepared in absolute ethanol, used at 20μg/ml final concentration

Kanamycin

25mg/ml working stock prepared in MQ water, used at 50μg/ml final concentration

Tetracycline

12.5mg/ml working stock prepared in 50% (v/v) ethanol, used at 12.5μg/ml final concentration

- **FISH ANALYSIS REAGENTS**

FISH analysis solutions were not sterilised prior to use, and were stored at -20°C unless otherwise stated.

2x hybridisation buffer

4x SSC, 0.2% (v/v) Tween 20, 2% (w/v) dextran sulphate in MQ water. Stored at 4°C

10x nick translation buffer

500mM Tris-HCl pH8.0, 50mM $MgCl_2$, 100mM β-mercaptoethanol and 100μg/ml BSA (fraction V)

Chromosome denaturing solution

50ml formamide, 7.1ml 20x SSC, made up to 70ml with MQ water. Prepared freshly on the day of use, and maintained at room temperature until required

DNase I working stock

4μl stock DNase I (Gibco BRL, amplification grade, 1U/μl), 100μl sterile glycerol, 20μl 10x nick translation buffer, in a total volume of 200μl. Prepared freshly each month

dNTP mix for labelling by nick translation

0.2mM dATP, dCTP and dGTP, 0.1mM dTTP (Pharmacia) and 0.1mM of either biotin-16-dUTP (Boehringer) or digoxigenin-11-dUTP (Boehringer) as required

Fixative

3:1 methanol:glacial acetic acid. Made no more than one hour prior to use, and maintained on ice

- **MISCELLANEOUS BUFFERS**

All buffers were sterilised by autoclave and stored at room temperature unless otherwise stated.

Acrylamide gel loading buffer

50mM EDTA, 50mg/ml dextran blue, plus five volumes of deionised formamide. Non-sterile, stored at -20°C

Agarose gel loading buffer (6x)

334mg/ml sucrose, 3.5mg/ml Orange G (Sigma) in MQ water. Non-sterile, stored at 4°C

GET buffer

50mM glucose, 10mM EDTA and 25mM Tris pH 8.0. Filter sterilised

Hybridisation buffer for radioactive screening (modified from Church and Gilbert, 1984)

1mM EDTA, 0.5M Na_2HPO_4 pH 7.2, 70g/l SDS, 10g/l BSA (fraction V) in MQ water. Adjusted to pH 7.2 with glacial acetic acid, non-sterile

LEI-PCR buffer (A. Jeffreys, pers. comm.)

45mM Tris-HCl pH8, 11mM ammonium sulphate, 4.5mM $MgCl_2$, 6.7mM 2-mercaptoethanol, 4.4μM EDTA pH8, 113μg/ml bovine serum albumin and 1mM each of dATP, dCTP, dGTP and dTTP. Filter sterilised, and stored at -20°C

2x magnetic bead binding buffer

10mM Tris HCl pH7.5, 1mM EDTA, 2M NaCl. Filter sterilised

Nucleon Reagent A

10mM Tris HCl, 320mM sucrose, 5mM $MgCl_2$, 1% Triton X100. Adjusted to pH8 with NaOH

5x PCR optimisation buffer

Buffers were prepared by the addition of $MgCl_2$ to Perkin Elmer PCR buffer II as indicated below. Buffers were diluted fivefold to generate the required final $MgCl_2$ concentration in the PCR,which is indicated in the first column.

Final $MgCl_2$ concentration (mM)	Quantity of Perkin Elmer buffer II (µl)	Quantity of 100mM $MgCl_2$ (µl)	Quantity of MQ water (µl)
1.0	100	10	90
1.25	100	12.5	87.5
1.5	100	15	85
2.0	100	20	80
3.0	100	30	70
4.0	100	40	60

3M sodium acetate

3M anhydrous sodium acetate (24.6g $C_2H_3O_2Na$ per 100ml MQ water), adjusted to pH 5.2 with glacial acetic acid. Filter sterilised and stored at 4°C.

10x SM buffer

1M NaCl, 0.1M $MgSO_4.7H_2O$, 0.35M Tris-HCl (pH 7.5), autoclaved and stored at 4°C. Prior to use, diluted to a 1x working concentration and 0.01% (v/v) sterile-filtered gelatin added. Stored at 4°C.

20x SSC:

3M NaCl, 0.3M tri-sodium citrate.

50x TAE

2M Tris-base ($C_4H_{11}NO_3$), 0.1M disodium EDTA, 57.1ml/l glacial acetic acid

10x TBE

890mM Tris-base, 890mM boric acid, 20mM disodium EDTA. Adjusted to pH 8.3 with HCl or NaOH as required

1x TNE

100mM NaCl, 10mM Tris-HCl (pH 8), 1mM EDTA (pH 8).

- **RH MAPPING REAGENTS**

RH mapping reagents were stored at 4°C, and were filter sterilised prior to use.

10x modified Cox buffer

450mM Tris-HCl, pH8.8, 0.05mM Na_2EDTA, 15mM $MgCl_2$, 100mM $(NH_4)_2SO_4$

420µg/ml cresol red (sodium salt)

PCR dilution buffer

0.1mM Tris-base pH8.8, 0.01mM Na_2EDTA, 2mM NaOH, 4.5µg/ml cresol red (sodium salt)

Appendix B:

Primer sequences

Locus name	Primer name	Primer sequence (5' to 3')	Original source
ACADM	ACADM F	GTTACCAGAGAGCAGCCTGG	Lyons *et al.* (1997)
ACADM	ACADM R	CTGATAGATCTTGGCGTCCC	Lyons *et al.* (1997)
ALB	077 Fa	CCACAAGCCTTTGGCACAAT	This study
ALB	077 Ra	GCTGCGCAGCACTTCTCTAC	This study
ALB	077 Fb	GCCTGTGCAGCTGAAGAGTC	This study
ALB	077 Rb	CAGGACATTCTCTCTGATTCAG	This study
ALB	077 Fc	GAGAGATTCAAGTGTGCCAGC	This study
ALB	077 Rc	CTCTAGTGTGGCTTCATATTCC	This study
ALB	077 Rd	TTCTGAGGCTCATCCACAAGAG	This study
ALDOA	ALDOA F	TGAGGCAACTGTTCTGGATCCTC	Priat *et al.* (1999)
ALDOA	ALDOA R	AGCCAAGATCTGATCTCCATGACAG	Priat *et al.* (1999)
APOA1	APOA1 F	CTGAAACTCCTGGACAACTGGGA	Jiang *et al.* (1998)
APOA1	APOA1 R	GCGTTCAGCTTCTTGGAGGC	Jiang *et al.* (1998)
APOC3	APOC3 F	CAGGAACAGAGGTGCCATGC	Venta *et al.* (1996)
APOC3	APOC3 R	TGCGCCACCTGGGACTCCTG	Venta *et al.* (1996)
APRT	APRT F	CCCAGGCGTGGTATTCAG	Lyons *et al.* (1997)
APRT	APRT R	TGCGATGTAGTCGATGCG	Lyons *et al.* (1997)
APRT	APRT F	GACTCCCGAGGCTTCCTCTTTG	Venta *et al.* (1996)
APRT	APRT R	ATACTGCAGGAGAGAGAAGAA	Venta *et al.* (1996)
ASPA	ASPA F	TTCTTCACTGCCAACAGCC	Lyons *et al.* (1997)
ASPA	ASPA R	GCTGCTTTTCTGCCGATC	Lyons *et al.* (1997)
CD3D	CD3D F	GACAAAGAATCTACCGTGCAAG	Lyons *et al.* (1997)
CD3D	CD3D R	GCTTGTGTGTCGGCAGC	Lyons *et al.* (1997)
CD3E	CD3E F	TCCAGGTCCTGCTACTCCCAGTAC	Priat *et al.* (1999)
CD3E	CD3E R	CGGTGAGTGGCCATGATGAG	Priat *et al.* (1999)
CTRB	CTRB F	AACGACATCACCCTGCTGTT	Venta *et al.* (1996)
CTRB	CTRB R	TGCAGGAGGAGACGCCACT	Venta *et al.* (1996)
CTRB1	CTRB1 F	CCTCGTCTGCCAGAAAGATG	Lyons *et al.* (1997)
CTRB1	CTRB1 R	TGGCTTCCAAGATCTGCTG	Lyons *et al.* (1997)
CYP2E	122 F	TCAGGAGCCCGTATCTCAGG	This study
CYP2E	122 R	GAATATTCTGCAGGTGGATATC	This study
DIO1	DIO1 F	ACAGGAGGGCTCCTCAAGTCCT	Priat *et al.* (1999)
DIO1	DIO1 R	CCCCAGCAGTATCCAGTGGG	Priat *et al.* (1999)
DPEP1	RDP F	GTGAAGGAGATGAACCGGCTG	Jiang *et al.* (1998)
DPEP1	RDP R	ACGTCATCCGGCACGTTG	Jiang *et al.* (1998)
DRD2	DRD2 F	GAGTGGAAATTCAGCAGGATTC	Lyons *et al.* (1997)
DRD2	DRD2 R	AGGAGATGGTGAAGGACAGG	Lyons *et al.* (1997)
ETS1	ETS1 F	AGCCGACTCTCACCATCATC	Lyons *et al.* (1997)
ETS1	ETS1 R	GATATCCCCCACAAAGTCTGG	Lyons *et al.* (1997)
FGB	053 F	CAGTGGAGGATGGACAGTGA	This study
FGB	053 R	CATTGCCAGCTGTTCCCTTA	This study
FGR	FGR F	AAGGACGCCTGGGAGATC	Lyons *et al.* (1997)
FGR	FGR R	GCTTCAGCGTCTTCACCG	Lyons *et al.* (1997)
FOS	FOS F	CCGTCAAGAGCATCGGCAG	This study
FOS	FOS R	CCTCAGGGTAGGTGAAGACT	This study
GOT2	GOT2 F	TTTAAGTTCAGCCGAGATGT	Venta *et al.* (1996)
GOT2	GOT2 R	CTTGGTAGGCCATGTCAAA	Venta *et al.* (1996)

Locus name	Primer name	Primer sequence (5' to 3')	Original source
GP1BA	gplb F	AGGCCTTAAGCACCTTTTCTG	Priat *et al.* (1999)
GP1BA	gplb R	CTCGTCTTGGTGCATCTCTTC	Priat *et al.* (1999)
GUCY2E	GUCY2E F	GTGATGCACTCGGTGCTG	A. Curson (pers. comm.)
GUCY2E	GUCY2E R	TAGTGGAGCGTGTCGAAGG	A. Curson (pers. comm.)
GUCY2E	GUCY2E F^{RH}	AGCAGCGCTGTCTGCTGCAG	This study
GUCY2E	GUCY2E R^{RH}	CATCGTGGGCCCTCCTCAGC	This study
H006	H006 F	AAAGCCCATGTAACAGCACC	This study
H006	H006 R	ACCAGGAGAGAAAAGTCAGTA	This study
H045	H045 F	GCCCCACTTGGTCCTAAGAAT	This study
H045	H045 R	AGAAACACGGGTTCAAAGAATG	This study
H068	H068 F	TGCTTATTCCTATGATGCTGC	This study
H068	H068 R	CTTTGTAAAGGGTGGGTTGAT	This study
H088	H088 F	ACCATAGGGAGACCCATTCAT	This study
H088	H088 R	TGCCATCCAGCACAGATAAGT	This study
H166	H166 F	GTGAATGTGAATGTGAATGCG	This study
H166	H166 R	ATGGCACCCAGTGATAGAGG	This study
H201	H201 F	CACTGCCCAGTTAGACCAGAC	This study
H201	H201 R	ACTTTTGCTTTAGAAGGCTCC	This study
H218	H218 F	GTCTGAATGAACTAAGAGCCACG	This study
H218	H218 R	GAGATGAATGTGATCGGGCCAAG	This study
H250	H250 F	GTGCACAGTAGTTCTGGGTGAG	This study
H250	H250 R	TGGCATGCTTGTTAGTACCCTTG	This study
H270	H270 F	AAGTGGGCTTCTATCCTCTGC	This study
H270	H270 R	CATGTGTGCTGAATAAGGTTC	This study
HP	HP F	CAGCGGATCATGGGTGGATCAG	This study
HP	HP R	CTTTTGCATCATCCTTGTGGCC	This study
HNRPA2B1	042 F	GGACTTCCTGCATTTCTTCTC	This study
HNRPA2B1	042 R	GCTGAAGCGACTGAGTCCG	This study
JUN	JUN F	CAGCGGAGCATTACCTCATC	Lyons *et al.* (1997)
JUN	JUN R	CCTGTGCGAACTGGTATGAG	Lyons *et al.* (1997)
K315	K315 F	CTCCTGCACCTCAACACAGATG	This study
K315	K315 R	CTGTTCCCTCTGCTTGCAATG	This study
λgt10	λgt10 F	AGCAAGTTCAGCCTGGTTAAC	Clontech
λgt10	λgt10 R	TTATGAGTATTTCTTCCAGGG	Clontech
MMP1	MMP1 F	ATGAAGCAGCCCAGATGTG	Jiang *et al.* (1998)
MMP1	MMP1 R	CTCTGAAATTTTTGGGTCCACCTT	Jiang *et al.* (1998)
MYH2	MYH2 F	GAACACCAGCCTCATCAACC	Lyons *et al.* (1997)
MYH2	MYH2 R	TGGTGTCCTGCTCCTTCTTC	Lyons *et al.* (1997)
NFIA	NFIA F	CCTACACCTGGTTCAACCTGCAG	Jiang *et al.* (1998)
NFIA	NFIA R	ATGACCAGGTCCAGCCGC	Jiang *et al.* (1998)
NPPA	NPPA F	CGAAGATAACAGCCAGGGAGG	Lyons *et al.* (1997)
NPPA	NPPA R	GTCCGTGGTGCTGAAGTTTATTC	Lyons *et al.* (1997)
NPPA	PND F	GCAGACCTGCTGGATTTCAAG	Venta *et al.* (1996)
NPPA	PND R	CAGTCCGCTCTGGGCTCCAAT	Venta *et al.* (1996)
NPPA	NPPA F(j)	GCCCAGAGAGATGGGGGTG	Jiang *et al.* (1998)
NPPA	NPPA R(j)	CGGAAGCTGTTACAGCCCAGTC	Jiang *et al.* (1998)
pBluescript SK-	$T3^{SK}$	AATTAACCCTCACTAAAGGG	Stratagene
pBluescript SK-	$T7^{SK}$	GTAATACGACTCACTATAGGGC	Stratagene
POLR2A	POLR2A F	TGGAGAGGAGATGGACAACAAG	Lyons *et al.* (1997)
POLR2A	POLR2A R	TCCTCATCTGAGATGCGTTTG	Lyons *et al.* (1997)
pWE15	$T3^{pWE15}$	AATTAACCCTCACTAAAGGG	Stratagene
pWE15	$T7^{pWE15}$	AATACGACTCACTATAGGG	Stratagene

Locus name	Primer name	Primer sequence (5' to 3')	Original source
RPE65	RPE65 F	TACCACCTGTTTGATGGGC	Lyons *et al.* (1997)
RPE65	RPE65 R	TATTCTTGCAGGGATCTGGG	Lyons *et al.* (1997)
RPE65	RPE-1	CAATGCCCTTGTTAATGTCTACCCAG	Aguirre *et al.* (1998)
RPE65	RPE-3	CCT GCTTAATTGTCTCCAAGGTCTCFA	Aguirre *et al.* (1998)
SLC2A4	SLC2A4 F(w)	CCGAGATCGAATCCCACTT	Werner *et al.* (1997)
SLC2A4	SLC2A4 R(w)	GGCTGCCCCTTATTTTTAT	Werner *et al.* (1997)
SLC2A4	F242	GATGACCATAGCCCTGCTTCT	This study
SLC2A4	F168	TTGTGGCCTTCTTTGAGATTG	This study
SLC2A4	R369	GTTGACCTGTTCGTTGAAGTAG	This study
SLC2A4	R398	CCAAAGGTCATATAGCGCCT	This study
SLC2A4	R485	GATTTTCACGGACTTTGGGC	This study
SLC2A4	SLC2A4 F(p)	TTGGTGTGACTGTGGCTTGA	Priat *et al.* (1999)
SLC2A4	SLC2A4 R(p)	ATGAGGATGCCAACAACGAT	Priat *et al.* (1999)
TAT	TAT F	CCCGGAGTCACACAGGAG	Priat *et al.* (1999)
TAT	TAT R	GTGCCCCAGCAAATGTATTAAT	Priat *et al.* (1999)
THY1	THY1 F	CCCATCCAGCATGAGTTCAG	Lyons *et al.* (1997)
THY1	THY1 R	GAACCAGCAGGCTTATGCC	Lyons *et al.* (1997)
THY1^{RH}	THY1^{RH} F	TACCCTTGAGCTGGGCTGGAG	This study
THY1^{RH}	THY1^{RH} R	CTGCTCTAGTAGAACCAACTA	This study
TP53	TP53 F (c)	GTGTAACAGTTCCTGCATGGG	Lyons *et al.* (1997)
TP53	TP53 R (c)	CACGCCCACGGATCTG	Lyons *et al.* (1997)
TP53	TP53 F(u)	TACAAGCAGTCACAGCACAT	Venta *et al.* (1996)
TP53	TP53 R(u)	TCTTCCAGTGTGATGATGGT	Venta *et al.* (1996)
TP53	TP53 F(p)	TCCGAGTGGAAGGAAATTTG	Priat *et al.* (1999)
TP53	TP53 R(p)	CATGCAGGAACTGTTACACATGTA	Priat *et al.* (1999)
TP53	TP53 Fλ	CACATGTAGTTGTAGTGG	M. Binns (pers. comm.)
TP53	TP53 Rλ	CCGAGTGGAAGGAAATTtGCG	M. Binns (pers. comm.)
UBB	UBB F	AGACCATCACCCTGGAAGTG	Jiang *et al.* (1998)
UBB	UBB R	GACTCCTTCTGGATGTTGTAGTCA	Jiang *et al.* (1998)
VCAM1	VCAM1 F(p)	CCTCAAAAGTGTAGCTAATGTTTG	Priat *et al.* (1999)
VCAM1	VCAM1 R(p)	CCTATGGCATAAATTTATTTTACA	Priat *et al.* (1999)

Appendix C:
Database similarity search results

The following two tables summarise the results obtained by searching the non-redundant database (table Ci) and the EST database (table Cii) with the cDNA sequences discussed in chapter three. Database searches were performed as described in section 2.1.9.
Clones are listed in numerical order. The insert size of each clone analysed is followed by a description of the most significant match reported by each database search, using the forward sequence data (upper row) and reverse data (lower row). The extent of the region of match is then provided, with the number of identical nucleotides shown as a proportion of the total length of the alignment. The probability value (P-value) reported for the match is shown in the final column.

Table Ci: BLAST search results for dog EST data against the non-redundant sequence database

Clone	Insert size (bp)	BLAST search results	Identity	P-value
001	1300	*Saccharomyces cerevisiae* chromosome XII cosmid	476/482 (98%)	0.0
008	1100	*Homo sapiens* chromosome 17, clone hRPK.180_P_8	53/59 (89%)	2e-12
009	900	Human DNA sequence from clone 360B4 on chromosome 16	23/24 (95%)	0.004
016	1200	Human DNA sequence from clone 522J7 on chromosome 22q13.3	20/20 (100%)	0.49
017	2200	*Saccharomyces cerevisiae* chromosome V cosmids 9163 and 9132	430/459 (93%)	0.0
020	600	Human aldehyde dehydrogenase II (ALDH-2) mRNA, 3'end Human mRNA for mitochondrial aldehyde dehydrogenase	398/460 (86%) 232/279 (83%)	e-127 3e-49
022	1400	*Arabidopsis thaliana* genomic DNA, chromosome 5	20/20 (100%)	0.2
023	900	Human DNA sequence from clone 1177I5 on chromosome 22q13.1	22/23 (95%)	0.009
025	1900	*Homo sapiens* Sec23-interacting protein p125	18/18 (100%)	1.3
036	1000	*Rattus norvegicus* mRNA for ABC transporter *Mus musculus* ABC transporter mRNA	76/88 (86%) 87/104 (83%)	1e-15 8e-14
039	900	*Homo sapiens* BAC clone GS489L14 from 7p14-p12	24/25 (96%)	0.032
042	800	*Homo sapiens* heterogeneous nuclear ribonucleoprotein A2/B1 *Homo sapiens* heterogeneous nuclear ribonucleoprotein A2/B1	365/436 (83%) 334/368 (90%)	e-146 e-128
049	1000	*Canis familiaris* LINE-1 element ORF2 mRNA	289/340 (85%)	e-106

Clone	Insert size (bp)	BLAST search results	Identity	P-value
052	2000	*S.cerevisiae* chromosome XV reading frame *S.cerevisiae* chromosome XV reading frame ORF YOL046c	444/472 (94%) 465/482 (96%)	0.0 0.0
053	700	Human fibrinogen beta-chain mRNA Bovine messenger RNA for the beta-chain of fibrinogen	402/451 (89%) 369/420 (87%)	e-150 e-115
054	1500	*Homo sapiens*, clone hRPK.67_A_1 *Homo sapiens* clone BAC 72m22 chromosome 8 map 8p21	27/28 (96%) 24/24 (100%)	3e-4 0.001
055	400	*Canis familiaris* mitochondrion	295/325 (90%)	e-115
069	2000	*Homo sapiens* chromosome 17, clone hRPK.651_L_9	97/111 (87%)	1e-19
071	2500	*Canis familiaris* microsatellite DNA, ZuBeCa2 *Crocus vernus* tandem repetitive DNA sequence	58/68 (85%) 21/21 (100%)	5e-13 0.081
072	1200	*S.scrofa* mRNA for alpha-1-antitrypsin Human Z type alpha-1-antitrypsin gene	407/474 (85%) 220/263 (83%)	e-105 2e-46
074	1300	Human DNA sequence from clone 366B10 on chromosome 22q12.2-12.3	19/19 (100%)	0.26
077	1900	*Canis familiaris* mRNA for albumin *Canis familiaris* mRNA for albumin	336/349 (96%) 466/469 (99%)	e-180 0.0
079	1800	Human mRNA FOR haptoglobin alpha 1S Rabbit flavin-containing monooxygenase FMO3 mRNA	390/449 (86%) 303/385 (78%)	e-116 3e-60
081	1200	*Homo sapiens* eukaryotic translation elongation factor 1 alpha 1 *O.cuniculus* mRNA for elongation factor 1 alpha	327/374 (87%) 450/476 (94%)	e-133 0.0
083	1000	*Homo sapiens* mRNA for 16G2	214/262 (81%)	8e-49
085	1200	*Homo sapiens* chromosome 5, P1 clone 1047D6	24/25 (96%)	0.12
088	1000	*S.cerevisiae* MRP-L9 gene for YmL9 mitochondrial ribosomal protein	24/25 (96%)	0.13

Clone	Insert size (bp)	BLAST search results	Identity	P-value
121	1400	*Mus musculus* minisatellite tandem repeat DNA	19/19 (100%)	1.4
122	1200	*Sus scrofa* mRNA for Cytochrome P-450-j *Bos taurus* mRNA for cytochrome P450 (CYP2E)	217/236 (91%) 391/476 (82%)	1e-86 3e-69
123	1000	*Homo sapiens* 12 BAC RPCI11-25E2	48/53 (90%)	5e-9
130	1400	Human mRNA for nuclear p68 protein Human ribosomal DNA complete repeating unit	447/468 (95%) 423/423 (100%)	0.0 0.0
131	1400	*Potos flavus* transthyretin intron 1	47/52 (90%)	1e-8
134	1400	Rat argininosuccinate lyase gene, exon 3 Human argininosuccinate lyase (ASL) gene, exon 3	80/92 (86%) 184/209 (88%)	4e-17 3e-54
135	1800	*Streptococcus pneumoniae* penicillin-binding protein Pbp2x	21/21 (100%)	0.097
136	1300	*H.sapiens* gene for arginase exon 8 and 3'-flanking region no match reported	238/267 (89%)	1e-75
139	1900	*S.cerevisiae* chromosome II reading frame ORF YBL004w	97/123 (78%)	0.002
140	2500	*Saccharomyces cerevisiae* serine hydroxymethyltransferase gene	429/467 (91%)	0.0
143	800	*Homo sapiens* PDZ domain containing 1 (PDZK1) mRNA *Homo sapiens* PDZ domain containing 1 (PDZK1) mRNA	424/474 (89%) 164/182 (90%)	e-153 7e-55
146	1000	*M.musculus* 45S pre rRNA gene *S.cerevisiae* DNA from CEN12 region	466/484 (96%) 19/19 (100%)	0.0 1.4
148	2000	Human P13-kinase associated p85 mRNA sequence Rabbit 18S rRNA	252/299 (84%) 420/421 (99%)	1e-85 0.0
149	1400	*Rattus norvegicus* serine/threonine kinase beta-PAK mRNA *Canis familiaris* MHC DLA Class II DRB pseudogene DRB2	143/150 (95%) 28/30 (93%)	2e-65 0.004

Clone	Insert size (bp)	BLAST search results	Identity	P-value
151	1300	*Homo sapiens* SKB1 (*S. cerevisiae*) homolog (SKB1) mRNA *Homo sapiens* SKB1 (*S. cerevisiae*) homolog (SKB1) mRNA	249/264 (94%) 441/472 (93%)	e-116 0.0
154	1000	*Canis familiaris* LINE-1 element DNA	236/284 (83%)	7e-43
155	800	*Homo sapiens* chromosome 17, clone 27_J_12	23/23 (100%)	0.002
156	1700	*S.cerevisiae* chromosome X reading frame ORF YJL105w	496/502 (98%)	0.0
157	2000	*S.cerevisiae* chromosome XII reading frame ORF YLR154c	227/255 (89%)	1e-98
159	1800	Human DNA sequence from PAC 170A21 on chromosome 22q12-qter	20/20 (100%)	0.3
161	1300	*Canis familiaris* MHC DLA Class II DRB pseudogene DRB2	22/22 (100%)	0.029
162	1600	*Canis familiaris* tyrosine aminotransferase gene	29/30 (96%)	2e-10
165	1100	*Homo sapiens* genomic DNA, chromosome 21q22.1	51/58 (87%)	1e-8
167	1100	*Canis familiaris* mitochondrion, complete genome	452/455 (99%)	0.0
177	2100	*Homo sapiens* clone UWGC:djs94 from 7p14-15	48/54 (88%)	4e-7
179	1000	*Caenorhabditis elegans* cosmid C35C5	21/21 (100%)	0.003
180	1200	*B.cereus* mreb gene	23/24 (95%)	0.52
181	1900	Human DNA sequence from clone 591B8 on chromosome 1p13.1	20/20 (100%)	0.046

Clone	Insert size (bp)	BLAST search results	Identity	P-value
183	900	*C.familiaris* mitochondrial 12S rRNA gene	447/457 (97%)	0.0
184	1800	*Sus scrofa* mRNA for Cytochrome P-450-j *Bos taurus* mRNA for cytochrome P450 (CYP2E)	212/248 (85%) 321/395 (81%)	8e-55 4e-48
187	2600	*Mycobacterium tuberculosis* H37Rv complete genome	18/19 (94%)	0.54
188	1100	*Canis familiaris* chymase gene	115/127 (90%)	2e-38
189	2200	Human DNA sequence from clone 125H2 on chromosome 22q11-12	22/22 (100%)	0.021
191	1400	*Erignathus barbatus* transthyretin intron 1	53/60 (88%)	3e-8
199	1100	Dog pancreatic colipase gene	129/142 (90%)	6e-47

Table Cii: BLAST search results for dog ESTs against the EST database

Clone	Insert size (bp)	BLAST search results	Identity	P-value
001	1300	Loblolly pine C *Pinus taeda* cDNA clone 1C4A	19/19 (100%)	1.6
008	1100	*Homo sapiens* cDNA clone DKFZp434N132 3', mRNA sequence	40/46 (86%)	2e-4
009	900	*Homo sapiens* cDNA clone IMAGE:1147471	19/19 (100%)	0.24
016	1200	Soares pregnant uterus *Homo sapiens* cDNA clone 470198 5'	19/19 (100%)	1.6
017	2200	*Mus musculus* cDNA clone 5' similar to *Mus musculus* mitogen-responsive 96 kDa phosphoprotein p96	19/19 (100%)	0.83
020	600	*Homo sapiens* cDNA clone DKFZp434C015 5', mRNA sequence *Homo sapiens* cDNA 5' end similar to aldehyde dehydrogenase 1, mitochondrial	258/299 (86%) 160/188 (85%)	2e-77 6e-39
022	1400	Soares normalized *Schistosoma mansoni* cDNA 3', mRNA sequence	20/20 (100%)	0.16
023	900	Soares parathyroid tumor *Homo sapiens* cDNA clone IMAGE:1403278 3'	21/22 (95%)	0.035
025	1900	*Homo sapiens* cDNA clone B006	247/267 (92%)	e-128
036	1000	*Homo sapiens* cDNA clone IMAGE similar to gb:M24194 GUANINE NUCLEOTIDE-BINDING PROTEIN BETA SUBUNIT-LIKE PROTEIN (HUMAN); mRNA sequence *Homo sapiens* cDNA clone IMAGE:2341405 3', mRNA sequence	19/19 (100%) 18/18 (100%)	1.2 5.0
039	900	Soares_total_fetus *Homo sapiens* cDNA clone IMAGE:2043820 3' mRNA sequence	19/19 (100%)	0.48

