

# **Analysis of minisatellites in humans and mice**

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

by  
John David Hadley Stead  
University of Leicester

January 2000



**University of  
Leicester**

UMI Number: U124006

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U124006

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.  
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against  
unauthorized copying under Title 17, United States Code.



ProQuest LLC  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106-1346

This thesis is dedicated to my grandfather, Pags,  
whose infectious enthusiasm for every aspect of the natural world  
continues to be such a major influence.

# Contents

<b>Abstract</b>	viii
<b>Acknowledgements</b>	ix
<b>Common abbreviations</b>	x
<b>Chapter 1: Introduction</b>	1
<b>Mutation and evolution</b>	1
<b>Population genetics of repetitive DNA</b>	2
<b>Classes of repetitive DNA</b>	3
Dispersed repeats	3
Short interspersed elements	3
Long interspersed elements	5
Multi-gene families and pseudogenes	6
Tandemly repetitive DNA	7
Satellite DNA	7
Minisatellites	9
Telomeres	10
Expanded simple tandem repeats	11
Simple tandem repeats	12
Cryptic tandem repeats	14
<b>Dynamic tandem repeat expansions and human disease</b>	15
<b>Phenotypic associations of minisatellites</b>	18
Coding minisatellites	18
Regulatory minisatellites	19
<b>Minisatellite mutation</b>	21
Mutation dynamics of AT-rich human minisatellites	21
Mutation dynamics of GC-rich human minisatellites	21
Patterns of population allele variation	21
Detection of <i>de novo</i> mutation events	22
Characteristics of minisatellite mutation in humans	23
Evidence for a recombination hotspot associated with MS32	26
<b>Mutation induction</b>	29
<b>Overview of this thesis</b>	32



<b>Chapter 2: Materials and methods</b>	35
<b>Materials</b>	35
<b>Methods</b>	39
1: Selection and growth of bacterial cultures	39
2: Agarose gel electrophoresis	40
3: DNA transfer to membrane	40
4: Hybridisation	41
5: Methods of DNA extraction and purification	42
6: Enzymatic manipulation of DNA	45
7: DNA amplification	46
8: Size-enrichment small pool PCR at the insulin minisatellite	47
9: Automated DNA sequence analysis	47
10: Subcloning techniques	48
<b>Chapter 3: Isolation and characterisation of mouse minisatellites</b>	51
<b>Summary</b>	51
<b>Introduction</b>	51
Potential applications of a mouse model of minisatellite instability	52
Effects of genomic context on instability	52
Determination of mutation rate variation	52
Characterising mutation processes from a range of tissues	53
Mechanisms of mutation induction	53
Characteristics of the ideal model locus	54
Transgenic mouse models of minisatellite instability	54
<b>Methods</b>	56
Strategies for the isolation of minisatellites	56
1: Synthetic tandem repeat probe cross-hybridisation to a cosmid library	56
2: Cross-hybridisation to size-selected charomid libraries	57
3: Cross-hybridisation to size-selected $\lambda$ libraries	57
4: Isolation of restriction endonuclease-resistant fragments	57
<b>Results</b>	59
Isolation of restriction endonuclease-resistant fragments from a cosmid library	59
Combined results of various approaches to identify mouse minisatellite	60
Analysis of VNTR length polymorphism	61
Chromosomal localisation of VNTR loci	62
Restriction mapping of mouse VNTR loci	63
Correction of allele sizes using restriction map data	64
Subcloning tandem repeat arrays for sequence analysis	65
Sequence analysis of VNTR loci	66
<b>Discussion</b>	67
Screening for VNTR loci	67
VNTR variability profiles	68
VNTR sequence analysis	69
Variant repeat distribution in three mouse minisatellites	70
Chromosomal localisation of mouse VNTRs	71

<b>Chapter 4: Variant repeat distribution in two mouse minisatellites</b>	73
<b>Summary</b>	73
<b>Introduction</b>	73
Computer analysis of putative DNA conformation	75
Analysis of variant repeat distribution	76
<b>Results</b>	77
PCR amplification of MMS57 and MMS58	77
Sequence analysis of minisatellite alleles	78
Computer analysis of DNA conformation	79
MMS57	79
MMS58	80
Variant repeat analysis by MVR-PCR	80
Variant repeat distribution at MMS57	80
Variant repeat distribution at MMS58	81
Problems with allele alignment and subdivision	81
<b>Discussion</b>	82
MMS57 may undergo dynamic expansions	82
MMS58 may be a recently expanded minisatellite	85
 <b>Chapter 5: The application of MVR-PCR to phylogenetic analysis</b>	87
<b>Summary</b>	87
<b>Introduction</b>	87
<b>Results and Discussion</b>	90
Subdivision of alleles using MultAlin software	90
Multidimensional scaling analysis using five mouse minisatellites	91
<b>Conclusion</b>	93
 <b>Chapter 6: Analysis of the <i>Hm-1</i> expanded simple tandem repeat</b>	94
<b>Summary</b>	94
<b>Introduction</b>	94
<b>Results and Discussion</b>	99
1: PCR amplification at <i>Hm-1</i>	99
2: Screens for variant repeats at <i>Hm-1</i>	100
3: Sequence analysis of the <i>Hm-1</i> flanking region	102
<b>Conclusion</b>	103
 <b>Chapter 7: Characterisation of human minisatellites with discontinuous allele length distributions</b>	104
<b>Summary</b>	104
<b>Introduction</b>	104
<b>Results</b>	107
Sequence analysis of candidate loci	107
PCR amplification of three minisatellites	108
MVR-PCR analysis of <i>D19S20</i>	109
MVR-PCR analysis of MS51	110
Relationship between allele lineage and size	112
<b>Discussion</b>	112

<b>Chapter 8: Allele diversity and germline mutation</b>	114
<b>at the insulin minisatellite</b>	
<b>Summary</b>	114
<b>Introduction</b>	115
<b>Results</b>	116
PCR amplification of the insulin minisatellite	116
Establishment of MVR-PCR at the insulin minisatellite	117
Allele diversity at the insulin minisatellite	119
Patterns of variation in the insulin minisatellite	120
Detection of <i>de novo</i> mutants	122
Mutation processes in the soma	123
Deletion processes in the germline	124
Expansion processes in the germline	125
PCR artefacts have simple structures	126
<b>Discussion</b>	127
 <b>Chapter 9: The insulin minisatellite and type 1 diabetes:</b>	131
<b>a literature review</b>	
<b>Introduction to diabetes</b>	131
<b>Type 1 diabetes is caused by genetic and environmental factors</b>	132
Familial clustering	132
Variation in the incidence of type 1 diabetes between populations	133
Animal experimentation	134
<b>i) The identification of type 1 diabetes susceptibility loci</b>	134
Principles for genetic mapping of type 1 diabetes susceptibility loci	134
Linkage and association analysis	135
Definition of the aetiological variant	136
Examples of the genetic mapping of type 1 diabetes susceptibility loci	137
Genome-wide scans for type 1 diabetes susceptibility loci	137
The functional candidate gene approach to the identification of	138
type 1 diabetes susceptibility loci	
<i>IDDM1</i> maps to the HLA complex	138
<i>IDDM2</i> maps to the insulin minisatellite	140
Defining the boundaries of <i>IDDM2</i>	141
Cross-match haplotype analysis of	142
the 4.1 kb region of susceptibility.	
<i>IDDM2</i> in different populations	144
Doubts over the identity of <i>IDDM2</i>	146
Redefining the boundaries of <i>IDDM2</i>	146
<i>IDDM2</i> may have a multi-locus aetiological basis	148
<b>ii) The development of type 1 diabetes</b>	149
The cell-mediated immune response	149
Antigen recognition	149
T cell selection in the thymus	151
Induction of T cell anergy	152
T cell activation	152
T cell response	153
The Pathology of environmental determinants of type 1 diabetes susceptibility	155

β-cell damage.....	155
T cell activation.....	155
The pathology of genetic determinants of type 1 diabetes susceptibility.....	156
<i>IDDM1</i> .....	156
<i>IDDM2</i> .....	158
Levels of gene expression correlate with <i>INS</i> VNTR identity.....	158
Putative mechanisms for <i>INS</i> VNTR enhancer activity.....	160
The effects of other type 1 diabetes susceptibility loci.....	163
Interaction between <i>IDDM1</i> and <i>IDDM2</i> .....	163
Type 1 diabetes and other autoimmune diseases.....	164
iii) Further investigations of <i>IDDM2</i> .....	164
<b>Chapter 10: Evidence that <i>IDDM2</i> has a multi-locus aetiology</b> .....	167
<b>Summary</b> .....	167
<b>Introduction</b> .....	168
<b>Results</b> .....	169
Analysis of flanking haplotypes.....	169
Class III allele haplotypes.....	169
Class I allele haplotypes.....	170
Allele distribution at <i>HUMTH01</i> .....	170
The insulin minisatellite and susceptibility to type 1 diabetes.....	171
Transmissions from class I/III heterozygous parents.....	171
Subdivision of class III alleles.....	172
Subdivision of class I alleles.....	173
Further analysis of the ID-/III paternal genotype.....	174
Transmissions from class I/I heterozygous parents.....	175
<b>Discussion</b> .....	175
A model for the aetiology of <i>IDDM2</i> .....	178
Future Directions.....	184
<b>Appendices</b>	
Appendix 1: Proximal restriction maps of mouse minisatellites.....	189
Appendix 2: Repeat unit sequence data from mouse VNTRs.....	196
Appendix 3: Sequences flanking two mouse minisatellites.....	212
Appendix 4: Sequences flanking the <i>Hm-1</i> repeat array.....	217
Appendix 5: Sequences flanking three human minisatellites.....	224
Appendix 6: Sequences flanking the insulin minisatellite.....	230
Appendix 7: Allele diversity at the insulin minisatellite.....	235
Appendix 8: <i>De novo</i> mutation at the insulin minisatellite.....	240
Appendix 9: Genotypes within each type 1 diabetes affected sib pair family.....	252
<b>References</b> .....	261

# Analysis of minisatellites in humans and mice

John David Hadley Stead

## Abstract

Mutation processes have previously been studied in human minisatellites with mutation rates of 1-15% per sperm. In this thesis, this work was expanded in two directions. The first was to characterise mouse minisatellites to generate a mouse model for the detailed analysis of minisatellite instability. The second was to analyse mutation in human minisatellites with lower mutation rates, to determine whether common mutation mechanisms operate at all human minisatellites.

A screen for endogenous mouse minisatellites identified 77 loci, the majority of which were structurally similar to human minisatellites. There was no evidence for the existence of mouse minisatellites with mutation rates above  $10^{-4}$  per gamete, suggesting a fundamental difference between germline tandem repeat instability in humans and mice.

To extend mutation analysis in humans to minisatellites with lower mutation rates, the insulin minisatellite was investigated. Allele diversity was characterised by mapping variant repeat distribution in 876 alleles, and *de novo* mutations were isolated from both germline and soma. Two forms of mutation were identified, the major resulting in simple deletions and duplications which were at least partially of premeiotic origin, and a minor form involving complex intra- and inter-allelic rearrangements of repeats, similar to mutation processes observed at hypermutable minisatellites, and almost certainly of meiotic origin. Homozygosity had no effect on either the rate or complexity of germline mutation.

The insulin minisatellite is the best known candidate for the type 1 diabetes susceptibility locus *IDDM2*. Variant repeat distribution and flanking haplotype were combined to identify five newly defined ancestral lineages which were used to further characterise association of the insulin-linked region with disease. This study found that type 1 diabetes risk was influenced by extended haplotype, raising doubts over the role of the minisatellite in disease susceptibility. A model to account for *IDDM2*-associated pathology is presented.

## Acknowledgements

First of all, I would like to thank my whole family for their love, support, pizza (Mum), curry (Dad), beer (both), and for always being there when I needed them. Thanks to Granny for her continued attempts to fatten me up and to Pags who, like my parents, continues to show so much interest in my work. Thanks also to my sister Rebecca and to Tim (cracking stag night - I think) for the London bolt-hole, and to my wonderful cousins Jane, Jack, and Lizzie, none of whom I see as much as I would like. A very big hug also to my friends Becky, Gaynor, Marika, Tanya, Laura, Sanne, Sarah, Jayne, and of course the gorgeous Louise (why are you still living in Dubai?). I am lucky to know any of you, let alone all of you. Here in Leicester, my thanks first and foremost to Matt for his friendship, for putting up with me for two whole years for which he deserves a medal, and for keeping me busy with constant washing up throughout the entire time. Huge thanks to every one in the lab; to Rita (how many hybridisation?), Ruth (for teaching me how to do thousands of PCR in a single day), Caroline (even at the *very* rare 6 a.m. I knew I would have company in the lab!), Angie and Julia (for constantly taking the mickey), Keiji (nobody can imitate computer noises like him, or would perhaps wish to), Helen (for ensuring I survived Denver), Ila (just how much did I pester you?), Gemma (for all the lifts to the climbing wall), Philippe (for his help when I was starting out), John (at least the lab was never quiet), Mark (for so much help), Nicola (for not objecting to all my questions), Jérôme (he actually smiled once! Really! I saw it happen!), and to everyone else who have made the lab such a good place to be. Thanks also to Esther and to Celia for putting up with my constant stream of questions while writing up. And of course, a drunken salute to the clan of the New Road Inn. To Carole for being a great friend, for the memory loss, and for the liver damage, to Yuri for friendship, four letter words, stats, cigarettes and vodka, and to Maria who is *still* desperately trying to get me to go clubbing. On the science front, my thanks to Prof. John Todd, Dr. Mark McCarthy and Dr. Michael Festing for their helpful discussions and to Dr. Raymond Dalgleish for web site advice. Finally, I would like to thank Alec with whom it has been a privilege to study, for his excellent supervision, and for ensuring that science can be so much fun.

## Common abbreviations

ASP	affected sib pair
BXD RI	C57BL/6JxDBA recombinant inbred mouse strain
CEPH	Centre d'Etude du Polymorphisme Humain
df	degrees of freedom
ESTR	expanded simple tandem repeat
HLA	human leukocyte antigen
<i>Hm-1</i>	hypervariable minisatellite-1
IBD	identical by descent
<i>IDDM2</i>	insulin-dependent diabetes mellitus susceptibility locus 2
<i>IGF2</i>	insulin-like growth factor 2
<i>INS</i>	insulin
LD	linkage disequilibrium
MDS	multi-dimensional scaling
MMS	mouse minisatellite
MS	minisatellite (human)
MVR	minisatellite variant repeat
PCOS	polycystic ovary syndrome
PCR	polymerase chain reaction
PH	protective haplotype
RR	relative risk
SDSA	synthesis-dependent strand annealing
SESP-PCR	size enrichment small pool PCR
SNP	single nucleotide polymorphism
SP-PCR	small pool PCR
STR	simple tandem repeat
TDT	transmission disequilibrium test
<i>TH</i>	tyrosine hydroxylase
UTR	untranslated region
VNTR	variable number of tandem repeats
VPH	very protective haplotype

# Chapter 1

## Introduction

### Mutation and evolution

"... I think it would be a most extraordinary fact if no variation ever had occurred useful to each being's own welfare, in the same way as so many variations have occurred useful to man. But if variations useful to any organic being do occur, assuredly individuals thus characterised will have the best chance of being preserved in the struggle for life; and from the strong principle of inheritance they will tend to produce offspring similarly characterised. This principle of preservation, I have called, for the sake of brevity, Natural Selection."

*From The Origin of Species.*

*Darwin, 1859.*

Without genetic variation, there could be no heritable variation and so no evolution. It is therefore difficult to overstate the fundamental importance of processes which generate variation. Variation in the mammalian genome is the result of two fundamental forces; mutation and recombination. Mutation generates genetic changes which are shuffled by recombination to produce novel combinations of variants. It will be seen that in some cases, mutation and recombination are intimately linked. Mutations can occur at a variety of levels from gross genomic rearrangements such as alterations in ploidy levels or chromosomal fusions, to minor changes such as alterations of single nucleotides. Within this thesis, I will be discussing variation, and the underlying mutation mechanisms which generate variation, at a specific class of tandemly repeated genetic elements in the mouse and human genomes: minisatellites.



## Population genetics of repetitive DNA

Whilst a variety of mutagenic processes introduce variation into populations, the persistence of variation depends on a number of forces; natural selection, genetic drift, and molecular drive. Selection acts on repetitive DNA. Telomeres function to maintain chromosomal integrity, while centromeric repeats serve as attachment sites for components of the mitotic and meiotic spindles (Alberts *et al.*, 1989). Rare illegitimate recombination events between interspersed Alu elements have been observed to cause deletions within the gene for the low density lipoprotein receptor causing familial hypocholesterolaemia (Lehrman *et al.*, 1987). Specific minisatellite alleles have been found to associate with predisposition to cancer and diabetes (Bennett *et al.*, 1995; Ding *et al.*, 1999), whilst dynamic expansions of a range of triplet repeat arrays can cause diseases such as mental retardation and neurodegenerative disorders (Richards and Sutherland, 1997).

However, most repeats in eukaryote genomes have no known function. There is an extraordinary amount of variation in total genome size between eukaryotes which bears little relation to differences in complexity of the organism, a phenomenon referred to as the C-value paradox. For example, the newt *Triturus cristatus* has ~6 times more DNA than humans, who have ~7.5 times more than the pufferfish *Fugu rubripes*. It is highly unlikely that such a large and variable amount of DNA is required for the maintenance of chromosome viability, and so the synthesis of large amounts of superfluous genetic material would appear to be selectively disadvantageous. However, selection does not generally appear to be a potent force acting to remove repetitive DNA, possibly due to the apparent inability of genomes to distinguish between essential and non-essential repetitive sequences. Furthermore, if at any one point in time there is little variation in total genome size within a population, levels of selection acting against slight increases in genome size would be minimal. Selection would therefore be essentially blind to a gradual increase in genome size. The primary forces acting on the majority of repetitive DNA are therefore likely to be mutation, genetic drift, and molecular drive.

Genetic drift is the random fluctuation in frequencies of polymorphic variants within populations caused by population sampling effects (Kimura, 1968). Some allelic variants will make no contribution to the next generation because by chance, their carriers fail to

survive to maturity, fail to mate, or the gametes carrying those variants are unsuccessful. The rate at which variation is lost from the population by genetic drift is dependent on three factors; population size, population substructure, and mutation rate. Molecular drive is proposed as the basis of concerted evolution which describes the observation that highly repeated sequences almost always show greater divergence between species than they do within species (Dover, 1982; Herman *et al.*, 1992). If a new mutation arises within a repeated element, the new variant appears either to be eliminated, or to spread to all other repeats both within an individual, and within the population/species. This process has been observed both within non-coding repeated DNA and multi-gene families such as the histones, histocompatibility genes, globins, rRNA genes, and snRNA genes (Dover and Flavell, 1984; Dover and Tautz, 1986). Although concerted evolution is a widely accepted phenomenon, the mechanisms which underlie the process of molecular drive remain obscure but are likely to include unequal exchange, transposition, RNA-mediated events, and gene conversion.

## **Classes of repetitive DNA**

### **Dispersed repeats**

#### ***Short interspersed elements***

Short interspersed elements (SINEs) are typically 80-400 bp long and in both humans and mice are preferentially located within R-bands of chromosomes with copy numbers in excess of 100000 (Boyle *et al.*, 1990). They display sequence homology to RNA genes and were probably inserted throughout the genome by a process of reverse transcription and integration, though they lack genes for reverse transcriptase. Full length SINEs typically contain internal promoters which can be recognised by RNA polymerase III. About 25 families of SINEs have been found in the genomes of mammals, reptiles, fishes, molluscs, insects, and plants (reviewed by Ohshima *et al.* (1993)). There are two main classes of SINEs. The first displays homology to the 7SL RNA component of the signal recognition particle and includes primate Alu elements and rodent-specific B1 elements (Krayev *et al.*, 1980). The second more extensive class includes SINEs which are derived from tRNA genes (usually tRNA<sup>Lys</sup> or tRNA<sup>Gly</sup>) and display extensive homology to these

genes at their 5' end. This second class includes rodent B2 elements (Kramerov *et al.*, 1979) and the MIR (mammalian-wide interspersed repeat) elements (Smit and Riggs, 1995).

The Alu element is the most abundant SINE in humans and was the first mammalian dispersed repeat to be isolated and sequenced (Houck *et al.*, 1979). Over 500000 copies of the 300 bp element are present throughout the human genome, representing 5-6% of the total genome. The Alu sequence has a dimeric structure composed of direct repeats of a region bearing homology to the 5' 100 bp, and 3' 45 bp of the 7SL RNA molecule and is flanked by short direct repeats, indicative of a transposon-like mechanism of chromosomal integration (Jelinek *et al.*, 1980; Van Arsdell *et al.*, 1981). Comparison of Alu elements among primates revealed that the majority of insertions were ancient events occurring prior to the major primate radiation. The ability to group Alu elements into subfamilies and the different levels of sequence divergence within subfamilies implies that the majority of Alu elements are transcriptionally and therefore transpositionally inert, and are derived from a limited number of sequentially activated master genes (Deininger *et al.*, 1992). More recent insertions have however been documented. For example, the HS/PV subfamily of elements is specific to humans and great apes, with copy numbers ranging from 2 in the gorilla (Lee flank *et al.*, 1993) to between 500 and 2000 in humans (Batzner and Deininger, 1991). MIR elements, with lengths of ~260 bp, are the second most abundant SINE in humans with an estimated 300000 copies in the human genome. They have been detected in all mammalian species analysed including marsupials. Their high level of sequence divergence and their presence at orthologous sites in different mammalian species are indicative of a very ancient origin and amplification (Smit and Riggs, 1995).

In the mouse genome, the B1 SINE is the most highly dispersed repeat with a copy number of 130000-180000, constituting 0.7-1.0% of the genome (Bennett and Hastie, 1984). Like Alu elements, they bear homology to 7SL RNA, but have a monomeric structure 130-150 bp in length (Sakamoto and Okada, 1985). B2 elements are 190 bp in length and are present in 80000-100000 copies, all of which display homology to tRNA<sup>Lys</sup> (Sakamoto and Okada, 1985). B2 elements in mice and rats can be further divided into two groups; 4.5s<sub>1</sub> and ID elements (Serdobova and Kramerov, 1998). The rat genome contains

20-500 times more ID elements than the mouse, indicative of a relatively recent SINE expansion. In multiple rodent species, the master gene for amplification has been identified as the BC1 RNA gene (probably derived from tRNA<sup>Ala</sup>) which is only found in the genomes of rodents, thus explaining the rodent specificity of the ID elements (Kass *et al.*, 1996). A further class of dispersed repeats in mice are the mouse transcript (MT) elements which have a consensus sequence of 400 bp, copy numbers of 40000-90000, and display structural features of retroposons (Bastien and Bourgaux, 1987). Whilst initially classified as SINE elements, the identification of a subpopulation of these elements found to flank a 1.1 kb internal sequence led to their re-classification as mammalian apparent LTR-retrotransposon elements (MaLR) (Kelly, 1994).

### ***Long interspersed elements***

Long interspersed elements (LINEs) can be 6-7 kb long and in both humans and mice are preferentially located within G-bands of chromosomes with copy numbers in excess of 100000 (Boyle *et al.*, 1990). They are generally highly heterogeneous in size, with most elements being truncated copies lacking 5' sequences. LINE expansion occurs through retrotransposition which requires LINE transcription, reverse transcription, and integration. Unlike SINEs, they may be able to replicate with greater autonomy as rare full length copies contain two open reading frames (ORFs) with both 5' and 3' untranslated regions (UTRs). The 5' UTR contains an internal RNA polymerase II promotor, presumably required for expression of both ORFs (DeBerardinis and Kazazian, 1998). ORF1 encodes a protein which binds LINE RNA in a sequence-specific manner, while ORF2 encodes a protein with reverse transcriptase and endonuclease activities. Both proteins are required for retrotransposition. LINE transcription from the RNA polymerase II promotor results in the newly inserted elements lacking promoters unless inserted adjacent to an RNA polymerase II promotor site (DeBerardinis and Kazazian, 1998).

The most abundant LINE family in humans is the L1 family which has a consensus of 6 kb although fewer than 5% are full length (Grimaldi *et al.*, 1984). The 5' UTR is conserved between L1 elements of both humans and gorillas (DeBerardinis and Kazazian, 1998). As with SINEs, they are flanked by direct repeats of <20 bp formed by duplication of the genomic site of integration. A predicted 40 active L1 elements are present in humans

(Sassaman *et al.*, 1997), the first of which (L1.2,<sup>1</sup>) was identified as the progenitor of a *de novo* L1 insertion into the factor VIII gene of a haemophilia A patient (Dombroski *et al.*, 1991).

In contrast to humans, the mouse genome contains at least five L1md (LINE 1 of *Mus domesticus*) families; A, F, V, T<sub>F</sub>, and Lx (DeBerardinis and Kazazian, 1998) with a total copy number of ~100000. Each family has a different 5' UTR consensus sequence, and at least three families are likely to have been active during the last 6 million years (Schichman *et al.*, 1992). The concerted evolution observed for mouse L1md elements is therefore due to the continual dispersion of new L1 elements carrying mutated sequences (Casavant *et al.*, 1988; Herman *et al.*, 1992). Examples of *de novo* retroviral insertions associating with disease phenotypes have also been observed in the mouse such as the dilute mutation (*d*) (Jenkins *et al.*, 1981), or a mutation in the hairless (*hr*) gene (Stoye *et al.*, 1988).

Both LINE and SINE sequences display an A-rich region of variable length at the 3' terminus, which is likely to be a relic of the polyadenylated RNA intermediate which forms during retrotransposition. The (A)<sub>n</sub> array is thought to allow the RNA intermediate to fold back on itself enabling self-priming for reverse transcription prior to integration (Hastie, 1996).

### **Multi-gene families and pseudogenes**

Large genomic deletions and duplications can occur through mechanisms of unequal and illegitimate crossing over, in some cases encouraged by the distribution of homologous dispersed repetitive elements. Duplications of genes can result in a degree of genetic redundancy and so allow functional divergence generating multi-gene families. Examples include the globin gene family and genes of the major histocompatibility complex (MHC). Alternatively, genetic redundancy can result in sequence degeneration of one of the duplicate copies with complete loss of function. The result are pseudogenes; loci which display sequence homology to functional genes but no longer produce active products. Pseudogenes can also be formed by reverse transcription and integration of processed mRNA gene transcripts as is evidenced by the lack of introns within many 'processed' pseudogenes (Lewin, 1994). This passive process of reverse transcription and integration is

likely to be similar to processes causing SINE expansions. Generally, such genes will be non-functional. However, integration near functional promoters may result in the gene becoming transcriptionally active, generating novel patterns of gene expression.

## **Tandemly repetitive DNA**

Historically, various classes of tandemly repetitive DNA have been defined by three criteria; size of repeat unit, total size of repeat array, and genomic location. Whilst these subdivisions are informative when describing the structure of each tandem repeat locus, it is becoming increasingly apparent that they may not precisely reflect the different mutational processes which generate polymorphism within different tandem repeat subclasses.

### **Satellite DNA**

Caesium chloride density gradient centrifugation of mouse DNA first led to the initial identification of genomic fractions, coined satellite DNA, which differed in GC content from the majority of the genome (Kit, 1961). Satellite DNA now refers to large arrays of centromeric tandem repeats (usually present in excess of 10000 copies), some of which differ in GC content from the rest of the genome hence their original detection (Singer, 1982). They are late replicating sequences and display characteristics of constitutive heterochromatin (Aker and Huang, 1996). Satellites can comprise up to 50% of the genomes of higher eukaryotes (Hastie, 1996). At least five classes of satellite sequences have been identified in humans (Tyler-Smith and Brown, 1987). Classes I-IV are composed of large arrays of short (5-25 bp) repeat units (Prosser *et al.*, 1986), whilst the major satellite, the  $\alpha$ -satellite, is composed of tandem repeats of a 171 bp sequence. Spanning up to 5 Mb at the centromeres of all human chromosomes, it is polymorphic both in length and repeat unit sequence, with observed clustering of homogeneous repeat types (Willard, 1990).

Satellite sequences in *Mus musculus* are relatively AT-rich and methylated at their CpG dinucleotides (Aker and Huang, 1996). Two different satellites predominate in the mouse. The major ( $\gamma$ ) satellite consists of 234 bp repeats composed of two smaller units of 118 and 116 bp, thought to be derived from three ancestral nonanucleotide repeats GAAAAATGA,

GAAAAAACT, and GAAAAACGT (Horz and Altenburger, 1981). The major satellite, present at about  $10^6$  copies, is located at the centromeres of all chromosomes except the Y chromosome (Hastie, 1996). The minor satellite is present in about 50000 copies in *Mus musculus* constituting up to 0.5% of the genome (Pietras *et al.*, 1983), and is composed of a tandemly repeated 120 bp monomer derived from smaller internal repeats similar to those found at the major satellite. It is located very close to the centromere of all but the Y chromosome and is flanked by the major satellite (Pietras *et al.*, 1983). Closely related species of *Mus* differ both in the amount and type of satellite DNA present. For example, *Mus spretus* has relatively little  $\gamma$  satellite, whilst *Mus caroli* has more  $\gamma$  satellite but totally lacks the minor satellite. *Mus musculus* has appreciable amounts of both sequences. High levels of polymorphism at the minor satellite exist even between closely related strains. However, this polymorphism is due to copy number as opposed to repeat sequence variations (Aker and Huang, 1996). This is in contrast to human  $\alpha$ -satellites which show sequence divergence between repeats of between 10% and 40%. (Choo *et al.*, 1991). Within a subset of repeats of both the human  $\alpha$ -satellite and the mouse minor satellite is found a 17 bp consensus binding site for CENP-B, one of the proteins found at active centromeres, indicative of a functional role for satellite sequences (Masumoto *et al.*, 1989).

Satellite DNA may promote chromosomal pairing during meiosis, act as specific binding sites for the mitotic and meiotic spindles, reduce the rate of recombination surrounding the centromere, or simply be an example of either 'selfish' or 'ignorant' DNA. Whatever its function, satellite DNAs are capable of rapid evolutionary change. One whale satellite is composed of 100000 copies of a 1.5 kb repeat unit (about 10% of the whale genome), whilst the homologous major satellite of dolphins is composed of repeat units 200 bp shorter. Since the divergence of whales and dolphins an estimated 200 million years ago, 100000 repeats must have undergone the same deletion/duplication event (Majerus *et al.*, 1996). This not only demonstrates the potentially rapid evolution of satellites, but also the phenomena of concerted evolution. The observation of extrachromosomal circular satellite DNA suggests that these repeats may be amplified by a process of rolling-circle replication, and reinsertion into the genome (Okumura *et al.*, 1987).

Satellite-like sequences have also been identified in non-centromeric locations such as a 250-500 kb array of a 40 bp repeat unit near the telomere of human chromosome 1 (Nakamura *et al.*, 1987), and a 10-50 kb array of a 61 bp tandem repeat in the pseudoautosomal region of the human sex chromosomes (Page *et al.*, 1987). These loci are collectively referred to as midisatellites.

### **Minisatellites**

Minisatellites are typically composed of 10-100 bp tandem repeat units within arrays of 0.5-50 kb (Armour *et al.*, 1990; Wong *et al.*, 1987). The first hypervariable locus to be described was *D14S1* (Balazs *et al.*, 1982; Wyman and White, 1980) and was later shown to be composed of short tandem repeats (Mulholland and Botstein, 1986). Further minisatellites were identified by chance due to their proximity to genes, such as a VNTR (Variable Number of Tandem Repeats) 5' of the insulin gene (Bell *et al.*, 1981; Bell *et al.*, 1982), a minisatellite 3' of the Harvey-Ras 1 gene (*HRAS1*) (Capon *et al.*, 1983), and minisatellites both within and near the  $\alpha$ -globin gene cluster (e.g. Proudfoot *et al.* (1982)).

Minisatellites include some of the most highly polymorphic loci in the human genome (Jeffreys *et al.*, 1999). This has been exploited with the development of various minisatellite-based DNA fingerprinting and profiling techniques which have been extensively used for forensic purposes, as genetic markers for linkage analysis, and to explore population structure (Armour *et al.*, 1996a; Jeffreys *et al.*, 1985b; Jeffreys *et al.*, 1986; Spurr *et al.*, 1994; Tamaki *et al.*, 1999a; Tamaki *et al.*, 1995). The first simultaneous detection of multiple polymorphic minisatellites (DNA fingerprinting) involved low stringency hybridisation to a genomic Southern blot of a probe derived from a short minisatellite composed of four tandem repeats of a 33 bp unit located within an intron of the human myoglobin gene (Jeffreys *et al.*, 1985a). The same probe was used to isolate further minisatellite loci from a human genomic library, including the clones  $\lambda$ 33.6 and  $\lambda$ 33.15 (Jeffreys *et al.*, 1985a), which are now most commonly used in DNA fingerprinting. Whilst in humans, PCR-based STR (Simple Tandem Repeat) typing has largely replaced DNA fingerprinting, the technique is still widely used for analysis of species where large numbers of microsatellite markers have not been developed (e.g. Signer *et al.* (1998); Yauk and Quinn (1996)).



There are an estimated 1500 minisatellites in both the mouse and human genomes (Bois *et al.*, 1998a; Jeffreys, 1987a). They are generally GC-rich and composed of tandemly repeated units of between 10 and 100 bp and in humans are clustered in proterminal regions of chromosomes. Polymorphism is due both to variation in the number of repeat units between alleles (length polymorphism), and small differences in repeat unit sequence within alleles. In contrast to STRs, variant repeats are generally interspersed throughout minisatellite alleles. These polymorphic interspersal patterns can be mapped by the technique of minisatellite variant repeat typing by PCR (MVR-PCR) (Figure 1.1) (Jeffreys *et al.*, 1991a) which greatly increases the informativity of minisatellite loci both for population studies, and the analysis of minisatellite mutational dynamics. AT-rich minisatellites show striking differences to GC-rich minisatellites with a typically modular structure where similar variant repeats are clustered into blocks within the repeat array (Bois and Jeffreys, 1999).

### **Telomeres**

Telomeres are a class of tandem repeat loci located at the termini of eukaryotic chromosomes and act to resist chromosomal fusions, degradation, and allow complete replication of chromosome ends. In both humans and mice, they are composed of tandem repeats of the hexameric consensus sequence TTAGGG which are added to chromosome ends by telomerase, a ribonucleoprotein complex which uses an RNA template for repeat unit addition by reverse transcription (Collins, 1996). This prevents the progressive and lethal shortening of chromosomes during replication in the germline. In addition to the germline, telomerase activity has been detected in many types of cancer and immortal cell lines but not in normal somatic tissues, and has therefore been implicated in cell senescence (Harley *et al.*, 1990). The action of telomerase to prevent telomere shortening specifically in the germline was recently highlighted by the detection of reduced telomere lengths in the cloned sheep Dolly (Shiels *et al.*, 1999).

In humans, telomeres are typically 5-10 kb in length, whilst in mice repeat arrays can be up to several hundred kb (Kipling and Cooke, 1990; Starling *et al.*, 1990). The acrocentric nature of all 20 *Mus musculus* chromosomes means many telomeres at the centromeric chromosomal ends lie within 1 Mb of minor satellite sequences (Kipling *et al.*, 1991).

## Figure 1.1

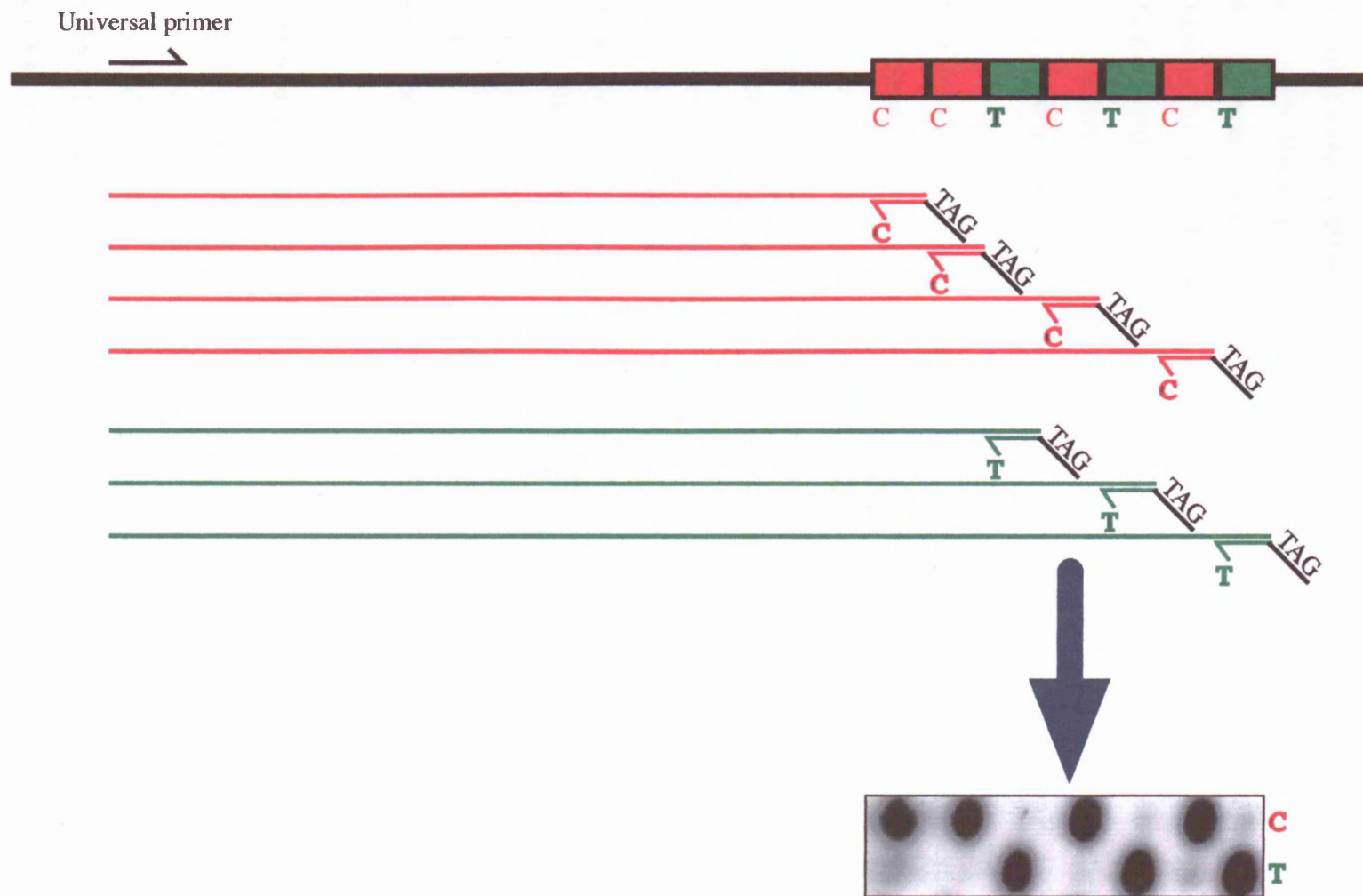
### ***Minisatellite variant repeat mapping by PCR (MVR-PCR)***

The system of MVR-PCR presented is based on the minisatellite *D19S20* (investigated in Chapter 7), and describes the technique in its simplest form. Two variant repeats are present at this locus named C and T (names here reflect the identity of the variant site within repeat units). PCR amplification between a universal flanking primer and a repeat-specific primer which binds to C-type repeats generates a population of amplicons the sizes of which reflect the positions of C-type repeats within the allele. A second PCR reaction using an T-type repeat-specific MVR primer similarly generates a population of amplicons which reflect locations of T-type repeats. Repeat-type distribution can be determined by electrophoresis of the PCR products and Southern blot hybridisation.

If only two primers were used, larger amplicons would be progressively lost during subsequent cycles of PCR as large amplicons contain binding sites for the repeat-specific primers. To reduce this loss, a 5' sequence (referred to as TAG) is added to the repeat specific primers. A third primer is used with sequence identity to TAG. Both TAG and the universal flanking primer are used at higher concentrations than the repeat-specific primer, and so subsequent cycles of amplification preferentially occur between the flanking and TAG primers, thereby preserving amplicon length.

A complementary approach (employed in Chapter 8) is to elevate the annealing temperature after a few cycles of PCR to prevent binding of the repeat-specific primer, whilst allowing the other primers to anneal.

Figure 1.1



Surprisingly high levels of telomere length variation have been observed both between and within inbred strains of mice (Starling *et al.*, 1990). The mechanisms which underlie this variation are unclear.

A number of studies have identified chromosomal ends as 'hotspots' for genome evolution (Macina *et al.*, 1995; Royle *et al.*, 1994; Trask *et al.*, 1998). Immediately adjacent to telomeres are subtelomeric domains of between a few hundred bp (human XpYpter (Baird and Royle, 1997)) to over one hundred kb (human 16pter (Wilkie *et al.*, 1991)). At 16pter, 4 different subtelomeric allele lengths have been described, some of which show greater homology to subtelomeric domains of other chromosomes than to other 16pter allelic variants, indicative of non-homologous chromosomal exchange (Harris and Thomas, 1992; Wilkie *et al.*, 1991). At XpYpter, three distinct haplotypes were identified over 19 polymorphisms in a subtelomeric region of 850 bp (Baird *et al.*, 1995). Intermediate haplotypes are apparently absent from modern human populations, despite divergence of the haplotypes occurring after the separation of *Pan* and *Homo* lineages (Baird and Royle, 1997). The mechanisms underlying the generation and maintenance of such highly diverged haplotypes are unclear. Whilst telomeres act to prevent chromosomal fusions, karyotypes can be highly diverged between species. For example two acrocentric chromosomes present in chimpanzees and gorillas, fused to form human chromosome 2 by a Robertsonian fusion. At the fusion point are located interstitial telomeric repeats, an apparent relic of this fusion event (Jdo *et al.*, 1991). Similarly, whilst the standard laboratory strains of *Mus musculus* have a karyotype of N=20 acrocentric chromosomes, closely related strains from W. Europe and N. Africa can have as low as N=11 due to the fusion of acrocentric chromosomes forming metacentric chromosomes (Nachman *et al.*, 1994).

### **Expanded simple tandem repeats**

Expanded simple tandem repeats (ESTRs) are common in the mouse but not human genome, and often display high levels of instability and polymorphism (Bois *et al.*, 1998b; Jeffreys *et al.*, 1987b; Kelly *et al.*, 1989; Kelly *et al.*, 1991). These loci were initially identified by DNA fingerprinting so were classed as minisatellites (Jeffreys *et al.*, 1987b), although further characterisation revealed them to be composed of very short repeat units and led to their re-classification as ESTRs (Bois and Jeffreys, 1999). Many ESTRs studied

to date apparently arose by expansion from within dispersed repetitive elements. Mouse locus *Hm-1* (originally called *Ms6-hm*) is composed of 1-16 kb of the perfect pentameric repeat GGGCA (Chapter 5; Kelly *et al.*, 1989) whilst *Hm-2* is composed of 2-25 kb of the tetramer GGCA (Gibbs *et al.*, 1993). Both repeats arose by expansions from precisely the same point within MT elements (Kelly, 1994). In addition, a large family of mouse ESTRs called the MMS10 family has been identified, composed of GGCAGA repeat units expanded from within B1 elements throughout the mouse genome (Bois *et al.*, 1998b). The MMS10 family is rodent-specific, reflecting the rodent-specificity of B1 elements.

ESTRs can display high levels of instability in both germline and somatic tissues. *Hm-1* and *Hm-2* display germline mutation rates of ~2.5% and ~3.6% per gamete respectively, with ~2.8% and ~30% of adult mice showing evidence of somatic mosaicism detectable by genomic Southern blot analysis. The identification of mice which are mosaic for the same alleles in both germline and soma indicates that mitotic instability arises during the first few divisions of the developing blastula (Gibbs *et al.*, 1993; Kelly *et al.*, 1989). Mutations are likely to arise from replication slippage possibly encouraged by the formation of unusual DNA conformations associated with the repeated array. For example, *Hm-1* is capable of *in vitro* formation of a hairpin structure, as well as two different intra-strand tetraplexes (Weitzmann *et al.*, 1998).

### ***Simple tandem repeats***

Simple tandem repeats (STRs) or microsatellites have been the focus of intensive study due to their utility in the construction of genome-wide linkage maps (e.g. Dib *et al.* (1996)), their informativity in the analysis of variation within and between populations (e.g. Tautz (1989)), their utility in human identification (e.g. Gill *et al.* (1994)), their use in the study of molecular evolution (e.g. Meyer *et al.* (1995)) and the association of instability in a subgroup of STRs with a variety of human genetic diseases (e.g. Stallings (1994)). They are typically short arrays composed of repeat units of 1-5 bp in length. Array lengths vary between species, for example 43% of rat STRs have array sizes in excess of 40 bp compared to only 12% in humans (Beckman and Weber, 1992). They are distributed throughout all eukaryote genomes with an average frequency in humans of one STR per 6 kb and display no tendency towards clustering within specific regions (Beckman and

Weber, 1992). In mice, they show high levels of polymorphism even between inbred strains of *Mus musculus* where ~50% of loci are variable, compared to ~90% variability between *Mus musculus* and either *Mus spretus* or *Mus castaneus* (Hearne *et al.*, 1992). In humans, the most abundant repeat types are A, AC, AAAN, AAN, and AG (in decreasing order of abundance) with ~80% of A, AAAN, and AAN loci located within 3' Alu sequences. In the rat the most common STRs are (AC)<sub>n</sub> and (AG)<sub>n</sub>, with a lower frequency of both (A)<sub>n</sub> and (AAAN)<sub>n</sub> loci (Beckman and Weber, 1992). In contrast to humans, studies in mice demonstrate that mononucleotide repeat arrays are not preferentially associated with SINE elements (Aitman *et al.*, 1991).

Heritable mutation rate at STRs is typically between  $10^{-3}$  and  $10^{-5}$  per generation and differs substantially between loci (Dietrich *et al.*, 1992; Heyer *et al.*, 1997; Weber and Wong, 1993). In humans, pedigree analysis demonstrated that dinucleotides have on average a four-fold lower mutation rate than tetranucleotides ( $5.6 \times 10^{-4}$  compared with  $2.1 \times 10^{-3}$  (Weber and Wong, 1993)). Studies in yeast indicate that this difference in mutation rate results from the tendency for larger repeat units to form larger repeat arrays as mutation rate was found to correlate with STR size more closely than repeat unit size (Pupko and Graur, 1999). There is also a strong correlation between mutation rate and the number of homogeneous repeats (Wierdl *et al.*, 1997) and studies in yeast and in human pedigrees both indicate that interspersed variant repeats have an inhibitory effect on mutation (Brinkmann *et al.*, 1998; Petes *et al.*, 1997).

About 90% of mutations are thought to involve the gain or loss of a single repeat unit (Brinkmann *et al.*, 1998). It is generally accepted that replication slippage is the major mechanism causing new mutations at STRs (Levinson and Gutman, 1987). Slippage implies the displacement of strands of a denatured duplex, followed by the mispairing of complementary bases within an existing STR array. To support this, the germline mutation rate in humans is ~6-fold higher in males than in females. Furthermore, mutation rate is significantly higher in older men than younger men (Brinkmann *et al.*, 1998). This difference reflects the number of cell divisions during germ-cell genesis: oogonia undergo ~22 divisions prior to meiosis, whereas for a 29-year-old man, the average sperm cell is

derived from a progenitor which has undergone an estimated 350 premeiotic cell divisions, which increases with age (Vogel and Motulsky, 1997).

Whilst replication slippage is likely to be the major source of mutation, deficiencies in DNA repair systems can also cause repeat instability as errors resulting from slippage will remain unrepaired. For example, abnormalities in the mismatch repair gene *MSH2* result in elevated microsatellite instability, observed in colorectal cancers (Leach *et al.*, 1993; Umar *et al.*, 1994). Instability in tumour cells has also been reported for other repeated elements such as Alu repeats (Ionov *et al.*, 1993) and the minisatellite MS1 (Armour *et al.*, 1989a).

### ***Cryptic tandem repeats***

A large number of tandem repeats also exist in cryptic form. To distinguish from STRs, cryptic repeats are defined (somewhat arbitrarily) as repeat arrays in which any one repeat unit sequence represents less than 2/3 of the total array (Jacobson *et al.*, 1993). As with STRs, cryptic repeats can be highly polymorphic. One example is repeats of alternate purines and pyrimidines (RY)<sub>n</sub> which are highly enriched in human and mouse genomes. An (RY)<sub>n</sub> repeat in the human factor IX gene displays length polymorphism which may predispose to deletions causing haemophilia B (see Ricke *et al.* (1995)). Cryptic (YRR)<sub>n</sub> arrays are also abundant in both humans and mice. Instability within (YRR)<sub>n</sub> arrays has been shown to account for a substantial fraction (~30%) of deletions and insertions detected within protein sequences between species (Ricke *et al.*, 1995).

It has been proposed both that STRs can evolve from cryptic repeats, and that cryptic repeats can evolve from STRs. An example of the former comes from phylogenetic analysis of the  $\eta$ -globin locus in primates. In the apes, a G → A transition at an ATGTGTGT sequence generated an (ATGT)<sub>2</sub> repeat which evolved to a (ATGT)<sub>5</sub> STR in humans, and an (ATGT)<sub>4</sub> STR in the African apes. In the lineage leading to the Owl monkey, an A → G transition at the same locus generated a (GT)<sub>4</sub> STR which expanded to the present day (GT)<sub>6</sub> allele (Messier *et al.*, 1996). The degeneration of STRs to form cryptic repeats has been proposed as a fundamental mechanism underlying protein evolution. In eukaryote proteins, glutamine is the most common amino acid in homogeneous repeat arrays of over 16 residues. Two genes in humans, involucrin and *GRP1*, may have both formed from

polyglutamine arrays which diverged by nucleotide substitution with the coding (CAG)<sub>n</sub> array. Higher order repeat homogenisation is observed at the involucrin gene, possibly reflecting the action of a form of molecular drive on the locus (see Djian (1998)).

### ***Dynamic tandem repeat expansions and human disease***

The last decade has seen a steady increase in the number of human genetic diseases found to be caused by dynamic expansions of tandemly repeated loci, primarily of triplet repeats (Djian, 1998; Mitas, 1997). For disease-associated triplet repeat loci, alleles display size polymorphism in the unaffected population. A subset of alleles show an increased number of perfect tandem repeats. These larger alleles are predisposed to germline expansions which associate with disease phenotypes. Events which lead to pre-expansions are likely to be rare, as haplotype analysis reveals that most expanded alleles have a common ancestral haplotype (Richards and Sutherland, 1997). In general, transmission of expanded alleles to subsequent generations results in progressive lengthening of the repeat array with increased severity of disease phenotype and/or earlier age-of-onset. This is referred to as anticipation. The probability of repeat array expansion depends not only on array length and repeat homogeneity but also on the parent-of-origin. Large expansions of fragile X arrays occur only in the female germline, whilst in (CAG)<sub>n</sub> expansions, germline mutations are more frequent and larger when transmitted from fathers (see Brock *et al.* (1999)). However, these parent-of-origin differences in triplet repeat expansions were detected by pedigree analysis which may not provide a representative picture of germline mutation dynamics, due to selection acting either pre- or post-fertilisation.

The largest subgroup of triplet repeat disorders identified to date are characterised by expansions of (CAG)<sub>n</sub> repeats within coding regions of genes, leading to large arrays of polyglutamine tracts in the mature protein. These unstable expansions have been identified as the mutation common to at least eight neurodegenerative diseases including spinal and bulbar muscular atrophy (SMBA), Huntington's disease (HD), spinocerebellar ataxia types 1, 2, 6, and 7 (SCA1, SCA2, SCA6, and SCA7), dentatorubral-pallidoluysian atrophy, and Machado-Joseph disease (MJD, or SCA3) (David *et al.*, 1997; The Huntington's Disease Collaborative Research Group, 1993; Imbert *et al.*, 1996; Kawaguchi *et al.*, 1994; Koide *et al.*, 1994; La Spada *et al.*, 1991; Orr *et al.*, 1993; Pulst *et al.*, 1996; Sanpei *et al.*, 1996;



Zhuchenko *et al.*, 1997). With the exception of the *SCA6* expansion, the normal and affected size ranges of polyglutamine tracts are common between all disease loci with unaffected individuals containing <30 repeats, and affected individuals with >50 repeats (Ashley and Warren, 1995). Expanded polyglutamine arrays may cause neurodegenerative disease by inducing apoptosis through enhanced activity of pro-apoptotic proteases such as apopain (see Djian (1998)). Cleavage of the HD locus-encoded protein huntingtin by apopain has been reported (Goldberg *et al.*, 1996b), and fragments of huntingtin have been detected in the brains of HD-affected patients (DiFiglia *et al.*, 1997).

Other triplet repeat expansions associated with disease occur in non-protein-coding regions. In Friedrich's ataxia, disease is associated with an intronic expansion of an (AAG)<sub>n</sub> array resulting in the inhibition of transcription (Campuzano *et al.*, 1996). Unlike other triplet repeat expansion disorders, no obvious anticipation is observed. This is due to the recessive nature of the disease as the consequences of copy number increases in successive generations are unlikely to be observed. Fragile X syndrome is the most common form of single locus mental retardation and results from the loss of function of the X-linked *FMR1* gene through hypermethylation and altered chromatin structure, caused by the massive expansion from ~60 to up to 3000 repeats of a (CGG)<sub>n</sub> motif in the 5'UTR (Green and Djian, 1998). It is this expansion which generates a fragile site on the X chromosome identifiable as a non-staining gap on the chromosome inducible by certain conditions of cell culture, after which the disease was named. Myotonic dystrophy (DM) is caused by the expansion of a (CTG)<sub>n</sub> motif within the 3'UTR of the myotonic dystrophy protein kinase gene (*DMPK*) (see Djian (1998)). This expansion may condense chromatin surrounding an enhancer element of the 3' myotonic dystrophy-associated homeodomain protein gene (*DMHAP*) which encodes a transcription factor normally expressed in all tissues affected by DM, resulting in *DMHAP* haploinsufficiency and disease (Heath *et al.*, 1997; Klesert *et al.*, 1997).

Whilst a stepwise mutation model can account for allele size distribution observed at disease-associated triplet repeat loci in the unaffected population, it clearly cannot incorporate the large dynamic expansions observed. As with the mutability of other STR loci, variant repeats within triplet arrays have been shown to stabilise arrays at *SCA1*, *SCA2*

and *FMRI* (Chung *et al.*, 1993; Imbert *et al.*, 1996; Zhong *et al.*, 1995) indicating that the mechanism of expansion may involve polymerase slippage during replication. The question remains as to why certain triplet repeat arrays are capable of dynamic expansions whilst other STRs such as dinucleotide repeats are not. One possible answer comes in part from analysis of an AT-rich minisatellite FRA16B. This locus is composed of variant tandem repeats with a 33 bp consensus sequence, and is capable of massive expansions similar to those witnessed at triplet repeat loci, giving rise to an array of up to 2000 perfect repeats which generates a fragile site. These expansions are again encouraged by repeat unit sequence homogeneity (Yu *et al.*, 1997). Sequence analysis of the repeat unit reveals the potential for the formation of a hairpin loop in which 26 A-T Watson-Crick base pair bonds may be formed from the 33 bases of a single repeat.

Similarly, biophysical and biochemical studies demonstrate that 5 of the 6 triplet repeat arrays associated with human diseases (CGG, CCG, CAG, CTG, GAA, but not CTT) form stable hairpin structures under physiological conditions (reviewed by Mitas (1997)). Other secondary structures such as triplexes or tetraplexes may also form. It has been proposed that these secondary structures form when a certain threshold number of perfect tandem repeats are present, and that these structures promote DNA recombination and/or interfere with DNA replication resulting in instability (Mitas, 1997). For example during DNA replication, lagging strand synthesis is dependent on the processing of Okazaki fragments by the FEN-1 protein (flap endonuclease and five prime exonuclease 1) (Lieber, 1997). One model for the action of FEN-1 is that synthesis of an Okazaki fragment displaces the 5' end of the downstream strand creating a flap which is cleaved by the endonuclease activity of FEN-1 (Lieber, 1997). DNA hairpin structures are resistant to FEN-1 activity (Harrington and Lieber, 1994), and this resistance may increase the frequency of replication errors and of strand slippage.

However, if hairpin formation was sufficient to cause instability, other tandem repeats such as (AT)<sub>n</sub> would be expected to display the same mutational characteristics. It has therefore been suggested that expansions only result from 'flexible' hairpins characterised by imperfect base pairings within the structure. Flexible hairpins would be able to denature more easily allowing the incorporation of additional repeats. The resulting structure would

have to be exceptionally unstable to account for observed mutation rates which can be as high as 98% in sperm (Leeftang *et al.*, 1999). A far more widespread role for secondary DNA conformation in tandem repeat instability has been proposed following *in vitro* observations that triplet repeats, ESTRs (*Hm-1*), minisatellites (insulin minisatellite), and satellites are capable of forming fold-back structures *in vitro* (Weitzmann *et al.*, 1997; Weitzmann *et al.*, 1998). However, it is unclear whether any such DNA conformations actually form *in vivo* so conclusions based upon unusual DNA structures must remain speculative.

*Cis*-acting sites adjacent to triplet repeat arrays have been implicated in instability, suggesting a role for DNA conformation outside the array, or the binding of *trans*-acting factors in mutation induction. Therefore a model of mutation based exclusively on the structure of tandemly repetitive DNA may be at best a simplification (Brock *et al.*, 1999). Furthermore, the observation that the identity of the normal allele at the *SCA3* locus affects the stability of the expanded allele in heterozygotes indicates that mechanisms of instability may have an inter-allelic component (Igarashi *et al.*, 1996).

### ***Phenotypic associations of minisatellites***

#### **Coding minisatellites**

Whilst the majority of minisatellites are located outside known genes, a number of loci have been identified within coding sequences. Variation in the size of coding VNTRs has, in some cases, been related to genetic disease. For example, the *apo(a)* gene is a member of the human lipoprotein family, all of which contain coding minisatellites composed of 33 or 66 bp tandem repeats (Boguski *et al.*, 1986). The *apo(a)* VNTR is polymorphic, ranging in size from 15-40 repeats resulting in variation in the size of its protein product from 300 to 800 kDa (Koschinsky *et al.*, 1990). There is an inverse correlation between repeat number and the concentration of the *apo(a)* protein product in the plasma. High levels of *apo(a)* lead to coronary vessel disease in addition to brain vascular and artery disorders (reviewed by Bliskovskii (1994)). Similarly, members of the mucin gene family (*MUC1-4*) have extensive coding minisatellites. *MUC1* contains 20-125 tandem repeats of a 60 bp unit (Gendler *et al.*, 1990). *MUC1* encodes episalin, a cell surface antigen

expressed by many epithelial cell types, which has an extracellular domain of between 1000 and 2200 residues depending on VNTR length (Lancaster *et al.*, 1990). *MUC1* overexpression has been observed in mammary carcinomas and some other adenocarcinomas, and may result in metastasis by the episalin molecule masking other cell surface adhesion molecules (Hilkens *et al.*, 1992). The D4 dopamine receptor (D4DR) contains a polymorphic 48 bp tandem repeat array, the length of which affects ligand binding affinity (Asghari *et al.*, 1994; Lichter *et al.*, 1993). D4DR is expressed at high levels in the limbic areas of the brain, and variation in its ligand binding capacity has been associated with cognitive and emotional disorders (Ebstein *et al.*, 1996). Coding minisatellites have also been described in genes for acrosomal protein sp-10 precursor, cholesterol esterase, neurofilament triplet H protein, coagulation factor V, and the loricrin and involucrin genes (reviewed by Nakamura *et al.* (1998)). However, not all coding minisatellites are known to be polymorphic or associated with disease.

### Regulatory minisatellites

A number of minisatellites located in putative gene regulatory regions have been associated with elevated susceptibility to disease. Two such loci are located at 11p15.5: the insulin minisatellite is 596 bp upstream of the insulin gene translation initiation site, with the *HRAS1* VNTR 1 kb 3' of the *HRAS1* proto-oncogene polyadenylation signal. The phenotypic associations of the insulin minisatellite will be discussed in detail in Chapters 9-10.

The *HRAS1* VNTR is composed of 30-100 copies of a 28 bp repeat which displays homology to the consensus sequence of the binding site for members of the rel/NF- $\kappa$ B family of transcription factors (TFs) (Kiaris *et al.*, 1995; Trepicchio and Krontiris, 1992). Binding of rel/NF- $\kappa$ B TFs to the VNTR activates transcription of a reporter gene in some cell lines, including cells derived from the FJ bladder carcinoma (Trepicchio and Krontiris, 1992). Five common alleles of the VNTR have been identified in addition to many rare variants (Phelan *et al.*, 1996). Rare alleles have been shown to associate with elevated risk of a number of forms of cancer, including carcinomas of the breast, colon, bladder, ovary, and acute leukaemia, although these associations are unconfirmed in several studies (reviewed by Krontiris (1995); Nakamura *et al.* (1998)). For example, women carrying

mutations in the *BRCA1* gene have a 2-fold elevated risk of developing ovarian cancer if they also have one or two rare alleles of the *HRAS1* VNTR, although no increased risk of breast cancer was detected (Phelan *et al.*, 1996). The mechanism of action of the VNTR in cancer susceptibility is unclear. Whilst it may act as a transcriptional regulator of the *HRAS1* gene through the binding of rel/NF- $\kappa$ B TFs, it has been proposed that the identification of rare alleles is an effect as opposed to a cause of cancer susceptibility. A study which genotyped lung cancer patients at the *HRAS1* VNTR, a minisatellite at *D17S4*, and 17 microsatellites, found that cancer patients with rare alleles at *HRAS1* also tended to have rare alleles at other loci (Lindstedt *et al.*, 1997). It was proposed that rare *HRAS1* alleles are indicators of increased genomic instability which may lead to cancer and so are not causally related to cancer predisposition (Lindstedt *et al.*, 1997).

Progressive myoclonus epilepsy of the Unverricht-Lundborg type (EPM1) is a rare autosomal recessive disorder characterised by generalised seizures of progressive and incapacitating myoclonus and is caused by mutations in the cystatin B gene (*CSTB*) which encodes a cysteine protease inhibitor (Virtaneva *et al.*, 1997). The majority of disease-associated mutations are due to expansion of a tandem repeat array upstream of *CSTB* (Laloti *et al.*, 1997). Normal alleles contain 2-3 copies of the 12 bp repeat. Rare alleles not associated with disease have been identified of 12-17 repeats, whilst repeat arrays from EPM1-affected individuals range in size from 30-75 repeats (Laloti *et al.*, 1998). In contrast to the dynamic expansions associated with triplet repeat diseases there is no evidence for anticipation, the repeat sequence of the EPM1 minisatellite is not compatible with hairpin formation, and even in expanded alleles there is a high degree of variant repeat sequence divergence. Furthermore, whilst expanded alleles show a high mutation rate (47%), they show no evidence for the gain or loss of large numbers of repeat units (Larson *et al.*, 1999). Expansion of the minisatellite results in a reduction in *CSTB* expression resulting in disease (Pennacchio *et al.*, 1996). Recent *in vitro* studies of *CSTB* transcription replacing expanded minisatellite alleles with a similar size fragment of non-repetitive DNA demonstrated that this reduction is due to changes in promoter organisation, as opposed to any direct effects of the tandemly repeated DNA (Laloti *et al.*, 1999).

Other VNTRs which may influence transcription of human genes include a minisatellite

composed of 50 bp tandem repeats within the  $D_H-J_H$  interval in the immunoglobulin heavy chain gene (Trepicchio and Krontiris, 1993), and VNTRs both in the 5' promotor region, and in the second intron of the serotonin transporter gene (Heils *et al.*, 1996; Ogilvie *et al.*, 1996).

## **Minisatellite Mutation**

### ***Mutation dynamics of AT-rich human minisatellites***

To date, five AT-rich minisatellites have been identified in humans; the autosomal loci *COL2A1* (Berg and Olaisen, 1993), *ApoB* (Buresi *et al.*, 1996), FRA16B (Yu *et al.*, 1997), FRA10B (Hewett *et al.*, 1998), and MSY1 located within the non-recombining region of the Y chromosome (Jobling *et al.*, 1998). The putative role of hairpin formation at FRA16B has already been discussed. It is therefore striking that each of these 5 AT-rich loci are composed of repeat unit sequences which display the potential for hairpin formation suggesting that each locus may have been generated by a common mutational process. MSY1 displays a high level of germline instability (5% per sperm) which, by definition, is the result of intra-allelic processes (Jobling *et al.*, 1998). MVR-PCR analysis at MSY1 reveals that variant repeats are clustered in homogeneous blocks, indicative of the linear diffusion of variants probably through slippage replication which may be encouraged by hairpin formation (Bouzekri *et al.*, 1998).

### ***Mutation dynamics of GC-rich human minisatellites***

#### **Patterns of population allele variation**

The high mutation rate of many minisatellites make them amongst the most variable loci in the human genome. This high mutation rate, and the capacity to map their internal structures by MVR-PCR, make them ideal loci for analysing processes underlying genome turnover. The comparison of allele structures allows inferences to be made as to the mechanisms which generate the observed variability. One approach is to characterise patterns of allele variation within populations which generally result from two opposing forces; *de novo* mutation which elevates variation, and genetic drift (and in some cases selection) which reduces variation. The alignment of allele structures allows indirect inferences to be made concerning mechanisms generating variation. For example, the minisatellites MS31, MS32, and MS205 display patterns of variation polarised towards one

end of the repeat array (Armour *et al.*, 1993; Jeffreys *et al.*, 1991a; Jeffreys *et al.*, 1990; Neil and Jeffreys, 1993), initially indicating either that properties within one end of the minisatellite predispose to mutation initiation, or that a *cis*-acting site within the flanking DNA at the variable end of the array drives mutation. Other minisatellites such as CEB1 (Buard and Vergnaud, 1994) show no obvious polar variation.

Characterisation of allele diversity both within and between populations can also shed light on the ancestry of populations. MVR-PCR mapping of MS205 allele structures revealed highly polarised allele variation (Armour *et al.*, 1993). Despite a mutation rate of 0.4% per generation, the 5' end has a high level of stability allowing alleles which diverged from a relatively distant common ancestor to be readily aligned into groups (Armour *et al.*, 1993). MVR-PCR mapping of alleles from both African and non-African populations detected clear differences in levels of allele diversity between populations. Non-African populations contained a limited range of 5' allele structures which formed a subset of the diversity found of African chromosomes, supporting a recent African origin for modern human diversity at this locus (Armour *et al.*, 1996a).