Clone	Insert size (bp)	BLAST search results	Identity	P-value
042	800	*Homo sapiens* cDNA clone IMAGE:1909349 3' similar to gb:M29064 HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEINS A2/B1 (HUMAN); mRNA sequence	284/344 (82%)	e-107
		Homo sapiens cDNA clone IMAGE:1960276 3' similar to gb:M29064 HETEROGENEOUS NUCLEAR RIBONUCLEOPROTEINS A2/B1 (HUMAN); mRNA sequence	314/343 (91%)	e-125
049	1000	*C.familiaris* mRNA; expressed sequence tag	94/107 (87%)	1e-19
052	2000	*Homo sapiens* cDNA clone DKFZp586P2423, mRNA sequence	20/20 (100%)	0.19
		Aspergillus nidulans 24hr asexual developmental and vegetative cDNA lambda zap library Emericella nidulans cDNA clone m6f09a1 5'	58/68 (85%)	4e-06
053	700	Gall bladder I *Homo sapiens* cDNA 5' end similar to fibrinogen, B beta polypeptide	255/284 (89%)	e-95
		Gall bladder I *Homo sapiens* cDNA 5' end similar to fibrinogen, B beta polypeptide	254/291 (87%)	2e-72
054	1500	*Homo sapiens* cDNA clone IMAGE:2347470 3' similar to gb:X53741_rna1 FIBULIN-1, ISOFORM A PRECURSOR (HUMAN), mRNA sequence	20/20 (100%)	0.09
		Mus musculus cDNA clone IMAGE:918842 3' similar to gb:M64863 Mouse cytochrome P450 17-alpha hydroxylse/C17-20 lyase mRNA, (MOUSE); mRNA sequence	22/23 (95%)	1.2
055	400	Normalized rat brain, Bento Soares Rattus sp. cDNA clone RBRAS34 3' end	99/117 (84%)	6e-18
069	2000	*Homo sapiens* cDNA clone 142499 3'	70/84 (83%)	0.001
071	2500	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp121	57/70 (81%)	3e-6
		Mus musculus 18-day embryo C57BL/6J Mus musculus cDNA clone 1110020B02, mRNA sequence	22/22 (100%)	0.017
072	1200	*Homo sapiens* cDNA clone 344107 5' similar to gb:X01683 ALPHA-1-ANTITRYPSIN PRECURSOR (HUMAN)	319/383 (83%)	4e-63
		Homo sapiens cDNA clone 202160 5' similar to gb:X01683 ALPHA-1-ANTITRYPSIN PRECURSOR (HUMAN);.	220/263 (83%)	4e-47
074	1300	*Homo sapiens* cDNA clone 898103 3' similar to contains Alu repetitive element;contains element PTR7 repetitive element	21/22 (95%)	0.94

Clone	Insert size (bp)	BLAST search results	Identity	P-value
077	1900	Fetal spleen *Homo sapiens* cDNA 3' end similar to albumin *Homo sapiens* library (Yu Y) *Homo sapiens* cDNA clone HA0142, mRNA sequence	252/300 (84%) 355/405 (87%)	2e-55 e-112
079	1800	*Homo sapiens* cDNA clone IMAGE:928826 similar to gb:K01763 HAPTOGLOBIN-1 PRECURSOR (HUMAN) *Homo sapiens* cDNA clone IMAGE:1673366 3' similar to gb:M83772 DIMETHYLANILINE MONOOXYGENASE (HUMAN)	356/407 (87%) 178/226 (78%)	e-108 3e-29
081	1200	*Homo sapiens* cDNA clone DKFZp566C2246 5', mRNA sequence *Homo sapiens* cDNA clone DKFZp586H1722 5', mRNA sequence	327/374 (87%) 438/476 (92%)	e-133 0.0
083	1000	Soares fetal liver spleen 1NFLS *Homo sapiens* cDNA clone 245266 3'	122/148 (82%)	9e-25
085	1200	Knowles Solter mouse E6 5d whole embryo *Mus musculus* cDNA clone 893124 3'	23/24 (95%)	0.39
088	1000	*Saccharomyces cerevisiae* cDNA 3' end	24/25 (96%)	0.11
092	1200	Soares fetal liver spleen 1NFLS *Homo sapiens* cDNA clone 245266 3'	127/148 (85%)	2e-27
093	1500	*Homo sapiens* cDNA clone IMAGE:2464823 3', mRNA sequence *Homo sapiens* cDNA clone IMAGE:2349582 3', mRNA sequence	55/60 (91%) 58/66 (87%)	3e-13 2e-09
094	3500	*Homo sapiens* cDNA clone DKFZp434C2323 5', mRNA sequence no match reported	330/369 (89%)	e-138
095	1600	*Saccharomyces cerevisiae* cDNA 3' end	198/241 (82%)	4e-48
096	1400	Colon carcinoma (HCC) cell line *Homo sapiens* cDNA 5' end similar to translation initiation factor 4AII *Homo sapiens* cDNA clone IMAGE:2271025 3' similar to gb:D13748 EUKARYOTIC INITIATION FACTOR 4A-II (HUMAN); mRNA sequence	379/387 (97%) 436/457 (95%)	0.0 0.0
098	1200	*Homo sapiens* cDNA clone IMAGE:1687873 3' similar to L1.t2 L1 L1 repetitive element	47/53 (88%)	9e-11
100	1400	*Lycopersicon esculentum* cDNA clone cLET12E22, mRNA sequence	20/20 (100%)	0.36

Clone	Insert size (bp)	BLAST search results	Identity	P-value
103	3100	LNCAP cells I *Homo sapiens* cDNA 5' end	205/248 (82%)	9e-54
104	1000	*Sus scrofa* cDNA clone S3.T3, mRNA sequence *Homo sapiens* cDNA 5', mRNA sequence	410/454 (90%) 469/488 (96%)	0.0 0.0
105	1000	no match reported		
109	1200	*Homo sapiens* cDNA clone 46000 5'	22/23 (95%)	1.6
113	1800	*Homo sapiens* cDNA clone IMAGE:2191602 3', mRNA sequence	19/19 (100%)	1.2
114	1300	Normalized rat placenta, Bento Soares Rattus sp. cDNA clone RPLAE14 3' end	22/22 (100%)	0.24
119	1200	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp144	27/29 (93%)	0.11
121	1400	Barstead mouse myotubes MPLRB5 *Mus musculus* cDNA clone 1049737 5' similar to gb:J04173 PHOSPHOGLYCERATE MUTASE	19/19 (100%)	1.2
122	1200	*Homo sapiens* cDNA clone 85299 5' similar to gb:J02625 CYTOCHROME P450 IIE1 (HUMAN) *Mus musculus* cDNA clone 1886219 5' similar to gb:J02625 CYTOCHROME P450 IIE1 (HUMAN);	159/179 (88%) 63/74 (85%)	6e-44 4e-07
123	1000	*Homo sapiens* cDNA clone IMAGE:2322740 3', mRNA sequence	25/26 (96%)	0.019
130	1400	Soares total fetus Nb2HF8 9w *Homo sapiens* cDNA clone 773346 5' similar to gb:X52104 P68 PROTEIN (HUMAN) *Sus scrofa* cDNA clone S3.T3, mRNA sequence	437/451 (96%) 417/424 (98%)	0.0 0.0
131	1400	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp2	43/49 (87%)	2e-4

Clone	Insert size (bp)	BLAST search results	Identity	P-value
134	1400	Sugano mouse liver *Mus musculus* cDNA cloneIMAGE:1923384 5' similar to gb:J03058 ARGININOSUCCINATE LYASE (HUMAN) *Homo sapiens* cDNA clone NHTBC_cn04c01 random, mRNA sequence	74/87 (85%) 175/199 (87%)	e-11 1e-50
135	1800	*Homo sapiens* cDNA clone DKFZp434I0516 5', mRNA sequence	22/22 (100%)	0.02
136	1300	Soares fetal lung NbHL19W *Homo sapiens* cDNA clone 308439 5' similar to gb:M14502 ARGINASE (HUMAN) no match reported	228/257 (88%)	3e-67
139	1900	*Dictyostelium discoideum* cDNA clone SSI851, mRNA sequence	21/21 (100%)	0.027
140	2500	*Aspergillus nidulans* 24hr asexual developmental and vegetative cDNA lambda zap library	104/126 (82%)	5e-12
143	800	Soares fetal liver spleen 1NFLS *Homo sapiens* cDNA clone 248651 5' *Homo sapiens* cDNA clone IMAGE:2436164 3' similar to TR:O60450 O60450 PDZ DOMAIN CONTAINING-PROTEIN; mRNA sequence	422/478 (88%) 164/182 (90%)	e-132 6e-55
146	1000	Rat cochlea outer hair cells Lambda Zap Express Library *Rattus norvegicus* cDNA clone 421c 5', mRNA sequence *T.cruzi* epimastigote normalized cDNA Library *Trypanosoma cruzi* cDNA clone 171 5'	403/426 (94%) 20/20 (100%)	0.0 0.30
148	2000	Fetal heart *Homo sapiens* cDNA clone F0905 5' end similar to PI3-kinase-associated p85 protein *Rattus norvegicus* cDNA clone UI-R-E0-dd-a-09-0-UI 3' similar to X82564 *M.musculus* 45S pre rRNA gene, mRNA sequence	176/191 (92%) 366/368 (99%)	5e-63 0.0
149	1400	Ko mouse embryo 11 5dpc *Mus musculus* cDNA clone 775190 5' similar to gb:U39738 *Mus musculus* P21 activated Knowles Solter mouse blastocyst B3 *Mus musculus* cDNA clone 1110500 5'	142/150 (94%) 23/24 (95%)	3e-60 0.003
151	1300	*Homo sapiens* cDNA clone DKFZp564B0978 5', mRNA sequence Stratagene HeLa cell s3 937216 *Homo sapiens* cDNA clone 612526 3' similar to WP:C34E10.5 CE01185	250/264 (94%) 401/436 (91%)	e-119 e-160

Clone	Insert size (bp)	BLAST search results	Identity	P-value
154	1000	Soares retina N2b4HR *Homo sapiens* cDNA clone 362959 5' similar to L1.b3 L1 repetitive element	102/118 (86%)	2e-21
155	800	*Brugia malayi* adult male cDNA (SAW94NL-BmAM) *Brugia malayi* cDNA clone SWAMCA802 5'	19/19 (100%)	0.61
156	1700	no match reported		
157	2000	*Rattus norvegicus* cDNA clone UI-R-C3-sk-a-09-0-UI 3', mRNA sequence	23/25 (92%)	0.58
159	1800	Stratagene fetal retina 937202 *Homo sapiens* cDNA clone 839382 5'	19/19 (100%)	1.0
161	1300	Soares mouse mammary gland NbMMG *Mus musculus* cDNA clone IMAGE:832221 5'	23/23 (100%)	0.006
162	1600	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp121	29/30 (96%)	1e-10
165	1100	*Homo sapiens* cDNA clone IMAGE:1607869 3'	49/54 (90%)	4e-11
167	1100	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp54	213/217 (98%)	e-102
177	2100	Infant brain, Bento Soares *Homo sapiens* cDNA 3'end	22/23 (95%)	1.4
179	1000	Mouse 7.5 dpc embryo ectoplacental cone cDNA library *Mus musculus* cDNA clone C0019D10 3'	17/17 (100%)	1.4
180	1200	Beddington mouse embryonic region *Mus musculus* cDNA clone IMAGE:539753 5' similar to gb:X52634 Murine tlm oncogene for tlm protein (MOUSE); mRNA sequence	19/19 (100%)	1.7
181	1900	Soares fetal lung NbHL19W *Homo sapiens* cDNA clone 300359 5'	19/19 (100%)	0.17

Clone	Insert size (bp)	BLAST search results	Identity	P-value
183	900	Soares mouse 3NbMS *Mus musculus* cDNA clone 777535 5'	221/255 (86%)	1e-64
184	1800	Normalized rat liver, Bento Soares Rattus sp. cDNA clone RLIAB20 3' end *Mus musculus* cDNA clone 1886219 5' similar to gb:J02625 CYTOCHROME P450 IIE1 (HUMAN); gb:L11650 Mouse cytochrome P-450 IIE1 mRNA, complete cds	157/185 (84%) 63/74 (85%)	9e-37 5e-07
187	2600	Sugano mouse liver *Mus musculus* cDNA clone 1450612 3' similar to TR:O35052 O35052 CDP-DIACYLGLYCEROL SYNTHASE, mRNA sequence [*Mus musculus*]	16/16 (100%)	2.3
188	1100	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp19	113/129 (87%)	3e-28
189	2200	*Homo sapiens* cDNA clone 29284 5'	19/19 (100%)	1.2
191	1400	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp144	26/26 (100%)	1e-04
199	1100	*C.familiaris* mRNA; expressed sequence tag; clone CR-ESTp142	117/134 (87%)	2e-32

Appendix D: Genotyping analysis at the RPE65 locus in the Briard

Abstract

Chapter four discussed the identification of a region of conserved synteny between CFA 5q24-q32 and HSA 1p32.3-p31.2 by reciprocal chromosome painting analysis. The RPE65 gene has previously been assigned to HSA 1p31 (Hamel *et al.*, 1994), and therefore lies within a cytogenetic band of which only part represents an ECCS shared with CFA 5. In order to refine the boundary of this shared ECCS, a dog BAC clone representing the RPE65 gene was isolated using primers designed to the dog RPE65 gene (Aguirre *et al.*, 1998). The clone was assigned by FISH analysis to CFA 6q25. Assuming that the assignment of human RPE65 is accurate (Hamel *et al.*, 1994), the assignment of the dog orthologue to CFA 6q25 therefore confirms that the HSA 1/CFA 5 ECCS does not extend proximally to include HSA 1p31.

Although the assignment of RPE65 precluded its use as a CFA 5 marker, the locus was considered worthy of further investigation due to its clinical interest to the dog genome mapping community. A deletion in this gene has previously been shown to be the cause of a hereditary visual disorder in the Briard breed population of Sweden (Veske *et al.*, 1999). The acquisition of DNA samples from the UK population allowed a preliminary analysis of whether this mutation exists in the UK breed stock. Genotyping analysis of a small number of individuals has confirmed that the same mutation is indeed present, at a frequency indicating that routine testing on a wider scale would be of value.

Introduction

Congenital stationary night blindness (CSNB) was first described in the Briard population in Sweden (Narfström *et al.*, 1989), as an analogue of the human disorder of the same name. The disease has more recently been described as hereditary retinal dystrophy, although it had not been proved without doubt that these represent the same disorder. Resulting effects of CSNB on vision vary from near-normal sight to complete day blindness. Visual impairment can be detected once the individual reaches approximately 5-6 weeks of age. Clinical and electrophysiological analysis suggest that the disorder is caused by a defect within the phototransduction process within the retinal pigment epithelium of the eye, and is most likely related to vitamin A metabolism (Gu *et al.*, 1997). Mutations in RPE65 are implicated in several other visual disorders, mainly childhood-onset retinal dystrophies (Gu *et al.*, 1997). Thus RPE65 is of great relevance to the dog mapping community, for which retinal disorders are a major source of interest. Veske *et al.* (1996) attempted to establish the genetic nature of the disorder, and were able to exclude as the cause three candidate genes that are known to cause inherited retinal dystrophies in other species. The changes apparent in the retinal pigment epithelium of affected animals, accompanied by the fact that RPE65 is solely expressed in this location, led to the analysis of RPE65 as a further candidate gene. Existing sequence data for the gene from other species were used to design consensus PCR primers to enable amplification of the cDNA from the retinal RNA of affected animals (GenBank entry Y16567). Subsequently, the cause of hereditary retinal dystrophy in the Swedish Briard was shown to be a four basepair deletion in the RPE65 gene (Veske *et al.*, 1999). RPE65 is widely thought to be involved in the metabolism of vitamin A in the eye (Redmond *et al.*, 1998), however the precise function of the gene product has not yet been established. Mutations in the RPE65 gene have been implicated in several human retinal degeneration disorders, including autosomal recessive retinitis pigmentosa and Leber congenital amaurosis (Morimura *et al.*, 1998).

A disease similar to CSNB has been recognised in European countries other than Sweden, and also in North America. Consequently, Aguirre *et al.* (1998) sought to establish whether the 4bp deletion in the RPE65 gene also exists in Briard dogs originating from countries outside Sweden, in an investigation parallel to that of Veske *et al.* (1999). Human, bovine and rat GenBank sequences for RPE65 were aligned and used to design primers within regions where sequence conservation was high. These enabled amplification of part of the 3' and 5' cDNA sequences of the dog RPE65 gene, from a

retinal cDNA library derived from a clinically normal animal. Further 3' data were obtained by reverse transcription from total retinal RNA using additional consensus primers. The identity of the resulting product was confirmed by sequencing analysis and a subsequent database similarity search. This showed that the coding region of the gene shares 88%-89% identity with the human and cattle sequence at the nucleotide level. Two of the consensus primers (RPE65-1 and RPE65-3) were designed within putative exon five of the gene, and spanned the site of the mutation previously described for CSNB. Ten affected Briards were used in the study, which originated from either Europe or North America. Five further animals were included, which were clinically normal but were related to the ten affected animals. PCR analysis was performed on RNA isolated from each individual using primers RPE65-1 and RPE65-3 on both homozygous normal and affected individuals. The resulting products were sequenced to confirm their identity, and that from affected individuals was found to be 4bp shorter than that from normal dogs. Sequencing analysis confirmed the presence of a 4bp deletion in this region of exon 5, resulting in mistranslation of the coding sequence and a non-functional gene product. This deletion was found to be present in Briards affected by CSNB originating from Canada and the USA.

Aguirre *et al.* (1998) also applied the same PCR primers to genomic DNA, and sequencing analysis of the resulting product from affected animals demonstrated it to be 4bp shorter than the 109bp product obtained from normal dogs. As expected, both products were present for those individuals that were heterozygous at this locus. These data demonstrated that the mutation found was identical to that previously reported in those Briards originating from Sweden that had been diagnosed with retinal dystrophy. Thus CSNB and the retinal dystrophy identified in the Swedish Briard appear to represent a single disorder. The existence of the same mutation in apparently unrelated animals originating in different countries would suggest that the aetiology of the disorder involves a founder effect which resulted in the transmission of a single mutated allele throughout the Briard population. Complete cosegregation of the mutation was found by genotyping analysis of a pedigree informative for the *csnb* locus. A partial dog mRNA sequence for RPE65 can be in found in the GenBank database (accession no. AF084537).

Results

D.1 Isolation of genomic DNA samples from buccal swabs

Buccal swabs were taken from 34 individuals from the UK Briard population. DNA was prepared from each sample using the method described in section 2.1.3. Additional swabs were taken from six dogs of other breeds to act as control samples.

D.2 Optimisation of genotyping analysis on unrelated individuals

PCR primers specific for exon five of the RPE65 gene (RPE65-1 and RPE65-3, Aguirre *et al.*, 1998, appendix B) were used in a standard 10µl amplification reaction with an annealing temperature of 60°C and 1.0mM $MgCl_2$. Initially, control reactions were performed by including 2µl of DNA isolated from the buccal swabs from the six unrelated dogs as the genomic template. Agarose gel electrophoresis confirmed the generation of a product of approximately 100bp. The PCR was then repeated with the inclusion of 0.05µl of Rhodamine-110 fluorescent dUTP in each sample, and genotyping analysis was performed as described in section 2.1.11. Results are summarised in figure Di. In each case, a single, well-defined product was observed at 105bp, demonstrating as expected that all six control individuals were homozygous for the normal RPE65 allele. The use of DNA derived from a buccal swab as the template did not appear to have a detrimental effect on the quality of the data obtained, generating comparable results to those using genomic DNA purified from whole blood. Previous sequencing analysis of the product has confirmed this as representing the normal allele (section 4.3.5). These conditions were therefore considered suitable for genotyping analysis at the RPE65 locus in the Briard.

D.3 Genotyping analysis of Briards

A total of 34 buccal swabs were obtained from the UK Briard population. Genotyping analysis was performed on all individuals as described in section 2.1.11. Examples of genotyping data are shown in figure Dii, and the disease status for all individuals tested is summarised in figure Diii. In 26 cases, genotyping data showed the presence of a single product of 105bp, corresponding to the wild-type allele seen in the unrelated animals. This was marginally smaller than expected, since the corresponding normal allele described by Aguirre *et al.* (1998) was 109bp. This is likely to reflect the small amount of variation in data typically obtained using different sequencing and genotyping

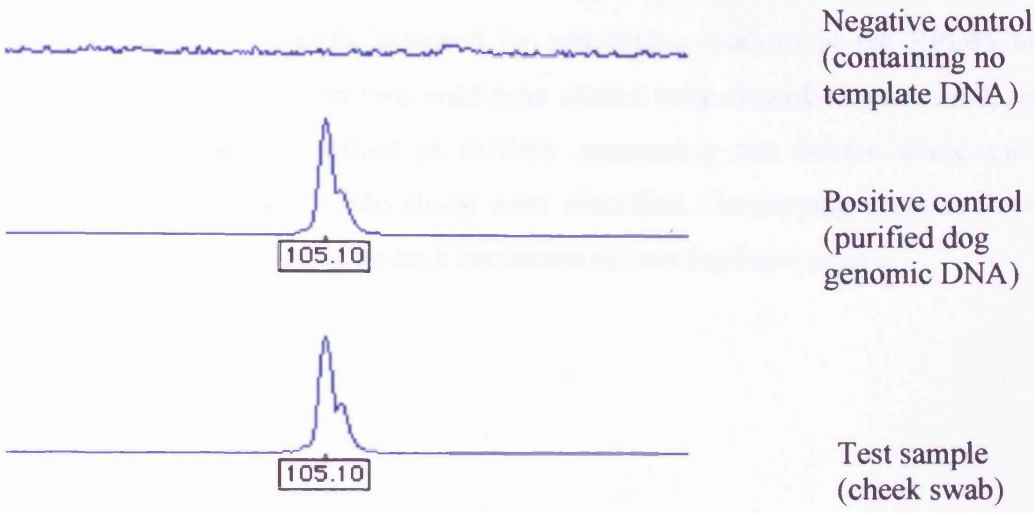


Figure Di:

Genotyping analysis of unrelated dogs at the RPE65 locus. Examples of genotyping traces are given, derived from two different types of DNA template, the nature of which is indicated to the right of each trace. Both individuals are homozygous for the normal allele, the size of which is given (in bp) below the peak.

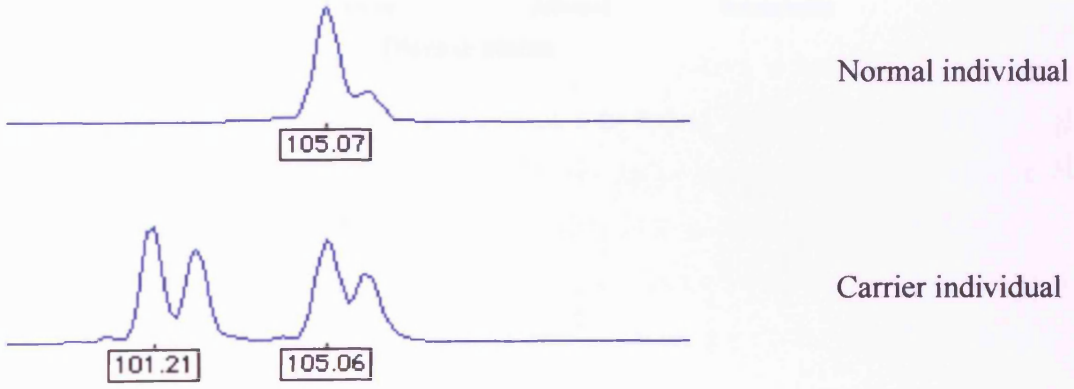
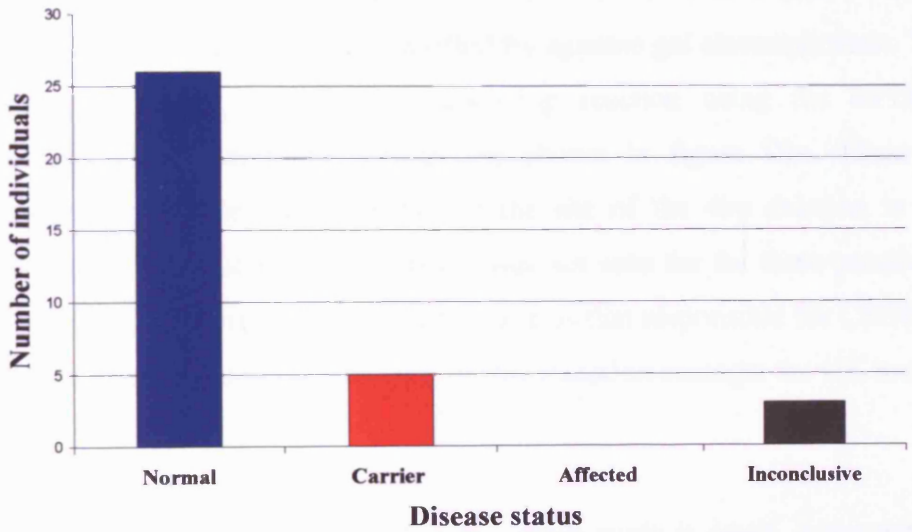


Figure Dii:

Genotyping analysis of two Briards at the RPE65 locus. The former is homozygous for the wild-type allele. The second individual is a carrier of the RPE65 mutation, with one allele demonstrating the four base pair deletion associated with the disease, and the other normal. The size of each allele is given (in bp) below the corresponding peak.

Figure Diii:

Disease status of 34 Briards assessed by genotyping analysis at the RPE65 locus. Twenty six individuals with two wild-type alleles were described as normal, whilst five individuals were identified as carriers, possessing one deleted allele and one normal allele. No affected individuals were identified. Genotyping analysis for three individuals was inconclusive under a minimum of two duplicate assays.



methods. Consequently, these Briards were classed as homozygous for the normal allele. Five further samples indicated the presence of a product of 101bp in addition to that at 105bp. The 101bp product was therefore hypothesised to correspond to the allele containing the four basepair deletion described by Veske *et al.* (1999). A further three samples produced inconclusive results despite repeated attempts, most likely reflecting the poor quality of the DNA template obtained from the initial swab sample. No individuals were identified as homozygous for the mutated allele.

D.4 Sequencing analysis of mutated alleles

To confirm the identity of the two products observed, the PCR was repeated for three putative carriers and three putative normal animals in a final volume of 100µl. The products were purified and quantified by agarose gel electrophoresis. The DNA samples were then used in a cycle sequencing reaction using the RPE65-1 primer. The corresponding electropherograms are shown in figure Div. These clearly show an interruption in the sequence data at the site of the 4bp deletion in the three putative carriers of the mutated allele, which was not seen for the three putative normal animals. This therefore confirmed the 101bp allele as that responsible for CSNB in the Briard, and in turn demonstrated the presence of this mutation amongst the UK breeding stock.

D.5 Discussion

Although the sample size used in the above study is small, it has demonstrated that the RPE65 deletion mutation exists within the UK Briard population. Application of the Hardy-Weinberg equation for allele frequency can be used to derive an estimate of the proportion of affected individuals expected should these data prove to be representative of the UK Briard population as a whole. This states that $p^2 + 2pq + q^2 = 1$, and also that $p + q = 1$, where in this situation p represents the frequency of the wild-type allele, and q the mutated allele, within the population. From this study:

p^2 =	number of normal individuals	= 26/31	= 0.839
p =	frequency of normal allele	$\sqrt{0.839}$	= 0.916

Since $p + q = 1$:

q =	frequency of deleted allele	= 1- 0.916	= 0.084
q^2 =	number of affected individuals	= $(0.084)^2$	= 0.007

The Hardy-Weinberg equation applies to populations where the frequency of each allele is independent from any other, which in cases such as this would imply conditions of

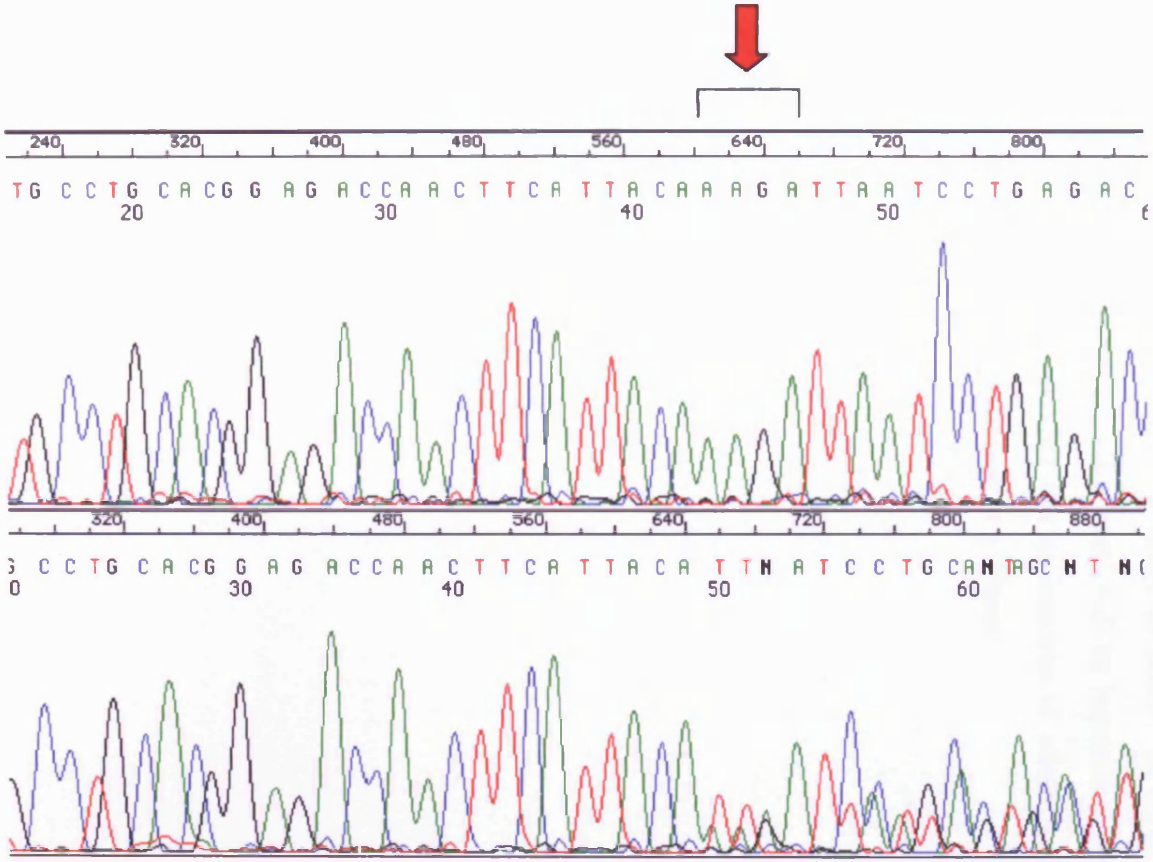


Figure Div:

Sequencing data for normal (upper lane) and carrier (lower lane) individuals at the RPE65 locus. The midregion of the RPE65 gene sequence is aligned for each allele. The arrow indicates the site of the 4bp deletion in the mutated allele, beyond which the quality of the sequence data deteriorates due to the presence of a mixed template. This 4bp region (sequence AAGA) is present in both alleles in the normal individual, and so sequence data are of high quality beyond this point.

random mating. However, the very nature of the pedigree dog is such that this is highly unlikely to be the case, since individuals demonstrating the desired characteristics of the breed will undoubtedly have been used in selected breeding programs to give rise to a disproportionate number of offspring. However, this study was intended primarily to demonstrate whether the mutation involved in CSNB in the Swedish Briard is present within the UK population of the breed. The overall frequency of homozygous affected animals can be estimated from this study as 0.007, or 0.7%. Similar analysis of a considerably larger number of individuals will be required in order to identify affected animals, and to establish whether the estimations of allele frequency obtained in this study are accurate for the population as a whole.

Appendix E: Genotyping analysis at the MMP1 and K315 loci

Abstract

Two markers, shown by FISH analysis to map to CFA 5, have been identified as amplifying microsatellite sequences from the dog. MMP1 was studied in chapter four using existing primers for this gene sequence developed by Jiang *et al.* (1998), and was shown to contain a $(TA)_{25}$ repeat motif in the individual analysed. A BAC clone for MMP1 was isolated and used to assign the gene to CFA 5q14.3. K315 is an anonymous cosmid clone that has been assigned to CFA 5q33dist. PCR primers for this marker have been developed (N. Holmes, pers. comm.) and were shown to amplify a region containing a $(CA)_{18}$ microsatellite repeat from the dog cosmid clone.

Genotyping analysis was performed at both the MMP1 and K315 loci, using the panel of reference family individuals described by Mellersh *et al.* (1997). This panel has been described in section 1.5.3, of which families four to seven were made available for analysis. Amplification of both marker products was performed using the conditions optimised in chapter four, using as the genomic template DNA from each of the first and second generation individuals within the reference family. Both markers demonstrated the presence of five distinct alleles within the individuals typed. Uninformative subsets of the pedigree were identified and excluded from further study. No exceptions to Mendelian inheritance were observed. All remaining individuals were typed and data were analysed to enable the MMP1 and K315 loci to be incorporated into the meiotic linkage map of CFA 5.

Introduction

One of the major aims of the dog genome mapping community is the production of an integrated map incorporating data from meiotic linkage and RH mapping with those derived from physical assignment of loci by FISH analysis. The map would be further strengthened by the extrapolation of data from other species, derived from approaches such as comparative chromosome painting techniques.

Chapters four and five have combined data from all of the above techniques with the exception of meiotic linkage analysis, since this was not amongst the most pertinent aims of the project. This is in part because a preliminary meiotic linkage map of CFA 5 has already been produced (Neff *et al.*, 1999; Werner *et al.*, 1999), comprising nine anonymous microsatellite markers and a single gene marker (Werner *et al.*, 1999). The current integrated RH and meiotic linkage map of CFA 5 (Mellersh *et al.*, 2000) consists of nine markers mapped by both methods, all of which represent anonymous microsatellite sequences. The scarcity of mapped genes on this chromosome, and within the dog genome as a whole, therefore meant that the focus of this project lay in the mapping of coding loci.

One microsatellite sequence was identified fortuitously in chapter four, in which primers described by Jiang *et al.* (1998) were used to amplify part of the dog MMP1 gene sequence. The product demonstrated the presence of a $(TA)_{25}$ microsatellite repeat sequence within the individual analysed. This gene was subsequently mapped to CFA 5q14.3, by the isolation and FISH analysis of a dog BAC clone containing part of the MMP1 gene sequence. The K315 cosmid clone was mapped in chapter five to CFA 5q33dist. by FISH analysis. A second primer pair has been designed for cosmid K315 (denoted K315 μF and μR, appendix B, N. Holmes, pers. comm.), amplifying a $(CA)_{18}$ microsatellite repeat from the clone. Both loci were also included in the RH map of CFA 5 discussed in chapter five.

Since MMP1 and K315 are well separated on CFA 5, the study was extended to genotyping analysis of these loci within the panel of reference family individuals described by Mellersh *et al.* (1997). The addition of these loci to the CFA 5 meiotic linkage map would enable correlation of this map with the physical map of the chromosome, since both loci have been assigned by FISH and RH analysis. In turn, this would provide a starting point for further integration of data from all mapping approaches available for the dog.

Results

E.1 Optimisation of conditions for genotyping analysis

Amplification conditions for the K315 repeat sequence were optimised (annealing 46°C, 3.0mM $MgCl_2$) as described in section 2.1.4. Amplification conditions for the MMP1 locus were established in chapter four. The appropriate conditions were used to amplify the K315 and MMP1 microsatellite loci from a panel of four randomly selected dog genomic DNA templates. The products were then analysed by polyacrylamide gel electrophoresis as described in section 2.1.11. In both cases, the optimised conditions generated products that could be utilised in genotyping analysis, with clearly defined allele peaks and limited spurious background amplification.

An example of the sequence data obtained for the microsatellite repeat sequence at both the K315 and MMP1 loci in the dog are shown in figures Ei and Eii. The MMP1 marker primers described by Jiang *et al.* (1998) were also found to robustly amplify from human genomic DNA. The PCR product was sequenced using marker primers and aligned with the dog sequence. It could be clearly seen that the corresponding region of the human gene does not contain a microsatellite. This could potentially be isolated to the single individual tested, and further analysis using human genomic DNA from a number of unrelated individuals would be required to establish whether absence of the repeat sequence is typical. An alignment of the human and dog gene sequence generated using the MMP1 forward primer is shown in figure Eii.

E.2 Genotyping analysis of reference family parents

Genotyping analysis was performed using a subset of the reference family described by Mellersh *et al.* (1997) in order to determine whether the MMP1 and K315 loci are polymorphic within this set of individuals. All members of the first and second generations of these families were subjected to genotyping analysis at both loci. Absence of, or limited, polymorphism amongst these individuals would preclude the markers from further study of the loci amongst the larger number of third generation individuals.

The structure of the four reference family pedigrees used is shown in figures Eiii-vi. Families four, five, six and seven respectively comprise six, nine, nine and twelve first and second generation individuals. The four families share a small number of common individuals, such that the total number of unique first and second generation individuals is 23. Genotyping analysis of these individuals showed the presence of five different alleles at both loci, which rendered both loci worthy of further analysis across the

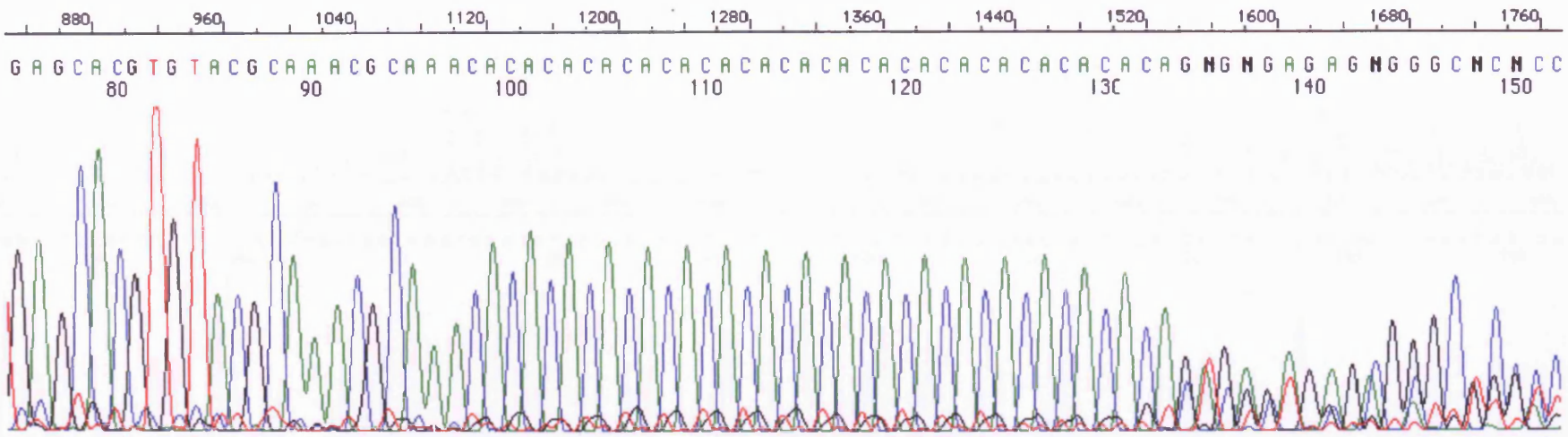


Figure Ei:

Sequence data from the microsatellite repeat region of the K315 locus. This individual appears to be heterozygous at this locus, since the sequence quality deteriorates markedly after the repeat region, indicative of a mixed template.

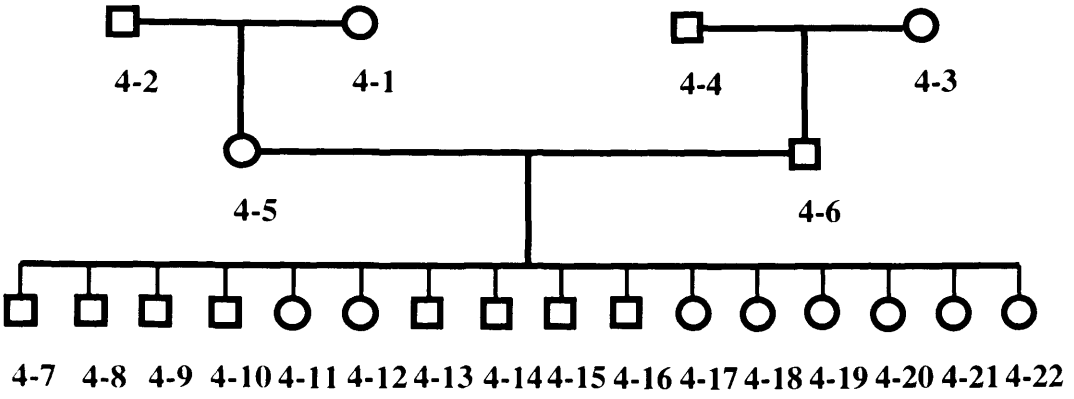


Figure Eiii:

Pedigree structure of family four of the dog reference panel.

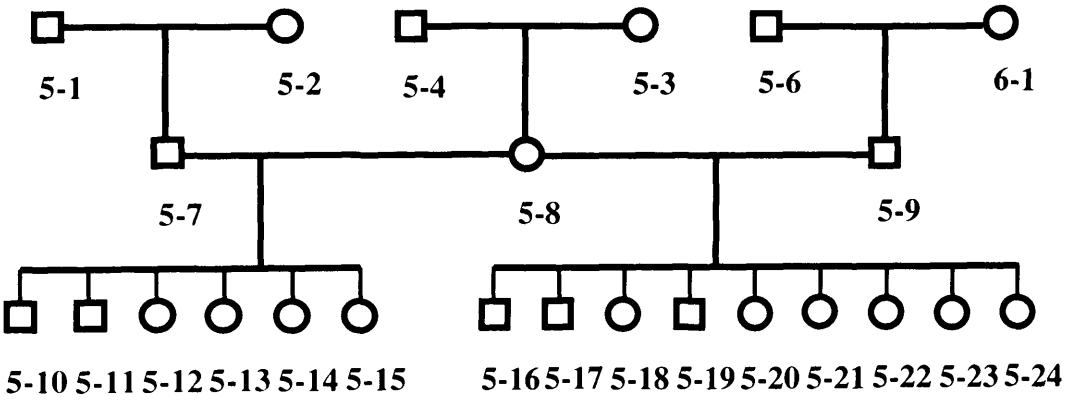


Figure Eiv:

Pedigree structure of family five of the dog reference panel.

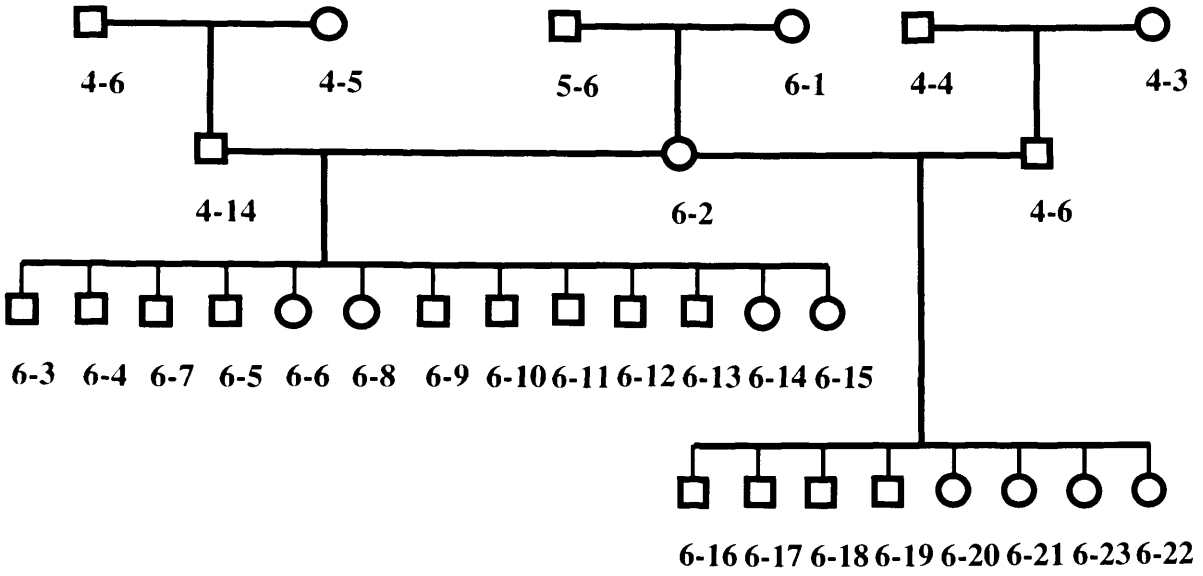


Figure Ev:

Pedigree structure of family six of the dog reference panel.

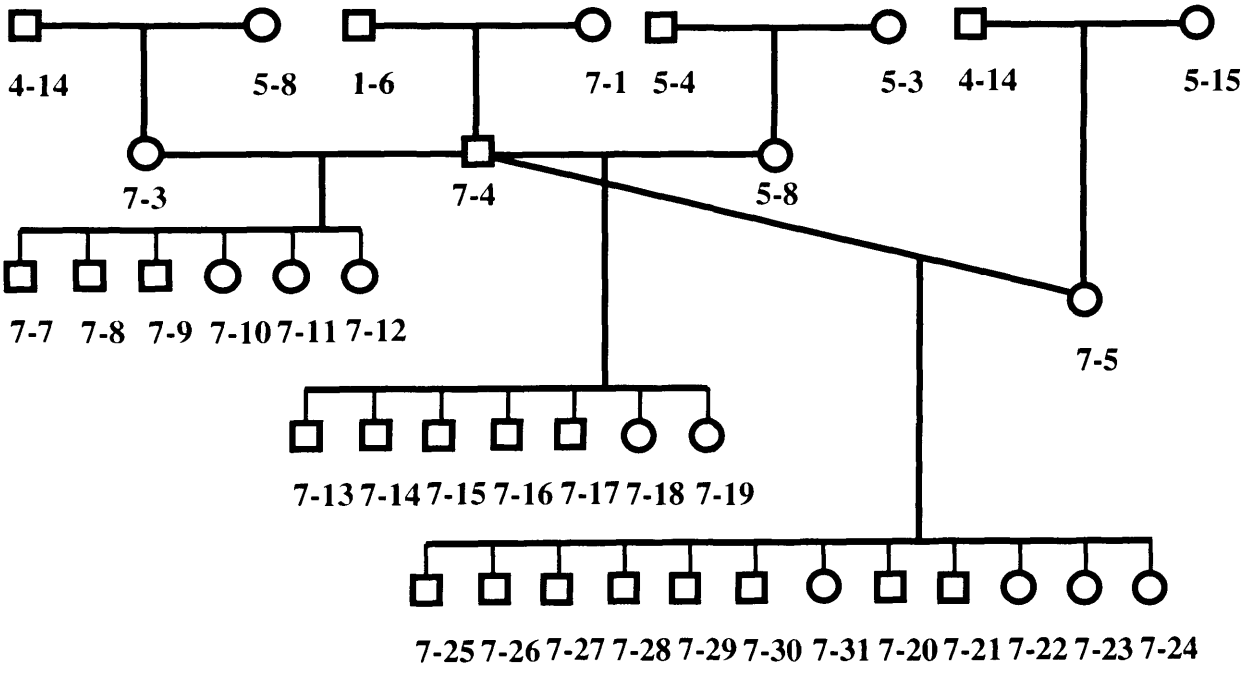


Figure Evi:

Pedigree structure of family seven of the dog reference panel.

remaining members of the pedigrees. K315 demonstrated alleles ranging from 256bp to 275bp, whilst those at the MMP1 locus ranged from 388bp to 416bp. Only a small subset of the panel was identified as being uninformative for the MMP1 locus, in which two second generation individuals (7-4 and 7-5) were found to be homozygous for the same allele. Genotyping analysis at this locus was therefore not pursued on the offspring of these individuals (7-20 to 7-31).

E.3 Genotyping analysis of the complete reference family pedigree

Genotyping analysis was performed in turn on the remaining members of each potentially informative family within the reference panel. In each case, amplified products derived from section E.2 were included so that parental alleles could be directly compared with those of their offspring. Allele data were initially scored anonymously, and then compared to the pedigree structure to ensure correlation with parental allele scores. No exceptions to Mendelian inheritance were identified.

Each of the 97 unique members of the reference panel were subjected to genotyping analysis at the K315 locus, of which 23 had previously been typed in section E.2. No alleles were identified that had not previously been observed within the first and second generation individuals. Products derived from the K315 locus were highly robust, generating well defined data that could be readily interpreted, with the exception of four individuals for which allele sizes could not be determined.

Having identified an uninformative subset of family seven of the reference family, genotyping analysis was performed on the remaining 62 individuals at the MMP1 locus. No additional alleles were identified within these individuals other than the five identified within the first and second generation members. Data for the MMP1 were less satisfactory than those for K315. The peaks corresponding to the different MMP1 alleles were not as well defined, but could still be readily interpreted. In instances of ambiguity, the relevant individual was retyped and the data from both duplicates were compared to ensure agreement. Where data remained ambiguous after both repetitions, that individual was excluded from further analysis. This applied to a total of ten individuals. Efforts were made to reoptimise the amplification conditions for the locus, using the method described in chapter two, but no improvement was found.

Genotyping scores for all individuals typed are shown in tables Ei and Eii.

Table Ei:

Genotyping scores at the K315 locus. Each individual is listed in numerical order within the relevant family followed by the corresponding allele sizes. Those individuals for which genotyping data could not be determined are denoted 'nd'.

Individual	Alleles	Individual	Alleles	Individual	Alleles	Individual	Alleles	Individual	Alleles
4-1	262/262	5-1	260/275	6-1	256/260	7-1	260/262	7-20	260/275
4-2	256/273	5-2	262/262	6-2	260/275	7-3	262/262	7-21	260/275
4-3	262/273	5-3	256/262	6-3	nd	7-4	260/262	7-22	262/262
4-4	256/262	5-4	256/273	6-4	260/262	7-5	262/275	7-23	262/262
4-5	256/262	5-6	256/275	6-6	260/262	7-7	nd	7-24	262/262
4-6	262/262	5-7	262/275	6-7	260/262	7-8	260/262	7-25	262/262
4-7	256/262	5-8	256/262	6-8	262/275	7-9	260/262	7-26	262/262
4-8	262/262	5-9	256/260	6-9	260/262	7-10	260/262	7-27	260/262
4-9	256/262	5-10	256/262	6-10	260/262	7-11	260/262	7-28	260/262
4-10	256/262	5-11	256/262	6-11	260/262	7-12	260/262	7-29	260/262
4-11	256/262	5-12	256/262	6-12	262/275	7-13	260/262	7-30	260/262
4-12	256/262	5-13	256/262	6-13	260/262	7-14	262/262	7-31	262/275
4-13	256/262	5-14	256/275	6-14	262/275	7-15	256/260		
4-14	262/262	5-15	262/275	6-15	260/262	7-16	256/260		
4-15	262/262	5-16	256/262	6-16	260/262	7-17	262/262		
4-16	262/262	5-17	nd	6-17	262/275	7-18	256/260		
4-17	262/262	5-18	256/262	6-18	260/262	7-19	256/260		
4-18	256/262	5-19	256/260	6-19	262/275				
4-19	262/262	5-20	260/262	6-20	262/275			1-6	260/260
4-20	256/262	5-21	256/262	6-21	262/275				
4-21	256/262	5-22	nd	6-22	260/262				
4-22	262/262	5-23	256/260	6-23	260/262				
		5-24	260/262						

Table Eii:

Genotyping scores at the MMP1 locus. Each individual is listed in numerical order within the relevant family followed by the corresponding allele sizes. Those individuals for which genotyping data could not be determined are denoted 'nd'. The allele size for the corresponding human locus (HSA) is also shown for comparison.

Individual	Alleles	Individual	Alleles	Individual	Alleles	Individual	Alleles
4-1	388/388	5-1	388/406	6-1	408/412	7-1	388/388
4-2	388/416	5-2	388/388	6-2	412/416	7-3	416/416
4-3	416/416	5-3	388/416	6-3	nd	7-4	388/388
4-4	388/406	5-4	406/416	6-4	nd	7-5	388/388
4-5	388/388	5-6	388/416	6-6	nd	7-7	388/416
4-6	388/416	5-7	388/388	6-7	388/412	7-8	388/416
4-7	388/416	5-8	406/416	6-8	388/416	7-9	388/416
4-8	388/416	5-9	408/416	6-9	388/412	7-10	388/416
4-9	388/416	5-10	388/416	6-10	388/412	7-11	388/416
4-10	388/388	5-11	388/406	6-11	388/412	7-12	388/416
4-11	388/388	5-12	388/406	6-12	nd	7-13	388/416
4-12	388/388	5-13	388/406	6-13	388/412	7-14	388/416
4-13	388/388	5-14	388/406	6-14	nd	7-15	388/406
4-14	388/416	5-15	388/406	6-15	nd	7-16	388/406
4-15	388/388	5-16	416/416	6-16	388/416	7-17	388/406
4-16	388/388	5-17	406/408	6-17	388/412	7-18	388/406
4-17	388/416	5-18	408/416	6-18	nd	7-19	388/406
4-18	388/416	5-19	406/416	6-19	416/416		
4-19	388/416	5-20	408/416	6-20	nd	1-6	388/412
4-20	388/388	5-21	nd	6-21	416/416		
4-21	388/416	5-22	408/416	6-22	nd		
4-22	388/388	5-23	406/416	6-23	388/412		
		5-24	406/416			HSA	373/373

E.4 Statistical analysis of genotyping data

Existing genotyping data for 11 markers from CFA 5 were provided by C. Mellersh (Fred Hutchinson Cancer Research Center, Seattle) or were generated internally (N. Holmes, pers. comm.). The CRI-MAP program was used for the analysis of data as described in section 2.1.11. Pairwise linkage data are shown in table Eiii. MMP1 demonstrates linkage to five other loci with a lod score of greater than three, and is most tightly linked to the SLC2A4 locus mapped to CFA 5 by Werner *et al.* (1997). This is consistent with data presented in chapters four and five, which confirmed that MMP1 lies proximal to SLC2A4, since both have been mapped in the present study by FISH analysis and RH analysis. Therefore data for these loci can be used to orientate the meiotic linkage map. All other markers showing pairwise linkage to MMP1 are anonymous microsatellite markers. K315 demonstrates linkage to four other loci with a lod score of greater than three, all of which are anonymous microsatellite markers. MMP1 and K315 are not significantly linked to each other, which is consistent with their distant location on CFA 5. Both loci are linked to three markers in common (CPH18, C05.771 and FH2383).

The locus order for the 11 existing CFA 5 markers (C. Mellersh and N. Holmes, pers. comm.) was used as a framework order for entry into the BUILD option of CRI-MAP, which was then used to integrate MMP1 and K315 into the map. This order is shown in figure Evii. The highest ranking order reported by BUILD demonstrated that the most likely location of the MMP1 locus lay between markers AHT141 and SLC2A4, and the relative distance between these markers indicated that MMP1 is situated closer to SLC2A4. This is consistent with the pairwise linkage data, which show a higher rate of recombination between AHT141 and MMP1 (θ = 0.08) than between MMP1 and SLC2A4 (θ = 0.02). K315 was placed between markers FH2383 and C05.377, closer to the former. This is also consistent with the pairwise linkage data, since a lower rate of recombination is observed between FH2383 and K315 (θ = 0.09) than between K315 and C05.377 (θ = 0.16).

The FLIPS option of CRI-MAP was used to swap loci in pairs and in groups of five to identify potentially improved locus orders. This confirmed that the proposed locus order was the most likely, since the likelihood of all other orders did not exceed that of the original. These data were used to generate an updated meiotic linkage map for CFA 5 incorporating these novel markers, which is presented in figure Evii.

Table Eiii:

Twopoint meiotic linkage analysis with markers a) MMP1 and b) K315, established using the TWOPOINT option of CRI-MAP. Pairwise linkages reported with a lod score of greater than three are listed below.

a) MMP1

Marker A	Marker B	Recombination fraction	Lod score
MMP1	SLC2A4	0.02	16.15
MMP1	AHT141	0.08	6.05
MMP1	CPH18	0.10	5.47
MMP1	C05.771	0.14	4.54
MMP1	FH2383	0.20	3.07

b) K315

Marker A	Marker B	Recombination fraction	Lod score
K315	FH2383	0.09	14.00
K315	C05.771	0.13	10.65
K315	C05.377	0.16	7.12
K315	CPH18	0.17	3.99

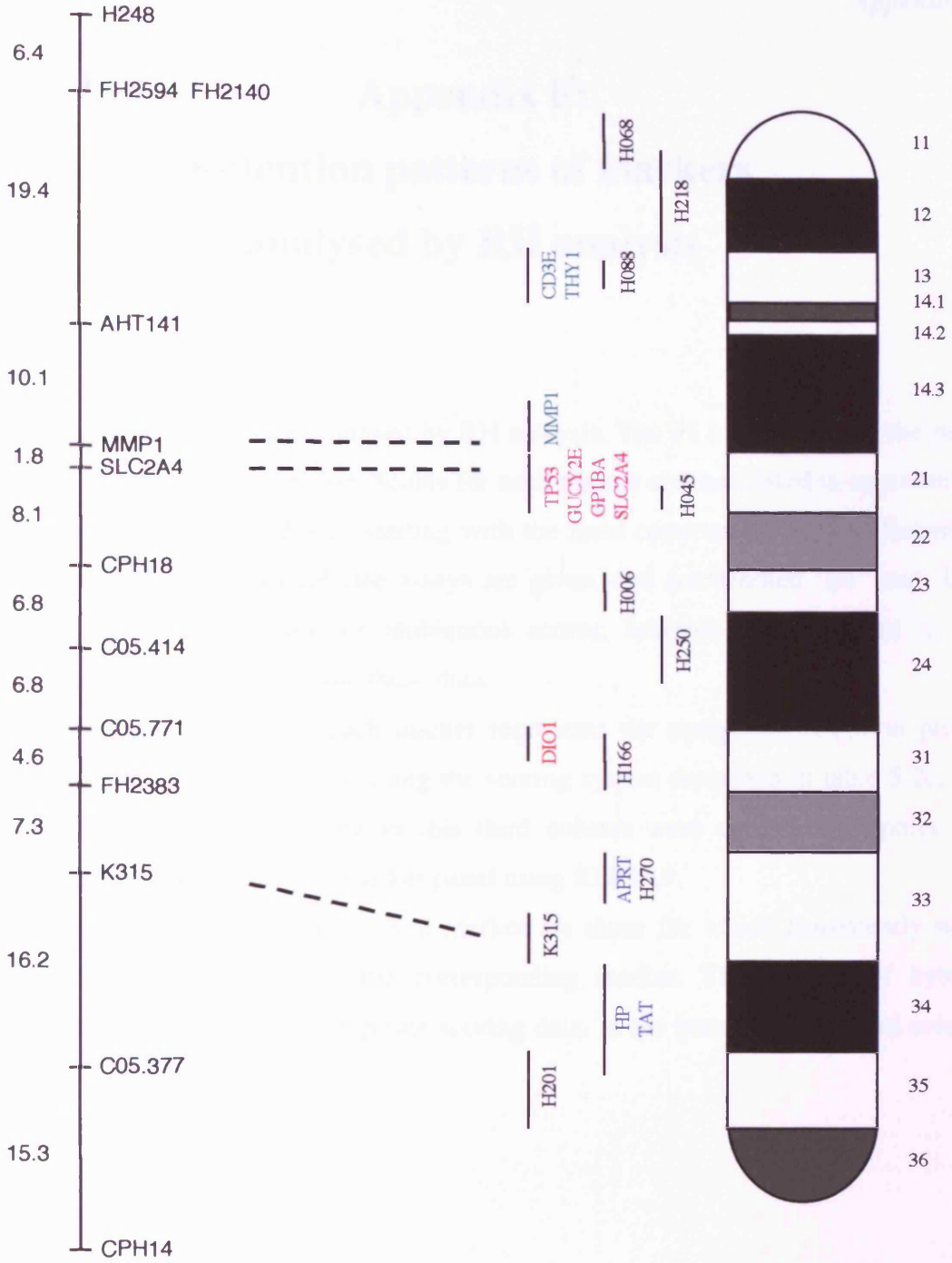


Figure Evii:

Meiotic linkage map of CFA 5 and integration of novel markers. Genotyping analysis of the K315 and MMP1 loci enabled these markers to be incorporated into the existing CFA 5 meiotic linkage map (Mellersh *et al.*, 2000; N. Holmes, pers. comm.). The orientation of the map can be determined on the basis of cytogenetic assignments of existing markers, shown on the right, which were generated in chapters four and five. Distances between markers on the meiotic map are shown to scale in centiMorgans.