#### Detection of *de novo* mutation events

An alternative approach to determine the basis of allele diversity is to identify and characterise *de novo* mutation events. The comparison of mutant allele structures with the structure of progenitor alleles can be highly informative as to mutation dynamics. *De novo* mutations at many minisatellite loci can be detected by pedigree analysis (Jeffreys *et al.*, 1994). However, this approach is labour intensive, and the small number of children within pedigrees prevents the comparison of mutation dynamics between different individuals. An alternative to pedigree analysis is the detection of mutants in sperm DNA. A single ejaculate can contain in excess of  $10^8$  sperm, each of which can be considered analogous to a potential offspring. For many minisatellites, analysis of sperm DNA can therefore provide an effectively unlimited number of mutants. However, characterisation of mutation events arising in the female germline must still rely on pedigree analysis.

Mutation events can typically be detected in sperm DNA by diluting genomic DNA and amplifying small pools of 50-100 sperm equivalents of DNA by PCR; a technique called small pool PCR (SP-PCR) (Jeffreys *et al.*, 1994). Mutant molecules are identified by changes of length from progenitor alleles. Further dilutions to a level where the average input of molecules is 0.5-0.8 molecules per PCR reaction allows the number of amplified molecules to be accurately estimated by Poisson analysis, and so a mutation rate can be calculated. Whilst SP-PCR has proved to be highly effective for the detection of mutants for loci with mutation rates in excess of  $10^{-3}$  per cell, it is not recommended at loci with lower mutation rates for two reasons. For a locus with mutation rate of  $10^{-5}$ , it would be necessary to perform an average of 1000 PCR reactions to detect a single mutant molecule, making the process highly labour intensive. Furthermore, jumping PCR and polymerase priming from slipped single-strand nicks in the template DNA can generate molecules of different lengths to the progenitor alleles (A. Jeffreys, pers. commun.), so a large proportion of observed length-changed alleles would be PCR artefacts as opposed to true mutant molecules. To overcome this deficiency, genomic DNA is digested with restriction endonucleases which target sites flanking the minisatellite, and length-changed mutants are separated by size from progenitor molecules by electrophoresis. The recovery of multiple size fractions of DNA depleted in progenitor molecules, and their amplification by SP-PCR, allows mutation detection at a frequency of as low as  $10^{-6}$  per sperm, although difficulties in estimating levels of enrichment mean this approach is only semi-quantitative (Jeffreys and Neumann, 1997). Furthermore, for each PCR reaction the expected size range of true mutants must correspond to the size fraction from which the DNA was derived, allowing the differentiation of true mutants from artefacts. This technique is called size-enrichment small pool PCR (SESP-PCR) (Jeffreys and Neumann, 1997).

### Characteristics of minisatellite mutation in humans

Germline mutation events have been extensively studied by SP-PCR at the minisatellites MS32, MS205, B6.7, and CEB1 which display mutation rates in sperm of 0.8%, 0.4%, 5%, and 13% respectively (Jeffreys *et al.*, 1994; May *et al.*, 1996; Tamaki *et al.*, 1999b; Buard *et al.*, 1998). Whilst there are substantial differences in the mutation profiles detected at these loci, a number of general rules can be seen to apply (reviewed by Jeffreys *et al.* (1997)). The majority of mutations involve small changes in allele length. There is also a



substantial bias towards mutations resulting in repeat gains, for example 74% of 761 germline mutants detected at MS32 were larger than the progenitor (Jeffreys *et al.*, 1994), with a similar bias detected at CEB1 (Buard and Vergnaud, 1994). Mutation rate is elevated in the male germline compared to females; at CEB1 mutation rate in sperm can be as high as 15%, whilst mutation rate in oocytes estimated from pedigree analysis is 0.2% (Vergnaud *et al.*, 1991). This is however not true for all loci, for example mutation rate at MS32 is the same for both male and female transmissions (Jeffreys *et al.*, 1994).

MVR mapping of mutants and their comparison to progenitor alleles reveals a highly complex mutation process involving both intra- and inter-allelic mechanisms resulting in extensive scrambling of repeat-type distribution (reviewed by Jeffreys *et al.* (1999)). In most cases, novel repeat types are not generated and repeat unit length is conserved. The division of mutations into intra- and inter-allelic events is in some cases difficult. Mutant structures have been observed in which either variant repeat types or repeat motifs present in one progenitor allele are transferred to the other allele. Such transfers are frequently in register between alleles, suggesting that alleles pair within the flanking DNA either prior to or during processing of the recombination complex. These apparent in-register transfers in other cases do not introduce novel repeat types into the mutated progenitor allele, but may be classified as inter-allelic events due to their in-register characteristics. Other mutations which have no characteristic repeat transfers between alleles, or which clearly result from repeat motif duplications and deletions are classified as intra-allelic. However, an inter-allelic component to the generation of such mutants cannot be excluded. The differences in mutation profile between loci are apparently due to differences in the relative frequencies of mutational processes and will be described with reference to the two most intensively studied loci, MS32 and CEB1.

At MS32, at least 80% of mutation events involve complex in-register inter-allelic transfers of repeats between progenitor alleles, frequently accompanied by duplications flanking the inserted repeats, and are highly polarised towards the 5' end of the array (Jeffreys *et al.*, 1994). Conversion is the dominant form of mutation, with the exchange of distal flanking polymorphic markers occurring at very low frequency (Wolff *et al.*, 1989). Mutation rate is variable between alleles and is independent of array length (Jeffreys *et al.*, 1994). In

contrast to MS32, on average 75% of mutations identified at locus CEB1 involve complex intra-allelic duplications which display no obvious polarity, hence the lack of polarity of variation observed in the population (Buard *et al.*, 1998; Buard and Vergnaud, 1994). Mutation rate varies with allele length (Buard *et al.*, 1998) with a proportional increase of mutation rate with allele size up to an array length of 40-60 repeats above which mutation rate reaches a plateau. While there is no obvious polarity of intra-allelic mutation, mutation events are not randomly distributed throughout alleles but occur at greater frequency in regions of homogeneous repeats (Buard *et al.*, 1998), reminiscent of the elevated mutation rate observed at homogeneous microsatellites. However, CEB1 has a highly heterogeneous population of variant repeats despite having the highest mutation rate of any minisatellite characterised to date, suggesting that mutation processes operating at CEB1 and at microsatellites are fundamentally different (Jeffreys *et al.*, 1999). Inter-allelic mutations also occur at a lower frequency at CEB1 and, similar to MS32, are polarised towards one end of the repeat array.

Complex intra- and inter-allelic mutations at minisatellites are germline specific and are probably of meiotic origin (Jeffreys *et al.*, 1999). Many models have been proposed to describe the process of recombination in eukaryotes (reviewed by Osman and Subramani (1998)). It has not been possible to develop a model of recombination which accounts for all rearrangements detected at human minisatellites. One model, developed to account for the complex gene conversion events which are frequently accompanied by target site duplications detected at human minisatellites, is presented in Figure 1.2.

The development of SESP-PCR allowed minisatellite mutation events to be characterised in somatic tissues. The isolation of *de novo* mutants of minisatellite MS32 from blood DNA demonstrated that somatic mutation is both quantitatively and qualitatively different from germinal mutation. In contrast to germline mutation rates of  $\sim 10^{-2}$  per sperm, somatic mutants at MS32 can display mosaicism and arise at frequencies of  $1-2 \times 10^{-5}$  per molecule (Jeffreys and Neumann, 1997). Somatic mutants have characteristically simple structures, generated by the simple intra-allelic deletion or duplication of repeats occurring apparently anywhere within the repeat array, probably due to unequal crossing over between sister chromatids, although a process of polymerase slippage during replication cannot be

## Figure 1.2

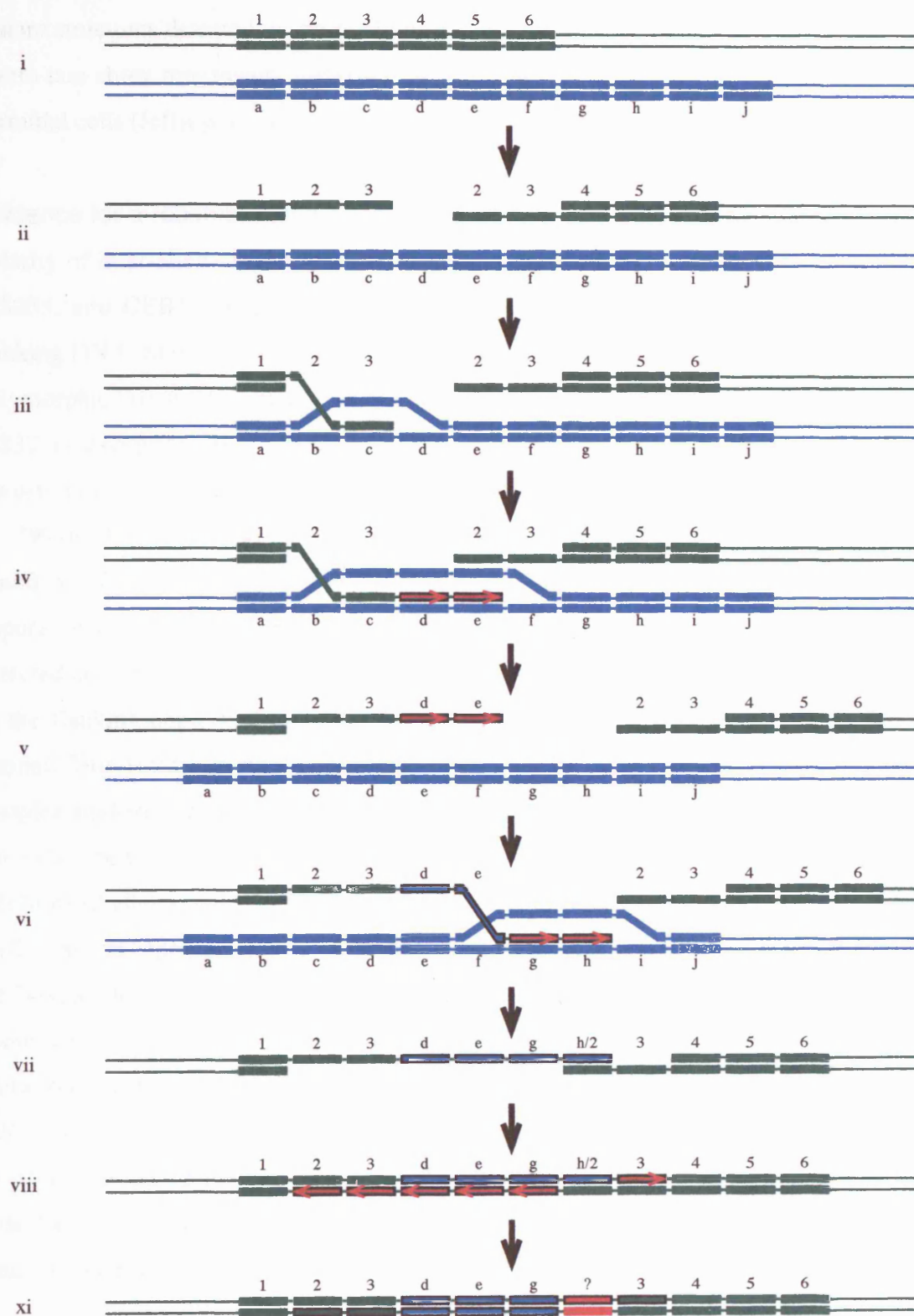
### ***A model for complex conversions at human minisatellites.***

The two progenitor alleles, and regions synthesised from each progenitor allele template, are shown in green and blue. Repeat elements are indicated by boxes and variant repeat identity shown by numbers (green progenitor allele) and letters (blue progenitor allele). DNA synthesis is indicated by red arrows, and repeats newly synthesised during the conversion process are highlighted by black borders. Red repeats indicate repeats of unknown identity formed from heteroduplex DNA repair.

- i) Two progenitor alleles are shown of 6 and 10 repeats. Alleles are paired within the 5' sequences flanking the minisatellite array.
- ii) A double strand break (DSB) is generated by the introduction of staggered nicks within the recipient allele.
- iii) The double strand break initiates recombination. The 3' end of the DSB from the recipient allele invades the donor allele creating a displacement loop.
- iv) The 3' end of the invading allele acts as a primer site for *de novo* synthesis using the donor allele as a template. This extends the displacement loop.
- v) The 3' extension is extruded leaving the donor allele unaltered.
- vi) Re-invasion of the same 3' region of the recipient allele occurs but out of register. A displacement loop is again formed and repeats are once more added to the recipient allele from the donor template by repair synthesis.
- vii) The 3' ends of the recipient allele re-anneal creating heteroduplex DNA between repeats h and 2.
- viii) Repair synthesis fills the gaps in the recipient allele.
- ix) The region of heteroduplex is repaired.

The mutant formed by this process has the structure of a complex in register inter-allelic conversion. The region inserted from the donor to recipient allele has a complex structure characterised by the apparent deletion of the f repeat, and possible generation of a novel repeat type by the process of heteroduplex repair. The inserted region is flanked by the duplication of repeat type 3. Figure 1.2 was adapted from Jeffreys *et al.* (1999).

**Figure 1.2**



excluded (Jeffreys and Neumann, 1997; May *et al.*, 1996). The highly complex mutations detected in sperm are not found in the soma. This qualitative difference, and the prevalence of germinal inter-allelic mutations, strongly suggest that germline mutation events occur during meiosis. A meiotic origin is further supported by the lack of mosaicism of complex mutant structures detected in sperm DNA. Rare large simple deletion mutations detected in sperm can show mosaicism, consistent with a low frequency of pre-meiotic mutation in germinal cells (Jeffreys *et al.*, 1990).

#### **Evidence for a recombination hotspot associated with MS32.**

Polarity of inter-allelic mutation events has been observed at minisatellites MS31, MS32, MS205, and CEB1, suggesting that mutations are initiated at sites located within the flanking DNA. Support for the existence of such a site at MS32 came from analysis of the polymorphic O1C/O1G variant located 48 bp 5' of the repeat array (Monckton *et al.*, 1994). MS32 is exceptionally variable in populations as a result of its high mutation rate. An estimated  $>10^8$  different alleles may exist in the current world population (Jeffreys *et al.*, 1991a). However, studies of allele diversity in African and Afro-Caribbean populations found a ~70-fold increase in homozygosity compared to Caucasian and Japanese populations. A major cause of this elevated homozygosity was an allele of 38 repeats detected at a frequency of ~10% in both Africans and Afro-Caribbeans. Sequence analysis of the flanking haplotype identified a base substitutional polymorphism called the O1C variant. This variant is rare in Caucasians (frequency ~0.004), absent from all Japanese samples studied (frequency <0.02) but present at frequencies of 10-20% in Africans and Afro-Caribbeans (Monckton *et al.*, 1994). Comparison of variability patterns of O1C- and O1G-linked alleles and of other polymorphic sites in the flanking DNA suggested that the O1C variant specifically associated with reduced variability. Mutation rate at O1C-associated alleles was analysed by SP-PCR in both O1C/O1G heterozygotes and O1C homozygotes and found to be suppressed in *cis*, supporting the existence of an initiator of mutation located within the 5' flanking region. Analysis of somatic mutations in O1C-associated alleles found mutations to be both quantitatively and qualitatively similar to O1G-associated alleles, providing further evidence that somatic and germline mutations arise by distinct mechanisms, and that the effects of the O1C variant were specific to meiosis (Jeffreys and Neumann, 1997).

Pedigree analysis at minisatellite MS31 identified a paternal mutant of the minisatellite which was apparently accompanied by the exchange of markers flanking the array, indicative of a true crossover event as opposed to gene conversion (Jeffreys *et al.*, 1998b). This was the first evidence for a link between minisatellite mutation processes and meiotic crossover. At MS32, nested allele-specific PCR on size-separated alleles using multiple heterozygous polymorphic sites in the DNA flanking the repeat array allowed the identification and isolation of the products of true inter-allelic crossover events. This led to the detailed analysis of the frequency and distribution of minisatellite-associated crossover events. True unequal crossovers were detected at a frequency of  $5 \times 10^{-5}$  per sperm, ~1% of the frequency of gene conversion events (Jeffreys *et al.*, 1998b). Unlike conversions, the majority of unequal crossovers were simple events involving the fusion of one allele with the other. These crossovers were not restricted to regions of sequence identity and could occur between highly (4%) diverged regions of repeats. This is in contrast to yeast where even low levels of sequence divergence inhibit crossovers (Borts and Haber, 1987). As with conversion mutation events, breakpoints of true unequal crossovers were polarised towards the 5' end of the array. The detection of isometric crossovers (crossover events which do not alter minisatellite array length) revealed similar mutant structures to unequal crossovers, with simple recombinant repeat arrays and breakpoints polarised towards the 5' end of the minisatellite array. These isometric crossovers occurred at 40% the frequency of unequal crossovers, far higher than would be expected if alleles paired at random sites within the repeat array, again indicating that alleles pair within the flanking DNA prior to both crossover and conversion (Jeffreys *et al.*, 1998b).

The polar distribution of both conversion and crossover events within the minisatellite suggested that both types of mutation process arose from a common mechanism. The existence of a conversion initiation site within the MS32 5' flanking DNA was indicated by the effects of the O1C/O1G variant, suggesting that the variant may affect the efficiency of a recombination hotspot located within the array flanking the minisatellite, and that this hotspot initiated a process (possibly by the induction of double strand breaks) which generated both conversion and crossover products. To identify the boundaries of this putative hotspot, crossovers were detected within the DNA flanking the minisatellite once more by nested allele-specific PCR and the position of recombination defined between

adjacent heterozygous polymorphic sites (Jeffreys *et al.*, 1998a). A high level of localised variation in recombination was detected, and an intense, highly localised hotspot with a maximum recombination frequency of 30-120 cM/Mb identified over a region of 1.5 kb. This compares with a genome average of ~1 cM/Mb. The peak was located ~200 bp 5' of the array extending 3' into the minisatellite and a short distance 5' of the hotspot. More distal sequences both 5' and 3' of the hotspot were recombinationally inert (<0.5 cM/Mb) (Jeffreys *et al.*, 1998a).

It was suggested that inter-allelic mutation at MS32 was driven by a recombination hotspot located within the 5' flanking DNA, and that the minisatellite may have been generated by the resulting recombinational activity. To further support this, recombination analysis of an O1C-linked MS32 allele found that the frequency of both equal and unequal crossover, the intensity of the recombination hotspot, and the frequency of conversions, was heavily suppressed (Jeffreys *et al.*, 1999; Jeffreys *et al.*, 1998a; Monckton *et al.*, 1994). Variation in hotspot activity between alleles (Jeffreys *et al.*, 1998a) may therefore result in the observed length-independent allele-specific differences in mutation rate. The hotspot can readily explain the polarised nature of mutation and variation at this locus. Furthermore, if the physical extent of the hotspot has remained constant and the bias towards length gain mutations observed in sperm is not countered by a bias towards deletion in the female germline, a picture is readily conceived of the minisatellite growing at the variable end and 'pushing' repeat units 3' outside of the hotspot thus accounting for the presence of tandem repeats within a recombinationally inert region 3' of the hotspot. The biological significance of areas of such localised recombination is unclear. It has been suggested that minisatellites mark chromosomal sites actively involved in homology searches required for meiotic chromosomal pairing (reviewed by Bois and Jeffreys (1999)). To support this, most human minisatellites (but not MS32) are clustered near the ends of chromosomes in regions that are involved in initiating chromosomal pairing (Royle *et al.*, 1988b).

The O1C variant indicates that the activity of recombination hotspots can be greatly affected by sequence variation. However, comparative sequence analysis of regions flanking human minisatellites failed to identify any significant sequence similarities between loci other than the prevalence of dispersed repeats (Armour *et al.*, 1989b),

suggesting that regions displaying intense recombination are not defined by a common primary DNA sequence (Murray *et al.*, 1999). Analysis of hotspots in yeast revealed that the majority are located within promoters, although some have been identified within coding sequences. It is thought that open chromatin conformation and, in some cases, the binding of transcription factors, determine hotspot activity. High resolution mapping in yeast has also demonstrated that double strand breaks which initiate recombination do not occur at specific sequences, but are distributed across a 50-200 bp region at a given locus (reviewed by Smith and Nicolas (1998)). In contrast to yeast, recombinational activity in humans was not found to correlate with associated gene promoters (Murray *et al.*, 1999). The factors which determine hotspot identity in humans have yet to be elucidated.

## **Mutation induction.**

Radiation and other natural and synthetic components of the environment have the potential to induce mutations. The most obvious mechanism of mutation induction is where an external factor directly causes damage to DNA. Minisatellites provide excellent systems for the detection of induced mutation for a number of reasons. The higher the spontaneous mutation rate of a locus, the smaller the sample size required to detect a significant shift in mutation rate induced by external mutagens. DNA fingerprint analysis allows the simultaneous detection of multiple minisatellite loci, and can be applied to species in which minisatellite loci are poorly characterised or uncharacterised. Minisatellites are also found on all mammalian autosomes (Jeffreys *et al.*, 1991b), so provide genome-wide indicators of mutation induction. In addition, the majority of minisatellite mutations are unlikely to be subject to selection. Detection of mutation events by pedigree analysis is therefore largely unbiased by the detrimental phenotypic consequences that result from many single base or chromosomal mutations.

The explosion at the Chernobyl nuclear power station in 1986 led to the release of  $\sim 5 \times 10^7$  Ci of assorted radionuclides. The immediate health risk was from the release of  $1 \times 10^7$  Ci of short-lived  $^{131}\text{I}$  which has led to a substantial increase in the incidence of thyroid cancer (Kazakov *et al.*, 1992; Likhtarev *et al.*, 1995). Longer term exposure was mainly due to contamination of the soil and food chain by  $^{90}\text{Sr}$ . A study using DNA fingerprinting techniques to analyse heritable mutation rates of minisatellites in pedigrees



from the Mogilev district of Belarus 300 km north of Chernobyl, identified a 2-fold higher mutation rate in this population compared to UK-based controls (Dubrova *et al.*, 1996; Dubrova *et al.*, 1997). Whilst this study has been criticised on the basis of potential population mismatching (e.g. Yauk (1998)) it provided evidence that exposure to radiation can induce heritable mutations in human populations. Further evidence came from dose-response analysis within the Belarus families which divided families by their levels of exposure to localised radiation, and found that families more heavily exposed to radioisotopes displayed significantly higher mutation rates than less exposed families (Dubrova *et al.*, 1996). Similar results were obtained from studies of the barn swallow *Hirundo rustica* breeding close to the Chernobyl site. Segregation analysis of two hypervariable microsatellite loci revealed a two- to ten-fold excess of mutation events in exposed swallows compared to control populations from the Ukraine and Italy (Ellegren *et al.*, 1997). In contrast, studies in humans looking for changes in inherited minisatellite mutation frequency in Japan (Kodaira *et al.*, 1995; Satoh and Kodaira, 1996) and amongst children whose fathers cleaned up after the Chernobyl nuclear power accident (Livshits *et al.*, 1999) failed to find significant evidence for radiation-induced minisatellite mutation. Furthermore, recent studies analysing germline mutation rates at human minisatellites CEB1 and B6.7 in the sperm of three seminoma patients following hemipelvic radiotherapy also found no evidence for mutation induction (C. May, *et al.*, manuscript in preparation).

Mutation induction in mammals can be studied more readily by using the mouse as a model organism. Typically, irradiated males are mated with non-irradiated females and rates of *de novo* mutation, detected by DNA fingerprint analysis, scored in the progeny (e.g. Dubrova *et al.* (1993)). Comparison of this rate with the mutation rate of non-irradiated males from the same strain allows mutation induction to be investigated in a more controlled environment. Studies of acutely irradiated mice found that they were extremely sensitive to  $\gamma$  irradiation, with a two-fold increase in mutation rate arising from exposure to just 0.33 Gy (the doubling dose) (Dubrova *et al.*, 1998a). This level of sensitivity is far too high for each mutation event to be induced by direct DNA damage and indicates that radiation activates an unidentified *trans*-acting factor which elevates mutation rate (Dubrova *et al.*, 1998a). Studies exposing males to X-irradiation at various time points prior to mating indicated that mutation induction occurs at premeiotic stages of

spermatogenesis (Dubrova *et al.*, 1998a). It is unknown whether the mechanism of mutation induction in mice are similar to those operating in humans as the loci analysed in mice were ESTRs as opposed to true minisatellites. The mechanism underlying spontaneous ESTR mutation is unclear, although polymerase slippage during replication is a strong candidate. Therefore the validity of using a mouse model to investigate the process of mutation induction in humans must be questionable.

The effects of chemical pollution on germline mutation rate have also been investigated. In a study of heritable mutations induced in herring gulls living near steel mills around the N. American Great Lakes, an elevated mutation rate was again detected by DNA fingerprint analysis for gulls living near the mills compared to gulls inhabiting less polluted regions (Yauk, 1998; Yauk and Quinn, 1996). Furthermore, mutation rate was found to correlate inversely with the distance of nesting site from the steel mills (C. Yauk *et al.*, manuscript in preparation). The steel mills release a wide array of pollutants, and so the identity of which chemical(s) induce this elevated mutation rate is unclear. The future use of mouse models may facilitate the identification of which chemicals induce mutation.

## Overview of this thesis.

Allele diversity and mutation dynamics have been extensively characterised at the minisatellites MS32, MS205, CEB1, and B6.7 (Buard *et al.*, 1998; Jeffreys and Neumann, 1997; Jeffreys *et al.*, 1994; May *et al.*, 1996; Tamaki *et al.*, 1999b). These loci may be unrepresentative of mammalian minisatellites in general for two reasons. First, they display unusually high germline mutation rates of  $>10^{-3}$  per generation. The majority (~90%) of human minisatellites display lower allele length heterozygosities (typically 70-80%), implying a mutation rate of perhaps  $3-5 \times 10^{-5}$  per gamete (Armour *et al.*, 1990; Nakamura *et al.*, 1987). Secondly, each of these loci are found in the human genome. Little is known about minisatellites or the mechanisms of minisatellite mutation that operate in other mammals.

The work described in chapters 3-8 investigates minisatellites in mice, and human minisatellites with estimated mutation rates of  $<10^{-3}$ . The primary objective to these studies was to determine whether the mechanisms of minisatellite mutation characterised at hypermutable human minisatellites also drives instability at less variable human minisatellites and at mouse minisatellites.

### ***Do mice have minisatellites similar to those analysed in humans?***

To further our understanding of mouse minisatellites, a systematic screen for mouse VNTRs is described in Chapter 3 to ascertain whether mouse counterparts of human minisatellites do exist. Little is known about the prevalence, genomic distribution, sequence composition, or variability of minisatellites in mice, and each of these points are addressed. Two loci identified in this screen displayed characteristics consistent with a form of dynamic instability similar to those detected at human triplet repeat expansions or the FRA16B AT-rich minisatellite. To gain further insight into the potential mechanisms underlying variation at these loci, variant repeat distribution within alleles was determined by MVR-PCR, and the repeat unit sequences analysed *in silico* for their potential to form secondary conformations, as described in Chapter 4.

In humans, individual minisatellites have been used as tools for phylogenetic analysis (Armour *et al.*, 1996a). To assess whether mouse minisatellites may be used in a similar capacity, five loci characterised in Chapters 3 and 4, and in Bois *et al.* (1998a) are analysed in Chapter 5, both individually and in combination to compare and contrast allele structures between different strains, subspecies, and species of mice.

The mouse ESTR *Hm-1* displays high levels of germline instability and has been extensively used to investigate the effects of radiation and other mutagens on mutation rate. However, little is known of the biological mechanisms underlying mutation at this locus. Some preliminary studies aimed at the further understanding of *Hm-1* mutation dynamics are described in Chapter 6.

### ***Do similar mutation processes operate at all human minisatellites?***

The majority of human minisatellites display lower levels of variability than any of the human loci analysed to date for *de novo* mutation events. The mutation mechanisms characterised at minisatellites with high ( $>10^{-3}$ ) germline mutation rates may be qualitatively different from those operating at less polymorphic loci. Alternatively, the same processes of tandem repeat turnover may affect all human minisatellites but vary between loci in their frequencies. The work described in Chapters 7 and 8 seeks to determine which of these hypotheses are correct. The first stage of this project was to identify loci suitable for mutation analysis. Three loci, described in Chapter 7, were analysed for their sequence composition and levels of allele length polymorphism. Variant repeat distribution at two of the loci was characterised by MVR-PCR to determine levels of allele variation in a CEPH cohort, from which inferences could be made concerning potential mutational mechanisms operating at each locus.

Ultimately, *de novo* mutation detection and analysis was not performed at any of the loci described in Chapter 7. Instead, the insulin minisatellite was selected for mutation analysis for two reasons. The first was that it has been intensively investigated due to its associations with type 1 diabetes (Bennett and Todd, 1996a), type 2 diabetes (Ong *et al.*, 1999), polycystic ovarian syndrome (Waterworth *et al.*, 1997), adult obesity (O'Dell *et al.*, 1999), and infant birth size (Dunger *et al.*, 1998), yet nothing was known of the mutational

mechanisms operating at the locus. The second reason was that a systematic survey of variant repeat distribution within alleles from a Caucasian cohort, described in this thesis, had been undertaken at this locus, which would allow a comparison to be made between patterns of allele variation and patterns of mutation detected at the minisatellite. Allele diversity and mutation analysis at the insulin minisatellite are described in Chapter 8.

### ***Does the insulin minisatellite predispose to type 1 diabetes?***

The final section of this thesis addresses a very different question. As reviewed in Chapter 9, classification of alleles of the insulin minisatellite by size and by flanking haplotype led to its identification as the type 1 diabetes susceptibility locus *IDDM2* (Bennett and Todd, 1996a). However allele size, at least at some loci (Armour *et al.*, 1996a), is relatively uninformative in identifying related minisatellite alleles. Variant repeat distribution was therefore analysed in 876 alleles from the parents of 219 families of type 1 diabetes affected sib pairs to investigate associations between the internal structures of minisatellite alleles and type 1 diabetes predisposition. The results of this work are presented in Chapter 10.

Due to the diversity of questions addressed within this thesis, no overall discussion is presented.

## **Chapter 2**

### **Materials and methods**

#### **Materials**

##### ***Chemical reagents***

Chemicals were obtained from Fisher Scientific (Loughborough, UK), Fisons (Loughborough, UK), Flowgen (Stafford, UK), FMC Bioproducts (Rockland, USA), Serva (Heidelberg, Germany), and Sigma Biochemical Company (Poole, UK). Molecular biology reagents were obtained from Advanced Biotechnologies (Leatherland, UK), Amersham International Plc. (Little Chalfont, UK), Bio-Rad (Hemel Hempstead, UK), Boehringer Mannheim (Lewes, UK), Clontech (Palo Alto, USA), Gibco-BRL (Paisley, UK), ICN Biochemicals Ltd. (High Wycombe, UK), Invitrogen (Leeke, The Netherlands), National Diagnostics (Hull, UK), NEN Life Sciences (Hounslow, UK), New England Biolabs (Hitchin, UK), Pharmacia (Milton Keynes, UK), Qiagen Ltd. (Dorking, UK), and Sigma Biochemical Company (Poole, UK). Specialised equipment was obtained from Advanced Biotechnologies (Surrey, UK), Bio-Rad (Hemel Hempstead, UK), Cecil Instruments (Cambridge, UK), Clare Chemical Research (Ross on Wye, UK), Corning (Maine, USA), Genetic Research Instrumentation (Dunmow, UK), Eppendorf Scientific Inc. (New York, USA), Heraeus Instruments (Hanau, Germany), Hybaid (Teddington, UK), MJ Research (Waltham, USA), Nalge Nunc International (Hereford, UK), New Brunswick Scientific Co. (Edison, USA), Perkin-Elmer/Applied Biosystems (Beaconsfield, UK), Shandon Southern (Runcorn, UK), and UVP Life Sciences (Cambridge, UK).

##### ***Oligonucleotides***

Oligonucleotides for PCR amplification were synthesised in-house (Dr. K. Lilley, Protein and Nucleic Acids Laboratory, University of Leicester, UK). Hexadeoxyribonucleotides for random primed labelling were obtained from Pharmacia.

## **Enzymes**

Restriction enzymes were supplied by Gibco-BRL, New England Biolabs, and Boehringer Mannheim. T4 ligase, *Pfu* polymerase, calf intestinal alkali phosphatase (CIAP) and REact™ buffers were obtained from Gibco-BRL. The Klenow fragment of DNA polymerase I of *E. coli* was obtained from Pharmacia. *Taq* polymerase was supplied by Advanced Biotechnologies. RNase and Proteinase K were supplied by Sigma.

## **Molecular weight markers**

1 kb ladder was supplied by Gibco-BRL.

## **Bacterial strains**

*Escherichia coli* strain Top10F', supplied by Invitrogen, was used in all cloning experiments. Top10F' is a recombination negative strain designed for stable replication of high-copy number plasmids. The F' episome, which carries the tetracycline resistance gene, allows isolation of single-stranded DNA from vectors that have an f1 origin of replication. The genotype of Top10F' allows blue-white screening of recombinant vectors by  $\alpha$ -complementation of  $\beta$ -galactosidase. In addition, Top10F' cells are *recA1* to increase the stability of inserts, *hsdR* to allow cloning without host restriction, and *endA1* to improve the quality of purified plasmid DNA.

Top10F' genotype:

F' { *labI*<sup>q</sup>, Tn10(Tet<sup>R</sup>) } *mcrA*  $\Delta$ (*mrr-hsdRMS-mcrBC*)  $\phi$ 80 *lacZ* $\Delta$ M15  $\Delta$ *lacX74* *deoR* *recA1* *araD139*  $\Delta$ (*ara-leu*)7697 *galU* *galK* *rpsL* (Str<sup>R</sup>) *endA1* *nupG*

## **Cloning vector**

DNA fragments were cloned into the *EcoRV* polylinker site of the pBluescript II SK<sup>+</sup> vector, supplied by Stratagene.

### **Human VNTR clones**

Human VNTR clones were selected from the Human Genome Mapping Project Resource Centre (HGMP-RC) probe bank, funded by the Medical Research Council, and supplied by the HGMP-RC (Cambridge, UK).

### **Mouse genomic library**

The commercial mouse genomic library was supplied by Clontech. The library was generated by the ligation of DNA fragments derived from the liver of an adult male BALB/c strain mouse into the pWE15 (Amp<sup>R</sup>) cosmid vector. Genomic DNA was mechanically sheared and *Bam*HI linkers added prior to cloning. The host strain for the library was NM554 (Genotype *recA13 araD139 Δ(ara-leu)7696 Δ(lac)l7A galU galK hsdR rpsL (Str<sup>R</sup>) mcrA mcrB*). The technique used to screen the library for VNTR loci was designed, developed, and patented by D. Gauguier, G Vergnaud, and J. Buard (Patents No. 93.12.923-France, and No. 5.573.912-USA). The full methodology is described by Amarger *et al.* (1998), with modified methodology described in Chapter 3.

### **Mouse DNA**

Inbred mouse DNA stocks were supplied care of M. Festing by A. Tomlinson from Harlan UK Ltd. Hillcrest (Loughborough, UK) or courtesy of the Division of Biomedical Services, University of Leicester. BXD recombinant inbred strains of mice were supplied by the Jackson Laboratory (Maine, USA). DNA from wild strains of mice was obtained from F. Bonhomme and A. Orth (Laboratoire Génome et Population, Université de Montpellier 2, Montpellier, France). Information about the strains of wild mice used in this thesis (BIK, MGT, 22MO, DOT, DBY, MAM, DJO, BZO, DMZ, DDO, MGL, CIM, MPR, DGA, BNC, BID, MPB, MBK, CTA, BFM, MBS, MDH, and BIR) is available elsewhere (Bonhomme and Guénet, 1996).

### **Human DNAs**

Genomic DNA from EBV-transformed lymphoblastoid cell lines was supplied by the Centre d'Etude du Polymorphisme Humain (CEPH, Paris, France). DNA from the



population of type 1 diabetic affected sib pair families was derived from EBV-transformed lymphoblastoid cell lines from the British Diabetic Association-Warren 1 Repository held at Porton Down (Wiltshire, UK), and supplied courtesy of Prof. John Todd (Cambridge Institute for Medical Research, Cambridge, UK). Sample collection and processing was funded by the British Diabetic Association and Wellcome Trust. Clinical criteria for the diagnosis of diabetes were as defined by Bain *et al.* (1992). Sperm DNA was derived from semen samples provided by Caucasian donors based in Leicester, UK.

### **Standard solutions**

Southern blot solutions (depurinating solution, denaturing solution, and neutralising solution), 20xSodium Chloride-Sodium Citrate (SSC) buffer, 10xTris-borate/EDTA (TBE) electrophoresis buffer, Luria-Bertani broth (LB) and Luria-Bertani agar (LUA) were as described by Sambrook *et al.* (1989), and were supplied by the media kitchen, Department of Genetics, University of Leicester. 11.1xPCR buffer was supplied by R. Neumann.

### **Computers**

This thesis was produced using a Power Macintosh G3 Minitower and a UMAX Vista-S6E scanner. It was printed on an HP4000 6MP LaserJet printer. DNA sequences were analysed by an IRIX Mainframe computer, operating the Genetics Computer Group (GCG) Sequence Analysis Software Package version 10.0 programs, developed at the University of Wisconsin (Devereux *et al.*, 1984). Data were stored, analysed, and presented using the software packages Adobe Acrobat, Adobe Photoshop, Autoassembler, Clarisdraw, EndNote, Fatura, Freehand, Microsoft Word, Microsoft Excel, and Microsoft Powerpoint all for Macintosh computers. Internet searches were performed using Microsoft Internet Explorer, the Sherlock search engine, and Netscape Navigator.

The accompanying web site to be found at <http://www.le.ac.uk/genetics/ajj> was produced using Adobe Acrobat and Dreamweaver software. All computer analyses were performed on Apple Macintosh computers.

## Methods

Due to the collaborative nature of much of the work in this thesis, detailed methods are only described for the techniques which I used. Techniques that I did not perform, but which were used in the projects presented, are either briefly summarised or referenced. The methods described in this section are general overviews of techniques. Where appropriate, detailed descriptions of specific applications of each technique are presented in the results chapters.

### ***1: Selection and growth of bacterial cultures***

All bacterial manipulations were performed under sterile conditions to minimise contamination of cultures.

#### **a: Growth of the mouse genomic DNA cosmid library**

A 100 µl aliquot of mouse cosmid genomic library at x100000 dilution was spread on Luria-Bertani Agar (LUA) containing ampicillin at 80 µg/ml and incubated for 16 hr at 37°C (all antibiotics were supplied by Sigma Biochemical company). Individual clones were transferred to 96-well microtitre plate arrays (Nunc, Nalge Nunc International) containing 100 µl per well of Luria-Bertani (LB) broth and Hogness Modified Freezing medium (HMFM; 44% glycerol with 5 mM Sodium Citrate, 3.6 mM  $\text{KH}_2\text{PO}_4$ , 1.3 mM  $\text{K}_2\text{HPO}_4$ , and 1 mM  $\text{MgSO}_4$ ) with 80 µg/ml ampicillin, covered with Mylar sealing film (Du Pont) and incubated for 16 hr at 37°C. Plates were frozen on dry ice and stored at -80°C. Aliquots of cultures were transferred individually to sterile test tubes containing 2.5 ml of freshly prepared Terrific Broth (TB) (Tartof and Hobbs, 1987) with 80 µg/ml ampicillin. Cultures were incubated at 37°C for 22-24 hr in a "Controlled Environment Incubator Shaker" (New Brunswick Scientific Co.) at 300 rpm. Cosmid cultures were pooled into four 50 ml Costar centrifugation tubes (Corning). The four pools corresponded to rows A/B, C/D, E/F, and G/H of the 96-well microtitre array. Where a single colony of interest had been identified, it was grown in 50 ml TB with 80 µg/ml ampicillin and incubated as described above.

#### **b: Growth of plasmid cultures**

Colonies containing plasmids were grown in LB broth with appropriate antibiotics.

### **2: Agarose gel electrophoresis**

#### **a: Electrophoresis conditions**

Unless stated otherwise, agarose gel electrophoresis was carried out using 1% LE (SeaKem™) agarose gels in horizontal submarine format with 0.5xTBE (44.5 mM Tris-borate pH 8.3, 1 mM EDTA) buffer containing 0.5 µg/ml ethidium bromide. Electrophoresis tanks were manufactured in-house and power packs were supplied by Bio-Rad and Shandon Southern. DNA was visualised using a UV wand (Chromato-vue UVM-57, UVP Life Sciences) or a UV transilluminator (UVP High Performance transilluminator, UVP Life Sciences), or for band excision using the Dark Reader System (Clare Chemical Research).

#### **b: Gel photography**

Photographic records of ethidium bromide-stained gels were generated by visualisation of the products on a UV transilluminator (UVP High Performance transilluminator) and photography using a Polaroid MP-4 camera with Kodak negative film (T-Professional 4052) or a Mitsubishi video copy processor (Genetic Research Instrumentation) with camera (UVP Life Sciences). Negatives and autoradiographs were developed using reagents recommended by Kodak.

### **3: DNA transfer to membrane**

#### **a: Southern blot**

Following electrophoresis, the region of agarose gel required for DNA transfer was excised and inverted into distilled water. The gel was depurinated in 0.25 M HCl for 2x5 min (depurinated DNA is cleaved more readily by NaOH), alkali-denatured in 0.5 M NaOH, 1 M NaCl for 8 min and 20 min (to cleave DNA into smaller fragments), and neutralised in

0.5 M Tris-HCl pH 7.5, 3 M NaCl for 8 min and 10 min. DNA was transferred to Hybond<sup>TM</sup>-N<sup>fp</sup> (Amersham International Plc.) nylon membrane (pre-soaked in 10xSSC) by the capillary transfer method using 20xSSC as the transfer buffer (Southern blotting (Southern, 1975)). Blotting was continued for 1-8 hr, depending on the agarose gel concentration. The membrane was washed in 2xSSC, dried at 80°C for 15 min, and the DNA covalently linked to the membrane by exposure to  $7 \times 10^4$  J/cm<sup>2</sup> of UV light in the RPN 2500 ultraviolet crosslinker (Amersham).

#### **b: Dot blot**

3 µl aliquots of DNA were pipetted onto dry Hybond<sup>TM</sup>-N<sup>fp</sup> membrane and dried at 80°C for 15 min. DNA was crosslinked to the membrane as described above.

#### **c: Preparation of 96-well plate replicate arrays**

Hybond<sup>TM</sup>-N<sup>fp</sup> of dimensions 110 mm x 90 mm was placed on the surface of LUA plates with 80 µg/ml ampicillin. Cell cultures were transferred to the membrane from cultures arrayed in 96-well plates using a sterile 'hedgehog' and incubated for 16 hr at 37°C. Cells were lysed by transferring membranes to 3MM Whatman chromatography paper soaked in 5% SDS, 2xSSC for 2 min. DNA from the cultures was fixed to the membrane by microwaving membranes at 850 W for 150 seconds. Membranes were vigorously washed in 0.1% SDS, 20xSSC to remove cell debris, rinsed in 2xSSC, and dried for 15 min at 80°C. The DNA was covalently linked to the membrane as described above.

### **4: Hybridisation**

Double stranded DNA (5-10 ng) generated (unless stated otherwise) by PCR amplification of the locus of interest was added to 10 ng of 1 kb ladder (Gibco) and labelled by the random primed labelling reaction (Feinberg and Vogelstein, 1983; Feinberg and Vogelstein, 1984) which involves the use of randomly generated hexamers and the *E. coli* DNA polymerase Klenow fragment to incorporate  $\alpha$ -<sup>32</sup>P-dCTP (supplied by Amersham International Plc.) into the DNA. Labelling reactions were performed in 30 µl reaction volume, and incubated at 37°C for 1-18 hr. The probe was recovered from unincorporated

deoxyribonucleotides by ethanol precipitation using 100 µg high molecular weight herring sperm DNA (Fluka, Sigma-Aldrich) as a carrier. Probes were dissolved in 0.5 ml distilled water and denatured by boiling for 3 min immediately prior to use. Membranes were pre-hybridised for at least 20 min at 65°C in 7% SDS, 0.5 M Na<sub>2</sub>PO<sub>4</sub> pH 7.2, 1 mM EDTA (modified from Church and Gilbert (1984)). Hybridisation was carried out at 65°C for 2-20 hr in a Maxi 14, or Mini 10 hybridisation oven (Hybaid). After hybridisation, the membrane was washed at 65°C in 3-5 changes of high stringency wash solution (0.1xSSC, 0.01% SDS). The pattern of hybridisation was visualised by autoradiography using Fuji RX100 X-ray film, either at room temperature for strong signals, or at -80°C with an intensifying screen.

Fluorescent in situ hybridisation was performed by Jill Williamson (Human Cytogenetics Laboratory, Imperial Cancer Research Fund, London). The technique is described elsewhere (Bois *et al.*, 1998a).

## **5: Methods of DNA extraction and purification**

### **a: Ethanol precipitation**

Unless stated otherwise, double-strand DNA was precipitated in microcentrifuge tubes by addition of 1/10 volumes 2 M NaAc pH 5.5, and 2 volumes 100% ethanol followed by incubation on ice for 10-30 min and centrifugation for 15 min at 15000 rpm in a Heraeus Septatech Biofuge 15 centrifuge. The pellet was washed in 80% ethanol and allowed to air dry. Single-strand DNA was precipitated in microcentrifuge tubes by addition of 1/10 volumes 2 M NaAc pH 5.5, and 2.5 volumes 100% ethanol followed by incubation at -80°C for 30 min and centrifugation for 30 min at 15000 rpm in a Heraeus Septatech Biofuge 15 centrifuge. The pellet was washed in 80% ethanol and allowed to air dry.

### **b: DNA extraction from agarose gels**

#### **i: Home-made spin column purification**

Spin column DNA extraction is as described by Heery *et al.* (1990). Spin columns were prepared by removing lids from 0.5 ml microcentrifuge tubes, punching a hole in the base

of the tube using a 1.1 mm x 40 mm hypodermic needle, and packing the base with synthetic filter wool ("SUPA" Aquatic supplies Ltd., Sheffield, UK) to a depth of 4-6 mm. Columns were placed inside carrier tubes (1.5 ml microcentrifuge tubes with lids removed) and rinsed by addition of 70 µl TE (10 mM Tris-HCl pH 7.5, 1 mM EDTA) and centrifugation at 6000 rpm for 6 min in a Heraeus Septatech Biofuge 15. Columns were transferred to fresh carrier tubes. Agarose bands excised from gels were transferred to the spin columns and centrifuged as before. DNA from within the eluate was used for subsequent reactions

## **ii: Electroelution**

The band was excised from the agarose gel and transferred to a slot cut within a second gel, slightly wider than the excised fragment. A piece of dialysis membrane was prepared by boiling for 10 min in TE and inserted into the gel slot curled under, and folded over the excised band. The gel was run at 4 V/cm allowing the DNA to electroelute onto the membrane. Electroelution was monitored using a UV wand. With continuous application of the current, the membrane was smoothly removed from the gel and placed into a microcentrifuge tube with a corner of the membrane trapped in the lid. Droplets of buffer containing the DNA fragment of interest were collected from the dialysis membrane by centrifugation at 15000 rpm for 30 seconds in a Heraeus Septatech Biofuge 15. DNA was recovered from the eluate by ethanol precipitation

## **iii: Freeze-thaw extraction**

The excised band was transferred to a microcentrifuge tube and frozen at -20°C. The band was allowed to partially thaw, and physically fragmented using a pipette tip. 50 µl of dilution buffer (5 mM Tris-HCl pH 7.5, 5 µg/ml carrier herring sperm DNA) was added, and the sample snap-frozen on dry ice/IMS. The sample was allowed to thaw in a 37°C water bath and vortexed for 10-20 seconds, followed by centrifugation at 15000 rpm for 2 min in a Heraeus Septatech Biofuge 15. This procedure was repeated 3 times. DNA in the solution was used for subsequent analysis.

#### **iv: Qiaex II kit purification**

DNA was recovered from excised bands using the Qiaex II purification kit (Qiagen) in accordance with the manufacturer's instructions.

#### **c: 'Maxiprep' alkali lysis extraction of cosmid DNA**

50 ml cultures of bacterial cells containing cosmid DNA were transferred to 50 ml Costar tubes (Corning) and centrifuged at 43000 rpm at 4°C in a Heraeus Septatech Megafuge 1.0R for 30 min and the supernatant removed. The pellet was thoroughly resuspended in 10 ml ice-cold P1 solution (50 mM Tris-HCl, pH 8.0, 10 mM EDTA) supplemented with 0.1 mg/ml RNase. Cells were lysed by addition of 10 ml of P2 solution (1% SDS, 200 mM NaOH) and incubation at room temperature for 5 min. Lysis was terminated and cell debris precipitated by addition of 10 ml ice-cold P3 solution (KAc and glacial acetic acid to a concentration of 3 M K<sup>+</sup>, 5 M Ac<sup>-</sup>, pH 5.5) and incubation on ice for 20 min. Debris was compacted by centrifugation at 43000 rpm at 4°C in a Heraeus Septatech Megafuge 1.0R for 10 min. Supernatant was removed and filtered through a mesh. DNA was purified via Maxi-Qiagen columns (Qiagen) according to manufacturer's instructions and precipitated by addition of 0.7 volumes of isopropanol followed by centrifugation at 43000 rpm at 4°C in a Heraeus Septatech Megafuge 1.0R for 30 min. The pellet was washed in 80% ethanol, redissolved in 200 µl TE, and transferred to a microcentrifuge tube.

#### **d: 'Miniprep' alkali lysis extraction of cosmid and plasmid DNA**

Alkali lysis of bacterial cultures was as described above, scaled down to the required cell culture volume. DNA was purified by addition of 1/2 volumes of a 24:23:1 mixture of phenol: chloroform: isoamyl alcohol, followed by vortexing, brief centrifugation, and removal of the aqueous phase. Phenol/chloroform extraction was repeated 2-3 times. DNA was ethanol precipitated as described above.

#### **e: DNA extraction from mouse tails**

Tail DNA from wild and inbred mice was prepared as described elsewhere (Allen *et al.*, 1987).

#### **f: DNA extraction from human sperm**

DNA extraction from sperm was performed in a category II laminar flow hood under conditions designed to minimise contamination, and in the absence of sharp objects to reduce the risk of infection. Frozen semen samples were thawed on ice for 1 hr. 250 µl of semen was transferred to a 1.5 ml screw top tube and diluted in 900 µl 1xSSC. Somatic cells were lysed by addition of 100 µl 1% SDS with immediate mixing and centrifugation. Supernatant was removed and lysis in 0.9xSSC, 0.1% SDS repeated three times. The pellet was washed in 1 ml 1xSSC and resuspended in 450 µl 0.2xSSC. Sperm heads were lysed by addition of 1% SDS and 1 M 2-mercaptoethanol and incubation at room temperature for 5 min. Proteinase K was added to a concentration of 200 µg/ml and incubated for 1 hr at 37°C with occasional mixing. Proteins were removed by phenol/chloroform extraction with gentle mixing for 5-10 min on a vertical rotor to allow emulsification. Extraction was repeated twice on both bulk and interphase. DNA was ethanol precipitated as described above, resuspended in distilled water, and re-precipitated. The pellet was air dried and resuspended for 16 hr at 4°C in 50 µl 5 mM Tris-HCl pH 7.5.

#### **6: *Enzymatic manipulation of DNA***

Enzymatic manipulation of DNA was carried out in the reaction buffer supplied with the enzyme according to the conditions recommended by the supplier, unless stated otherwise.



## 7: DNA amplification

### a: PCR buffer

11.1xPCR buffer (Jeffreys *et al.*, 1990) was produced in Leicester by R. Neumann with the following components. dNTPs and BSA were supplied by Pharmacia.

Component	Concentration of Stock Solution	Volume (arbitrary units)	Final Concentration in PCR Reaction
Tris-HCl pH 8.8	2 M	167	45 mM
Ammonium Sulphate	1 M	83	11 mM
MgCl <sub>2</sub>	1 M	33.5	4.5 mM
2-mercaptoethanol	100%	3.6	6.7 mM
EDTA pH 8.0	10 mM	3.4	4.4 $\mu$ M
dATP	100 mM	75	1 mM
dCTP	100 mM	75	1 mM
dGTP	100 mM	75	1 mM
dTTP	100 mM	75	1 mM
BSA	10 mg/ml	85	113 $\mu$ g/ml
Total Volume		676	

### b: PCR conditions

#### i: General PCR

DNA was amplified using the Polymerase Chain Reaction (PCR) (Saiki *et al.*, 1988) on a PTC-225 DNA Engine Tetrad Peltier thermal cycler with heated lid (MJ Research). PCR reactions were performed, unless stated otherwise, in 10  $\mu$ l reactions with 0.9  $\mu$ l of 11.1xPCR buffer as described above, supplemented with 12 mM Tris base, 1  $\mu$ g/ml carrier herring sperm DNA, plus 0.4  $\mu$ M of each primer, 0.07 U/ $\mu$ l *Taq* polymerase and 0.007 U/ $\mu$ l *Pfu* polymerase. *Pfu* removes base mismatches, allowing *Taq* to resume extension of the new strand. Additional Tris increases the pH of the reaction, reducing the risk of template depurination at high temperatures. To minimise contamination, precautions were taken to ensure that the reagents and materials used in the PCR reaction were kept separate from general laboratory chemicals, and PCR reactions were set up in a category II laminar flow

hood. Details of specific PCR thermal cycling conditions are included in relevant chapters. Primer sequences are listed in Table 2.1.

## **ii: MVR-PCR**

Minisatellite variant repeat mapping by PCR (MVR-PCR) uses variant repeats within minisatellite loci to generate internal maps of minisatellite alleles by a simple PCR assay (Figure 1.1; Jeffreys *et al.* (1991a)). MVR-PCR reactions were performed, unless stated otherwise, in 7  $\mu$ l reactions with 1  $\mu$ l DNA from previously separated alleles at a concentration of 0.1-1.0 pg/ $\mu$ l, 0.63  $\mu$ l of 11.1xPCR buffer supplemented with 12 mM Tris base, 1  $\mu$ g/ml carrier herring sperm DNA, plus 0.25  $\mu$ M of TAG and flanking primer, 0.035 U/ $\mu$ l *Taq* polymerase and 0.0035 U/ $\mu$ l *Pfu* polymerase. Unless stated otherwise, repeat-specific MVR primers were at concentrations of 10 nM.

## **8: Size-enrichment small pool PCR at the insulin minisatellite**

Genomic DNA was digested to completion with *Hinf*I. Size enrichment and small pool PCR were performed as described elsewhere (Jeffreys and Neumann, 1997).

## **9: Automated DNA sequence analysis**

Sequencing was carried out using a PE Applied Biosystems Model 377 DNA Sequencing System, with the ABI PRISM BigDye™ Terminator Cycle Sequencing Ready Reaction Kit, in accordance with the manufacturers instructions. Sequencing reactions were cycled at 96°C for 10s, 60°C for 4 min for 25 cycles using ~20 ng/kb of DNA template. Reactions were purified by ethanol precipitation and dissolved in 2  $\mu$ l 83% de-ionised formamide, 8.3 mM EDTA, prior to loading onto the sequencing gel. Gel running and analysis was carried out in the Protein and Nucleic Acid Laboratory (PNACL), University of Leicester.

Sequence data generated on mouse minisatellite loci has been deposited with the GenBank/EMBL Data Libraries under Accession Nos. AJ002239-57 and AJ002934-39.

## Table 2.1

### ***List of primers***

Primers are listed in 5'-3' orientation. Regions of MVR-PCR primers with sequence identity to TAG are presented in lower case. Primers 51-A, 51-B, MINS-C, and TAG were designed by A. Jeffreys. Hm1pA, HMA, and HMB were designed by R. Kelly. KS and SK were commercially available primers, obtained from Stratagene. With the exception of primers KS and SK, all oligonucleotides were synthesised in-house (Dr. K. Lilley, Protein and Nucleic Acids Laboratory, University of Leicester, UK). Prior to use, primers were ethanol precipitated and dissolved in distilled water. The concentration was determined measuring  $A_{260}$  using a Cecil Instruments 2040 UV spectrophotometer, and adjusted to a working stock concentration of 10  $\mu$ M.

### List of primers

Chapter 2 Table 2.1

## **10: Subcloning techniques**

### **a: Digestion and dephosphorylation of vector**

20 µg of uncut pBluescript II SK<sup>+</sup> vector was digested with 50 units of *EcoRV* in REact™ 2 buffer for 2 hr at 37°C, and dephosphorylated by the addition of 50 units of calf intestinal alkali phosphatase (CIAP, Gibco) with incubation for 30 min at 37°C. The reaction was stopped by addition of SDS and EDTA to concentrations of 0.5% and 5 mM respectively, followed by incubation at 65°C for 20 min for enzyme denaturation. The vector was ethanol precipitated, washed in 80% ethanol, and redissolved in 50 µl TE. Concentration of the stock solution was assayed by measuring A<sub>260</sub> using a Cecil Instruments 2040 UV spectrophotometer, and adjusted to 100 µg/ml.

### **b: End-filling and ligation**

The cosmid fragment for subcloning was isolated by digestion of 6 µg of cosmid DNA with appropriate enzymes, agarose gel electrophoresis and excision of the band of interest, generating ~200 ng of the fragment for cloning. The band was gel purified by electroelution, ethanol precipitated and redissolved in 5 µl of distilled water. For purposes of simplicity, all ligations were designed for blunt ended fragments. Where necessary, fragments for ligation were end-filled by addition of 5 mM of each dNTP with 2 units of Klenow in a 20 µl reaction volume with REact1™ buffer (Gibco-BRL) and incubation at 37°C for 30 min. End-filled fragments were purified by gel electrophoresis, electroelution, and precipitation as described above.

Ligations were performed in a 10 µl reaction volume containing 7.5 ng of *EcoRV*-digested and dephosphorylated pBluescript II SK<sup>+</sup>, 75 ng of insert DNA, 1 unit T4 DNA ligase and T4 DNA ligase buffer (Gibco-BRL). Reactions were incubated for 16 hr at 16°C in a PTC-225 DNA Engine Tetrad Peltier thermal cycler (MJ Research). Ligation products were recovered by ethanol precipitation and redissolved in 10 µl of distilled water.

#### c: Preparation of electrocompetent Top10F' cells

A single colony of Top10F' cells, grown at 37°C on LUA supplemented with 25 µg/ml tetracycline, was transferred to 5 ml of LB with 25 µg/ml tetracycline and incubated for 16 hr at 37°C with shaking at 300 rpm. The culture was transferred to 600 ml of LB with 25 µg/ml tetracycline and incubated for 4-6 hr as above until cells reached a density of  $OD_{600}=0.5-0.6$ . Unless stated otherwise, all further stages of electrocompetent cell preparation were performed at 4°C in a temperature controlled room using equipment pre-cooled to 4°C. Cells were chilled for 15 min in an iced water bath then transferred to centrifugation bottles and centrifuged at 4000 rpm at 2°C for 20 min in a Sorvall RC-5B centrifuge (Du Pont Instrumentation) using a pre-cooled GS-3 rotor. The pellet was gently resuspended in 5 ml ice cold distilled water followed by addition of a further 200 ml ice cold water and gentle mixing. Cells were centrifuged as before, and the wash repeated 3 times. The pellet was resuspended in 25 ml 10% glycerol and transferred to 50 ml Costar tubes (Corning). Cells were pelleted by centrifugation at 4000 rpm at 2°C for 10 min in a Heraeus Sepatech Megafuge 1.0R and resuspended in 0.5 ml 10% glycerol. The suspension was transferred in 40 µl aliquots to individual tubes pre-chilled to -80°C, immediately frozen on dry ice/IMS, and stored at -80°C. Cells were used for transformations within 3 months of preparation.

#### d: Electroporation and selection of transformants

4 µl of ligation products were added to a 40 µl aliquot of electrocompetent cells thawed on ice and transferred to an electroporation cuvette (Bio-Rad) cooled to 4°C. Cells were transformed using a Bio-Rad Gene Pulser™ electroporation unit at 1.5 kV with capacitance of 25 mF. 1 ml of ice cold LB was immediately added and the cell suspension transferred to a test tube. Isopropyl thio-b-D-galactoside (IPTG, Sigma) was added to a concentration of 1 mM and cells incubated in a shaker for 45 min at 37°C and 300 rpm. Cells were diluted x20 and x100 in LB with IPTG, and 200 µl of culture plated onto LUA supplemented with 80 µg/ml ampicillin, 25 µg/ml tetracycline, and previously spread with 100 µl of 20 mg/ml X-gal solution (Sigma). Plates were incubated for 16 hr at 37°C. For each ligation, 5 white colonies were selected, cultured in 2.5 ml LB, and DNA extracted by alkali lysis. Inserts were excised by digestion with *Bss*HIII, and electrophoresed to check for correct insert size.

The identity of the insert was confirmed by Southern blot hybridisation using the probes which originally identified the locus to be of interest (see Chapter 3).

#### e: Ligation reaction and electroporation controls

For each newly prepared stock of either pBluescript II SK<sup>+</sup> vector or electrocompetent cells, the following control electroporations were performed.

Control	Vector Present	Vector Digested	Vector Dephosphorylated	Ligated	Control Insert
1	×	×	×	×	×
2	✓	×	×	×	×
3	✓	✓	×	×	×
4	✓	✓	×	✓	×
5	✓	✓	✓	✓	×
6	✓	✓	✓	✓	✓

#### Expected results from controls:

- 1) No cells should be detected due to the absence of vector, and therefore the Amp<sup>R</sup> gene.
- 2) A large number of blue colonies should be detected due to the presence of both antibiotic resistance genes and an undisturbed LacZ operon.
- 3) Linearised vector should not be viable so no colonies should be present. Any colonies can be assumed to be the result of incomplete digestion.
- 4) When ligase is added, the cut vector should re-ligate giving results similar to control 2.
- 5) The dephosphorylated vector should not re-ligate. Any blue colonies present are indicative of either incomplete digestion or incomplete dephosphorylation.
- 6) The 1.35 kb band from the  $\phi$ x *Hae*III ladder was selected as the control insert fragment. Most colonies should be white due to LacZ disruption by the insert DNA. Blue colonies should occur at similar frequency to control 5.

## Chapter 3

# Isolation and characterisation of mouse minisatellites

### Summary

Minisatellites provide the most informative system for analysing processes of tandem repeat turnover in humans. However, little is known about minisatellites and the mechanisms by which they mutate in other species. Furthermore, a mouse model of minisatellite mutation would allow the detailed investigation of mechanisms of minisatellite mutation. To this end, 77 endogenous mouse VNTRs were isolated and characterised. A correlation was identified between allele variability and mean array length. Fifty-one loci have been localised on mouse chromosomes and, unlike in humans, show no clustering in proterminal regions. One minisatellite was identified within intron 16 of the mouse hairless gene. Sequence analysis of 28 loci revealed the majority to be authentic minisatellites with GC-rich repeat units ranging from 14 to 47 bp in length. In contrast to humans, there was no evidence for the existence of mouse loci with mutation rates  $\geq 10^{-3}$ . The mouse is therefore not a good model organism for the analysis of mechanisms of minisatellite mutation.

### Introduction

Attempts to isolate and characterise endogenous mouse minisatellites held two long-term objectives. The first was to characterise mutation processes operating at mouse minisatellites. Whilst minisatellite mutation has been intensively studied at a number of human loci (Jeffreys *et al.*, 1999), it was unknown whether similar processes of tandem repeat turnover operate in other organisms. Mutation analysis in the mouse would allow comparisons to be made between mutation dynamics in two different mammalian species. The second objective was to generate a mouse model to facilitate the detailed investigation of mechanisms of minisatellite repeat turnover. If mouse minisatellites mutate by similar mechanisms and at similar rates to human loci, a mouse model would permit the



investigation of generic mechanisms of mutation using methodologies which, for either technical or ethical reasons, cannot be used in humans.

### ***Potential applications of a mouse model of minisatellite instability***

#### **Effects of genomic context on instability**

Levels of heterozygosity may affect minisatellite mutation. The recombinational nature of minisatellite mutation in humans (Jeffreys *et al.*, 1999) raises the possibility that complex mutation events are triggered in heterozygotes either by mismatches in allele length or in heteroduplexes arising from strand invasions between alleles that are thought to be involved in generating recombination initiation complexes (Nassif *et al.*, 1994; Paques *et al.*, 1998). In contrast, even low levels of heterozygosity greatly suppress local recombination in yeast (Borts and Haber, 1987; Borts *et al.*, 1990). Whilst the degree of sequence mismatch between alleles does not appear to suppress recombination at human minisatellite MS32 (Jeffreys *et al.*, 1998b) or within the mouse major histocompatibility complex (Yoshino *et al.*, 1995), heterozygosity may nevertheless affect mouse minisatellite mutation processes. Selective breeding would allow the effects on mutation of both heterozygosity at a specific locus and genome-wide levels of heterozygosity to be systematically investigated. The genomic location of minisatellites may also effect mutation. The creation of transgenic strains with unstable mouse loci inserted at ectopic locations would permit position effects to be analysed. Mutation analysis in mutant strains of mice, such as mice with defects in genes involved in double-strand break repair (e.g. Essers *et al.* (1997)), could also facilitate the identification of components of the mutation pathway.

#### **Determination of mutation rate variation**

At human minisatellite MS32, inter-allelic mutation rate is likely to be determined by the intensity of the flanking recombination hotspot (Jeffreys *et al.*, 1998a), whilst at CEB1 intra-allelic mutation rate is a function of array length and repeat homogeneity (Buard *et al.*, 1998). Mouse models may further clarify the basis of mutation rate heterogeneity between alleles, between individuals, and between loci. Inter-allelic mutation rate variation may result from *cis*-acting factors either within the minisatellite array or the flanking DNA.