Appendix F: Retention patterns of markers analysed by RH analysis

Table Fi: (overleaf)

Retention patterns of markers analysed by RH analysis. The 91 hybrids within the panel are represented within column one. Scores for each marker are then listed in approximate order of their location on CFA 5, starting with the most centromeric marker. Retention patterns observed for both duplicate assays are given, and are denoted 'D1' and 'D2'. The number of discrepancies, or ambiguous scores, between data obtained in the duplicate assays is indicated below these data.

The third column of data for each marker represents the composite retention profile derived from the duplicate assays, using the scoring system described in table 5.2c, and are denoted 'F' (final). The data in this third column were used for twopoint and multipoint analysis of the CFA 5 marker panel using RHMAP.

Hybrids highlighted in blue for a given marker are those for which consistently weak amplification was observed for the corresponding marker. The number of hybrids concerned is given below the composite scoring data, at the bottom of the third column of data for each marker.

Hybrid	H068	H068	H068	H218	H218	H218	H006	H006	H006	H250	H250	H250
	D1	D2	F	D1	D2	F	D1	D2	F	D1	D2	F
1	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-
3	+	+	+	-	-	-	+	+	+	-	-	-
4	-	-	-	+	+	+	-	-	-	+	+	+
5	-	-	-	-	-	-	+	+	+	+	+	+
6	-	-	-	-	-	-	-	-	-	+	+	+
7	+	+	+	+	+	+	-	-	-	-	-	-
8	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	+	+	+	-	-	-	-	-	-
10	-	-	-	-	-	-	+	+	+	-	-	-
11	+	+	+	+	+	+	-	-	-	-	-	-
12	+	+	+	+	+	+	-	-	-	-	-	-
13	+	+	+	+	+	+	-	-	-	-	-	-
14	-	-	-	-	-	-	-	-	-	-	-	-
15	+	+	+	+	+	+	+	+	+	+	+	+
16	-	-	-	-	-	-	-	-	-	-	-	-
17	+	+	+	+	+	+	+	+	+	+	+	+
18	-	-	-	-	-	-	+	+	+	+	+	+
19	-	-	-	+	+	+	-	-	-	-	-	-
20	-	-	-	w	-	?	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-
22	?	?	?	+	+	+	-	-	-	-	-	-
23	+	+	+	+	+	+	-	-	-	-	w	?
24	+	+	+	+	+	+	-	-	-	-	-	-
25	?	?	?	w	w	+	-	-	-	-	-	-
26	+	+	+	+	+	+	-	-	-	-	-	-
27	+	+	+	+	+	+	-	-	-	-	-	-
28	+	+	+	+	+	+	-	-	-	-	-	-
29	-	-	-	-	-	-	-	-	-	-	-	-
30	+	+	+	-	-	-	+	+	+	+	+	+
31	-	-	-	-	-	-	-	-	-	-	-	-
32	-	-	-	-	-	-	-	-	-	-	-	-
33	-	-	-	-	-	-	-	-	-	-	-	-
34	+	+	+	+	+	+	w	w	+	+	+	+
35	+	+	+	+	+	+	-	-	-	-	-	-
36	+	+	+	-	-	-	-	-	-	-	-	-
37	+	+	+	+	+	+	-	-	-	-	-	-
38	+	+	+	-	-	-	-	-	-	-	-	-
39	+	+	+	-	-	-	+	+	+	-	-	-
40	+	+	+	+	+	+	+	+	+	+	+	+
41	-	-	-	-	-	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-	w	-	?
44	-	-	-	-	-	-	+	+	+	+	+	+
45	-	-	-	-	-	-	+	+	+	+	+	+
46	-	-	-	-	-	-	-	-	-	-	-	-
47	-	-	-	-	-	-	+	+	+	+	+	+
48	+	+	+	+	+	+	+	+	+	+	+	+
49	+	+	+	-	-	-	-	-	-	-	-	-
50	+	+	+	+	+	+	-	-	-	-	-	-
51	+	+	+	+	+	+	-	-	-	-	-	-
52	+	+	+	+	+	+	+	+	+	+	+	+
53	+	+	+	+	+	+	-	-	-	-	-	-
54	+	+	+	+	+	+	+	+	+	+	+	+
55	w	w	+	+	+	+	-	-	-	-	w	?
56	-	-	-	-	-	-	-	-	-	-	-	-
57	+	+	+	+	+	+	-	-	-	-	-	-
58	-	-	-	-	-	-	-	-	-	-	-	-
59	-	-	-	-	-	-	-	-	-	w	-	?
60	+	+	+	+	+	+	-	-	-	-	-	-
61	+	+	+	+	+	+	-	-	-	-	-	-
62	-	-	-	-	-	-	-	-	-	-	-	-
63	+	+	+	+	+	+	-	-	-	-	-	-
64	-	-	-	-	-	-	+	+	+	+	+	+
65	-	-	-	-	-	-	-	-	-	-	-	-
66	+	+	+	+	+	+	-	-	-	-	-	-
67	+	+	+	+	+	+	-	-	-	-	-	-
68	-	-	-	-	-	-	-	-	-	-	-	-
69	-	-	-	-	-	-	-	-	-	-	-	-
70	-	-	-	-	-	-	-	-	-	-	-	-
71	-	-	-	-	-	-	-	-	-	-	-	-
72	+	+	+	+	+	+	-	-	-	-	-	-
73	-	-	-	-	-	-	-	-	-	-	-	-
74	-	-	-	-	-	-	-	-	-	-	-	-
75	+	+	+	+	+	+	-	-	-	-	-	-
76	-	-	-	-	-	-	-	-	-	w	-	?
77	-	-	-	-	-	-	-	-	-	-	-	-
78	-	-	-	w	-	?	+	+	+	+	+	+
79	-	-	-	w	-	?	+	+	+	+	+	+
80	-	-	-	-	-	-	-	-	-	+	+	+
81	-	-	-	-	-	-	+	+	+	+	+	+
82	w	w	+	+	+	+	-	-	-	-	-	-
83	w	w	+	-	-	-	-	-	-	-	-	-
84	w	w	+	-	-	-	-	-	-	-	w	?
85	+	+	+	+	+	+	-	-	-	-	-	-
86	+	+	+	+	+	+	+	+	+	+	+	+
87	-	-	-	-	-	-	-	-	-	-	-	-
88	+	+	+	+	+	+	-	-	-	-	-	-
89	+	+	+	+	+	+	-	-	-	-	-	-
90	+	+	+	+	+	+	-	-	-	-	-	-
91	-	-	-	-	-	-	-	-	-	-	-	-
	2		4	3		1	0		1	6		0

Hybrid	THY1 D1	THY1 D2	THY1 F	CD3E D1	CD3E D2	CD3E F	H088 D1	H088 D2	H088 F	MMP1 D1	MMP1 D2	MMP1 F
1	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	+	+	+
3	-	+	?	+	+	+	+	+	+	-	-	-
4	-	-	-	-	-	-	-	-	-	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	-	-
7	w	+	+	+	+	+	-	-	-	+	+	+
8	-	-	-	-	-	-	-	-	-	-	-	-
9	+	+	+	+	+	+	+	+	+	-	-	-
10	+	+	+	+	+	+	+	+	+	-	-	-
11	-	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-	-
13	-	+	?	+	+	+	w	w	+	+	+	+
14	-	-	-	+	+	+	+	+	+	+	+	+
15	-	-	-	-	-	-	-	-	-	-	-	-
16	-	-	-	-	-	-	-	-	-	-	-	-
17	+	+	+	+	+	+	w	w	+	+	+	+
18	w	+	+	+	+	+	w	w	+	+	+	+
19	-	-	-	-	-	-	-	-	-	w	-	?
20	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	w	?	?
23	-	-	-	-	-	-	-	-	-	-	-	-
24	+	+	+	+	+	+	+	+	+	-	-	-
25	-	-	-	-	-	-	-	-	-	-	-	-
26	w	w	+	-	-	-	-	-	-	-	-	-
27	w	w	+	-	-	-	-	-	-	-	-	-
28	w	+	+	+	+	+	w	-	?	+	+	+
29	-	-	-	-	-	-	-	-	-	+	+	+
30	-	-	-	-	-	-	-	-	-	-	-	-
31	-	-	-	-	-	-	-	-	-	-	-	-
32	-	-	-	-	-	-	-	-	-	-	-	-
33	+	+	+	-	-	-	-	-	-	-	-	-
34	+	+	+	+	+	+	?	?	?	+	?	?
35	-	-	-	-	-	-	-	-	-	-	-	-
36	-	-	-	-	-	-	-	-	-	-	-	-
37	+	+	+	-	-	-	-	-	-	-	-	-
38	-	-	-	-	-	-	-	-	-	w	+	+
39	-	-	-	-	-	-	-	-	-	w	+	+
40	-	-	-	-	-	-	-	-	-	-	-	-
41	-	-	-	-	-	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-	-	-	-
44	w	+	+	-	-	-	-	-	-	+	+	+
45	-	-	-	-	-	-	-	-	-	-	-	-
46	-	-	-	-	-	-	-	-	-	-	?	?
47	+	+	+	+	+	+	?	?	?	-	-	-
48	+	+	+	+	+	+	?	?	?	+	+	+
49	-	-	-	-	-	-	-	-	-	-	-	-
50	-	w	?	+	+	+	-	-	-	-	-	-
51	-	-	-	w	+	+	-	-	-	-	-	-
52	-	+	+	+	+	+	-	-	-	+	+	+
53	-	+	+	+	+	+	-	-	-	-	-	-
54	-	-	-	-	-	-	-	-	-	-	-	-
55	-	w	?	-	w	?	-	-	-	-	-	-
56	-	-	-	-	-	-	-	-	-	-	-	-
57	-	+	?	+	+	+	-	-	-	-	-	-
58	-	-	-	-	-	-	-	-	-	-	-	-
59	-	-	-	-	-	-	-	-	-	-	-	-
60	-	-	-	-	-	-	-	-	-	-	-	-
61	-	+	?	-	w	?	-	-	-	-	-	-
62	-	-	-	-	-	-	-	-	-	-	-	-
63	-	-	-	-	-	-	-	-	-	-	-	-
64	-	-	-	+	+	+	-	-	-	-	-	-
65	-	-	-	-	-	-	-	-	-	-	-	-
66	-	-	-	-	-	-	-	-	-	-	-	-
67	-	-	-	-	-	-	-	-	-	-	-	-
68	-	-	-	-	-	-	-	-	-	-	-	-
69	-	-	-	-	-	-	-	-	-	-	-	-
70	+	+	+	+	+	+	-	-	-	-	-	-
71	-	-	-	-	-	-	-	-	-	-	-	-
72	-	-	-	-	-	-	-	-	-	-	-	-
73	-	-	-	-	-	-	-	-	-	-	-	-
74	-	-	-	-	-	-	-	-	-	-	-	-
75	-	-	-	-	-	-	-	-	-	+	+	+
76	-	-	-	-	-	-	-	-	-	-	-	-
77	-	-	-	-	-	-	-	-	-	-	-	-
78	+	+	+	+	+	+	+	+	+	+	+	+
79	-	+	?	+	+	+	+	+	+	+	+	+
80	+	+	+	-	+	?	+	+	+	-	-	-
81	+	+	+	+	+	+	+	+	+	-	-	-
82	w	w	+	-	w	?	-	-	-	+	+	+
83	-	-	-	-	-	-	w	?	?	+	+	+
84	-	-	-	-	-	-	-	-	-	-	-	-
85	-	-	-	-	-	-	-	-	-	+	+	+
86	w	w	+	w	w	+	-	-	-	+	+	+
87	-	-	-	-	-	-	-	-	-	-	-	-
88	+	+	+	+	+	+	w	?	?	+	+	+
89	+	+	+	+	+	+	w	?	?	+	+	+
90	w	+	+	w	w	+	w	w	+	+	+	+
91	-	-	-	-	-	-	-	-	-	-	-	-
	7		4	4		2	7		4	4		0

Hybrid	GUCY	GUCY	GUCY	TP53	TP53	TP53	H045	H045	H045	GP1B	GP1B	GP1B
	D1	D2	F	D1	D2	F	D1	D2	F	D1	D2	F
1	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	-	-	-	+	w	+	-	-	-
5	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	-	-
7	+	+	+	+	+	+	+	+	+	+	+	+
8	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	-	-	-
11	-	-	-	-	-	-	-	-	-	+	-	?
12	-	-	-	-	-	-	-	-	-	-	-	-
13	+	+	+	+	+	+	+	+	+	+	+	+
14	-	-	-	-	-	-	-	-	-	-	-	-
15	-	-	-	-	-	-	w	w	+	w	w	+
16	-	-	-	-	-	-	-	-	-	-	-	-
17	+	+	+	+	+	+	+	+	+	+	+	+
18	+	+	+	+	+	+	+	+	+	+	+	+
19	-	w	?	w	w	+	w	w	+	w	w	+
20	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-	-	-	-	-	-
24	-	-	-	-	-	-	-	+	?	-	-	-
25	-	-	-	-	-	-	-	-	-	-	-	-
26	-	-	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-	-	-	-
28	+	+	+	+	+	+	+	+	+	+	+	+
29	-	-	-	-	-	-	-	-	-	w	-	?
30	-	-	-	-	-	-	-	-	-	-	-	-
31	-	-	-	-	-	-	-	-	-	-	-	-
32	-	-	-	-	-	-	+	-	?	-	-	-
33	-	-	-	-	-	-	-	-	-	-	-	-
34	w	+	+	w	w	+	+	+	+	+	+	+
35	-	-	-	-	-	-	-	-	-	-	-	-
36	-	-	-	-	-	-	+	-	?	-	-	-
37	-	-	-	-	-	-	-	-	-	-	-	-
38	+	+	+	+	+	+	+	+	+	+	+	+
39	+	+	+	+	+	+	+	+	+	+	+	+
40	-	-	-	-	-	-	-	-	-	-	-	-
41	-	-	-	-	-	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-	-	-	-
44	+	+	+	+	+	+	+	+	+	+	+	+
45	-	-	-	-	-	-	-	-	-	-	-	-
46	-	-	-	-	-	-	-	-	-	-	-	-
47	w	w	+	w	w	+	w	+	+	+	+	+
48	+	+	+	+	+	+	+	+	+	+	+	+
49	-	-	-	-	-	-	-	-	-	-	-	-
50	+	+	+	+	+	+	+	+	+	+	+	+
51	+	+	+	+	+	+	+	+	+	+	+	+
52	+	+	+	+	+	+	+	+	+	+	+	+
53	-	-	-	-	-	-	-	-	-	-	-	-
54	-	-	-	-	-	-	-	-	-	-	-	-
55	-	-	-	-	-	-	-	-	-	-	-	-
56	-	-	-	-	-	-	-	-	-	-	-	-
57	-	-	-	-	-	-	-	-	-	-	-	-
58	-	-	-	-	-	-	-	-	-	-	-	-
59	+	+	+	+	+	+	+	+	+	+	+	+
60	-	-	-	-	-	-	-	-	-	-	-	-
61	-	-	-	-	-	-	-	-	-	-	-	-
62	-	-	-	-	-	-	-	-	-	-	-	-
63	-	-	-	-	-	-	-	-	-	-	-	-
64	-	-	-	-	-	-	-	-	-	+	+	+
65	-	-	-	-	-	-	-	-	-	-	-	-
66	-	-	-	-	-	-	-	-	-	-	-	-
67	-	-	-	-	-	-	-	-	-	-	-	-
68	-	-	-	-	-	-	-	-	-	-	-	-
69	-	-	-	-	-	-	-	-	-	-	-	-
70	-	-	-	-	-	-	-	+	?	-	-	-
71	-	-	-	-	-	-	-	-	-	-	-	-
72	-	-	-	-	-	-	-	-	-	-	-	-
73	-	-	-	-	-	-	-	-	-	-	-	-
74	-	-	-	-	-	-	-	-	-	-	-	-
75	-	-	-	-	-	-	-	-	-	-	-	-
76	-	-	-	-	-	-	-	-	-	-	-	-
77	-	-	-	-	-	-	-	-	-	-	-	-
78	w	-	?	-	-	-	w	-	?	-	-	-
79	-	-	-	-	-	-	-	-	-	-	-	-
80	-	-	-	-	-	-	-	-	-	-	-	-
81	-	-	-	-	-	-	-	-	-	-	-	-
82	+	+	+	+	+	+	+	+	+	+	+	+
83	-	-	-	-	-	-	-	-	-	-	-	-
84	-	-	-	-	-	-	-	-	-	-	-	-
85	-	-	-	-	-	-	-	-	-	-	-	-
86	+	+	+	+	+	+	+	+	+	+	+	+
87	-	-	-	-	-	-	-	-	-	+	+	+
88	-	-	-	-	-	-	-	-	-	-	-	-
89	-	-	-	-	-	-	-	-	-	-	-	-
90	-	-	-	-	-	-	-	-	-	-	-	-
91	-	-	-	-	-	-	-	-	-	-	-	-
	2		1	0		3	5		2	2		2

Hybrid	SLC2	SLC2	SLC2	DIO1	DIO1	DIO1	H166	H166	H166	APRT	APRT	APRT
	D1	D2	F	D1	D2	F	D1	D2	F	D1	D2	F
1	-	-	-	-	-	-	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-
4	-	-	-	w	?	?	-	-	-	+	+	+
5	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	-	-
7	-	+	?	+	+	+	-	-	-	+	+	+
8	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-	-
10	-	-	-	-	-	-	+	+	+	+	+	+
11	-	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	-	-	-	-	-	-
13	-	+	?	w	-	-	-	-	-	-	-	-
14	-	-	-	+	+	+	+	+	+	+	+	+
15	-	w	?	-	-	-	-	-	-	-	-	-
16	-	-	-	-	-	-	w	+	+	+	+	+
17	+	+	+	+	+	+	-	-	-	-	-	-
18	+	+	+	+	+	+	+	+	+	+	+	+
19	-	w	?	-	-	-	-	-	-	-	-	-
20	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-
23	-	-	-	+	+	+	-	-	-	-	-	-
24	-	-	-	-	-	-	-	-	-	-	-	-
25	-	-	-	-	-	-	-	-	-	-	-	-
26	-	-	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-	-	-	-
28	+	-	-	-	-	-	-	-	-	-	-	-
29	-	-	-	-	-	-	+	+	+	-	-	-
30	-	-	-	-	-	-	-	-	-	-	-	-
31	-	-	-	-	-	-	-	w	?	-	-	-
32	-	-	-	-	-	-	-	-	-	-	-	-
33	-	-	-	+	+	+	+	+	+	-	-	-
34	-	w	?	-	-	-	-	w	?	-	-	-
35	-	-	-	-	-	-	-	-	-	-	-	-
36	-	-	-	+	+	+	-	+	+	-	-	-
37	-	-	-	-	-	-	-	-	-	-	-	-
38	+	+	+	-	-	-	-	-	-	-	-	-
39	w	+	+	-	-	-	-	-	-	-	-	-
40	-	-	-	-	-	-	-	-	-	-	-	-
41	-	-	-	-	-	-	-	-	-	-	-	-
42	-	-	-	-	-	-	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-	-	-	-
44	+	+	+	+	+	+	-	-	-	-	-	-
45	-	-	-	+	+	+	+	+	+	+	+	+
46	-	-	-	-	-	-	-	-	-	-	-	-
47	w	w	+	-	-	-	-	w	?	-	-	-
48	+	+	+	+	+	+	+	+	+	+	+	+
49	-	-	-	-	-	-	-	w	?	-	-	-
50	+	+	+	+	+	+	+	+	+	+	+	+
51	+	+	+	+	+	+	+	+	+	+	+	+
52	w	w	+	+	+	+	+	+	+	+	+	+
53	-	-	-	-	-	-	-	-	-	-	-	-
54	-	-	-	w	-	?	-	-	-	+	+	+
55	-	-	-	-	-	-	-	-	-	-	-	-
56	-	-	-	-	-	-	-	-	-	-	-	-
57	-	-	-	-	-	-	-	-	-	-	-	-
58	-	-	-	+	+	+	+	+	+	w	w	+
59	+	+	+	-	-	-	-	-	-	w	+	+
60	-	-	-	+	+	+	+	+	+	+	+	+
61	-	-	-	w	-	?	-	-	-	-	-	-
62	-	-	-	+	+	+	+	+	+	-	-	-
63	-	-	-	+	+	+	+	+	+	+	+	+
64	-	+	?	-	-	-	-	-	-	+	+	+
65	-	-	-	-	-	-	-	-	-	-	-	-
66	-	-	-	+	+	+	+	+	+	+	+	+
67	-	-	-	+	+	+	+	+	+	+	+	+
68	-	-	-	-	-	-	-	-	-	-	-	-
69	-	-	-	+	+	+	-	-	-	w	-	?
70	-	-	-	-	-	-	-	-	-	-	-	-
71	-	-	-	-	-	-	-	-	-	-	-	-
72	-	-	-	+	+	+	+	+	+	-	-	-
73	-	-	-	-	-	-	-	-	-	w	-	?
74	-	-	-	-	-	-	-	-	-	-	-	-
75	-	-	-	-	-	-	-	-	-	-	-	-
76	-	-	-	-	-	-	-	-	-	-	-	-
77	-	-	-	-	-	-	-	-	-	-	-	-
78	-	-	-	-	-	-	+	+	+	w	?	?
79	-	-	-	-	-	-	+	+	+	+	+	+
80	-	-	-	-	-	-	-	-	-	-	-	-
81	-	-	-	-	-	-	-	-	-	-	-	-
82	+	+	+	+	+	+	-	-	-	-	-	-
83	-	-	-	-	-	-	-	-	-	-	-	-
84	-	-	-	+	-	?	-	-	-	-	-	-
85	-	-	-	-	-	-	-	-	-	-	-	-
86	w	+	+	+	+	+	+	+	+	-	-	-
87	+	+	+	+	+	+	+	+	+	+	+	+
88	-	-	-	+	-	?	-	-	-	-	-	-
89	-	-	-	-	-	-	-	-	-	-	-	-
90	-	-	-	+	?	?	-	-	-	-	-	-
91	-	-	-	-	-	-	-	-	-	-	-	-
	6		2	6		0	4		0	3		1

Hybrid	H270	H270	H270	K315	K315	K315	TAT	TAT	TAT	HP	HP	HP
	D1	D2	F	D1	D2	F	D1	D2	F	D1	D2	F
1	-	-	-	-	w	?	-	-	-	-	-	-
2	-	-	-	-	-	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-	-	-	-	-
4	+	+	+	+	+	+	-	-	-	w	+	+
5	-	-	-	-	-	-	-	-	-	-	-	-
6	-	-	-	-	-	-	-	-	-	-	-	-
7	+	+	+	+	+	+	-	-	-	-	-	-
8	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-	-	-	-
10	+	+	+	+	+	+	+	+	+	-	+	?
11	-	-	-	-	-	-	-	-	-	-	-	-
12	-	-	-	-	-	-	+	+	+	+	+	+
13	-	-	-	-	-	-	-	-	-	-	-	-
14	+	+	+	-	w	?	-	-	-	-	-	-
15	-	-	-	w	w	+	+	+	+	+	+	+
16	+	+	+	+	+	+	-	-	-	-	-	-
17	+	+	+	+	+	+	+	+	+	+	+	+
18	+	+	+	+	+	+	+	+	+	+	+	+
19	-	-	-	w	?	?	-	-	-	-	-	-
20	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-	-	-	-	-	-
24	-	-	-	-	-	-	-	-	-	-	-	-
25	-	-	-	-	-	-	-	-	-	-	-	-
26	-	-	-	-	-	-	-	-	-	-	-	-
27	-	-	-	-	-	-	-	-	-	-	-	-
28	-	-	-	-	-	-	-	-	-	-	-	-
29	w	w	+	+	+	+	-	-	-	-	-	-
30	-	-	-	-	-	-	-	-	-	-	-	-
31	-	-	-	-	-	-	-	-	-	-	-	-
32	-	-	-	-	-	-	-	-	-	-	-	-
33	-	-	-	-	-	-	-	-	-	-	-	-
34	-	-	-	-	-	-	+	+	+	+	+	+
35	-	-	-	-	?	?	-	-	-	-	-	-
36	-	-	-	-	-	-	-	-	-	-	-	-
37	-	-	-	-	-	-	-	-	-	-	-	-
38	-	-	-	-	-	-	-	-	-	-	-	-
39	+	+	+	+	+	+	+	+	+	+	+	+
40	-	-	-	+	+	+	-	-	-	-	-	-
41	-	-	-	-	-	-	-	-	-	-	-	-
42	-	-	-	-	?	?	-	-	-	-	-	-
43	-	-	-	-	-	-	-	-	-	-	-	-
44	-	-	-	-	-	-	-	-	-	-	-	-
45	+	+	+	-	-	-	-	-	-	-	-	-
46	-	-	-	-	-	-	-	-	-	-	-	-
47	-	-	-	-	-	-	+	+	+	+	+	+
48	+	+	+	+	+	+	+	+	+	+	+	+
49	-	-	-	-	w	?	-	-	-	-	-	-
50	+	+	+	+	+	+	+	+	+	+	+	+
51	+	+	+	+	+	+	+	+	+	+	+	+
52	+	+	+	+	+	+	w	w	+	+	+	+
53	-	-	-	-	-	-	-	-	-	-	-	-
54	+	+	+	-	-	-	+	+	+	+	+	+
55	-	-	-	-	-	-	+	+	+	w	+	+
56	-	-	-	-	-	-	-	-	-	-	-	-
57	+	+	+	+	+	+	-	-	-	-	-	-
58	+	+	+	+	+	+	-	-	-	-	-	-
59	+	+	+	+	+	+	-	-	-	-	-	-
60	+	+	+	+	+	+	+	+	+	+	+	+
61	-	-	-	-	-	-	-	-	-	-	-	-
62	-	-	-	w	w	+	-	-	-	?	?	?
63	+	+	+	+	+	+	+	w	+	+	+	+
64	+	+	+	+	+	+	-	-	-	+	+	+
65	-	-	-	-	-	-	-	-	-	-	-	-
66	+	+	+	+	+	+	+	+	+	+	+	+
67	+	+	+	+	+	+	+	+	+	+	+	+
68	-	-	-	-	-	-	-	-	-	-	-	-
69	+	+	+	+	+	+	-	-	-	-	-	-
70	-	-	-	-	w	?	+	+	+	+	+	+
71	-	-	-	-	w	?	-	-	-	-	-	-
72	-	-	-	-	w	?	+	+	+	+	+	+
73	+	+	+	+	+	+	-	-	-	-	-	-
74	-	-	-	-	w	?	-	-	-	-	-	-
75	-	-	-	-	-	-	-	-	-	-	-	-
76	-	-	-	-	-	-	-	-	-	-	-	-
77	-	-	-	-	w	?	-	-	-	-	-	-
78	+	+	+	+	+	+	-	-	-	-	-	-
79	+	+	+	+	+	+	-	-	-	-	-	-
80	-	-	-	w	w	+	-	-	-	-	-	-
81	-	-	-	w	-	-	-	-	-	-	-	-
82	+	+	+	+	+	+	-	-	-	-	-	-
83	-	-	-	-	-	-	-	-	-	-	w	?
84	-	-	-	-	-	-	+	-	?	+	+	+
85	-	-	-	-	-	-	-	-	-	-	-	-
86	+	+	+	+	+	+	-	-	-	-	-	-
87	+	+	+	+	+	+	+	+	+	+	+	+
88	-	-	-	-	-	-	-	-	-	-	-	-
89	-	-	-	-	w	?	-	-	-	w	-	?
90	-	-	-	-	-	-	-	-	-	-	-	-
91	-	-	-	-	-	-	-	-	-	-	-	-
	0		1	12		3	1		1	4		0

Hybrid	H201	H201	H201
	D1	D2	F
1	-	-	-
2	-	-	-
3	-	-	-
4	-	-	-
5	-	-	-
6	-	-	-
7	+	+	+
8	-	-	-
9	-	-	-
10	+	+	+
11	-	-	-
12	-	-	-
13	-	-	-
14	-	-	-
15	+	+	+
16	-	-	-
17	+	+	+
18	+	+	+
19	-	-	-
20	-	-	-
21	-	-	-
22	w	-	?
23	-	-	-
24	-	-	-
25	-	-	-
26	-	-	-
27	-	-	-
28	-	-	-
29	-	-	-
30	-	-	-
31	-	-	-
32	-	-	-
33	-	-	-
34	-	-	-
35	-	-	-
36	-	-	-
37	-	-	-
38	-	-	-
39	-	-	-
40	-	-	-
41	-	-	-
42	-	-	-
43	-	-	-
44	-	-	-
45	-	-	-
46	-	-	-
47	+	+	+
48	+	+	+
49	-	-	-
50	+	+	+
51	+	+	+
52	+	+	+
53	-	-	-
54	+	+	+
55	-	-	-
56	-	-	-
57	+	+	+
58	-	-	-
59	-	-	-
60	w	-	-
61	-	-	-
62	w	w	+
63	-	-	-
64	+	+	+
65	-	-	-
66	-	-	-
67	-	-	-
68	-	-	-
69	w	-	?
70	+	+	+
71	w	-	?
72	+	+	+
73	-	-	-
74	-	-	-
75	-	-	-
76	-	-	-
77	-	-	-
78	-	-	-
79	-	-	-
80	w	-	?
81	w	-	?
82	+	+	+
83	-	-	-
84	-	-	-
85	-	-	-
86	-	-	-
87	-	-	-
88	-	-	-
89	-	-	-
90	-	-	-
91	-	-	-
	5		1

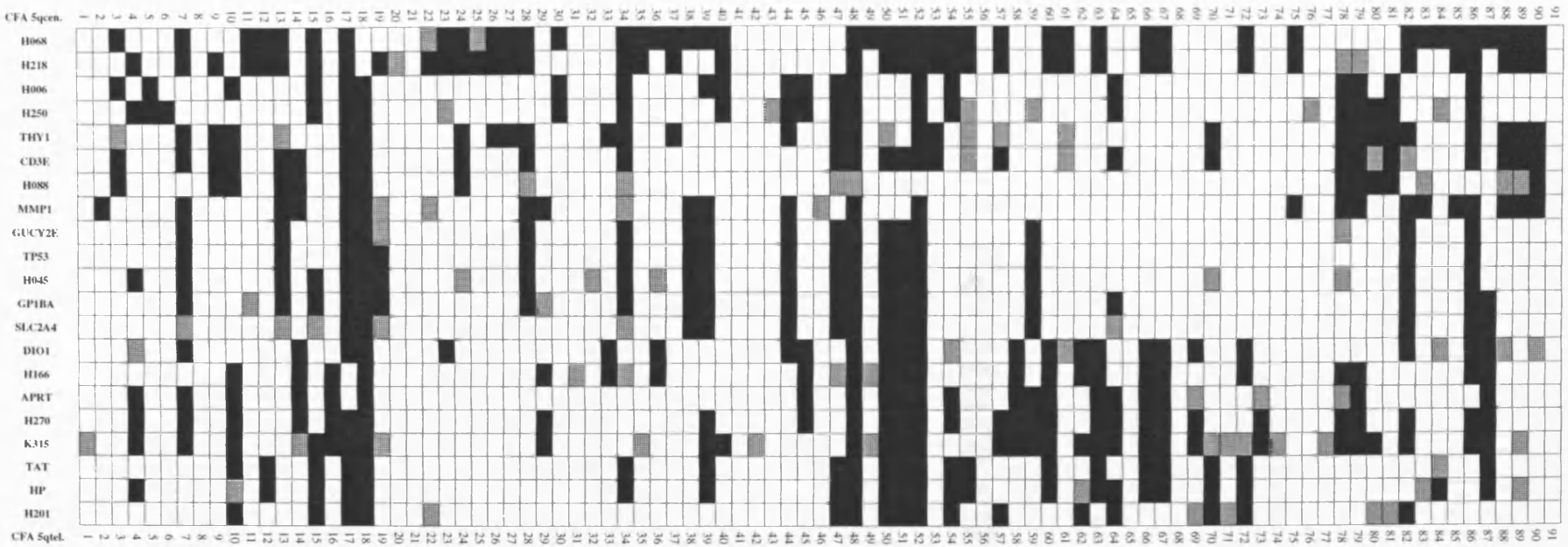


Figure Fi:

Schematic representation of the retention of CFA 5 markers within the WG-RH panel. Markers are listed vertically on the left of the figure, in order of their physical location on CFA 5, beginning with the most proximal marker. Each of the 91 hybrids within the WG-RH panel is represented along the horizontal axis. Where a given locus was identified as present within a hybrid, the corresponding block within the chart is filled in black. Blocks are filled in grey if the scoring of that hybrid for the corresponding locus was ambiguous (see chapter five text), otherwise blocks are left blank. This figure therefore provides an assessment of the proportion of CFA 5 retained by each hybrid cell line.

Literature cited

Aaronson, J.S., Eckman, B., Blevins, R.A., Borkowski, J.A., Myerson, J., Imran, S. and Elliston, K.O. (1996). Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Research* **6** 829-845.

Adams, M.D., Bento Soares, M., Kerlavage, A.R., Fields, C. and Venter, J.C. (1993b). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genetics* **4** 373-380.

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F. (2000). The genome sequence of *Drosophila melanogaster*. *Science*.**287** 2185-95.

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R.F., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C. and Venter, J.C. (1992a). Sequence identification of 2,375 human brain genes. *Nature* **355** 632-634.

Adams, M.D., Fields, C. and Venter, J.C. (1992b). Introns in sequence tags-reply. *Nature* **357** 367-368.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., *et al.* (1991). Complementary DNA sequencing: expressed sequence tags and the human genome project. *Science* **252** 1651-1656.

Adams, M.D., Kerlavage, A.R., Fields, C. and Venter, J.C. (1993a). 3400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature Genetics* **4** 256-267.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377 (supp)** 3-17.

Aguirre, G.D., Baldwin, V., Pearce-Kelling, S., Narfström, K., Ray, K. and Acland, G.M. (1998). Congenital stationary night blindness in the dog- common mutation in the RPE65 gene indicates founder effect. *Molecular Vision* **4** 23-29.

Aleyasin, A. and Barendse, W. (1997). Novel conserved synteny between human chromosome 22 and cattle chromosome 22 established by linkage mapping of transducin alpha-1 subunit (GNAZ). *Mammalian Genome* **8** 458-459.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology* **215** 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25** 3389-3402.

Anon (1993). Genome databases worry about yeast (and other) infections. *Science* **259** 1685.

Antonarakis, S.E. (1994). Genome linkage scanning: systematic or intelligent? *Nature Genetics* **8** 211-212.

Antonucci, T.K., Von Deimling, O.H., Rosenblum, B.B., Skow, L.C. and Meisler, M.H. (1984). Conserved linkage within a 4-cM region of mouse chromosome 9 and human chromosome 11. *Genetics* **107** 463-75.

Archer, B.T., Ozçelik, T., Jahn, R., Francke, U. and Südhof, T.C. (1990). Structures and chromosomal localizations of two human genes encoding synaptobrevins 1 and 2. *Journal of Biological Chemistry* **265** 17267-17273.

Austruy, E., Jeanpierre, C., Antignac, C., Whitmore, S.A., Van Cong, N., Bernheim, A., Callen, D.F. and Junien, C. (1993). Physical and genetic mapping of the dipeptidase gene DPEP1 to 16q24.3. *Genomics* **15** 684-687.

Barnes, W.M. (1994). PCR amplification of up to 35kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proceedings of the National Academy of Sciences of the United States of America* **91** 2216-2220.

Barton, D.E., Yang-Feng, T.L. and Francke, U. (1986). The human tyrosine aminotransferase gene mapped to the long arm of chromosome 16 (region 16q22-q24) by somatic cell hybrid analysis and *in situ* hybridization. *Human Genetics* **72** 221-224.

Benham, F., Hart, K., Crolla, J., Bobrow, M., Francavilla, M. and Goodfellow, P.N. (1989). A method for generating hybrids containing nonselected fragments of human chromosomes. *Genomics* **4** 509-517.

Bentolila, S., Bach, J.M., Kessler, J.L., Bordelais, I., Cruaud, C., Weissenbach, J. and Panthier, J.J. (1999). Analysis of major repetitive DNA sequences in the dog (*Canis familiaris*) genome. *Mammalian Genome* **10** 699-705.

Biamonti, G., Ruggiu, M., Saccone, S., Della Valle, G. and Riva, S. (1994). Two homologous genes, originated by duplication, encode the human hnRNP proteins A2 and A1. *Nucleic Acids Research* **22** 1996-2002.

Bielec, P.E., Gallagher, D.S., Womack, J.E. and Busbee, D.L. (1998). Homologies between human and dolphin chromosomes detected by heterologous chromosome painting. *Cytogenetics and Cell Genetics* **81** 18-25.

Bishop, J.O., Morton, J.G., Rosbash, M. and Richardson, M. (1974). Three abundance classes in HeLa cell messenger RNA. *Nature* **250** 199-204.

Boehnke, M., Lange, K. and Cox, D.R. (1991). Statistical methods for multipoint radiation hybrid mapping. *American Journal of Human Genetics* **49** 1174-1188.

Boeke, J.D. (1997). LINEs and *Alus*: the poly A connection. *Nature Genetics* **16** 6-7.

Boguski, M.S., Lowe, T.M.J. and Tolstoshev, C.M. (1993). dbEST- database for "expressed sequence tags". *Nature Genetics* **4** 332-333.

Boguski, M.S. and Schuler, G.D. (1995). ESTablishing a human transcript map. *Nature Genetics* **10** 369-370.

Breen, M., Bullerdiek, J. and Langford, C.F. (1999b). The DAPI-banded karyotype of the domestic dog (*Canis familiaris*) generated using chromosome-specific paint probes. *Chromosome Research* **7** 401-406.

Breen, M., Downs, P., Irvin, Z. and Bell, K. (1994). Intrageneric amplification of horse microsatellite markers with emphasis on the Przewalski's horse (*E. przewalskii*). *Animal Genetics* **25** 401-405.

Breen, M., Reimann, N., Bosma, A.A., Landon, D., Zijlstra, S., Bartnitzke, S., Switonski, M., Long, S.E., de Haan, N.A., Binns, M.M., *et al.* (1998). Standardisation of the chromosome nos. 22-38 of the dog (*Canis familiaris*) with the use of chromosome painting probes. *Proceedings of the 13th European Colloquium on Cytogenetics of Domestic Animals (Budapest)* .

Breen, M., Thomas, R., Binns, M.M., Carter, N.P. and Langford, C.F. (1999a). Reciprocal chromosome painting reveals detailed regions of conserved synteny between the karyotypes of the domestic dog (*Canis familiaris*) and human. *Genomics* **61** 145-155.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B. and Aparicio, S. (1993). Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* **366** 265-268.

Bürglin, T.R. and Barnes, T.M. (1992). Introns in sequence tags. *Nature* **357** 367.

Callen, D.F., Doggett, N.A., Stallings, R.L., Chen, L.Z., Whitmore, S.A., Lane, S.A., Nancarrow, J.K., Apostolou, S., Thompson, A.D., Lapsys, N.M., *et al.* (1992). High-resolution cytogenetic-based physical map of human chromosome 16. *Genomics* **13** 1178-1185.

Carter, N.P. (1994). Cytogenetic analysis by chromosome painting. *Cytometry* **18** 2-10.

Ceccherini, I., Romeo, G., Lawrence, S., Breuning, M.H., Harris, P.C., Himmelbauer, H., Frischauf, A.M., Sutherland, G.R., Germino, G.G., Reeders, S.T., *et al.* (1992). Construction of a map of chromosome 16 by using radiation hybrids. *Proceedings of the National Academy of Sciences of the United States of America* **89** 104-108.

Chaudhary, R., Kijas, J., Raudsepp, T., Guan, X.-Y., Zhang, H. and Chowdhary, B.P. (1998a). Microdissection of pig chromosomes: dissection of whole chromosomes, arms and bands for construction of paints and libraries. *Hereditas* **128** 265-271.

Chaudhary, R., Raudsepp, T., Guan, X.Y., Zhang, H. and Chowdhary, B.P. (1998b). Zoo-FISH with microdissected arm specific paints for HSA 2, 5, 6, 16, and 19 refines known homology with pig and horse chromosomes. *Mammalian Genome* **9** 44-49.

Chaudhary, R., Winterø, A.K., Fredholm, M. and Chowdhary, B.P. (1997). FISH mapping of seven cDNA sequences in the pig. *Chromosome Research* **5** 545-9.

Chen, Z.Q., Lautenberger, J.A., Lyons, L.A., McKenzie, L. and O'Brien, S.J. (1999). A human genome map of comparative anchor tagged sequences. *Journal of Heredity* **90** 477-484.

Chen-Liu, L.W., Huang, B.C., Scalzi, J.M., Hall, B.K., Sims, K.R., Davis, L.M., Siebert, P.D. and Hozier, J.C. (1995). Selection of hybrids by affinity capture (SHAC): a method for the generation of cDNAs enriched in sequences from a specific chromosome region. *Genomics* **30** 388-392.

Cho, K.-W., Okuda, M., Endo, Y., Satoh, H., Kang, C.-B., Watari, T., Tsujimoto, H. and Hasegawa, A. (1997b). Assignment of the cat p53 tumour suppressor gene (TP53) to cat chromosome E1p14-p13 by fluorescence *in situ* hybridisation. *Cytogenetics and Cell Genetics* **79** 145-146.

Cho, K.-W., Youn, H.-Y., Watari, T., Tsujimoto, H., Hasegawa, A. and Satoh, H. (1997a). A proposed nomenclature of the domestic cat karyotype. *Cytogenetics and Cell Genetics* **79** 71-78.

Chowdhary, B.P., Frönicke, L., Gustavsson, I. and Scherthan, H. (1996). Comparative analysis of the cattle and human genomes: Detection of Zoo-FISH and gene mapping-based chromosomal homologies. *Mammalian Genome* **7** 297-302.

Chowdhary, B.P. and Raudsepp, T. (2000). HSA 4 and GGA4: Remarkable conservation despite 300-Myr divergence. *Genomics* **64** 102-105.

Chowdhary, B.P., Raudsepp, T., Frönicke, L. and Scherthan, H. (1998). Emerging patterns of comparative genome organisation in some mammalian species as revealed by Zoo-FISH. *Genome Research* **8** 577-589.

Christian, A.T., Garcia, H.E. and Tucker, J.D. (1999). PCR *in situ* followed by microdissection allows whole chromosome painting probes to be made from single microdissected chromosomes. *Mammalian Genome* **10** 628-631.

Christodoulou, K., Tsingis, M., Deymeer, F., Serdaroglu, P., Ozdemir, C., Al-Shehab, A., Bairactaris, C., Mavromatis, I., Mylonas, I., Evoli, A., *et al.* (1997). Mapping of the familial infantile myasthenia (congenital myasthenic syndrome type Ia) gene to chromosome 17p with evidence of genetic homogeneity. *Human Molecular Genetics* **6** 635-40.

Chu, L.L., Rutteman, G.R., Kong, J.M.C., Ghahremani, M., Misdorp, W., van Garderen, E. and Pelletier, J. (1998). Genomic organisation of the canine p53 gene and its mutational status in canine mammary neoplasia. *Breast Cancer Research and Treatment* **50** 11-25.

Chung, D.W., Harris, J.E. and Davie, E.W. (1991). Nucleotide sequences of the three genes coding for human fibrinogen. In: Fibrinogen, Thrombosis, Coagulation and Fibrinolysis. C. Y. Liu and S. Chien (eds.) New York, Plenum Press, 39-48.

Church, G.M. and Gilbert, W. (1984). Genomic sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **81** 1991-1995.

Cox, D.R. (1992). Radiation hybrid mapping. *Cytogenetics and Cell Genetics* **59** 80-81.

Cox, D.R., Burmeister, M., Royden Price, E., Kim, S. and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell method for constructing high-resolution maps of mammalian chromosomes. *Science* **250** 245-250.

Craig, J.M. and Bickmore, W.A. (1994). The distribution of CpG islands in mammalian chromosomes. *Nature Genetics* **7** 376-382.

Cremer, T., Lichter, P., Borden, J., Ward, D.C. and Manuelidis, L. (1988). Detection of chromosome aberrations in metaphase and interphase tumor cells by *in situ* hybridization using chromosome-specific library probes. *Human Genetics* **80** 235-246.

Cybulsky, M.I., Fries, J.W., Williams, A.J., Sultan, P., Eddy, R., Byers, M., Shows, T., Gimbrone, M.A.J. and Collins, T. (1991). Gene structure, chromosomal location, and basis for alternative mRNA splicing of the human VCAM1 gene. *Proceedings of the National Academy of Sciences of the United States of America* **88** 7859-7863.

Das, M., Chu, L.L., Ghahremani, M., Abrams-Ogg, T., Roy, M.S., Housman, D. and Pelletier, J. (1998). Characterization of an abundant short interspersed nuclear element (SINE) present in *Canis familiaris*. *Mammalian Genome* **9** 64-69.

de Fatima Bonaldo, M., Lennon, G. and Bento Soares, M. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome* **6** 791-806.

Dean, M. and Allikmets, R. (1995). Contamination of cDNA libraries and expressed sequence tags databases. *American Journal of Human Genetics* **57** 1254-1255.

Del Mastro, R.G., Wang, L., Simmons, A.D., Gallardo, T.D., Clines, G.A., Ashley, J.A., Hilliard, C.J., Wasmuth, J.J., McPherson, J.D. and Lovett, M. (1995). Human chromosome-specific cDNA libraries: new tools for gene identification and genome annotation. *Genome Research* **5** 185-94.

Deloukas, P., Schuler, G., Gyapay, G., Beasley, E., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T., McKusick, K., Beckmann, J., *et al.* (1998). A physical map of 30,000 human genes. *Science* **282** 744-746.

DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A. and Trent, J.M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics* **14** 457-460.

Deschenes, S.M., Puck, J.M., Dutra, A.M., Somberg, R.L., Felsburg, P.J. and Henthorn, P.S. (1994). Comparative mapping of canine and human proximal Xq and genetic analysis of canine X-linked severe combined immunodeficiency. *Genomics* **23** 62-68.

Diatchenko, L., Lau, Y.-F.C., Campbell, A.P., Chenchik, A., Moqadam, F., Huang, B., Lukyanov, S., Lukyanov, K., Gurskaya, N., Sverdlov, E.D., *et al.* (1996). Suppression subtractive hybridisation: a method for generating differentally regulated or tissue-specific cDNA probes and libraries. *Proceedings of the National Academy of Sciences of the United States of America* **93** 6025-6030.

Dib, C., Fauré, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., *et al.* (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380** 152-154.

Don, R.H., Cox, P.T., Wainwright, B.J., Baker, K. and Mattick, J.S. (1991). 'Touchdown' PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Research* **19** 4008.

Dunham, I., Shimizu, N., Roe, B.A., Chissoe, S., Hunt, A.R., Collins, J.E., Bruskiewich, R., Beare, D.M., Clamp, M., Smink, L.J., *et al.* (1999). The DNA sequence of human chromosome 22. *Nature* **402** 489-495.

Durkin, A.S., Nierman, W.C., Zoghbi, H., Jones, C., Kozak, C.A. and Maglott, D.R. (1994). Chromosome assignment of human brain expressed sequence tags (ESTs) by analysing fluorescently labelled PCR products from hybrid cell panels. *Cytogenetics and Cell Genetics* **65** 86-91.

Dutra, A.S., Mignot, E. and Puck, J.M. (1996). Gene localisation and syntenic mapping by FISH in the dog. *Cytogenetics and Cell Genetics* **74** 113-117.

Dutrillaux, B. and Couturier, J. (1983). The ancestral karyotype of Carnivora: comparison with that of platyrrhine monkeys. *Cytogenetics and Cell Genetics* **35** 200-208.

Ellegren, H., Fredholm, M., Edfors-Lilja, I., Winterø, A.K. and Andersson, L. (1993). Conserved synteny between pig chromosome 8 and human chromosome 4 but rearranged and distorted linkage maps. *Genomics* **17** 599-603.

Fan, Y.-S., Eddy, R.L., Byers, M.G., Haley, L.L., Henry, W.M., Kayano, T., Shows, T.B. and Bell, G.I. (1989). Assignment of genes encoding three human glucose transporter/transporter-like proteins (GLUT4, GLUT5 and GLUT6) to chromosomes 17, 1 and 5, respectively. (Abstract). *Cytogenetics and Cell Genetics* **51** 997.

Fanning, T.G. (1989). Molecular evolution of centromere-associated nucleotide sequences in two species of canids. *Gene* **85** 559-563.

Fanning, T.G. and Singer, M.F. (1987). LINE-1: a mammalian transposable element. *Biochimica et Biophysica Acta* **910** 203-212.

Fields, C., Adams, M.D., White, O. and Venter, J.C. (1994). How many genes in the human genome? *Nature Genetics* **7** 345-349.

Fischer, P.E., Holmes, N.G., Dickens, H.F., Thomas, R., Binns, M.M. and Nacheva, E.P. (1996). The application of FISH techniques for physical mapping in the dog (*Canis familiaris*). *Mammalian Genome* **7** 37-41.

Florijn, R.J., van de Rijke, F.M., Vrolijk, H., Blonden, L.A.J., Hofker, M.H., den Dunnen, J.T., Tanke, H.J., van Ommen, G.-J.B. and Raap, A.K. (1996). Exon mapping by fibre-FISH. *Genomics* **38** 277-282.

Foster, J.W., Schafer, A.J., Critcher, R., Spillett, D.J., Feakes, R.W., Walter, M.A., Dominguez-Steglich, M., Guiolo, S., Brook, J.D. and Goodfellow, P.N. (1996). A high-resolution whole genome radiation hybrid map of human chromosome 17q22-q25.3 across the genes for GH and TK. *Genomics* **33** 185-192.

Fournier, R.E. and Ruddle, F.H. (1977). Microcell-mediated transfer of murine chromosomes into mouse, Chinese hamster, and human somatic cells. *Proceedings of the National Academy of Sciences of the United States of America* **74** 319-323.

Francke, U. (1994). Digitised and differentially shaded human chromosome ideograms for genomic applications. *Cytogenetics and Cell Genetics* **65** 206-219.

Francke, U. and Taggart, R.T. (1980). Comparative gene mapping: order of loci on the X chromosome is different in mice and humans. *Proceedings of the National Academy of Sciences of the United States of America* **77** 3595-3599.

Fratini, A., Simmers, R.N., Callen, D.F., Hyland, V.J., Tischfield, J.A., Stambrook, P.J. and Sutherland, G.R. (1986). A new location for the human adenine phosphoribosyltransferase gene (APRT) distal to the haptoglobin (HP) and fra(16)(q23)(FRA16D) loci. *Cytogenetics and Cell Genetics* **43** 10-13.

Fredholm, M. and Winterø, A.K. (1995). Variation of short tandem repeats within and between species belonging to the *Canidae* family. *Mammalian Genome* **6** 11-18.

Fridolfsson, A.K., Hori, T., Winterø, A.K., Fredholm, M., Yerle, M., Robic, A., Andersson, L. and Ellegren, H. (1997). Expansion of the pig comparative map by expressed sequence tags (EST) mapping. *Mammalian Genome* **8** 907-912.

Frigerio, J.-M., Berthezene, P., Garrido, P., Ortiz, E., Barthellemy, S., Vasseur, S., Sastre, B., Seleznieff, I., Dagorn, J.-C. and Iovanna, J.L. (1995). Analysis of 2166 clones from a human colorectal cancer cDNA library by partial sequencing. *Human Molecular Genetics* **4** 37-43.

Frönicke, L., Chowdhary, B.P., Scherthan, H. and Gustavsson, I. (1996). A comparative map of the porcine and human genomes demonstrates Zoo-FISH and gene mapping-based chromosomal homologies. *Mammalian Genome* **7** 285-290.

Frönicke, L., Müller-Navia, J., Romanakis, K. and Scherthan, H. (1997). Chromosomal homeologies between human, harbor seal (*Phoca vitulina*) and the putative ancestral carnivore karyotype revealed by Zoo-FISH. *Chromosoma* **106** 108-113.

Gallagher, D.S.J. and Womack, J.E. (1992). Chromosome conservation in the Bovidae. *Journal of Heredity* **83** 287-298.

Gantz, I., Yamada, T., Tashiro, T., Konda, Y., Shimoto, Y., Miwa, H. and Trent, J.M. (1994). Mapping of the gene encoding the melanocortin-1 (alpha-melanocyte stimulating hormone) receptor (MC1R) to human chromosome 16q24.3 by fluorescence *in situ* hybridization. *Genomics* **19** 394-5.

Gilley, J. and Fried, M. (1999). Extensive gene order differences within regions of conserved synteny between the *Fugu* and human genomes: implications for chromosomal evolution and the cloning of disease genes. *Human Molecular Genetics* **8** 1313-1320.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., *et al.* (1996). Life with 6000 genes. *Science* **274** 563-567.

Gold, D.P., van Dongen, J.J., Morton, C.C., Bruns, G.A., van den Elsen, P., Geurts van Kessel, A.H. and Terhorst, C. (1987). The gene encoding the epsilon subunit of the T3/T-cell receptor complex maps to chromosome 11 in humans and to chromosome 9 in mice. *Proceedings of the National Academy of Sciences of the United States of America* **84** 1664-1668.

Goodman, L. (1997). More BLAST for the buck. *Genome Research* **7** 858-859.

Gorski, J.L., Boehnke, M., Reyner, E.L. and Burright, E.N. (1992). A radiation hybrid map of the proximal short arm of the human X chromosome spanning incontinentia pigmenti 1 (IP1) translocation breakpoints. *Genomics* **14** 657-665.

Goss, S.J. and Harris, H. (1975). New method for mapping genes in human chromosomes. *Nature* **255** 680-684.

Goureau, A., Yerle, M., Schmitz, A., Riquet, J., Milan, D., Pinton, P., Frelat, G. and Gellin, J. (1996). Human and porcine correspondence of chromosome segments using bidirectional chromosome painting. *Genomics* **36** 252-262.

Gracia, E., Fischer, U., ElKahloun, A., Trent, J.M., Meese, E. and Meltzer, P.S. (1996). Isolation of genes amplified in human cancers by microdissection-mediated cDNA capture. *Human Molecular Genetics* **5** 595-600.

Gracia, E., Ray, M.E., Polymeropoulos, M.H., Dehejia, A., Meltzer, P.S. and Trent, J.M. (1997). Isolation of chromosome-specific ESTs by microdissection-mediated cDNA capture. *Genome* **7** 100-107.

Graphodatsky, A.S., Beklemisheva, V.R. and Dolf, G. (1995). High-resolution GTG-banding patterns of dog and silver fox chromosomes: description and comparative analysis. *Cytogenetics and Cell Genetics* **69** 226-231.

Graves, J.A. (1996). Mammals that break the rules: genetics of marsupials and monotremes. *Annual Review of Genetics* **30** 233-260.

Gu, S.M., Thompson, D.A., Srikumari, C.R., Lorenz, B., Finckh, U., Nicoletti, A., Murthy, K.R., Rathmann, M., Kumaramanickavel, G., Denton, M.J., *et al.* (1997). Mutations in RPE65 cause autosomal recessive childhood-onset severe retinal dystrophy. *Nature Genetics* **17** 194-197.

Gu, W.K. and Ray, K. (1997). A polymorphic (TTTA)n tandem repeat in an intron of the canine factor IX gene. *Animal Genetics* **28** 370.

Gubler, U. and Hoffman, B.J. (1983). A simple and very efficient method for generating cDNA libraries. *Gene* **25** 263-269.

Guevara-Fujita, M.L., Loechel, R., Venta, P.J., Yuzbasiyan-Gurkan, V. and Brewer, G.J. (1996). Chromosomal assignment of seven genes on canine chromosomes by fluorescence *in situ* hybridisation. *Mammalian Genome* **7** 268-270.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M. and Weissenbach, J. (1994). The 1993-94 Généthon human genetic linkage map. *Nature Genetics* **7** 246-339.

Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillet, D., Muselet, D., Prud'Homme, J.-F., Dib, C., Auffray, C., *et al.* (1996). A radiation hybrid map of the human genome. *Human Molecular Genetics* **5** 339-346.

Haldane, J. (1919). The combination of linkage values, and the calculation of distance between the loci of linked factors. *Journal of Genetics* **8** 299-309.

Hameister, H., Klett, C.H., Bruch, J., Dixkens, C.H., Vogel, W. and Christensen, K. (1997). Zoo-FISH analysis: the American Mink (*Mustela vison*) closely resembles the cat karyotype. *Chromosome Research* **5** 5-11.

Hamel, C.P., Jenkins, N.A., Gilbert, D.J., Copeland, N.G. and Redmond, T.M. (1994). The gene for the retinal pigment epithelium-specific protein RPE65 is localized to human 1p31 and mouse 3. *Genomics* **20** 509-512.

Hawkins, J.D. (1988). A survey on intron and exon lengths. *Nucleic Acids Research* **16** 9893-9905.

Hayes, H. (1995). Chromosome painting with human chromosome-specific DNA libraries reveals the extent and distribution of conserved segments in bovine chromosomes. *Cytogenetics and Cell Genetics* **71** 168-174.

Hayes, P.D., Schmitt, K., Jones, H.B., Gyapay, G., Weissenbach, J. and Goodfellow, P.N. (1996). Regional assignment of human ESTs by whole-genome radiation hybrid mapping. *Mammalian Genome* **7** 446-450.

Hediger, R., Ansari, H.A. and Stranzinger, G.F. (1991). Chromosome banding and gene localizations support extensive conservation of chromosome structure between cattle and sheep. *Cytogenetics and Cell Genetics* **57** 127-134.

Heiskanen, M., Karhu, R., Hellsten, E., Peltonen, L., Kallioniemi, O.P. and Palotie, A. (1994). High resolution mapping using fluorescence *in situ* hybridisation to extended DNA fibres prepared from agarose-embedded cells. *BioTechniques* **17** 928-933.

Heiskanen, M., Peltonen, L. and Palotie, A. (1996). Visual mapping by high resolution FISH. *Trends in Genetics* **12** 379-380.

Helou, K., Das, A.T., Lamers, W.H., Hoovers, J.M., Szpirer, C., Szpirer, J., Klinga-Levan, K. and Levan, G. (1997). FISH mapping of three ammonia metabolism genes (Glul, Cps1, Glud1) in rat, and the chromosomal localization of GLUL in human and Cps1 in mouse. *Mammalian Genome* **8** 362-364.

Henthorn, P.S., Somberg, R.L., Fimiani, V.M., Puck, J.M., Patterson, D.F. and Felsburg, P.J. (1994). IL-2Rgamma gene microdeletion demonstrates that canine X-linked severe combined immunodeficiency is a homologue of the human disease. *Genomics* **23** 69-74.

Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., *et al.* (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Research* **6** 807-28.

Holmes, N.G., Dickens, H.F., Parker, H.L., Binns, M.M., Mellersh, C.S. and Sampson, J. (1995). Eighteen canine microsatellites. *Animal Genetics* **26** 132-133.

Holmes, N.G., Mellersh, C.S., Humphreys, S.J., Binns, M.M., Holliman, A., Curtis, R. and Sampson, J. (1993). Isolation and characterisation of microsatellites from the canine genome. *Animal Genetics* **24** 289-292.

Horelli-Kuitunen, N., Aaltonen, J., Yaspo, M.L., Eeva, M., Wessman, M., Peltonen, L. and Palotie, A. (1999). Mapping ESTs by fiber-FISH. *Genome Research* **9** 62-71.

Howard-Peebles, P.N. and Pryor, J.C. (1980). The R-banding pattern of the canine karyotype. *Journal of Heredity* **71** 361-362.

Hozier, J., Graham, R., Westfall, T., Siebert, P. and Davis, L. (1994). Preparative *in situ* hybridization: selection of chromosome region-specific libraries on mitotic chromosomes. *Genomics* **19** 441-447.

Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., Silva, J., Slonim, D.K., Baptista, R., Kruglyak, L., Xu, S.H., *et al.* (1995). An STS-based map of the human genome. *Science* **270** 1945-1954.

The Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72** 971-83.

Hwang, D.M., Fung, Y.W., Wang, R.X., Laurenssen, C.M., Ng, S.H., Lam, W.Y., Tsui, K.W., Fung, K.P., Waye, M., Lee, C.Y., *et al.* (1995). Analysis of expressed sequence tags from a fetal human heart cDNA library. *Genomics* **30** 293-298.

Iannuzzi, L., Di Meo, G.P., Perucatti, A. and Incarnato, D. (1999). Comparison of the human with the sheep genomes by use of human chromosome-specific painting probes. *Mammalian Genome* **10** 719-723.

Inazawa, J., Fukunaga, R., Seto, Y., Nakagawa, H., Misawa, S., Abe, T. and Nagata, S. (1991). Assignment of the human granulocyte colony-stimulating factor receptor gene (CSF3R) to chromosome 1 at region p35-p34.3. *Genomics* **10** 1075-1078.

Isobe, M., Emanuel, B.S., Givol, D., Oren, M. and Croce, C.M. (1986). Localization of gene for human p53 tumour antigen to band 17p13. *Nature* **320** 84-85.

Jakobs, T.C., Koehler, M.R., Schmutzler, C., Glaser, F., Schmid, M. and Köhrle, J. (1997). Structure of the human type I iodothyronine 5'-deiodinase gene and localization to chromosome 1p32-p33. *Genomics* **42** 361-363.

Jauch, A., Wienberg, J., Stanyon, R., Arnold, N., Tofanelli, S., Ishida, T. and Cremer, T. (1992). Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. *Proceedings of the National Academy of Sciences of the United States of America* **89** 8611-8615.

Jay, P., Diriong, S., Taviaux, S., Roeckel, N., Mattéi, M.-G., Audit, M., Bergé-Lefranc, J.-L., Fontès, M. and Berta, P. (1997). Isolation and regional mapping of cDNAs expressed during early human development. *Genomics* **39** 104-108.

Jeremiah, S.J., Povey, S., Burley, M.W., Kielty, C., Lee, M., Spowart, G., Corney, G. and Cook, P.J. (1982). Mapping studies on human mitochondrial glutamate oxaloacetate transaminase. *Annals of Human Genetics* **46** 145-152.

Jiang, Z., Priat, C. and Galibert, F. (1998). Traced orthologous amplified sequence tags (TOASTs) and mammalian comparative maps. *Mammalian Genome* **9** 577-587.

Jones, H.B. (1996). Pairwise analysis of radiation hybrid mapping data. *Annals of Human Genetics* **60** 351-357.

Jones, H.B. (1997). Estimating physical distances from radiation hybrid mapping data. *Genomics* **43** 258-66.

Jorgensen, C.B., Winterø, A.K., Yerle, M. and Fredholm, M. (1997). Mapping of 22 expressed sequence tags isolated from a porcine small intestine cDNA library. *Mammalian Genome* **8** 423-427.

Joshi, A.R., Mullen, L. and Small, K.W. (1997). The retina: genetic studies of several retinopathies located on the short arm of chromosome 17. *Current Opinion in Neurology* **10** 31-35.

Kallioniemi, A., Kallioniemi, O.P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F. and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258** 818-821.

Khan, A.S., Wilcox, A.S., Polymeropoulos, M.H., Hopkins, J.A., Stevens, T.J., Robinson, M., Orpana, A.K. and Sikela, J.M. (1992). Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genetics* **2** 180-185.

Ko, M.S. (1990). An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs. *Nucleic Acids Research* **18** 5705-5711.

Koehler, U., Bigoni, F., Wienberg, J. and Stanyon, R. (1995). Genomic reorganization in the concolor gibbon (*Hylobates concolor*) revealed by chromosome painting. *Genomics* **30** 287-292.

Korenburg, J.R., Chen, X.-N., Adams, M.D. and Venter, J.C. (1995). Toward a cDNA map of the human genome. *Genomics* **29** 364-370.

Korn, B., Sedlacek, Z., Manca, A., Kioschis, P., Konecki, D., Lehrach, H. and Poustka, A. (1992). A strategy for the selection of transcribed sequences in the Xq28 region. *Human Molecular Genetics* **1** 235-242.

Korstanje, R., O'Brien, P.C., Yang, F., Rens, W., Bosma, A.A., van Lith, H.A., van Zutphen, L.F. and Ferguson-Smith, M.A. (1999). Complete homology maps of the rabbit (*Oryctolagus cuniculus*) and human by reciprocal chromosome painting. *Cytogenetics and Cell Genetics* **86** 317-322.

Kozu, T., Henrich, B. and Schafer, K.P. (1995). Structure and expression of the gene (HNRPA2B1) encoding the human hnRNP protein A2/B1. *Genomics* **25** 365-371.

Kukita, A., Yoshida, M.C., Fukushige, S., Sakakibara, M., Joh, K., Mukai, T. and Hori, K. (1987). Molecular gene mapping of human aldolase A (ALDOA) gene to chromosome 16. *Human Genetics* **76** 20-26.

Lahbib-Mansais, Y., Dalias, G., Milan, D., Yerle, M., Robic, A., Gyapay, G. and Gellin, J. (1999). A successful strategy for comparative mapping with human ESTs: 65 new regional assignments in the pig. *Mammalian Genome* **10** 145-153.

Lalley, P.A., Minna, J.D. and Francke, U. (1978). Conservation of autosomal gene synteny groups in mouse and man. *Nature* **274** 160-163.

Lander, E.S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84** 2363-2367.

Lange, K., Boehnke, M., Cox, D.R. and Lunetta, K.L. (1995). Statistical methods for polyploid radiation hybrid mapping. *Genome Research* **5** 136-150.

Langford, C.F., Fischer, P.E., Binns, M.M., Holmes, N.G. and Carter, N.P. (1996). Chromosome-specific paints from a high-resolution flow karyotype of the dog. *Chromosome Research* **4** 115-123.

Langston, A.A., Mellersh, C.S., Neal, C.L., Ray, K., Acland, G.M., Gibbs, M., Aguirre, G.D., Fournier, R.E.K. and Ostrander, E.A. (1997). Construction of a panel of canine-rodent hybrid cell lines for use in the partitioning of the canine genome. *Genomics* **46** 317-325.

Larsen, N.J., Hayes, H., Bishop, M., Davis, S.K., Taylor, J.F. and Kirkpatrick, B.W. (1999). A comparative linkage and physical map of bovine chromosome 24 with human chromosome 18. *Mammalian Genome* **10** 482-487.

Lennon, G., Auffray, C., Polymeropoulos, M. and Bento Soares, M. (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33** 151-152.

Levitt, R.C. (1991). Polymorphisms in the transcribed 3' untranslated region of eukaryotic genes. *Genomics* **11** 484-489.

Li, R., Mignot, E., Faraco, J., Kodatani, H., Cantanese, J.J., Zhao, B.H., Lin, X., Hinton, L., Ostrander, E.A., Patterson, D.F., *et al.* (1999). Construction and characterisation of an eightfold redundant dog genomic bacterial artificial chromosome library. *Genomics* **58** 9-17.

Lin, C.-T. and Sargan, D.R. (1997). A method for generating subtractive cDNA libraries retaining clones containing repetitive elements. *Nucleic Acids Research* **25** 4427-4428.

Lin, L., Faraco, J., Li, R., Kadotani, H., Rogers, W., Lin, X., Qiu, X., de Jong, P.J., Nishino, S. and Mignot, E. (1999). The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* **98** 365-76.

Lingaas, F., Sorensen, A., Juneja, R.K., Johansson, S., Fredholm, M., Winterø, A.K., Sampson, J., Mellersh, C.S., Curzon, A., Holmes, N.G., *et al.* (1997). Towards construction of a canine linkage map: establishment of 16 linkage groups. *Mammalian Genome* **8** 218-221.

Lisitsyn, N., Lisitsyn, N. and Wigler, M. (1993). Cloning the differences between two complex genomes. *Science* **259** 946-951.

Little, P. (1999). The book of genes. *Science* **402** 467-468.

Lovett, M. (1994a). Fishing for complements: finding genes by direct selection. *Trends in Genetics* **10** 352-357.

Lovett, M. (1994b). Direct selection of cDNAs using genomic contigs. In: Current Protocols in Human Genetics, N. C. Dracopoli (ed.) New York, John Wiley & Sons, 6.3.1-6.3.15.

Lovett, M., Kere, J. and Hinton, L.M. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proceedings of the National Academy of Sciences of the United States of America* **88** 9628-9632.

Lyons, L.A., Laughlin, T.F., Copeland, N.G., Jenkins, N.A., Womack, J.E. and O'Brien, S.J. (1997). Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics* **15** 47-56.

Ma, R.Z., van Eijk, M.J., Beever, J.E., Guérin, G., Mummery, C.L. and Lewin, H.A. (1998). Comparative analysis of 82 expressed sequence tags from a cattle ovary cDNA library. *Mammalian Genome* **9** 545-549.

MacEwen, E.G. (1990). Spontaneous tumors in dogs and cats: models for the study of cancer biology and treatment. *Cancer and Metastasis Reviews* **9** 125-36.

Mancini, M., Sala, C., Rivella, S. and Toniolo, D. (1996). Selection and fine mapping of chromosome-specific cDNAs: application to human chromosome 1. *Genomics* **38** 149-154.

Manolache, M., Ross, W.M. and Schmid, M. (1976). Banding analysis of the somatic chromosomes of the domestic dog (*Canis familiaris*). *Canadian Journal of Genetics and Cytology* **18** 513-518.

Matise, T.C., Wasmuth, J.J., Myers, R.M. and McPherson, J.D. (1999). Somatic cell genetics and radiation hybrid mapping. In: Genome Analysis: A Laboratory Manual. B. Birren, E. D. Green, P. Hieter, S. Klapholz, R. M. Myers, H. Riethman and J. Roskans (eds.) Cold Spring Harbor, Cold Spring Harbor Laboratory Press, 258-302.

Matthey, R. (1968). General survey of the chromosome numbers by eutherian mammals. *Mammalian Chromosomes Newsletter* **9** 17-18.

Mayr, B., Schweizer, D. and Schleger, W. (1983). Characterisation of the canine karyotype by counterstain-enhanced chromosome banding. *Canadian Journal of Genetics and Cytology* **25** 616-621.

McBride, O.W., Merry, D. and Givol, D. (1986). The gene for human p53 cellular tumor antigen is located on chromosome 17 short arm (17p13). *Proceedings of the National Academy of Sciences of the United States of America* **83** 130-134.

McCarthy, L. and Soderland, C. (1998). Radiation hybrid mapping. In: Guide to Human Genome Computing. M. J. Bishop (ed.) London, Academic Press, 113-149.

McCombie, W.R., Adams, M.D., Kelley, J.M., FitzGerald, M.G., Utterback, T.R., Khan, M., Dubnick, M., Kerlavage, A.R., Venter, J.C. and Fields, C. (1992). *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genetics* **1** 124-131.

McPherson, J.D., Apostol, B., Wagner-McPherson, C.B., Hakim, S., del Mastro, R.G., Aziz, N., Baer, E., Gonzales, G., Krane, M.C., Markovich, R., *et al.* (1997). A radiation hybrid map of human chromosome 5 with integration of cytogenetic, genetic and transcript maps. *Genome Research* **7** 897-909.

Mellersh, C.S., Hitte, C., Richman, M., Vignaux, F., Priat, C., Jouquand, S., Werner, P., André, C., DeRose, S., Patterson, D.F., *et al.* (2000). An integrated linkage-radiation hybrid map of the canine genome. *Mammalian Genome* **11** 120-130.

Mellersh, C.S., Langston, A.A., Acland, G.M., Fleming, M.A., Ray, K., Wiegand, N.A., Francisco, L.V., Gibbs, M., Aguirre, G.D. and Ostrander, E.A. (1997). A linkage map of the canine genome. *Genomics* **46** 326-336.

Miller, C., Mohandas, T., Wolf, D., Prokocimer, M., Rotter, V. and Koeffler, H.P. (1986). Human p53 gene localized to short arm of chromosome 17. *Nature* **319** 783-4.

Miller, L.M., Hegreberg, G.A., Prieur, D.J. and Hamilton, M.J. (1984). Inheritance of congenital myasthenia gravis in smooth fox terrier dogs. *Journal of Heredity* **75** 163-166.

Minghetti, P.P., Ruffner, D.E., Kuang, W.J., Dennison, O.E., Hawkins, J.W., Beattie, W.G. and Dugaiczyk, A. (1986). Molecular structure of the human albumin gene is revealed by nucleotide sequence within q11-22 of chromosome 4. *Journal of Biological Chemistry* **261** 6747-6757.

Minnick, M.F., Stillwell, L.C., Heineman, J.M. and Steigler, G.L. (1992). A highly repetitive DNA sequence possibly unique to Canids. *Gene* **110** 235-238.

Moore, S.S., Sargeant, L.L., King, T.J., Mattick, J.S., Georges, M. and Hetzel, D.J. (1991). The conservation of dinucleotide microsatellites among mammalian genomes allows the use of heterologous PCR primer pairs in closely related species. *Genomics* **10** 654-660.

Morgan, J.G., Dolganov, G.M., Robbins, S.E., Hinton, L.M. and Lovett, M. (1992). The selective isolation of novel cDNAs encoded by the regions surrounding the human interleukin 4 and 5 genes. *Nucleic Acids Research* **20** 5173-5179.

Morimura, H., Fishman, G.A., Grover, S.A., Fulton, A.B., Berson, E.L. and Dryja, T.P. (1998). Mutations in the RPE65 gene in patients with autosomal recessive retinitis pigmentosa or Leber congenital amaurosis. *Proceedings of the National Academy of Sciences of the United States of America* **95** 3088-3093.

Morrell, V. (1997). The origin of dogs: running with the wolves. *Science* **276** 1647-1648.

Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C. and Bernardi, G. (1991). The distribution of genes in the human genome. *Gene* **100** 181-187.

Muraro, T., Stephan, D., Tiso, N., Zimbello, R., Danieli, G.A., Hoffman, E.H., Valle, G. and Lanfranchi, G. (1997). Chromosome assignment of 115 expressed sequence tags (ESTs) from human skeletal muscle. *Cytogenetics and Cell Genetics* **76** 144-152.

Murphy, W.J., Menotti-Raymond, M., Lyons, L.A., Thompson, M.A. and O'Brien, S.J. (1999a). Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. *Genomics* **57** 1-8.

Murphy, W.J., Sun, S., Chen, Z.Q., Pecon-Slattery, J. and O'Brien, S.J. (1999b). Extensive conservation of sex chromosome organization between cat and human revealed by parallel radiation hybrid mapping. *Genome Research* **9** 1223-1230.

Nadeau, J.H. (1989). Maps of linkage and synteny homologies between mouse and man. *Trends in Genetics* **5** 82-86.

Nadeau, J.H. and Taylor, B.A. (1984). Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **81** 814-818.

Nagase, T., Ishikawa, K., Nakajima, D., Ohira, M., Seki, N., Miyajima, N., Tanaka, A., Kotani, H., Nomura, N. and Ohara, O. (1997). Prediction of the coding sequences of unidentified human genes. VII. The complete sequences of 100 new cDNA clones from brain which can code for large proteins *in vitro*. *DNA Research* **4** 141-150.

Narfström, K., Wrigstad, A. and Nilsson, S.E. (1989). The Briard dog: a new animal model of congenital stationary night blindness. *British Journal of Ophthalmology* **73** 750-756.

Nash, W.G. and O'Brien, S.J. (1982). Conserved regions of homologous G-banded chromosomes between orders in mammalian evolution: carnivores and primates. *Proceedings of the National Academy of Sciences of the United States of America* **79** 6631-6635.

Nash, W.G. and O'Brien, S.J. (1987). A comparative chromosome banding analysis of the *Ursidae* and their relationship to other carnivores. *Cytogenetics and Cell Genetics* **45** 206-212.

Neff, M.W., Bromann, K.W., Mellersh, C.S., Ray, K., Acland, G.M., Aguirre, G.D., Ziegle, J.S., Ostrander, E.A. and Rine, J. (1999). A second-generation genetic linkage map of the domestic dog, *Canis familiaris*. *Genetics* **151** 803-820.

Nelson, D.L., Ledbetter, S.A., Corbo, L., Victoria, M.F., Ramírez-Solis, R., Webster, T.D., Ledbetter, D.H. and Caskey, C.T. (1989). *Alu* polymerase chain reaction: a method for rapid isolation of human-specific sequences from complex DNA sources. *Proceedings of the National Academy of Sciences of the United States of America* **86** 6686-6690.

Nelson, P.S., Ng, W.-L., Schummer, M., True, L.D., Liu, A.Y., Bumgarner, R.E., Ferguson, C., Dimak, A. and Hood, L. (1998). An expressed-sequence-tag database of the human prostate: sequence analysis of 1168 cDNA clones. *Genomics* **47** 12-25.

Newton, J.M., Wilkie, A.L., He, L., Jordan, S.A., Metallinos, D.L., Holmes, N.G., Jackson, I.J. and Barsh, G.S. (2000). Melanocortin 1 receptor variation in the domestic dog. *Mammalian Genome* **11** 24-31.

Nickerson, E. and Nelson, D.L. (1998). Molecular definition of pericentric inversion breakpoints occurring during the evolution of humans and chimpanzees. *Genomics* **50** 368-372.

Nishino, S. and Mignot, E. (1997). Pharmacological aspects of human and canine narcolepsy. *Progress in Neurobiology* **52** 27-78.

Noguchi, M., Yi, H., Rosenblatt, H.M., Filipovich, A.H., Adelstein, S., Modi, W.S., McBride, O.W. and Leonard, W.J. (1993). Interleukin-2 receptor gamma chain mutation results in X-linked severe combined immunodeficiency in humans. *Cell* **73** 147-157.

O'Brien, S.J. (1986). Molecular genetics in the domestic cat and its relatives. *Trends in Genetics* 137-142.

O'Brien, S.J., Cevario, S.J., Martenson, J.S., Thompson, M.A., Nash, W.G., Chang, E., Graves, J.A.M., Spencer, J.A., Cho, K.-W., Tsujimoto, H., *et al.* (1997). Comparative gene mapping in the domestic cat (*Felis catus*). *Journal of Heredity* **88** 408-414.

O'Brien, S.J. and Nash, W.G. (1982). Genetic mapping in mammals: chromosome map of domestic cat. *Science* **216** 257-265.

O'Brien, S.J., Womack, J.E., Lyons, L.A., Moore, K.J., Jenkins, N.A. and Copeland, N.G. (1993). Anchored reference loci for comparative mapping in mammals. *Nature Genetics* **3** 103-112.

Ohno, S. (1973). Ancient linkage groups and frozen accidents. *Nature* **244** 259-262.

Ohno, S. (1984). Linkage group conservation and the notion of 24 primordial vertebrate linkage groups. *Chromosomes Today* **8** 268-278.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics* **2** 173-179.

Ostrander, E.A., Galibert, F. and Patterson, D.F. (2000). Canine genetics comes of age. *Trends in Genetics* **16** 117-124.

Ostrander, E.A., Mapa, F.A., Yee, M. and Rine, J. (1995). One hundred and one new simple sequence repeat-based markers for the canine genome. *Mammalian Genome* **6** 192-195.

Ostrander, E.A., Sprague, G.F. and Rine, J. (1993). Identification and characterisation of dinucleotide repeat (CA)n markers for genetic mapping in dog. *Genomics* **16** 207-213.

Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D. and Weissman, S.M. (1991). cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* **88** 9623-7.

Park, J.P. (1996). Shared synteny of human chromosome 17 loci in Canids. *Cytogenetics and Cell Genetics* **74** 133-137.

Patanjali, S.R., Parimoo, S. and Weissman, S.M. (1991). Construction of a uniform-abundance (normalised) cDNA library. *Proceedings of the National Academy of Sciences of the United States of America* **88** 1943-1947.

Patterson, D.F. (1993). Understanding and controlling inherited diseases in dogs and cats. *Tijdschrift voor Diergeneeskunde* **118** 23S-27S.

Patterson, D.F. (in press). Canine Genetic Disease Information System: a computerised knowledgebase of genetic diseases in dogs. St. Louis, Mosby-Harcourt.

Patterson, D.F., Haskins, M.E. and Jezyk, P.F. (1982). Models of human genetic disease in domestic animals. *Advances in Human Genetics* **12** 263-339.

Pendás, A.M., Santamaría, I., Alvarez, M.V., Pritchard, M. and López-Otín, C. (1996). Fine physical mapping of the human matrix metalloproteinase genes clustered on chromosome 11q22.3. *Genomics* **37** 266-268.

Pinkel, D., Landegent, J., Collins, C., Fuscoe, J., Segraves, R., Lucas, J. and Gray, J. (1988). Fluorescence *in situ* hybridization with human chromosome-specific libraries: detection of trisomy 21 and translocations of chromosome 4. *Proceedings of the National Academy of Sciences of the United States of America* **85** 9138-9142.

Piumi, F., Schibler, L., Vaiman, D., Oustry, A. and Cribiu, E.P. (1998). Comparative cytogenetic mapping reveals chromosome rearrangements between the X chromosomes of two closely related mammalian species (cattle and goats). *Cytogenetics and Cell Genetics* **81** 36-41.

Pol, S., Bousquet-Lemercier, B., Pavé-Preux, M., Bulle, F., Passage, E., Hanoune, J., Mattei, M.G. and Barouki, R. (1989). Chromosomal localization of human aspartate aminotransferase genes by *in situ* hybridization. *Human Genetics* **83** 159-164.

Polymeropoulos, M.H., Xiao, H., Glodek, A., Gorski, M., Adams, M.D., Moreno, R.F., Fitzgerald, M.G., Venter, J.C. and Merril, C.R. (1992). Chromosomal assignment of 46 brain cDNAs. *Genomics* **12** 492-496.

Polymeropoulos, M.H., Xiao, H., Sikela, J.M., Adams, M.D., Venter, J.C. and Merril, C.R. (1993). Chromosomal distribution of 320 genes from a brain cDNA library. *Nature Genetics* **4** 381-386.

Poulson, B.S., Shibasaki, Y., Ikeuchi, T. and Ronne, M. (1990). Banding studies in *Canis familiaris*. I. Replication patterns in karyotypes from lymphocyte cultures. *Cytobios* **62** 161-165.

Priat, C., Hitte, C., Vignaux, F., Renier, C., Jiang, Z., Jouquand, S., Chéron, A., André, C. and Galibert, F. (1998). A whole-genome radiation hybrid map of the dog genome. *Genomics* **54** 361-378.

Priat, C., Jiang, Z.H., Renier, C., André, C. and Galibert, F. (1999). Characterisation of 463 Type I markers suitable for dog genome mapping. *Mammalian Genome* **10** 803-813.

Puck, J.M., Deschênes, S.M., Porter, J.C., Dutra, A.S., Brown, C.J., Willard, H.F. and Henthorn, P.S. (1993). The interleukin-2 receptor gamma chain maps to Xq13.1 and is mutated in X-linked severe combined immunodeficiency, SCIDX1. *Human Molecular Genetics* **2** 1099-1104.

Qian, F., Kruse, U., Lichter, P. and Sippel, A.E. (1995). Chromosomal localisation of the four genes (NFIA, B, C and X) for the human transcription factor nuclear factor I by FISH. *Genomics* **28** 66-73.

Qumsiyeh, M.B. (1994). Evolution of number and morphology of mammalian chromosomes. *Journal of Heredity* **85** 455-465.

Raudsepp, T., Frönicke, L., Scherthan, H., Gustavsson, I. and Chowdhary, B.P. (1996). Zoo-FISH delineates conserved chromosomal segments in horse and man. *Chromosome Research* **4** 218-225.

Redmond, T.M., Yu, S., Lee, E., Dok, D., Hamasaki, D., Chen, N., Goletz, P., Ma, J.-X., Crouch, R.K. and Pfeifer, K. (1998). Rpe65 is necessary for production of 11-cis-vitamin A in the retinal visual cycle. *Nature Genetics* **20** 344-351.

Reimann, N., Bartnitzke, S., Bullerdiek, J., Schmitz, U., Rogalla, P., Nolte, I. and Ronne, M. (1996). An extended nomenclature of the canine karyotype. *Cytogenetics and Cell Genetics* **73** 140-144.

Reiner, O., Carrozzo, R., Shen, Y., Wehnert, M., Faustinella, F., Dobyns, W.B., Caskey, C.T. and Ledbetter, D.H. (1993). Isolation of a Miller-Dieker lissencephaly gene containing G protein beta-subunit-like repeats. *Nature* **364** 717-721.

Rettenberger, G., Klett, C., Zechner, U., Bruch, J., Just, W., Vogel, W. and Hameister, H. (1995a). ZOO-FISH analysis: cat and human karyotypes closely resemble the putative ancestral mammalian karyotype. *Chromosome Research* **3** 479-486.

Rettenberger, G., Klett, C., Zechner, U., Kunz, J., Vogel, W. and Hameister, H. (1995b). Visualisation of the conservation of synteny between humans and pigs by heterologous chromosomal painting. *Genomics* **26** 372-378.

Riggins, G.J., Lokey, L.K., Chastain, J.L., Leiner, H.A., Sherman, S.L., Wilkinson, K.D. and Warren, S.T. (1997). Human genes containing polymorphic trinucleotide repeats. *Nature Genetics* **2** 186-191.

Robic, A., Seroude, V., Jeon, J.T., Yerle, M., Wasungu, L., Andersson, L., Gellin, J. and Milan, D. (1999). A radiation hybrid map of the RN region in pigs demonstrates conserved gene order compared with the human and mouse genomes. *Mammalian Genome* **10** 565-568.

Rosenberg, H.F. (1993). Bacterial chromosomal contamination of a eukaryotic cDNA library. *BioTechniques* **15** 200-201.

Rouquier, S., Trask, B.J., Taviaux, S., van den Engh, G., Diriong, S., Lennon, G.G. and Giorgi, D. (1995). Direct selection of cDNAs using whole chromosomes. *Nucleic Acids Research* **23** 4415-4420.

Rowley, J.D., Diaz, M.O., Espinosa, R., Patel, Y.D., van Melle, E., Ziemin, S., Taillon-Miller, P., Lichter, P., Evans, G.A., Kersey, J.H., *et al.* (1990). Mapping chromosome band 11q23 in human acute leukemia with biotinylated probes: identification of 11q23 translocation breakpoints with a yeast artificial chromosome. *Proceedings of the National Academy of Sciences of the United States of America* **87** 9358-9362.

Ryder, E.J., Holmes, N.G., Suter, N., Sampson, J., Dolf, G., Gerlach, J.A. and Binns, M.M. (1999). Seven new linkage groups assigned to the DogMap reference families. *Animal Genetics* **30** 63-65.

Ryder, O.A., Epel, N.C. and Benirschke, K. (1978). Chromosome banding studies of the Equidae. *Cytogenetics and Cell Genetics* **20** 332-350.

Saccone, S., Federico, C., Solovei, I., Croquette, M.F., Della Valle, G. and Bernardi, G. (1999). Identification of the gene-richest bands in human prometaphase chromosomes. *Chromosome Research* **7** 379-386.

Sanz, M.M., Jenkins, E.C., Brown, W.T., Davisson, M.T., Kevin, M.J., Roderick, T.H., Silverman, W.P. and Wisniewski, H.M. (1986). Mouse chromosome fragility. *American Journal of Medical Genetics* **23** 491-509.

Sargan, D.R., Yang, F., Milne, B., O'Brien, P.C.M., Solanky, N., Squire, M., Rens, W., Ferguson-Smith, M.A. (1999) The integration of canine genetic maps with the canine karyotype using specific gene amplification of chromosome-specific DNA. *Animal Biotechnology* **10** 159-164.

Savakis, C. and Doelz, R. (1993). Contamination of cDNA sequences in databases. *Science* **259** 1677-1678.

Sawyer, J.R. (1991). Highly conserved segments in mammalian chromosomes. *Journal of Heredity* **82** 128-133.

Schatzberg, S.J., Olby, N.J., Breen, M., Anderson, L.V., Langford, C.F., Dickens, H.F., Wilton, S.D., Zeiss, C.J., Binns, M.M., Kornegay, J.N., *et al.* (1999). Molecular analysis of a spontaneous dystrophin 'knockout' dog. *Neuromuscular Disorders* **9** 289-295.

Scherthan, H., Cremer, T., Arnason, U., Weier, H.U., Lima-de-Faria, A. and Frönicke, L. (1994). Comparative chromosome painting discloses homologous segments in distantly related mammals. *Nature Genetics* **6** 342-347.

Schibler, L., Vaiman, D., Oustry, A., Giraud-Delville, C. and Cribiu, E.P. (1998). Comparative gene mapping: a fine-scale survey of chromosome rearrangements between ruminants and humans. *Genome Research* **8** 901-915.

Schläpfer, J., Switonski, M., Schelling, C., Stranzinger, G., Gaillard, C. and Dolf, G. (1998). The highly polymorphic canine microsatellite ZuBeCa2 is localized on canine chromosome 1q210-q211. *Animal Genetics* **29** 329-330.

Schrock, E., du Manoir, S., Veldman, T., Schoell, B., Weinberg, J., Ferguson-Smith, M.A., Ning, Y., Ledbetter, D.H., Bar-Am, I., Soenksen, D., *et al.* (1996). Multicolour spectral karyotyping of human chromosomes. *Science* **273** 494-497.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tomé, P., Aggarwal, A., Bajorek, E., *et al.* (1996). A gene map of the human genome. *Science* **274** 540-546.

Searle, A.G., Peters, J., Lyon, M.F., Evans, E.P., Edwards, J.H. and Buckle, V.J. (1987). Chromosome maps of man and mouse, III. *Genomics* **1** 3-18.

Sedlacek, Z., Konecki, D.S., Siebenhaar, R., Kioschis, P. and Poustka, A. (1993). Direct selection of DNA sequences conserved between species. *Nucleic Acids Research* **21** 3419-25.