*Cis*-acting sites could be defined by creating transgenic strains with deletions of repeats or of flanking sequences and characterising their effects on mutation. Alternatively, by replacing repeat arrays of specific minisatellites by homologous recombination with either single copy DNA or repeat arrays from other tandem repeat loci, the relative contributions of repetitive DNA and flanking DNA to instability could be elucidated. Genes which encode putative *trans*-acting factors that affect mutation rate could theoretically be identified through classical linkage analysis. Hybridisation of two strains which display substantially different levels of genome-wide minisatellite instability may allow the identification of genetic components of the mutation pathway by segregation analysis, using the phenotype of minisatellite instability as a marker.

### Characterising mutation processes from a range of tissues

In humans, the detection of female germline mutation is restricted to pedigree analysis. A mouse model would allow the isolation of oocytes, greatly increasing the number of potential germline mutants available from a single female mouse. Furthermore, the availability of tissues such as testes would facilitate the biochemical identification of components of the mutation initiation and processing complex, whilst separation of cells at various stages of gametogenesis would further elucidate the timing of mutation events.

### Mechanisms of mutation induction

The effects of exposure to radiation and other mutagens have been discussed (Chapter 1). The mouse loci previously analysed for induced mutation were ESTRs as opposed to true minisatellites (Bois *et al.*, 1998b; Dubrova *et al.*, 1993; Dubrova *et al.*, 1998a). The mechanisms for either spontaneous or induced mutation operating at ESTRs are unknown. The large array lengths and lack of known variant repeats in the most unstable ESTRs have made the detailed analysis of patterns of repeat turnover intractable. The application of MVR-PCR to the analysis of spontaneous and induced mutation at true mouse minisatellites would allow more detailed investigations of the mechanisms of mutation induction. Furthermore, studies in humans of the effects of radiation on tandem repeat mutation have analysed true minisatellites (Dubrova *et al.*, 1996; Dubrova *et al.*, 1997). It is unknown whether the mechanisms of mutation induction operating at human minisatellites

and at mouse ESTRs are the same, or indeed whether radiation definitely induces minisatellite mutation in humans (Yauk, 1998). The analysis of induced instability at true mouse minisatellites would therefore provide a more appropriate model for characterising the effects of mutagens on humans.

### ***Characteristics of the ideal model locus***

The ideal mouse model for analysing minisatellite mutation may be defined by four criteria:

i) The minisatellite should be detected by a single-locus probe. If the minisatellite is a member of a multi-locus family, the identification of specific loci is not possible using genomic Southern blot hybridisation; ii) To easily isolate *de novo* mutation events, to detect differences in mutation rates between mice, and to identify statistically significant changes in mutation rate due to exposure to environmental mutagens using relatively small numbers of mice, a high spontaneous mutation rate is essential; iii) To detect *de novo* mutation events in both the germline and soma using SP-PCR, alleles must be amplifiable by PCR so ideally would be less than ~4 kb in length; iv) The characterisation of allele variation and the structures of mutants requires the establishment of systems of MVR-PCR, and so the model locus should display patterns of variant repeat dispersion throughout alleles.

Two approaches to establish a mouse model of minisatellite mutation have been attempted; the identification of endogenous mouse minisatellites (as described in this chapter), and the generation of mice transgenic for human hypermutable minisatellites.

### ***Transgenic mouse models of minisatellite instability***

The first attempts to generate mice transgenic for human minisatellites used a short construct of an MS32 allele with a few hundred bases of flanking genomic DNA resulting in both single-copy and multi-copy integrants (Allen *et al.*, 1994; Collick *et al.*, 1994). Multi-copy integrants did display instability, but this was due to the palindromic nature of inserted multi-copy transgenes, as opposed to being a direct consequence of the minisatellite itself (Collick *et al.*, 1996). Single-copy integrants could not be bred to homozygosity preventing the analysis of inter-allelic mutation processes, but the

hemizygous strains displayed no signs of germline instability in pedigrees (Allen *et al.*, 1994).

Further transgenic lines were created using 29 kb constructs of DNA (Bois *et al.*, 1997) which included the MS32 recombination hotspot later characterised in humans (Jeffreys *et al.*, 1998a). Somatic instability was detected for both single- and multi-copy integrants with similar rates and patterns of mutation to those detected in the human soma (Bois *et al.*, 1997; Jeffreys and Neumann, 1997). However, no mutations were detected in the germline corresponding to a >600-fold suppression of mutation rate compared to humans. The total lack of mutations in sperm DNA indicated that the mouse germline was protected from mitotic instability and supported other work indicating that germline and somatic mutations in humans arise by distinct mechanisms (Jeffreys and Neumann, 1997). Furthermore, there appears to be a major barrier to the transfer of minisatellite germline instability from humans to mice, potentially mediated by incompatibilities between the species in any *cis*- or *trans*-acting factors involved in the mutation mechanism (Bois *et al.*, 1997). Alternative explanations for the lack of instability include insufficient flanking DNA within the transgenic construct, position effects suppressing mutation at the various integration sites, the failure to organise the chromatin conformation at the transgene, and either locus-specific or genome-wide homozygosity in transgenic lines inhibiting mutation initiation.

In contrast to minisatellites, mice transgenic for triplet repeat loci associated with human disease do in some cases display instability. Early attempts to transfer disease-associated (CAG)<sub>n</sub> repeats from cDNAs of the human androgen receptor gene and from the genes for HD, SCA1, and MJD/SCA3 displayed no evidence of instability (Bingham *et al.*, 1995; Burright *et al.*, 1995; Goldberg *et al.*, 1996a; Ikeda *et al.*, 1996), indicating that the mechanisms underlying tandem repeat instability in humans may not operate in mice. However, instability has since been reported in transgenic models of the (CAG)<sub>n</sub> repeats which cause HD, DM, and DRPLA (Gourdon *et al.*, 1997; Lia *et al.*, 1998; Mangiarini *et al.*, 1997; Monckton *et al.*, 1997; Sato *et al.*, 1999; Wheeler *et al.*, 1999). The reasons for the success of some models and the failure of others are unclear. The average length of STRs is greater in mice than humans (Beckman and Weber, 1992) indicating that mice may have a higher length threshold for instability. However, the size threshold effects cannot

account for all differences between stable and unstable constructs (Bates *et al.*, 1997). Other hypotheses include variation in *cis*- and *trans*-acting factors, variation in expression of the construct, and position effects due to the site of transgene integration (Bates *et al.*, 1997). Nevertheless, the presence of triplet repeat instability in at least some transgenic mice compared to the apparent total absence of germline mutation in mice transgenic for human minisatellites further indicates that germline mutation processes which operate at human minisatellites are fundamentally different from those operating at STR loci.

## Methods

A large-scale screen of the mouse genome for minisatellites is described in this chapter. In total, four approaches were employed to identify minisatellites in a collaborative project spanning many years. Whilst my own contribution was to apply a single strategy to isolate loci, the results of the entire project are here presented.

## Strategies for the isolation of minisatellites

### ***1: Synthetic tandem repeat probe cross-hybridisation to a cosmid library***

Four thousand bacterial cosmid clones from a commercial mouse genomic library were screened with synthetic tandem repeat (SyTR) probes 14C2, 16C2, and 16C24 (Mariat and Vergnaud, 1992; Vergnaud, 1989; Vergnaud *et al.*, 1991) known to detect polymorphic loci in rodent genomes (Mariat and Vergnaud, 1992). DNA was prepared from positively hybridising bacterial clones by alkaline SDS lysis (Birnboim and Doly, 1979) and digested with *AluI* (AGCT), *HaeIII* (GGCC), and *MseI* (TTAA) double-digest combinations. These three restriction enzymes were chosen for their different recognition sites to reduce any sequence bias. Tandemly repetitive DNA has a lower sequence complexity, and therefore a lower diversity of endonuclease restriction sites than non-repetitive DNA. Digestion therefore enriches for repetitive DNA. The largest digestion-resistant fragments were isolated using home-made spin columns (Heery *et al.*, 1990).

## **2: Cross-hybridisation to size-selected charomid libraries**

Two charomid libraries were constructed as described by Armour *et al.* (1990) using 3-14 kb size-selected fractions of *Mbo*I- (GATC) or *Tsp*509I- (AATT) digested genomic DNA. Minisatellites can be difficult to clone using standard vectors in *E. coli* hosts (Wong *et al.*, 1986; Wyman *et al.*, 1985) and are prone to gross rearrangements, commonly deletions of the repeat array (Brutlag *et al.*, 1977). Charomid vectors (Saito and Stark, 1986) are less sensitive to a reduction in insert size caused by repeat loss (Armour *et al.*, 1990). Two different enzymes were used to reduce any sequence bias during cloning, and therefore increase the number of clonable mouse tandem repeats. An estimated 1200 clones represented one haploid genome equivalent for these size selected fractions. Five thousand clones from each library were grown in 96-well ordered arrays and replicated onto filters as described in Chapter 2. Filters were hybridised with a series of multi-locus probes ( $\lambda$ 33.15,  $\lambda$ 33.6, MS1, and *Hm*-2) known to detect large numbers of polymorphic tandem repeat loci in mouse DNA (Jeffreys *et al.*, 1987b; Kelly *et al.*, 1989). Positively hybridising inserts were recovered using home-made spin columns.

## **3: Cross-hybridisation to size-selected $\lambda$ libraries**

MMS80, a locus previously named Mm1 (Jeffreys *et al.*, 1997), was isolated from a genomic library of 4-9 kb size-selected *Sau*3AI fragments cloned into the  $\lambda$ L47.1 vector (Loenen and Brammar, 1980), by hybridisation with probe  $\lambda$ 33.6 (Jeffreys *et al.*, 1985a; Kelly *et al.*, 1989; R. Kelly unpublished data).

## **4: Isolation of restriction endonuclease-resistant fragments from a cosmid library**

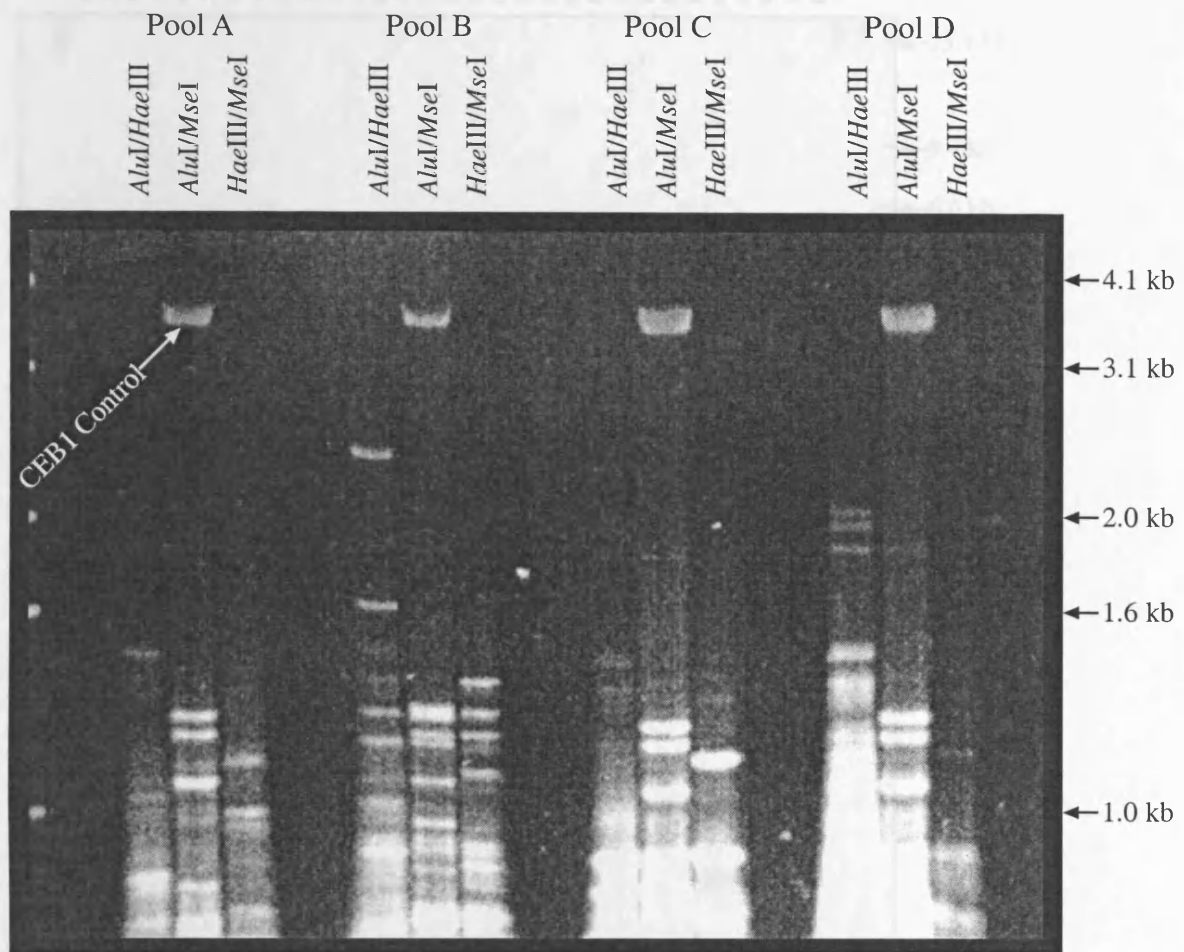
The strategy which I employed for the identification of mouse minisatellites was adapted from a procedure designed, developed, and patented by D. Gauguier, G. Vergnaud, and J. Buard (Patents No. 93.12.923-France, and No. 5.573.912-USA), and described by Amarger *et al.* (1998). Individual cosmids were selected at random from a mouse genomic cosmid library and cultured in a 96-well microtitre plate. The ordered cosmid library was replicated onto nylon filters as described in Chapter 2. Each colony was individually grown in 2.5 ml TB and pools of 24 cultures, each representing two rows of the microtitre array,

were pooled into 50 ml Costar tubes. Within each pool of 24 was included a CEB1-containing cosmid as an internal positive control. Growing cultures separately ensured that each cosmid within the pool was at a relatively uniform concentration. Cosmid DNA was extracted from each pool (Qiagen extraction, Chapter 2) and 15 µg of DNA digested with each *AluI*, *HaeIII*, and *MseI* double-digest combination (enzyme selection was as described above). Digestion-resistant fragments were separated by agarose gel electrophoresis (Figure 3.1) and isolated from the gel by home-made spin columns (Heery *et al.*, 1990). CEB1 is digested by *HaeIII* but not by *AluI* or *MseI* so provided an internal control for both the quality of DNA purification, and the efficiency of endonuclease digestion (Figure 3.1).

The isolated fragments were used as probes on *MboI*-digested genomic Southern blots of DNAs from six inbred strains of mice, and 23 laboratory stocks of mice recently derived from wild colonies (Figure 3.2). Descriptions of wild mice strains are provided in Table 3.1 All genomic Southern blots and hybridisations were performed by R. Neumann. Initially, probes were hybridised to Southern blots of genomic DNA from inbred strains of mice only. The majority of minisatellites identified were monomorphic between strains and so these blots were not informative as to the differences in the levels of variability between loci. Analysis of wild mice DNA greatly increased the observable levels of polymorphism between strains allowing variability to be compared between loci. VNTR loci were identified both by patterns of allele length variability between strains, and by the hybridisation signal intensity (probes for repetitive loci produce signals of greater intensity than probes for non-repetitive DNA). To identify the cosmid clone from which the VNTR probe was derived, the same probe was hybridised to the 96-well replicate of the ordered cosmid library (Figure 3.3a).

MMS10 is a multi-locus family of ESTRs (Bois *et al.*, 1998b) which was frequently isolated using this procedure. The ideal mouse minisatellite for analysis would be detected by a single-locus probe. To exclude MMS10 family members, all probes were initially transferred to membrane by dot blot and screened with an MMS10 probe prior to their hybridisation to genomic Southern blots (Figure 3.3b).

**Figure 3.1**



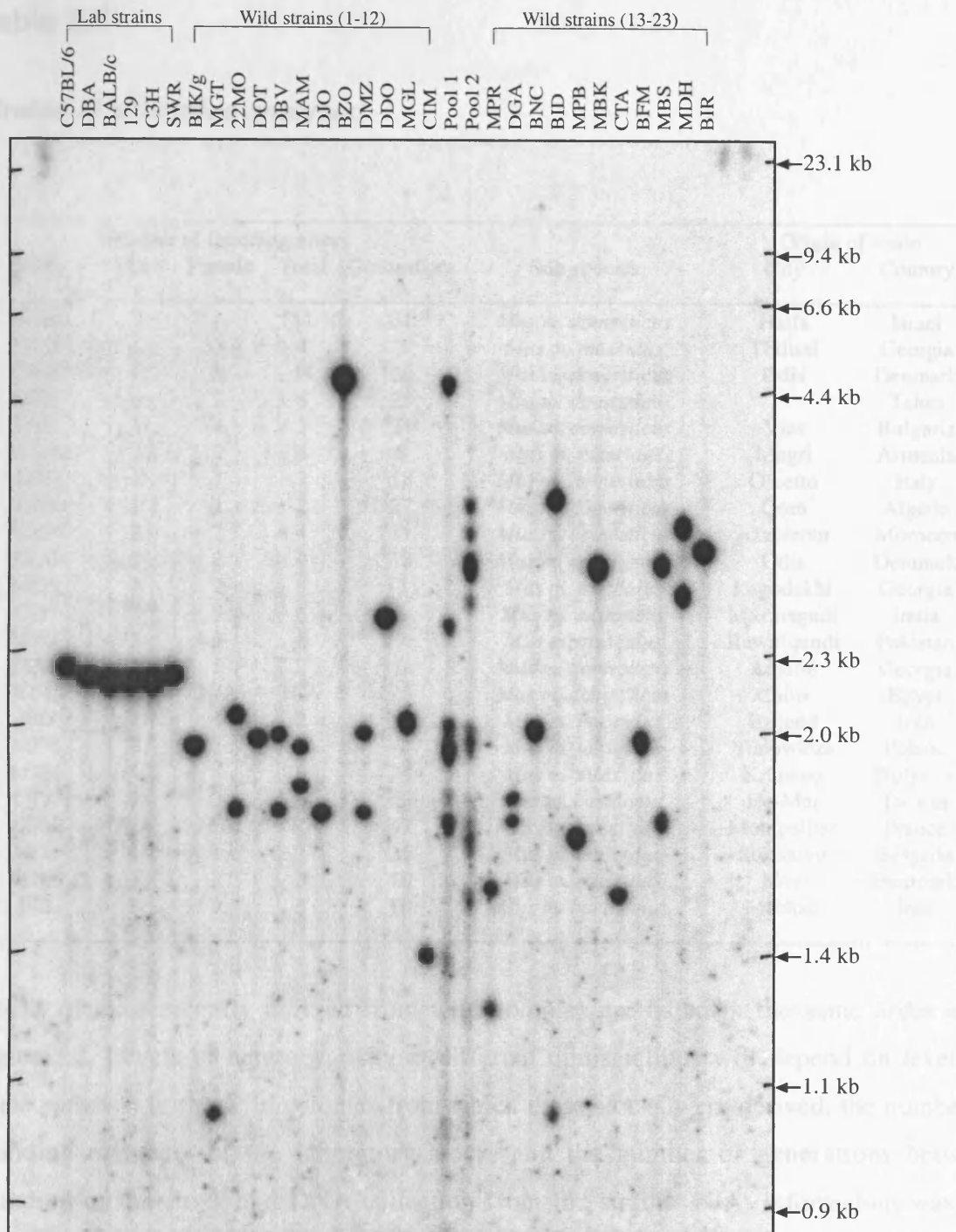
**Isolation of digestion-resistant fragments from 96 cosmid clones**

Each double-digest combination using enzymes *AluI*, *HaeIII*, and *MseI* was performed on 15 µg of DNA extracted from a pool of 24 cosmids. The four pools shown in Figure 3.1 represent a single 96-well ordered cosmid array, with for example pool A corresponding to cosmids within rows A and B of the 96-well plate. Within each pool, one cosmid containing the human minisatellite locus CEB1 was included as an internal control. CEB1 is resistant to digestion by *AluI* and *MseI*, but is cut by *HaeIII* resulting in a single band of ~3.5 kb visible in the *AluI*/*MseI* double-digest. Digestion-resistant fragments such as the 1.6 and 2.6 kb bands in pool B were extracted from the gel and used as probes on *MboI*-digested mouse genomic DNA Southern blots (Figure 3.2). Depending on the quality of DNA extraction and digestion, digestion-resistant fragments of as small as 0.9 kb could be isolated. Some of the bands below 1.3 kb in length were derived from the pWE15 cosmid vector so were present in each pool (e.g. compare *AluI*/*MseI* digest patterns between pools).

*Southern blots and hybridizations were performed by R. Neumann.*



**Figure 3.2**



***MMS64 allele variability in inbred and wild mice***

Probe cMMS64 was hybridised to *Mbo*I-digested genomic DNA from 6 inbred laboratory strains and 23 wild strains of mice. Pools of DNA from wild mice 1-12 (Pool 1) and 13-23 (Pool 2) were used to facilitate comparison of allele sizes between strains. Information about wild mice is presented in Table 3.1. No length polymorphism was observed between the 6 inbred strains. High levels of polymorphism (18 different alleles) and heterozygosity (10 heterozygotes) were observed amongst the 23 strains of wild mice. All *Mbo*I-digested Southern blots and hybridisations were performed by R. Neumann.

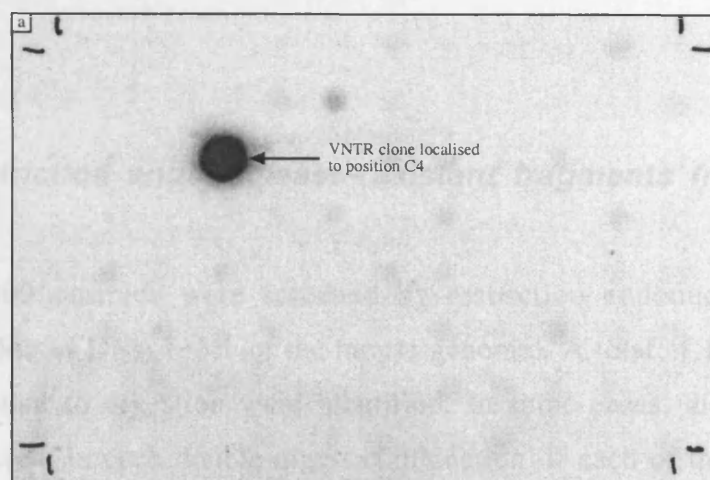
**Table 3.1*****Strains of wild mice analysed***

Strain	Number of founding mice			Generations	Sub-species	Origin of strain	
	Male	Female	Total			City	Country
BIK/g	5	6	11	31	<i>Mus m. domesticus</i>	Haifa	Israel
MGT	1	3	4	9	<i>Mus m. musculus</i>	Tbilissi	Georgia
22MO	5	9	14	26	<i>Mus m. domesticus</i>	Odis	Denmark
DOT	4	2	6	25	<i>Mus m. domesticus</i>		Tahiti
DBV	1	4	5	21	<i>Mus m. domesticus</i>	Vlas	Bulgaria
MAM	2	2	4	8	<i>Mus m. musculus</i>	Megri	Armenia
DJO	1	1	2	18	<i>Mus m. domesticus</i>	Orcetto	Italy
BZO	1	1	2	27	<i>Mus m. domesticus</i>	Oran	Algeria
DMZ	2	2	4	11	<i>Mus m. domesticus</i>	Azzemour	Morocco
DDO	2	2	4	18	<i>Mus m. domesticus</i>	Odis	Denmark
MGL	3	2	5	7	<i>Mus m. musculus</i>	Lagodekhi	Georgia
CIM	3	3	6	8	<i>Mus m. castaneus</i>	Masinagudi	India
MPR	3	3	6	9	<i>Mus m. musculus</i>	Rawalipindi	Pakistan
DGA	1	1	2	6	<i>Mus m. domesticus</i>	Adjarie	Georgia
BNC	N/A	N/A	N/A	29	<i>Mus m. domesticus</i>	Cairo	Egypt
BID	1	1	2	5	<i>Mus m. musculus</i>	Birjand	Iran
MPB	2	3	5	8	<i>Mus m. musculus</i>	Bialowieza	Poland
MBK	4	4	8	24	<i>Mus m. musculus</i>	Kranevo	Bulgaria
CTA	3	3	6	8	<i>Mus m. castaneus</i>	He-Mei	Taiwan
BFM	N/A	N/A	N/A	47	<i>Mus m. domesticus</i>	Montpellier	France
MBS	1	1	2	25	<i>Mus m. musculus</i>	Sokolovo	Bulgaria
MDH	3	2	5	10	<i>Mus m. musculus</i>	Hov	Denmark
BIR	0	3	3	10	<i>Mus m. bactrianus</i>	Machad	Iran

Stocks of mice recently derived from wild colonies are listed in the same order as in Figure 3.2. Levels of heterozygosity at different minisatellites will depend on levels of allele variation in the wild colonies from which these stocks were derived, the number of founding members of the laboratory stock, and the number of generations between founding of the stock and DNA collection from the strains. N/A: information was not available. Further information on these strains is available elsewhere (Bonhomme and Guénet, 1996).

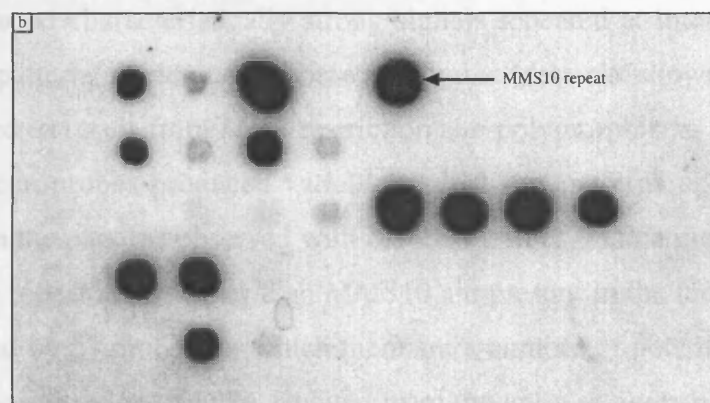
### Figure 3.3

#### **Hybridisation screens of ordered arrays**



a: Localisation of a VNTR clone on a 96-well plate replicate

MMS45 was identified as a VNTR locus by hybridisation to a genomic Southern blot as in Figure 3.2. The probe was generated by digestion of pool B of cosmid DNA derived from the 96-well plate, so was known to have originated from a cosmid clone contained within rows C and D of the ordered array. Hybridisation of the VNTR probe to a filter replicate of the plate (Chapter 2) identified the location of the clone (in this case location C4), from which a stock of DNA from a single cosmid could be generated. The membrane was marked at each corner to facilitate alignment of the autoradiograph with the 96-well plate.



b: Screen of probes for MMS10 loci

MMS10 is a multi-locus family of ESTRs present in the mouse genome with an estimated copy number of 1000-3000 (Bois *et al.*, 1998b). The screen for mouse minisatellites was aimed at the isolation of loci detected by single-locus probes, and so MMS10 loci were excluded from analysis at an early stage. Each probe derived from the restriction digestion-resistant assay was dot-blotted onto membrane (Chapter 2). Hybridisation with a known MMS10 probe identified any additional MMS10 loci, which were excluded from further analysis.

In the next section, a summary of my own results is initially presented, followed by a description of the results from the completed project.

## Results

### ***Isolation of restriction endonuclease-resistant fragments from a cosmid library***

Approximately 3000 cosmids were screened by restriction endonuclease digestion representing ~100 Mb of DNA (~3% of the mouse genome). A total of 188 fragments of sizes >0.9 kb resistant to digestion were identified. In some cases, digestion-resistant fragments were present in each double-digest combination. If each of the bands were of similar intensity, it was assumed that they were derived from a single locus and so only one band was selected for further analysis. The remaining 140 fragments were gel-purified using home-made spin columns (Heery *et al.*, 1990). Pre-screening of each DNA with MMS10 identified 13 positively hybridising loci. All other 127 fragments were radiolabelled with  $\alpha$ -<sup>32</sup>P-dCTP and hybridised to genomic Southern blots (Figures 3.2).

Minisatellites are only one of several classes of tandemly repetitive DNA which could be identified by this method. Satellite or midisatellite sequences were detected by 13 probes. These probes produced characteristically strong signals repeated at intervals of 0.5-1 kb with hybridisation patterns monomorphic between strains (data not shown). The observed pattern was assumed to result from *Mbo*I restriction site polymorphisms between satellite repeat units. Fourteen probes produced variable multi-locus patterns upon hybridisation which differed from the patterns observed with MMS10 probes, indicating that multi-locus families of variable repeat arrays other than MMS10 are present in the mouse genome. No signal was generated by 57 probes for which there are a number of potential explanations: i) The quality or concentration of DNA isolated from the gel was poor; ii) The probe may have formed DNA conformations which prevented efficient labelling or hybridisation; iii) The probe may be AT-rich and so labels poorly with  $\alpha$ -<sup>32</sup>P-dCTP; iv) The probe detects a VNTR which is digested by *Mbo*I so will not be detectable by Southern blot hybridisation to *Mbo*I-digested DNA; and v) The probe was single-copy DNA which was resistant to digestion. Monomorphic VNTRs would be distinguishable from monomorphic single copy

DNA due to the greater signal intensity generated by repetitive probes upon hybridisation. The remaining 43 probes identified 38 different VNTR loci, the majority of which displayed length polymorphism between strains of wild mice. The remaining chapter will consider the results of the entire screen for mouse VNTRs. In general, my own further experiments were specific to the 38 loci described above.

### ***Combined results of various approaches to identify mouse minisatellites***

Two main approaches have been described for the isolation of mouse minisatellites. First, three different mouse DNA libraries (cosmid, charomid and  $\lambda$ ) were screened with a range of tandem repeat probes known to detect multiple polymorphic mouse loci. The main limitation of this approach is that the number of loci detected is limited by the probes used, and will show an inherent sequence bias. The second cloning approach used the frequent-cutting endonucleases *AluI*, *HaeIII*, and *MseI*, to identify restriction digestion-resistant putative tandem repeat loci. This approach largely avoids the sequence bias inherent in cross-hybridisation strategies. However, limitations to this approach remain: i) Tandem repeat arrays within restriction fragments shorter than ~0.9 kb in the BALB/c mouse from which the cosmid library was derived will be undetectable against the background of cosmid vector restriction fragments (Figure 3.1); ii) Arrays containing sites either for *MboI* or for two of the three test enzymes (*AluI*, *HaeIII*, and *MseI*) will be lost; iii) The labelling and hybridisation conditions do not favour detection of AT-rich loci; and iv) Any sequences prone to collapse in cosmids will be under-represented or undetectable in an amplified cosmid library. The results of each screening approach are presented in Table 3.2.

A total of 77 different mouse VNTRs were identified (Table 3.3). Neither the results of the charomid screen nor the SyTR screen can be readily used to estimate the total number of VNTR loci present throughout the mouse genome due to the bias for larger VNTR arrays inherent in the size-selection strategy, and the sequence bias due to the hybridisation probes selected. Based on the restriction digestion screen and correcting for repeated isolations of the same locus, an estimated 700 loci are present throughout the mouse genome (Bois *et al.*, 1998a). This is an underestimate due to the limitations inherent in this cloning strategy described above.

**Table 3.2*****Comparison of screens for mouse VNTRs***

	Charomid	SyTR	Digestion	Total
Number of clones screened by hybridisation	5000	4000		9000
Number of clones selected for probe isolation	212	122	6700	7034
Number of probes isolated	132	34	251	418
Number of probes detecting:				
No signal	97	18	136	251
Satellites	6	0	32	38
MMS10 loci	4	10	22	36
Single locus VNTRs	25	6	61	92
Different single locus VNTRs	23	6	51	77

A total of 5000 charomid clones from a size-selected library (representing four haploid genome equivalents) were screened by hybridisation with tandem repeat probes. The SyTR (synthetic tandem repeat probe) screen of a commercial mouse library analysed 4000 cosmids representing ~5% of the mouse genome. In a joint project between P. Bois, S. Bakshi, and myself, 6700 cosmids from the same library (an estimated 235 Mb representing 12% of the mouse genome) were screened for fragments resistant to endonuclease digestion. For the charomid screen, a number of positive clones did not contain large fragments of mouse DNA, suggesting collapse of repeat arrays within the vector. The high proportion of MMS10 loci isolated using the SyTR screen was due to cross-hybridisation with the 14C2 SyTR probe under low stringency hybridisation conditions. A total of 77 different single locus probes for VNTR loci were identified.

### Table 3.3

#### ***Characteristics of cloned mouse VNTRs***

Where available, the chromosomal localisation is provided for each locus, together with the distance in cM from the centromere. *In situ* localisation data are presented in parentheses. The cloning strategies employed to isolate mouse VNTRs are listed as follows: Char. (cross-hybridisation to size-selected charomid library), SyTR (cross-hybridisation to cosmid library),  $\lambda$ L47.1 (cross-hybridisation to size-selected  $\lambda$ L47.1 library), and Dig. (digestion-resistance screen of cosmid library). The number of different alleles detected in inbred and wild mice is shown, with the number of heterozygotes (H) identified in wild mice. *Mbo*I-digested allele size range and mean size (including an unknown amount of flanking DNA) are given in kb. Some alleles were too small to be included within the size range of the Southern blots. In these cases, allele numbers are underestimated and are signified by \* and a  $\leq$  in the allele size range column. ND, not determined; these probes were either too small or contained internal *Mbo*I cleavage sites and could not therefore be hybridised to the mouse *Mbo*I-digested genomic Southern blots. <sup>a</sup> Cloned twice with the same methodology. <sup>b</sup> Cloned twice using two independent methodologies. <sup>c</sup> cMMS10 is a member of a novel family of unstable mouse VNTRs (Bois *et al.*, 1998b). <sup>d</sup> Cloned with all three methodologies. To clarify my own contribution, all VNTR loci isolated using the digestion-resistance screen of a cosmid library which were not identified by me are indicated by Dig.†.

**Table 3.3*****Characteristics of cloned mouse VNTRs***

Probe	Localisation		Cloning method	Number of alleles			Size range	Mean Size
	Chr	cM(FISH)		Inbred	Wild	H		
cMMS2 <sup>a,b</sup>	16	31	Char.+SyTR	4	12	5	1.3-6.6	3.0
cMMS3	7	29	Char.	2	4	1	≤2.0	1.5
cMMS4	ND		Char.	2	3	0	1.5-3.0	2.9
cMMS5	11	47	Char.	ND				
cMMS6	15	62	Char.	3	15*	5*	≤6.6	2.3
cMMS7	4	59	Char.	2	5	1	0.8-1.5	1.2
cMMS9	12	10	Char.	3	15	7	3.0-6.6	4.0
cMMS10 <sup>c</sup>			Char.					
cMMS11	11	31	Dig.†	3*	8*	1*	≤1.3	0.7
cMMS12 <sup>a</sup>	7	74	Char.	ND				
cMMS13	4	70	Char.	ND				
cMMS15	4	54	Char.	ND				
cMMS16	7	22	Char.	ND				
cMMS18 <sup>d</sup>	7	22	Char.+SyTR+ Dig.†	5	9	3	1.4-1.9	1.6
cMMS19	4	49	Char.	2	4	0	0.5-1.3	1.0
cMMS20	6	51	SyTR	ND				
cMMS21	2	217	SyTR	ND				
cMMS22	12	66	Char.	2	17	7	1.2-4.4	2.6
cMMS24 <sup>a</sup>	7	22	Dig.	4	11	6	1.4-4.5	2.5
cMMS25 <sup>a</sup>	15	23	Dig.†	2	12	5	1.0-10.0	3.0
cMMS26 <sup>a</sup>	9	68 (F3)	Dig.†	2	19	9	0.9-4.5	2.0
cMMS28	16	(A3-B1)	Dig.†	1	15	6	1.5-3.0	2.0
cMMS30	X	43	Dig.	3	8	1	1.0-2.5	1.8
cMMS34	4	(C1)	Dig.†	ND				
cMMS35 <sup>a</sup>	12	(E)	Dig.	1	17	8	1.2-4.5	2.8
cMMS36	2	(B)	Dig.†	ND				
cMMS37	16	(C3-4)	Dig.†	2	8	3	0.6-4.5	1.4
cMMS38 <sup>a</sup>	17	(C)	Dig.†	1*	7*	2*	≤1.4	0.6
cMMS39	14	(D3)	Dig.†	2	8	6	1.2-1.6	1.4
cMMS40	18	(D)	Dig.†	3	11	4	1.9-2.5	2.1
cMMS41	19	1 (A-B)	Dig.†	3	11	6	0.9-2.3	1.4
cMMS42	15	(E)	Dig.†	3	11	5	0.8-1.5	1.0
cMMS44	15	(F)	Dig.	2	9+	6*	≤1.3	0.8
cMMS45 <sup>a</sup>	5	(A2)	Dig.	3	14	1	0.8-2.5	1.5
cMMS46	8	(E2)	Dig.	2	12	7	0.6-2.5	1.5
cMMS47	1	(E1)	Dig.	2	8*	1*	≤2.4	1.1
cMMS48	4	59	Dig.	5	9	3	0.8-2.0	1.5
cMMS49 <sup>b</sup>	2	35	SyTR+Dig.	4	12	3	0.9-3.0	1.8



**Table 3.3 (continued)**

Probe	Localisation		Cloning method	Number of alleles			Size range	Mean Size
	Chr	cM(FISH)		Inbred	Wild	H		
cMMS52 <sup>a</sup>	2	(H3)	Dig.	1	10	6	0.9-4.0	1.9
cMMS53	2	(E8)	Dig.	1	3+	1*	≤0.9	0.5
cMMS54	ND		Dig.	1	17	9	0.9-5.0	2.1
cMMS55	ND		Dig.	3	12	5	1.3-3.0	1.9
cMMS56	1	(H5-6)	Dig.	2	6+	6*	≤2.3	1.5
cMMS57	8	(E2)	Dig.	2	16	7	1.0-11.0	2.4
cMMS58	3	(G)	Dig.	2	7	0	1.4-2.3	1.4
cMMS59	ND		Dig.	2	16	8	1.6-9.4	3.4
cMMS60	ND		Dig.	1	5	1	1.9-2.2	2.1
cMMS61	9	(A5)	Dig.	1	3	1	1.3-3.0	1.4
cMMS62	12	(F1)	Dig.	1	11	7	1.3-3.0	1.7
cMMS63	16	(B4-5)	Dig.	2	10	6	0.9-2.4	1.7
cMMS64 <sup>a</sup>	1	(D)	Dig.	2	18	10	0.5-4.0	2.2
cMMS65 <sup>a,b</sup>	19	2 (B)	Char.+Dig.	3	10	5	1.3-4.0	1.8
cMMS66 <sup>a</sup>	9	66	Dig.	2	11	6	1.4-4.0	2.3
cMMS67	5	21	Char.	2	6	3	2.0-2.3	2.2
cMMS69	5	51	Dig.	3	13	3	0.5-2.0	1.2
cMMS71	1	50	Dig.	2	12	5	1.5-10.0	2.6
cMMS73	ND		Dig.	4	16	10	1.3-4.0	2.9
cMMS74	ND		Char.	2	24	11	1.0-20.0	4.2
cMMS75	ND		Char.	1	8	3	4.0-6.6	5.0
cMMS76	ND		Char.	1	3	0	4.6-5.0	4.8
cMMS77	ND		Char.	1	15*	6*	≤10.0	2.9
cMMS78	ND		Dig.	3	16	4	1.1-2.5	1.4
cMMS80	9	79	λL47.1	3	18	8	1.4-6.6	2.4
cMMS82	ND		SyTR	2*	8*	4*	≤1.8	1.2
cMMS85	ND		Dig.	2	2	0	0.8-0.9	0.9
cMMS86	ND		Dig.	2	3	0	1.2-1.3	1.3
cMMS87	ND		Dig.	2	3	1	2.1-2.4	2.2
cMMS88	ND		Dig.	1	2	0	0.7-0.8	0.8
cMMS89	ND		Dig.	1	1	0	1.4	1.4
cMMS90	ND		Dig.	1	1	0	1.0	1.0
cMMS91	ND		Dig.	1	1	0	0.8	0.8
cMMS92	ND		Dig.	2	4	0	0.7-0.8	0.7
cMMS93	ND		Dig.	3	4	1	2.0-2.2	2.1
cMMS94	ND		Dig.†	2	14	7	0.9-6.0	2.0
cMMS95	ND		Dig.†	0*	4*	1*	≤1.1	1.0
cMMS96	ND		Char.	1	2	0	2.9-3.0	3.0
cMMS97	ND		Char.	2	2	0	2.7-2.8	2.8

### ***Analysis of VNTR length polymorphism***

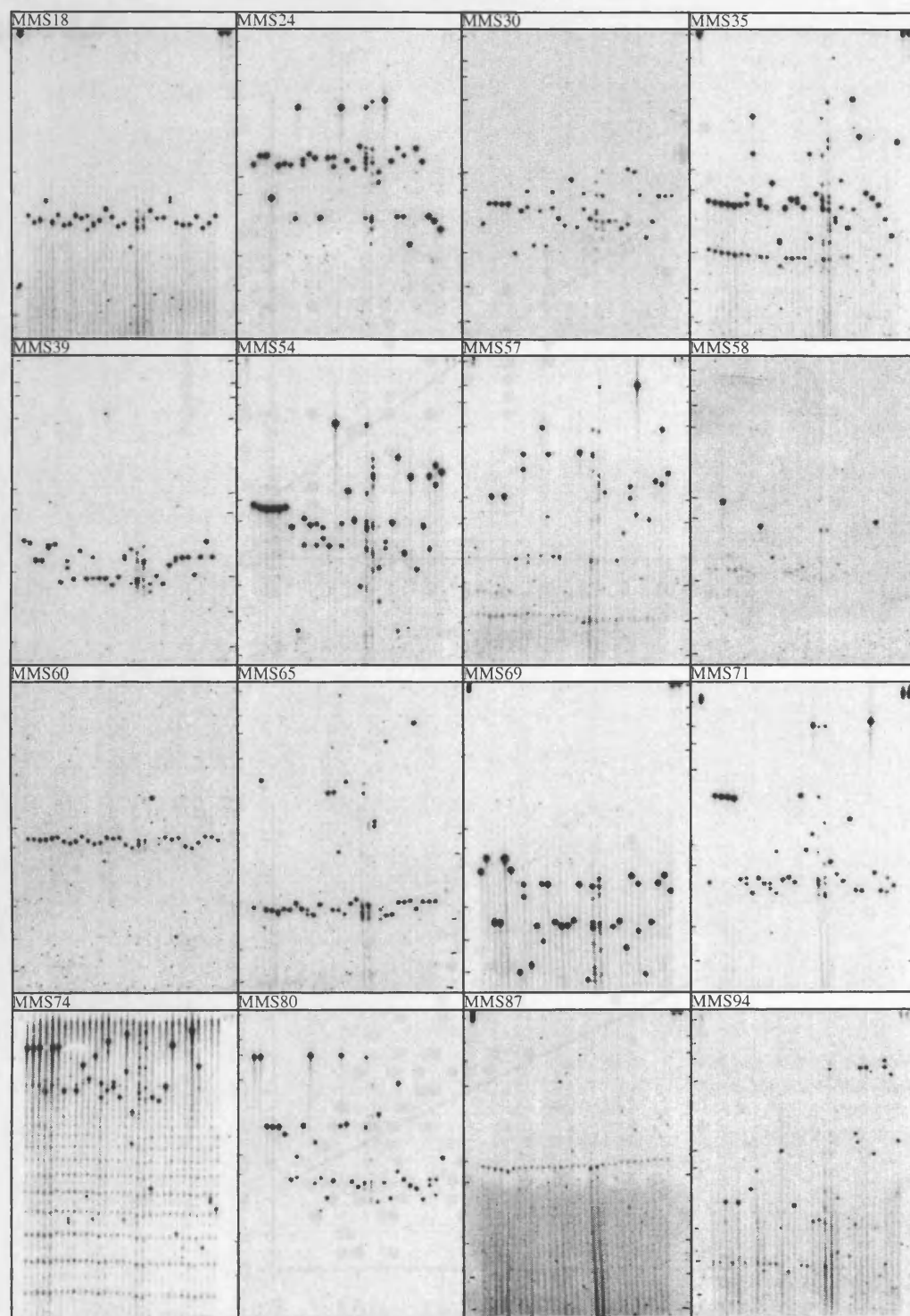
All VNTR probes were hybridised to *Mbo*I-digested Southern blots as in Figure 3.2. A selection of allele length diversity profiles of 16 VNTR loci is presented in Figure 3.4. A range of variabilities were displayed by different loci. As expected, the 23 strains of wild mice revealed considerably greater levels of allele diversity than seen in inbred strains. A few loci such as MMS87 were largely monomorphic between both wild and inbred strains of mice whilst other loci such as MMS80 displayed high levels of length polymorphism. Some probes (e.g. cMMS35) detected >2 bands in some strains of mice. In this example, hybridisation of the same probe to an *Alu*I-digested genomic Southern blot demonstrated that the multiple bands were the result of an internal *Mbo*I restriction site within alleles. Different loci also showed a diversity of allele size ranges. Despite relatively high allele diversity, MMS39 alleles are confined within a small size range from 1.2-1.6 kb. In contrast, alleles of MMS57 range in size from 1.0-11 kb. At this stage of analysis, all allele sizes represent the length of the VNTR repeat array plus an unknown amount of flanking DNA. An interesting variability profile was obtained at locus MMS58. In the majority of wild and inbred strains, alleles were composed of very short repeat arrays (as is apparent from the low signal intensity upon hybridisation). However, in two wild and one inbred strain, much larger alleles were identified. Whilst the process of VNTR expansion was unclear, it may reflect a form of dynamic expansion similar to that observed at a number of triplet repeat loci (Richards and Sutherland, 1997), or at some AT-rich minisatellites such as FRA16B (Yu *et al.*, 1997).

Levels of heterozygosity in the wild strains of mice were strongly and linearly correlated with the number of different alleles (Figure 3.5a). Assuming an effective population size ( $N_e$ ) for wild mice of  $10^4$  (Kimura and Crow, 1964), heterozygosity (H) yielded a mutation rate ( $\mu$ ) for the most variable locus (MMS74) of  $2.3 \times 10^{-5}$  per gamete, as estimated from

$$H = \frac{4N_e\mu}{1 + 4N_e\mu} \text{ (Falconer, 1960), and corresponding lower rates for other loci. It therefore}$$

appears that unstable minisatellites with germline mutation rates in the range of 0.1%-2% per gamete which have been the subject of intensive analysis in humans, do not occur in mice. However, these estimates of instability are crude and will be perturbed by demographic factors in the populations from which the wild mice were derived, and by the

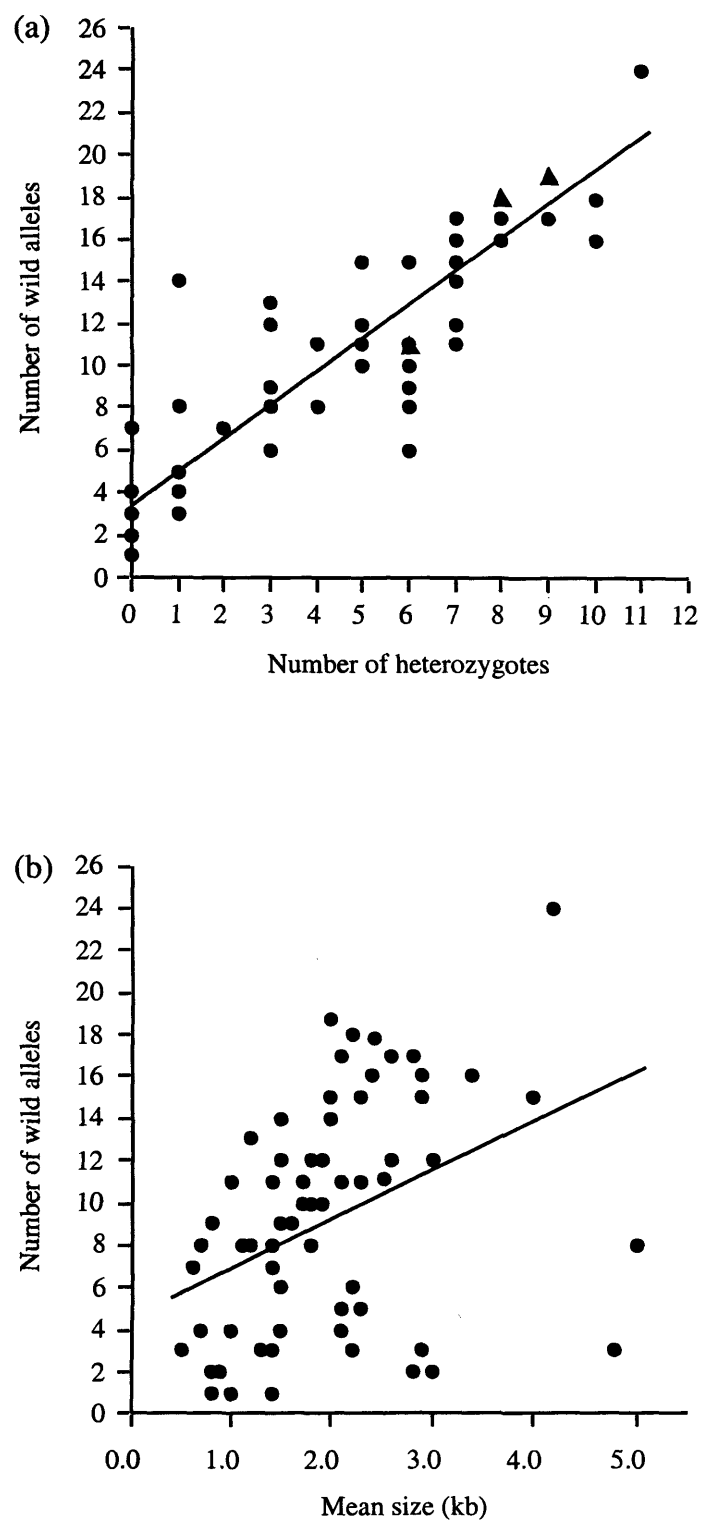
**Figure 3.4**



***Variability profiles of 16 mouse VNTR loci***

Each mouse genomic Southern blot is as described in Figure 3.2, with similar size ranges depicted. A wide range of variabilities are identified, from the monomorphic MMS87 (identified as a VNTR by the characteristically strong hybridisation signal) to the highly variable MMS57 with an allele size range from 1-11 kb. MMS39 displays relatively high levels of polymorphism, but alleles are restricted to a narrow size range. Loci MMS57 and MMS58 are considered in detail in Chapter 4.

**Figure 3.5**



***Allele variability and allele size at mouse minisatellite loci***

A positive correlation was observed between the number of different alleles observed in wild mice and both the number of heterozygous wild mice (a) and the mean size of the mouse VNTR (b). The mean minisatellite array length contains an undetermined length of DNA flanking the repeat array. Figure 3.5 was adapted from Bois *et al.* (1998a)

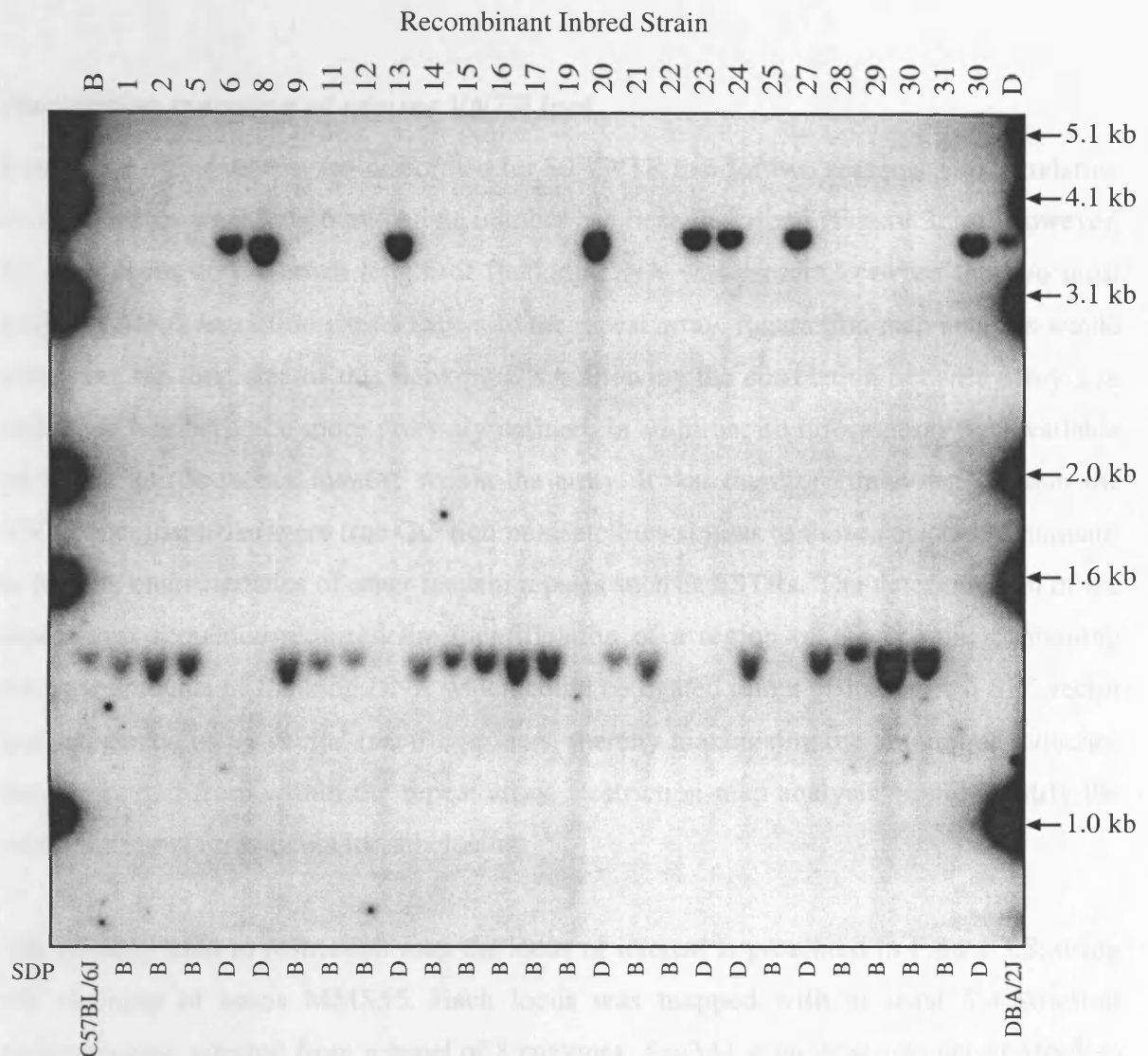
inevitable inbreeding during propagation of the wild strains (although great care was taken to minimise inbreeding; A. Orth and F. Bonhomme, Montpellier, pers. commun.). There was also a positive correlation between the mean size of mouse VNTR loci and the number of different scorable alleles per locus in wild mice strains (Figure 3.5b). This observation strongly suggests that germline instability at mouse VNTR loci is influenced by the array size.

A more direct approach used to estimate mutation rate of specific mouse minisatellites was to use the VNTR clones as probes on Southern blots of *Mbo*I- or *Tsp*509I-digested genomic DNA from the 26 BXD recombinant inbred (RI) strains of mice (Taylor, 1978). BXD mice were generated by hybridisation of the progenitor strains C57BL/6J and DBA/2J, crossing the F1 sibs, then systematically inbreeding isolated lines of sib-pairs to homozygosity. The entire BXD RI set represents 2800 generations of breeding (Ben Taylor, the Jackson Laboratory, Bar Harbour, ME, pers. commun.). A total of 48 VNTRs were hybridised to the BXD genomic Southern blots. For all loci, allele sizes detected in the RI strains were identical to allele sizes within the BXD progenitor strains, demonstrating a total lack of mutation to electrophoretically distinguishable new length alleles during the breeding of the RI lines. This corresponds to a germline mutation rate for each locus of  $\leq 10^{-3}$  (no observed mutations in 2800 generations carries a 95% upper confidence limit of 3 mutations corresponding to a rate of  $10^{-3}$ ) (Bois *et al.*, 1998a).

### **Chromosomal localisation of VNTR loci**

Any locus which displays polymorphism between inbred mouse strains C57BL/6J and DBA/2J can be mapped genetically using BXD RI strains. The segregation patterns of a large number of strain-specific markers have been analysed in each BXD strain. Comparison of strain distribution patterns (SDPs) of alleles of a VNTR locus dimorphic between C57BL/6J and DBA/2J strains with SDPs of other markers with known chromosomal locations using the MapManager software (Manly, 1993) allows the localisation of a VNTR locus. A total of 29 VNTRs were localised in this manner, an example of which is presented in Figure 3.6. Minisatellite MMS80 was localised using the EUCIB backcross mice (as described by Bois *et al.* (1998a)). An additional 22 loci were assigned to chromosomal positions by cosmid fluorescent *in situ* hybridisation

**Figure 3.6**



#### ***MMS65* hybridisation to 26 BXD recombinant inbred strains**

Alleles of locus *MMS65* are dimorphic between strains C57BL/6J and DBA/2J (first and last samples depicted). Hybridisation to a Southern blot of *Mbo*I-digested genomic DNA from the 26 BXD recombinant inbred (RI) lines generated a characteristic pattern of allele sizes between strains. Comparison of this strain distribution pattern (SDP) with SDPs of loci at known chromosomal locations using MapManager software (Manly, 1993) allowed *MMS65* to be localised to chromosome 19. Furthermore, no alleles of sizes detectably different from the C57BL/6J and DBA/2J progenitor strains were observed amongst the RI strains indicative of the absence of germline mutation at the minisatellites during the breeding of the RI strains.

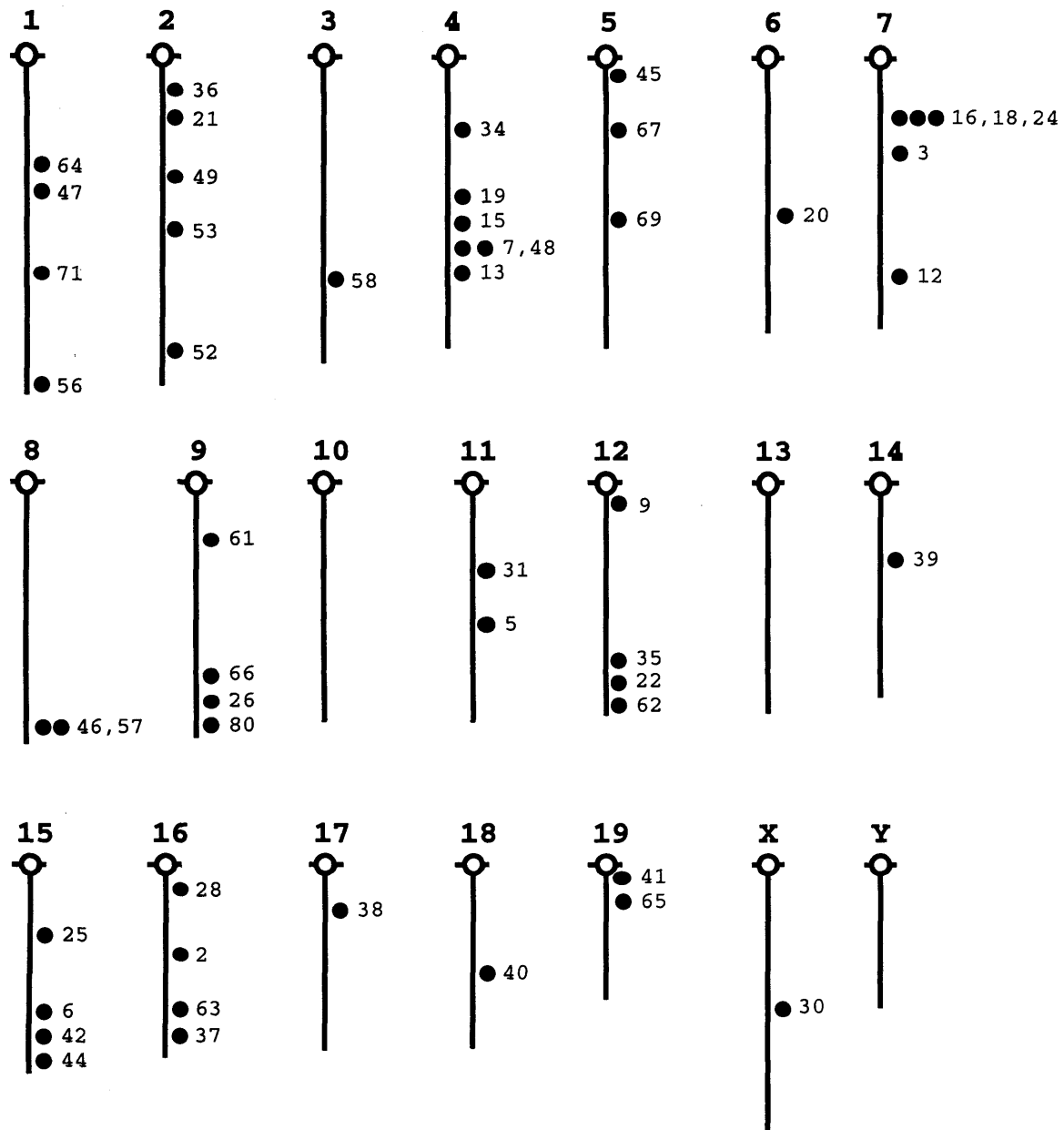
(FISH analysis by J. Williamson, ICRF, London) (Bois *et al.*, 1998a). The cloned loci were dispersed over all autosomes except 10 and 13, with one locus (MMS30) located on the X chromosome (Figure 3.7).

### ***Restriction mapping of mouse VNTR loci***

Restriction map analysis was performed for 30 VNTR loci for two reasons. The correlation between mean array length and allele number has been described (Figure 3.5b). However, for each locus an unknown length of flanking DNA was present between the two most proximal *Mbo*I restriction sites adjacent to the repeat array. Restriction map analysis would determine the total size of this flanking DNA allowing the correlation between array size and allele number to be more precisely defined. In addition, no information was available on repeat unit sequence identity within the array. It was therefore unknown whether the VNTR loci identified were true GC-rich minisatellites similar to those detected in humans, or had the characteristics of other tandem repeats such as ESTRs. The determination of the repeat unit sequence required the identification of a region of the cosmid containing minimal amounts of flanking DNA which could be ligated into a pBluescript II SK<sup>+</sup> vector and sequenced using vector-specific primers, thereby maximising the amount of sequence data generated from within the repeat array. Restriction map analysis would identify the most appropriate fragments for subcloning.

The strategy used to restriction map the locus of interest is presented in Figure 3.8, using the example of locus MMS55. Each locus was mapped with at least 5 restriction endonucleases selected from a panel of 8 enzymes. *Sau*3AI is an isoschizomer of *Mbo*I so was included to allow the amount of flanking DNA present on the locus variability profile detected by hybridisation to the *Mbo*I-digested Southern blot (Figure 3.2) to be determined. The panel of 8 enzymes were selected for their compatibility of buffers allowing efficient double-digest combinations as required for restriction mapping, and for their target site recognition sequences. The generation of high resolution restriction maps surrounding the VNTR would facilitate identification of the boundaries of the repeat array. Each of the 8 enzymes therefore had 4-5 bp recognition sites; *Alu*I (AGCT), *Ava*II (GG<sup>A</sup>/TCC), *Dde*I (CTNAG), *Hae*III (GGCC), *Hinf*I (GANTC), *Mse*I (TTAA), *Rsa*I (GTAC), and

**Figure 3.7**



***Distribution of 51 cloned VNTRs localised in the mouse genome***

Each chromosome is represented as a vertical line with a circle denoting centromeric location. Loci which were dimorphic between strains C57BL/6J and DBA/2J were mapped to chromosomal locations by Southern blot hybridisation to BXD RI strains (Figure 3.6). Loci monomorphic between these strains were localised by fluorescent *in situ* hybridisation (FISH analysis by Jill Williamson, ICRF, London). Locus MMS80 was localised using the EUCIB backcross (Bois *et al.*, 1998a). Loci are distributed along the lengths of chromosomes, with evidence of interstitial clustering (e.g. MMS16, MMS18, MMS24 on chromosome 7). Figure 3.7 was adapted from Bois *et al.* (1998a).



## Figure 3.8

### ***Proximal restriction map analysis of MMS55***

#### **a: Selection of enzymes for restriction map analysis**

A panel of 8 enzymes was screened for their suitability for the restriction mapping of locus MMS55 by performing single digests of the MMS55 cosmid, followed by electrophoresis of the digestion products and Southern blot hybridisation with the VNTR probe. Enzymes which cut within the repeat array (e.g. *RsaI*) were excluded. Of the remaining enzymes, those generating the largest fragments (to maximise the size of the mapped region) and more importantly the smallest fragments (to increase map resolution proximal to the repeat array) were selected. From the 8 enzymes screened, 5 (*AluI*, *AvaII*, *DdeI*, *HaeIII*, and *MseI*) were selected for restriction map analysis. *Sau3AI* was included at a later stage (data not shown).