Selden, J.R., Moorhead, P.S., Oehlert, M.L. and Patterson, D.F. (1975). The Giemsa banding pattern of the canine karyotype. *Cytogenetics and Cell Genetics* **15** 380-387.

Sevelius, E., Andersson, M. and Jönsson, L. (1994). Hepatic accumulation of alpha-1-antitrypsin in chronic liver disease in the dog. *Journal of Comparative Pathology* **111** 401-412.

Shibuya, H., Nonneman, D.J., Huang, T.H., Ganjam, V.K., Mann, F.A. and Johnson, G.S. (1993). Two polymorphic microsatellites in a coding segment of the canine androgen receptor gene. *Animal Genetics* **24** 345-348.

Shizuya, H., Birren, B., Kim, U.-J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. (1992). Cloning and stable maintenence of 300-kilobase-pair fragments of human DNA in *Eschericia coli* using an F-factor-based vector. *Proceedings of the National Academy of Sciences of the United States of America* **89** 8794-8797.

Short, J.M., Fernandez, J.M., Sorge, J.A. and Huse, W.D. (1988). Lambda ZAP: a bacteriophage lambda expression vector with *in vivo* excision properties. *Nucleic Acids Research* **16** 7583-7600.

Skvorak, A.B., Weng, Z., Yee, A.J., Robertson, N.G. and Morton, C.C. (1999). Human cochlear expressed sequence tags provide insight into cochlear gene expression and identify candidate genes for deafness. *Human Molecular Genetics* **8** 439-452.

Smit, A.F.A. (1996). The origin of interspersed repeats in the human genome. *Current Opinion in Genetics and Development* **6** 743-748.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L. and Efstratiadis, A. (1994). Construction and characterization of a normalized cDNA library. *Proceedings of the National Academy of Sciences of the United States of America* **91** 9228-9232.

Solinas-Toldo, S., Lengauer, C. and Fries, R. (1995). Comparative genome map of human and cattle. *Genomics* **27** 489-496.

Speicher, M.R., Gwyn Ballard, S. and Ward, D.C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics* **12** 368-375.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E. and Moyzis, R.K. (1991). Evolution and distribution of (GT)n repetitive sequences in mammalian genomes. *Genomics* **10** 807-815.

Stanyon, R., Arnold, N., Koehler, U., Bigoni, F. and Wienberg, J. (1995). Chromosomal painting shows that marked chromosomes in lesser apes and Old World monkeys are not homologous and evolved by convergence. *Cytogenetics and Cell Genetics* **68** 74-78.

Stanyon, R., Yang, F., Cavagna, P., O'Brien, P.C., Bagga, M., Ferguson-Smith, M.A. and Wienberg, J. (1999). Reciprocal chromosome painting shows that genomic rearrangement between rat and mouse proceeds ten times faster than between humans and cats. *Cytogenetics and Cell Genetics* **84** 150-155.

Stewart, E.A., McKusick, K.B., Aggarwal, A., Bajorek, E., Brady, S., Chu, A., Fang, N., Hadley, D., Harris, M., Hussain, S., *et al.* (1997). An STS-based radiation hybrid map of the human genome. *Genome Research* **7** 422-433.

Stewart, G.J., Savioz, A. and Davies, R.W. (1997). Sequence analysis of 497 mouse brain ESTs expressed in the substantia nigra. *Genomics* **39** 147-153.

Stone, D.M., Jacky, P.B. and Prieur, D.J. (1991). The Giemsa banding pattern of canine chromosomes, using a cell synchronisation technique. *Genome* **34** 407-412.

Strachan, T., Abitbol, M., Davidson, D. and Beckmann, J.S. (1997). A new dimension for the human genome project: towards comprehensive expression maps. *Nature Genetics* **16** 126-132.

Strausberg, R.L., Feingold, E.A., Klausner, R.D. and Collins, F.S. (1999). The mammalian gene collection. *Science* **286** 455-457.

Su, Y.A., Trent, J.M., Guan, X.-Y. and Meltzer, P.S. (1994). Direct isolation of genes encoded within a homogeneously staining region by chromosome microdissection. *Proceedings of the National Academy of Sciences of the United States of America* **91** 9121-9215.

Sun, H.-F.S., Ernst, C.W., Yerle, M., Pinton, P., Rothschild, M.F., Chardon, P., Rogel-Gaillard, C. and Tuggle, C.K. (1999a). Human chromosome 3 and pig chromosome 13 show complete synteny conservation but extensive gene-order differences. *Cytogenetics and Cell Genetics* **85** 273-278.

Sun, H.-F.S., Yerle, M., Pinton, P., Chardon, P., Rogel-Gaillard, C., Lyons, L.A., Laughlin, T.F. and Tuggle, C.K. (1999b). Physical assignments of human chromosome 13 genes on pig chromosome 11 demonstrate extensive synteny and gene order conservation between pig and human. *Animal Genetics* **30** 304-308.

Sutcliffe, J.G. (1988). mRNA in the mammalian central nervous system. *Annual Review of Neuroscience* **11** 157-198.

Switonski, M., Reimann, N., Bosma, A.A., Long, S., Bartnitzke, S., Pienkowska, A., Moreno-Milan, M.M. and Fischer, P. (1996). Report on the progress of standardisation of the G-banded canine (*Canis familiaris*) karyotype. *Chromosome Research* **4** 306-309.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Research* **17** 6463-6471.

Tautz, D. and Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Research* **12** 4127-4138.

Telenius, H., Carter, N.P., Bebb, C.E., Nordenskjöld, M., Ponder, B.A.J. and Tunnucliffe, A. (1992a). Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13** 718-725.

Telenius, H., Pelmear, A.H., Tunnucliffe, A., Carter, N.P., Behmel, A., Ferguson-Smith, M.A., Nordenskjöld, M., Pfragner, R. and Ponder, B.A.J. (1992b). Cytogenetic analysis by chromosome painting using DOP-PCR amplified flow-sorted chromosomes. *Genes, Chromosomes and Cancer* **4** 257-263.

Thomas, R., Breen, M., Langford, C.F. and Binns, M.M. (1999). Zoo-FISH analysis of dog chromosome 5: Identification of conserved synteny with human and cat chromosomes. *Cytogenetics and Cell Genetics* **87** 4-10.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22** 4673-4680.

Thomsen, P.D., Winterø, A.K. and Fredholm, M. (1998). Chromosomal assignment of 19 porcine cDNA sequences by FISH. *Mammalian Genome* **9** 394-396.

Toder, R., Xia, Y. and Bausch, E. (1998). Interspecies comparative genome hybridisation and interspecies representational difference analysis reveal gross DNA differences between humans and great apes. *Chromosome Research* **6** 487-494.

Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., *et al.* (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388** 539-547.

Trask, B.J. (1991). Fluorescence *in situ* hybridisation:applications in cytogenetics and gene mapping. *Trends in Genetics* **7** 149-154.

Tunnacliffe, A. and McGuire, R.S. (1990). A physical linkage group in human chromosome band 11q23 covering a region implicated in leukocyte neoplasia. *Genomics* **8** 447-453.

Tweardy, D.J., Anderson, K., Cannizzaro, L.A., Steinman, R.A., Croce, C.M. and Huebner, K. (1992). Molecular cloning of cDNAs for the human granulocyte colony-stimulating factor receptor from HL-60 and mapping of the gene to chromosome region 1p32-34. *Blood* **79** 1148-1154.

Umeno, M., McBride, O.W., Yang, C.S., Gelboin, H.V. and Gonzalez, F.J. (1988). Human ethanol-inducible P450IIE1: complete gene sequence, promoter characterization, chromosome mapping, and cDNA-directed expression. *Biochemistry* **27** 9006-9013.

van de Sluis, B.J., Breen, M., Nanji, M., van Wolferen, M., de Jong, P., Binns, M.M., Pearson, P.L., Kuipers, J., Rothuizen, J., Cox, D.W., *et al.* (1999). Genetic mapping of the copper toxicosis locus in Bedlington terriers to dog chromosome 10, in a region syntenic to human chromosome region 2p13- p16. *Human Molecular Genetics* **8** 501-507.

Van Poucke, M., Tornsten, A., Mattheeuws, M., Van Zeveren, A., Peelman, L.J. and Chowdhary, B.P. (1999). Comparative mapping between human chromosome 3 and porcine chromosome 13. *Cytogenetics and Cell Genetics* **85** 279-284.

VanDevanter, D.R., Choongkittaworn, N.M., Dyer, K.A., Aten, J., Otto, P., Vehler, C., Bryant, E.M. and Rabinovitch, P.S. (1994). Pure chromosome-specific PCR libraries from single sorted chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* **91** 5858-5862.

Veldhoen, N. and Milner, J. (1998). Isolation of canine p53 cDNA and detailed characterisation of the full length canine p53 protein. *Oncogene* **16** 1077-1084.

Venta, P.J., Brouillette, J.A., Yuzbasiyan-Gurkan, V. and Brewer, G.J. (1996). Gene-specific universal mammalian sequence-tagged sites: application to the canine genome. *Biochemical Genetics* **34** 321-341.

Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O. and Hunkapiller, M. (1998). Shotgun sequencing of the human genome. *Science* **280** 1540-1542.

Veske, A., Narfström, K., Finckh, U., Sargan, D.R., Nilsson, S.E.G. and Gal, A. (1996). Isolation of canine retinal arrestin cDNA and exclusion of three candidate genes for Swedish Briard retinal dystrophy. *Current Eye Research* **16** 270-274.

Veske, A., Nilsson, S.E., Narfström, K. and Gal, A. (1999). Retinal dystrophy of Swedish briard/briard-beagle dogs is due to a 4-bp deletion in RPE65. *Genomics* **57** 57-61.

Vezuli, A., Hauwy, M., Warter, S. and Rumpler, Y. (1997). Chromosome painting: a method for testing chromosomal changes in lemur evolution. *Cytogenetics and Cell Genetics* **78** 147-152.

Vignaux, F., Hitte, C., Priat, C., Chuat, J.C., André, C. and Galibert, F. (1999a). Construction and optimization of a dog whole-genome radiation hybrid panel. *Mammalian Genome* **10** 888-894.

Vignaux, F., Priat, C., Jouquand, S., Hitte, C., Jiang, Z., Chéron, A., Renier, C., André, C. and Galibert, F. (1999b). Toward a dog radiation hybrid map. *Journal of Heredity* **90** 62-67.

Vilà, C., Savolainen, P., Maldonado, J.E., Amorim, I.R., Rice, J.E., Honeycutt, R.L., Crandall, K.A., Lundeberg, J. and Wayne, R.K. (1997). Multiple and ancient origins of the domestic dog. *Science* **276** 1687-1689.

Walter, M.A., Spillet, D.J., Thomas, P., Weissenbach, J. and Goodfellow, P.N. (1994). A method for constructing radiation hybrid maps of whole genomes. *Nature Genetics* **7** 22-28.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280** 1077-1082.

Wang, Y., Kudoh, J., Kubota, R., Asakawa, S., Minoshima, S. and Shimizu, N. (1996). Chromosomal mapping of a family of human glutamine synthetase genes: functional gene (GLUL) on 1q25, pseudogene (GLULP) on 9p13, and three related genes (GLULL1, GLULL2, GLULL3) on 5q33, 11p15, and 11q24. *Genomics* **37** 195-199.

Waterston, R., Martin, C., Craxton, M., Huynh, C., Coulson, A., Hillier, L., Durbin, R., Green, P., Shownkeen, R., Halloran, N., *et al.* (1992). A survey of expressed genes in *Caenorhabditis elegans*. *Nature Genetics* **1** 114-123.

Wayne, R.K. (1993). Molecular evolution of the dog family. *Trends in Genetics* **9** 218-224.

Wayne, R.K., Nash, W.G. and O'Brien, S.J. (1987a). Chromosomal evolution of the *Canidae*: Divergence from the primitive carnivore karyotype. *Cytogenetics and Cell Genetics* **44** 134-141.

Wayne, R.K., Nash, W.G. and O'Brien, S.J. (1987b). Chromosomal evolution of the *Canidae*: Species with high diploid numbers. *Cytogenetics and Cell Genetics* **44** 123-133.

Wayne, R.K. and Ostrander, E.A. (1999). Origin, genetic diversity, and genome structure of the domestic dog. *Bioessays* **21** 247-257.

Weber, J.L. and May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* **44** 388-396.

Weimer, J., Kiechle, M., Senger, G., Wiedemann, U., Ovens-Raeder, A., Schuierer, S., Kautza, M., Siebert, R. and Arnold, N. (1999). An easy and reliable procedure of microdissection technique for the analysis of chromosomal breakpoints and marker chromosomes. *Chromosome Research* **7** 355-362.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. and Lathrop, M. (1992). A second-generation linkage map of the human genome. *Nature* **359** 794-801.

Wenger, R.H., Wicki, A.N., Kieffer, N., Adolph, S., Hameister, H. and Clemetson, K.J. (1989). The 5' flanking region and chromosomal localization of the gene encoding human platelet membrane glycoprotein Ib alpha. *Gene* **85** 517-524.

Werner, P., Mellersh, C.S., Raducha, M.G., DeRose, S., Acland, G.M., Prociuk, U., Wiegand, N., Aguirre, G.D., Henthorn, P.S., Patterson, D.F., *et al.* (1999). Anchoring of canine linkage groups with chromosome-specific markers. *Mammalian Genome* **10** 814-823.

Werner, P., Raducha, M.G., Prociuk, U., Henthorn, P.S. and Patterson, D.F. (1997). Physical and linkage mapping of human chromosome 17 loci to dog chromosomes 9 and 5. *Genomics* **42** 74-82.

White, O., Dunning, T., Sutton, G., Adams, M., Venter, J.C. and Fields, C. (1993). A quality control algorithm for DNA sequencing projects. *Nucleic Acids Research* **21** 3829-3838.

Wienberg, J., Jauch, A., Stanyon, R. and Cremer, T. (1990). Molecular cytotaxonomy of primates by chromosomal *in situ* suppression hybridization. *Genomics* **8** 347-350.

Wienberg, J., Stanyon, R., Nash, W.G., O'Brien, P.C., Yang, F., O'Brien, S.J. and Ferguson-Smith, M.A. (1997). Conservation of human vs. feline genome organization revealed by reciprocal chromosome painting. *Cytogenetics and Cell Genetics* **77** 211-217.

Wilcox, A.S., Khan, A.S., Hopkins, J.A. and Sikela, J.M. (1991). Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Research* **19** 1837-1843.

Williamson, A.R., Elliston, K.O. and Sturchio, J.L. (1995). The Merck Gene Index, a public resource for genomics. *Journal of NIH Research* **7** 61-63.

Winterø, A.K., Fredholm, M. and Davies, W. (1996). Evaluation and characterisation of a porcine small intestine cDNA library: analysis of 839 clones. *Mammalian Genome* **7** 509-517.

Winterø, A.K., Jorgensen, C.B., Robic, A., Yerle, M. and Fredholm, M. (1998). Improvement of the porcine transcription map: localisation of 33 genes, of which 24 are orthologous. *Mammalian Genome* **9** 366-372.

Wolfsberg, T.G. and Landsman, D. (1997). A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Research* **25** 1626-1632.

Wurster-Hill, D.H. and Centerwall, W.R. (1982). The interrelationships of chromosome banding patterns in canids, mustelids, hyena, and felids. *Cytogenetics and Cell Genetics* **34** 178-192.

Yang, F., O'Brien, P.C.M., Milne, B.S., Graphodatsky, A.S., Solanky, N., Trifonov, V., Rens, W., Sargan, D. and Ferguson-Smith, M.A. (1999). A complete comparative chromosome map for the dog, red fox, and human and its integration with canine genetic maps. *Genomics* **62** 189–202.

Yang, R.B., Fülle, H.J. and Garbers, D.L. (1996). Chromosomal localization and genomic organization of genes encoding guanylyl cyclase receptors expressed in olfactory sensory neurons and retina. *Genomics* **31** 367-372.

Yang, Y.-P., Rexroad III, C.E., Schläpfer, J. and Womack, J.E. (1998). An integrated radiation hybrid map of bovine chromosome 19 and ordered comparative mapping with human chromosome 17. *Genomics* **48** 93-99.

Yang, Y.-P. and Womack, J.E. (1995). Human chromosome 17 comparative anchor loci are conserved on bovine chromosome 19. *Genomics* **27** 293-297.

Yang, Y.-P. and Womack, J.E. (1998). Parallel radiation hybrid mapping: a powerful tool for high-resolution genomic comparison. *Genome Research* **8** 731-736.

Yang-Feng, T.L., Floyd-Smith, G., Nemer, M., Drouin, J. and Francke, U. (1985). The pronatriodilatin gene is located on the distal short arm of human chromosome 1 and on mouse chromosome 4. *American Journal of Human Genetics* **37** 1117-1128.

Yunis, J.J., Jones, C., Madden, M.T., Lu, D. and Mayer, M.G. (1989). Gene order, amplification, and rearrangement of chromosome band 11q23 in hematologic malignancies. *Genomics* **5** 84-90.

Yuzbasiyan-Gurkan, V., Blanton, S.H., Cao, Y., Ferguson, P., Li, J., Venta, P.J. and Brewer, G.J. (1997). Linkage of a microsatellite marker to the canine copper toxicosis locus in Bedlington terriers. *American Journal of Veterinary Research* **58** 23-27.

Zeiss, C.J., Trepanier, L.A., Aguirre, G.D. and Ray, K. (1998). A highly conserved microsatellite in the dystrophin gene of diverse mammalian species. *Animal Genetics* **29** 224-227.

Cytogenet Cell Genet 87:4–10 (1999)

Cytogenetics and Cell Genetics

Zoo-FISH analysis of dog chromosome 5: identification of conserved synteny with human and cat chromosomes

R. Thomas,[a,b] M. Breen,[a] C.F. Langford,[c] and M.M. Binns[a]

[a] Centre for Preventive Medicine, Animal Health Trust, Newmarket, Suffolk;
[b] Department of Biochemistry, University of Leicester, Leicester; and
[c] The Sanger Centre, Wellcome Trust Genome Campus, Cambridge (UK)

Abstract. Conserved segments of synteny between the human genome and chromosome 5 (CFA 5) of the domestic dog (*Canis familiaris*) have been identified by reciprocal chromosome painting analysis. A CFA 5 paint probe was applied to human metaphase spreads, revealing distinct hybridisation sites on human (HSA) chromosomes 1, 11, 16, and 17. Paint probes for these human chromosomes were then hybridised to dog metaphase spreads, identifying the regions of CFA 5 with which homology is shared with the corresponding human chromosome. Application of the CFA 5 paint probe to metaphase spreads of the domestic cat (*Felis catus*, FCA) demonstrated hybridisation to cat chromosomes C1, D1, E1, and E2. Dog PCR primers for type 1 markers known to lie in the corresponding regions of HSA 11, 16, and 17 were used to isolate dog BAC clones representing four genes. Fluorescence in situ hybridisation analysis confirmed their localisation to CFA 5 and suggested that two of the conserved segments lie in opposing orientations on CFA 5, compared to the human chromosome concerned. A third segment appears to lie in the same orientation on both human and dog chromosomes. No suitable gene markers were available for analysis of the fourth segment. The significance of these findings is discussed with reference to current and future dog genome mapping efforts.

Genome maps for the domestic dog (*Canis familiaris*, CFA) have recently been developed (Lingaas et al., 1997; Mellersh et al., 1998; Neff et al., 1999), based predominantly on linkage analysis of type II markers. To date, however, relatively few type I loci have been physically assigned to the dog genome (e.g., Deschenes et al., 1994; Dutra et al., 1996; Guevara-Fujita et al., 1996).

Efforts to improve this relatively gene-poor map have included the development and application of panels of PCR primers designed to permit amplification from conserved gene sequences in multiple species (Venta et al., 1996; Lyons et al., 1997; Jiang et al., 1998). These resources provide sequence tagged sites (STSs) for sets of genes that can be mapped by diverse methods in a variety of species. Their application has allowed identification of regions of the dog genome where patterns of gene localisation appear to be highly conserved compared to those of other mammalian species. This observation is a result of evolution from a common ancestral genome, where, despite the occurrence of chromosome breakages and rearrangements, it is possible to identify sets of genes that have remained linked on a single chromosome. These "syntenic groups" can be identified within a genome by a variety of molecular and/or cytogenetic techniques, allowing the location of a given gene to be predicted on the basis of knowledge of its location within the corresponding group in another species.

The current status of the dog genome map

Recently, the identification of putative regions of conserved synteny between the dog genome and those of other species has been achieved by radiation hybrid mapping of over 200 type I loci and subsequent comparison of the physical assignments of these coding sequences in the dog with those in other genomes (Priat et al., 1998). In addition, the interspecific application of whole-chromosome paints has been shown to be a rapid and

The dog BAC library was gridded by the Human Genome Mapping Resource Centre, Hinxton, Cambridge, supported by a grant from the Wellcome Trust (Project Grant No. 052908/Z/97). Dog genome analysis at the Animal Health Trust and the University of Leicester is generously supported by the Guide Dogs for the Blind Association. R.T. is the recipient of a BBSRC postgraduate studentship. C.F.L. is supported by a grant from the Wellcome Trust.

Received 29 March 1999; revision accepted 28 June 1999.

Request reprints from R. Thomas, Centre for Preventive Medicine, Animal Health Trust, Lanwades Park, Kentford, Newmarket, Suffolk, CB8 7UU (UK); telephone: +44(0)1638-750659; fax: +44(0)1638-750794; e-mail: rthomas@hgmp.mrc.ac.uk.

KARGER Fax + 41 61 306 12 34 E-mail karger@karger.ch www.karger.com

0301–0171/99/0872–0004$17.50/0

Accessible online at: www.karger.com/journals/ccg

comprehensive method for the identification of syntenic blocks within different genomes. Such studies have initially focused on comparisons between closely related primate species, such as apes and gibbons (Jauch et al., 1992), and subsequently has been extended to nonprimate species, such as rodents and ungulates (Scherthan et al., 1994), using human chromosome-specific libraries as probes. Probes constructed from flow-sorted chromosomes are now more commonly used and have been developed for a wide range of species. It is therefore possible to perform this type of fluorescence in situ hybridisation (FISH) analysis in a bidirectional, reciprocal manner to any given species for which such resources are available, allowing identification of syntenic regions at the subchromosomal level. This cross-specific reciprocal chromosome painting technique ("Zoo-FISH") has been used for comparison between a range of genomes, including cattle and human (Chowdhary et al., 1998), pig and human (Goureau et al., 1996), and cat and human (Rettenberger et al., 1995; Wienberg et al., 1997). The application of human chromosome paints specifically allows the transfer of data from the more advanced human genome mapping work to less well-studied species. However, reciprocal chromosome painting has not yet been described in the dog. This is perhaps largely due to the complexity of the dog karyotype, which comprises a large number of small, highly similar acrocentric autosomes. Consequently, high-quality resources are required for reliable cytogenetic analysis, such as chromosome-specific paint probes (Langford et al., 1996) and/or chromosome-specific unique sequence probes (Breen et al., 1999b). Here, we report on detailed reciprocal chromosome painting analysis with specific reference to dog chromosome 5 and its regions of conserved synteny with the genomes of two more widely studied species, human and the domestic cat.

Comparative mapping of dog chromosome 5

Dog chromosome 5 (CFA 5) is an ideal candidate for the development of comparative chromosome painting analysis in this species. Its combination of size and G+C content results in the production of a well-defined peak in the bivariate flow karyotype of the dog (Langford et al., 1996) and, in turn, an extremely high-quality chromosome-specific paint probe.

In an earlier study, the application of an human (HSA) chromosome 17 paint probe to dog metaphase spreads, in combination with the physical mapping of canine homologues of two human genes, has shown that the midportion of CFA 5 shares homology with HSA 17p (Werner et al., 1997). We now extend this analysis with the identification of regions of conserved synteny between CFA 5 and three additional human chromosomes, as well as with four chromosomes of the domestic cat (*Felis catus*, FCA).

Materials and methods

Metaphase chromosome preparation and FISH analysis

Metaphase chromosomes were produced by standard mitogenic stimulation of peripheral blood lymphocytes, followed by conventional harvesting and fixation treatments. Slide preparations were air dried and stored at room temperature for 4 d before use. Prior to application of the probe, slides were dehydrated through an ethanol series (70%, 90%, 100%), and the chromosomes were denatured in 70% formamide, 2 × SSC for 2 min at 65 °C, passed through the ethanol series a second time, and air dried.

In all cases, posthybridisation stringency washes were performed as described by Breen et al. (1999b). Immunological detection of biotinylated probes was performed with Texas Red Avidin DCS (Vector Laboratories), and digoxigenin-labelled probes were detected with FITC-conjugated mouse anti-digoxin (Sigma). Chromosomes were counterstained with 4′,6-diamidino-2-phenylindole (DAPI, Serva), mounted with Vectashield (Vector Laboratories), and sealed with a cover slip. Images were captured using a fluorescence microscope (Axiophot, Zeiss) equipped with an FITC/Texas Red/DAPI filter set and a cooled CCD camera (Photometrics, KAF1400), both driven by SmartCapture software (Vysis). At least 20 metaphase chromosome spreads were examined for each experiment. Dog chromosomes are described using the nomenclature proposed by Breen et al. (1999a). Unless otherwise stated, assignments of human genes are based on those reported in the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene).

Hybridisation of the CFA 5 paint probe to human chromosomes

For the identification of those human chromosomes that share homology with CFA 5, 150 ng of biotinylated CFA 5 paint probe (Langford et al., 1996) were combined with 1 μg of human Cot-1 DNA (GIBCO BRL) and 10 μg of sonicated salmon sperm DNA (GIBCO BRL) in 15 μl of a hybridisation buffer containing 40% formamide, 10% dextran sulphate, and 2 × SSC. The probe mixture was denatured at 70 °C for 10 min and allowed to preanneal for 40 min at 37 °C prior to application to denatured human metaphase spreads. Hybridisation at 37 °C for 7 d was shown to be optimal for FISH analysis of dog paint probes on human chromosomes.

Hybridisation of human chromosome paints onto dog chromosomes

Human paint probes were generated from flow-sorted chromosomes following a previously described method (Carter, 1994). For reciprocal hybridisation of human paints onto dog chromosomes, 150 ng of each biotinylated paint probe were mixed separately with 10 μg of sonicated dog genomic DNA in 15 μl of hybridisation buffer. Probe mixes were denatured and preannealed as described above and applied separately to denatured dog metaphase spreads. Hybridisation of human paint probes onto dog chromosomes was found to be optimal at 37 °C for 2 d. Analysis of hybridisation signals enabled identification of the regions of CFA 5 that are homologous to their corresponding human counterparts.

Hybridisation of CFA 5 paint to cat chromosomes

The CFA 5 paint probe was applied to chromosomes of the domestic cat (FCA). One hundred and fifty nanograms of the CFA 5 paint were denatured and preannealed as described above, in the presence of 10 μg sonicated cat genomic DNA. The probe was hybridised to cat metaphases for 2 d at 37 °C. Hybridisation signals were then compared with existing data for conserved synteny between the human and feline genomes (Wienberg et al., 1997).

Mapping of type I markers onto dog chromosomes

PCR primers for genes located in the regions of those human chromosomes that demonstrated conserved synteny with CFA 5 were selected (Venta et al., 1996; Lyons et al., 1997; Priat et al., unpublished data available at http://www-recomgen.univ-rennes1.fr./doggy.html). Conditions were optimised for PCR on dog genomic DNA. Reactions were performed in 10-μl volumes consisting of 80 pmol of each PCR primer, 50 ng dog genomic DNA, 1 U AmpliTaq Gold (Perkin Elmer), 1.3 μl of a 1.5 mM mix of each dNTP (Pharmacia), 1 μl AmpliTaq Gold 10× PCR buffer (Perkin Elmer), and 1.5 mM $MgCl_2$. PCR was performed on a PTC-225 thermal cycler (MJ Research), commencing with a 25-min denaturation step at 95 °C, followed by 30 cycles of 1 min at 94 °C, 1 min at the optimised annealing temperature, and 1 min at 72 °C. PCR products were purified using a commercial kit (Qiagen), and approximately 50 ng of each product were sequenced using dRhodamine dye terminator chemistry and analysed on an ABI 377 sequencer. The identity of each product was confirmed by comparison with existing sequence data using the BLAST search tool with default parameters (Altschul et al., 1997).