#### **b: Cosmid digestion products**

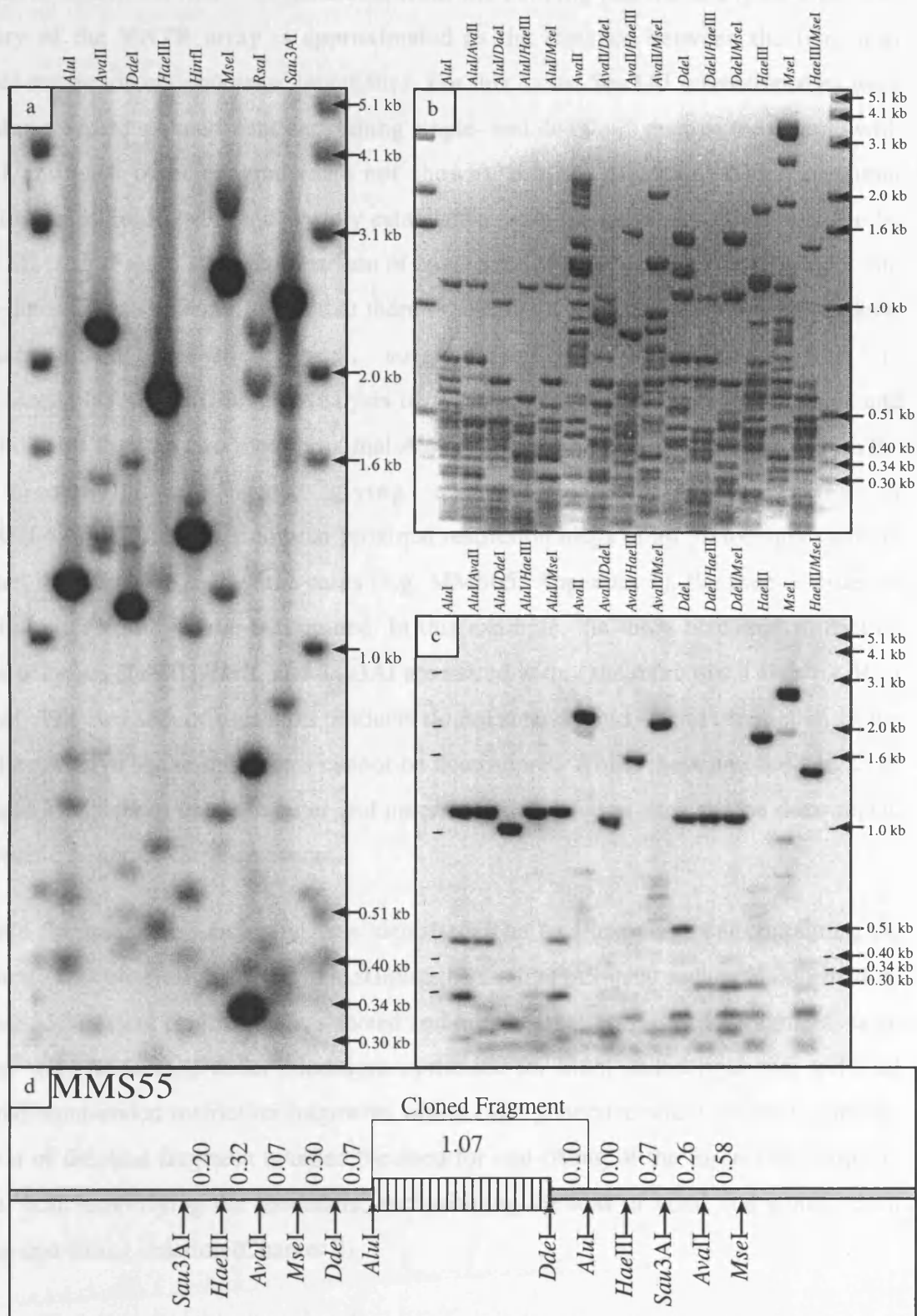
Each single- and double-digest combination of the cosmid was performed using the five enzymes selected from Figure 3.8a. Electrophoresis of the digest products revealed multiple bands from digestion of the vector and insert.

#### **c: Identification of cosmid digestion products containing the VNTR**

The gel from Figure 3.8b was Southern blotted and hybridised using a probe for the VNTR. Cosmid digestion products containing the repeat array were identified, from which a proximal restriction map surrounding the VNTR locus could be generated. The size of each positively hybridising band was determined from the distance migrated during electrophoresis using software modified by Y. Dubrova from Schaffer and Sederoff (1981), which was based on the linear relationship between the size of DNA fragment and the reciprocal of fragment mobility (Schaffer and Sederoff, 1981; Southern, 1979).

Figure 3.8 (continued)

Figure 3.8



## Figure 3.8 (continued)

### d: Schematic proximal restriction map of MMS55

A proximal restriction map was generated from the banding pattern of Figure 3.8c. The boundary of the VNTR array is approximated as the distance between the two most proximal restriction endonuclease target sites. For this locus, *Sau3AI* restriction sites were mapped in a separate experiment performing single- and double-digests of the cosmid with *Sau3AI* and each other enzyme (data not shown). For the majority of loci, proximal restriction maps could be unequivocally established from the band sizes observed. In the case of MMS55 (Figure 3.8d), comparison of band sizes of the *HaeIII* and *MseI* single- and double-digests clearly demonstrates that there is overlap at each end between the products of each single enzyme digest, so the order of restriction sites is *HaeIII-MseI-VNTR-HaeIII-MseI*. Analysis of digestion fragments generated by *AluI* and *AvaII* indicates that the two most proximal *AluI* restriction sites both lie nested within the most proximal *AvaII* sites, giving an order of restriction sites of *AvaII-AluI-VNTR-AluI-AvaII*. Similar proximal restriction maps of all 30 loci analysed are presented in Appendix 1. In some cases (e.g. MMS65, Appendix 1), the precise order of restriction sites could not be determined. In this example, the more proximal restriction sites for enzymes *HaeIII*, *HinfI*, and *Sau3AI* are nested within the more distal sites for *MseI* and *RsaI*. The two sets of digestion products do not overlap and so the orientation of the outer sites relative to the inner sites cannot be determined. Whilst the sum (denoted  $\Sigma$ ) of the 5' and 3' distances between outer and inner nested restriction sites can be determined, each specific 5' and 3' distance cannot.

A cosmid fragment for subcloning was identified. The smallest fragment containing the repeat array was generally selected to maximise the amount of repeat sequence information. For MMS55, the *AluI* fragment was selected and not the smaller *AluI/DdeI* fragment, as all ligations were performed under conditions optimised for blunt-ended ligations, and *AluI* generated blunt-ended restriction fragments whilst *DdeI* generated sticky ended fragments. Selection of the *AluI* fragment avoided the need for end-filling of the insert DNA prior to ligation, both simplifying the procedure, and avoiding the loss of yield that would result from the end-filling reaction (Chapter 2).

*Sau3AI* (GATC). Proximal restriction maps of all 30 loci analysed are presented in Appendix 1.

### ***Correction of allele sizes using restriction map data***

Using the proximal restriction maps containing *Sau3AI* sites, the amount of flanking DNA surrounding the VNTR on the *MboI*-digested genomic Southern blots (Figures 3.2 and 3.4) could be more accurately estimated. For three loci (MMS24, MMS46, and MMS59) for which restriction maps had been generated, the amount of flanking DNA determined by proximal restriction map analysis was greater than the size of some alleles on the *MboI*-digested Southern blot, indicating that a polymorphic *MboI* site was present in these strains near the VNTR, but was absent from the BALB/c strain from which the genomic library and therefore restriction map was derived. To support this assumption, the signal intensity of some small alleles of the VNTR detected by genomic Southern blot was greater than for larger alleles, indicative of larger arrays of repetitive DNA despite smaller *MboI* fragment size (e.g. MMS24, Figure 3.4)

Excluding these three loci, the correlation between corrected array length and allele number was greater than for the uncorrected array lengths for all 77 loci (corrected array length for 27 loci,  $r=0.55$ ; uncorrected array length for 77 loci,  $r=0.37$ , Figures 3.9a and 3.9b). Surprisingly, this improved correlation was not due to the array size correction as the same loci showed a correlation between allele number and uncorrected array length of  $r=0.54$  (Figure 3.9c). The loci selected for restriction map analysis were amongst the most variable minisatellites. This selection apparently identified two groups of loci which displayed different linear regressions between allele size and variability (Figure 3.9a). Each set of loci considered separately therefore displayed stronger correlations than when all loci were combined (selected loci,  $r=0.54$ ; unselected loci,  $r=0.41$ , combined,  $r=0.37$ ). Therefore, not only did the correction of mean array size for the amount of flanking DNA by restriction map analysis fail to improve the correlation between array length and allele variability, but it is possible that *MboI* sites polymorphic between strains are present adjacent to other loci which would lead to inaccuracies in the calculated mean array size.

## Figure 3.9

### ***Correlation between allele variability and allele size corrected for flanking DNA***

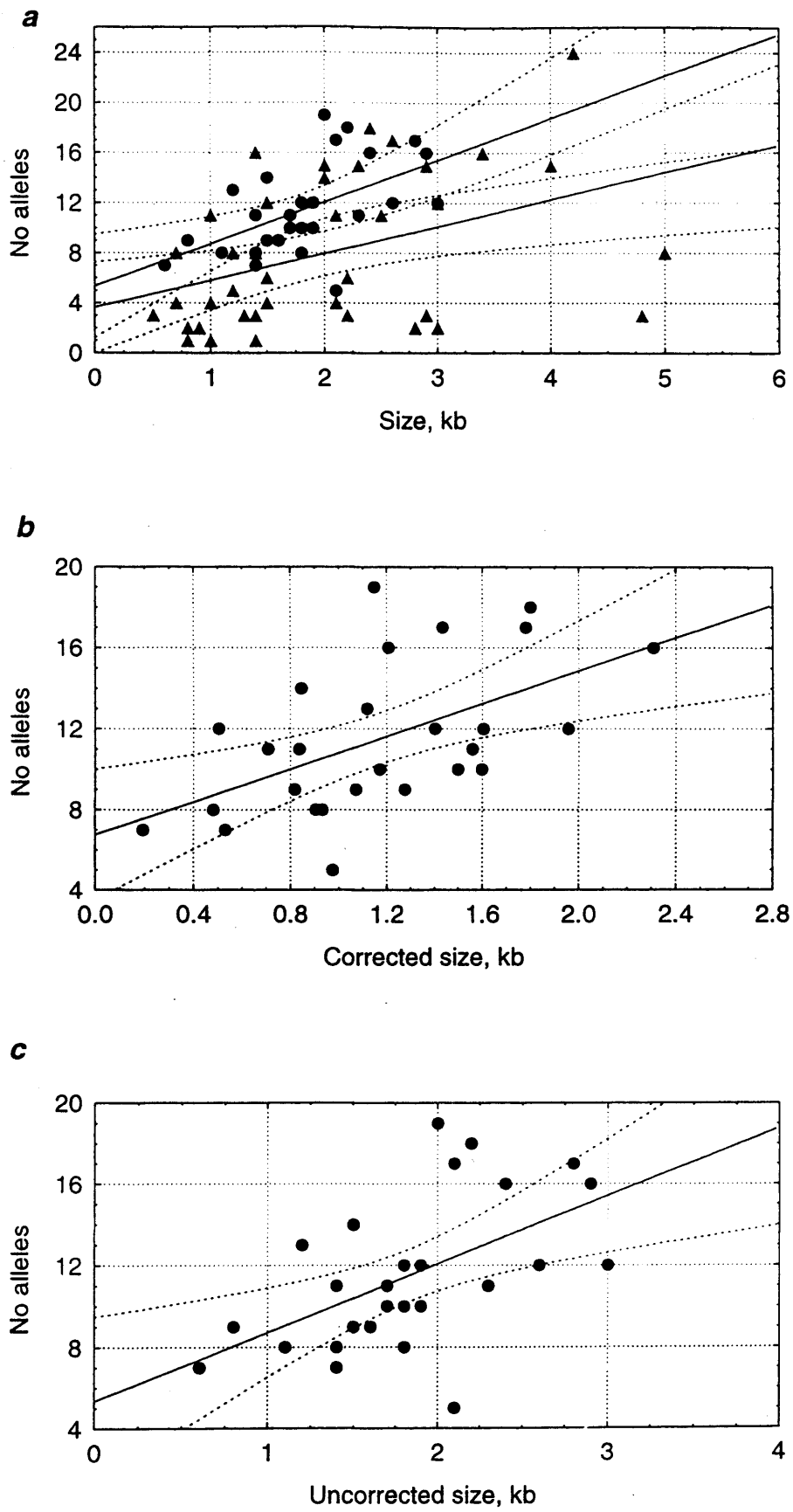
All 77 mouse VNTRs were divided into loci which were selected for restriction map analysis and those which were not selected. Restriction mapping allowed mean array sizes to be corrected for the amount of flanking DNA present on *Mbo*I-digested Southern blots. For all loci combined, mean uncorrected allele size positively correlates with allele number (correlation coefficient;  $r=0.37$ ).

(a) Mean uncorrected array size is plotted against allele number for all loci. Loci selected for restriction map analysis (represented as ●) have a higher regression coefficient than loci which were not selected (represented as ▲).

(b) Mean corrected array size of restriction mapped loci correlates better with allele number than for all loci combined (corrected loci,  $r=0.55$ ; all loci uncorrected,  $r=0.37$ ).

(c) The improved correlation was due to locus selection as opposed to the size correction as the same loci display a similar correlation when corrected and uncorrected for flanking DNA (corrected,  $r=0.55$ ; uncorrected,  $r=0.54$ ). Statistical analysis and preparation of Figure 3.9 was by Y. Dubrova.

**Figure 3.9**

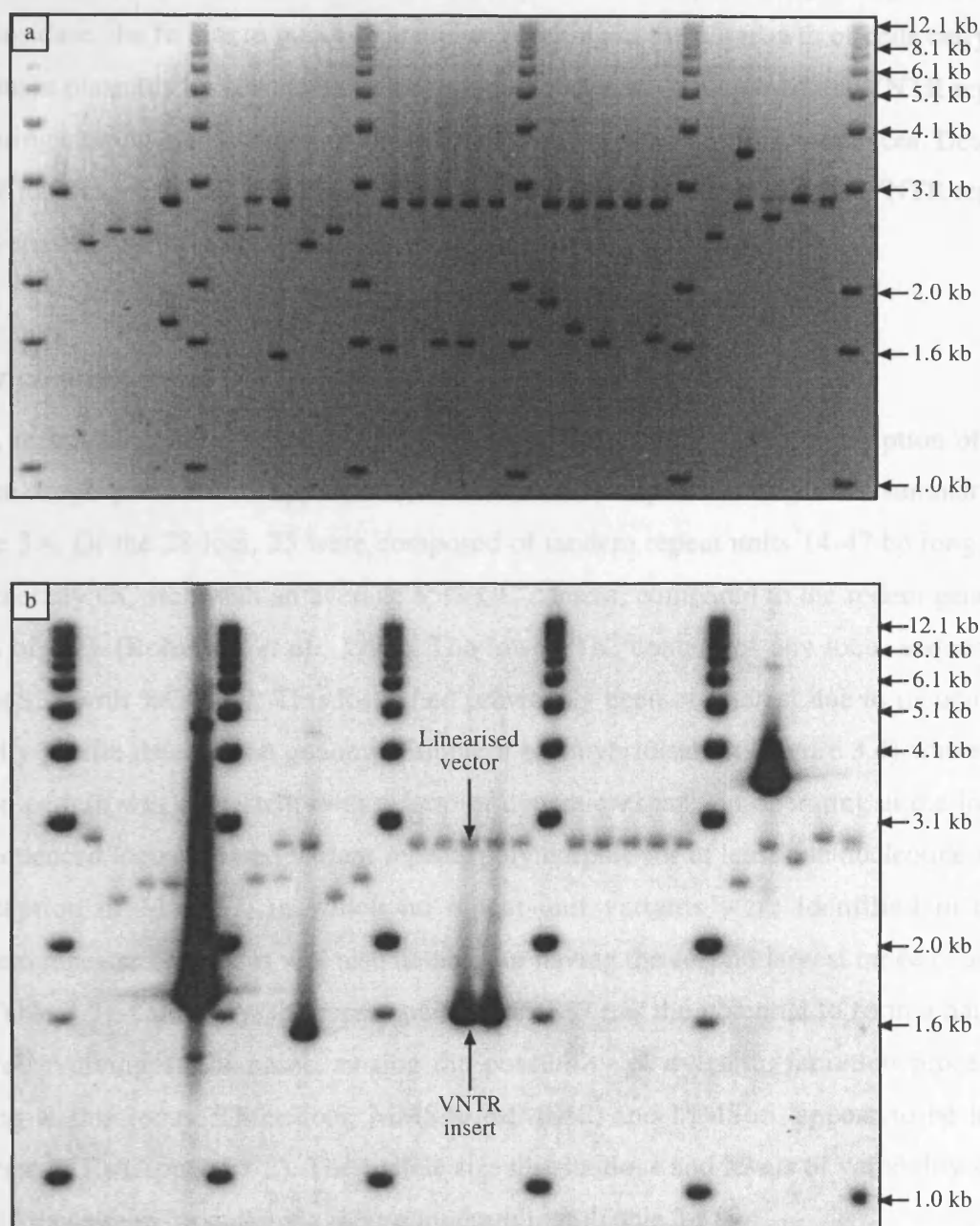


### **Subcloning tandem repeat arrays for sequence analysis**

The cosmid fragments containing the VNTR which were subcloned were selected using three criteria. The first, as mentioned above, was to minimise the amount of flanking DNA thereby maximising the amount of sequence generated from within the repeat array. The second reflected the ease with which the band to be digested could be excised from a gel. In the case of MMS55 (Figure 3.8), the obvious band to select based on the first criterion is the product of the *AluI/DdeI* double digest. This could be readily gel-purified as the VNTR-containing fragment was easily identifiable on an ethidium bromide-stained gel (Figure 3.8b). However, a fragment such as the *AvaII* digestion product would not be selected due to the obvious difficulty in identifying and purifying the band of interest. The third criterion related to the ease with which the subcloning ligation reactions could be performed. A large number of cosmid fragments surrounding different loci generated by digestion with different enzyme combinations were to be ligated into the pBluescript II SK<sup>+</sup> vector. Fragments of insert DNA would therefore have multiple different combinations of sequence termini, some blunt ended (*AluI*-, *HaeIII*-, or *RsaI*-digested termini), others sticky-ended (*AvaII*-, *DdeI*-, *HinfI*-, *MseI*-, or *Sau3AI*-digested termini). To generate a uniform procedure for VNTR ligations, all sticky-ended termini were end-filled and all ligations performed under conditions to maximise blunt-ended ligations into the *EcoRV*-digested pBluescript II SK<sup>+</sup> vector (Chapter 2). To avoid end-filling and the consequential loss of yield during fragment re-purification, enzymes generating blunt-ended fragments were sometimes preferentially selected (e.g. MMS55, Figure 3.8d). All subcloned fragments are indicated in Appendix 1.

Successfully transformed clones were identified by blue-white selection of colonies grown in the presence of X-gal and IPTG (Chapter 2). To confirm the identity of the insert prior to sequence analysis, the recombinant plasmids of 5 colonies for each cloned VNTR locus were purified and the insert excised with *BssHII* which cuts either side of the *EcoRV* insertion site within the vector. Electrophoresis confirmed insert size (Figure 3.10a), whilst Southern blot hybridisation using the original VNTR probe confirmed insert identity (Figure 3.10b). It is clear from Figure 3.10 that the number of false positives was surprisingly high and were characterised by either clones with no insert, inserts of approximately the correct size but lacking VNTR sequences, or inserts of both incorrect

**Figure 3.10**



### ***Verification of size and sequence of the subcloned fragment***

Colonies transformed with recombinant pBluescript II SK<sup>+</sup> vector containing insert DNA were identified by blue-white selection. For each locus, five colonies were selected and grown individually. Plasmid DNA was purified by alkali SDS lysis (Chapter 2) and digested with *Bss*HII (GCGCGC) to excise the cloned fragment in accordance with manufacturer's instructions. Electrophoresis of digestion products allowed cloned fragments to be identified by size (Figure 3.10a) whilst Southern blot hybridisation with the VNTR probe confirmed the identity of the insert DNA (Figure 3.10b). Successful transformants were sequenced as in Chapter 2.



size and sequence. The reasons for the false positives are unclear but may include a failure of blue-white selection of colonies due to low efficiency of X-gal digestion by  $\beta$ -galactosidase, the failure to pick single clones resulting in the outgrowth of cells carrying recombinant plasmids by cells lacking recombinant plasmids, collapse of the VNTR repeat array during cloning, and contamination of insert DNA with non-VNTR sequences. Despite these difficulties, of the 30 loci for which restriction maps were generated, 25 VNTR repeat arrays were successfully subcloned into the pBluescript II SK<sup>+</sup> vector.

### **Sequence analysis of VNTR loci**

In total, repeat arrays of 28 cloned VNTR loci were sequenced. A full description of the sequence data is provided in Appendix 2, with the VNTR repeat unit identities summarised in Table 3.4. Of the 28 loci, 25 were composed of tandem repeat units 14-47 bp long and were generally GC-rich with an average 55% GC content, compared to the rodent genome average of 42% (Robinson *et al.*, 1997). The lowest GC content of any locus sequenced was MMS58 with %GC=37. This locus had previously been of interest due to its unusual variability profile detected on genomic Southern blot hybridisation (Figure 3.4) where the variation pattern was consistent with a form of dynamic expansion operating at the locus. Each sequenced locus showed variant repeats polymorphic for at least one nucleotide with the exception of MMS57, in which no repeat unit variants were identified in over 40 tandem repeats. This locus was also notable for having the second largest range of allele sizes (Table 3.3). Curiously, the repeat unit of MMS57 has the potential to form a hairpin structure involving 16/24 bases, raising the possibility of dynamic mutation processes operating at this locus. Three loci, MMS49, MMS52, and MMS65, appear to be large degenerate STRs (Appendix 2). Their allele size distributions and levels of variability were similar to those seen for authentic mouse minisatellites (Table 3.3).

Despite efforts to minimise the amount of mouse genomic DNA flanking the VNTR that was cloned and sequenced, many subcloned loci did contain flanking sequences. In each case, homology searches were performed using the BLAST and FASTA software packages within the Genetics Computer Group (GCG) Sequence Analysis Software Package version 10.0 programs (Devereux *et al.*, 1984). A single positive result was identified. Sequences flanking MMS39 displays perfect homology to cDNA from the mouse hairless (*hr*) gene

**Table 3.4*****Mouse VNTR repeat unit sequences and polymorphisms***

Locus	Repeat unit size (bp)	%GC	Repeat unit sequence
MMS5	24	63	CTT <b>A</b> RG <b>R</b> TC <b>Y</b> GTGGGCAGGCTCA <b>V</b>
MMS9	20	65	GG <b>R</b> G <b>Y</b> AGGGTASGAGAGTGA
MMS18	14–17	67	GGGTGACA[ <b>V</b> ] [ <b>H</b> ]G <sub>1–4</sub> A[ <b>C</b> ]DG <b>R</b> T
MMS24	28	64	TGTGAGCAC <b>R</b> T <b>G</b> MCTGCAGTGTCTGC <b>Y</b> <b>V</b>
MMS25	17	70	GGGTCCCT <b>T</b> MCTCC <b>Y</b> CA <b>T</b>
MMS26	46	59	GCACACTGCTGCTTCTC <b>Y</b> GCAGTG <b>K</b> TCTC <b>M</b> TGCCCATAGTCTCCAT
MMS30	39	47	AGGAGATT <b>C</b> <b>M</b> S <b>T</b> TCACACTATACAGAAGATGGTGT <b>C</b> AG <b>C</b>
MMS35	28	57	GGCCATGCCAGTGGT <b>C</b> CTTT <b>C</b> ACW <b>C</b> T <b>C</b> A
MMS38	47	62	GGGGATTCCACAGGG <b>K</b> GCCTGTGGTCCAGCACCTGGACAACATGGCT
MMS39	30	60	GATGT <b>Y</b> SCWGT <b>G</b> YGTGCTCC <b>C</b> ACCTCCTGT
MMS41	42	71	GGCCACACACAGGGGCTGACTCC <b>S</b> AGGAGCAGGCTGGGAGCA
MMS44	39	43	CCTGCTG <b>A</b> S <b>A</b> SCATCTTCTGTATAGTGTGA <b>S</b> KGAATCT
MMS45	40	44	GGGTAGGG <b>T</b> RGAGATACTCAGTTGTTACACTGTCATCTAA
MMS47	38	45	TT <b>T</b> YCTGACCTAGCTTACCTTTGGTGT <b>T</b> AGAGCGTGT <b>G</b>
MMS48	19	68	GGC <b>A</b> SAGGG <b>A</b> NG <b>R</b> SAGCAG
MMS49	3–17	58	C <sub>1–4</sub> ( <b>T</b> <b>S</b> ) <sub>1–15</sub>
MMS52	4–9	57	C <sub>1–4</sub> T <sub>1–2</sub> CA[ <b>T</b> ] [ <b>Y</b> <b>T</b> ]
MMS54	20	60	ARGA[ <b>T</b> <b>G</b> <b>C</b> <b>T</b> <b>G</b> ]GTGAGYACAGC
MMS55	44	48	GGKGAGGGCAS <b>A</b> T <b>K</b> CTGAG(TG) <sub>2–4</sub> [ <b>T</b> ]ACATG(T) <sub>3–5</sub> GCATGAT
MMS57	24	50	GCTGTGTAGACAGAGCAGTAGAGT
MMS58	27	37	TTGGTTAG <b>R</b> TAGTTGATACATGCTCAC
MMS60	25	60	CCTCC[ <b>T</b> <b>C</b> <b>C</b> ][ <b>A</b> ]TGTGCTCC <b>Y</b> [ <b>Y</b> ]TGT[ <b>T</b> <b>C</b> <b>T</b> ]
MMS63	31	55	TCCCCAGTCTGACCTC <b>R</b> TAGTCTATCTGTCC
MMS65		62	GC-rich STR
MMS69	41	54	GGG <b>T</b> CCAGCAT <b>Y</b> CCCAGCTCTATCTGAGCACACTCTCTAT
MMS71	20	45	C <b>Y</b> KGCT <b>R</b> TAGATG <b>W</b> TGACTT
MMS73	31–39	47	(TG) <sub>3</sub> TA(TG) <sub>4–8</sub> [ <b>C</b> ][ <b>A</b> ][ <b>T</b> ][ <b>G</b> ]CACTATAS <b>C</b> <b>C</b> <b>Y</b>
MMS80	40	53	CCAGCCCATGGGACAGACTGTA[ <b>T</b> <b>A</b> ]RCACTAGGTCAGT <b>C</b> <b>T</b>

The sequence of the common repeat for each VNTR locus is given, together with the size of repeat and GC content (%GC). Polymorphic bases are highlighted in bold and described using international degeneracy codes. Square brackets indicate bases, or groups of bases, that are deleted in some variant repeats. Internal dinucleotide repeats are indicated by parentheses. In general, variant sites are only indicated if variants were detected in  $\geq 2$  repeat units. Poor sequence quality or very high levels of variant repeat diversity mean that some repeat unit sequences are simplifications or approximations. A full description of VNTR sequences is provided in Appendix 2.

(Figure 3.11) identifying a new VNTR located within intron 16 of the gene (hairless cDNA sequence: GDB accession number AA760207).

## Discussion

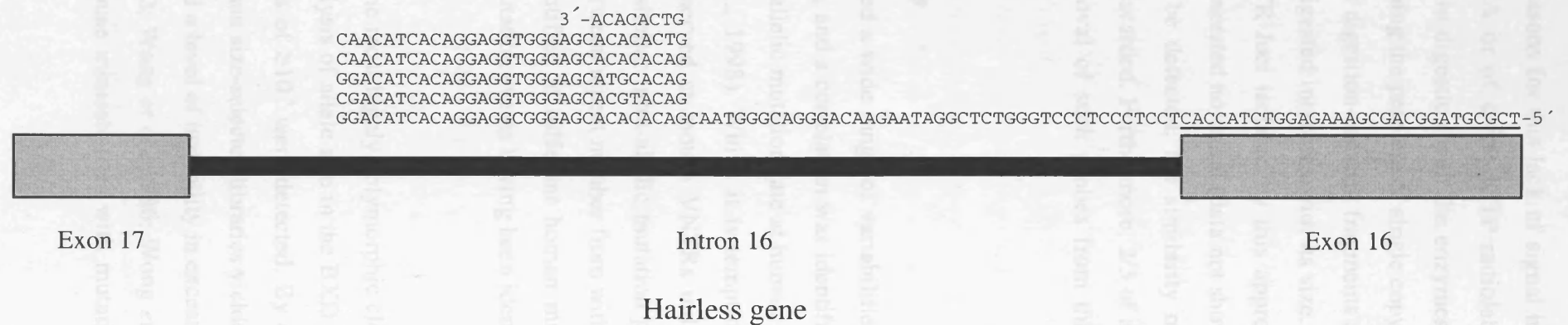
### ***Screening for VNTR loci***

Two independent approaches have been employed to isolate mouse minisatellites. The first mainly used the cross-hybridising properties of previously isolated VNTR probes and of synthetic tandem repeat probes to screen either a size-selected charomid library or a commercial genomic cosmid library. The screening of 9000 clones led to the generation of 166 probes of which 31 hybridised to single VNTR loci (Table 3.2). The second approach relied on the restriction digestion-resistant characteristics of tandem repeat arrays. From 6700 clones of the genomic cosmid library, 251 probes were generated of which 61 hybridised to single VNTR loci (Table 3.2). A total of 77 different VNTRs were identified (Table 3.3).

In general, the digestion-resistance screen may be considered a preferable approach towards the identification of VNTRs for a number of reasons. i) Whilst there are inherent biases within each cloning strategy, the digestion screen, by avoiding most of the sequence bias of the hybridisation strategy, will identify a more representative selection of mouse VNTRs. ii) To identify one single-locus VNTR probe, an average of 110 colonies were screened by digestion-resistance compared to 290 clones for the cross-hybridisation approach. This difference is not surprising as the largely sequence-independent digestion strategy would be capable of identifying a wider range of tandem repeat sequences, and the insert size within the cosmid library was greater than the charomid library. iii) The ideal locus for analysis would have all alleles of sizes below 4 kb. The digestion resistant screen identified loci with significantly shorter alleles (defined from *Mbo*I-digested genomic Southern blots) than the hybridisation screen (median test,  $\chi^2=4.20$ , 1 df,  $p<0.05$ ). This difference is most likely due to the size-selection procedure employed to generate the charomid library.

When data from each cloning strategy are combined, 60% of all probes used for Southern blot hybridisation generated no signal (Table 3.2), severely hampering the efficiency of

**Figure 3.11**



***MMS39 is located within intron 16 of the hairless gene***

The sequence of the repeat array and flanking DNA of locus MMS39 is presented in a 3' to 5' orientation with the region of homology to exon 16 of the hairless gene underlined. The sequence contained 93 bp of complete identity to the published cDNA sequence (data not shown). The minisatellite repeat begins 42 bp from the 5' end of intron 16. The 5' sequence of the clone was not characterised.

each technique. Possible reasons for this lack of signal have been discussed and include poor quality of probe DNA or of  $\alpha$ -<sup>32</sup>P-dCTP radiolabelling, and the possibility that single-copy DNA resistant to digestion with the enzymes selected had been isolated. An obvious solution to overcoming the problem of single copy DNA for the digestion-resistant strategy is to discard smaller digestion-resistant fragments as the probability that a region of single-copy DNA will be digested increases with its size. The mean size of probes which detected single locus VNTR loci isolated by this approach was 1.59 kb, compared to 1.46 kb for probes which generated no signal (data not shown). Whilst a small difference in probe sizes can therefore be detected, the similarity of the sizes argues against any recoverable probes being discarded. Furthermore, 2/3 of all VNTRs were identified using probes of  $\leq 1.5$  kb, so removal of such probes from this screen would greatly reduce screening efficiency.

### ***VNTR variability profiles***

The 77 VNTR loci displayed a wide range of variabilities between both inbred and wild strains of mice (Figure 3.4), and a correlation was identified between mean size of alleles and locus variability. Intra-allelic mutation rate at human minisatellite CEB1 is dependent on allele size (Buard *et al.*, 1998). Whilst it is tempting to speculate that the length-dependent variabilities observed at mouse VNTRs indicates that mouse minisatellite mutation is dominated by similar intra-allelic mutation processes, Armour *et al.* (1994) identified a similar trend between repeat number from within the original VNTR clone and the number of alleles identified at different human minisatellite loci, despite a high frequency of inter-allelic mutation events having been identified at a number of human loci (Jeffreys *et al.*, 1999).

Forty-eight loci including the most highly polymorphic clones were screened for evidence of *de novo* mutation by analysis of allele size in the BXD RI strains of mice. Surprisingly, no loci with mutation rates of  $\geq 10^{-3}$  were detected. By comparison, cross-hybridisation methods used to screen human size-selected libraries yielded 32 independent minisatellites, at least 5 of which displayed a level of instability in excess of 0.5% per gamete (Armour *et al.*, 1990; Royle *et al.*, 1992; Wong *et al.*, 1986; Wong *et al.*, 1987). Whilst it is possible that a small number of mouse minisatellites with mutation rates of  $\geq 10^{-3}$  do exist, this

difference suggests a systematic shift in instability profiles between the human and mouse genomes, and raises the possibility that mouse minisatellites are processed in the germline differently from their human counterparts. Due to the apparent absence of hypermutable mouse minisatellites, no further attempts to identify endogenous minisatellites from mice are planned.

### ***VNTR sequence analysis***

To confirm that authentic minisatellites had been identified, proximal restriction maps were generated for 30 loci (Appendix 1) and the repeat arrays of 28 loci were subcloned and sequenced (Table 3.4, Appendix 2). Sequences of the majority of loci were very similar to true human minisatellites (Armour *et al.*, 1993; Buard and Vergnaud, 1994; Jeffreys *et al.*, 1991a; Neil and Jeffreys, 1993), with GC-rich repeat units of 14-47 bp in length, and the presence of sequence polymorphisms between repeats in each array with the single exception of MMS57. Three loci were large degenerate GC-rich STRs. This was not surprising given that the screening strategies tend not to select for repeat unit size within the array, but rather for the base composition, repetitive nature, and overall size of the cloned locus. It was not possible to compare sequences of loci isolated by the digestion-resistance screen with those from the hybridisation screen to determine whether a difference existed between the loci identified using the alternate approaches as only 3 of the sequenced loci were identified exclusively using the hybridisation strategy.

Three minisatellites, MMS39, MMS57, and MMS58, were of particular interest due to their unusual characteristics defined both by sequence analysis (Table 3.4) and allele size variability profiles (Figure 3.4). MMS39 displays moderate levels of polymorphism despite the restriction of allele sizes to a range of just 400 bp (representing 10-15 30 bp repeats). Only two other loci, MMS60 and MMS67 showed similar levels of variability over such a small size range. Sequence analysis of the DNA flanking MMS39 identified the locus to be situated within intron 16 of the hairless gene (*hr*). *hr* encodes a putative transcription factor with a single zinc-finger domain, which is highly expressed in both brain and skin (Ahmad *et al.*, 1998a; Cachon-Gonzalez *et al.*, 1994; Thompson, 1996). Homozygous recessive mice with the *hr* mutation or the more severe rhino mutation (*hr<sup>rh</sup>*) in the hairless gene become essentially naked by the third week of life (Mann, 1971). Furthermore, immune

defects associate with the *hr* mutation with *hr/hr* mice of the HRS/J strain having a higher incidence of leukaemia than their normal sibs and a lower cellular immune response against viruses (Heiniger *et al.*, 1976; Hiai *et al.*, 1977). Analysis of its human homologue (*HR*) revealed high levels of conservation of both sequence and genomic organisation between the species (Ahmad *et al.*, 1998b; Ahmad *et al.*, 1999). Mutations in the human hairless gene have been implicated in congenital atrichia, a rare form of hereditary hair loss characterised by the complete shedding of hair shortly after birth (Ahmad *et al.*, 1998a; Nothen *et al.*, 1998). The *hr* mutation in the HRS/J strain was caused by the insertion of a murine leukaemia provirus into intron 6 of the gene which was thought to interfere with mRNA splicing (Cachon-Gonzalez *et al.*, 1994; Stoye *et al.*, 1988), whilst the more severe *rhino* phenotype is due to a nonsense mutation within exon 4 (Ahmad *et al.*, 1998b). The intronic location of MMS39 may explain its narrow size window of allele variation as large alleles may interfere with either gene transcription or mRNA splicing and so be selected against. It is unknown whether any hairless strains of mice carry MMS39 alleles of unusual sizes, or whether any phenotype would be associated with VNTR alleles outside of the size window described in Table 3.3. No further work was undertaken at MMS39 due to time constraints. The intronic minisatellite was instead brought to the attention of Dr. Angela Christiano who has worked extensively on the mouse hairless gene (e.g. Ahmad *et al.* (1998a); Ahmad *et al.* (1998b); Ahmad *et al.* (1999)).

Minisatellites MMS57 and MMS58 display characteristics consistent with a form of dynamic mutation operating at the loci. Both loci will be considered in greater detail in Chapter 4.

### ***Variant repeat distribution in three mouse minisatellites***

Systems of MVR-PCR were established for three of the most variable mouse minisatellite loci (MMS24, MMS26, and MMS80) to investigate allele structures (MVR-PCR was performed by P. Bois) (Bois *et al.*, 1998a). All loci showed human-like interspersions of variant repeats with considerable structural allelic variability between strains of wild mice (data not shown). Alleles could be readily divided into closely related groups by aligning regions of MVR map similarity (Bois *et al.*, 1998a). This ability to define allele lineages means that mouse minisatellites may be used as tools for the study of population diversity

and evolution. The application of mouse minisatellites to phylogenetic analysis will be considered in Chapter 5.

Comparison of allele structures also provided indirect evidence for the mechanistic basis of mouse minisatellite instability. Structural differences between closely related alleles tended to involve small numbers of repeats, with evidence of local reduplications and no clear signs of polarity or inter-allelic exchange in contrast to the polarised inter-allelic processes observed to operate at high frequency at a number of human minisatellites. While this indicates that mouse minisatellites are likely to mutate mainly by intra-allelic mechanisms, an inter-allelic component to mutation cannot be excluded. The modest allelic diversity observed suggested a rate of germline instability well below  $10^{-3}$  per gamete. Mutation analysis by SESP-PCR at locus MMS80 indicated that the mutation rate is likely to be well below  $10^{-4}$  as a screen of  $10^6$  sperm from a BALB/c mouse homozygous for a 47 repeat allele failed to detect any mutations (Jeffreys *et al.*, 1997).

### ***Chromosomal localisation of mouse VNTRs***

In humans, minisatellites tend to be clustered in proterminal chromosomal regions (Royle *et al.*, 1988b). In contrast, previous studies in mice have demonstrated that large VNTRs have interstitial distributions (Jeffreys *et al.*, 1988; Julier *et al.*, 1990). Similarly, we found no evidence for clustering towards telomeres, although there was evidence of interstitial clustering. This distribution of VNTRs is interesting given the recombination-based instability of human minisatellites, which has been proposed to be a result of their possible involvement in pairing between homologous chromosomes at meiosis. In the mouse genome, meiotic chromosomes pair at multiple interstitial locations, whilst in humans, pairing is initiated near chromosomal termini reflecting the differential distribution of VNTRs observed between the species (Carpenter, 1987).

A screen for minisatellite loci from the human, pig, and rat genomes identified similar differences in VNTR distribution patterns with 90% of human VNTRs clustered in subtelomeric regions compared to 66% and 30% in pigs and rats respectively. We detected the localisation of only 15% of mouse VNTRs in these regions. Comparative mapping of the locations of minisatellite-containing regions in the human, pig and rat genomes



demonstrated that the VNTR-rich interstitial sites in pig and rat often correspond to terminal cytogenetic bands in humans (Amarger *et al.*, 1998). For example, 3 of the 11 interstitial VNTRs detected in pigs are clustered on chromosome 6q2.1-q2.2 which corresponds to the position of an ancestral chromosomal fusion event and contains remnants of the telomere array (Gu *et al.*, 1996). The homologous regions in humans are present on chromosomes 1 and 19 (Goureau *et al.*, 1996; Yerle *et al.*, 1997). It was therefore suggested that minisatellites were created near telomeres, and that their internalisation arose from cytogenetic rearrangements involving chromosomal ends (Amarger *et al.*, 1998).

## Chapter 4

# Variant repeat distribution in two mouse minisatellites

### Summary

Variant repeat distribution was analysed at two endogenous mouse minisatellites, MMS57 and MMS58, both of which were identified in Chapter 3 with characteristics suggesting the possible operation of dynamic mutation processes. MMS57 displays the second greatest range of allele sizes of all true mouse minisatellites characterised to date. Sequence analysis of the MMS57 clone revealed complete uniformity in tandem repeat sequence, and the potential for the repeat array to form a hairpin conformation involving 16/24 bases within each repeat. In this chapter, analysis of alleles from a number of mouse strains identified variant repeats which differed in their capacity to form hairpin conformations. The purine-rich strand of a variant termed 'G' formed predicted hairpin conformations with thermodynamic stability comparable to the FRA16B repeat array. MVR-PCR analysis identified large homogeneous arrays within large alleles composed entirely of G-type repeats, consistent with the operation of hairpin-mediated dynamic expansions at this locus. Large homogeneous arrays of variant repeats which were not predicted to form stable hairpins were not detected. MMS58 was the most AT-rich mouse minisatellite identified. Despite the majority of MMS58 alleles being small, large alleles were identified in a small number of strains, suggesting that MMS58 undergoes dynamic expansions. However, MVR-PCR and DNA conformation analyses failed to identify evidence for dynamic mutation processes operating at this locus. Instead, MMS58 is likely to be a recently expanded minisatellite, which may have undergone independent expansions in both *Mus m. domesticus* and *Mus spretus*.

### Introduction

Three of the most variable loci isolated in Chapter 3, MMS24, MMS26, and MMS80, had been further analysed by MVR-PCR (Bois *et al.*, 1998a). Mutation screening at MMS80 by

SESP-PCR failed to identify any *do novo* mutation events in  $10^6$  sperm analysed (Jeffreys *et al.*, 1997), demonstrating the severe limitations of using these loci to analyse mechanisms of minisatellite mutation in mice. While such systems of MVR-PCR have potential use in mouse phylogenetic analysis (discussed further in Chapter 5), the detailed analysis of further loci displaying similar levels of variability was unlikely to improve our understanding of mutation processes operating at mouse minisatellites. However two loci, MMS57 and MMS58, displayed characteristics consistent with mutation mediated by a form of dynamic instability and so warranted further investigation.

Dynamic mutation refers to the large expansions of tandem repeat arrays which occur within a single generation at a subset of triplet repeat loci (Mitas, 1997; Richards and Sutherland, 1997) and AT-rich minisatellites (Bois and Jeffreys, 1999; Yu *et al.*, 1997). Typically, these expansions occur from within an array of homogeneous repeat units above a certain size threshold (Richards and Sutherland, 1997). Dynamic instability may occur by polymerase slippage during DNA replication where the nascent strand dissociates from the template allowing the two strands to slip relative to each other. Successful priming from the slipped strand results in a change in repeat number. The putative formation of hairpin structures, triplexes, or tetraplexes by both unstable triplet repeats and AT-rich minisatellites such as FRA16B may mediate dynamic expansion events at these loci (Bois and Jeffreys, 1999; Mitas, 1997; Yu *et al.*, 1997) as unusual DNA conformations may stabilise the slipped strand intermediate, thereby increasing the probability that synthesis would re-initiate from the slipped strand position (Weitzmann *et al.*, 1997). Variant repeats which disrupt secondary structure formation would reduce the probability of dynamic expansions occurring (Richards and Sutherland, 1997). However, despite the standard dogma that dynamic mutation occurs through intra-allelic mechanisms of slippage, there is evidence that flanking sequences (Brock *et al.*, 1999) and even the identity of the allelic homologue (Igarashi *et al.*, 1996) may affect mutation rate.

MMS57 displays the second largest range of allele sizes (1-11 kb) of all loci identified in Chapter 3, suggesting that mutation events may involve unusually large changes in repeat array length. Analysis of over 40 repeats from the cloned minisatellite found complete sequence identity between all repeats. The repeat array also displayed the potential for

hairpin formation in which 16/24 bases within each repeat could form Watson-Crick base pairings. These characteristics are all consistent with mutation processes mediated by dynamic instability.

The allele size distribution of MMS58 (Figure 3.4) revealed that whilst alleles in most strains were very short, a minority of mice displayed much longer alleles again consistent with large changes in array length. Furthermore, the locus is the most AT-rich minisatellite identified in the mouse, although in contrast to the human AT-rich minisatellite FRA16B which is subject to dynamic mutation, the MMS58 repeat unit displays no apparent palindromic characteristics.

If both MMS57 and MMS58 do undergo dynamic expansions, two testable predictions can be made. The first is that the repeat array would show the potential for the formation of secondary conformations from single-stranded DNA. The second is that large (and therefore potentially dynamically expanded) alleles would be composed of large arrays of homogeneous hairpin-forming repeat units. Both predictions are tested in this chapter, the first by computer analysis of the repeat unit sequences for their potential to form secondary structures, the second by the characterisation of variant repeat type distribution by MVR-PCR.

### ***Computer analysis of putative DNA conformation***

Two programs were employed to determine the potential for DNA secondary structure formation. The first is the stemloop algorithm within the GCG molecular biology software package (Devereux *et al.*, 1984) which searches for inverted repeats (therefore putative DNA hairpins) within the selected sequence. The algorithm compares the input sequence with the inverted input sequence at each base, and in every register, to identify the most stable putative hairpin loops. A more sophisticated algorithm for the identification of putative hairpin structures is provided by the Mfold version 3.0 software, available at <http://mfold.wustl.edu/~folder/dna/form1.cgi>. The software, described by SantaLucia (1998), is based on the nearest-neighbour (NN) model for nucleic acid binding, which calculates the stability of a given base pairing modified by the identity and orientation of neighbouring nucleotides. The binding energies assigned to each NN pair were derived

from seven studies on natural polymers, synthetic polymers, oligonucleotide dumbbells, and oligonucleotide duplexes (Allawi and SantaLucia, 1997; Breslauer *et al.*, 1986; Delcourt and Blake, 1991; Doktycz *et al.*, 1992; Gotoh and Tagashira, 1981; SantaLucia *et al.*, 1996; Sugimoto *et al.*, 1996; Vologodskii *et al.*, 1984) and were empirically corrected for salt dependencies (SantaLucia, 1998; SantaLucia *et al.*, 1999). Salt concentrations were set at 10 mM Na<sup>+</sup>, 0.5 mM Mg<sup>2+</sup>, reflecting the intra-cellular concentrations of each ion in mammalian cells (Alberts *et al.*, 1989). The stability of hairpin structures was calculated at 37°C.

### ***Analysis of variant repeat distribution***

The analysis of minisatellite loci based purely on allele size provides a very limited picture of allelic variation. Analysis of variant repeat distribution, most typically by MVR-PCR, allows relationships between the structures of different alleles to be compared. MVR-PCR can rapidly determine repeat type distribution throughout regions in excess of 100 repeats. Division of alleles by MVR code into distinct subgroups also allows minisatellites to be used as tools to study phylogeny at the locus, from which population structure may be inferred (see Chapter 5). Patterns of variation both within and between allele groups may be informative as to the mutational mechanisms which underlie the observed variation.

The design of a system of MVR-PCR can be broadly divided into 3 stages. Initially, systems must be established for the amplification by PCR of the locus which allows: i) The accurate sizing of alleles; ii) The physical separation and purification by electrophoresis of amplified alleles from a heterozygous subject, from which MVR codes may be determined unequivocally for each allele, and; iii) The sequencing of amplified alleles without the need for cloning. The second stage is to generate sequence data from alleles derived from different strains of mice, in order to identify variant repeats which may be distinguished by MVR-PCR. Some sequence data for repeat-type distribution of both MMS57 and MMS58 has been described in Chapter 3. However, these sequences were derived from cosmid subclones which may have undergone rearrangements during the cloning process. Furthermore, variant repeat identity can vary substantially between strains of mice (for example at locus MMS26 (Bois *et al.*, 1998a)) so repeat types common in many strains may have been absent from the BALB/c mouse from which the cosmid library screened in

Chapter 3 was derived. Finally, the system of MVR-PCR needs to be designed and optimised with the criteria of maximum informativity and efficiency (the ability to detect and distinguish as many variant repeats as possible, whilst minimising the number of MVR primers and therefore the number of reactions required to type each allele).

No variant repeats had been identified at MMS57. This was highly unusual for both human and mouse minisatellites. Initially, MVR-PCR at this locus was considered as a technique by which large numbers of repeats from within a range of alleles could be screened for variants, which would be potentially identifiable as null repeats (unamplifiable repeats due to the presence of sequence variants which prevented annealing of the MVR primer). An alternative approach to MVR-PCR analysis was considered for locus MMS58. Genomic Southern blot hybridisation (Figure 3.4) and restriction map analysis (Appendix 1) of this locus demonstrated that the repeat array of the largest allele detected was only ~1 kb and so all alleles could potentially be analysed directly by sequencing of PCR products. However, where multiple alleles were to be analysed, MVR-PCR was considered to be a quicker and cheaper approach than direct sequence analysis.

## **Results**

### ***PCR amplification of MMS57 and MMS58***

Prior to the investigation of either putative DNA conformations by computer analysis, or variant repeat distribution by MVR-PCR, methods were established for the amplification of alleles at each locus for the reasons described above. Sequence data for both MMS57 and MMS58 had been previously generated from cloned fragments designed to include minimal amounts of flanking DNA to maximise information available from within the repeat array (Chapter 3). As a result, <50 bp of flanking sequence were available for each locus (Appendix 2). Further regions of each cosmid were therefore selected from the restriction maps (Appendix 1) to include either 5' or 3' flanking regions and the repeat array, as indicated in Appendix 1. For MMS57, 807 bp of 5' and 385 bp of 3' flanking sequence were generated, whilst for MMS58, 956 bp of 5' and 185 bp of 3' sequence were produced. (The designation of 5' and 3' in these non-coding regions is arbitrary.) Primers for PCR amplification were designed either side of each locus. Sequence data and primer location

for each locus are presented in Appendix 3. The products of PCR amplification are presented for both MMS57 (Figure 4.1) and MMS58 (Figure 4.2).

Variability profiles for both loci have been described. Alleles including the ~10 kb MMS57 allele from strain MBK were successfully amplified (Figure 4.1). Levels of size variation amongst the smaller alleles of both loci are more clearly apparent when typed by PCR compared with genomic Southern blot hybridisation. A number of additional strains of mice were amplified, including two strains of *Mus spretus* which diverged from *Mus musculus* approximately 1-2 million years ago (Boursot *et al.*, 1996). MMS57 displays a fairly continuous allele size distribution with alleles of between 0.5 kb and 10 kb. In contrast, the majority of MMS58 alleles are small and composed of between 2 and 5 repeat units. Larger alleles are present, most notably in inbred strain BALB/c (35 repeats), and the *Mus m. domesticus* wild strains 22MO and BFM (both alleles were identical by size at 22 repeats). As with the mouse loci MMS24, MMS26, and MMS80 (Bois *et al.*, 1998a), MMS57 alleles in *Mus spretus* are shorter than in most strains of *Mus musculus*, indicating expansion of the minisatellite to have occurred after the *Mus spretus/Mus musculus* divergence, reminiscent of the minisatellite differences observed between primates and humans (Gray and Jeffreys, 1991). However, in contrast to MMS57 and all other mouse minisatellites analysed by MVR-PCR (Bois *et al.*, 1998a), alleles of MMS58 in *Mus spretus* are larger than in most strains of *Mus musculus* (Figure 4.2) indicating that either the ancestral state was an expanded form of the locus which had suffered deletions in most *Mus musculus* strains, or that the locus has expanded independently in both *Mus spretus* and a subset of *Mus musculus* strains.

### **Sequence analysis of minisatellite alleles**

Sequencing of the cloned MMS57 and MMS58 loci derived from a single BALB/c mouse had demonstrated that for MMS57, no variant repeats were present, while there were two variants of MMS58 which displayed repeat sequence divergence at a single nucleotide (Chapter 3). To extend the range of alleles from which sequence information on the repeat array was available, a selection of alleles of each locus was amplified and sequenced from a number of mouse strains. Each allele selected was <1.6 kb to facilitate clean (i.e. without PCR collapse) amplification to levels detectable on ethidium bromide-stained gels prior to

## Figure 4.1

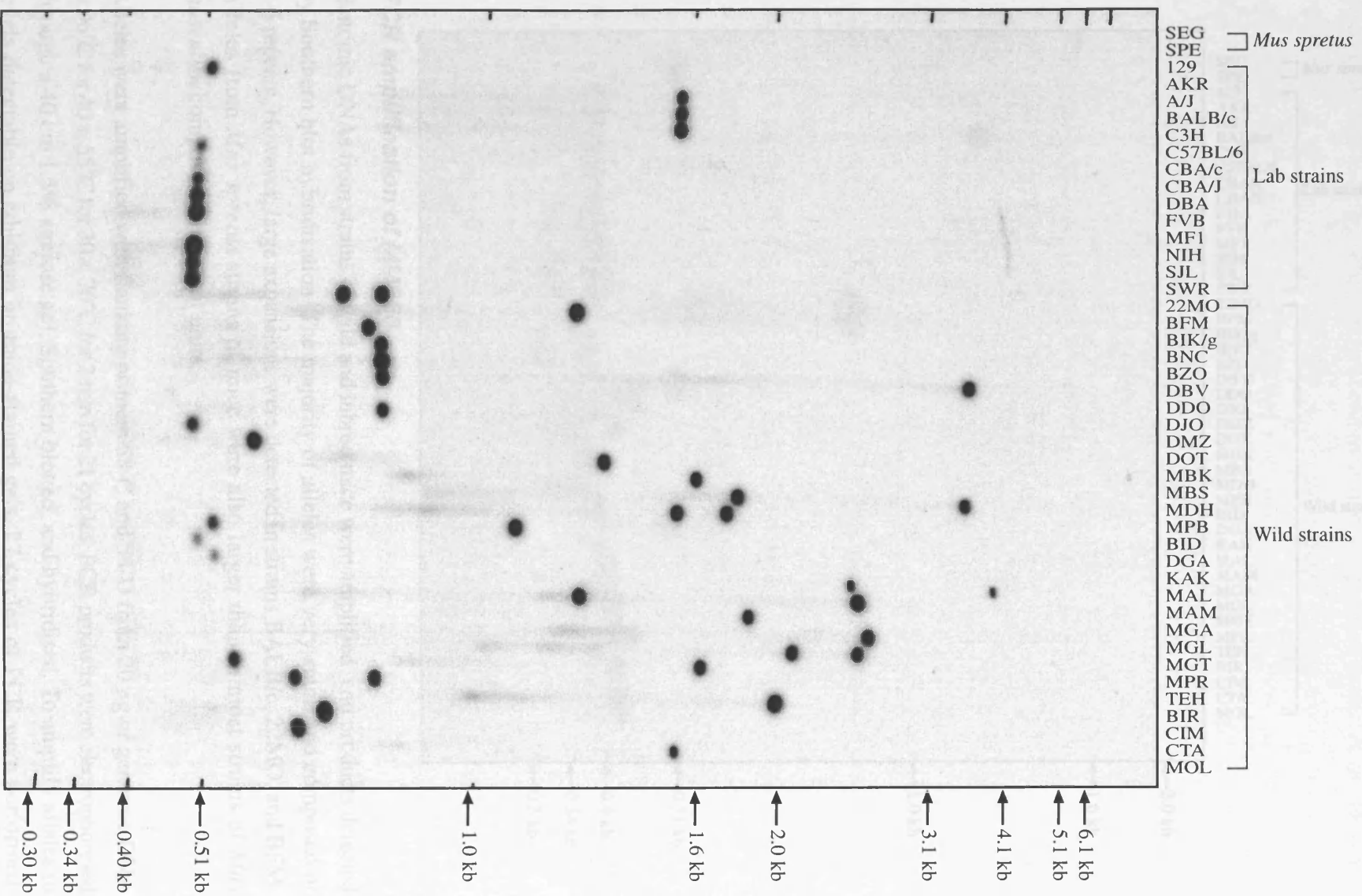
### ***PCR amplification of MMS57***

Genomic DNAs from strains of wild and inbred mice were amplified, and products detected by Southern blot hybridisation. PCR amplification clarifies the level of variation between the smaller alleles at this locus compared with hybridisation of the same probe to genomic Southern blots (Figure 3.4). The first two samples from *Mus spretus* failed to amplify on the autoradiograph presented, suggesting that polymorphisms may exist between *Mus spretus* and *Mus musculus* within the primer sites. However, *Mus spretus* alleles were successfully amplified in further experiments and were shown to be very short, composed of just three repeat units.

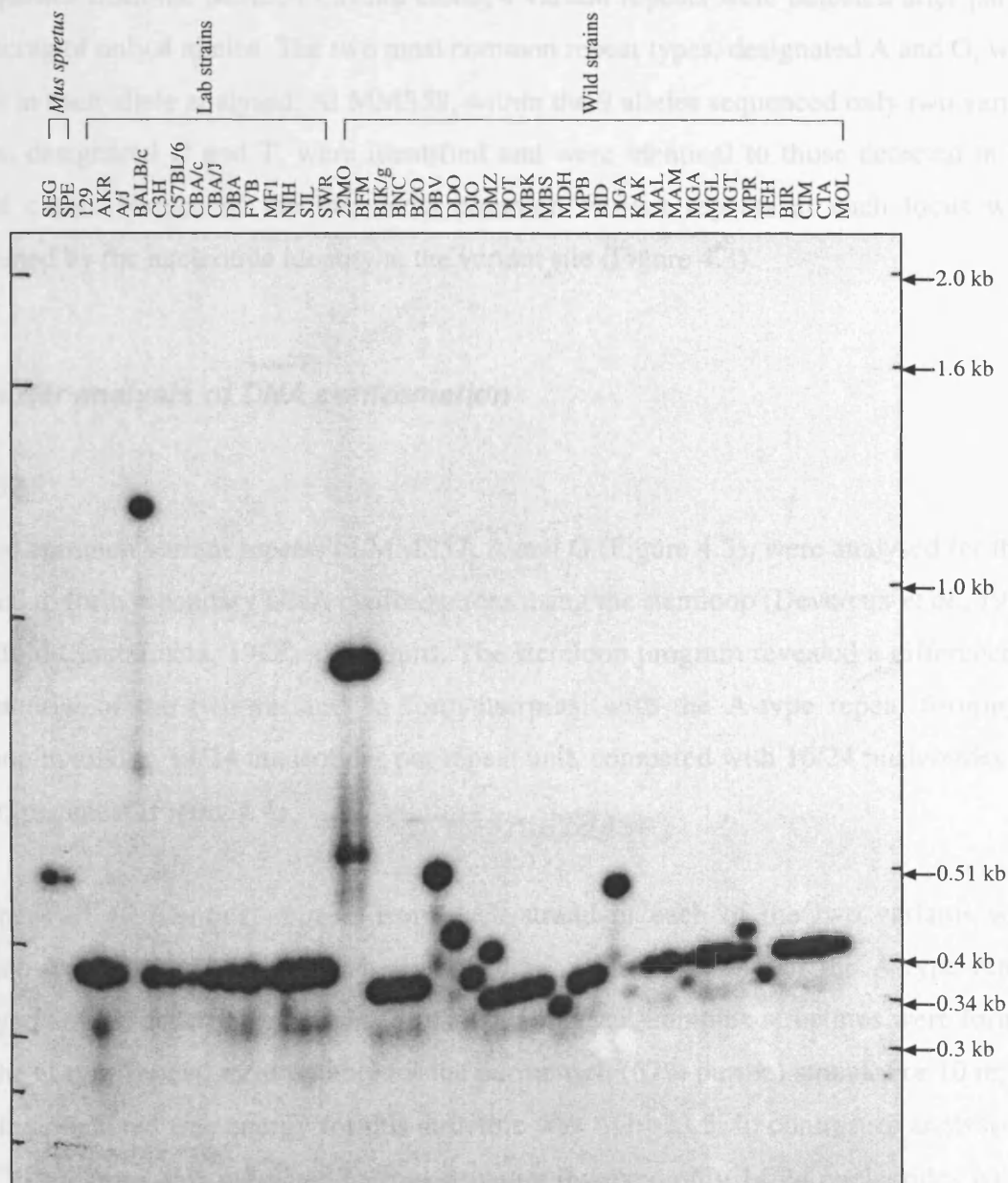
General PCR conditions were as described in Chapter 2. Alleles were amplified with flanking primers 57-C and 57-D from 20 ng of genomic DNA at 96°C for 40 s, 60°C for 30 s, 70°C for 3 min for 22 cycles. PCR products were electrophoresed through a 40 cm 1% agarose gel, Southern blotted, and hybridised. To amplify alleles to levels detectable on ethidium bromide-stained gels, 28 cycles of PCR were performed.



Figure 4.1



**Figure 4.2**



### **PCR amplification of MMS58**

Genomic DNAs from strains of wild and inbred mice were amplified, and products detected by Southern blot hybridisation. The majority of alleles were very small and composed of 2-5 repeats. However, large expansions were detected in strains BALB/c, 22MO, and BFM. Alleles from *Mus spretus* strains of mice were also larger than in most strains of *Mus musculus*, composed of 9 repeat units.

Alleles were amplified with flanking primers 58-C and 58-D from 20 ng of genomic DNA at 96°C for 40 s, 55°C for 30 s, 70°C for 2 min for 21 cycles. PCR products were electrophoresed through a 40 cm 1.5% agarose gel, Southern blotted, and hybridised. To amplify alleles to levels detectable on ethidium bromide-stained gels, 27 cycles of PCR were performed.

sequencing. The results of this sequence analysis are presented in Figure 4.3. In contrast to the sequence from the MMS57 cosmid clone, 4 variant repeats were detected after partial sequencing of only 4 alleles. The two most common repeat types, designated A and G, were present in each allele analysed. At MMS58, within the 9 alleles sequenced only two variant repeats, designated C and T, were identified and were identical to those detected in the cosmid clone. Names of the two most common variant repeats at each locus were determined by the nucleotide identity at the variant site (Figure 4.3).

### **Computer analysis of DNA conformation**

#### **MMS57**

The two common variant repeats of MMS57, A and G (Figure 4.3), were analysed for their potential to form secondary DNA conformations using the stemloop (Devereux *et al.*, 1984) and Mfold (SantaLucia, 1998) algorithms. The stemloop program revealed a difference in the potential of the two variants to form hairpins, with the A-type repeat forming a stemloop involving 14/24 nucleotides per repeat unit, compared with 16/24 nucleotides for the G-type repeat (Figure 4.4).

Sequences of 10 identical repeats from each strand of each of the two variants were analysed using the Mfold software (Figure 4.5a). Neither strand of the A-type repeat displayed any propensity for hairpin formation. However, complex structures were formed from the G-type repeat, most notably for the purine-rich (67% purine) strand. For 10 repeat units, the predicted free energy for this structure was  $\Delta G = -23.5$ . In contrast to analysis by the GCG program, this predicted hairpin structure involved only 14/24 nucleotides within each repeat with the discrepancy most likely attributable to the NN factors and steric hindrance taken into account by this more sophisticated form of analysis.

To evaluate whether a DNA conformation with an associated  $\Delta G = -23.5$  could potentially have an impact on mutation dynamics, Mfold analysis was performed on each strand of each of the two common repeat units from the FRA16B locus (Yu *et al.*, 1997) (Figure 4.5b). Sequences representing  $7\frac{9}{33}$  repeats at FRA16B (total length identical to 10 repeats of MMS57) were analysed. Structural analysis with Mfold demonstrates that each strand of

**Figure 4.3**

**(a) MMS57**

Variant	Sequence
G	CA <u>CT</u> CTACTG <u>CT</u> CTGTCTACACAG
A	CA <u>CT</u> CTACTG <u>CT</u> CTGTCTACACAA
D	CA--CTACTG <u>CT</u> CTGTCTACACAA
E	<u>ACT</u> CTACTG <u>TT</u> CTGTCTACACAA

Strain	Variant repeat distribution 5' -> 3'
C3H	GAAGGGG . . .
MBK	GAAAGGG . . . GGGGGGGGE
C57BL/6	GAGGGAA
DOT	GAAGGGAADAA

**(b) MMS58**

Variant	Sequence
C	<u>C</u> CTAACCAAGTGAGCATGTATCAACTA
T	<u>T</u> CTAACCAAGTGAGCATGTATCAACTA

Strain	Variant repeat distribution 3' -> 5'
BALB/c	. . TTTTCCTTTTTTCTTTTT
22MO	CTTTTCCCTTTTTTCTTTTT
DBV	CTTTTCTTT
DGA	CTTCTTT
DDO	CCTCCT
C57BL6	CTTTT
DJO	CTTT
DOT	CTT
MDH	CT

***Minisatellite sequences from inbred and wild strains of mice***

Sequence data were generated from the repeat arrays of 4 alleles of MMS57 (Figure 4.3a) and 9 alleles of MMS58 (Figure 4.3b). Four and two variant repeats were identified for MMS57 and MMS58 respectively, the sequence identities of which are presented. Sites polymorphic between variant repeats are underlined. Repeat-type distribution within sequenced alleles is represented as codes of variant repeats. Dots represent breaks in allele code due to incomplete sequence data.

## Figure 4.4

### G-type repeat

CACTCTACTGCTCTGTCTACACAG

```
TCTGTCTACACAGCACTCTACTGCTCTGTCTACACAGCACTCTACTGCTCTGTC
|  | | |  | | | | | | | | | | | | | | | | | | | | | | | | | | | |
ACACATCTGTCTCGTCATCTCACGACACATCTGTCTCGTCATCTCACGACACAT
```

16/24 base pairs formed per repeat

### A-type repeat:

CACTCTACTGCTCTGTCTACACAA

```
TCTGTCTACACAACACTCTACTGCTCTGTCTACACAACACTCTACTGCTCTGTC
|  | | |  | | | | | | | | | | | | | | | | | | | | | | | | | | | |
ACACATCTGTCTCGTCATCTCACAAACACATCTGTCTCGTCATCTCACAAACACAT
```

14/24 base pairs formed per repeat

### ***Stemloop analysis of MMS57 variant repeats***

Arrays of (G)<sub>n</sub> and (A)<sub>n</sub> tandem repeats were analysed with the stemloop single-strand DNA folding program within the GCG package of molecular biology software (Devereux *et al.*, 1984). Single copies of the G and A repeats are indicated in grey, and sites polymorphic between the variant repeats underlined. Positions for the putative formation of Watson-Crick base pairs are indicated with vertical lines. G-type repeats may form hairpins involving 16/24 bases per repeat, whilst A-type repeats may form hairpins with 14/24 bases per repeat.

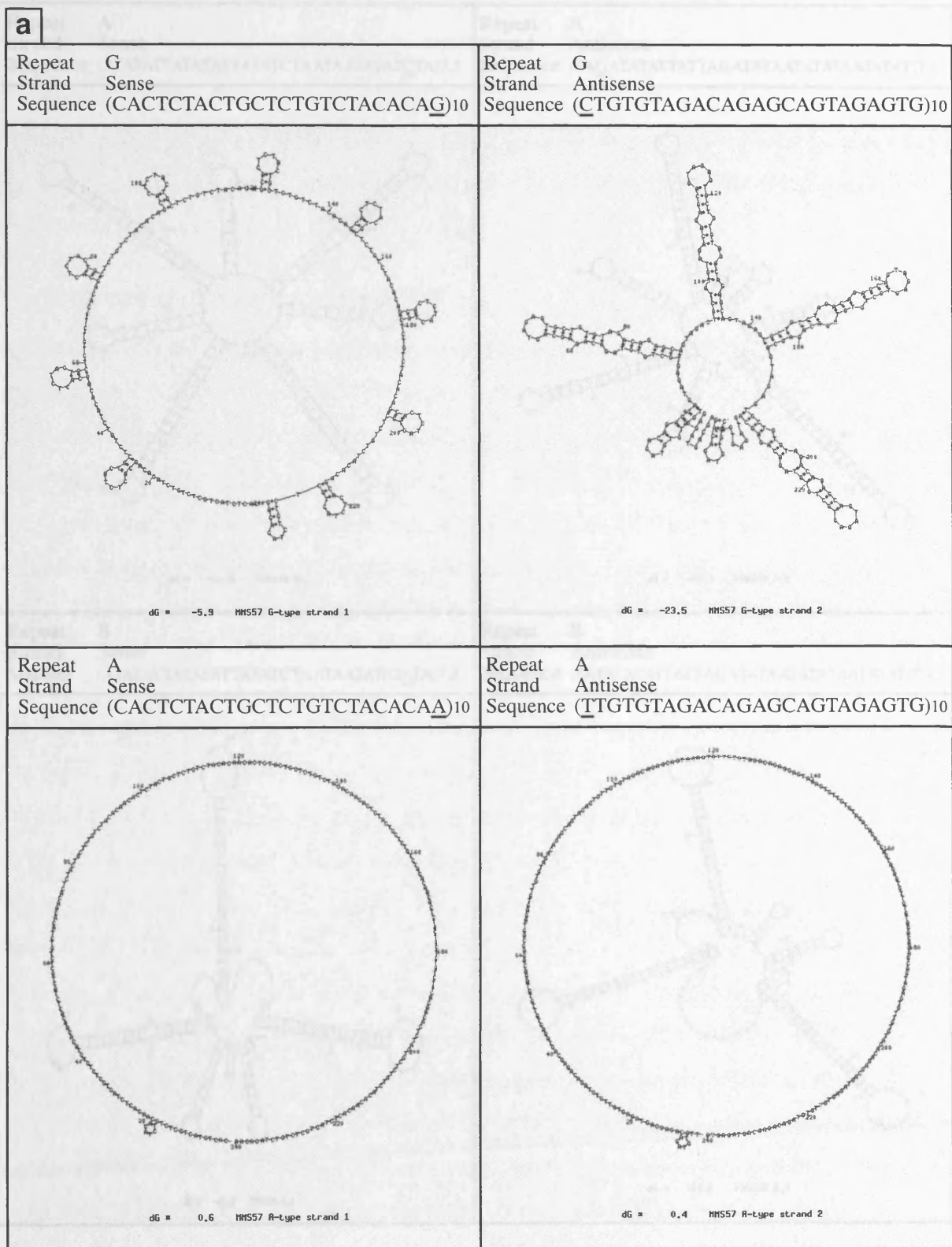
## Figure 4.5

### ***Mfold identification of putative variant repeat DNA conformations***

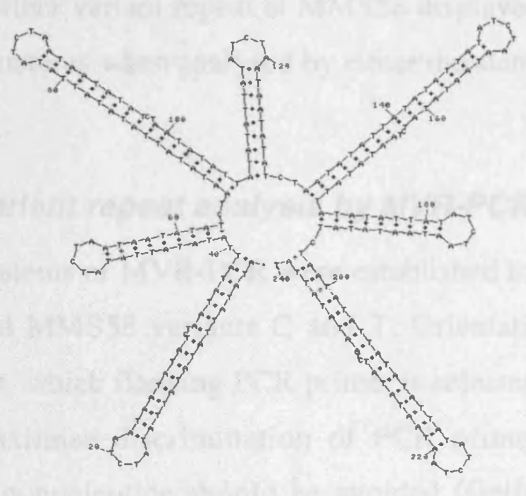
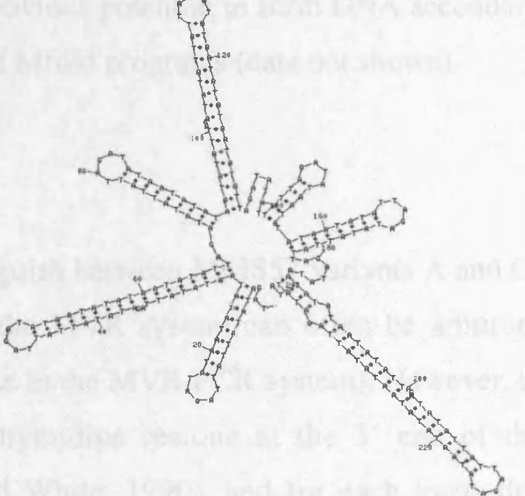
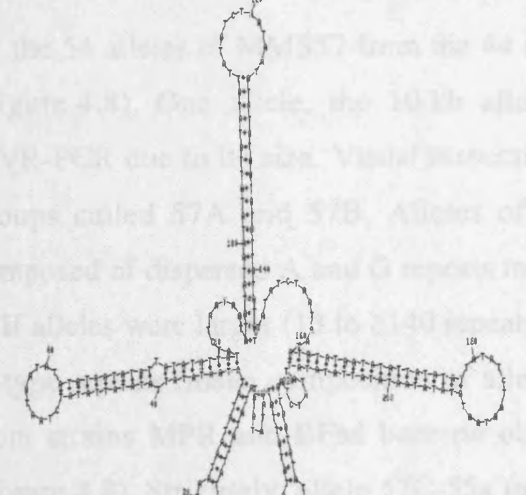
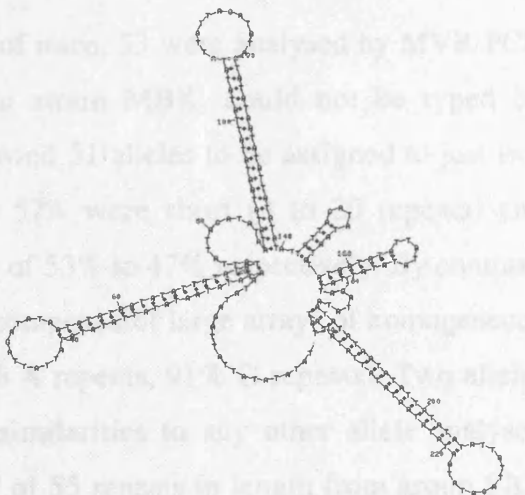
The Mfold program (SantaLucia, 1998), available at <http://mfold.wustl.edu/~mfold/dna/form1.cgi>, uses the nearest-neighbour model for nucleic acid binding which assumes that the stability of a given base pair is dependent on the identity and orientation of neighbouring base pairs (Crothers and Zimm, 1964). Sequences of a single repeat unit, and the (arbitrary) designation of sense and antisense strands are presented. In each case, the structure with the lowest associated free energy ( $\Delta G$ ) is shown.

(a) Ten tandem copies of each strand of each of the two MMS57 variant repeats were analysed under conditions of 37°C, 10 mM Na<sup>+</sup>, 0.5 mM Mg<sup>2+</sup> (Alberts *et al.*, 1989). There is clearly a striking difference between the propensity for G-type repeats to form hairpin structures compared with A-type repeats, most notably within the purine-rich (designated antisense) strand. Neither strand of the A-type repeat formed any secondary conformation, with the exception of a small loop at the start of the sequence which results from end effects associated with the analysis of a linear sequence. (b) Variant repeats of FRA16B (Yu *et al.*, 1997) were also analysed for comparison between the values of  $\Delta G$  associated with MMS57 secondary structures, and the secondary structures of a locus known to undergo dynamic expansions. For FRA16B, 7<sup>9</sup>/<sub>33</sub> (7.3) repeats were analysed to keep the total array lengths analysed (240 bp) consistent between both MMS57 and FRA16B.

**Figure 4.5**



**b**

Repeat Strand Sequence	A Sense (ATATATTATATATTATATCTAATAATATAT <u>CTA</u> )7.3	Repeat Strand Sequence	A Antisense (TAGATATATTATTAGATATAATATATAATATAT)7.3
 <p>dg = -21.8 FRR168 A.1</p>		 <p>dg = -20.3 FRR168 A.2</p>	
Repeat Strand Sequence	B Sense (ATATATTATATATTATATCTAATAATATAT <u>ATA</u> )7.3	Repeat Strand Sequence	B Antisense (TATATATATTATTAGATATAATATATAATATAT)7.3
 <p>dg = -25.3 FRR168 B.1</p>		 <p>dg = -27.2 FRR168 B.2</p>	



each variant is capable of forming secondary structures with a range of  $\Delta G$  from  $\Delta G=-20.3$  to  $\Delta G=-27.2$  (Figure 4.5b). The  $\Delta G$  of the MMS57 putative hairpin lies within this range, consistent with the hypothesis that hairpins may form at this locus and result in dynamic mutation.

## MMS58

Neither variant repeat at MMS58 displayed any obvious potential to form DNA secondary structures when analysed by either the stemloop of Mfold programs (data not shown).

### ***Variant repeat analysis by MVR-PCR***

Systems of MVR-PCR were established to distinguish between MMS57 variants A and G, and MMS58 variants C and T. Orientation of the MVR system can often be arbitrary (i.e. which flanking PCR primer is selected for use in the MVR-PCR system). However, to maximise discrimination of PCR primers, a thymidine residue at the 3' end of the oligonucleotide should be avoided (Gelfand and White, 1990), and for each locus this criterion determined the MVR orientation (Appendix 3). Examples of alleles analysed by MVR-PCR are presented for both MMS57 (Figure 4.6) and MMS58 (Figure 4.7).

### Variant repeat distribution at MMS57

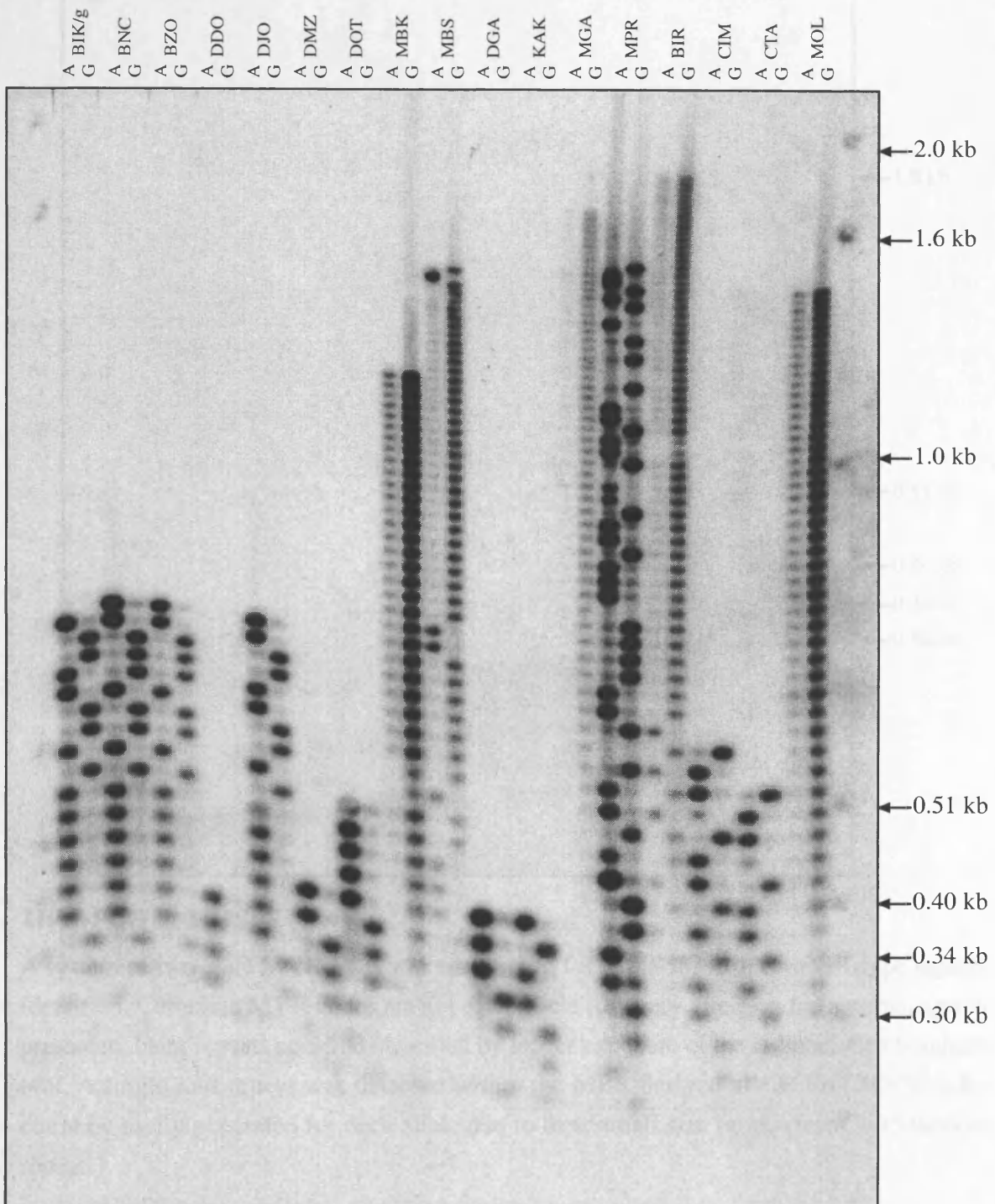
Of the 54 alleles of MMS57 from the 44 strains of mice, 53 were analysed by MVR-PCR (Figure 4.8). One allele, the 10 kb allele from strain MBK, could not be typed by MVR-PCR due to its size. Visual inspection allowed 51 alleles to be assigned to just two groups called 57A and 57B. Alleles of group 57A were short (3 to 20 repeats) and composed of dispersed A and G repeats in a ratio of 53% to 47% respectively. By contrast, 57B alleles were larger (13 to  $\geq 140$  repeats) and composed of large arrays of homogeneous G-type repeats (mean composition of alleles; 9% A repeats, 91% G repeats). Two alleles from strains MPR and BFM bore no obvious similarities to any other allele analysed (Figure 4.8). Strikingly, allele 57C-55a (allele 'a' of 55 repeats in length from group 57C) contained 10 null repeats due to the presence of novel repeat sequence variants preventing annealing of the MVR primer. The greatest number of null repeats in any other allele was 1.

## Figure 4.6

### ***MVR-PCR at MMS57***

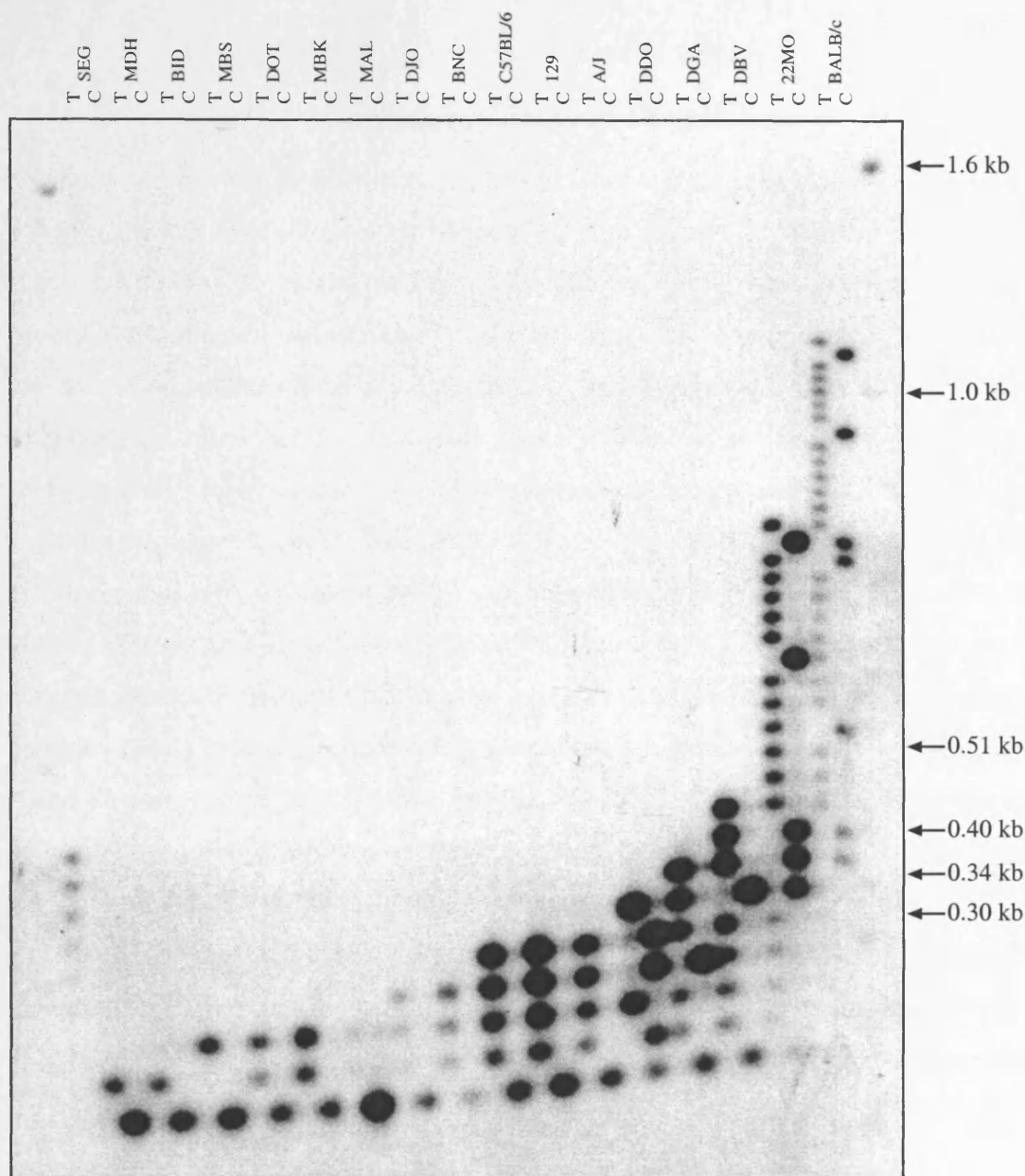
A system of two-state MVR-PCR was established for MMS57, with A- and G-type repeats identified. Perfect discrimination between repeat types was not achieved, but was sufficient to distinguish between the two repeats. In general, small alleles were composed of interspersed repeats whilst large alleles were composed of long uninterrupted arrays of G-type repeats. The presence of repeat types other than the A and G repeats is apparent from either gaps in repeat code (such as the multiple null repeats within the MPR allele) or faint repeats such as two weak G-type repeats (classified as g-typed repeats) within the BIR allele. Up to 140 repeats within an allele could be typed. General MVR-PCR conditions were as described in Chapter 2. To separate alleles in heterozygous mice, samples were amplified to levels visible on ethidium bromide-stained gels as described in Figure 4.1 and bands excised. DNA was gel purified using home-made spin columns and diluted x500 prior to MVR-PCR analysis. For homozygous mice, MVR-PCR was performed directly on stocks of genomic DNA. Samples were amplified with flanking primer 57-C and MVR primers 57-MA or 57-MG at 96°C for 40s, 55°C for 30 s, 70°C for 2 min for 23 cycles. PCR products were electrophoresed through a 40 cm 1.2% agarose gel, Southern blotted, and hybridised.

Figure 4.6



DNA was prepared for MVR-PCR analysis as described in Figure 4.6. The PCR products were amplified with flanking primer 58-D and MVR primers 58-MC or 58-MT at 94°C for 45 s, 55°C for 30 s, 70°C for 1 min for 21 cycles. PCR products were electrophoresed through a 40 cm 1.5% agarose gel, Southern blotted, and hybridised.

**Figure 4.7**



#### ***MVR-PCR at MMS58***

A system of two-state MVR-PCR was established for MMS58, with C- and T-type repeats identified. Complete MVR codes are not discernible for every allele on the autoradiograph presented. Faint repeats could be identified by longer exposure of the radiolabelled Southern blot. A single null repeat was detected within the MBS-derived allele. Full MVR codes could be easily generated for each allele due to their small size range (from 2-35 tandem repeats)

DNA was prepared for MVR-PCR analysis as described in Figure 4.6. Samples were amplified with flanking primer 58-D and MVR primers 58-MC or 58-MT at 96°C for 40s, 55°C for 30 s, 70°C for 1 min for 21 cycles. PCR products were electrophoresed through a 40 cm 1.5% agarose gel, Southern blotted, and hybridised.

## Figure 4.8

### ***MMS57 MVR codes aligned by visual inspection***

MVR codes were read 5' to 3' from the bottom to the top of the autoradiograph. Repeat distribution patterns of all alleles were aligned by eye. G- and A-type repeats are as described in Figure 4.3. Alleles were divided into four groups, 57A-57D. Groups 57C and 57D each contain a single representative allele. Allele names denote allele group, repeat number, and a discriminating letter. The strain from which each allele was derived is indicated. (s) and (l) indicates the small and large alleles from heterozygous mice respectively. The *Mus* species or subspecies is described as: bact., *Mus m. bactrianus*; cast., *Mus m. castaneus*; dom., *Mus m. domesticus*; mol., *Mus m. molossinus*; mus., *Mus m. musculus*; spr., *Mus spretus*; I, Inbred strain. MVR primers at this locus only bind to a 17 bp region within the 24 bp repeat unit. Consequently, variants outside of this 17 bp region will not be detectable. 'o' denotes null repeats (unidentified repeats due to the presence of additional unknown polymorphic sites) whilst 'a' and 'g' denote repeats which are detected only weakly with MVR primers specific for A-type and G-type repeats respectively. This weak signal is assumed to result from novel polymorphisms at positions towards the 5' end of the repeat-specific region of the MVR primer. >> indicates incomplete MVR code generated for an allele, and ? denotes repeat types which were untyped due to a weak signal resulting from conditions of MVR-PCR (the start of MVR codes is often faint from larger alleles, see Figure 4.6). Hyphens were inserted to improve MVR code alignments.



### Figure 4.8

[illegible]

## Variant repeat distribution at MMS58

All 48 alleles of MMS58 derived from 44 strains of mice were analysed by MVR-PCR (Figure 4.9). Visual inspection divided all alleles into three groups, 58A, 58B, and 58C. Of the 48 alleles, 40 were within groups 58A and  $\leq 5$  repeats in length. With the single exception of allele 58A-3b (allele 'b' of 3 repeats in length from group 58A) which contained a null repeat, all differences between 58A alleles were attributable exclusively to the number of tandemly repeated T-type repeats. Group 58B contained the largest alleles ranging from 7 to 35 repeats. Structurally, they were similar to 58A alleles being apparently composed of imperfect duplications of entire 58A arrays. MMS58 alleles from *Mus spretus* were longer than in most strains of *Mus musculus* and were composed entirely of homogeneous T-type repeats, so were assigned to a distinct allelic group, 58C.

## Problems with allele alignment and subdivision

A consistent deficiency in the analysis of data generated by MVR-PCR is the criteria used to subdivide alleles into groups. Most commonly, alleles are divided by visual inspection. However, this approach is subjective as it requires decisions to be made as to what differences between alleles are sufficient to generate novel allele subgroups. This difficulty can be clearly illustrated with the subdivisions of alleles of both MMS57 and MMS58 described above. For MMS57, the decision to include alleles 57B-13a and 57B-15a within group 57B is certainly questionable as alignments between these alleles and alleles of group 57A are possible. Their inclusion into group 57B was based on the presence of a 3' G repeat, and the regular interspersion of G and A repeat types at the 5' end of alleles which is common to 57B alleles but is generally only observed towards the 3' of 57A alleles. Similarly, for alleles of MMS58, 58B-9a and 58B-7a could be included within group 58A if the difference between groups 58A and 58B was defined as the presence of (C)<sub>22</sub> tandem repeats, as opposed to the difference between 1 and >1 C-type repeats. Alternative strategies to subdivide and align MVR codes will be considered in Chapter 5.

## Figure 4.9

### ***MMS58 MVR codes aligned by visual inspection***

MVR codes were read 3' to 5' from the bottom to the top of the autoradiograph. Repeat distribution patterns of all alleles were aligned by eye. C- and T-type repeats are as described in Figure 4.3. Alleles are divided into three groups, 58A-58C. Allele names denote allele group, repeat number, and a discriminating letter. The strain from which each allele was derived is indicated. (s) and (l) denote the small and large alleles from heterozygous mice respectively. The *Mus* species or subspecies is as described in Figure 4.8. MVR primers at this locus only bind to a 17 bp region within the 27 bp repeat unit. Consequently, variants outside of this 17 bp region will not be detectable. 'o' denotes null repeats (unidentified repeats due to the presence of additional unknown polymorphic sites). Hyphens were inserted to improve MVR code alignments.



**Figure 4.9**

Strain	Species	Allele	MVR code (3'→5' orientation)
129	I	58A-5a	CTTTT
AKR	I	58A-5a	CTTTT
A/J	I	58A-5a	CTTTT
SJL	I	58A-5a	CTTTT
SWR	I	58A-5a	CTTTT
C3H	I	58A-5a	CTTTT
C57BL/6J	I	58A-5a	CTTTT
CBA/C	I	58A-5a	CTTTT
MF1	I	58A-5a	CTTTT
CBA/J	I	58A-5a	CTTTT
DBA	I	58A-5a	CTTTT
NIH	I	58A-5a	CTTTT
DMZ (l)	dom.	58A-5a	CTTTT
FVB	I	58A-5a	CTTTT
MPR (l)	mus.	58A-4a	CT-TT
BIK/g	dom.	58A-4a	CT-TT
BNC	dom.	58A-4a	CT-TT
BZO	dom.	58A-4a	CT-TT
DJO	dom.	58A-4a	CT-TT
BIR	bact.	58A-3a	CT--T
CIM	cast.	58A-3a	CT--T
CTA	cast.	58A-3a	CT--T
DOT	dom.	58A-3a	CT--T
MPB	mus.	58A-3a	CT--T
MAM	mus.	58A-3a	CT--T
MBK	mus.	58A-3a	CT--T
MOL	mol.	58A-3a	CT--T
MGA (l)	mus.	58A-3a	CT--T
MGL	mus.	58A-3a	CT--T
MAC	mus.	58A-3a	CT--T
DMZ (s)	dom.	58A-3a	CT--T
KAK (l)	mus.	58A-3a	CT--T
MGT	mus.	58A-3a	CT--T
MBS	mus.	58A-3b	C-o-T
BID	mus.	58A-2a	C---T
KAK (s)	mus.	58A-2a	C---T
MDH	mus.	58A-2a	C---T
MGA (s)	mus.	58A-2a	C---T
MPR (s)	mus.	58A-2a	C---T
TEH	mus.	58A-2a	C---T
DBV	dom.	58B-9a	CTTTTC-TTT
DGA	dom.	58B-7a	CTT--C-TTT
22MO	dom.	58B-22a	CTTTTCC-----CTTTTTCTTTTTCT
BFM	dom.	58B-22a	CTTTTCC-----CTTTTTCTTTTTCT
BALB/c	I	58B-35a	CCTTTTCCTTCTTTTTTTCTTTTTCTTTTTCT
DDO	dom.	58B-6a	CC---TCCT
SEG	spr.	58C-9a	TTTTTTTTT
SPE	spr.	58C-9a	TTTTTTTTT

## Discussion

### ***MMS57 may undergo dynamic expansions***

The wide range of allele sizes, absence of variant repeats, and potential for the formation of hairpin structures within the repeat array as described in Chapter 3, suggested that MMS57 may undergo a form of dynamic mutation similar to triplet repeats (Mitas, 1997) or FRA16B (Yu *et al.*, 1997). If dynamic expansions had occurred at MMS57, the formation of complex DNA secondary structures and the presence of homogeneous arrays of repeats would be predicted within large alleles. Both predictions were tested in this chapter and found to be consistent with mutation by dynamic expansion.

Sequence analysis of MMS57 identified two common variant repeats, A and G, which differed by a single base transition. Simulations of single-strand DNA structures predicted to form under physiological conditions demonstrated that complex hairpins could form exclusively from G-type repeats, and primarily within the purine-rich strand. The thermodynamic stability of the putative hairpin was comparable to that of the FRA16B AT-rich minisatellite which undergoes dynamic expansions *in vivo* to form a fragile site. The strand asymmetry observed in the potential for secondary structure formation at the MMS57 G-type repeat indicates that if hairpin formation does mediate dynamic expansion, it is likely to do so only if the purine-rich strand is the nascent strand during DNA replication (hairpin formation in the template strand would most likely result in deletion mutations). If instability was due to hairpin formation in Okazaki fragments leading to resistance to FEN-1 digestion (Chapter 1; Lieber (1997)), levels of instability may be at least partially dependent on the orientation of DNA replication. However, whether any such secondary structures are capable of forming at MMS57 *in vivo*, or affect mutational mechanisms, is unknown. Both components of chromatin, and single strand DNA binding proteins, may interfere with formation of the predicted hairpin structure.

Dynamic mutation is thought to be mediated by, and result in, homogeneous arrays of repeat units with the potential to form secondary conformations (Mitas, 1997). If dynamic instability occurred at MMS57, long homogeneous arrays of G-type repeats would be predicted within large minisatellite alleles. Similar arrays of A-type repeats would not be

expected to form. MVR-PCR analysis demonstrated that precisely this pattern of variant repeat distribution is found. The largest uninterrupted array of A-type repeats was only 6 repeats long, and found within many 57A alleles. These short alleles displayed patterns of variant repeat interspersions characteristic of most minisatellites studied to date. In contrast, the largest homogeneous array of G-type repeats successfully typed was 90 repeats long (~2.2 kb) within allele 57B-140a from strain MDH. Similar (G)<sub>n</sub> arrays were observed in many strains of *Mus m. bactrianus*, *Mus m. molossinus*, and *Mus m. musculus* mice, but were largely absent from *Mus m. domesticus*, *Mus m. castaneus*, and *Mus spretus*. The 5' repeat interspersion patterns observed in alleles with long 3' homogeneous G-type repeat arrays are similar to the distribution of variant repeats observed at expanded alleles of the FRA16B minisatellite (Yu *et al.*, 1997), further supporting a common mutational process operating at these two loci.

Allele 57C-55a from the *Mus m. musculus* MPR strain of wild mice contained 10-fold more null repeats than any other allele analysed. The allele is likely to represent an ancient lineage highly diverged from the other lineages detected, and which was present at a low frequency in the mice analysed. High levels of repeat sequence divergence may have been generated by point mutation within the lineage resulting in the accumulation of novel repeat types. A more probable explanation (due to the low frequency of point mutation) is that a single novel variant repeat arose and spread throughout the allele by intra-allelic duplication.

The presence of a high frequency of null repeats within a single allele lineage is, at least superficially, reminiscent of a recent study of the minisatellite MSY1 (*DYF155S1*) (Bouzekri *et al.*, 1998). MSY1 is the only known polymorphic minisatellite located on the Y chromosome (Bouzekri *et al.*, 1998). It consists of ~50-115 repeats of an AT-rich 25 bp unit which is predicted to form stable hairpin structures and displays a mutation rate of 2-11% per generation (Jobling *et al.*, 1998). Within an African-specific Y chromosome lineage (haplogroup 8), MVR-PCR analysis identified a high frequency of null repeats. Sequence analysis demonstrated that these null repeats were identical to variants detected in other lineages, with the exception of a single base transition mutation which had become homogenised throughout the repeat array (Bouzekri *et al.*, 1998). This homogenisation was

position-specific within the repeat unit as other variant sites within repeats, common to variants in other lineages, were not homogenised (Bouzekri *et al.*, 1998). It was suggested that this transition had spread throughout the allele by either slippage replication or unequal sister chromatid exchange, with mispaired intermediates within heteroduplexes repaired in a biased manner, and with heteroduplex repair restricted to a single position within the repeat unit (Bouzekri *et al.*, 1998).

While a similar process of hairpin-mediated base-specific homogenisation may have resulted in the high frequency of null repeats identified in allele 57C-55a, the MSY1 repeats containing the novel variants occur within uninterrupted blocks of repeats, indicative of the linear diffusion of the novel variant between adjacent repeats (M. Jobling, pers. commun.). In contrast, null repeats within 57C-55a are generally separated by known repeat types. While this suggests that at MMS57, the spread of null repeats is driven by a different mechanism than that operating at MSY1, it is possible that differences in hairpin conformation between the loci result in homogenisation at MMS57 between non-adjacent repeat units brought into apposition by hairpin formation. The apparent restriction of this high frequency of null variants to one allele is indicative of a low frequency of inter-allelic mutation processes operating at MMS57.

On balance, current evidence favours the operation of dynamic mutation processes at MMS57. However, the predicted germline mutation rate at this locus is low. For example, identical alleles containing 48 perfect tandem G-type repeats were identified in the inbred strains A/J, BALB/c, and C3H, and in the *Mus m. molossinus* strain MOL. Furthermore, the analysis of this locus in BXD RI strains failed to identify any new mutant alleles, demonstrating the mutation rate to be  $<10^{-3}$  ( $p>0.95$ ) (Bois *et al.*, 1998a).

It is currently not possible to demonstrate beyond doubt that the putative MMS57 hairpin forms *in vivo*, or affects mutation dynamics at the locus. However, further support could be derived from two sources. The first would be the detailed characterisation of *in vitro* DNA conformation of both A-type and G-type repeats by, for example, nuclear magnetic resonance (NMR) analysis (e.g. Catasti *et al.* (1996)). The second would involve mutation detection and analysis of a variety of alleles at MMS57. Mutation studies of two alleles

with similar sizes but very different internal structures such as 57C-55a and 57B-51b (Figure 4.8) would allow mutation rates between the alleles to be compared. If long regions of homogeneous G-type repeats induce mutation, a higher mutation rate would be expected for 57B-51b compared with 57C-55a. However, if such a difference in mutation rate was detected, alternative explanations for this difference such as a role for flanking sequences would be possible. A similar approach would be to compare mutation rates between alleles containing short and long (G)<sub>n</sub> repeat arrays. Mutation rate would be predicted to rise with allele size if mutation induction was a property of the repeat array.

### ***MMS58 may be a recently expanded minisatellite***

Whilst variation in allele sizes between strains was consistent with the operation of dynamic mutation processes at MMS58, the absence of any apparent DNA secondary conformations, and the interspersed patterns of variant repeat units within the largest alleles, do not support this model for minisatellite instability. Instead, some MMS58 alleles may have undergone relatively recent expansion. All of the larger MMS58 alleles (>5 repeats) are within lineages 58B and 58C. With the exception of the BALB/c inbred strain, groups 58B and 58C include alleles exclusively from strains of *Mus m. domesticus* and *Mus spretus* respectively. The shorter alleles (group 58A) are found in strains of *Mus m. domesticus*, *Mus m. musculus*, *Mus m. castaneus*, *Mus m. bactrianus*, and *Mus m. molossinus*. It can therefore be concluded that either the ancestral state of MMS58 was an expanded form which collapsed in *Mus m. domesticus*, *Mus m. musculus*, *Mus m. castaneus*, *Mus m. bactrianus*, and *Mus m. molossinus*, or that the ancestral alleles were small and expanded in both *Mus spretus* and in some strains of *Mus m. domesticus*. Analysis of variant repeat distribution indicates that the putative *Mus spretus* expansion may have been independent from the putative *Mus m. domesticus* expansion, as variant repeat distribution differs substantially between 58B and 58C alleles, consistent with the divergence of *Mus m. domesticus* from *Mus spretus* earlier than the divergence of *Mus m. domesticus* from the other *Mus musculus* subspecies of mice analysed (Bonhomme and Guénet, 1996).

Whilst it is perhaps more likely that MMS58 expanded in the two mouse lineages *Mus spretus* and *Mus m. domesticus* than collapsed in most strains from all five of the *Mus*

*musculus* subspecies, confirmation of the ancestral state would require the analysis of MMS58 alleles in a range of other diverged species of *Mus* such as *Mus macedonius* and *Mus caroli*. Similarly, the short alleles of mouse minisatellites MMS24, MMS26, MMS57, and MMS80 characterised in *Mus spretus* were assumed to represent ancestral states of the loci each of which are both expanded and polymorphic between strains of *Mus musculus* (Bois *et al.*, 1998a), although it is possible that the ancestral state was the expanded form which collapsed in strains of *Mus spretus*.

It is not possible to make any substantive inferences as to the mutation dynamics of MMS58 due to the low level of informativity of the 2-state system of MVR-PCR, the short lengths of most alleles, and the lack of qualitative differences between repeat-type composition in the majority of alleles. However, the apparent expansions generating alleles of group 58B may have resulted from relatively complex intra-allelic duplications which, due to the distribution of C-type variants centrally within 58B alleles, would have incorporated the 3' region of a putative 58A progenitor repeat array.

Finally, it is possible that mouse minisatellites showing relatively recent expansions could be more prevalent than would be expected from the allele size distributions of the loci described in Chapter 3. Most mouse minisatellites characterised to date were derived ultimately from the DNA of a BALB/c mouse. For MMS58, this was the only inbred strain of *Mus musculus* analysed in this study with alleles larger than 5 repeats (135 bp), and had the largest allele identified to date (35 repeats, 945 bp). Other loci which display similar expansions in strains other than BALB/c would not have been isolated. This strain-specific bias towards expanded alleles could be avoided by screening a cosmid library derived from multiple diverged strains of *Mus musculus*.

## Chapter 5

# The application of MVR-PCR to phylogenetic analysis

### Summary

MVR-PCR analysis has been carried out at a total of five mouse minisatellites in wild mice from each of the five *Mus musculus* subspecies, and in a range of inbred strains. Alleles at each locus may be divided by eye into lineage groups. An alternative method of allele subdivision using MultAlin software (Corpet, 1988) was investigated, but found to be inappropriate for MVR data. Lineage data from each locus and each mouse strain were simultaneously analysed by multidimensional scaling (MDS) to test for consistent genetic similarities and differences between strains. Four of the five *Mus musculus* subspecies had very similar minisatellite allele lineages, but differed substantially from *Mus m. domesticus*. Inclusion of MVR data derived from inbred strains of mice demonstrated that the major genetic component of the inbred strains was from the *Mus m. domesticus* subspecies. Mouse MVR data can therefore be analysed on two different levels; simultaneous analysis of multiple loci provides a general overview of the genetic similarities and differences between strains, whilst the consideration of single loci in isolation indicates patterns of gene flow between populations.

### Introduction

The analysis of polymorphisms in mitochondrial DNA (Boursot *et al.*, 1996) and proteins (Din *et al.*, 1996) have indicated that the *Mus musculus* species originated from within the Indian subcontinent (Bonhomme and Guénet, 1996). Geographical separation led to the divergence of subspecies from the *Mus musculus* ancestor, creating the subspecies *Mus m. bactrianus*, *Mus m. castaneus*, *Mus m. domesticus*, *Mus m. molossinus*, and *Mus m. musculus*. However, these subspecies exchange genes whenever they come into contact, hence their designation as subspecies as opposed to distinct species (Bonhomme and Guénet, 1996). For example, genetic exchanges have been characterised in some detail

between European populations of *Mus m. domesticus* and *Mus m. musculus* (Bonhomme *et al.*, 1983; Selander and Yang, 1969), and between Japanese populations of *Mus m. musculus* and *Mus m. castaneus* (Moriwaki, 1987; Yonekawa *et al.*, 1988). *Mus spretus* represents a distinct species which diverged from *Mus musculus* 1-2 million years ago (Bonhomme and Guénet, 1996). The two species can be hybridised, but F1 males exhibit sterility hence the classification of *Mus spretus* as a distinct species (Copeland *et al.*, 1993). Inbred strains of laboratory mice are thought to have been developed with contributions from more than one species/subspecies of wild mice. For example, some strains carry the *Mus m. domesticus* Y chromosome, whilst others carry the *Mus m. musculus* Y chromosome (Bonhomme and Guénet, 1996). However, there is no record of the original breeding patterns by which most laboratory strains were ultimately derived from wild mice stocks (Festing, 1996).

MVR-PCR at mouse minisatellites MMS24, MMS26, and MMS80 has been described elsewhere (Bois *et al.*, 1998a). It was found that closely related alleles were usually derived from the same *Mus musculus* subspecies. Similar alleles were also identified in different subspecies. For example *Mus m. molossinus* alleles of MMS24 were structurally similar to those of *Mus m. musculus*, and MMS80 alleles from *Mus m. bactrianus* were closely related to *Mus m. musculus* alleles (Bois *et al.*, 1998a). However, phylogenetic analysis based on individual loci in a small number of mice from a limited number of wild mice populations is open to sampling effects. A more robust method to determine genetic differences between different populations would involve the simultaneous analysis of multiple loci in a variety of strains.

Variant repeat distribution has now been analysed at a total of five mouse minisatellites, MMS24, MMS26, MMS57, MMS58, and MMS80, in a range of inbred and wild strains of mice (Chapter 4; Bois *et al.* (1998a)). These loci are distributed across 4 autosomes; chromosomes 7, 9, 8, 3, and 9 respectively (Chapter 3). Within this chapter, patterns of variant repeat distribution at all loci are combined to provide a picture of the overall genetic similarities and differences between different strains. This work had two objectives. The first was to detect genetic difference between mice from different *Mus musculus*



subspecies. The second was to use any potential differences identified to attempt to define which subspecies of mice were the dominant founders of a range of inbred strains.

The approach to this analysis is simple. Alleles of each locus can be divided into subgroups based upon similarities and differences between MVR codes. These allele subdivisions are performed blind (i.e. without reference to the strain from which each allele mapped by MVR-PCR was derived). The assignation of a number to each lineage group generates a multi-state matrix including data from each of the five minisatellites for each mouse strain analysed. Similarities and differences between each strain may then be determined by multidimensional scaling (MDS) analysis of the multi-state matrix. MDS here uses pairwise comparisons between each strain and across all loci to determine the levels of divergence between strains. This divergence is converted to a distance which may be displayed graphically. The combination of the distances between every pairwise combination of strains converts the multi-state matrix to an N-dimensional graph (N may range from 1 to n-1 where n is the number of strains analysed) from which genetically similar strains may be observed as clusters.

The main problem with this approach lies in the criteria used to subdivide alleles into lineage groups. Typically, alleles typed by MVR-PCR are aligned by eye and then divided into groups. This is clearly a subjective approach to allele division as the criteria for the extent to which similarities and difference must exist between alleles for them to be assigned to different groups is a matter of personal judgement. In addition, short alleles may by chance display regions of similarity to alleles of several groups, and again the assignation of such alleles to specific lineages would be questionable. An alternative approach, described by Jeffreys *et al.* (1991a) was to use the dotplot program within the GCG software package (Devereux *et al.*, 1984) to perform each combination of pairwise alignments of MVR codes. The degree of similarity between different codes may then be used to determine allele groups. However, the definition of the degree of similarity required for two alleles to be grouped together was again subjective. One program which has previously been successfully applied to the alignment of MVR codes at mouse minisatellites (P. Bois, pers. commun.) is the MultAlin program developed by F. Corpet (INRA, Toulouse, France) (Corpet, 1988) and available at

<http://www.toulouse.inra.fr/multalin.html>. This program was designed for the alignment of sequences of both nucleic acids and proteins, and employs pairwise alignments of multiple sequences to establish hierarchical subgroupings of similar sequences, which are then aligned into a single group. Its application to MVR data will be assessed in this chapter.

## Results and Discussion

### ***Subdivision of alleles using MultAlin software***

MVR codes of both MMS57 (data not shown) and MMS58 (Figure 5.1) were analysed with MultAlin software. From Figure 5.1, it is clear that the software is unsuitable for the alignments of MVR codes. For example, MVR codes from alleles in strains 22MO and BFM are identical, yet the BALB/c allele is positioned between the two codes despite a difference in length between the alleles of 13 repeats. Furthermore, several examples are readily apparent where identical alleles (such as the CTT alleles of strains DOT and BIR, or the (T)<sub>9</sub> alleles of the *Mus spretus* strains SEG and SPE) are presented with different MVR code alignments. Similar inappropriate alignments were obtained for analysis of MVR codes of MMS57 (data not shown).

There appear to be two main features of the program which make it unsuitable for its application to MVR codes. The first is the implicit assumption that all sequences are meaningfully alignable. Whilst it may be assumed that all alleles at a given locus arose from a common progenitor, many of the intermediate alleles between the progenitor and the present population of alleles are likely to have been lost by genetic drift. Allele lineages may therefore be so distantly diverged that meaningful alignments of all alleles are not possible. The second deficiency is that all alignments are based on positive criteria. The similarity between two sequences is defined by the total number of matches between sequence units, whilst the number of mismatches is not considered as a contributing factor to the alignments. While it may be possible to modify the program to incorporate penalties for both mismatches and gaps introduced to improve alignments, any decision of the relative importance of different forms of variation between two codes will be either completely subjective, or based on empirical evidence obtained at loci previously investigated by mutation analysis. Unless mutation has been analysed at the locus of

## Figure 5.1

### ***MMS58 MVR codes aligned by MultAlin and by eye***

The MultAlin program (Corpet, 1988) uses pairwise comparisons of sequence codes to align and group alleles based on the number of matching sequence units between alleles. Sequence mismatches and gaps are not penalised. MVR codes of MMS58 were aligned using MultAlin. A comparison with alleles aligned by visual inspection (as previously depicted in Figure 4.9) is presented. MultAlin does not produce reasonable alignments or subgroupings of alleles. Whilst this conclusion is a subjective judgement of a method used to generate objective subgroupings, the presence of identical alleles such as C(T)<sub>2</sub> and (T)<sub>9</sub>, which are neither grouped together nor presented with identical alignments, clearly supports this judgement.

# Figure 5.1

Aligned by MultAlin

Strain	Species	Allele	MVR code
129	I	58A-5a	C-TT-----TT
AKR	I	58A-5a	C-TT-----TT
A/J	I	58A-5a	C-TT-----TT
SWR	I	58A-5a	C-TT-----TT
SJL	I	58A-5a	C-TT-----TT
C3H	I	58A-5a	C-TT-----TT
C57BL/6J	I	58A-5a	C-TT-----TT
CBA/C	I	58A-5a	C-TT-----TT
MF1	I	58A-5a	C-TT-----TT
CBA/J	I	58A-5a	C-TT-----TT
DBA	I	58A-5a	C-TT-----TT
NIH	I	58A-5a	C-TT-----TT
DGA	dom.	58B-7a	C-TT-----CTT
DMZ (1)	dom.	58A-5a	C-TT-----TT
FVB	I	58A-5a	C-TT-----TT
22MO	dom.	58B-22a	C-TTTTCC---CTTTTTT-C-TTTT-CT
BALB/c	I	58B-35a	CCTTTTCCTTTCTTTTTTTCCTTTTTCTTTTCT
BFM	dom.	58B-22a	C-TTTTCC---CTTTTTT-C-TTTT-CT
SPE	spr.	58C-9a	TTTT-----TTTT
DBV	dom.	58B-9a	CTTT-----C-TTT
SEG	spr.	58C-9a	TTTT-----TTTT
MPR (1)	mus.	58A-4a	C-TT-----T
BIK/g	dom.	58A-4a	C-TT-----T
DJO	dom.	58A-4a	C-TT-----T
BNC	dom.	58A-4a	C-TT-----T
BZO	dom.	58A-4a	C-TT-----T
DOT	dom.	58A-3a	C-TT-----T
MAM	mus.	58A-3a	C-TT-----T
CTA	cast.	58A-3a	C-TT-----T
CIM	cast.	58A-3a	C-TT-----T
MBK	mus.	58A-3a	C-TT-----T
MPB	mus.	58A-3a	C-TT-----T
MAC	mus.	58A-3a	C-TT-----T
KAK (1)	mus.	58A-3a	C-TT-----T
MOL	mol.	58A-3a	C-TT-----T
MBS	mus.	58A-3b	CoT-----T
BIR	bact.	58A-3a	C-T-----C-----CT
DDO	dom.	58B-6a	CCCT-----C-----CT
DMZ (s)	dom.	58A-3a	C-T-----T
MGA (1)	mus.	58A-3a	C-T-----T
MGL	mus.	58A-3a	C-T-----T
MGT	mus.	58A-3a	C-T-----T
BID	mus.	58A-2a	C-T-----T
KAK (s)	mus.	58A-2a	C-T-----T
MDH	mus.	58A-2a	C-T-----T
MGA (s)	mus.	58A-2a	C-T-----T
MPR (s)	mus.	58A-2a	C-T-----T
TEH	mus.	58A-2a	C-T-----T

Aligned by visual inspection

Strain	Species	Allele	MVR code
129	I	58A-5a	CTTTT
AKR	I	58A-5a	CTTTT
A/J	I	58A-5a	CTTTT
SJL	I	58A-5a	CTTTT
SWR	I	58A-5a	CTTTT
C3H	I	58A-5a	CTTTT
C57BL/6J	I	58A-5a	CTTTT
CBA/C	I	58A-5a	CTTTT
MF1	I	58A-5a	CTTTT
CBA/J	I	58A-5a	CTTTT
DBA	I	58A-5a	CTTTT
NIH	I	58A-5a	CTTTT
DMZ (1)	dom.	58A-5a	CTTTT
FVB	I	58A-5a	CTTTT
MPR (1)	mus.	58A-4a	CT-TT
BIK/g	dom.	58A-4a	CT-TT
BNC	dom.	58A-4a	CT-TT
BZO	dom.	58A-4a	CT-TT
DJO	dom.	58A-4a	CT-TT
BIR	bact.	58A-3a	CT--T
CIM	cast.	58A-3a	CT--T
CTA	cast.	58A-3a	CT--T
DOT	dom.	58A-3a	CT--T
MPB	mus.	58A-3a	CT--T
MAM	mus.	58A-3a	CT--T
MBK	mus.	58A-3a	CT--T
MOL	mol.	58A-3a	CT--T
MGA (1)	mus.	58A-3a	CT--T
MGL	mus.	58A-3a	CT--T
MAC	mus.	58A-3a	CT--T
DMZ (s)	dom.	58A-3a	CT--T
KAK (1)	mus.	58A-3a	CT--T
MGT	mus.	58A-3a	CT--T
MBS	mus.	58A-3b	C-o-T
BID	mus.	58A-2a	C---T
KAK (s)	mus.	58A-2a	C---T
MDH	mus.	58A-2a	C---T
MGA (s)	mus.	58A-2a	C---T
MPR (s)	mus.	58A-2a	C---T
TEH	mus.	58A-2a	C---T
DBV	dom.	58B-9a	CTTTTCTTT
DGA	dom.	58B-7a	CTT--C-TTT
22MO	dom.	58B-22a	CTTTTCC-----CTTTTTTCTTTTCT
BFM	dom.	58B-22a	CTTTTCC-----CTTTTTTCTTTTCT
BALB/c	I	58B-35a	CCTTTTCCTTTCTTTTTTTCCTTTTTCTTTTCT
DDO	dom.	58B-6a	CC---TCCT
SEG	spr.	58C-9a	TTTTTTTTT
SPE	spr.	58C-9a	TTTTTTTTT

interest, this would introduce the assumption that mutation processes operating at all minisatellites are the same.

### ***Multidimensional scaling analysis using five mouse minisatellites***

In the apparent absence of alternatives, alleles at MMS24, MMS26, MMS57, MMS58, and MMS80, were divided into lineage groups by eye as described by Bois *et al.* (1998a) and in Chapter 4. A number was assigned to each group (group A becomes 1, etc.) and a matrix constructed of MVR code lineages for each strain (Table 5.1). Twenty-three strains had been typed at all 5 minisatellites generating 21 different combinations of alleles. These 21 strains represented 7 inbred strains, 7 *Mus m. domesticus* wild strains, and 1-2 wild strains each of *Mus m. bactrianus*, *Mus m. castaneus*, *Mus m. molossinus*, *Mus m. musculus*, and the *Mus spretus* species. The results of MDS analysis of the multi-state matrix are presented in Figure 5.2. All MDS analysis was performed by Y. Dubrova.

The distribution of points on the graph in Figure 5.2 along dimension 1 represents the greatest component of the observed variation between strains. A clear distinction was identified between minisatellite allele lineages present in *Mus m. domesticus* compared with all other wild strains of mice as is apparent from the gap between the two groups in dimension 1 (Figure 5.2). The relatively diffuse cluster of *Mus m. domesticus* strains generated by MDS analysis indicates much higher levels of genetic heterogeneity between these wild strains than between each of the other four *Mus musculus* strains analysed. The high levels of similarity between four of the five *Mus musculus* subspecies may be interpreted in two ways. Either the four subspecies diverged more recently than all five subspecies, or that more recent interbreeding occurred between the subspecies resulting in their genetic similarities. With the current data, it is not possible to differentiate between these two hypotheses. However, expansion of the study to incorporate more loci and more wild mice from a diversity of populations might allow levels of allele diversity both within and between subspecies to be compared. If the similarities between subspecies were due to a more recent common origin after which relatively little interbreeding occurred, diversity of minisatellite allele lineages within a subspecies would be much lower than the diversity between subspecies. However, if the similarities were due to recent interbreeding, a range

## Table 5.1

### ***Matrix of MVR code lineages analysed by MDS***

Five minisatellites have been analysed by MVR-PCR in 23 wild and inbred strains of mice. Alleles at each locus were divided by eye into lineage groups (Chapter 4; Bois *et al.* (1998a)) and each lineage assigned a number (e.g. group A corresponds to 1). Each previously ungrouped allele (Bois *et al.*, 1998a) was designated with a unique lineage number. For at least one minisatellite, strains 22MO, DBV, and MGT were heterozygous for alleles of different lineages. In these strains, each combination of allele lineages at the five loci were treated as a distinct strain during multidimensional scaling (MDS) analysis. Strains MPB and CBA had identical allele lineages at each locus to strains BIR and AKR respectively, so were not included in MDS analysis. The species or subspecies from which each strain was derived is indicated as follows. bact., *Mus m. bactrianus*; cast., *Mus m. castaneus*; dom., *Mus m. domesticus*; mol., *Mus m. molossinus*; mus., *Mus m. musculus*; spr., *Mus spretus*; I, Inbred strain

MDS uses pairwise combinations between allele lineages to determine the distance between two strains (the more differences, the greater the distance). The number assigned to each lineage was arbitrary so that the difference between alleles from lineages 1 and 2 was identical to the difference between lineages 1 and 12.

**Table 5.1*****Matrix of MVR code lineages analysed by MDS***

Strain	Species	MMS24	MMS26	MMS57	MMS58	MMS80
BIR	bact.	1	3	2	1	1
CIM	cast.	5	7	2	1	2
CTA	cast.	5	2	2	1	2
22MO/1	dom.	4	4	1	2	7
22MO/2	dom.	4	6	1	2	7
BFM	dom.	3	4	4	2	4
BIK/g	dom.	4	1	1	1	3
DBV/1	dom.	2	4	1	2	5
DBV/2	dom.	2	4	1	2	7
DBV/3	dom.	2	4	2	2	5
DBV/4	dom.	2	4	2	2	7
DDO	dom.	4	4	1	2	7
DJO	dom.	2	5	1	1	4
DOT	dom.	4	4	1	1	5
MOL	mol.	1	2	2	1	6
MBS	mus.	1	8	2	1	1
MGT/1	mus.	8	2	2	1	1
MGT/2	mus.	9	2	2	1	1
MGT/3	mus.	8	9	2	1	1
MGT/4	mus.	9	9	2	1	1
SPE	spr.	12	13	1	3	11
A/J	I	13	4	2	1	7
AKR	I	3	4	1	1	5
BALB/c	I	13	4	2	2	7
C3H	I	3	4	2	1	7
DBA	I	13	4	1	1	5
FVB	I	3	4	1	1	7
SWR	I	3	4	1	1	6

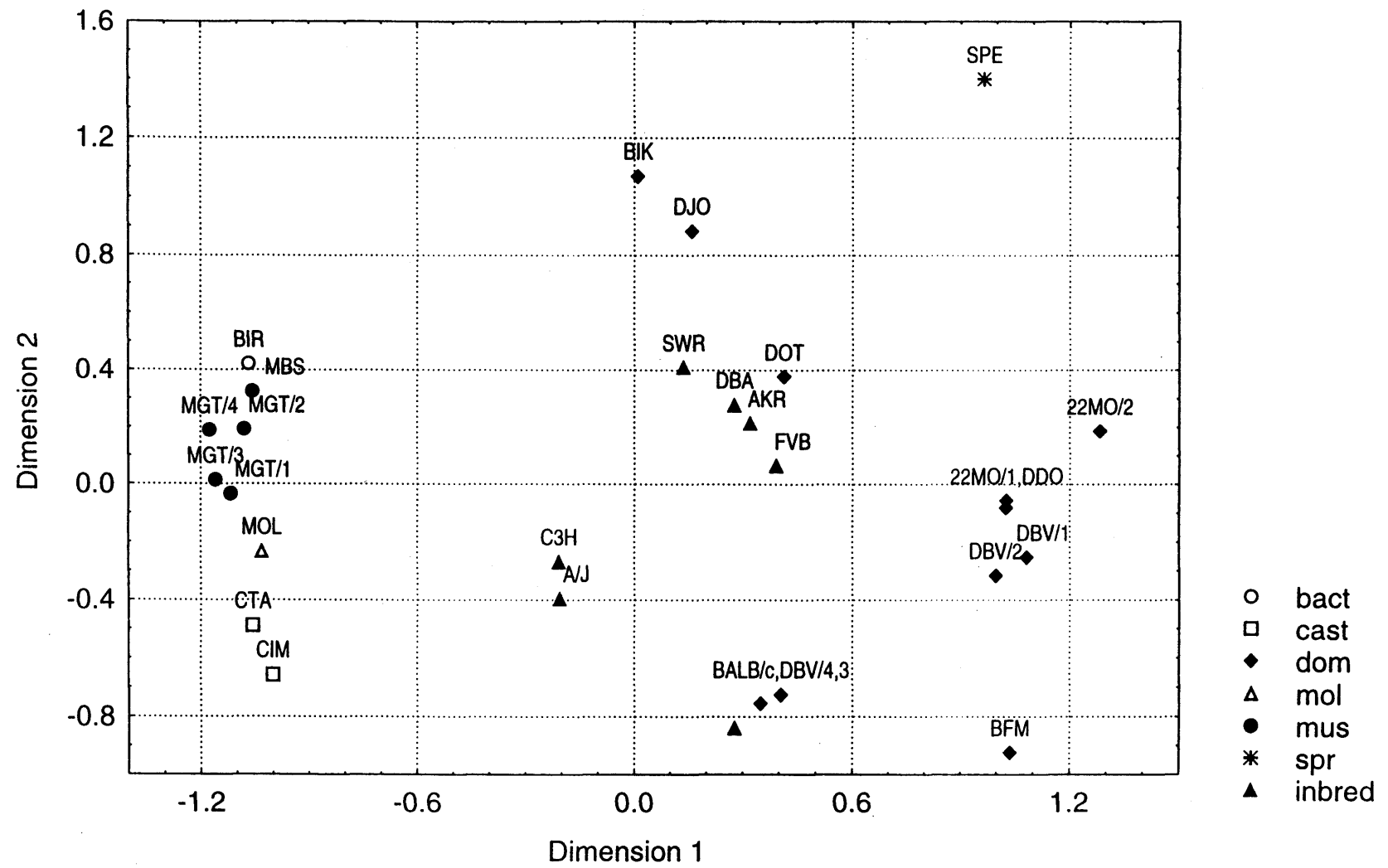
## Figure 5.2

### ***Multidimensional scaling analysis of 5 minisatellites in 21 strains of mice***

The principles of multidimensional scaling are described in the text. Distances between each strain are represented in two dimensions. The greatest similarities and differences between strains are represented by dimension 1. The four *Mus musculus* subspecies *Mus m. bactrianus*, *Mus m. castaneus*, *Mus m. molossinus*, and *Mus m. musculus* form a single cluster which is distinct from *Mus m. domesticus*. There is a high level of diversity within the *Mus m. domesticus* group. Inbred strains of mice display greatest similarity to *Mus m. domesticus* strains. Clusters of inbred strains are apparent such as C3H and A/J, suggesting that the two strains may be derived from a more recent ancestry than for all inbred strains combined. MDS analysis and Figure 5.2 were by Y. Dubrova.



Figure 5.2



of diverged allele lineages both within each subspecies and common to each of the four subspecies would be expected.

The simultaneous analysis of MVR data from multiple mouse minisatellites can therefore be used to generate an overall picture of the genetic similarities and differences between multiple strains. In contrast, analysis of the distributions of individual alleles at individual minisatellites can be used to examine patterns of gene flow between different populations (Bois *et al.*, 1998). For example, the subspecies *Mus m. domesticus* and *Mus m. musculus* are known to have interacted and inbred along a hybrid zone ranging from Denmark (Selander and Yang, 1969) to Bulgaria (Bonhomme *et al.*, 1983). The *Mus m. domesticus* strain DBV analysed in this study was derived from Denmark (Table 3.1) and found to be heterozygous for MMS57 allele lineages 57A and 57B (Figure 4.8). Lineage 57A is most commonly associated with *Mus m. domesticus*, whilst lineage 57B is common in *Mus m. musculus* strains, consistent with a recent hybridisation between the two subspecies.

Incorporation of MVR data from inbred strains of mice allows their origins to be investigated. MDS analysis suggests that the major contribution to inbred stocks was from *Mus m. domesticus* strains (Figure 5.2). However, MDS also reveals a shift in the position of the inbred lines towards the other *Mus musculus* subspecies, consistent with other evidence that laboratory mice have been developed from more than one subspecies of mouse (Festing, 1996). Consideration of individual loci supports this multi-strain origin, for example at MMS57 the inbred stocks A/J, BALB/c, and C3H all had alleles of the 57B lineage which are common to *Mus m. musculus* but which, with the exception of the single 57B allele from strain DBV described above, are absent from the *Mus m. domesticus* mice.

## **Conclusion**

Genome-wide similarities between strains of mice may be determined by the simultaneous analysis of a variety of different minisatellite loci. These similarities may be the result of either recent interbreeding between populations, or a recent common ancestry of the populations. In contrast, the consideration of individual alleles at single loci may allow levels of interbreeding between two diverged populations to be assessed. In this way, the analysis of one set of data in two distinct ways may allow the population histories of different groups of mice to be investigated on two different levels.

## Chapter 6

# Analysis of the *Hm-1* expanded simple tandem repeat

### Summary

The mouse expanded simple tandem repeat (ESTR) locus *Hm-1* has a spontaneous germline mutation rate of  $>10^{-2}$  and has been extensively used as a tool for analysing mutation induction in response to external mutagens (e.g. Dubrova *et al.* (1993); Dubrova *et al.* (1998a); Dubrova *et al.* (1998b); Fennelly *et al.* (1997)). To further understand the mechanisms of mutation operating at this locus, attempts were made to identify variant repeat units within the *Hm-1* array which could be used to analyse patterns of mutation within the ESTR. *Hm-1* alleles were screened for repeat variants using multiple restriction endonucleases which would cut within putative variant repeats of the locus. No variant repeats were identified. As an early step towards recombination analysis at the locus using multiple rounds of nested allele-specific PCR, strategies were developed for the systematic sequencing of regions flanking *Hm-1* to identify polymorphisms. This work proceeded no further than the correction and completion of previously published sequence data.

### Introduction

DNA fingerprint analysis of *HinfI*-digested genomic DNA from the BXD recombinant inbred (RI) strains of mice using probe  $\lambda$ 33.6 identified a highly polymorphic locus present as a 7 kb allele in strain C57BL/6J, but varying in length from 5-13 kb across the 25 BXD RI strains (Jeffreys *et al.*, 1987b). It was proposed that these fragments were allelic and represented a locus so unstable that new mutant alleles had arisen in all BXD strains (Jeffreys *et al.*, 1987b). The locus was initially named *Ms6-hm* (minisatellite detected by probe  $\lambda$ 33.6, hypermutable) (Jeffreys *et al.*, 1987b), and is here referred to as *Hm-1* (hypervariable minisatellite-1). Southern blot analysis of *Hm-1* in a variety of inbred strains identified a size range from 2 kb (AKR) to 16 kb (C57BL/6J), with many of the inbred

lines showing alleles heterozygous by size, further supporting a high mutation rate (Jeffreys *et al.*, 1987b).

The 7 kb C57BL/6J allele was cloned by size fractionation of *Sau*3AI-digested genomic DNA followed by ligation of the fractionated products into vector  $\lambda$ L47.1 and screening of the resulting library with probe  $\lambda$ 33.6 (Kelly *et al.*, 1989). Many of the positively hybridising clones did not contain the *Hm-1* fragment, but instead a true minisatellite termed Mm1, which was later renamed MMS80 (Chapter 3; Bois *et al.* (1998a)). One clone contained an insert of 2 kb. It was unlikely that this fragment would have survived size fractionation, and it was concluded that it must have collapsed from a larger fragment during cloning. Upon hybridisation of the clone to a mouse genomic Southern blot, the *Hm-1* locus was detected, in addition to a number of other weakly hybridising loci. Transfer of the 2 kb insert to pUC13 and propagation of the clone resulted in further collapse generating an insert of 0.4 kb. Sequence analysis of this clone identified a residual tandem repeat array composed of 19 identical GGGCA repeats located within a region with homology to a mouse transcript (MT) dispersed repeat (Kelly *et al.*, 1989). This short repeat unit and the lack of apparent repeat sequence variation led to its later reclassification as an expanded simple tandem repeat (ESTR) as opposed to a true minisatellite (Bois and Jeffreys, 1999).