Radiolabelled gene probes were generated by the inclusion of 1.1 Mbq [α-^{32}P]dATP (specific activity, 22 Tbq/mmol, Amersham) into a PCR reaction as described previously except that all volumes were doubled to give a final reaction volume of 20 μl, and the dATP added was reduced to 25% of that used in the unlabelled reaction. The probe was purified using a G50 Sephadex column (Pharmacia) following the manufacturer's instructions and

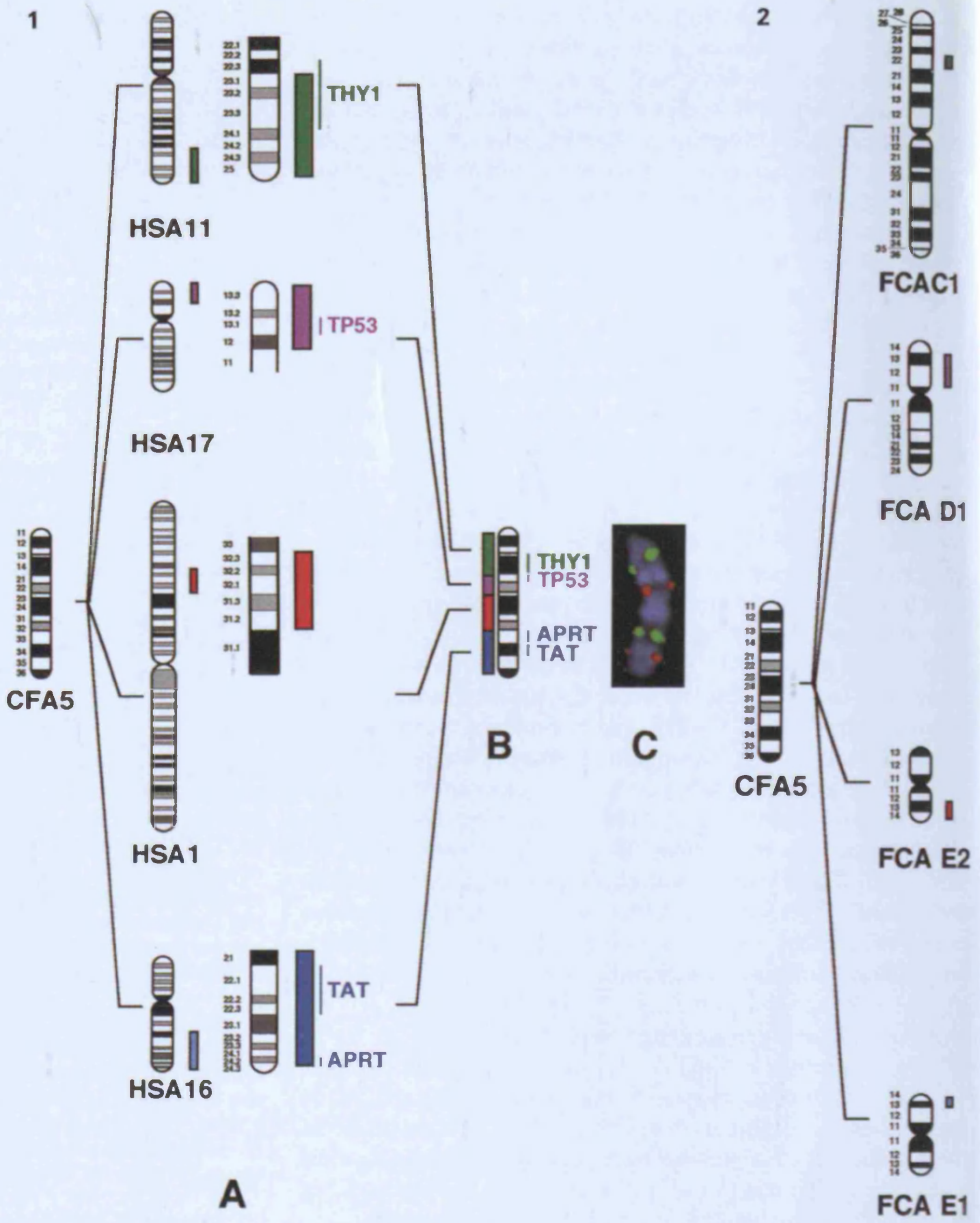


Fig. 1. Reciprocal Zoo-FISH analysis and mapping of type 1 markers to CFA 5. (**A**) Application of the CFA 5 paint probe to human chromosomes demonstrated hybridisation to HSA 1, 11, 16, and 17. The extent of the hybridisation is indicated by a vertical bar adjacent to each ideogram. This region of interest is then enlarged to the right. The location of the human gene corresponding to the STS used for orientation of conserved segments is also given. Human ideograms are adapted from Francke (1994). (**B**) Chromosome paints representing HSA 1, 11, 16, and 17 were applied separately to dog metaphase chromosomes, indicating the region of CFA 5 to which each hybridised. The physical location of the type 1 markers on CFA 5 is also given. (**C**) FISH analysis of the type 1 markers shown on an isolated CFA 5.

Fig. 2. Application of the CFA 5 paint probe to cat chromosomes. Hybridisation sites are indicated against the corresponding ideogram as before. Cat ideograms are adapted from Cho et al. (1997b).

was used to screen a dog genomic DNA BAC library with an average insert size of 155 kb (Li et al., 1999). Briefly, filters containing the gridded BAC library were prehybridised for at least 1 h at 60 °C in Church buffer (Church and Gilbert, 1984). The probe was denatured at 95 °C for 5 min, added to the prehybridisation buffer, and hybridisation was carried out at 60 °C overnight. Posthybridisation washes were performed at 60 °C for 30 min in two changes of 2 × SSC, followed by a further 30 min in two changes of 1 × SSC, 0.1 % SDS. Filters were exposed to BioMax FS film (Kodak) overnight at −70 °C with a BioMax intensifying screen. Clones generating positive hybridisation signals were verified by PCR directly from a crude bacterial suspension, using the relevant STS primers to ensure that a single product of the expected size was produced. This was repeated with BAC DNA purified from selected clones using a standard alkaline lysis method. The PCR product was sequenced as described earlier to confirm its identity with that obtained from genomic DNA.

One microgram of purified BAC DNA was labelled with biotin-16-dUTP (Boehringer) or digoxigenin-11-dUTP (Boehringer) by a modified nick translation method in which the DNAse I concentration had been optimised to generate fragments of 100–300 bp in length. Approximately 50 ng of each labelled BAC DNA were denatured and preannealed in the presence of 10 μg of sonicated dog genomic DNA, as described above. Probes were hybridised to denatured dog metaphase chromosomes overnight at 37 °C. Following immunological detection as described above, assignment of each clone was made on the basis of image-enhanced linear filtration of the DAPI banding pattern.

Results

Reciprocal chromosome painting

The application of the CFA 5 chromosome paint to human metaphase spreads identified consistent regions of hybridisation on HSA 1p32.3→p31.2, 11q23→q25, 16q21→q24, and 17p13→p12. Data from at least 20 metaphase chromosome spreads were analysed and are summarised in Fig. 1a. In addition, hybridisation signals were observed infrequently on the

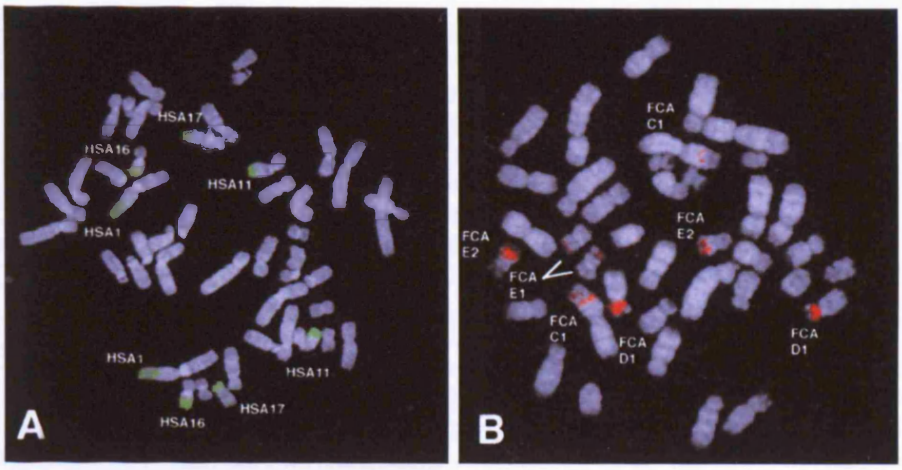


Fig. 3. Examples of hybridisation of the CFA 5 chromosome paint probe onto (**A**) human and (**B**) cat metaphase chromosomes. In each case, fluorescent signals are detected on four pairs of homologous chromosomes, as indicated in Figs. 1 and 2. Additionally, these examples demonstrate the weak hybridisation signals observed on HSA 1pter and its homologue in the cat, FCA C1pter (see text for discussion).

Table 1. Gene primers used for verification and orientation of synteny groups

Locus	Reference	Location in *Homo sapiens*[a]	Location in *Canis familiaris*	Sequence identity[b]
APRT	Venta et al. (1996)	16q24.2→qter	5q33	82/96 (85%) Human adenine phosphoribosyltransferase gene (M16446)
TAT	Priat et al. (unpublished)[c]	16q22	5q34	70/71 (98%) Dog tyrosine aminotransferase gene (L47165)
THY1	Lyons et al. (1997)	11q22.3→q23	5q13→q14	200/236 (84%) Human Thy1 glycoprotein gene (M11749)
TP53	Lyons et al. (1997)	17p13.1	5q21	334/346 (96%) Dog tumor suppressor protein (p53) gene (U51857)
NPPA	Lyons et al. (1997)	1p36	2q34→q35	294/296 (99%) Dog atrial natriuretic polypeptide (M12045)

[a] Human gene assignments are based on that given in the paper cited, with one exception: the assignment of the TAT gene was based on that given in the UniGene database (http://inhouse.ncbi.nlm.nih.gov/cgi-bin/UniGene).
[b] Number of identical bases out of total bases in closest match (using BLASTN at NCBI), followed by overall percentage match at the nucleotide level; the accession number of the highest matching GenBank entry is also given.
[c] Priat et al. (unpublished results, available at http://www-recomgen.univ-rennes1.fr./doggy.html).

most distal region of HSA 1 at 1p36, but they were not sufficiently consistent to be considered as representing a further conserved segment.

Chromosome paints representing these four human chromosomes were applied to dog metaphase chromosome spreads in separate experiments. This enabled identification of the regions of CFA 5 which correspond to the segments of the four human chromosomes with which conserved synteny had been established. Again, at least 20 metaphases were examined for each experiment, and the data are summarised in Fig. 1b.

The application of the CFA 5 paint to cat metaphase chromosome spreads revealed blocks of signal on four chromosome pairs. These regions were identified as FCA C1p22, D1p13→p12, E1p14→p13, and E2q13→q14 on the basis of their DAPI-banding patterns, as compared to the G-banded ideogram of Cho et al. (1997b). The data are presented in Fig. 2.

All reciprocal chromosome painting experiments were repeated at least once and compared with previously obtained data. No significant variations were found between duplications. Examples of the hybridisation of the CFA 5 paint probe to human and cat metaphase chromosomes are shown in Fig. 3.

Mapping of gene markers to CFA 5

Confirmation of the chromosome assignments of conserved segments was provided by the mapping to CFA 5 of gene markers located within the corresponding human synteny groups (Fig. 1c). Four such markers were selected for which PCR conditions could be optimised. Sequencing of the PCR products and comparison to existing data verified their identity as the expected gene in each case (Table 1). Additional potentially useful markers were rejected either due to difficulties in optimisation of PCR conditions or because sequencing of the resulting product indicated that its identity was not consistent with that expected.

Screening of the dog BAC library with radiolabelled PCR products corresponding to gene markers led to the identification of BAC clones representing each of the four gene probes used. Purified BAC DNA was used as a template for PCR to confirm the generation of a product of the expected size, which was sequenced. Identity with that generated from dog genomic DNA was demonstrated in each case, confirming isolation of the correct BAC clone.

Following hybridisation of labelled BAC probes to dog metaphase spreads, chromosome assignments were made on the basis of digitally enhanced DAPI-banding analysis and are summarised in Table 1. In each case, the location of the gene marker on CFA 5 confirmed the chromosome assignments of the human synteny groups identified by application of the CFA 5 paint probe, and also where the corresponding conserved segment exists on the dog chromosome.

The physical mapping of these gene markers was used to predict the relative orientation of the synteny groups in both

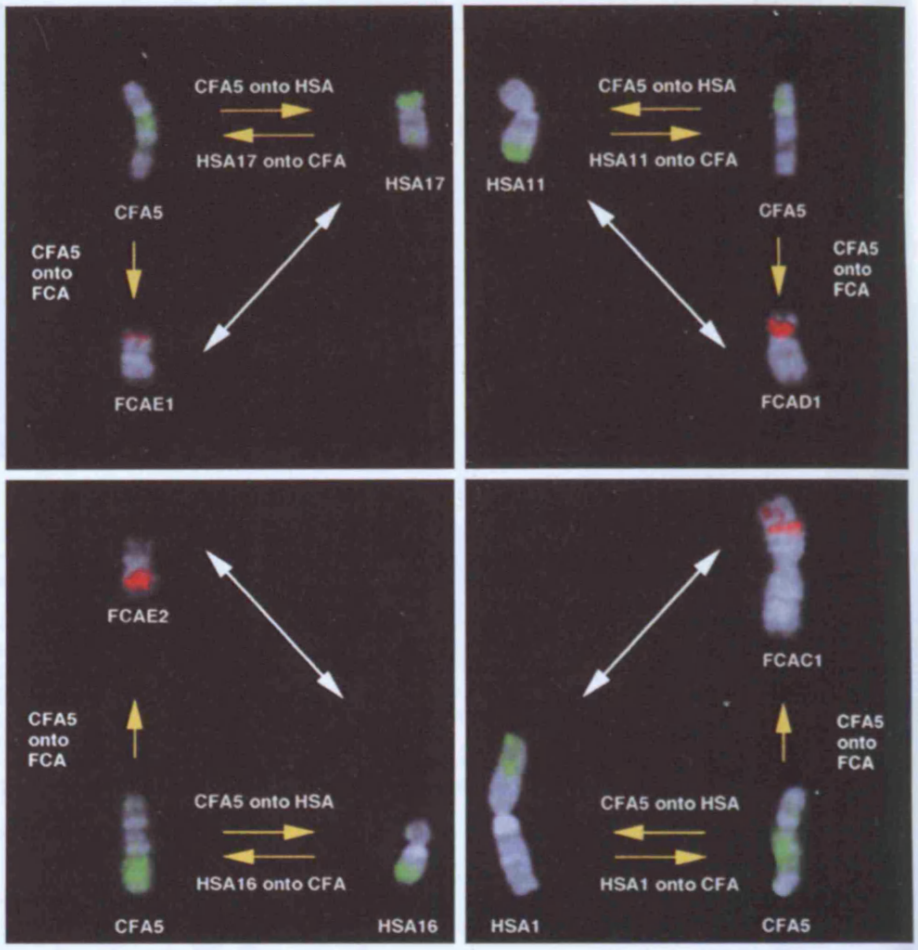


Fig. 4. Chromosome painting analysis using dog, cat, and human chromosome paint probes allows a three-way comparison between these species with respect to CFA 5. Cat/human Zoo-FISH data (indicated by white arrows) are taken from Wienberg et al. (1997).

genomes. The TP53 gene STS (Lyons et al., 1997) was used as a marker for HSA 17p13. The canine homologue was shown to map to CFA 5q21. These data, combined with existing mapping information for HSA 17p loci on CFA 5 (Werner et al., 1997), suggest that the HSA 17/CFA 5 synteny groups lie in the same orientation on both chromosomes. The THY1 gene STS (Lyons et al., 1997) was assigned to CFA 5q13→q14, at the distal region of the conserved segment. The THY1 gene has been mapped to HSA 11q22.3→q23 and represents a proximal marker for this synteny region. As such, this suggests that the HSA 11/CFA 5 synteny groups lie in opposing orientations.

Two suitable gene STSs were available for analysis of the HSA 16/CFA 5 conserved segment, namely APRT (Venta et al., 1996) and TAT (Priat et al., unpublished data available at http://www-recomgen.univ-rennes1.fr./doggy.html). The assignment of the former gene to CFA 5q33 and of the latter to CFA 5q34 suggests a reversal of orientation with respect to the centromere as compared to the reported locations of these genes on HSA 16q24→qter and 16q22, respectively.

No suitable gene STSs were available for analysis of the HSA 1/CFA 5 conserved segment. However, a marker was available which was used to further investigate the infrequent hybridisation signals observed on the distal end of HSA 1p. The NPPA gene has been assigned to HSA 1p36, and the corresponding STS (Lyons et al., 1997) was used in a similar fashion to that described above in order to isolate a BAC clone for this gene. The dog homologue was thus assigned to CFA2 (data not shown). This provides further evidence to suggest that HSA 1p36 and CFA 5 do not represent a further conserved segment, and that the hybridisation signals observed at this location on HSA 1p are indeed artefactual.

Discussion

Reciprocal chromosome painting analysis has been used to demonstrate that dog chromosome 5 is composed of conserved segments corresponding to regions of human chromosomes 1p, 11q, 16q, and 17p. Analysis of hybridisation signals on DAPI-banded dog metaphases has allowed refinement of these regions to the single-band level. The subsequent hybridisation of paint probes for these human chromosomes revealed the regions of CFA 5 to which each of the human conserved segments correspond, permitting a direct two-way comparison of two genomes with respect to CFA 5. Patterns of hybridisation were highly consistent between replications of these experiments, reflecting the high quality of the flow-sorted material used.

Due to the close proximity of flow peaks representing HSA 9, 10, 11, and 12 in the human bivariate flow karyotype, it was necessary to use a paint probe that contained material from all

four of these chromosomes to obtain data for the hybridisation of HSA 11 to dog metaphase chromosome spreads. The concentration of this probe was therefore increased four fold to 600 ng. As expected, the resulting hybridisation signals were more complex due to the lower signal-to-noise ratio produced by the increased probe complexity. However, the CFA 5 paint showed no detectable hybridisation to HSA 9, 10, and 12, and so it can be concluded with confidence that CFA 5 does not contain any major homology to the additional three human chromosomes in this pool. In addition, further studies have clearly demonstrated that CFA 5 shares no detectable synteny with any of the other 18 human autosomes (Breen et al., 1999c).

Further verification of these data has been provided by the application of the CFA 5 paint probe to cat chromosomes. The specificity of the hybridisation signals generated was again high and is likely to be a reflection of the close evolutionary relationship between the cat and the dog. FISH analysis revealed distinct blocks of hybridisation signal on cat chromosomes C1, D1, E1, and E2. Previously, reciprocal cat/human painting analysis has demonstrated regions of synteny between these four cat chromosomes and HSA 1, 11, 17, and 16, respectively (Wienberg et al., 1997). It can be seen that these data are consistent with the hybridisation patterns observed from the application of the CFA 5 paint probe to cat chromosomes. It is therefore possible to draw a three-way comparison between four synteny groups in these three species, as indicated in Fig. 4.

In this study, four genes have been mapped in the dog whose human orthologue lies in an evolutionarily conserved segment corresponding to CFA 5. The assignment of these genes to CFA 5 has confirmed the HSA/CFA conserved segments identified by the reciprocal Zoo-FISH analysis. In addition, it has permitted a prediction of whether the orientation of the conserved segments is the same or inverted between these species with respect to the centromeres of the chromosomes concerned. Mapping data suggest that both the HSA 11/CFA 5 and HSA 16/CFA 5 syntenic groups lie in opposing orientations, whilst the HSA 17/CFA 5 segments appear to lie in the same orientation. No suitable gene markers were available for the HSA 1/CFA 5 segment for which a reliable PCR assay was possible. This is, in part, a reflection of the limited size of this region.

High-resolution comparative mapping analysis is clearly reliant on comprehensive gene mapping data for the species involved, exemplified by the data available for well-studied systems such as the human genome. The cat genome has been more comprehensively mapped than that of the dog with respect to type I loci (O'Brien and Nash, 1982; O'Brien et al., 1993, 1997), many of which have also been physically assigned to human chromosomes. Identification of conserved regions between these genomes (Wienberg et al., 1997) is at a more advanced stage than it is in the dog. As such, the inclusion of the cat in this study has been a useful intermediary for comparison of the data obtained for human and dog Zoo-FISH analysis.

The mapping of type 1 markers in the dog is itself at a relatively early stage. Guevara-Fujita et al. (1996) mapped the first six autosomal genes in the dog by FISH analysis. This included the assignment of the APRT gene to the dog chromosome 3 defined by Stone et al. (1990). This corresponds to CFA 5 of the karyotype proposed by Breen et al. (1999a) used here. In the same study, Guevara-Fujita and co-workers (1996) also assigned the tumour suppressor gene TP53 (located on HSA 17p13.1) to dog chromosome 5 (Stone et al., 1990), which corresponds to CFA 7 (Breen et al., 1999a). However, we believe that TP53 also lies on CFA 5 (Breen et al., 1999a), in band 5q21. This has been demonstrated in this study by the physical mapping to CFA 5 of a dog BAC clone containing the TP53 gene and is additionally supported by the reciprocal chromosome painting analysis of CFA 5 and HSA 17. The TP53 gene has previously been mapped to cat chromosome E1 (Cho et al., 1997a), which is also in accordance with data obtained in the present study. It is therefore possible to draw direct comparisons between FCA E1p14→p13, HSA 17p13.1, and CFA 5q21.

Prior to this study, two further genes have been reported to map to CFA 5. The application of an HSA 17 chromosome paint to dog metaphase chromosome spreads has previously demonstrated that a region of conserved synteny exists between CFA 5 and the small arm of HSA 17 (Werner et al., 1997). This was confirmed by the mapping of two gene markers known to lie on HSA 17p. The SLC2A4 (formerly GLUT4) gene, located on HSA 17p13, was mapped to CFA 5 by FISH analysis, which is clearly supported by the present study. The PMP22 locus, located at HSA 17p12→p11.2, was similarly mapped by Werner et al. (1997) to the midportion of CFA 5q. The present study suggests that the conserved region that HSA 17 shares with CFA 5 is restricted to HSA 17p13→p12, as the CFA 5 paint probe showed no consistent hybridisation to HSA 17p11. The region involved is small, and at this resolution, it is not unexpected for some variability to exist between studies in the sub-band boundaries of hybridisation signal blocks. Alternatively, this may indicate that the human PMP22 locus is located specifically in HSA 17p12.

Anomalies between the hybridisation signals in both directions of Zoo-FISH analysis have been described previously in other species, e.g., cat and human (Wienberg et al., 1997). Here, it was proposed that incidences of chromosome-painting signals being observed in only one direction of the reciprocal process may relate to the hybridisation conditions used. This may explain the above observations, particularly since the HSA 17 paint probes used in the two studies are of different origin. It is also a potential explanation for the infrequent observation of hybridisation signals on HSA 1p36 following the application of the CFA 5 paint to human metaphase spreads. This region of HSA 1 is also the site of weak hybridisation signals of a number of other dog chromosome paints (data not shown). Anomalies such as these reinforce the value of the reciprocal nature of such studies, which provides an opportunity for verification of data obtained from the two components of the procedure. The resolution limit of Zoo-FISH has previously been estimated as 7 Mb (Scherthan et al., 1994). Conserved segments that are smaller than 7 Mb may therefore only be revealed by more comprehensive gene mapping in the species of interest, which would also allow further confirmation of the boundaries and relative orientations of the conserved segments in different genomes. In addition, higher resolution gene mapping is necessary to establish the degree to which gene order is conserved when such segments are compared between species. It is possible that internal rearrangements may have occurred within segments during evolution

from the ancestral karyotype. As such, orientation on the basis of mapping of small numbers of type I markers must be treated as tentative in the absence of more detailed mapping data.

This study has therefore led to the identification of three conserved segments between the human genome and CFA 5, in addition to that previously identified by Werner et al. (1997). These data have been shown to be concordant with existing data from the cat genome. Two genes (THY1 and TAT) have been newly assigned to CFA 5, and the mapping of a further two genes (APRT and TP53) has been refined. The orientation of three conserved segments in the dog has been proposed compared to that on the corresponding human chromosome. Such an approach should be valuable in the search for genes of interest in the dog where the location of the human homologue is known. This will require the complete identification of all conserved segments throughout the dog genome. The fact that many genes exist as members of larger gene families, complicating both library screening and FISH analysis, reduces the scope for selection of suitable genes for comparative mapping. New resources, such as the canine whole-genome radiation hybrid panel developed by Priat et al. (1998), are likely to be vital for circumventing some of these complications. The continuing efforts to place type I markers on the dog genome map will therefore be critical in the production of high-resolution comparative mapping data for the complete canine genome.

Acknowledgements

We thank Dr. N. Holmes for helpful discussions during the preparation of the manuscript.

Note added in proof

A marker for the DIO1 gene (Priat et al. 1999) has now been used to isolate a BAC clone that has been mapped to CFA5q31. Since DIO1 has previously been assigned to HSA1p33 → p32, this would suggest that the corresponding regions of conserved synteny between HSA1 and CFA5 lie in the same orientation with respect to the centromere.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389–3402 (1997).

Breen M, Bullerdiek J, Langford CF: A DAPI-banded karyotype of the domestic dog *(Canis familiaris)* using chromosome-specific paint probes. Chrom Res 7:401–406 (1999a).

Breen M, Langford CF, Carter NP, Holmes NG, Dickens HF, Thomas R, Suter N, Ryder EJ, Pope M, Binns MM: FISH mapping and identification of canine chromosomes. J Hered 90:27–30 (1999b).

Breen M, Thomas R, Binns MM, Carter NP, Langford CF: Reciprocal chromosome painting reveals detailed regions of synteny between the karyotypes of the domestic dog *(Canis familiaris)* and man. Genomics (1999c, in press).

Carter NP: Cytogenetic analysis by chromosome painting. Cytometry 18:2–10 (1994).

Cho K-W, Okuda M, Endo Y, Satoh H, Kang C-B, Watari T, Tsujimoto H, Hasegawa A: Assignment of the cat p53 tumour suppressor gene (TP53) to cat chromosome E1p14-p13 by fluorescence in-situ hybridisation. Cytogenet Cell Genet 79:145–146 (1997a).

Cho K-W, Youn H-Y, Watari T, Tsujimoto H, Hasegawa A, Satoh H: A proposed nomenclature of the domestic cat karyotype. Cytogenet Cell Genet 79:71–78 (1997b).

Chowdhary BP, Raudsepp T, Fronicke L, Scherthan H: Emerging patterns of comparative genome organisation in some mammalian species as revealed by Zoo-FISH. Genome Res 8:577–589 (1998).

Church GM, Gilbert W: Genomic sequencing. Proc natl Acad Sci, USA 81:1991–1995 (1984).

Deschenes SM, Puck JM, Dutra AM, Somberg RL, Felsburg PJ, Henthorn PS: Comparative mapping of canine and human proximal Xq and genetic analysis of canine X-linked severe combined immunodeficiency. Genomics 23:62–68 (1994).

Dutra AS, Mignot E, Puck JM: Gene localisation and syntenic mapping by FISH in the dog. Cytogenet Cell Genet 74:113–117 (1996).

Francke U: Digitised and differentially shaded human chromosome ideograms for genomic applications. Cytogenet Cell Genet 65:206–219 (1994).

Goureau A, Yerle M, Schmitz A, Riquet J, Milan D, Pinton P, Frelat G, Gellin J: Human and porcine correspondence of chromosome segments using bidirectional chromosome painting. Genomics 36:252–262 (1996).

Guevara-Fujita ML, Loechel R, Venta PJ, Yuzbasiyan-Gurkan V, Brewer J: Chromosomal assignment of seven genes on canine chromosomes by fluorescence in-situ hybridisation. Mammal Genome 7:268–270 (1996).

Jauch A, Wienberg J, Stanyon R, Arnold N, Tofanelli S, Ishida T, Cremer T: Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. Proc natl Acad Sci, USA 89:8611–8615 (1992).

Jiang Z, Priat C, Galibert F: Traced orthologous amplified sequence tags (TOASTs) and mammalian comparative maps. Mammal Genome 9:577–587 (1998).

Langford CF, Fischer PE, Binns MM, Holmes NG, Carter NP: Chromosome-specific paints from a high-resolution flow karyotype of the dog. Chrom Res 4:115–123 (1996).

Li R, Mignot E, Faraco J, Kadotani H, Cantanese JJ, Zhao BH, Lin X, Hinton L, Ostrander EA, Patterson DF, de Jong PJ: Construction and characterisation of an eight fold redundant dog genomic bacterial artificial chromosome library. Genomics 58:9–17 (1999).

Lingaas F, Sorensen A, Juneja RK, Johansson S, Fredholm M, Wintero AK, Sampson J, Mellersh CS, Curzon A, Holmes NG, Binns MM, Dickens HF, Ryder EJ, Gerlach J, Baumle E, Dolf G: Towards construction of a canine linkage map: establishment of 16 linkage groups. Mammal Genome 8:218–221 (1997).

Lyons LA, Laughlin TF, Copeland NG, Jenkins NA, Womack JE, O'Brien SJ: Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nature Genet 15:47–56 (1997).

Mellersh CS, Langston AA, Acland GM, Fleming MA, Ray K, Wiegand NA, Francisco LV, Gibbs M, Aguirre GD, Ostrander E: A linkage map of the canine genome. Genomics 46:326–336 (1998).

Neff MW, Bromann KW, Mellersh CS, Ray K, Acland GM, Aguirre GD, Ziegle JS, Ostrander EA, Rine J: A second-generation genetic linkage map of the domestic dog, *Canis familiaris*. Genetics 151:803–820 (1999).

O'Brien SJ, Cevario SJ, Martenson JS, Thompson MA, Nash WG, Chang E, Graves JAM, Spencer JA, Cho K-W, Tsujimoto H, Lyons LA: Comparative gene mapping in the domestic cat *(Felis catus)*. J Hered 88:408–414 (1997).

O'Brien SJ, Nash WG: Genetic mapping in mammals: chromosome map of domestic cat. Science 216:257–265 (1982).

O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA, Copeland NG: Anchored reference loci for comparative mapping in mammals. Nature Genet 3:103–112 (1993).

Priat C, Hitte C, Vignaux F, Renier C, Jiang Z, Jouquand S, Cero A, Andre C, Galibert F: A whole-genome radiation hybrid map of the dog genome. Genomics 54:361–378 (1998).

Priat C, Jiang ZH, Renier C, Andre C, Galibert F: Characterization of 463 type I markers suitable for dog genome mapping. Mammal Genome 10:803–813 (1999).

Rettenberger G, Klett C, Zechner U: Zoo-FISH analysis: cat and human karyotypes closely resemble the putative ancestral mammalian karyotype. Chromosome Res 3:479–486 (1995).

Scherthan H, Cremer T, Arnason U, Weier H-U, Lima-de-Faria A, Fronicke L: Comparative chromosome painting discloses homologous segments in distantly related mammals. Nature Genet 6:342–347 (1994).

Stone DM, Jacky PB, Prieur DJ: The Giemsa banding pattern of canine chromosomes, using a cell synchronisation technique. Genome 34:407–412 (1990).

Venta PJ, Brouillette JA, Yuzbasiyan-Gurkan V, Brewer GJ: Gene-specific universal mammalian sequence-tagged sites: application to the canine genome. Biochem Genet 34:321–341 (1996).

Werner P, Raducha MG, Prociuk U, Henthorn PS, Patterson DF: Physical and linkage mapping of human chromosome 17 loci to dog chromosomes 9 and 5. Genomics 42:74–82 (1997).

Wienberg J, Stanyon R, Nash WG, O'Brien PCM, Yang F, O'Brien SJ, Ferguson-Smith MA: Conservation of human vs. feline genome organisation revealed by reciprocal chromosome painting. Cytogenet Cell Genet 77:211–217 (1997).