The same clone was used as a probe to construct a restriction map of the *Hm-1* region from genomic DNA, and led to the identification of two polymorphisms between strains C57BL/6J and DBA/2J. The first was an insertion/deletion of a 2 kb fragment of unknown identity located 2-3 kb 5' of *Hm-1* (Kelly *et al.*, 1991). The second was an *AluI* polymorphism located immediately 3' of the repeat array and present in DBA/2J but absent from C57BL/6J mice. This site was used to map *Hm-1* to an interstitial location on chromosome 4 near the murine interferon- $\alpha$  gene cluster using the BXD RI strains (Kelly *et al.*, 1989). Interestingly, this locus is near a breakpoint between regions showing synteny with human chromosomes 1p and 9p (Kelly, 1990), raising the possibility that both the *Hm-1* ESTR and true mouse minisatellites may have a subtelomeric origin (Amarger *et al.*, 1998). Furthermore, the *AluI* polymorphism allowed each allele from the BXD RI strains to be traced to either a C57BL/6J or DBA/2J ancestor and demonstrated that the sizes of the

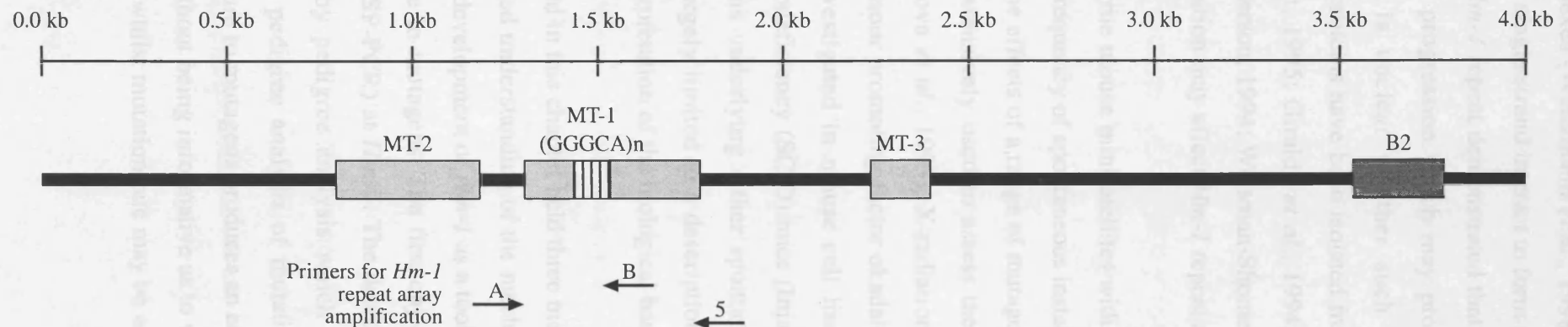
mutant alleles were more similar to the progenitor alleles from which they were ultimately derived, indicating that most *Hm-1* mutation events involved relatively small changes in allele length (Kelly *et al.*, 1989).

To characterise the sequences either side of the *Hm-1* array, a 15 kb *Bam*HI fragment from a C57BL/6J mouse containing ~1.4 kb of 5' and ~2.4 kb of 3' DNA flanking the repeat array was selected for cloning into a charomid 9-36 vector (Kelly *et al.*, 1991). The resulting clone contained a 4.4 kb fragment due once again to collapse of the *Hm-1* repeat array. Complete 5' and incomplete 3' sequence data were generated from the clone, including 60 *Hm-1* repeat units all of which were identical to the GGGCA consensus sequence (Kelly *et al.*, 1991). Within the flanking DNA, three regions with homology to MT elements, and a 3' B2 dispersed repeat were identified (Figure 6.1). The *Hm-1* array had expanded from within the central MT element (Kelly *et al.*, 1991).

Germline mutation rate was determined by pedigree analysis at 2.5% per gamete (Kelly *et al.*, 1989), although this was likely to be an underestimate as changes in length of a just few repeats would be undetectable by Southern blot hybridisation. Mutants displayed gains and losses in length at similar frequencies and, as predicted from BXD analysis of allele size (Jeffreys *et al.*, 1987b), generally involved small changes in length. Mutation rate was biased towards the male germline (Kelly *et al.*, 1989). In contrast to human minisatellites, a high incidence of somatic mutation has also been detected at *Hm-1* with 2.8% of mice detectably mosaic by Southern blot hybridisation with the intensity of the mutant allele indicating that between 8% and 60% of cells in somatic mosaic tissues contained the mutant allele (Kelly *et al.*, 1989). Analysis of allele transmissions from two mosaic mice revealed three-way non-Mendelian segregation of *Hm-1* demonstrating mosaicism in both germline and somatic tissues (Kelly *et al.*, 1989). This indicated that somatic instability occurred during early development prior to the separation of the germline and somatic progenitor cells.

The mechanism of mutation at *Hm-1* is unclear. It has been suggested that the formation of complex DNA secondary structures may initiate mutation by slippage replication. Under physiological conditions, the *Hm-1* repeat forms a hairpin as well as two different

**Figure 6.1**



### ***Region flanking the Hm-1 ESTR***

The *Hm-1* (GGGCA)<sub>n</sub> repeat array expanded from within a member of the mouse transcript (MT) family of dispersed repeat elements. Three MT-elements are associated with *Hm-1*, and a B2 repeat is located 3' of the repeat array. The region depicted represents the area previously sequenced (Kelly *et al.*, 1991). The *Hm-1* repeat was amplified either using primers Hm1pA and Hm1pB (Kelly, 1990), or Hm1pA and Hm1p5 (this chapter).

intra-strand tetraplexes (Weitzmann *et al.*, 1998). Intra-strand tetraplexes form when four G-rich motifs on a single strand interact to form a series of tetrads. *In vitro* DNA replication assays using the *Hm-1* repeat demonstrated that tetraplex formation is capable of blocking DNA polymerase progression which may promote strand slippage (Weitzmann *et al.*, 1997). Whilst it is unclear whether such structures form *in vivo*, a variety of tetraplex-binding proteins have been isolated from eukaryotic cells (Fang and Cech, 1993; Frantz and Gilbert, 1995; Giraldo *et al.*, 1994; Liu *et al.*, 1993; Liu and Gilbert, 1994; Schierer and Henderson, 1994; Weisman-Shomer and Fry, 1993), supporting the hypothesis that DNA conformation may affect *Hm-1* repeat array stability.

In the absence of true mouse minisatellites with high mutation rates (Chapter 3; Bois *et al.* (1998)), the high frequency of spontaneous instability at *Hm-1* has made it an attractive tool for investigating the effects of a range of mutagens on mutation rate (Dubrova *et al.*, 1993). *Hm-1* has been extensively used to assess the impact on mutation rate of for example  $\gamma$ -radiation (Dubrova *et al.*, 1993), X-radiation (Dubrova *et al.*, 1998a; Fennelly *et al.*, 1997), and the tumour promoting factor okadaic acid (Nakagama *et al.*, 1997). Instability has also been investigated in mouse cell lines such as cultures derived from severe combined immunodeficiency (SCID) mice (Imai *et al.*, 1997). However, as little is known of the mechanisms underlying either spontaneous or induced mutation at this locus, conclusions are largely limited to a description of observed changes in mutation rate as opposed to an interpretation of the biological basis of the mutation induction.

The work described in this chapter held three main objectives all of which would contribute both to an increased understanding of the mechanisms of mutation operating at the locus, and to the further development of *Hm-1* as a tool for investigating mutation induction in the mouse in response to mutagens. The first objective was the establishment of systems of small pool PCR (SP-PCR) at *Hm-1*. The detection of *de novo* mutations had previously been performed by pedigree analysis which carries a number of disadvantages. The determination by pedigree analysis of mutation rate and changes in mutation rate in response to exposure to mutagens produces an estimate of the mean mutation rate across all mice analysed, without being informative as to variation in mutation rate between different mice. In addition, whilst mutation rate may be accurately assessed by analysing a relatively



small cohort of mice due to the high spontaneous mutation rate of *Hm-1*, SP-PCR analysis of mutation in sperm DNA would reduce the numbers of mice that need to be analysed to generate a statistically meaningful result, thereby reducing both the labour intensity of mutation detection, and the number of mice to be culled for any mutation analysis study.

The second objective was to screen *Hm-1* alleles for the presence of variant repeats. The informativity of mutation analysis at true minisatellites has been greatly enhanced by MVR-PCR, and the establishment of analogous systems at *Hm-1* was therefore a major objective. To date, no variants of the GGGCA pentameric consensus repeat sequence have been identified. However, only ~300 bp (60 repeats) of the repeat array had been sequenced. *Hm-1* alleles can be up to 16 kb in length, and so the presence of variant repeat types could not be excluded.

The third long term objective was to analyse recombination frequency and distribution surrounding the *Hm-1* locus. Analysis of human minisatellites has demonstrated a fundamental link between minisatellite mutation and recombination. Recombination events surrounding MS32 have been analysed using multiple rounds of nested allele-specific PCR to screen for recombinant combinations of polymorphic markers flanking the repeat array (Jeffreys *et al.*, 1998a; Jeffreys *et al.*, 1998b). Whilst it is widely believed that STRs and ESTRs mutate by intra-allelic slippage-like mechanisms, there is no conclusive evidence to support this dogma. To investigate whether recombination has a role in *Hm-1* mutation, the long-term objective was to screen sperm DNA for crossover events using a similar strategy to that employed at human minisatellites (Jeffreys *et al.*, 1998a; Jeffreys *et al.*, 1998b). This form of analysis requires the identification of multiple polymorphisms between different strains of mice surrounding the locus of interest. The creation of an F1 hybrid between diverged strains would generate mice heterozygous at polymorphic sites which could be used to screen for the products of recombination, and to determine whether crossovers localise to within or near the *Hm-1* array. The first stage of this project was therefore to screen different strains of mice for polymorphisms flanking the *Hm-1* repeat.

The analyses described within this chapter are preliminary as my involvement with *Hm-1* was terminated at an early stage due to commitments to other projects. Studies on *Hm-1* are currently being continued at Leicester University by C. Yauk.

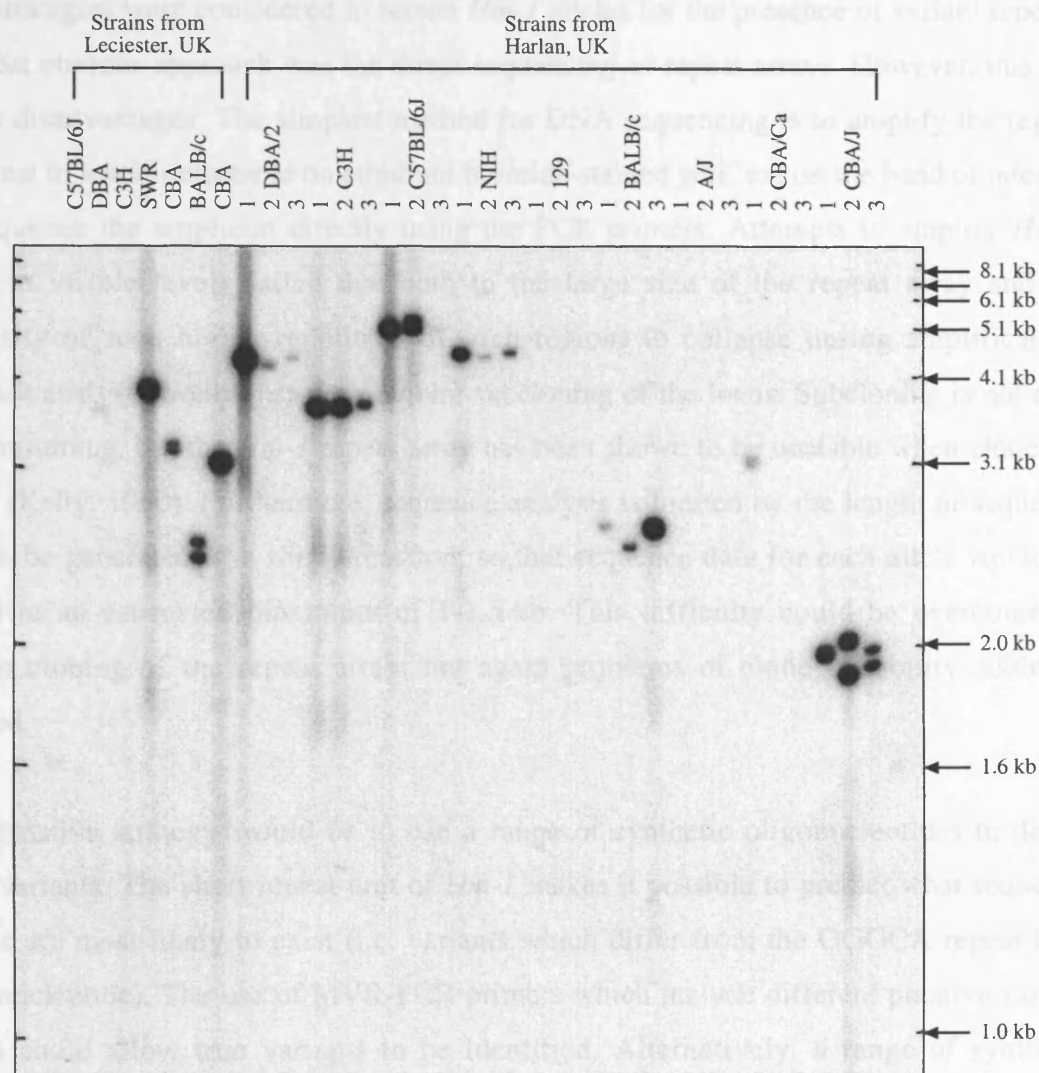
## Results and Discussion

### 1: PCR amplification at *Hm-1*

PCR analysis of *Hm-1* had been previously established by R. Kelly with alleles of  $\leq 3.7$  kb successfully amplified. Alleles of 6 kb and 9 kb from strain C57BL/6J were not amplifiable to levels detectable by Southern blot hybridisation. However, this work was done prior to 1990 and an elevated PCR efficiency would obviously be expected using current resources. The original system of PCR used the 5' Hm1pA primer in conjunction with a primer (Hm1pB) specific to a site 3' of the repeat array and within the MT-1 elements (Figure 6.1, primer sequences are described in Chapter 2). Whilst specificity was apparently achieved using this primer combination, it was decided to move the 3' primer to a site outside any known dispersed repeat. Primers Hm1pA and Hm1p5 were therefore chosen to amplify *Hm-1*. Examples of *Hm-1* alleles amplified from stocks of mice held either at the Division of Biomedical Services, University of Leicester, or supplied by Harlan UK Ltd., Loughborough are provided in Figure 6.2.

Alleles from strain C57BL/6J of 5-6 kb were readily amplifiable by PCR. However, no product was obtained from amplifications of Harlan strains A/J and 129, or the Leicester C57BL/6J and C3H strains. Whilst the reasons for these failures are unclear, they may include either the *Hm-1* repeat being too large for PCR amplification, the presence of polymorphisms within these strains generating mismatches at the primer sites, or simply low quality or low concentration of input DNA. Amplification of the same strains from both the Leicester and Harlan stocks demonstrated that alleles from a single strain were of similar size despite being derived from different stocks, consistent with previous studies demonstrating that whilst the mutation rate of *Hm-1* may be high, the majority of mutations result in small changes in array length (Kelly, 1990). Due to time constraints, no attempts at establishing SP-PCR at *Hm-1* were made.

**Figure 6.2**



### **PCR amplification of Hm-1**

DNA from a variety of inbred strains of mice from either the Division of Biomedical Services, University of Leicester, or from Harlan UK Ltd., Loughborough was amplified. Three mice (denoted 1, 2, and 3) from each of the mice stocks obtained from Harlan UK Ltd. were amplified. There is substantial variation in signal intensity between strains. This is likely to result from variation in the quality and concentration of input DNA used for PCR amplification, although the presence of polymorphisms between strains at primer sites affecting PCR efficiency cannot be excluded.

Amplification was performed with primers Hm1pA and Hm1p5 at 96°C for 40 s, 56°C for 30 s, 70°C for 3 min for 22 cycles using 20 ng of input genomic DNA. Samples were electrophoresed through a 40 cm 0.8% agarose gel, Southern blotted and hybridised using a probe generated by PCR amplification using oligonucleotides (HMA and HMB) of overlapping and complementary sequence, each of which contained 4 complete repeat units (Kelly, 1990).

## **2: Screens for variant repeats at *Hm-1***

Three strategies were considered to screen *Hm-1* alleles for the presence of variant repeats. The most obvious approach was the direct sequencing of repeat arrays. However, this had various disadvantages. The simplest method for DNA sequencing is to amplify the region of interest to levels detectable on ethidium bromide-stained gels, excise the band of interest, and sequence the amplicon directly using the PCR primers. Attempts to amplify *Hm-1* alleles to visible levels failed due both to the large size of the repeat array and the propensity of such highly repetitive GC-rich regions to collapse during amplification. Sequence analysis would therefore require subcloning of the locus. Subcloning is not only time consuming, but the *Hm-1* repeat array has been shown to be unstable when cloned in *E. coli* (Kelly, 1990). Furthermore, sequence analysis is limited by the length of sequence that can be generated in a single reaction, so that sequence data for each allele would be limited to an estimated maximum of 1-1.5 kb. This difficulty could be overcome by shotgun cloning of the repeat array, but again problems of clone instability could be expected.

An alternative strategy would be to use a range of synthetic oligonucleotides to detect repeat variants. The short repeat unit of *Hm-1* makes it possible to predict what sequence variants are most likely to exist (i.e. variants which differ from the GGGCA repeat by a single nucleotide). The use of MVR-PCR primers which include different putative variant repeats could allow true variants to be identified. Alternatively, a range of synthetic oligonucleotides could be used as probes of the *Hm-1* array with strong hybridisation indicative of the presence of variants. However, this approach was excluded due to the expense of synthesising large numbers of oligonucleotides.

The third strategy was to screen for variants using restriction endonuclease digestion. No restriction endonucleases cleave within (GGGCA)<sub>n</sub> repeat arrays. However, the introduction of a range of variant sites within the GGGCA repeat would generate restriction enzyme target sites. By screening alleles with appropriate enzymes, such variants could be detected. This method also carries a number of disadvantages. Not all putative variants generate restriction sites and so could not be detected. Furthermore, most enzymes that cut any one variant will cut several putative variants, so even if the technique detects variant

sites, it may not directly identify the variant sequence. Despite these disadvantages, the digestion approach was considered the most rapid and cost-effective method to screen for variants.

To select the appropriate restriction enzymes for the detection of variant repeats, every putative variant which differed from the GGGCA sequence by a single base addition, substitution, or deletion was inserted *in silico* in multiple copies into a (GGGCA)<sub>n</sub> array, and the synthetic sequences analysed using the restriction mapping (map) program within the GCG molecular biology software package (Devereux *et al.*, 1984). Of the 33 putative variants, 25 generated new restriction sites (Table 6.1). Nine enzymes (*AccI*, *BfaI*, *BsgI*, *Bsp1286I*, *Cac8I*, *HaeIII*, *HhaI*, *MnII*, and *NlaIII*) were selected which together cut within 20 of the 25 variants. Criteria for enzyme selection included the number of different variants detected by each enzyme, the sequence diversity of variants cleaved by the combination of all enzymes, and the cost and availability of each endonuclease. With the exception of *BfaI*, each enzyme cleaves >1 different putative repeats. However, 15 of the 20 putative repeats potentially identifiable by these enzymes have unique digestion profiles (a unique combination of different enzymes that either cut or fail to cut the variant) so digestions with additional enzymes would permit the exact sequence identity of these variants to be determined (for example of the two repeats cut by *BsgI*, only the GTCGA variant is also cut by *BstI*). The five exceptions are digested exclusively by either *MnII* or *NlaIII* (Table 6.1).

Four different strains of mice were selected for analysis (C57BL/6J, C3H, BALB/c, and CBA/J) to provide a wide size range of amplifiable alleles. To eliminate any possible variation in PCR efficiency between reactions, *Hm-1* was amplified in all strains and PCR products pooled and precipitated prior to digestion. The digestion products are shown in Figure 6.3. Of the nine enzymes selected, seven had known restriction endonuclease target sites within the amplified DNA flanking the repeat array (data not shown) resulting in changes in size of the product detected (Figure 6.3). Digestions were partial for four of the nine enzymes. To control for digestion efficiency, other amplicons adjacent to the *Hm-1* repeat array which contained known endonuclease target sites were amplified and digested

**Table 6.1**

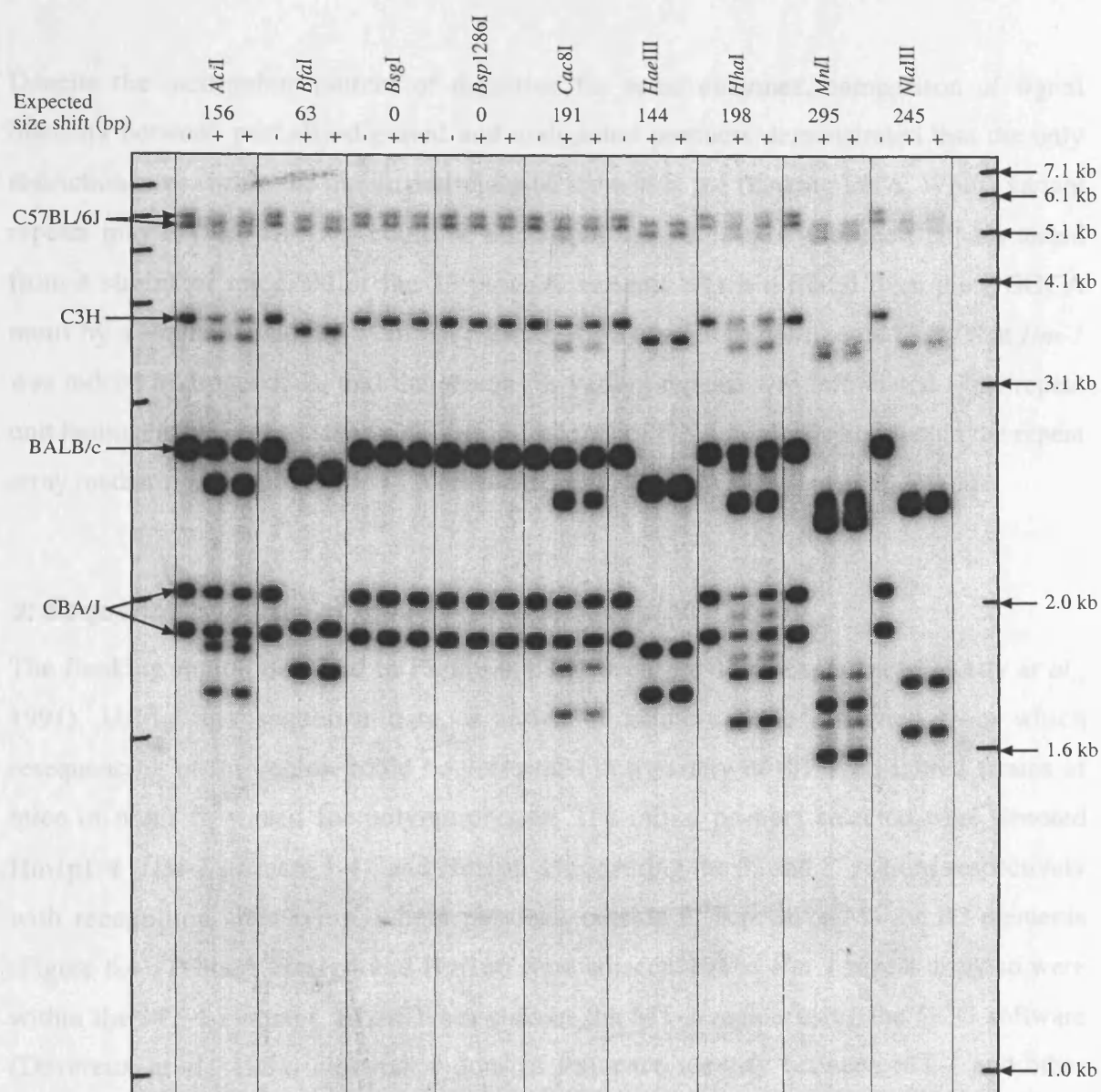
***Putative variant repeats generate novel restriction sites***

Every putative variant repeat which differed from the GGGCA consensus by a single base addition, substitution, or deletion is listed at left, with variant sites underlined. The three classes of variant are separated by bold horizontal lines with the GGGCA consensus at the top. Each variant was embedded *in silico* within a (GGGCA)<sub>n</sub> array and the sequence analysed for restriction endonuclease target sites using the map program of the GCG software package (Devereux *et al.*, 1984). Every enzyme which cuts at least one repeat is listed. Black boxes denote which variant repeats are cut by which enzymes. The 9 enzymes selected are indicated in grey, as are the 20 putative variant repeats cut by these enzymes. Enzyme selection was based on the total number and diversity of variant sites which could be detected, and on the relative costs and availabilities of enzymes.

Table 6.1

	<i>AccI</i>	<i>BbvI</i>	<i>BceFI</i>	<i>BfaI</i>	<i>BmgI</i>	<i>BsaII</i>	<i>BscGI</i>	<i>BsgI</i>	<i>BsII</i>	<i>BsmFI</i>	<i>Bsp1286I</i>	<i>BspMI</i>	<i>BtsI</i>	<i>Cac8I</i>	<i>CviJI</i>	<i>CviRI</i>	<i>EcoRII</i>	<i>FauI</i>	<i>Fnu4HI</i>	<i>HaeI</i>	<i>HaeIII</i>	<i>HhaI</i>	<i>MnlI</i>	<i>MspI</i>	<i>MspAII</i>	<i>MwoI</i>	<i>NciI</i>	<i>NlaIII</i>	<i>Sau96I</i>	<i>ScrFI</i>	<i>SimI</i>	<i>TauI</i>	<i>TseI</i>	<i>TspRI</i>	<i>UbaEI</i>	<i>UbaOI</i>	
GGGCA																																					
GGGCAA																																					
GGGACA																																					
GGAGCA																																					
GAGGCA																																					
GGGCAT																																					
GGGCTA																																					
GGGTCA																																					
GGTGCA																																					
GTGGCA																																					
GGGCAC																																					
GGGCCA																																					
GGCGCA																																					
GCGGCA																																					
GGGCAG																																					
GGGCGA																																					
GGGAA																																					
GGACA																																					
GAGCA																																					
AGGCA																																					
GGGCT																																					
GGGTA																																					
GGTCA																																					
GTGCA																																					
TGGCA																																					
GGGCC																																					
GGCCA																																					
GCGCA																																					
CGGCA																																					
GGGCG																																					
GGGGA																																					
GGGC																																					
GGGA																																					
GGCA																																					

**Figure 6.3**



**Digestion screen of *Hm-1* for variant repeats**

*Hm-1* alleles from the strains CBA/J, BALB/c, C3H, and C57BL/6J were amplified by PCR and all PCR products pooled and precipitated. Samples were digested with each of the 9 selected enzymes to screen for 20 putative variant repeats. Enzyme selection is described in Table 6.1, and in the text. For each enzyme, digestions were performed in duplicate (digestion reactions are indicated by +) and accompanied by a negative control containing digestion buffer but without enzyme (designated -). Sequence analysis of the *Hm-1* amplicon identified restriction sites for seven of the enzymes selected within the flanking DNA of the repeat array. The expected size shift due to digestion at these sites is indicated. No evidence for digestion within the repeat array was detected, indicating that amongst the 6 alleles screened, none of the 20 putative variant repeats were present.



with the same enzymes. Each enzyme cleaved at least partially at the expected sites (data not shown).

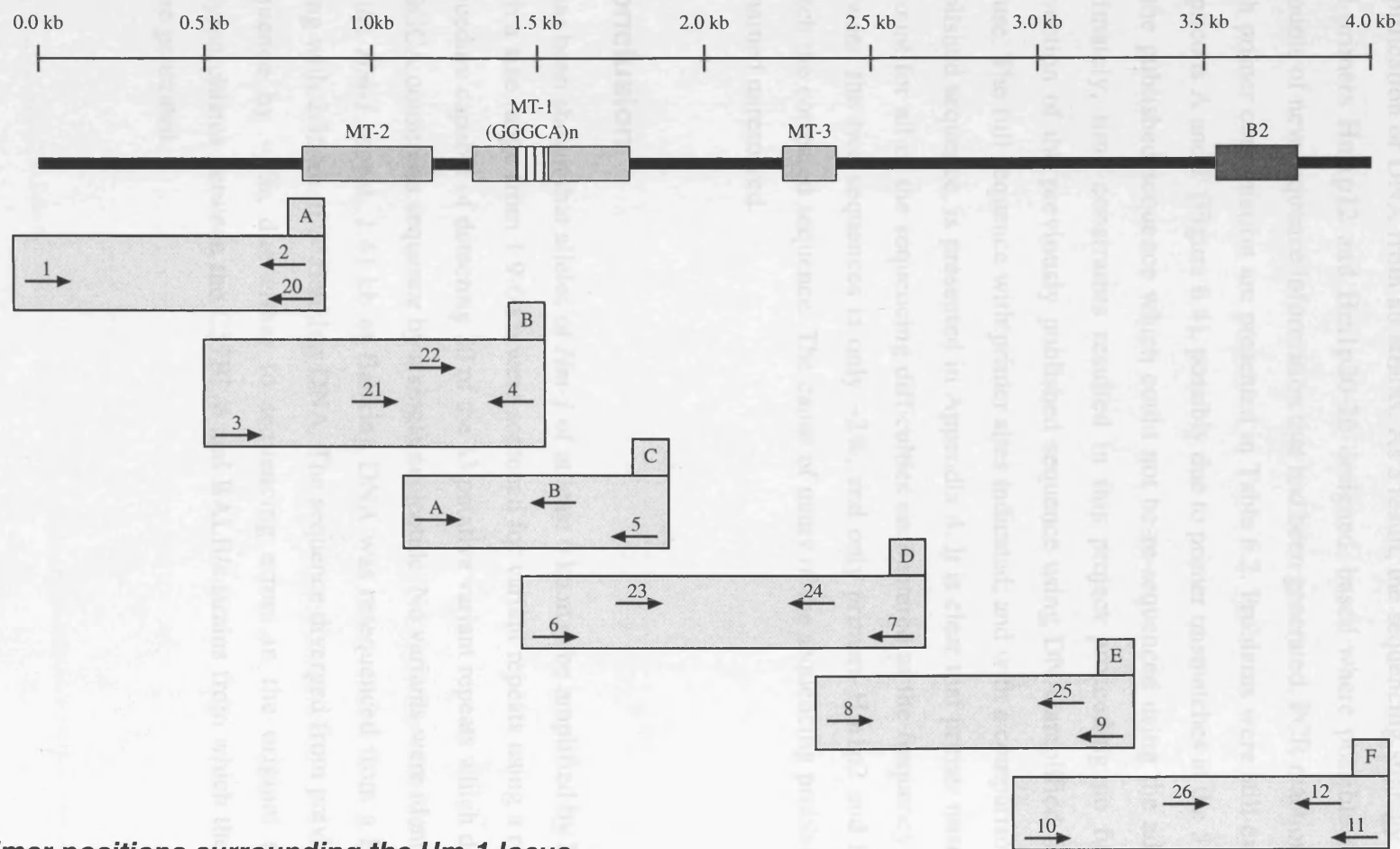
Despite the incomplete pattern of digestion for some enzymes, comparison of signal intensity between partially-digested and undigested products demonstrated that the only restriction sites within the amplicons analysed lay within the flanking DNA. Whilst variant repeats may exist at *Hm-1* it could be concluded that, at least within the 6 alleles tested from 4 strains of mice, 20 of the 33 putative variants which differed from the GGGCA motif by a single nucleotide were not present. It was therefore considered likely that *Hm-1* was indeed homogeneous, and the screen for variant repeats was terminated. This repeat unit homogeneity is consistent with a proposed role of DNA conformation within the repeat array mediating allele instability (Weitzmann *et al.*, 1997; Weitzmann *et al.*, 1998).

### **3: Sequence analysis of the *Hm-1* flanking region**

The flanking region depicted in Figure 6.1 had been previously sequenced (Kelly *et al.*, 1991). Using this sequence data, a series of primers were designed from which resequencing of the region could be performed in a variety of diverged inbred strains of mice in order to screen for polymorphisms. The initial primers selected were denoted Hm1p1-4 (*Hm-1* primers 1-4), and Hm1p6-11, covering the 5' and 3' regions respectively with recognition sites lying, where possible, outside of repetitive MT or B2 elements (Figure 6.4). Primers Hm1p4 and Hm1p6 were adjacent to the *Hm-1* repeat array so were within the MT-1 element. BLAST searches of the MT-1 region using the GCG software (Devereux *et al.*, 1984) allowed regions of sequence identity between MT-1 and other MT elements to be defined and avoided during primer design (data not shown).

Eight strains of mice (129, A/J, BALB/c, C3H, C57BL/6, CBA/J, DBA, and NIH) were selected in consultation with Dr. M. Festing, Leicester University, for sequence analysis based on their mating compatibilities, the ease with which stocks could be maintained, and to maximise divergence between strains and therefore elevate the expected frequency of sequence polymorphisms between selected mice. Serious difficulties were experienced at an early stage in the amplification and sequencing of the region of interest (data not shown). The most obvious explanation was that errors existed in the previously published

**Figure 6.4**



#### **Primer positions surrounding the Hm-1 locus**

Primer positions are indicated to scale with the centre of each primer representing its position within the flanking sequence. Primers 1 (Hm1p1) to 11 (Hm1p11) were initially designed based on the sequence data of Kelly *et al.* (1991). Primers 12 (Hm1p12) and 20-26 (Hm1p20-Hm1p26) were designed where possible from within resequenced regions. The resulting sequence is presented in Appendix 4.

sequence at sites where the primers had been designed. It was unlikely that strain-specific polymorphisms within primer sites were responsible for the problems, as the original sequence was derived from a C57BL/6 mouse, and difficulties were experienced in the amplification of DNA from all strains. As a result, the sequencing strategy was reassessed and primers Hm1p12 and Hm1p20-26 designed, based where possible on the minimal amounts of new sequence information that had been generated. PCR reaction conditions for each primer combination are presented in Table 6.2. Problems were still experienced with amplicons A and F (Figure 6.4), possibly due to primer mismatches at the 5' and 3' termini of the published sequence which could not be re-sequenced using the adopted strategy. Ultimately, time constraints resulted in this project proceeding no further than the correction of the previously published sequence using DNA amplified from a BALB/c mouse. The full sequence with primer sites indicated, and with a comparison to previously published sequence, is presented in Appendix 4. It is clear that primer mismatches cannot account for all of the sequencing difficulties encountered as the frequency of mismatches between the two sequences is only ~2%, and only primers Hm1p2 and Hm1p9 did not match the corrected sequence. The cause of many of the sequencing problems experienced remained unresolved.

## Conclusion

It has been shown that alleles of *Hm-1* of at least 6 kb may be amplified by PCR. Six alleles with a size range from 1.9-6 kb were screened for variant repeats using a restriction digest procedure capable of detecting 20 of the 33 putative variant repeats which differed from the GGGCA consensus sequence by a single nucleotide. No variants were identified. Upstream of the *Hm-1* repeat, 1.41 kb of flanking DNA was resequenced from a BALB/c mouse, along with 2.35 kb of 3' flanking DNA. The sequence diverged from previously published sequence by ~2%, due either to sequencing errors in the original sequence or to polymorphisms between the C57BL/6 and BALB/c strains from which the two sequences were generated.

**Table 6.2*****PCR conditions for Hm-1 primer pairs***

Region	5' primer	3' primer	T <sub>A</sub>	Cycle Number
A	1	2	53°C	27
	1	20	58°C	26
B	3	4	61°C	28
	21	4	62°C	24
	22	4	54°C	29
C	A	5	56°C	22*
	A	B	60°C †	24†*
D	6	7	61°C	28
	6	24	N/A	N/A
	23	7	N/A	N/A
	23	24	60°C	24
E	8	9	51°C	31
	8	25	54°C	26
F	10	11	PCR failed	PCR failed
	10	12	54°C	35
F	26	11	PCR failed	PCR failed
	26	12	54°C	35

Each combination of 5' and 3' primer from each of the 6 regions presented in Figure 6.4 are listed. Annealing temperatures (T<sub>A</sub>) and cycle numbers for PCR amplifications are described. PCR products were amplified to levels visible on ethidium bromide-stained gels, with the exception of \* where products were amplified to levels detected by Southern blot hybridisation. † denotes PCR amplifications described by Kelly (1990). N/A denotes not available, where primer combinations listed were never used.

## Chapter 7

# Characterisation of human minisatellites with discontinuous allele length distributions

### Summary

Mutation dynamics and processes have been extensively characterised at human minisatellites which display mutation rates of  $>10^{-3}$  per sperm, have continuous allele size distributions, and population heterozygosities of  $>98\%$ . However, the majority of human minisatellites display lower levels of heterozygosity (70-80%) indicative of a mutation rate of  $\sim 10^{-5}$  per gamete. It is unknown whether mutation mechanisms which operate at hypermutable minisatellites also operate at minisatellites with lower mutation rates. As a result, three minisatellites (MS51, *D19S20* and *D17S74*) with reported heterozygosities of 70%-90% were characterised as potential candidates for mutation analysis. *D17S74* was found to be unsuitable due to high levels of allele size diversity and population heterozygosity. In contrast, both MS51 and *D19S20* displayed heterozygosities of  $\sim 80\%$ , trimodal allele size distributions, and patterns of variant repeat interspersions making them good candidates for the characterisation of *de novo* mutation. Systems of MVR-PCR were established to further characterise patterns of allele diversity. There was little evidence for polar variation at either locus indicating that if polar mutational processes do operate, they occur at a lower frequency than non-polar mechanisms. Patterns of variation consistent with intra-allelic duplications were observed at both loci. The highly informative system of MVR-PCR established at MS51 made it a good candidate for future mutation analysis. However, due to time constraints these experiments were not continued.

### Introduction

Germline mutation processes have been extensively characterised directly from sperm DNA at the human minisatellites MS205, MS32, B6.7 and CEB1 (Buard *et al.*, 1998; Jeffreys *et al.*, 1994; May *et al.*, 1996; Tamaki *et al.*, 1999b). These loci all display high levels of allele diversity both in length and internal structure due to the operation of

complex mutation processes generating mutants in the male germline at rates of 0.4-15% per sperm. Characterisation of patterns of allele variation by MVR-PCR in Caucasian populations at MS32, MS31, and MS205 demonstrated that most variability was concentrated towards one end of the repeat array (Armour *et al.*, 1993; Jeffreys *et al.*, 1990; May *et al.*, 1996; Monckton *et al.*, 1993; Neil and Jeffreys, 1993; Tamaki *et al.*, 1993; Tamaki *et al.*, 1992). Analysis of germline mutants both in pedigrees and in sperm confirmed that polarised variation reflected polarity of mutation, which at MS32 has been shown to be due to a localised recombination hotspot within the flanking DNA (Jeffreys *et al.*, 1998a; Jeffreys *et al.*, 1998b). Most mutations at this locus result from the conversion-like transfer of repeats usually occurring in register between alleles, suggesting an allele pairing function in DNA flanking the unstable end of the repeat array. Polarised variation in allele structure was not observed at the hypermutable loci CEB1, MS621 and D7S22 (Andreassen and Olaisen, 1998; Armour *et al.*, 1996b; Buard *et al.*, 1998; Buard and Vergnaud, 1994). However at CEB1, polarised inter-allelic repeat transfers have been characterised in sperm DNA, and the lack of polar variation was due to the high frequency of complex non-polar intra-allelic mutation at this locus (Buard *et al.*, 1998). Therefore while polar variation is indicative of polar mutation potentially due to a flanking recombination hotspot, a lack of polarity does not demonstrate the absence of such a hotspot.

Similar mechanisms of mutation operate at all hypermutable loci characterised to date. In addition, most germline mutations result in small changes in allele size, with a bias towards an increase in repeat number (Buard *et al.*, 1998; Jeffreys *et al.*, 1994; May *et al.*, 1996; Tamaki *et al.*, 1999b). Mutation increases allele diversity, whilst factors such as genetic drift and (at some loci) selection reduce allele diversity. A high mutation rate at minisatellites will maintain high levels of allele diversity both in allele length and internal structure, and multiple consecutive incremental changes in array length by mutation could readily result in the continuous distribution of allele sizes observed at hypermutable loci. For minisatellites with lower mutation rates, a subset of alleles would be expected to increase in frequency due to the effects of drift resulting in a discontinuous allele size distribution, as is seen at minisatellite D2S44 (Holmlund and Lindblom, 1998). Analysis of variant repeat distribution at this locus by MVR-PCR revealed that its bimodal allele size distribution corresponded to three distinct allele lineages, with small alleles derived from a

single lineage and large alleles derived from two lineages (Holmlund and Lindblom, 1998), consistent with a low mutation rate and population sampling effects. Alternatively, mutation rate may not be low, but instead there may be selection for alleles of specific sizes. However, it can be assumed that discontinuous size distributions and low levels of length heterozygosity both indicate a low mutation rate.

The majority of human minisatellites display much lower levels of allele length heterozygosity (typically 70-80%) than the hypermutable loci characterised by mutation analysis (Armour *et al.*, 1990; Nakamura *et al.*, 1987). Heterozygosity of 70-80% implies a mutation rate of  $\sim 5 \times 10^{-5}$  per gamete by  $H = \frac{1}{1 + 4N_e\mu}$ , where  $H$  is heterozygosity, the effective population size ( $N_e$ ) is 20000, and  $\mu$  is the mutation rate. It is unknown whether all human minisatellites mutate by similar mechanisms, or whether mechanisms operating at hypermutable loci are specific to that class of minisatellites. However, there are indications that complex mutation mechanisms do operate at minisatellites with low mutation rates as extensive sequence analysis of the *HRAS1* minisatellite found evidence for conversion-like mutation processes (Ding *et al.*, 1999), although confirmation would require the characterisation of *de novo* mutation events.

Mutation analysis at minisatellites with low mutation rates and variabilities would also allow the effects of minisatellite homozygosity to be investigated. In heterozygotes, mismatches in length and repeat array sequence between interacting alleles may serve to trigger the abortion of recombination initiation complexes and instead result in complex rearrangements without crossover. The low frequency of true homozygotes at hypermutable loci has prevented analysis of the effects of homozygosity on mutation. At minisatellite MS32, true crossovers between interacting minisatellites generate simpler mutant structures than the products of conversion (Jeffreys *et al.*, 1998b). In homozygotes, the absence of sequence mismatches between alleles interacting in register could be expected to reduce gene conversion and increase the frequency of true crossovers, therefore reduce the complexity of mutant structures generated. If inter-allelic recombination in homozygotes led to a high frequency of in-register crossover without rearrangement, the products of recombination would be identical to progenitor alleles, so a reduced mutation rate could be expected.

The ultimate objective of the work within this chapter was to analyse mutation dynamics at human minisatellites with lower levels of allele diversity and heterozygosity, in order to compare the mutation mechanisms that operate at minisatellites with low and high mutation rates, and to determine the effects of homozygosity on mutation. The first stage was to identify and characterise suitable candidate loci for analysis. The ideal locus would display allele sizes of <4 kb, and be composed of interspersed variant repeats allowing MVR-PCR analysis. Target heterozygosities were between 70% and 90% corresponding to mutation rates of between  $2 \times 10^{-5}$  and  $1 \times 10^{-4}$  per gamete, substantially lower than the hypermutable loci characterised to date, but sufficiently high to allow the detection of *de novo* germline mutants. One obvious candidate was MS51 (*D11S97*) (Royle *et al.*, 1988a) which was isolated as a *Sau3AI-EcoRI*-digested DNA fragment cloned from a DNA fingerprint (Jeffreys *et al.*, 1988) and displayed an estimated heterozygosity of 77% with an allele size range of 1.3-4.3 kb (Jeffreys *et al.*, 1988). Furthermore, 113 bp of 5' and 193 bp of 3' DNA flanking the repeat array had been previously sequenced, and PCR established for amplification of the locus (Jeffreys *et al.*, 1988). In order to identify additional candidate loci, the VNTRs *D17S74* (pCMM86) and *D19S20* (pJCZ3) (Nakamura *et al.*, 1988; Nakamura *et al.*, 1987) were selected from the Human Genome Mapping Project (HGMP) probe bank on the basis of the reported range of allele sizes (1.0-3.5 kb and 1.5-4.0 kb respectively) and heterozygosities (90% and 75% respectively). (Information about probes available from the HGMP was derived from the HGMP probe bank web site which is no longer in existence.)

## Results

### ***Sequence analysis of candidate loci***

Sequence analysis of MS51 identified multiple variant repeats of 25 and 33 bp in length (Jeffreys *et al.*, 1988). This, combined with an allele size range of between 1.3 kb and 4.3 kb, made the locus amenable to MVR-PCR analysis. To determine the suitability for variant repeat typing at *D17S74* and *D19S20*, each locus was sequenced. However for reasons unknown, attempts to amplify these loci using primers specific to the pUC18 vector into which the inserts were cloned were unsuccessful. This work was undertaken simultaneously with the subcloning of mouse loci described in Chapter 3, so it was decided



to transfer the repeat regions of each locus to a pBluescript II SK<sup>+</sup> vector for sequence analysis. The repeat regions were identified by restriction endonuclease-resistance as described in Chapter 3, and were used as probes to generate proximal restriction maps surrounding the VNTRs from each of the clones obtained from the HGMP, as described in Figure 3.8.

Further problems were encountered for locus *D19S20*. Sequence analysis of the cloned region revealed that the original insert in the pJCZ3 clone (Nakamura *et al.*, 1988) lacked flanking DNA on one side of the repeat array. Without flanking sequences, PCR amplification of the locus from genomic DNA would not be possible. Fortunately, a cosmid clone (cMCOB19) (Nakamura *et al.*, 1987) covering the same region was also available from the HGMP probe bank. Proximal restriction map analysis of the cosmid using the repeat region derived from the corresponding plasmid as a probe led to the identification of appropriate regions for subcloning and sequencing. Restriction maps of the *D17S74* plasmid and the *D19S20* cosmid, with subcloned regions indicated, are presented in Figure 7.1a. Repeat composition of *D17S74*, *D19S20*, and MS51 are presented in Figure 7.1b. Each of the three loci were GC-rich with high levels of purine-pyrimidine strand asymmetry and variant sites present between repeat units making each locus amenable to MVR-PCR analysis. Repeat units of both MS51 and *D19S20* were strikingly similar, composed of multiple A(G)<sub>n</sub>A degenerate motifs, with at least some repeat units of 33 bp in length. Known sequences flanking each locus are presented in Appendix 5.

### **PCR amplification of three minisatellites**

PCR amplification has been previously performed at MS51 using primers 51-A and 51-B (Table 2.1; Appendix 5; Jeffreys *et al.* (1988)). These primers were located at or near the outer limits of known sequence flanking the locus. For mutation detection and analysis, nested primer pairs are required to avoid contamination between consecutive experiments. Initial amplification of samples uses outer primers for a limited number of cycles only. These reactions are set up in a dedicated single-molecule clean flow hood. Mutants are then reamplified outside of the flow hood to detectable levels using internal primers. This combination of spatial separation and primer nesting serves to minimise contamination between experiments. With the potential for future single molecule mutation detection

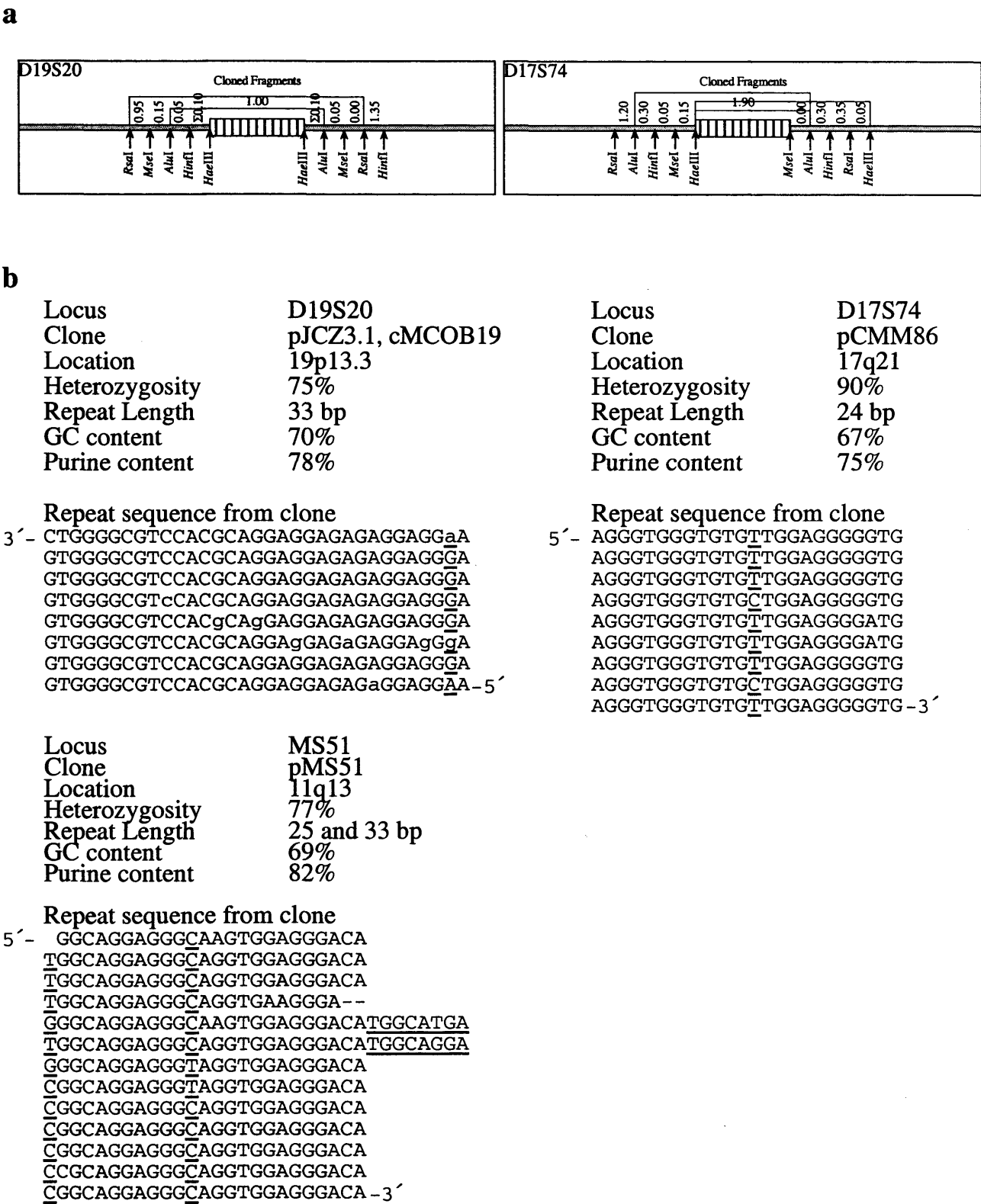


## Figure 7.1

### ***Primary sequence characteristics of three human minisatellites***

Proximal restriction maps were generated as described in Figure 3.8 for *D19S20* and *D17S74*, from the clones cMCOB19 and pCMM86 respectively (Nakamura *et al.*, 1988; Nakamura *et al.*, 1987) (Figure 7.1a). Regions containing the tandem repeat array plus flanking DNA were subcloned into pBluescript II SK<sup>+</sup> vectors and sequenced. Full sequences are presented in Appendix 5. Sequence data from the repeat array of each locus and of MS51 are presented (Figure 7.1b). MS51 sequences were as described in Jeffreys *et al.* (1988). The repeat units for each locus are GC-rich and display purine-pyrimidine strand asymmetry. Polymorphic sites, where each variant was identified in at least two repeats, are underlined. 5' and 3' orientations are designated as in Appendix 5.

Figure 7.1



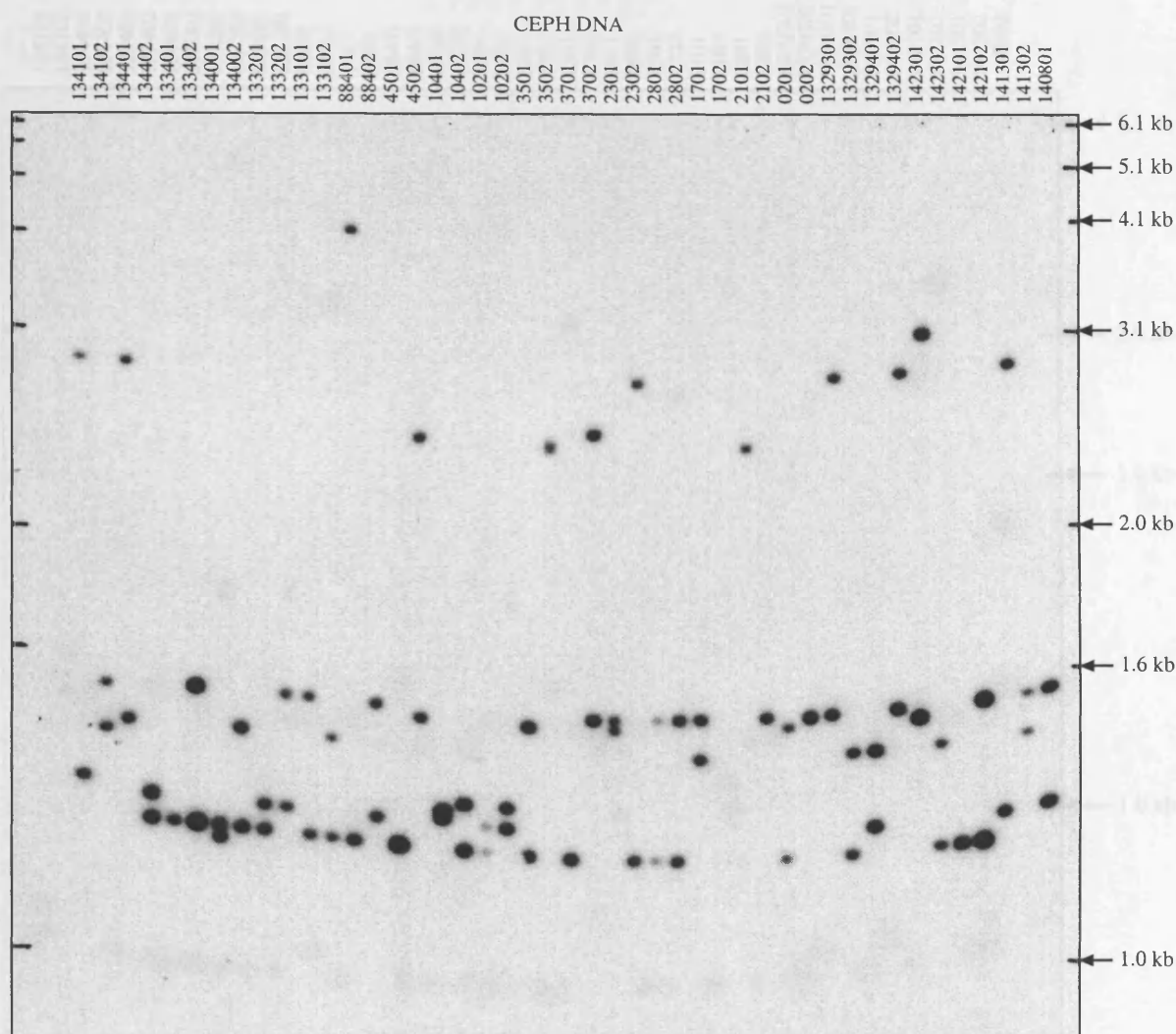
studies at MS51, it was necessary to avoid use of the outer primers. New primers for amplification of the locus were designed, named 51-C and 51-D. Primers were also designed for the amplification of both *D17S74* and *D19S20*. Sequences of primers are presented in Table 2.1, and positions indicated in Appendix 5. Each locus was amplified in the same panel of 45 unrelated CEPH DNAs. Allele size distributions detected by PCR are presented for loci MS51 (Figure 7.2), *D19S20* (Figure 7.3) and *D17S74* (Figure 7.4).

Both MS51 and *D19S20* display strikingly similar trimodal allele size distributions as well as repeat unit sequences. Heterozygosities for the two loci were determined from Figures 7.2 and 7.3 at 82% and 80% respectively, corresponding to an estimated mutation rate of  $\sim 5 \times 10^{-5}$  per gamete ( $N_e = 20000$ ). All alleles at both loci were below  $\sim 4$  kb in size so could potentially be analysed completely by MVR-PCR. In contrast, *D17S74* displayed much higher levels of allele size variability with alleles showing a continuous size distribution between 1.0 kb and 6.0 kb. Heterozygosity was determined from Figure 7.4 at 93%, corresponding to an estimated mutation rate of  $2 \times 10^{-4}$  per gamete. This mutation rate was higher than the target range. Furthermore, with a size range from  $\sim 30$ -250 repeats, many alleles were too large to be analysed along their entire length by MVR-PCR. The loci MS51 and *D19S20* were therefore suitable candidates for the characterisation of *de novo* mutation events, whilst *D17S74* was excluded from further analysis.

### ***MVR-PCR analysis of D19S20***

As described in Chapter 4, the first stage in establishing a system of MVR-PCR was to sequence alleles from genomic DNA to ascertain levels of variant repeat sequence diversity. Diversity within the cloned allele of *D19S20* was very low with polymorphism restricted to a single dimorphic site (Figure 7.1). A single 0.8 kb allele from CEPH DNA 133202 was sequenced, and is presented in Figure 7.5. Repeat size heterogeneity was observed within this allele, with two 28 bp and eight 33 bp repeat units. MVR primers were designed to distinguish two repeat types called C and T, which differed by a C/T transition at a single site (Figure 7.5). The 5 bp deletion/insertion polymorphism which was apparently restricted to a subset of C-type repeats was not detected by the MVR primers. Examples of alleles of *D19S20* analysed by MVR-PCR are presented in Figure 7.6.

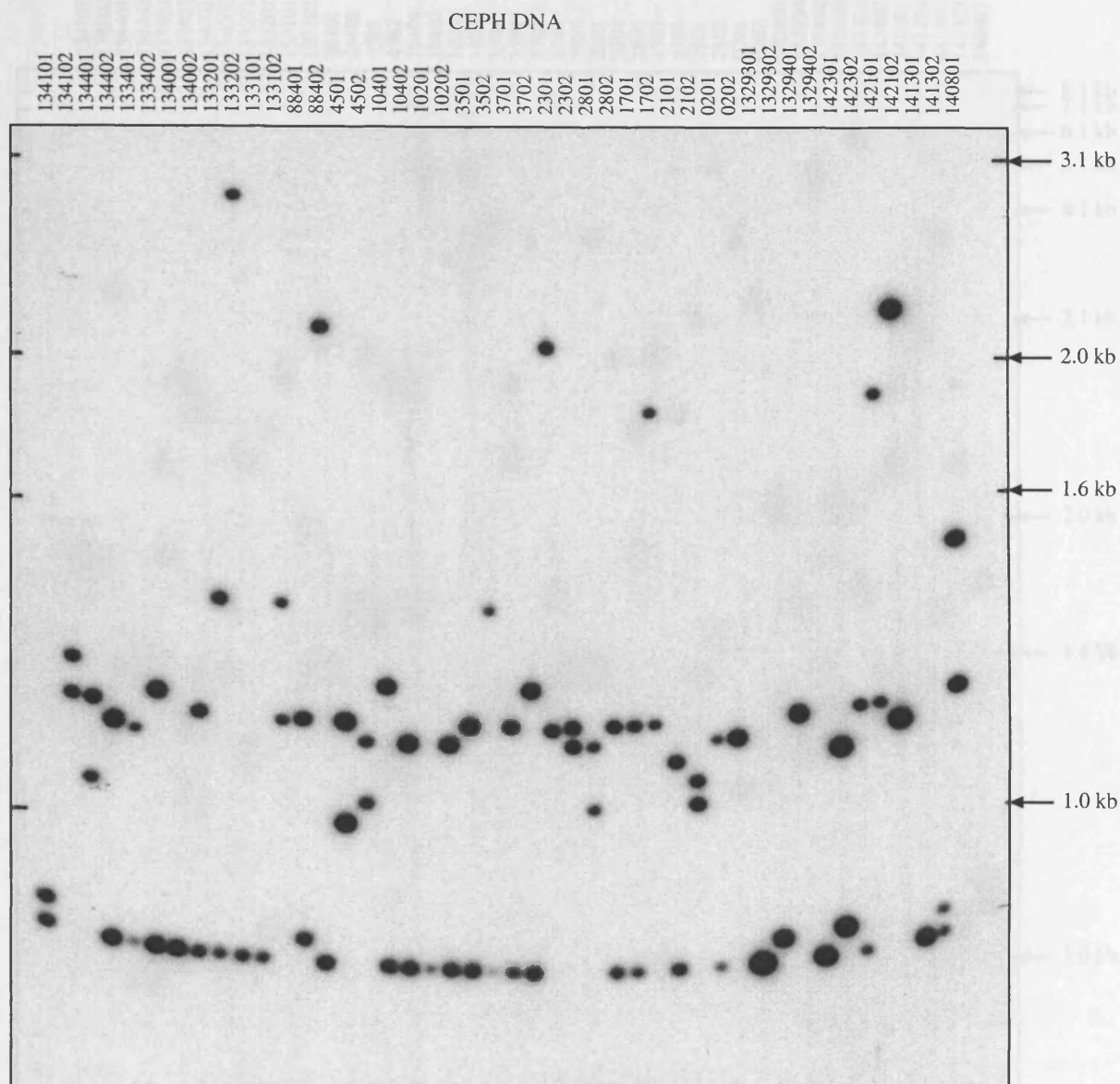
**Figure 7.2**



### ***PCR amplification of MS51***

MS51 was amplified in a panel of 45 unrelated CEPH DNAs with flanking primers 51-C and 51-D from 20 ng of genomic DNA at 96°C for 40s, 68°C for 30 s, 70°C for 3 min for 18 cycles. PCR products were electrophoresed through a 40 cm 1% agarose gel, Southern blotted, and hybridised. To amplify alleles to levels detectable on ethidium bromide-stained gels, 26 cycles of PCR were performed. Amplification efficiency was found to be greatest when reducing the input of 11.1xPCR buffer to 80% of the standard concentration, and elevating the concentration of Tris base by 20% to 14.4 mM. Alleles display a broadly trimodal size distribution ranging from ~1.3-4.1 kb with allele length heterozygosity of 82% (37/45 heterozygous individuals).

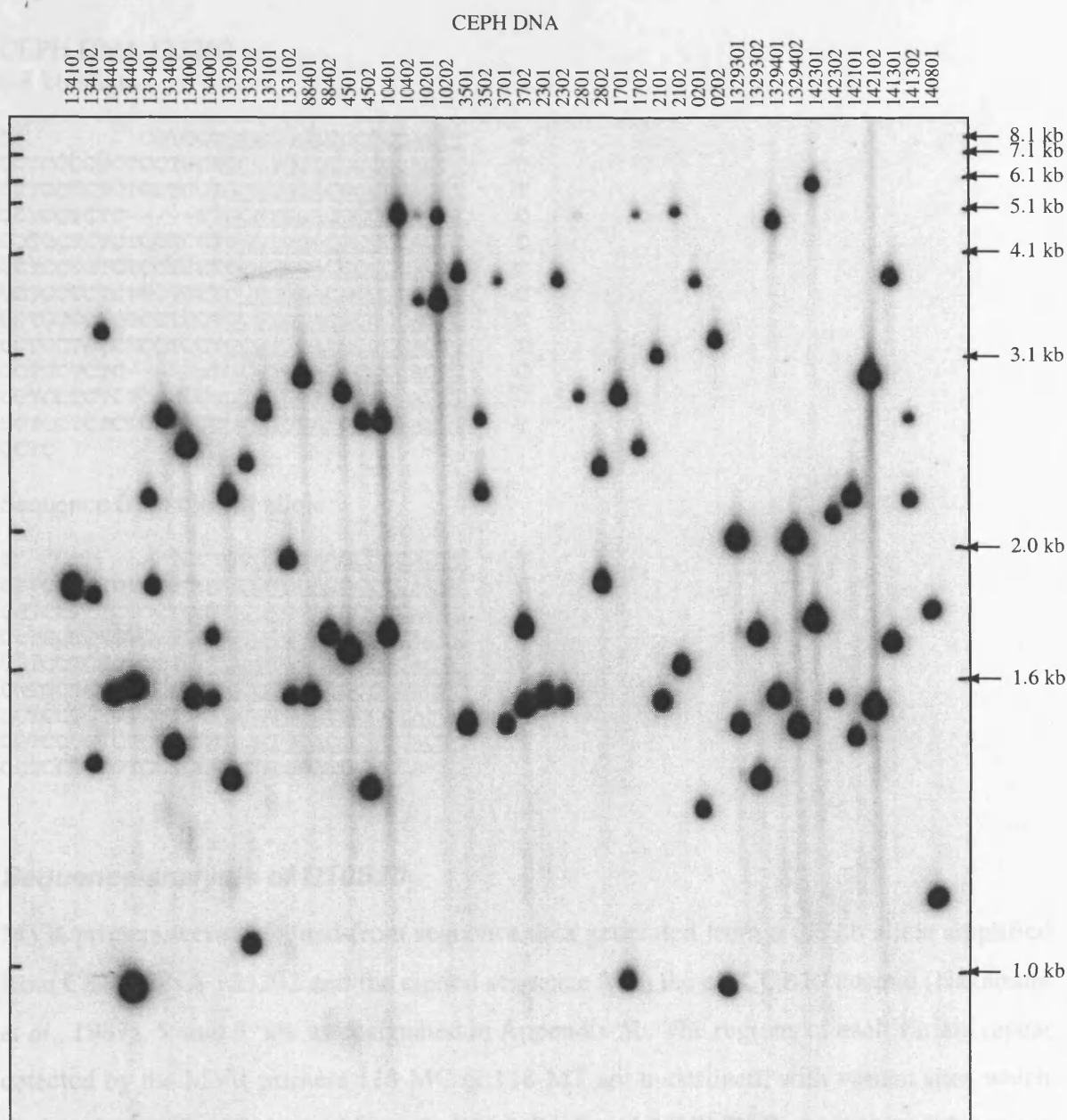
**Figure 7.3**



### **PCR amplification of D19S20**

*D19S20* was amplified in the DNA samples described in Figure 7.2 under standard conditions with primers 118-C and 118-D (118 refers to the probe identification number used by the HGMP probe bank) from 20 ng of genomic DNA at 96°C for 40s, 62°C for 30 s, 70°C for 3 min for 20 cycles. PCR products were electrophoresed through a 40 cm 1% agarose gel, Southern blotted, and hybridised. To amplify alleles to levels detectable on ethidium bromide-stained gels, 26 cycles of PCR were performed. As with MS51, alleles display a trimodal size distribution ranging from ~0.8-2.8 kb with allele length heterozygosity of 80% (36/45 heterozygous individuals).

**Figure 7.4**



#### ***PCR amplification of D17S74***

*D17S74* was amplified in the DNA samples described in Figure 7.2 under standard conditions with primers 426-C and 426-D (426 refers to the probe identification number used by the HGMP probe bank) from 20 ng of genomic DNA at 96°C for 40 s, 60°C for 30 s, 70°C for 3 min for 18 cycles. PCR products were electrophoresed through a 40 cm 1% agarose gel, Southern blotted, and hybridised. To amplify alleles to levels detectable on ethidium bromide-stained gels, 25 cycles of PCR were performed. In contrast to both MS51 and *D19S20*, *D17S74* displays a continuous allele size distribution between ~1.0 kb and ~6.0 kb, with allele length heterozygosity of 93% (42/45 heterozygous individuals). This continuous size distribution and high heterozygosity suggest a high germline mutation rate.



## Figure 7.5

CEPH DNA 133202  
0.8 kb allele

5'	CCTCCTG <u>T</u> GTGGACGCCCCACTT	O
	CCTCCCCTCTCCTCCTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTT	T
	CCTCCTCTC-----CTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTNCTCCTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTT	T
	CCTCCTCTC-----cTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTT	T
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTT	T
	CCTC	-3'

### Sequence from cloned allele

5'	TCCTGyGTGGACGCCCCACTT	T
	CCTCCtCTCTCCTCCTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTc	C
	CcTCCTCtCTCcTCCTGCGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCcTGcGTGGACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGgACGCCCCACTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCAcTC	C
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCACTt	T
	CCTCCTCTCTCCTCCTGCGTGGACGCCCCA-3'	

### Sequence analysis of D19S20

MVR primers were designed from sequence data generated from a 0.8 kb allele amplified from CEPH DNA 133202 and the cloned sequence from the cMCOB19 cosmid (Nakamura *et al.*, 1987). 5' and 3' are as designated in Appendix 5b. The regions of each variant repeat detected by the MVR primers 118-MC or 118-MT are underlined, with variant sites which distinguish repeat types double underlined. Predicted MVR-PCR repeat-type identity is listed at right.

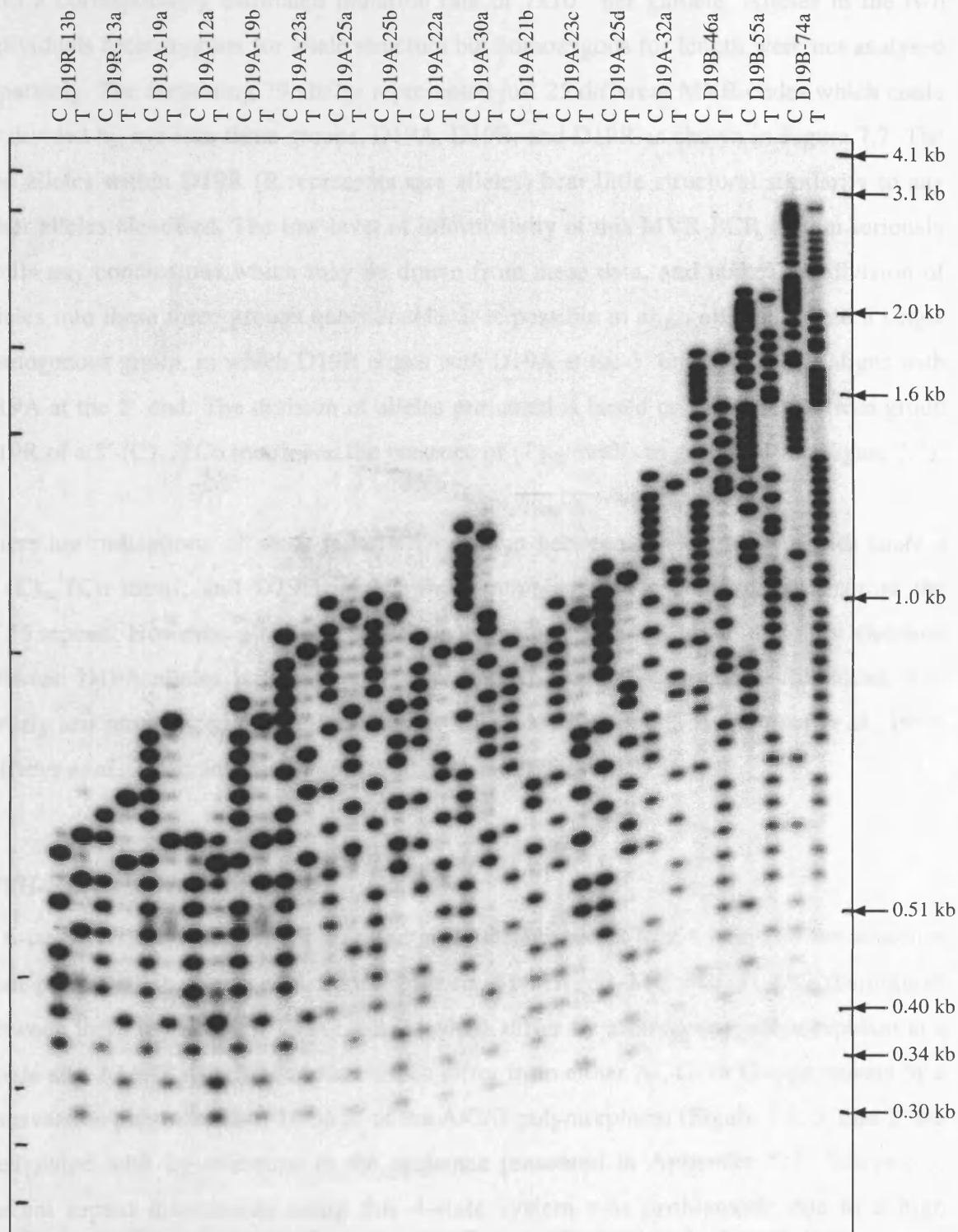
## Figure 7.6

### ***MVR-PCR at D19S20***

A system of two-state MVR-PCR was established at *D19S20* with C- and T-type repeats identified. General MVR-PCR conditions were as described in Chapter 2. To separate alleles in heterozygotes, samples were amplified to levels visible on ethidium bromide-stained gels as described in Figure 7.3 and bands excised from the agarose gel. DNA was gel purified using home-made spin columns and diluted x10 prior to MVR-PCR analysis. Alleles were MVR mapped with flanking primer 118-D and MVR primers 118-MC or 118-MT at 96°C for 40 s, 65°C for 30 s, 70°C for 2 min for 15 cycles. PCR products were electrophoresed through a 40 cm 1.2% agarose gel, Southern blotted, and hybridised.

MVR-PCR analysis of 17 alleles is presented, and includes both the smallest (D19A-12a) and largest (D19B-74a) alleles detected in this study. Allele names reflect allele group (defined in Figure 7.7), repeat number, and a further discriminating letter so that D19R-13b is the second (b) allele of 13 repeats identified in group D19R. Null repeats (unamplifiable or poorly amplifiable repeats due to additional polymorphisms which prevent efficient annealing of the MVR primers) can be clearly distinguished in many alleles as faint bands, for example at the top of alleles D19R-13b and D19A-19a.

**Figure 7.6**



All 81 alleles of *D19S20* defined by size from the 45 CEPH DNAs amplified previously (Figure 7.3) were typed by MVR-PCR. Nine individuals were homozygous for allele length. MVR analysis revealed two of these nine homozygotes to be heterozygous for allele structure resulting in a corrected heterozygosity of 84% (38/45 heterozygous individuals) with a corresponding estimated mutation rate of  $7 \times 10^{-5}$  per gamete. Alleles in the two individuals heterozygous for allele structure but homozygous for length were not analysed separately. The remaining 79 alleles represented just 25 different MVR codes which could be divided by eye into three groups, D19A, D19B, and D19R as shown in Figure 7.7. The two alleles within D19R (R represents rare alleles) bear little structural similarity to any other alleles identified. The low level of informativity of this MVR-PCR system seriously limits any conclusions which may be drawn from these data, and makes the division of alleles into these three groups questionable. It is possible to align all alleles into a single homogenous group, in which D19R aligns with D19A at the 3' end, and D19B aligns with D19A at the 5' end. The division of alleles presented is based on the absence from group D19R of a 5' (C)<sub>5,6</sub>TCo motif, and the presence of (T)<sub>7,8</sub> motifs in group D19B (Figure 7.7).

There are indications of weak polarised variation between alleles. Most alleles share a 5' (C)<sub>5,6</sub>TCo motif, and D19B alleles share complete MVR code identity across the 5' 35 repeats. However, all alleles also share a common 3' (T)<sub>2,3</sub>C motif and most variation between D19A alleles is located centrally. If there is polar variation at this locus, it is clearly less pronounced than at loci such as MS32, MS205, or MS31 (Armour *et al.*, 1993; Jeffreys *et al.*, 1991; Jeffreys *et al.*, 1990; Neil and Jeffreys, 1993).

### ***MVR-PCR analysis of MS51***

A 6-state system of MVR-PCR was ultimately established at MS51 based on the sequence data presented in Figure 7.8. MVR primers 51-MA, 51-MC and 51-MG distinguish between three repeat types (A, C and G) which differ by a three-way polymorphism at a single site. 51-MT identifies repeats which differ from either A-, C- or G-type repeats by a transversion polymorphism 14 bp 3' of the A/C/G polymorphism (Figure 7.8, 5' and 3' are designated with by reference to the sequence presented in Appendix 5c). Analysis of variant repeat distribution using this 4-state system was problematic due to a high frequency of null repeats. MS51 repeats are 25 and 33 bp in length. A gap of ~100 bp in the

## Figure 7.7

### ***MVR-PCR codes at D19S20***

MVR codes were read 3' to 5' from the bottom to top of the autoradiographs. Seventy-nine alleles were typed by 2-state MVR-PCR with the identification of 25 different MVR codes. Repeat distribution patterns of all alleles were aligned by eye, and hyphens were inserted to improve alignments. Alleles were divided by eye into three groups, D19A, D19B, and D19R. D19R is a heterogeneous group of **R**are alleles which bear comparatively little structural similarity to any other allele analysed. Allele names are as described in Figure 7.6. The number of copies of each allele detected in this sample is presented.

**Figure 7.7**

Allele	Copies	MVR code (3' -> 5' orientation)
D19R-13b	1	TTCCCCCTCTCToo
D19R-13a	2	TTCTCCCCCTCCT
D19A-12a	27	TTc-----TCCCCCTCo
D19A-13a	1	TTTc-----TCCCCCTCo
D19A-19a	1	TTCT-CCCCC-----TCCCCCToo
D19A-20a	1	TTCTCCCCCCC-----TCCCCCTCo
D19A-18a	1	TTc---CCCC---TCT-----CCCCCTCo
D19A-21a	1	TTcCT-CCCC---TCT-----CCCCCTCo
D19A-19b	3	TTc-T-CCCC---TCT-----CCCCCTCo
D19A-23a	2	TTc-T-CCCCCTCC-TCT-----CCCCCTCo
D19A-24a	1	TTc--CTCCCCCTCTTCT-----CCCCCTCo
D19A-25a	1	TTTc--CTCCCTCTCTTCT-----CCCCCTCo
D19A-23b	2	TTc--CCCTCTCTCTT-----CCCCCTCo
D19A-25b	1	TTTc--CCCCTTCTCTTCT-----CCCCCTCo
D19A-30a	4	TTc-TCCCCCTTCCCTTCTCTCTT-CCCCCTCo
D19A-22a	8	TTc-----CCTTCTCTCTT-CCCCCTCo
D19A-25c	1	TTc-----CCCCTTCTCTCTTCCCCCTCo
D19A-23c	12	TTc-----CCTTCTCTCTTCCCCCTCo
D19A-25d	1	TTc-----TCCCTTCTCTCTTCCCCCTCo
D19A-32a	1	TTCCCTTCTCTCCCTTCTCTCTTCCCCCTCo
D19A-21b	1	TTT-----TTCTCTCTTCCCCCTCo
D19B-46a	2	TTc-----TTTTTTTc-----TTCCCTTCTCTCCCCCTTCTCTCTCTCCCCCTCo
D19B-53a	1	TTc-----TTTTTTTCT-----CCTTTTTTCCCTTCTCTCCCCCTTCTCTCTCTCCCCCTCo
D19B-55a	2	TTc-----TC-TTTTTTTCCCCCTTTTTTCCCTTCTCTCCCCCTTCTCTCTCTCCCCCTCo
D19B-74a	1	TTCCCTTCTTCTTTTTTTTTCTTCTTTTTTTTCCCCCTTTTTTCCCTTCTCTCCCCCTTCTCTCTCTCCCCCTCo

## Figure 7.8

### ***Sequence analysis of MS51***

Partial sequence data were generated from 3 small (1.2-1.5 kb) and one large (~3.0 kb) alleles of MS51. 5' and 3' are designated as in Appendix 5c. The 3.0 kb allele was too large for the central region to be sequenced from primers flanking the repeat array. The allele was therefore amplified to levels visible on an ethidium bromide-stained gel using 0.4  $\mu$ M flanking primer 51-C and MVR primer 51-MC. Bands were excised and sequenced using the TAG primer. Regions of each repeat detected exclusively with either 51-MA, 51-MC, 51-MG, or 51-MT are underlined. Regions detected exclusively by 51-MB are underlined in bold. Regions detected by 51-MB and another MVR primer are underlined twice. A subset of A-type repeats would not be detected by 51-MB due to a transversion polymorphism 18 bp from the 5' end of the repeat unit. These repeats were denoted 'a'. The predicted repeat type as defined by MVR-PCR is presented (separated from the repeat by ' . . . . . ').

Heterogeneity in repeat unit length is likely to have arisen from duplication of part of a repeat unit. For example in the 1.3 kb allele from CEPH DNA 134101, the first 33 bp repeat from the 5' end contains an additional TGGCAGGA sequence which is identical to the first 8 bases of an A-type repeat. This sequence identity continues for a total of 15 bp making the definition of the start and end of each repeat problematic. The division of the sequence into repeat units presented is based on the identity of the residue 16 bp from the 5' end of the repeat unit.

**Figure 7.8**

**CEPH DNA 134101**  
1.3 kb allele

T...5'-GGGCAGGAGGGCAAGTGAgGGACA  
o....TGGCAGGAGGGCAGGTGAaGGACA  
B....CAGCACGAGGGCAGGTGGAGGGACA  
G....CGGCAGGAGGGCAGGTGAGGGACA  
A....TGGCAGGAGGGCAGGTGAGGGACA  
A....TGGCAGGAGGGCAGGTGAGGGACA  
a....TGGCAGGAGGGCAGGTGAAGGGACA  
G....CGGCAGGAGGGCAGGTGAGGGACA  
G....CGGCAGGAGGGCAGGTGAGGGACATGGCAGGA  
C....GGGCAGGAGGGCAGGTGAGGGACG  
A....TGGCAGGAGGGCAGGTGAGGGACA  
A....TGGCAGGAGGGCAGGTGAGGGACA  
A....TGGCAGGAGGGCAGGTGAGGGACATGGCACGA-3'

3'-TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....B  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....C  
TCcTGCCATgTccCTCCACCTGCCCTCCTGCCG....C  
TCATGCCATGTCCcCTCCACTTGCCCTCCTGCCG....T  
TCATGCCATGTCCCTCCACTTGCCCTCCTGCCG-5'..T

**CEPH DNA 4502**  
1.5 kb allele

a..5'-TGGCAGGAGGGCAGGTGAAaGGACA  
B....CaGcAGGAGGGCAGGTGGAGGGACA  
G....CGGCAGGAGGGCAGGTGAAGGGACA  
A....TGGCAGGAGGGCAGGTGAAGGGACA  
a....TGGCAGGAGGGCAGGTGAAaGGACA  
B....CAGCAGGAGGGCAGGTGGAGGGACA  
G....CGGCAGGAGGGCAGGTGAAGGGACATGGCAGGA  
C....GGGCAGGAGGGCAGGTGAAGGGATG  
A....TGGCAGGAGGGCAGGTGAAGGGACA  
A....TGGCAGGAGGGCAGGTGAAGGGACA  
A....TGGCAGGAGGGCAGGTGAAaGGACA  
G....CGGCAGGAGGGCAGGTGAAGGGACA  
G....CGGCAGGAGGGCAGGTGAAGGG-3'

3'-TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....B  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG-5'..G

**CEPH DNA 2302**  
1.2 kb allele

A..5'-TGGCAGGAGGGCAGGTGAGGGACA  
A....TGGCAGGAGGGCAGGTGAGGGACA  
a....TGGCAGGAGGGCAGGTGAAGGGACA  
G....CGGCAGGAGGGCAGGTGAGGGACA  
G....CGGCAGGAGGGCAGGTGAGGGACATGGCAGGA  
C....GGGCAGGAGGGCAGGTGAGGGACG  
A....TGGCAGGAGGGCAGGTGAGGGACA  
A....TGGCAGGAGGGCAGGTGAGGGACA  
A....TGGCAGGAGGGCAGGTGAGGGACATGGCAGGA  
C....GGGCAGGAGGGCAGGTGAGGGACATGGCAAGA-3'

3'-TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....B  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACCTGCCCTCCTGCCG....C  
TCCTcGCCATGTCCCTCCACCTGCCCTCCTGCCG....C  
TCATGCCATGTCCCTCCACTTGCCCTCCTGCCG-5'..T

**CEPH DNA 134101**  
3.0 kb allele (sequenced with the TAG primer)

3'-TCGTGCCATGTCCCTCCACTTGCCCTCCTGCCG....T  
TCCTGCCATGTCCCTCCACTTGCCCTCCTGCCG....T  
TcCTGCCATGTCCCTCCACTTGCCCTCCTGCCG....A  
TGTCCCTCCACCTGCCCTCCTGCCG....G  
TGTCCCTCCACTTGCCCTCCTGCCG....T  
TcCTGTGTGTGTGTGTCCACTTTCCTCCTGCCG....T  
CTcCTGCCATGTCCCTCCACTTGCCCTCCTGCCG....o  
TGNCCCTCCACTTGCCCTCCTGCCG....B  
NGTCTGTCTCACTTGCCCTCCTGCCG....T  
TGTCCCTTCACTTGCCCTCCTGCCG....a  
TGTCCCTCCACTTGCCCTCCTGCCG....A  
CGTNCCTCCACTTGCCCTCCTGCCG....C  
TCCTGCCATGTCCCTTCAcCTGcCTCTGCCG....o  
CGTTCCTCCACTTGCCCTCCTGCCG-5'..C

3'-TCATGCCATGTCCCTCCACTTGCCCTCCTGCCG....T  
TCATGCCATGTCCCTCCACTTGCCCTCCTGCCG....T  
TcCTGCCATGTCCCTCCACTTGCCCTCCTGCCG....T  
TGCCCTCCACTTGCCCTCCTGCCG....B  
TGTCCCTCCACTTGCCCTCCTGCCG....G  
TGTCCCTCCACTTGCCCTCCTGCCG....A  
CGTCCCTCCACTTGCCCTCCTGCCG....C  
TCCTGCCATGTCCCTCCACTTGCCCTCCTGCCG-5'..C



MVR code could therefore represent either 3 or 4 consecutive null repeats. In addition, without careful measurement it was difficult to determine the size of a region containing null repeats by comparing band positions between lanes on the autoradiograph. A fifth primer, 51-MB, was therefore designed to detect the majority of repeats to facilitate the scoring of consecutive repeat types.

Predicted MVR codes of sequenced alleles are presented in Figure 7.8. 51-MB detects all C- and G-type repeats, and most A-type repeats. The transition polymorphism 18 bp from the 5' end of the repeat which prevents 51-MB binding was only associated with the A-residue which defines all A-type repeats. Such co-association of specific variants at different polymorphic sites within repeats appears to be unusual for minisatellites (Tamaki *et al.*, 1999b; Tamaki *et al.*, 1993). This polymorphism divided A-type repeats into two further repeat types named A and a. Repeats detected exclusively by 51-MB were named as B-type repeats. Due to potential difficulties in using an MVR primer which detected too many repeats (possibly resulting from interaction between amplicons during the simultaneous amplification of multiple amplicons of different sizes), 51-MB was designed to not detect T-type repeats (Figure 7.8). Examples of alleles of MS51 analysed by MVR-PCR are presented in Figure 7.9.

All 82 alleles from the 45 CEPH DNAs analysed in Figure 7.2 were typed by MVR-PCR. No individuals homozygous for allele length but heterozygous for MVR code were identified, so the true heterozygosity remained unchanged at 82% (37/45 heterozygous individuals). From the 82 alleles, 29 different structures were defined by MVR-PCR, and alleles were aligned by eye and divided into 4 groups, 51A, 51B, 51C and 51D (Figure 7.10a). The six variant repeats identified produced highly informative allele structures. Variation within allele groups was mainly due to either switches in repeat identity within regions containing higher order repeat motifs, or the deletion or duplication of repeat motifs.

There was substantial evidence for duplications, especially amongst the larger group 51A alleles. For example, allele 51A-103a (allele 'a' of 103 repeats in group 51A) differed from other 51A alleles by the complex reduplication of a 5' motif (Figure 7.10a). 51A alleles

## Figure 7.9

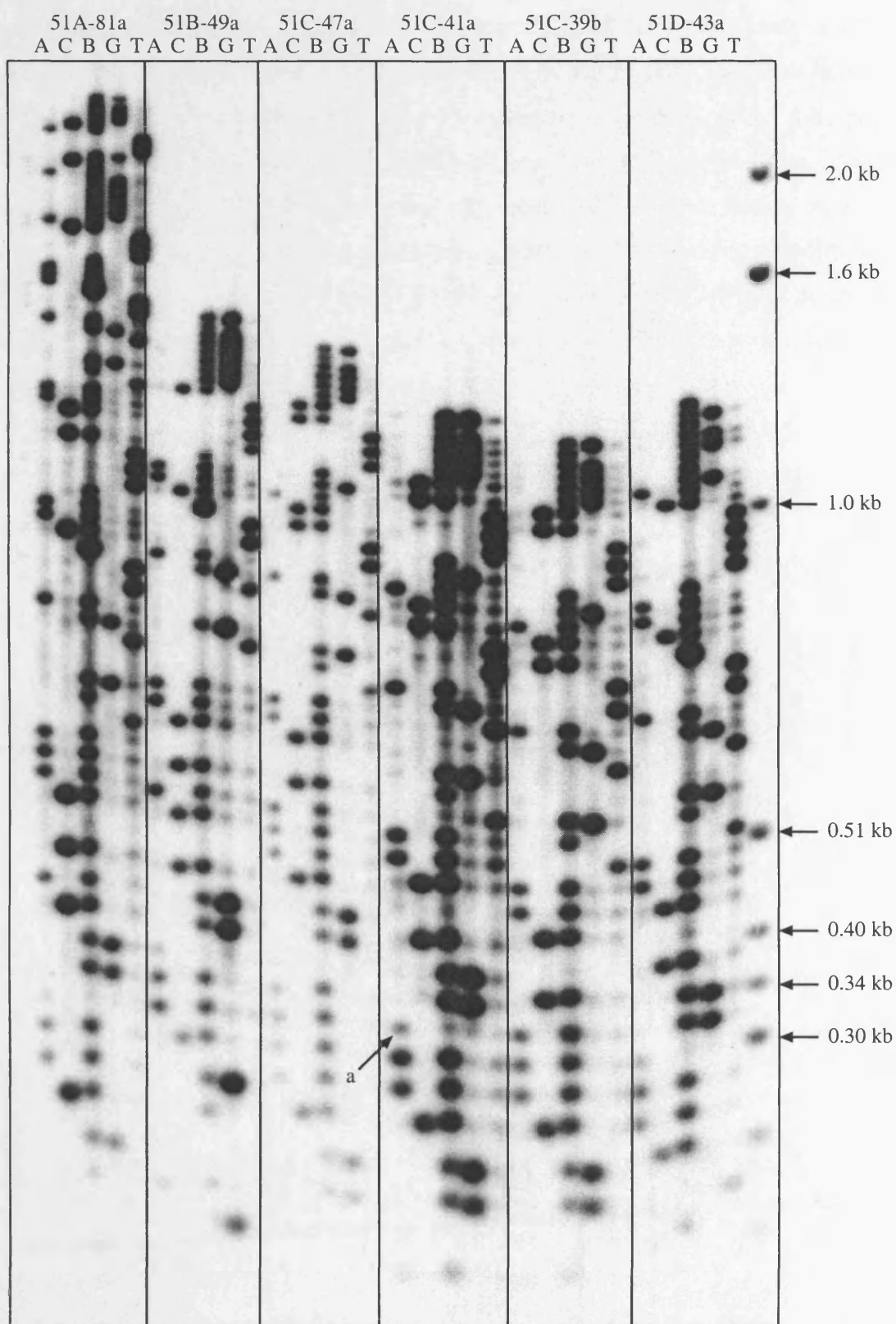
### ***MVR-PCR at MS51***

A system of 6-state MVR-PCR was established at MS51 using 5 MVR primers. Prior to MVR-PCR, alleles were separated as in Figure 7.6, and samples diluted x200. Due to variation between the regions within variant repeats detected by different MVR primers, no one set of MVR-PCR conditions could be attained for optimal specificity of all primers. Whilst MVR-PCR could be performed at 96°C for 40 s, 65°C for 30 s, 70°C for 2 min for 21 cycles (using the same modified buffer as described for flanking amplifications; see Figure 7.2), greater specificity was obtained when adjusting annealing temperature, cycle number, and primer concentration independently for each primer as described below. MVR-PCR was performed using flanking primer 51-C.

Primer	Annealing temperature	Cycle number	Primer concentration (nM)
51-MA	67°C	22	2
51-MB	65°C	22	20
51-MC	67°C	20	3
51-MG	67°C	20	5
51-MT	65°C	21	5

MVR-PCR at 6 alleles is presented. Allele names reflect allele group (defined in Figure 7.10), repeat number, and a further discriminating letter so that 51A-81a is the first (a) allele of 81 repeats detected in group 51A. All repeats detected by primers 51-MC and 51-MG were also detected by 51-MB. Repeats detected by 51-MA but not 51-MB were denoted 'a', and example of which is indicated. Only repeats detected exclusively by 51-MB were called B-type repeats. The first 1-3 repeats of each allele were not visible on the autoradiograph presented.

**Figure 7.9**



## Figure 7.10

### ***MVR-PCR codes at MS51***

MVR codes were read 5' to 3' from bottom to top of the autoradiographs. Eighty-two alleles were analysed with the identification of 29 different MVR codes. Repeat distribution patterns of all alleles were aligned by eye, and hyphens were inserted to improve alignments. Alleles 51A-103a and 51A~140a are divided over several lines to highlight regions which are likely to have been generated by duplication. Allele names are as described in Figure 7.9. The repeat number of allele 51A~140a was estimated as the allele was too large to be completely analysed by MVR-PCR, as is indicated by >>. '?' denotes repeats which could not be identified as they represent small, weakly hybridising bands upon Southern blot hybridisation. 'o' denotes null repeats. Alleles were divided by eye into four groups, 51A, 51B, 51C and 51D (Figure 7.10a). The number of copies of each allele is presented. Alleles of each group could be meaningfully aligned as is indicated in Figure 7.10b where the most common alleles from each lineage are aligned by the insertion of hyphens and division of alleles over multiple lines.

**Figure 7.10**

<b>a</b>		
Allele	Copies	MVR code (5' -> 3' orientation)
51A-103a	1	BGCAAAaGG CAAa BGCAAAaG B--AAa BGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATT-----BAAABTTT CAGGGGGGGGAB--CGTTT A-C---GGGGBG
51A-83a	1	AaBGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATT-----BAAABTTT CAGGGGGGGGAB--oCGTTT A-C---GGGGBG
51A-81a	1	AAaBGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATT-----BAAABTTT CAGGGGGGGG--A--CGTTT A-C---GGGGBG
51A-82a	1	?AA-BGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATT-----BAAABTTT CAGGGGGGGG-A--CGTTT A-C---GGGGBG
51A-87a	1	AaBGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATT-----BAAABTTT CAGGGGGGGG-ACCGGTTT AACCGGGGGGBG
51A-93a	1	AaBBCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATTTCAAAABTTT CBAAABTTT CAGGGGGGGGAB--CGTTT A-C---GGGGBG
51A-93b	1	AaBGCAAAaGGCACoCAAA BCo GATTBCAABTTTCoCAA BGo CATTCAAAABTTT CBAAABTTT CAGGGGGGGGA--CGTTT A-C---GGGGBG
51A-93c	2	AAaBGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATTTCAAAABTTT CBAAABTTT CAGGGGGGGG-A--CGTTT A-C---GGGGBG
51A-92a	1	???BGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATTTCAAAABTTT C-AAAABTTT CAGGGGGGGG-A--CGTTT A-C---GGGGBG
51A-91a	1	AaBGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATTTCAAAABTTT C-AAAABTTT CAGGGGGGGG-A--CGTTT A-C---GGGGBG
51A-140a*	1	AaBGCAAAaGGCACoCAAA BGo GATTBCAABTTTCoCAA BGo GATTTCAAAABTTT CBAAABTTT CAGGGG BTTTCoCAA-BGo GATTTCAAA-BTTT CBA>>
51B-49a	4	?BGAaBGCAAAaGGCoCACoCAAAo Go GATTBCAABTTTTCGGGGGGBG
51B-49b	1	?BGAaBGCAAAaGGCoCACoCAAAo Go GATTBCAABTTTTCGGGGGGBG
51B-48a	15	BGAaBGCAAAaGGCoCACoCAAAo Go GATTBCAABTTTTCGGGGGGBG
51B-48b	1	BGAaBGCAAAaGGCoCACoCAAAo Go GATTBCAABTTTTCGGGGGGBG
51B-48c	1	BGAaBGCAAAaGGCoCACoCAAAo Go GATTBCAABTTTTCGGGGGGBG
51B-47a	1	BGAaBGCAAAa-GCoCACoCAAAo Go GATTBCAABTTTTCGGGGGGBG
51C-41a	3	?AaGGC--AAaGG---CoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-40a	4	AaGGC--AAaGG---CoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-39a	1	AAaGGC--A--GG---CoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-38a	1	AaGGC--A--GG---CoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-47a	4	AAaGGCAAAAaGGCAAACoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-46a	5	AaGGCAAAAaGGCAAACoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-38b	1	AAaGGCAAACoCAA BGo GATTCC-GBTTTTCGGGGGBG
51C-42a	1	???AAaGGCAAACoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-39b	16	AAaGGCAAACoCAA BGo GATTCCAGBTTTTCGGGGGBG
51C-38c	8	AaGGCAAACoCAA BGo GATTCCAGBTTTTCGGGGGBG
51D-43a	2	AaB CAAaGGCoCAA BGo GATTBCAABTTTTCAGBBGGGBG
51D-42a	1	??B CAAaGGCoCAA BGo GATTBCAABTTTTCAGBBGGGBG
<b>b</b>		
51A93c		AAaBGCAAAaGG--CACoCAAA BGo GATTBCAA--BTTT CoCAA--BGo GATTTCAAAABTTT CBAAABTTT CAGGGGGGGG ACGTTT-A C--GGGGBG
51B48a		BGAaBGCAAAaGGCoCACoCAAAo Go GATTBCAA--BTTTTCGGGGGBG
51C39b		AAaGGCAAACoCAA BGo GATTCCA--GBTTTTC--GGGGBG
51D43a		AaB CAAaGG---CoCAA BGo GATTBCAA--BTTTTCAGBBGGGBG

could also be split into those with and those without a central duplication of an (A)<sub>3</sub>B(T)<sub>3</sub>C motif. The similarity of this motif to a region immediately 3' strongly argues in favour of this polymorphism arising by duplication and not deletion.

It is also striking that alleles from all groups contained all six variant repeats detected in this study, and that all allele structures could be meaningfully aligned (Figure 7.10b). Group 51D is perhaps the most distantly related group of alleles. While they share common central motifs, they differ substantially from all other alleles at both the 5' and 3' ends. Overall, there is little evidence for polarity of variation at this locus with most alleles sharing a common 3' (G)<sub>4-6</sub>BG motif, and variations of a common 5' AaGGC motif.

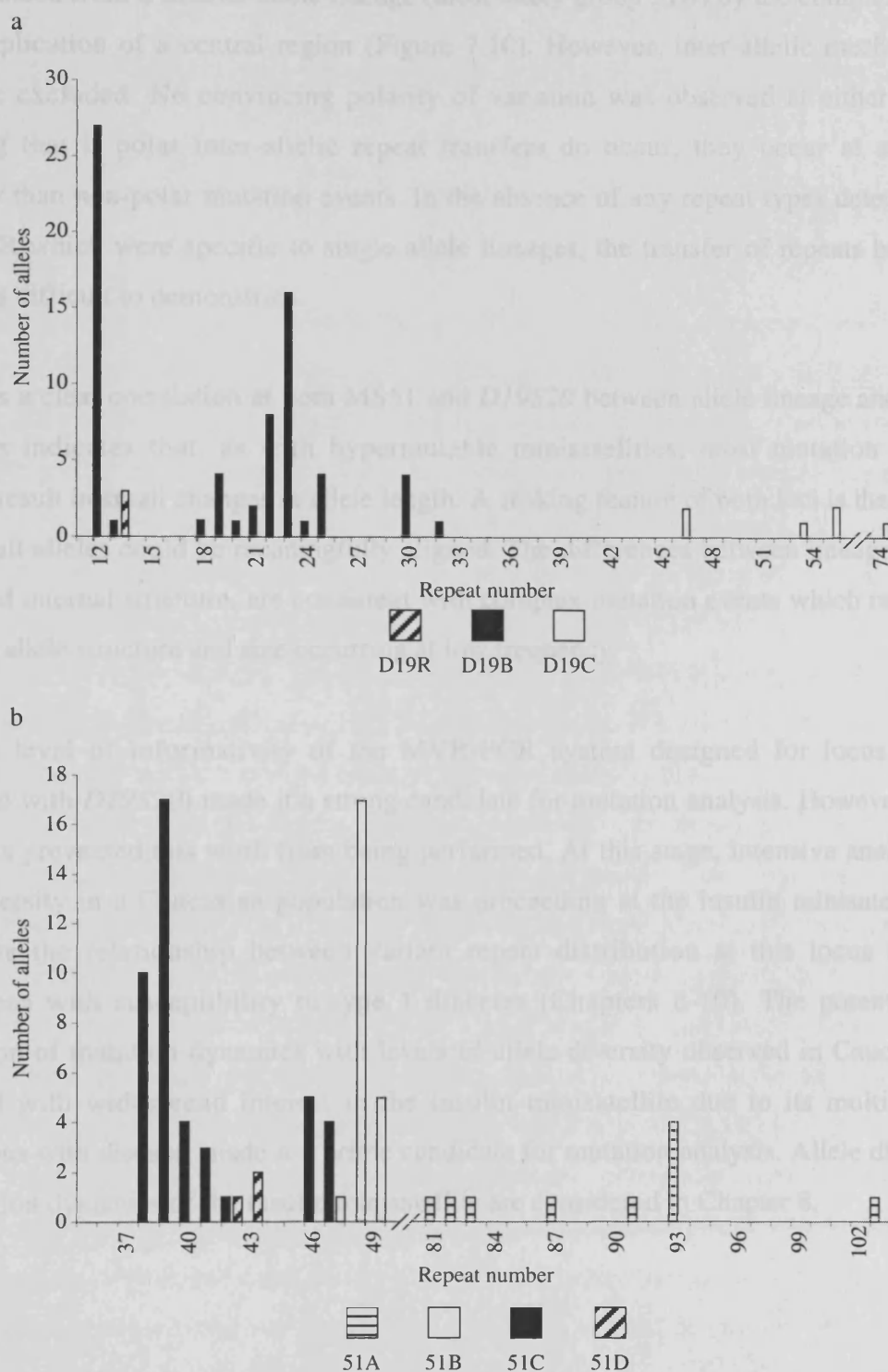
### ***Relationship between allele lineage and size***

Both *D19S20* and MS51 display broadly trimodal allele size distributions which reflect different allele lineages as defined by MVR code (Figure 7.11). For example at MS51, the smallest group of alleles represents almost exclusively allele lineage 51C, with the medium size group dominated by 51B. Large alleles at this locus all lie within group 51A (Figure 7.11a). Similarly, at *D19S20* the largest alleles represent a structurally distinct group, whilst the two clusters of smaller alleles mainly represent different alleles from group D19B. Of the 31 alleles within the smallest size group, 27 have identical MVR structures (D19B-12a; Figure 7.7). This restriction of alleles within each lineage to a small size range indicates that most mutations at both loci involve small changes in repeat number.

## **Discussion**

Differences between patterns of variant repeat distribution in alleles of a minisatellite from a population reflect historical mutation events. Without knowledge of the identity of the progenitor alleles from which mutant structures were derived, any inferences as to the mechanisms which generated observed levels of variation must in general be treated with caution. Nevertheless, analysis of both MS51 and *D19S20* found evidence of duplication events which were most likely of intra-allelic origin. For example at *D19S20* multiple copies of a (T)<sub>7-8</sub> motif were identified within most alleles of group D19B, but was absent

**Figure 7.11**



**Correlation between allele size and lineage at D19S20 and MS51**

Alleles of *D19S20* and *MS51* were divided by eye into groups as described in Figure 7.7 and Figure 7.10 respectively. Allele size generally correlates with lineage for both *D19S20* (Figure 7.11a) and *MS51* (Figure 7.11b). On each bar chart, the x-axis is split as indicated to exclude size windows within which no alleles were identified.

from any other lineage (Figure 7.7), whilst at MS51 the largest group of alleles (51A) may have expanded from a smaller allele lineage (most likely group 51B) by the complex intra-allelic duplication of a central region (Figure 7.10). However, inter-allelic mechanisms cannot be excluded. No convincing polarity of variation was observed at either locus, indicating that if polar inter-allelic repeat transfers do occur, they occur at a lower frequency than non-polar mutation events. In the absence of any repeat types detected by MVR-PCR which were specific to single allele lineages, the transfer of repeats between lineages is difficult to demonstrate.

There was a clear correlation at both MS51 and *D19S20* between allele lineage and allele size. This indicates that, as with hypermutable minisatellites, most mutation events probably result in small changes in allele length. A striking feature of both loci is that MVR codes of all alleles could be meaningfully aligned. The differences between lineages, both in size and internal structure, are consistent with complex mutation events which radically alter both allele structure and size occurring at low frequency.

The high level of informativity of the MVR-PCR system designed for locus MS51 (compared with *D19S20*) made it a strong candidate for mutation analysis. However, time constraints prevented this work from being performed. At this stage, intensive analysis of allele diversity in a Caucasian population was proceeding at the insulin minisatellite to investigate the relationship between variant repeat distribution at this locus and its associations with susceptibility to type 1 diabetes (Chapters 8-10). The potential for comparison of mutation dynamics with levels of allele diversity observed in Caucasians, combined with widespread interest in the insulin minisatellite due to its multifarious associations with disease, made it a prime candidate for mutation analysis. Allele diversity and mutation dynamics of the insulin minisatellite are considered in Chapter 8.



## Chapter 8

# Allele diversity and germline mutation at the insulin minisatellite

### Summary

The insulin minisatellite has been intensively studied due to its associations with type 1 diabetes, type 2 diabetes, polycystic ovary syndrome, adult obesity and infant birth size. However, little is known of levels of structural diversity at this locus. Furthermore, its bimodal allele size distribution in Caucasian populations, combined with strong linkage disequilibrium surrounding the minisatellite, both indicate that mutation processes operating at the insulin minisatellite may be very different to those which operate at previously characterised hypermutable loci. A system of MVR-PCR was established to analyse allele diversity within a Caucasian cohort. Complete MVR maps were generated for 876 alleles with the identification of 189 different codes, almost all of which could be assigned to one of three very distinct lineages. The dominant form of mutation appeared to involve small insertions and deletions of repeats primarily in homogeneous regions of the repeat array, occurring at a frequency of perhaps  $10^{-3}$  per gamete. There was also evidence for more complex mechanisms of repeat turnover involving inter-allelic transfers and complex intra-allelic duplications. Mutation mechanisms were analysed directly by the detection of mutants in blood and sperm DNA. Two forms of mutation were identified, the major resulting in simple deletions and duplications which were at least partially of premeiotic origin, and a minor form occurring at a frequency of  $\sim 2 \times 10^{-5}$  per sperm which results in complex intra- and inter-allelic rearrangement of repeats similar to mutation processes observed at hypermutable minisatellites, and almost certainly of meiotic origin. Characterisation of germline mutation in an individual homozygous for both allele length and MVR code demonstrated that homozygosity reduced neither the rate nor complexity of germline mutation.

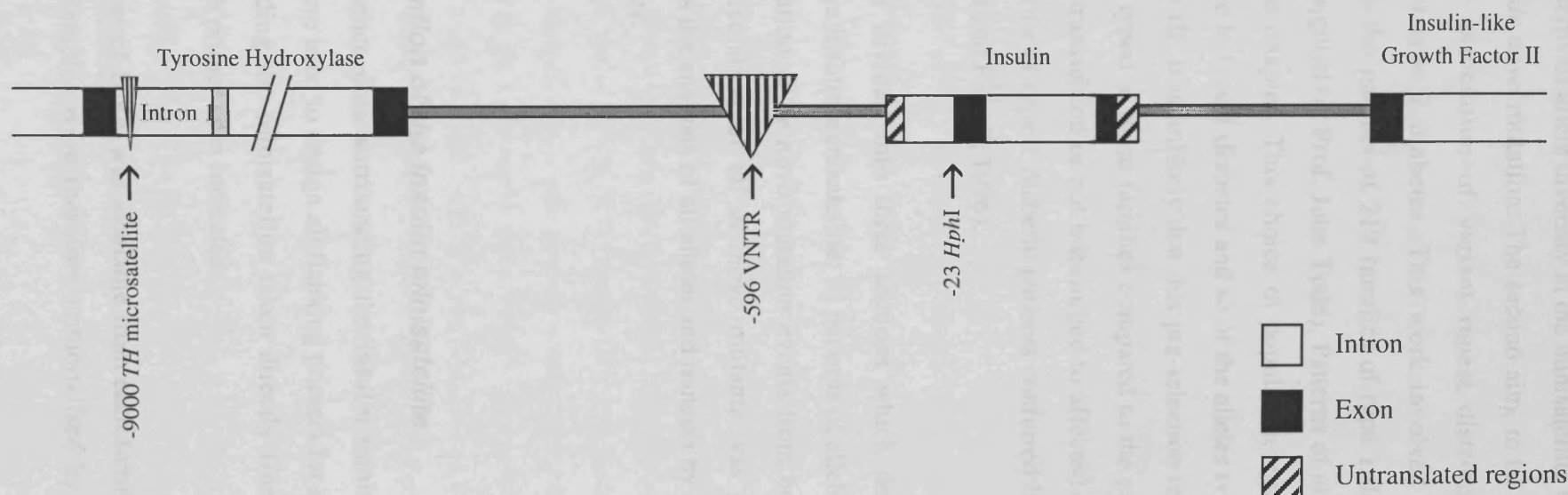
## Introduction

The insulin minisatellite, located 596 bp 5' of the insulin gene translation initiation site on chromosome 11p15.5 (Figure 8.1), was the first minisatellite to be sequenced (Bell *et al.*, 1982). It displays a bimodal size distribution in Caucasian populations with small class I alleles (28-44 repeats) and large class III alleles (138-159 repeats) representing ~70% and ~30% of all alleles respectively (Bell *et al.*, 1984). Class II alleles of intermediate size are rare in Caucasian populations (Bennett and Todd, 1996a). Fourteen variant repeats of 14-15 bp in length (named a-n) based on the consensus sequence ACAGGGGTGTGGGG have been identified (Bennett and Todd, 1996a). Of these 14 variants, 2 (l and m) were found exclusively in an allele of African-American origin (Rotwein *et al.*, 1986), whilst variant n has only been found in chimpanzees (Seino *et al.*, 1992).

In 1984, an association was identified between the insulin minisatellite and susceptibility to type 1 diabetes (Bell *et al.*, 1984). Class I alleles associate with predisposition to type 1 diabetes whilst class III alleles associate with dominant protection. More recently, the locus has also been associated with predisposition to type 2 diabetes (Ong *et al.*, 1999), polycystic ovary syndrome (PCOS) (Waterworth *et al.*, 1997), adult obesity (O'Dell *et al.*, 1999), and variation in birth size (Dunger *et al.*, 1998). Transmission ratio distortion has also been reported, with class I alleles transmitted to unaffected offspring from class I/III heterozygous parents at an average frequency of 54% (Eaves *et al.*, 1999). It has been proposed that the minisatellite acts as a transcriptional regulator both of the insulin gene and *IGF2* (Paquette *et al.*, 1998), and that this regulation may be affected by the composition of different variant repeats within alleles (Awata *et al.*, 1997; Bennett *et al.*, 1995; Catasti *et al.*, 1996; Kennedy *et al.*, 1995). (The phenotypic associations of the insulin minisatellite will be discussed in detail in Chapter 9.) However despite intensive analysis of the locus, variant repeat distribution has only been characterised by sequence analysis of a small number of alleles (Bennett and Todd, 1996a).

The bimodal size distribution, combined with strong linkage disequilibrium surrounding the minisatellite (Lucassen *et al.*, 1993), suggest that mutation processes operating at the insulin minisatellite may be fundamentally different from those operating at hypermutable loci. In addition, the medical interest in the locus made it a prime candidate for analysis.

**Figure 8.1**



### **Location of the insulin minisatellite**

The insulin minisatellite is located on chromosome 11p15.5, 596 bp upstream of the insulin gene (*INS*) translation start site. *INS* is flanked 5' by the tyrosine hydroxylase (*TH*) gene which contains the *TH* microsatellite (*HUMTH01*) within intron 1 ~9 kb upstream of *INS*, and 3' by the gene for insulin-like growth factor II (*IGF2*). An *HphI* restriction site polymorphism is located 23 bp upstream of the insulin gene translation start site (-23*HphI*) within intron 1 and is in tight linkage disequilibrium with the two classes of minisatellite allele as defined by size. Other features of the *INS* region will be discussed in Chapters 9 and 10.

The work presented in this thesis on the insulin minisatellite had two aims. The first was to characterise patterns of allele diversity at the minisatellite by MVR-PCR, and to investigate mechanisms of *de novo* mutation. The second aim, to be discussed in Chapter 10, was to investigate the association of variant repeat distribution at the minisatellite with susceptibility to type 1 diabetes. This work involved the analysis by MVR-PCR of 876 alleles from the parents of 219 families of type 1 diabetes affected sib pairs (DNA samples were supplied by Prof. John Todd). Patterns of allele diversity between alleles are described in this chapter. This choice of population resulted in pre-selection for alleles which predispose to type 1 diabetes and so of the alleles typed 704 (80%) were class I, with 164 (19%) class III. It is unlikely that this pre-selection resulted in a qualitative difference between alleles typed in these families compared to the general population, as all parental alleles whether transmitted or not transmitted to affected offspring were analysed, and the relative risk to sibs of type 1 diabetic patients conferred by the insulin minisatellite is low ( $\lambda_s=1.25$ ) (Todd and Farrall, 1996).

This chapter is divided into three sections which describe i) the establishment of MVR-PCR at the insulin minisatellite; ii) patterns of allele diversity, and; iii) the detection and characterisation of *de novo* mutation events from both blood and sperm DNA. The detection and isolation of all *de novo* mutants was performed by A. Jeffreys; my contribution was the analysis of all alleles and mutants by MVR-PCR and the interpretation of their structures.

## Results

### ***PCR amplification of the Insulin minisatellite***

Extensive sequence data surrounding the insulin minisatellite (GENBANK accession No. L15440) were used to design all flanking primers for amplification of the locus. A 3 kb region surrounding the minisatellite taken directly from this sequence is presented in Appendix 6 with primer sites indicated.

A limited amount of DNA was available from each family of type 1 diabetic affected sib pair families. Samples were therefore immortalised by pre-amplification with primers

INS-5 and INS-3 (Table 2.1, Appendix 6). A base substitution polymorphism which generates an *HphI* restriction site 23 bp upstream of the insulin gene translation start site (-23*HphI*; Figure 8.1) (Ullrich *et al.*, 1980) is in almost complete linkage disequilibrium with the two major minisatellite classes as defined by size. To determine allele size within each family, the minisatellite was amplified from immortalised DNA stocks with primers INS-1296 and allele specific primers INS-23+ or INS-23- (Figure 8.2). With the exception of a single class I allele, INS-23+ amplified all class I alleles with INS-23- amplifying all class III alleles.

### ***Establishment of MVR-PCR at the insulin minisatellite***

Eight alleles (five class I alleles, one class II allele and two class III alleles) for which sequence data were available were selected from the literature (Awata *et al.*, 1997; Bell *et al.*, 1982; Rotwein *et al.*, 1986) to determine which variant repeats would be appropriate for detection by MVR-PCR. There is substantial heterogeneity in the literature between definitions of variant repeat sequences (Table 8.1), so repeats were renamed to be consistent with Table 2 of Bell *et al.* (1982) prior to the determination of their relative frequencies. The six most frequent variants (a-f) were selected as was the h-type repeat which, by *in vitro* transcription studies, had been implicated as a putative enhancer of insulin gene transcription (Kennedy *et al.*, 1995). Previous definitions of repeat unit sequence had located the most polymorphic region of the repeat centrally. In order to maximise specificity of MVR-PCR primers, the repeat unit was redefined as described in Table 8.2. Redefined repeats are distinguished from previous definitions by the use of capital letters. Six MVR primers were initially designed to distinguish the seven variant repeats with D- and F-repeats both detected with primer INS-MD. Upon completion of the allele diversity study described in this chapter, 36 alleles were reanalysed with primer INS-MF which distinguished between D and F repeats. All 166 variant repeats detected in these alleles by INS-MD had the sequence of F-type repeats.

Prior to MVR-PCR, alleles from heterozygotes were separated either by allele-specific PCR in -23*HphI*<sup>+/+</sup> heterozygotes, or by amplification to levels detectable on ethidium bromide-stained gels and band excision for -23*HphI*<sup>+/+</sup> homozygotes. In 23*HphI*<sup>+/+</sup> heterozygotes (therefore class I/III heterozygotes), class I and class III alleles were amplified as described

## Figure 8.2

### ***Allele-specific PCR amplification of the insulin minisatellite***

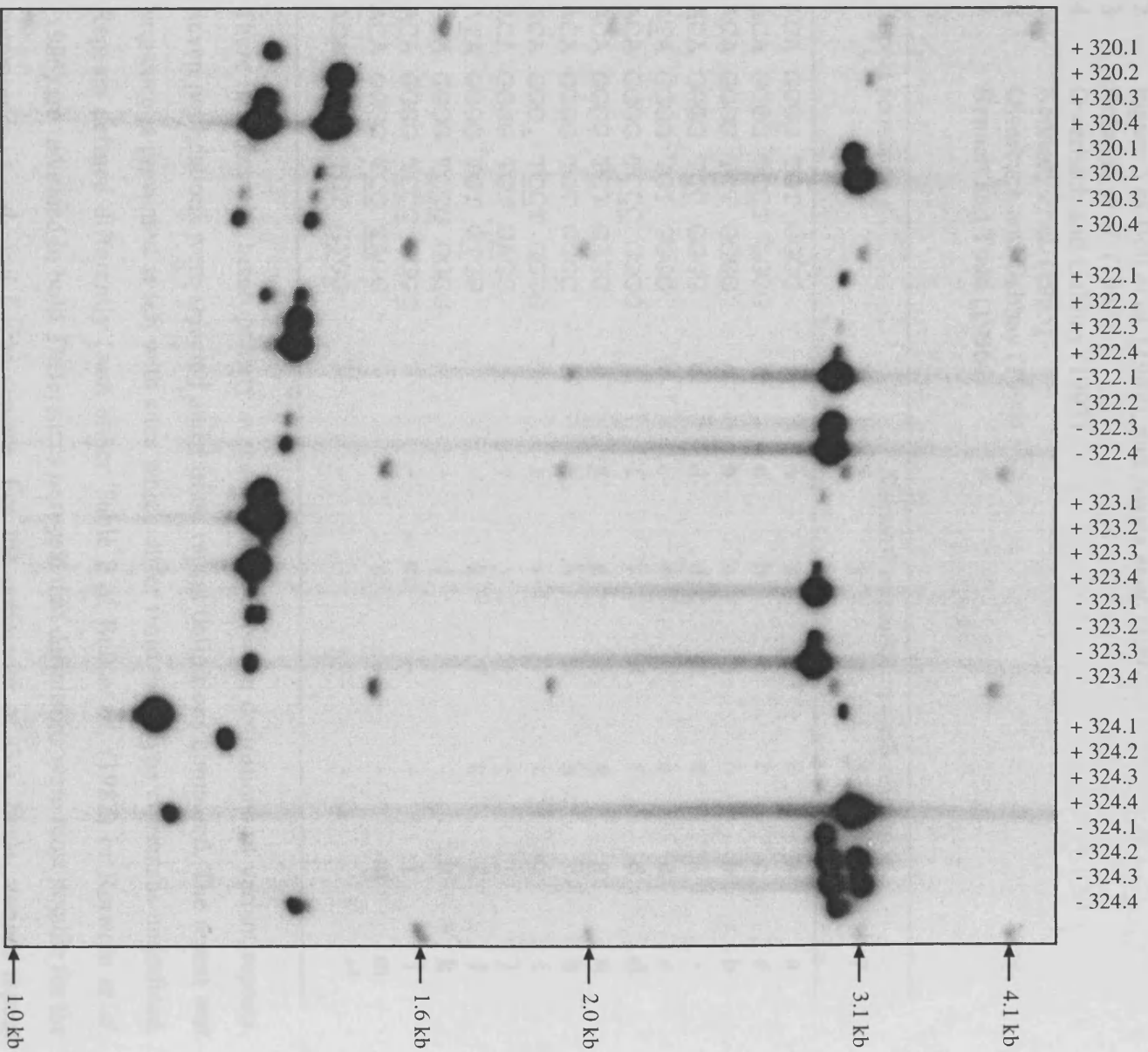
Alleles of the insulin minisatellite are shown amplified by allele-specific PCR in each of four families denoted 320, 322, 323, and 324. Numbers 1-4 represent family members with 1 being fathers, 2 mothers, and 3-4 the children. +/- indicates which allele-specific primer was used (INS-23+ amplifies class I alleles, INS-23- amplifies class III alleles). Both offspring in each family were affected by type 1 diabetes. DNA was supplied courtesy of Prof. John Todd.

DNA samples were preamplified due to limited DNA resources under standard PCR conditions from 10 ng genomic DNA with primers INS-5 and INS-3 at 96°C 40 sec, 60°C 30 sec, 70°C 3 min for 10 cycles and PCR products were diluted x128 in dilution buffer (5 mM Tris-HCl pH 7.5, 5 µg/ml carrier herring sperm DNA). Cycle number was limited to 10 as in class I/III heterozygotes there would be preferential amplification of the small class I alleles. Samples were reamplified with primers INS-1296 and allele specific primers INS-23+ or INS-23-. Class I alleles were amplified at 96°C 40 sec, 61°C 30 sec, 70°C 3 min for 20 cycles. Class III alleles were amplified under the same conditions for 22 cycles. PCR products were detected by Southern blot hybridisation as shown. To amplify alleles to levels visible on ethidium bromide-stained gels, both class I and class III alleles were amplified as above for 32 cycles.

Table 8.1

*Definitions of variant repeat sequences*

Figure 8.2



**Table 8.1**

**Definitions of variant repeat sequences**

**Reference**

- 1 Table 2 of Bell *et al.* (1982)
- 2 Figure 3 of Bell *et al.* (1982), and Awata *et al.* (1997)
- 3 Rotwein *et al.* (1986)
- 4 Owerbach and Gabbay (1993)
- 5 Kennedy *et al.* (1995)
- 6 Owerbach and Gabbay (1996)
- 7 Bennett and Todd (1996a)

Repeat sequence	Name of sequence in each reference						
	1	2	3	4	5	6	7
ACA GGGG TGT GGGG	a	a	a	a	a	a	a
ACA GGGG <u>TCCT</u> GGGG	b	<b>c</b>	b	<b>c</b>	<b>c</b>	<b>c</b>	<b>c</b>
ACA GGGG <u>TCT</u> GGGG	c	<b>b</b>	c	<b>b</b>	<b>b</b>	<b>b</b>	<b>b</b>
ACA GGGG <u>TCC</u> GGGG	d	d	d	d	d	-	-
<u>ATA</u> GGGG TGT GGGG	e	e	e	e	e	e	e
ACA GGGG <u>TCCC</u> GGGG	f	f	f	<b>d</b>	f	<b>d</b>	<b>d</b>
ACA GGGG <u>TCT</u> <u>GAGG</u>	g	g	g	g	g	g	<b>q</b>
ACA GGGG TGT <u>GGGC</u>	h	h	h	h	h	h	h
ACA GGGG <u>TCCT</u> GGGG	i	i	i	i	i	i	i
ACA GGGG TGT <u>GAGG</u>	j	j	j	j	j	j	j
<u>ATA</u> GGGG TGT <u>GTGG</u>	-	-	k	<b>f</b>	k	<b>f</b>	<b>f</b>
ACA GGGG <u>TCCG</u> GGGG	-	-	l	-	-	<b>k</b>	<b>k</b>
ACA GGGG <u>TCCC</u> <u>GGGT</u>	-	-	m	-	-	<b>l</b>	<b>l</b>
ACA GGGG <u>TCT</u> <u>TAGG</u>	-	-	n	-	-	<b>m</b>	<b>m</b>
ACA GGGG <u>TCT</u> <u>GTGG</u>	-	-	-	-	-	-	* <sup>a</sup>

There is substantial heterogeneity in the literature between definitions of variant repeats. Seven publications were selected and variant repeat definitions compared. The repeat unit sequence is presented at left with sites which differ from the a-type consensus underlined. Repeats defined differently from either Table 2 of Bell *et al.* (1982) or Rotwein *et al.* (1986) are indicated in bold. Differences in repeat unit definitions were most notable for the common b-, c-, d- and f-type repeats. For the remainder of this thesis, variant repeat definitions will be based on Table 2 of Bell *et al.* (1982).



**Table 8.2*****Re-definition of variant repeat sequences at the insulin minisatellite***

Previous repeat definition	Redefined repeats	MVR primers
a ACAGGGGTGT·GGGG	A ·GTGGGGACAGGGGT	INS-MA
b ACAGGGGT <u>CCT</u> TGGGG	B <u>CCT</u> TGGGGACAGGGGT	INS-MB and INS-MC
c ACAGGGGT <u>CT</u> ·GGGG	C · <u>CT</u> TGGGGACAGGGGT	INS-MC
d ACAGGGGT <u>CC</u> ·GGGG	D · <u>CC</u> GGGGACAGGGGT	INS-MD
e A <u>T</u> AGGGGTGT·GGGG	E ·GTGGGGAT <u>A</u> GGGGT	INS-ME
f ACAGGGGT <u>CCC</u> GGGG	F <u>CCC</u> GGGGACAGGGGT	INS-MD and INS-MF
h ACAGGGGTGT·GGG <u>C</u>	H ·GTGGG <u>C</u> ACAGGGGT	INS-MH

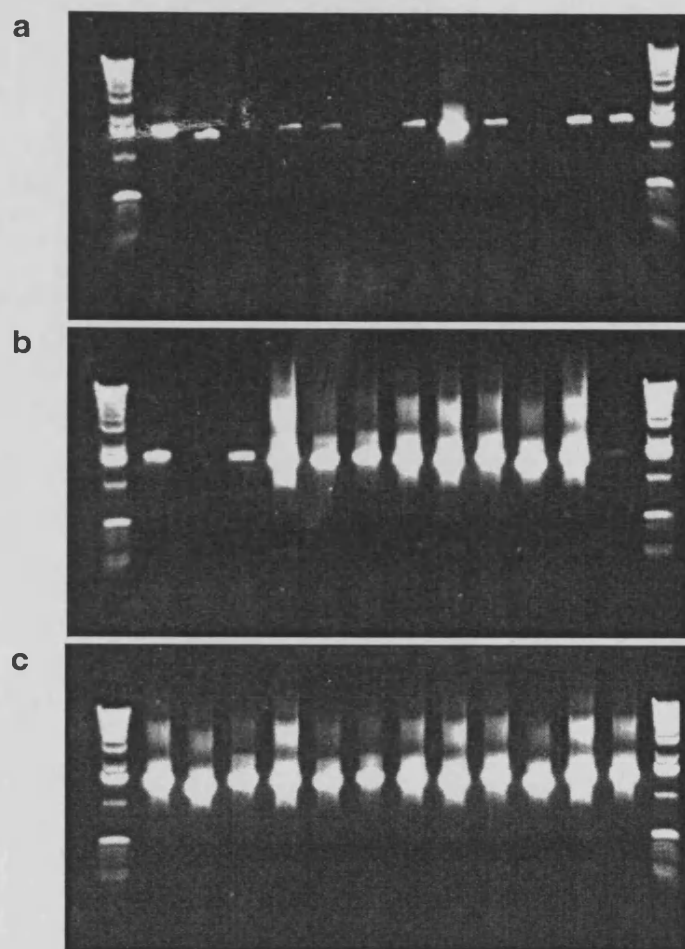
Repeat register was redefined to maximise MVR-PCR primer specificity. The identity of variant repeats was based on Table 2 of Bell *et al.* (1982), and named with capital letters to distinguish from previous definitions. Positions of repeat sequence divergence from the A-type repeat consensus are underlined. MVR primers which detect each variant are listed.

in Figure 8.2 for 20 and 22 cycles respectively and diluted 500-fold and 100-fold respectively in dilution buffer (5 mM Tris-HCl pH 7.5, 5 µg/ml carrier herring sperm DNA). Class I alleles separated by gel electrophoresis were gel-purified by freeze-thaw and diluted 1000-fold in dilution buffer. Class III alleles separated by gel electrophoresis were gel purified using the Qiaex II gel purification kit (Qiagen) and diluted 100-fold in dilution buffer. The necessity for dilutions in buffer containing carrier DNA is described in Figure 8.3. Examples of class I alleles analysed by six-state MVR-PCR are presented in Figure 8.4.

While class I alleles could be readily analysed using this technique, difficulties were experienced when analysing class III alleles. Alleles of up to 213 repeats were identified with  $\geq 50\%$  of repeats in each allele being A-type repeats. Full MVR-PCR analysis of alleles of this size would therefore require simultaneous amplification of  $>100$  different amplicons in a single reaction, followed by their resolution by electrophoresis (requiring the resolution of two amplicons of  $\sim 3$  kb which differed by just 14 bp). The MVR system described above uses a universal primer which anneals to flanking DNA 5' of the minisatellite. As described in Figure 8.5, amplification of the repeat array using a 3' universal primer and a reverse MVR primer created a series of deletion amplicons containing only the 3' repeats from the minisatellite allele. This reduced both the number of repeat units within the amplicon, and the size of the amplicon compared to full length alleles. Forward MVR mapping of deletion fragments therefore allowed large alleles to be accurately typed along their entire length, with full allele codes assembled from overlapping codes generated from the whole allele and each deletion amplicon.

The system of reverse MVR-PCR was designed using a single MVR primer which detected E-type repeats (primer INS-MER; **INSulin MVR** primer detecting **E**-type repeats in **Reverse**). INS-MER is a composite primer with the 3' sequence specific to E-type repeats and the 5' sequence identical to primer INS-1296. PCR generated a population of amplicons starting with the INS-1296 primer sequence and extending from each E-type repeat to the 3' flanking site (Figure 8.5b). Amplicons were separated by gel electrophoresis (Figure 8.5c) and DNA from 1-2 amplicons (depending on allele size) was gel purified by freeze-thaw and diluted 100-fold in dilution buffer. The INS-1296 5'

**Figure 8.3**



***All Dilutions are performed in the presence of carrier DNA***

The greatest specificity of variant repeat discrimination for the system of MVR-PCR established at the insulin minisatellite was obtained when diluting template DNA to very low concentrations (c. 0.1 pg/ $\mu$ l of pre-amplified DNA) and elevating PCR cycle number accordingly (data not shown). DNA from bands excised from agarose gels was initially diluted up to 5000-fold in water. If amplifications of diluted samples were done immediately, consistent results were obtained (data not shown). However, after storage of diluted samples for ~1 month, amplifications were sporadic (Figure 8.3a; all amplifications were of DNA from class I alleles to levels visible on ethidium bromide-stained gels). It was suggested that at such low dilutions, DNA is adsorbed onto the sides of the polypropylene storage tubes, so all samples should be diluted in a buffer containing high molecular weight carrier DNA (A. Jeffreys, pers. commun.). To test this hypothesis, the same samples were diluted from stock solutions by a factor of x2000 either in water (Figure 8.3b) or in dilution buffer (5 mM Tris-HCl pH 7.5, 5  $\mu$ g/ml carrier herring sperm DNA) (Figure 8.3c) and dilutions incubated at room temperature for 30 minutes. Subsequent PCR amplification demonstrated that amplification of stocks diluted in water were sporadic, whilst amplification of samples diluted in dilution buffer were consistent, indicating that carrier DNA was necessary in all dilutions.

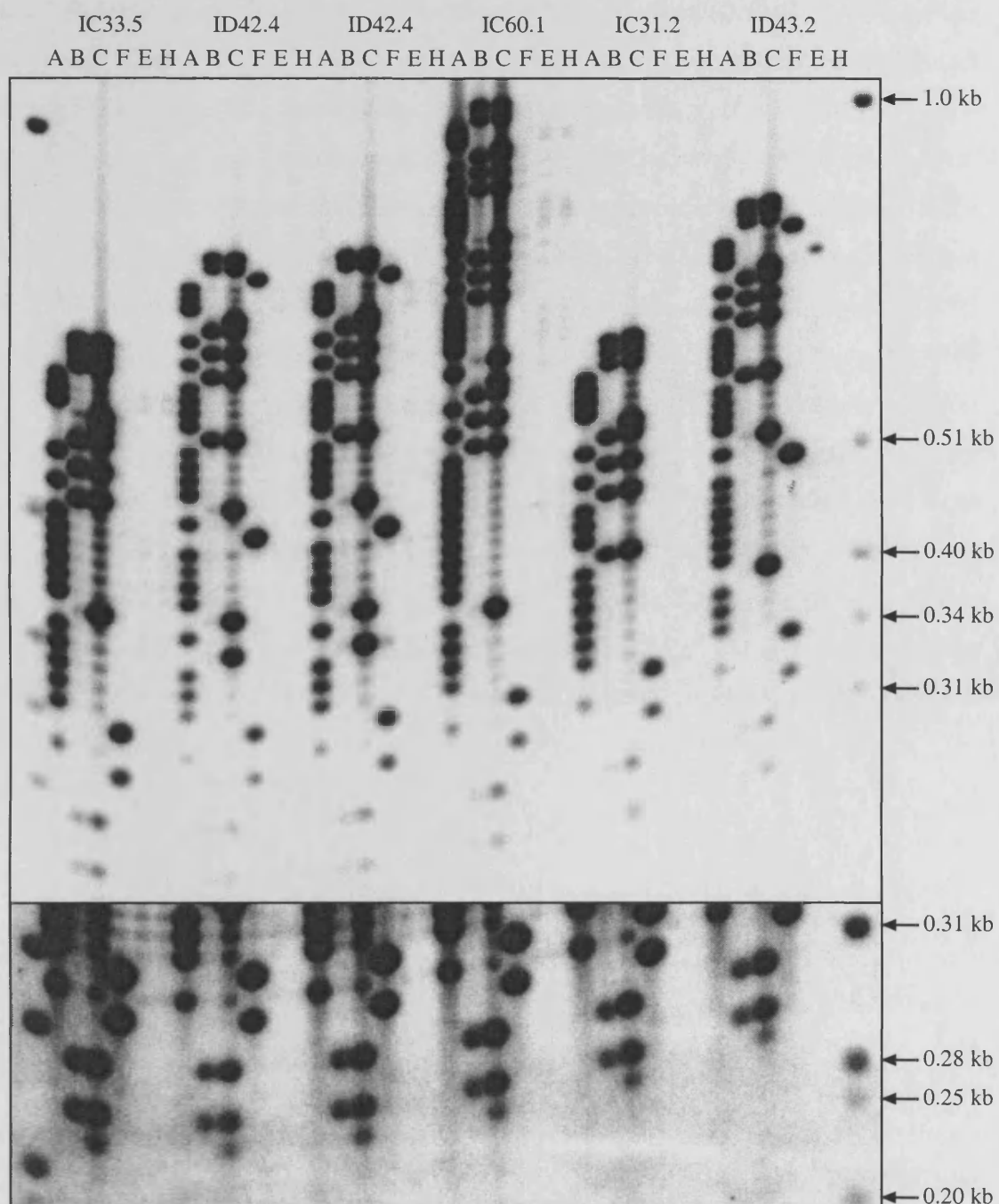
## Figure 8.4

### ***MVR-PCR of class I alleles***

All class I alleles were MVR mapped with primers INS-MA, INS-MB, INS-MC, INS-MD, INS-ME, and INS-MH. INS-MC detects both B- and C-type repeats (Table 8.2). INS-MB discriminates between these two variants. Similarly, INS-MD should detect both D- and F-type repeats. However, further analysis of a selection of alleles with primer INS-MF which detects only F-type repeats indicated that all repeats detected by INS-MD were F-type repeats. No E- or H-type repeats were detected in class I alleles. The first few repeats were often not detectable after short (14 hr at room temperature) exposure of the autoradiograph. Autoradiographs of each radiolabelled Southern blot were therefore also produced after 36 hr exposure at -80°C with intensifier screen (longer exposure is presented at the bottom). A subdivision of class I alleles is apparent from the autoradiograph presented, most readily defined by the presence or absence of F-type repeats located at the centre and top of alleles. This division will be considered further in Figure 8.7.

MVR-PCR reactions were performed under standard conditions with primers INS-1296, TAG, and one MVR primer, and PCRs cycled at 96°C 40 sec, 58°C 30 sec, 70°C 2 min for 8 cycles followed by 96°C 40 sec, 65°C 30 sec, 70°C 2 min for 12 cycles. MVR primer concentrations were INS-MA, 8 nM; INS-MB, 10 nM; INS-MC, 25 nM; INS-MD, 15 nM; INS-ME, 50 nM; INS-MF, 10 nM; INS-MH, 3 nM. Samples were electrophoresed through a 40 cm 1.5% LE agarose gel at 3 V/cm for 18 hr and detected by Southern blot hybridisation. The MVR system was very sensitive to input DNA concentration, so a test MVR of each allele using only INS-MA was initially performed and input DNA subsequently adjusted according to signal strength and quality.

**Figure 8.4**

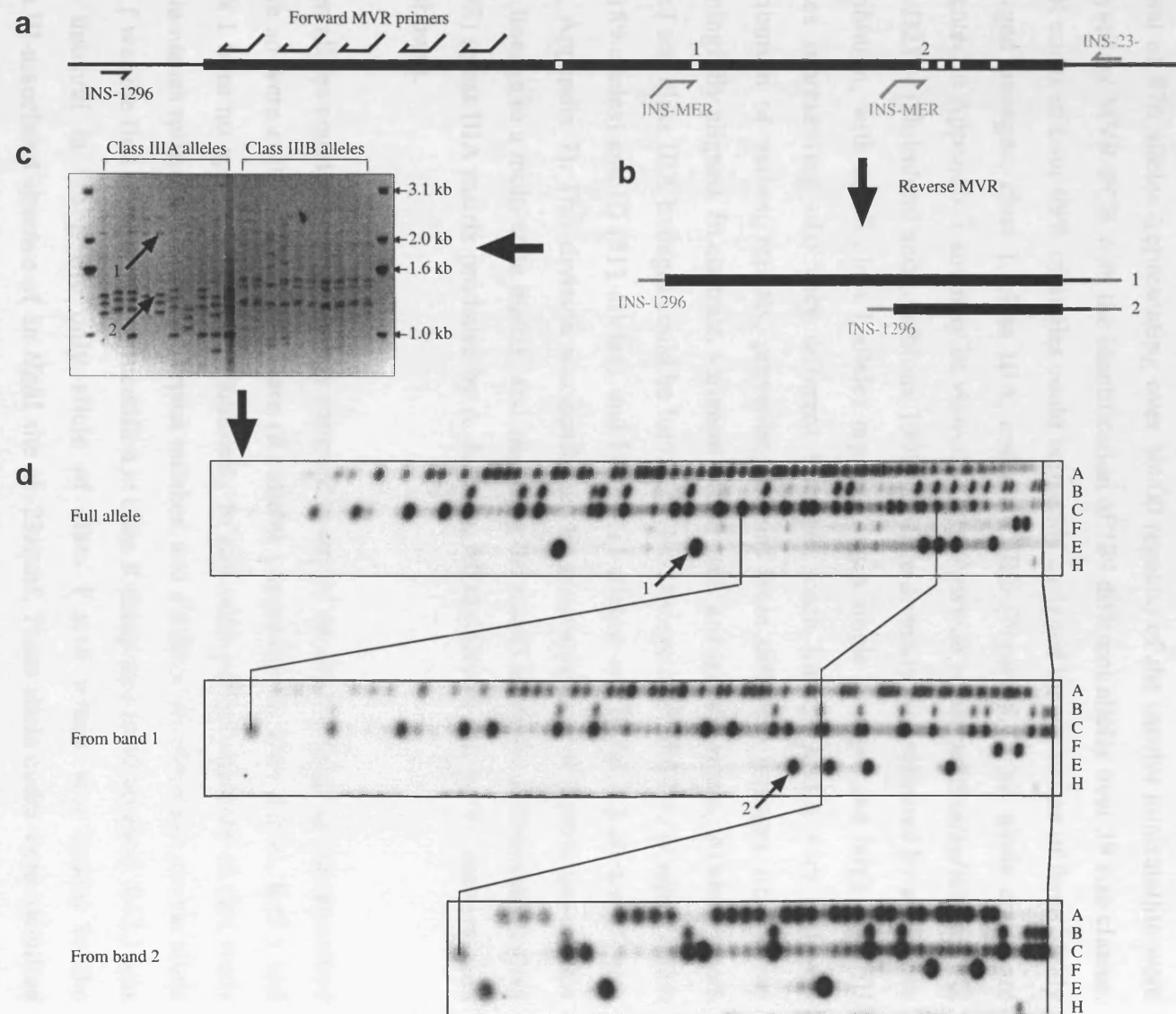


## Figure 8.5

### ***Analysis of large class III alleles by "reverse-forward" MVR-PCR***

Class III alleles were too large to be accurately mapped by MVR-PCR along their full length. A system of reverse MVR-PCR was therefore established using a single MVR primer INS-MER, which detects E-type repeats (indicated as white squares), and the class III allele-specific primer INS-23- (shown in grey; Figure 8.5a). INS-MER is a composite primer in which the 5' sequence is identical to flanking primer INS-1296. PCR generates a series of amplicons between the INS-23- flanking site and each E-type repeat, each of which carries a 5' extension with INS-1296 sequence identity. Reverse MVRs were performed on separated class III alleles in 10 µl reactions with 0.4 µM primers INS-23- and INS-MER at 96°C 40 sec, 60°C 30 sec, 70°C 2 min for 1 cycle followed by 96°C 40 sec, 65°C 30 sec, 70°C 2 min for 26 cycles. Two of these bands (1 and 2) selected for MVR mapping are presented (Figure 8.5b). The deletion amplicons are separated by agarose gel electrophoresis (Figure 8.5c), and gel purified by freeze-thaw extraction. Based purely on the reverse MVR with the single primer, two structurally distinct classes of class III alleles can be identified (Figure 8.5c) which were called IIIA and IIIB (these lineages are described in detail in Figure 8.7). MVR mapping of the entire allele and each of the two deletion amplicons (Figure 8.5d) allowed full MVR codes of each class III alleles to be unequivocally determined.

**Figure 8.5**



extension on the deletion amplicon allowed them to be analysed by forward MVR-PCR as described in Figure 8.4. Two examples of class III alleles mapped by MVR-PCR are presented in Figure 8.6.

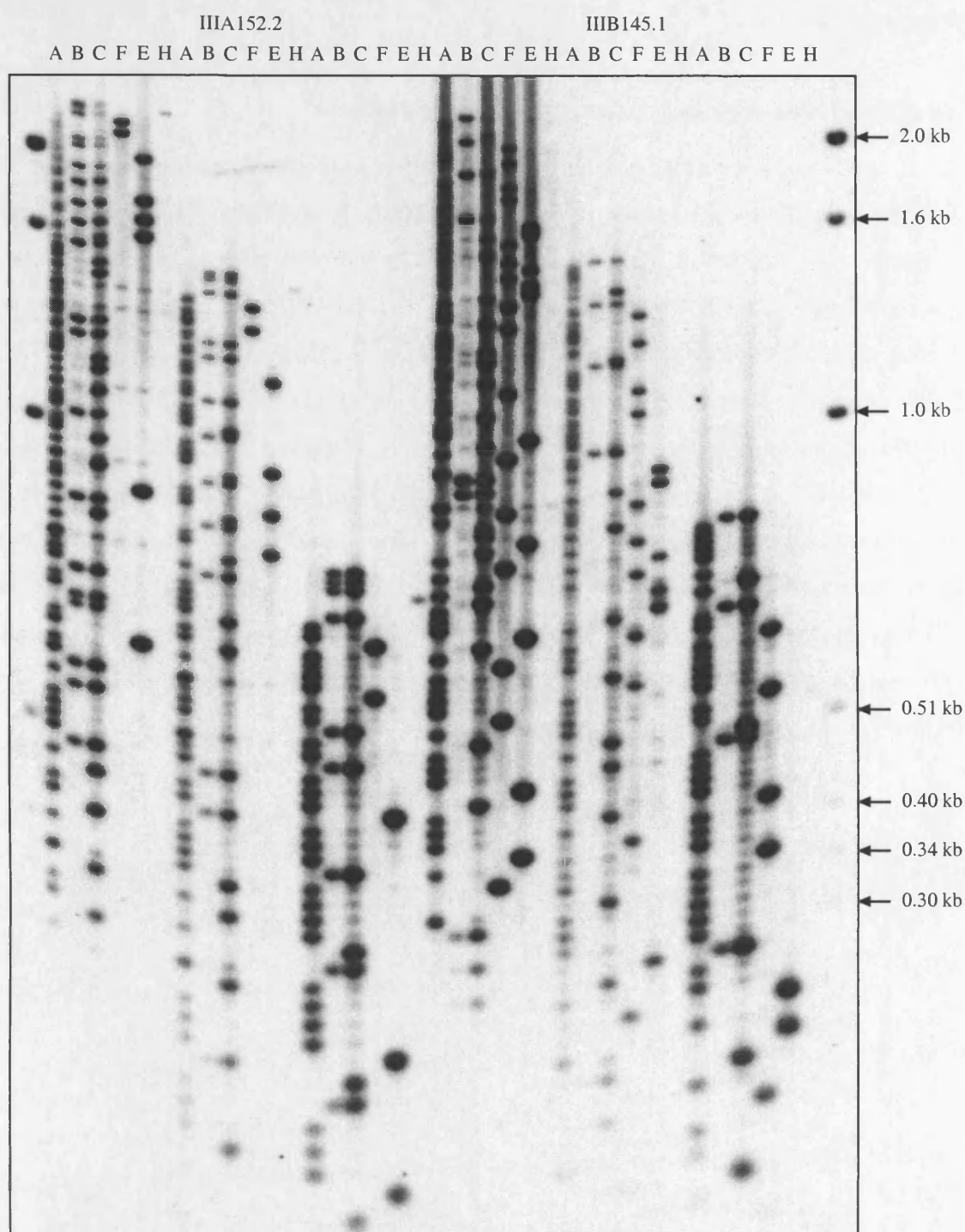
### ***Allele diversity at the insulin minisatellite***

A total of 876 alleles (representing over 50000 repeats) of the insulin minisatellite were analysed by MVR-PCR with the identification of 189 different alleles over 39 size classes. MVR codes of over 99% of alleles could be readily assigned by eye to one of three highly diverged lineages; class I, class IIIA, and class IIIB (Figure 8.7a, all allele codes are presented in Appendix 7 and may be viewed at <http://www.le.ac.uk/genetics/ajj/insulin>). As with *D2S44* (Holmlund and Lindblom, 1998), lineage diversity was reflected by allele size distribution, with small class I alleles representing a single lineage, and large class III alleles representing two very different lineages. Each lineage had a very different distribution of variant repeats, preventing alleles from different lineages from being meaningfully aligned. In contrast, variation within each lineage was minor. Alleles of both class I and class IIIA lineages could be further divided by eye each into two groups, classes IC (189 alleles) and ID (511 alleles), and IIIAi (111 alleles) and IIIAii (15 alleles) (Figure 8.7a, Appendix 7). This division was confirmed by the conversion of variant sites within each lineage to a multi-state matrix, and analysing the matrix by multi-dimensional scaling (MDS) (class IIIA matrix produced by A. Jeffreys, MDS analysis was by Y. Dubrova, data not shown).

Seven alleles could not be confidently assigned to any of the allele lineage groups described above so were collectively called **Rare (R)** alleles (Appendix 7). Two alleles, R42.1 and R188.1 bore no apparent structural similarity to any other alleles analysed in this study (allele names reflect allele lineage, repeat number, and a further discriminator so that allele R42.1 was the first allele structure identified in class R composed of 42 repeats). R42.1 was also unusual in being the only allele of class I size which was linked to the class III-associated absence of an *Hph*I site at -23*Hph*I. Three allele codes were identified (IC60.1, IIIB110.1 and IIIB128.1) of class II allele sizes. These alleles could all be readily aligned with one of the major defined lineages indicating that, at least in Caucasians, class II alleles do not represent a distinct lineage.



**Figure 8.6**



***MVR-PCR of class III alleles***

Two class III alleles, IIIA152.2 and IIIB145.1, were MVR mapped from the full allele and two deletion amplicons generated as described in Figure 8.5. MVR-PCR conditions were as described in Figure 8.4. Allele names reflect allele lineage, repeat number, and a further discriminator so that IIIA152.2 is the second allele in lineage IIIA identified of 152 repeats in length (allele lineages are described in detail in Figure 8.7). Substantial structural differences between the two alleles are readily apparent. The H-type repeat is present as a single copy in allele IIIA152.2, and is absent from IIIB145.1.

## Figure 8.7

### ***Structural diversity in insulin minisatellite alleles***

MVR codes are presented in a 5' to 3' orientation with the insulin gene to the right. 'o' denotes null repeats (unamplifiable repeats due to additional unknown variants). Hyphens were inserted to improve alignments. Alleles were initially divided by eye into three lineages, class I, IIIA and IIIB (Figure 8.7a). Visual inspection and multi-dimensional scaling further divided class I alleles into classes IC and ID, and class IIIA into IIIAi and IIIAii (IIIAii indicated by \*). All allele codes are presented in Appendix 7, and at <http://www.le.ac.uk/genetics/ajj/insulin>. Three rare alleles which may have arisen by complex mutation processes are also shown aligned to potential progenitor alleles detected in the diversity survey (Figure 8.7b). The complex duplication product IC60.1 is split across several lines to facilitate alignment of the mutant with the putative progenitor. Regions showing apparent inter-allelic transfers are underlined in the allele and the potential donor allele. A-type repeats are shown in green, B in red, C in blue, E in cyan, F in yellow, H in pink, and null repeats in black.

Figure 8.7

a

### Class IC

```
IC28.1 CBoBo A AAAAAA-----BABABC--AAABBB
IC31.1 CBoBo A AAAAAA-----BABABC--AAABBB
IC31.2 CBoBo A AAAAAA-----BAAA-BABABC-AAAABBB
IC33.1 CBoBo A AAAAAA-----BAAA-BABABC-AAAABBB
IC34.1 CBoBo A AAAAAA-----BAAA-BABABC-AAAABBB
IC30.6 CBoBo A AAAA-----CAAAAA-----BABABC--AAABBB
IC32.6 CBoBo A AAAA-----CAAAAA-----BABABC--AAABBB
IC32.7 CBoBo A AAAA-----CAAAAA-----BABABC--AAAB-B
IC35.4 CBoBo A AAAA-----CAAAAA--BAAA-BABABC--AAABBB
IC37.1 CBoBo A AAAA-----CAAAAA--BAAA-BABABC--AAABBB
IC38.3 CBoBo A AAAACAAA-CAAAA--BAAA-BABABC--AAABBB
IC41.2 CBoBo A AAAACAAAACAAAA--BAAA-BABABC-AAAAB-B
```

### Class ID

```
ID37.1 CBoBo A AAAC--AAA--ACAAAABAAA--BABCAA BBB
ID38.2 CBoBo A AAACACAAAA-ACAAAABAAA--BABCAA-B-B
ID39.1 CBoBo A AAACACAAAA-ACAAAABAAA--BABCAA-BBB
ID39.4 CBoBo A AAACACAAAA-ACAAA-BAAA--BABCAA BBB
ID39.6 CBoBo A AAACACAAAA-ACAAAABAAA--BABCAA B-B
ID40.2 CBoBo A AAACACAAAA-ACAAAABAAA--BABCAA BBB
ID41.5 CBoBo A AAACACAAAA-ACAAAABAAA--BABCAA BBB
ID41.1 CBoBo A AAAC--AAAA- A AAAABAAAABABACAAA BBB
ID42.2 CBoBo A AAAC-CAAAA-ACAAAABAAAABABACAAA BBB
ID42.4 CBoBo A AAACACAAAA-ACAAAABAAAABABACAAA B-B
ID43.9 CBoBo A AAACACAAAA-ACAAAABAAAABABACAAA B-B
ID44.1 CBoBo A AAACACAAAA-ACAAAABAAAABABACAAA BBB
```

### Class IIIA

```
IIIA138.1 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA143.2 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA146.2 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA147.2 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA149.4 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA150.3 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA150.5 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA152.2 CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA158.3* CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA159.2* CBoBACAACoA--CAACABAAAABAAAABBAABAAAA-oCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
```

### Class IIIB

```
IIIB110.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-----CAAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB128.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA ACAA-----CAAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB141.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB142.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB143.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB143.4 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB144.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB144.3 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB145.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
IIIB145.3 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA A-CAAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
```

b

### Complex duplication

```
IC30.6 CBoBo A AAAAAAABABCAAAABBB
IC60.1 CBoBo A AAAAAAABAB
ACAAAABABAB
ACAAAABABABCAABBB
```

### Inter-allelic in register conversion

```
IIIA .....AAAABCAAAAABAAAABAAA AAAA ABoHBBB
IC38.3 CBoBo A AAAAAACAAACAAABAAAA-----BABABCAABBB
R44.1 CBoBo A AAAAAACAAACAAABAAAAAAA AABABCAABBB
```

### Inter-allelic conversion

```
IIIB144.1 CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA ACAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
R213.1 5' CBoBoBo AAAAAACAAACA AAAA ACIAAACACA ACICCA AoBoBo ACIAAACAAA ACAAACACAAAA CAAA-ACICCA AFA ACA ACIAAABAAAA- AAA AAAABCAA AAAA ABoACoAAAAAB
R213.1 3' oBACAACoACAAABAAAABAAAABBAABAAAAoCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
IIIA149.4 CBoBACAACoACAAABAAAABAAAABBAABAAAAoCABAACAAoACAAABAAACoACACAAAABABAAAoAAA-----oACAAACAAACAAABACAAABCAAABCAAAA-----BAA AAAA AABoHBBB
```

Levels of MVR code diversity varied substantially between different allele lineages. For example the 126 alleles analysed within the class IIIA lineage were represented by 73 different MVR codes, whereas in group ID, 511 alleles had only 36 different codes. Three allele structures in group ID accounted for over 70% of all ID alleles with ID40.2, ID42.4 and ID44.1 being present in 112, 184, and 69 copies respectively. 91% of parents were heterozygous determined by MVR-PCR analysis (compared with 89% allele length heterozygosity) suggesting a mutation rate of  $1 \times 10^{-4}$  by  $H = \frac{1}{1 + 4N_e\mu}$ . In contrast, mutation rate was estimated at  $9 \times 10^{-4}$  using Ewens' distribution (Ewens, 1972) which considers the number of different alleles detected within the population (Table 8.3) (estimations of mutation rate by Ewens distribution were by A. Jeffreys). A possible explanation for the 9-fold difference in mutation rate obtained using these two approaches is described in Chapter 10. Mutation rate estimated by Ewens' distribution was greatest for the large class IIIA alleles, although there was no consistent relationship between estimated mutation rate and allele size (Table 8.3). However, these estimations of mutation rate include assumptions which are questionable for this locus such as an infinite allele model (which is invalid if *de novo* mutation generates alleles already present in the population), mutation-drift equilibrium (low levels of allele lineage diversity suggest mutation-drift equilibrium has not been achieved in Caucasians), and the absence of selection (selection will act upon the insulin-linked region due to its disease associations).

### ***Patterns of variation in the insulin minisatellite***

Patterns of variation between alleles within a lineage provided clues about the mechanisms underlying mutation at the insulin minisatellite. Most variation within a lineage was minor, and due to small deletions and duplications of repeats most notably within uninterrupted arrays of A-type repeats. For  $A_n$  arrays of 3-5 repeats, most changes involved the gain or loss of a single A-type repeat. Arrays of over 5 A-type repeats were generally located in regions that were interrupted with variant repeats at equivalent sites in other alleles of the same lineage, suggesting that large  $A_n$  arrays may be formed by the deletion of variants causing the apposition of two shorter  $A_n$  arrays. Higher order repeated motifs were also identified such as the ABABAB motif towards the 3' end of many class I alleles. There was also evidence for imperfect duplications, such as variations of a ABACEA motif duplicated

**Table 8.3*****Allelic variability at the insulin minisatellite in Caucasians***

Allele class	Mean number of repeats (range)	Number of alleles	Number of different alleles defined by:		Estimated mutation rate x10 <sup>-4</sup>
			length	MVR code	
IC	33.6 (28-60)	189	14	59	17
ID	41.6 (37-44)	511	8	36	2
IIIA	149.1 (138-159)	126	16	73	63
IIIB	140.2 (110-145)	43	7	14	17
All	60.5 (28-213)	876	39	189	9

Germline mutation rates were estimated from Ewens' distribution (Ewens, 1972), where the expected number of different alleles defined by MVR structure,  $n_a$ , in a sample of  $n$  alleles of a given class is given by  $n_a = \sum_{i=1}^n \frac{\theta}{\theta + i - 1}$ ;  $\theta = 4N_e f \mu$ , where  $N_e$  is the effective population size for Caucasians, assumed to be 20000,  $f$  is the proportion of all alleles that belong to the class being tested, and  $\mu$  is the mutation rate. Mutation rates were estimated by A. Jeffreys.

towards the 3' end of class IIIA alleles. In addition to large duplications, there was evidence for large deletion events. Allele IIIB110.1 is identical to many other class IIIB alleles at both 3' and 5' ends, but contains the simple deletion of a region of a central ~35 repeats (Figure 8.7a). Furthermore, the first allele ever sequenced (clone  $\lambda$ H1.1) was composed of 34 repeats but closely resembles a class IIIA allele with a large internal deletion (Bell *et al.*, 1982). However, this allele was not detected in this study and whilst there is evidence that this allele exists at low frequency in the general population (Bennett *et al.*, 1995), it is possible that the deletion event arose during generation and propagation of the library from which the clone was derived (Bell *et al.*, 1981).

While most differences between aligned alleles could be explained by small and simple deletions and duplications, there was evidence for more complex turnover of repeats. Allele IIIB128.1 shows a deletion relative to other class IIIB alleles plus the insertion of 20-repeat motif at the site of deletion which was not present in any other allele typed (Figure 8.7a). Indications of complex intra-allelic duplications come from allele IC60.1 which is identical to IC30.1 at the 5' end but has undergone an apparent triplication event towards the 3' end of the array, doubling allele size (Figure 8.7b). The largest allele detected (R213.1) was identical to many class IIIB alleles at the 5' end, and displayed complete identity to allele IIIA149.4 over 147 repeats at the 3' end (Figure 8.7b). These regions of allele code identity were separated by the complex reduplication of a motif located at the 5' end of class IIIA alleles. Analysis of allele length at the tyrosine hydroxylase microsatellite (*HUMTH01*) 9 kb upstream of the insulin gene identified an allele commonly associated with class IIIA haplotypes (Bennett *et al.*, 1995) (the relationship between MVR code and flanking haplotype will be considered in detail in Chapter 10) suggesting that R213.1 was produced by a complex inter-allelic conversion-like event between a class IIIA allele and a class IIIB allele. Further evidence for inter-allelic repeat transfers comes from allele R44.1 which is similar to the class IC allele IC38.3 except for the insertion of an F-type repeat 14 repeats from the 3' end of the array (Figure 8.7b). Most class IIIA alleles have an F-type repeat at the same position suggesting that R44.1 was produced by an in-register gene conversion-like transfer between a IC allele and a IIIA allele. However, major differences in repeat-type composition do exist between allele lineages, for example E-type repeats are

present only in class III alleles, whilst H-type repeats are restricted to the class IIIA lineage, suggesting that inter-allelic repeat transfers only occur at a very low frequency.

### ***Detection of de novo mutants***

The detection and isolation of *de novo* mutants at the insulin minisatellite was performed by A. Jeffreys, so the technique will only be described in brief. Germline mutation rate at this locus was too low for mutants to be detected by small pool PCR (SP-PCR) of sperm DNA (Jeffreys *et al.*, 1994). Sperm DNA was therefore digested to completion with *HinfI* and digestion products electrophoresed to enrich for mutant molecules which differ in array length from the progenitor (Jeffreys and Neumann, 1997). *HinfI* was selected as it cleaves at sites close to the minisatellite (255 bp 5' and 144 bp 3' of the repeat array; Appendix 6). Furthermore, the more distal *HinfI* sites were located 1104 bp 5' and 825 bp 3' of the most proximal sites (Appendix 6) so DNA fractions selected to size-enrich for mutations of the class I progenitor allele would not be contaminated by partially digested DNA, unless enriching for mutants of over twice the size of the progenitor. Mutants were identified from the size fractions by SP-PCR using primers MINS-A and MINS-B and isolated by amplification to levels visible on an ethidium bromide-stained gel using the nested primers INS-1296 and MINS-C (Table 2.1, Appendix 6).

The small size of repeat unit at the insulin minisatellite prevented the separation of mutants which differed in size from class I progenitor alleles by a single repeat. Only mutants from the class I progenitor alleles which gained or lost 3 or more repeats could be quantitatively recovered. The detection of mutants derived from the larger class III alleles would therefore have only been possible for mutants which differed from the progenitor by substantially larger changes in length. Furthermore, electrophoresis effectively separates small molecules from large molecules, but not vice versa. This is due to a low frequency of molecules which migrate more slowly than expected during electrophoresis. This aberrant retardation of the migration of smaller molecules would result in the contamination by class I progenitor molecules of gel fractions which enrich for class III-derived mutants. Therefore only mutants generated from class I progenitor alleles were targeted, within a size range of 1 to 110 repeats. Sperm DNA from three non-diabetic Caucasian donors was screened for mutants. Donor 1 was a ID/IIIA heterozygote (ID40.2/IIIA152.2), donor 2 a

ID/IIIB heterozygote (ID42.4/IIIB145.1), and donor 3 a ID/ID homozygote with both alleles identical for size and MVR code (ID44.1/ID44.1) (Figure 8.7a). Mutants were also analysed in blood DNA of donor 1. Among these donors, the three most common alleles detected in the Caucasian cohort were all represented (Appendix 7). The structures of all mutants were determined by MVR-PCR. Variant repeat distributions in all molecules are presented in Appendix 8, and at <http://www.le.ac.uk/genetics/ajj/insulin>.

Mutation rates of class I alleles were determined for length changes of at least 3 repeats from surveys of  $1.3\text{--}2.6 \times 10^6$  progenitor DNA molecules from blood or sperm (Table 8.4). Deletions occurred at a similar frequency of  $4\text{--}8 \times 10^{-6}$  in both blood and sperm. In contrast, expansions were far more common in sperm ( $8\text{--}28 \times 10^{-6}$  in sperm compared with  $1 \times 10^{-6}$  in blood). Germline mutation rates were similar for all three donors, and highest for the class I homozygote. In both blood and sperm, most mutations resulted in length changes of just a few repeats from the progenitor allele (Figure 8.8) with deletions of 2 repeats detected at the highest frequency in both tissues, despite an estimated 50% loss of these deletion mutants during size fractionation compared with mutants which changed in length by  $\pm 3$  repeats. Some very large changes in allele size were also detected especially in the germline. For example, a 15 repeat mutant molecule was detected in sperm DNA of donor 1 which was generated by a simple deletion of over 90% of the 152 repeat class III progenitor allele, whilst a complex duplication of the 44 repeat progenitor allele of donor 3 more than doubled allele size to 109 repeats. Class I and class III allele sizes were therefore interchangeable in a single mutational step. Major expansions were further investigated in sperm DNA of donor 3 by screening  $15 \times 10^6$  progenitor molecules for expansions of 40–200 repeats; of the eight additional mutants detected, none had gained more than 65 repeats indicating that there is a ceiling to allele expansion.

### ***Mutation processes in the soma***

The majority of somatic mutants arose from simple intra-allelic deletions and duplications of the progenitor allele (Figure 8.9). One mutant (B1-38.2) was detected three times, suggesting somatic mosaicism for this mutant. (Nomenclature of mutants reflects the donor and tissue of origin, repeat number, and a further discriminator, so mutant B1-38.2 was the second mutant of 38 repeats detected from blood DNA of donor 1.) Only one mutant

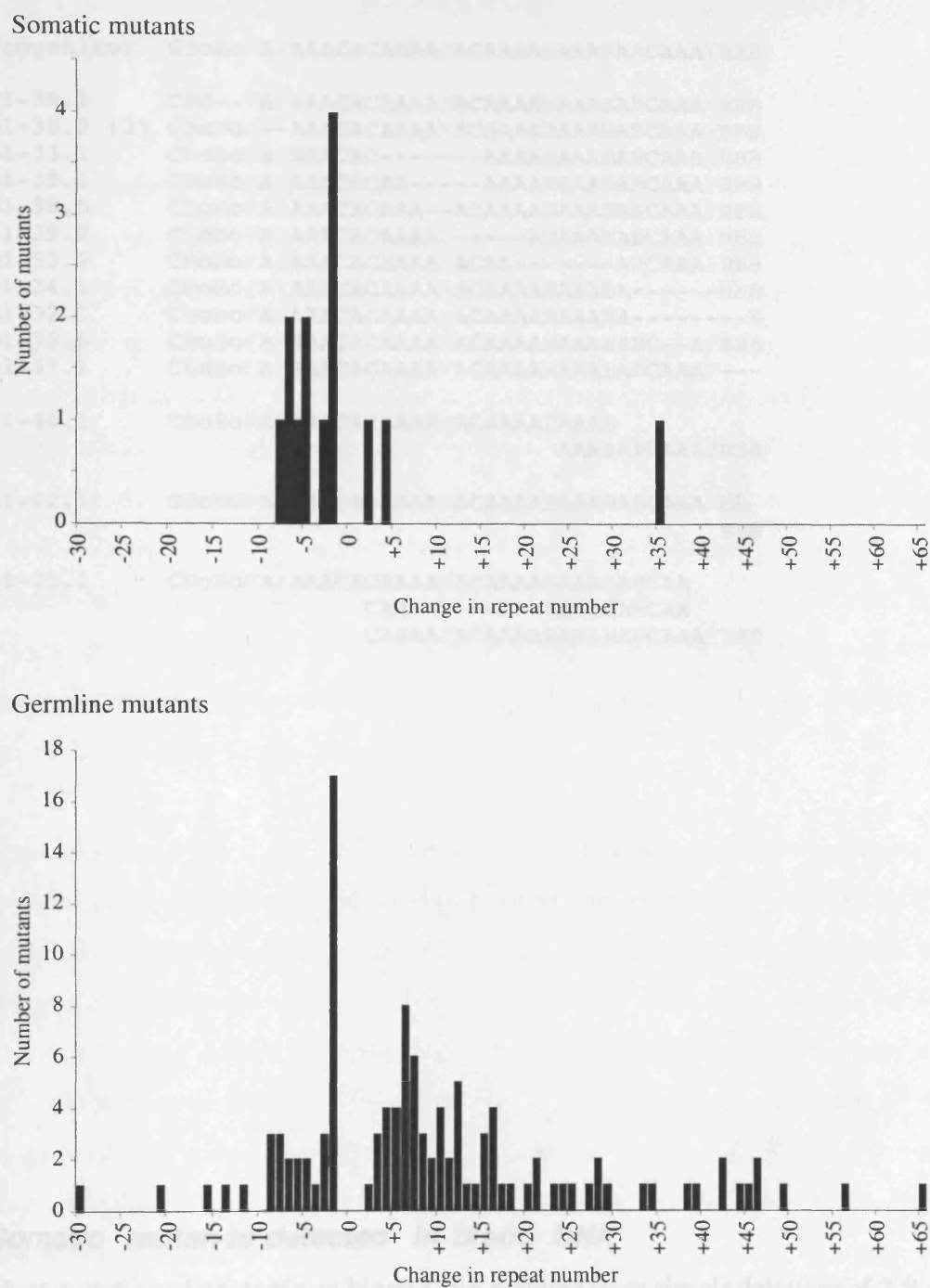


**Table 8.4*****Germline and somatic mutation rates for class I alleles of the insulin minisatellite***

Sample	Donor genotype	Progenitor molecules screened $\times 10^6$		Number of mutants Total number	Number different	Number of different mosaics	Rate $\times 10^{-6}$
Donor 1 Blood	ID/IIIA	2.0	Gains	2	2	0	1
			Losses	7	7	0	4
Donor 1 Sperm	ID/IIIA	1.3	Gains	10	10	0	8
			Losses	11	9	2	8
Donor 2 Sperm	ID/IIIB	2.6	Gains	29	28	1	11
			Losses	7	4	2	3
Donor 3 Sperm	ID>ID	1.4	Gains	34	31	2	24
			Losses	9	9	0	6

Data is presented for mutants of between 1 and 110 repeats in length showing a change of at least 3 repeats relative to the progenitor allele. Mutation rates per amplifiable progenitor molecule were calculated from the total number of mutations detected. Smaller changes in size could not be recovered quantitatively and were excluded. Data was supplied by A. Jeffreys.

**Figure 8.8**



### ***Spectrum of size changes of de novo mutations***

The difference in size between mutant molecule and progenitor allele is presented for both somatic and germline mutants. For germline mutants, data from all three donors were combined. Each group of mosaic alleles was considered as a single mutant. Mutants which differ from the progenitor by a single repeat could not be isolated. Size changes of  $\pm 2$  repeats could only be isolated at an estimated 50% efficiency compared with larger size shifts.

**Figure 8.9**

Progenitor	CBoBoFAFAAACACAAAAFACAAAABAAABABCAAAFBBB
B1-38.1	CBo--FAFAAACACAAAAFACAAAABAAABABCAAAFBBB
B1-38.2 (3)	CBoBoF--AAACACAAAAFACAAAABAAABABCAAAFBBB
B1-33.1	CBoBoFAFAAACAC-----AAAABAAABABCAAAFBBB
B1-35.1	CBoBoFAFAAACACAA-----AAAABAAABABCAAAFBBB
B1-38.5	CBoBoFAFAAACACAAA--ACAAAABAAABABCAAAFBBB
B1-35.2	CBoBoFAFAAACACAAAAF-----ABAAABABCAAAFBBB
B1-33.2	CBoBoFAFAAACACAAAAFACAA-----ABCAAFFBBB
B1-34.1	CBoBoFAFAAACACAAAAFACAAAABAAABA-----BBB
B1-32.1	CBoBoFAFAAACACAAAAFACAAAABAAABA-----B
B1-38.6	CBoBoFAFAAACACAAAAFACAAAABAAABABC--AFBBB
B1-37.1	CBoBoFAFAAACACAAAAFACAAAABAAABABCAAAF---
B1-44.1	CBoBoFAFAAACACAAAAFACAAAABAAAB AAABABCAAAFBBB
B1-42.1	CBoBoFAFAAACACAAAAFACAAAABAAABABCAAAFBB BBB
B1-75.1	CBoBoFAFAAACACAAAAFACAAAABAAABABC CAA-----BAAABABC CAAAAFACAAAABAAABABCAAAFBBB

### ***Somatic mutations detected in blood DNA***

Most mutations detected from blood DNA resulted from simple deletions of 2-8 repeats from the progenitor allele. Mutant B1-38.2 was detected as 3 identical copies as is indicated in parentheses. Mutant names reflect DNA source, repeat number, and a further discriminator so that B1-38.2 was the second mutant of 38 repeats identified from blood DNA of donor 1. The three expansion mutants are split onto more than one line to clarify the nature of the duplication, and hyphens were inserted by eye to improve allele alignments.

(B1-75.1) was more complex and involved the triplication and internal deletion of a 3' motif (Figure 8.9). There were no indications for any mutants being generated by inter-allelic mechanisms.

### ***Deletion processes in the germline***

Germline deletion mutations were predominantly simple events and displayed mosaicism, similar to deletion mutants detected in blood (Figure 8.10). Common with somatic deletions, all sperm samples also showed evidence of mosaicism indicating that at least some of these deletions had a premeiotic origin. In each sperm sample, multiple copies of a mutant allele with a two repeat deletion within a CAC motif 12-14 repeats from the beginning of the allele were identified, suggesting that this motif may act as a hotspot for deletion in the soma. In two of the three sperm samples, approximately half of these CAC deletion mutants were associated with an identical single repeat-type switch converting an AFACA motif to AFAFA close to the 3' end of the deletion, raising the possibility that both classes of mutant (2-repeat deletion with or without distal repeat switch) may have been generated by the same somatic mutation event.

A two repeat deletion at the CAC motif could result in the generation of heteroduplex DNA if the deletion arose by either unequal sister chromatid exchange followed by branch migration of the Holliday junction, or by a two repeat single strand slippage event during DNA replication. This would create a C/F repeat mismatch at the site at which half of the mutants show a repeat-type switch from C to F. If all other heteroduplex sites were subject to biased mismatch repair but the C/F mismatch was not repaired, the two types of mutant molecule detected (2 repeat deletion with or without the C->F switch) could be formed. Segregation of these structures into daughter germline stem cells followed by stem cell proliferation would therefore generate the multiple copies of the two different mutant structures detected in the sperm DNA of donors 1 and 2. It is perhaps likely that the initial 2-repeat deletion was generated by polymerase slippage during replication as opposed to unequal sister chromatid exchange as the reciprocal 2-repeat gain mutant was detected only once (S1-42.1, Appendix 8).

## Figure 8.10

### ***Deletion mutants detected in sperm DNA***

All deletion mutants detected in sperm DNA of donors 1-3 are presented as described in Figure 8.9. Mosaic mutants S1-38.2 and S1-38.6 share the same deletion but differ by a C/F repeat-type switch 7 repeats 5' of the deletion, as do mosaic mutants S2-40.2 and S2-40.5. Allele ID41.1 detected in Caucasians (Figure 8.7) shows evidence of similar C/F repeat switching.

**Figure 8.10**

### Donor 1

Progenitor	CBoBoFAFAAACACAAAAFACAAAABAAAABABCAAAFB
S1-10.1	C-----BCAAAFBBB
S1-24.1	CB-----FACAAAABAAAABABCAAAFB
S1-38.1	CBoB--AFAAACACAAAAFACAAAABAAAABABCAAAFB
S1-37.1 (2)	CBoBoF--AACACAAAAFACAAAABAAAABABCAAAFB
S1-38.2 (4)	CBoBoFAFAAAC--AAAAFACAAAABAAAABABCAAAFB
S1-38.6 (3)	CBoBoFAFAAAC--AAAAFAFAAAAABAAAABABCAAAFB
S1-38.9	CBoBoFAFAAAC--AAAAFAAAAAABAAAABABCAAAFB
S1-38.10	CBoBoFAFAAACACAAAA--CAAAAABAAAABABCAAAFB
S1-34.1	CBoBoFAFAAACACAAAAFACA-----oBABCAAAFB
S1-36.1	CBoBoFAFAAACACAAAAFACAAA-----ABABCAAAFB
S1-38.11	CBoBoFAFAAACACAAAAFACAAAA--AAABABCAAAFB
S1-33.1	CBoBoFAFAAACACAAAAFACAAAABA-----AAFBB
S1-34.2	CBoBoFAFAAACACAAAAFACAAAABAA-----AAFBB
S1-35.1	CBoBoFAFAAACACAAAAFACAAAABAAA-----AAFBB
S1-32.1 (2)	CBoBoFAFAAACACAAAAFACAAAABAAAAB-----B

### Donor 2

Progenitor	CBoBoFAFAAACACAAAAFACAAAABAAAABABABCAAAFB
S2-40.1	CBoBoF--AAACACAAAAFACAAAABAAAABABABCAAAFB
S2-39.2	CBoBoFAFAA---CAAAAFACAAAABAAAABABABCAAAFB
S2-40.2 (2)	CBoBoFAFAAAC--AAAAFACAAAABAAAABABABCAAAFB
S2-40.4	CBoBoFAFAAACACAA--FACAAAABAAAABABABCAAAFB
S2-40.5 (2)	CBoBoFAFAAAC--AAAAFAFAAAAABAAAABABABCAAAFB
S2-40.7	CBoBoFAFAAACAC--AAFAFAAAAABAAAABABABCAAAFB
S2-40.8	CBoBoFAFAAACACAAAA--CAAAAABAAAABABABCAAAFB
S2-35.1 (2)	CBoBoFAFAAACACAAAA-----BAAAABABABCAAAFB
S2-37.1 (3)	CBoBoFAFAAACACAAAAFACAAAA-----BABABCAAAFB
S2-33.1	CBoBoFAFAAACACAAAAFACAAAA-----BCAAAFBB
S2-40.9	CBoBoFAFAAACACAAAAFACAAAABAA--BABABCAAAFB
S2-40.10 (2)	CBoBoFAFAAACACAAAAFACAAAABAAA--BABCAAAFB

### Donor 3

Progenitor	CBoBoFAFAAACACAAAAFACAAAABAAAABABABCAAAFB
S3-36.1	CBoBoFAFAAA-----FACAAAABAAAABABABCAAAFB
S3-24.1	CBoBoFAFAAA-----BABABCAAAFB
S3-42.1 (6)	CBoBoFAFAAAC--AAAAFACAAAABAAAABABABCAAAFB
S3-30.1	CBoBoFAFAAAACA-----AAAABABABCAAAFB
S3-41.1	CBoBoFAFAAAACA?A--AFACAAAABAAAABABABCAAAFB
S3-42.7	CBoBoFAFAAACACAA--AFACAAAABAAAABABABCAAAFB
S3-23.1	CBoBoFAFAAACACAAAAAF-----BBB
S3-36.2	CBoBoFAFAAACACAAAAAFACAAA-----BABCAAAFB
S3-35.1	CBoBoFAFAAACACAAAAAFACAAAA-----BCAAAFBB
S3-32.1	CBoBoFAFAAACACAAAAAFACAAAA-----AAFBB
S3-42.8	CBoBoFAFAAACACAAAAAFACAAAABA--ABABABCAAAFB
S3-35.2	CBoBoFAFAAACACAAAAAFACAAAABAAAAB-----BBB

### ***Expansion processes in the germline***

Simple duplications and triplications similar to those observed in the soma were also detected in sperm DNA. Three expansions (a duplication in donor two and a duplication and triplication in donor 3) displayed mosaicism suggesting a premeiotic origin for at least some length-gain mutants (Figure 8.11). Most of the simple duplications were relatively small (2-10 repeat) with the largest involving 21 repeats. In contrast, approximately 2/3 of all length gain mutants displayed complex rearrangements (Figure 8.12). These highly complex mutants were only detected in the germline and there was no evidence of mosaicism, consistent with a meiotic origin. For the I/III heterozygous donors, about 55% of complex expansions appeared to involve intra-allelic rearrangements of the class I progenitor, although an inter-allelic component to the mutation mechanism cannot be excluded. Drastic reshuffling of repeats was observed at some mutants. For example at mutant S3-93.1 the centre of the allele has been profoundly re-modelled so that the origins of the re-organised repeats are unclear.

The remaining 45% of complex expansions detected in the sperm of the I/III heterozygotes displayed evidence for the inter-allelic transfer of repeats between alleles (Figure 8.12). For example, S1-51.1 contained two E-type repeats which were present in the class III allele, but absent from the class I progenitor. The position of the 3' E-type repeat within the mutant matches the position of an E repeat in the class III progenitor when alleles are aligned at the 3' end. Complex duplication of the resulting EAAAC motif would generate the observed mutant structure. Similarly, mutant 2S-51.1 contained the in-register transfer of an extended ACAoAAAAA motif from the class III progenitor allele. While this may have been the product of a gene conversion-like transfer of repeats, the observed structure is also compatible with unequal inter-allelic crossover. The division of complex mutants into intra- and inter-allelic mutants is in some cases questionable as apparent in-register inter-allelic transfer of repeat-types already present in the class I progenitor could be the result of intra-allelic duplication. With a single exception (S2-48.5), all putative inter-allelic repeat transfers were in-register between progenitor alleles when aligned at the 3' end, indicating a possible allele-pairing function in the 3' flanking DNA during meiosis.

## Figure 8.11

### ***Simple duplications detected in sperm DNA***

All simple duplications detected in sperm DNA of donors 1-3 are presented. Mutants are divided over multiple lines to clarify the nature of the duplication. The number of copies of each mosaic mutant detected is presented in parentheses.



Progenitor S1-44.1	CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAA</del> BABC <del>AAA</del> BBB CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA AAAA <del>BAAA</del> BABC <del>AAA</del> BBB
Progenitor S1-42.1	CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAA</del> BABC <del>AAA</del> BBB CB <del>o</del> B <del>o</del> A AAACAC ACAAAA ACAAAA <del>BAAA</del> BABC <del>AAA</del> BBB
Progenitor S2-54.1 (2)	CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BB CB <del>o</del> B <del>o</del> A AAACACAAAA ACA CACAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BB
Progenitor S2-48.1	CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BB CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAAA</del> ABAAAA <del>BABABC</del> AAA BB
Progenitor S2-46.1	CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BB CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAAA</del> BAB ABABC <del>AAA</del> BB
Progenitor S2-45.2	CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BB CB <del>o</del> B <del>o</del> A AAACACAAAA ACAAAA <del>BAAAA</del> BABA ABABC <del>AAA</del> BB
Progenitor S3-65.1	CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> BABAB <del>CAAA</del> BBB CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA A AAACACAAAAA ACAAAA <del>BAAAA</del> BABAB <del>CAAA</del> BBB
Progenitor S3-62.1	CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> CAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB
Progenitor S3-54.1 (2)	CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> B AAAA AAAA <del>BABABC</del> AAA BBB
Progenitor S3-53.1	CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA AA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB
Progenitor S3-51.1 (3)	CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> B AAAA <del>BABABC</del> AAA BBB
Progenitor S3-50.1	CB <del>o</del> B <del>o</del> A AAACACAAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB CB <del>o</del> B <del>o</del> A AAACAC AAACACAAAAA ACAAAA <del>BAAAA</del> BABABC <del>AAA</del> BBB

Progenitor CBoBo A AAACACAAAA ACAAAAABAAABABCAAA BBB  
S1-43.1 CBoBo A AAACACAAAA ACAAAAABAAA  
AAABABCAAA BBB

Progenitor S2-49.1 CBoBo A AAACACAAAA ACAAABAAAAABABCAAA BB  
CBoBo A AAACACAAAA ACAAABAAAAABABCAAA BB  
AAAAABABCAAA BB

Progenitor S2-47.1 CBoBo A AAACACAAAA ACAAABAAAAABABCAAA BB  
CBoBo A AAACACAAAA ACAAABAAAAABABCAAA BB  
CAAA BB

Progenitor S2-45.1 CBoBo A AAACACAAAA ACAAABAAAAABABCAAA BB  
CBoBo A AAACACAAAA ACAAAB  
AAAAABABABCAAA BB

<b>Progenitor</b>	CbBo	A	AAACACAAAAA	ACAAAABAAAABABBCAAA	BBB
S3-65.2	CbBo	A	AAACACAAAAA	AC	
	BoBo	A	AAACACAAAAA	ACAAAABAAAABABBCAAA	BBB
<b>Progenitor</b>	CbBo	A	AAACACAAAAA	ACAAAABAAAABABBCAAA	BBB
S3-59.1	CbBo	A	AAACACAAAAA	ACAAAABAAAABABAB	
				CAAAAABAAAABABBCAAA	BBB
<b>Progenitor</b>	CbBo	A	AAACACAAAAA	ACAAAABAAAABABBCAAA	BBB
S3-54.3	CbBo	A	AAACACAAAAA	ACAAAABAA	
				ACAAAABAAAABABBCAAA	BBB
<b>Progenitor</b>	CbBo	A	AAACACAAAAA	ACAAAABAAAABABBCAAA	BBB
S3-52.1	CbBo	A	AAACACAAAAA	ACAAAABAAAABABBCAA	
				BABAABC AAA	BBB
<b>Progenitor</b>	CbBo	A	AAACACAAAAA	ACAAAABAAAABABBCAAA	BBB
S3-51.4	CbBo	A	AAACACAAAAA	ACAAAAB	
				ACAAAABAAAABABBCAAA	BBB
<b>Progenitor</b>	CbBo	A	AAACACAAAAA	ACAAAABAAAABABBCAAA	BBB
S3-49.1	CbBo	A	AAACACAAAAA	ACAAAABAAAAB	
				AAAABABBCAAA	BBB
<b>Progenitor</b>	CbBo	A	AAACACAAAAA	ACAAAABAAAABABABBCAAA	BBB
S3-49.2	CbBo	A	AAACACAAAAA	ACAAAABAAAABAB	
				AAABAABCAA	BBB

## Figure 8.12

### ***Complex expansion mutants detected in sperm DNA***

A selection of complex expansion mutants detected in sperm are presented. A full list of complex mutants is presented in Appendix 8. Duplicated regions are split between lines to align mutants with the progenitor, though alternative alignments are often possible due to the complexity of the duplications. Probable or definite inter-allelic transfers from the class III donor allele are underlined in the mutant and donor alleles. '.....' indicates where class III alleles have been shortened due to their size.

# Figure 8.12

## Complex Intra-allelic mutants

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABACAAA BBB  
 S1-68.1 CBoBo A AAACACAAAA AC AAAABAAAABACAA  
 AAACACAAAA  
 CACAAAA-----BAAAABACAAA BBB

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABACAAA BBB  
 S1-60.1 CBoBo A AAACACAAAA  
 AAAAA AC AAAA  
 AAAAAA AC AAAABAAAABACAAA BBB

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABACAAA BBB  
 S1-52.1 CBoBo A AAACACAAAA AC AAAAB--BAB  
 ABA--BAB-AA  
 AAAABACAAA BBB

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABABA--BCAAA BB  
 S2-80.1 CBoBo A AAACACAAAA AC AAAABAAAABABABC  
 AAAABABABABC  
 AAAA AC AAAABAAAABABABCAAAA BB

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABABABCAAAA BB  
 S2-57.1 CBoBo A AAACACAAAA AACAA  
 AA AA AA  
 CACAAAAAAAABABABCAAAA BB

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABABABCAAAA BB  
 S2-48.4 CBoBo A AAACACAAAA AC AAAABAA  
 CAAAAABA--BABABCAAAA BB

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABABABCAAAA BBB  
 S3-109.1 CBoBo A AAACACAAAA AC AAAABAAAABABABCAAAA B  
 ABoBo A AAACACAAAA AC AAAABAAAABABABCAAAA B  
 AAAAAAAAABABABCAAAA BBB

Progenitor CBoBo A AAACACAAAA AC AAAABAAAABABABCAAAA BBB  
 S3-93.1 CBoBo A AAACACAAAA  
 AAA AC AAAABAAA  
 CAAAA  
 CAAAAAAC  
 A  
 AAAA AAAA  
 CAAAAAA  
 AC AAAABAAAABABABCAAAA BBB

## Complex Inter-allelic mutants

Progenitor .....CEAAABCEAAAABCAAAA BAAAEAAA BAAABA AAAA AABoHBBB  
 Progenitor CBoBo A AAACACAAAA A-----CAAAA BAAAABACAAA BBB  
 S1-51.1 CBoBo A AAACACAAAA AAEEAAACEAAACAAA BAAAABACAAA BBB

Progenitor .....BCEAAAABCAAAA BAAAEAAA BAAABA AAAA AABoHBBB  
 CBoBo A AAACACAAAA AC AAAABAAAABA-----BCAAA BBB  
 S1-46.1 CBoBo A AAACACAAAA AC AAAABAAAABA AAAABCAAAA BBB

Progenitor .....BoEA AEA AC AA ACEAEAA BAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAABAAAAB-----ABABCAAAA BB  
 S2-66.1 CBoBo A AAACACAAAA AC AAAABAAAABAAA AC AAAABAAAABAAA AABABABCAAAA BB

Progenitor .....ACAA ACEAEAA BAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAABAAAAB-----BABABCAAAA BB  
 S2-58.1 CBoBo A AAACACAAAA AC AAAABAAAAB AACAAAABAAA AABABABCAAAA BB

Progenitor .....CAA ACEAEAA BAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAA-----BAAAABABABCAAAA BB  
 S2-56.1 CBoBo A AAACACAAAA AC AAAAAAAAABCAAAA BAAAABABABCAAAA BB

Progenitor .....AA ACEAEAA BAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAABAAAAB-----BABABCAAAA BB  
 S2-55.1 CBoBo A AAACACAAAA AC AAAABAAAABAAACAAAAA AABABABCAAAA BB

Progenitor .....A ACEAEAA BAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAA-----BAAAABABABCAAAA BB  
 S2-54.5 CBoBo A AAACACAAAA AC AAAA AC AAAAAA AABABAAAABABABCAAAA BB

Progenitor .....CEAEAA BAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAABAAAABABABCAAAA B-----B  
 S2-51.1 CBoBo A AAACACAAAA AC AAAABAAAABABABCAAAA AACaOAAAAAB

Progenitor CBoBo BAO AEAAACEAAACA AAAA ACEAAACACA ACECC.....  
 Progenitor CBoBo-----A AAACACAAAA AC AAAABAAAABABABCAAAA BB  
 S2-48.5 CBoBo BAAAAA A AAACACAAAA AC AAAABAAAABABABCAAAA BB

Progenitor .....AABAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAABAAAAB-----BABABCAAAA BB  
 S2-47.2 CBoBo A AAACACAAAA AC AAAABAAAABAAA AABABABCAAAA BB

Progenitor .....ABAAAAA AAA AAAABCAA AAAAA AABoACaOAAAAAB  
 Progenitor CBoBo A AAACACAAAA AC AAAABAAAABABAB-----CAAA BB  
 S2-46.2 CBoBo A AAACACAAAA AC AAAABAAAABABABoACaOAAAAA BB

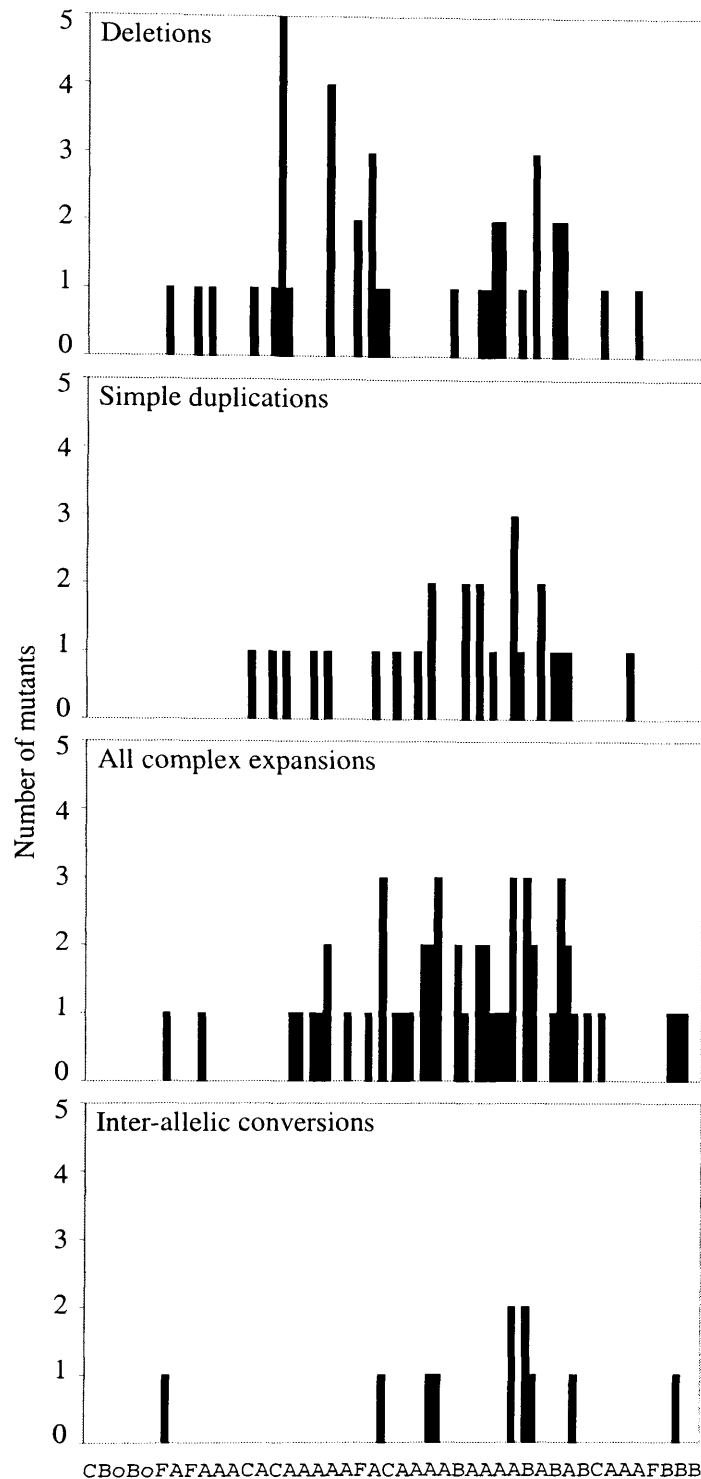
Mutational polarity was investigated further by locating breakpoints at which rearrangements occurred in germline mutants (Figure 8.13). In addition to mild polarity for putative inter-allelic mutants, a similar distribution of breakpoints was observed for both simple and complex duplications. This mild polarity is far less apparent than at minisatellites such as MS32 and MS205 (Jeffreys *et al.*, 1994; May *et al.*, 1996), and is primarily characterised by an avoidance of mutation at the 5' end of the repeat array.

Germline mutations characterised from sperm DNA of the true homozygote showed similar patterns of rearrangements to those seen in class I/III heterozygotes, with 56% of length gain mutants showing complex or very complex rearrangements. Whilst it is not possible to identify inter-allelic transfers of repeats in the homozygote, the frequency and characteristics of mutation in this individual suggest that homozygosity has no obvious effect on sperm mutation rate or process.

### ***PCR artefacts have simple structures***

Size enrichment for mutant molecules involves the collection of multiple size fractions from the enrichment gel, each covering a size range of between 4 and 15 repeats. Molecules detected by SP-PCR analysis of the size fractions which were outside the expected size window could therefore be identified as PCR artefacts. In total, 159 size-enriched mutants in the correct size window were identified, plus 37 additional artefacts of similar hybridisation intensity to authentic mutants, but which lay outside (usually substantially outside) the expected size range. All artefacts were characterised by MVR-PCR and structures presented in Figure 8.14. Thirty-four of the 37 artefacts differed from the progenitor allele by either simple deletion or duplication. However, there were three exceptions. Artefact AS2-44.2 (nomenclature is as described for true mutants, but prefixed by 'A' to denote artefact) was generated by a simple duplication and deletion, whilst AS3-59.1 had a duplication with repeat-type switch forming a null repeat. This change in repeat type was presumably the result of PCR misincorporation. Surprisingly, artefact AB1-45.1 displayed an apparent inter-allelic unequal crossover. The artefact however showed no repeat unit rearrangements of the type seen for true inter-allelic transfers and presumably arose by the annealing of single stranded DNA derived from a class I progenitor allele and a broken class III progenitor which co-migrated with the class I

**Figure 8.13**



***Distribution of sperm mutation breakpoints in class I alleles***

Breakpoint distributions are shown for deletions, simple duplications, all complex expansions, and inter-allelic conversions. Alignment of class I progenitor alleles allowed data to be combined for all three sperm donors, except for conversions which were combined for individuals 1 and 2 only. Breakpoints were determined by aligning each mutant with the progenitor class I allele to determine the location of the end of the region of 5' MVR identity between mutant and progenitor, and position of the beginning of the region of 3' MVR identity; the breakpoint was defined as the mean of these two positions. Each set of identical mosaic mutants are treated as a single event.

## Figure 8.14

### ***Structures of PCR artefacts***

Size enrichment for mutants excises DNA fractions within known size ranges from the enrichment gel. PCR artefacts were distinguished from true mutants by a difference between the size of molecule and the size range of the DNA fraction from which the molecule was amplified. Of the 37 artefacts detected from the one blood and three sperm samples, 34 were either simple deletions or duplications of the progenitor allele. Three artefacts of slightly greater complexity were identified and are discussed in the text. Nomenclature of artefacts is as for true mutant alleles, but prefixed 'A' to denote artefact.



**Figure 8.14**

**Simple deletions (class III)**

Progenitor	CB <sup>o</sup> BACAA <sup>o</sup> oACACACABAAAABABAA <sup>o</sup> EAA <sup>o</sup> BBAA <sup>o</sup> BAAAA <sup>o</sup> oACAB <sup>o</sup> BAAC <sup>o</sup> oACAA <sup>o</sup> BAAAC <sup>o</sup> oACACAAAA <sup>o</sup> BA <sup>o</sup> BA <sup>o</sup> AA <sup>o</sup> AAAA <sup>o</sup> oACAA <sup>o</sup> CAACAA <sup>o</sup> BA <sup>o</sup> CEAA <sup>o</sup> BC <sup>o</sup> AAAA <sup>o</sup> BCAAAA <sup>o</sup> BA <sup>o</sup> AA <sup>o</sup> AA <sup>o</sup> BA <sup>o</sup> BA <sup>o</sup> AAAA <sup>o</sup> AA <sup>o</sup> BoHBBB
AB1-105	CB <sup>o</sup> BACAA <sup>o</sup> oACACACABAAAABABAA <sup>o</sup> EAA <sup>o</sup> -----BA <sup>o</sup> AA <sup>o</sup> oACAA <sup>o</sup> CAACAA <sup>o</sup> BA <sup>o</sup> CEAA <sup>o</sup> BC <sup>o</sup> AAAA <sup>o</sup> BCAAAA <sup>o</sup> BA <sup>o</sup> AA <sup>o</sup> AA <sup>o</sup> BA <sup>o</sup> BA <sup>o</sup> AAAA <sup>o</sup> AA <sup>o</sup> BoHBBB
AS1-98	C-----CAAB <sup>o</sup> BAAC <sup>o</sup> oACACAAAA <sup>o</sup> BA <sup>o</sup> BA <sup>o</sup> AA <sup>o</sup> AAAA <sup>o</sup> oACAA <sup>o</sup> CAACAA <sup>o</sup> BA <sup>o</sup> CEAA <sup>o</sup> BC <sup>o</sup> AAAA <sup>o</sup> BCAAAA <sup>o</sup> BA <sup>o</sup> AA <sup>o</sup> AA <sup>o</sup> BA <sup>o</sup> BA <sup>o</sup> AAAA <sup>o</sup> AA <sup>o</sup> BoHBBB

**Simple deletions (class I)**

Progenitor	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AB1-32.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> B-----AA <sup>o</sup> BBB
AB1-34.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> AC-----AAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS1-37.1	CB <sup>o</sup> Bo <sup>o</sup> ---AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS1-38.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> --ACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS1-36.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> ---BAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS1-37.2	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ---AAA <sup>o</sup> BBB
Progenitor	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BB
AS2-36.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> -----ABAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BB
AS2-35.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> -----BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BB
AS2-39.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> ---AAB <sup>o</sup> ABABCAAA <sup>o</sup> BB
AS2-37.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> -----BAB <sup>o</sup> ABABCAAA <sup>o</sup> BB
Progenitor	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB
AS3-32.1	CB-----AAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB
AS3-25.1	CB <sup>o</sup> Bo <sup>o</sup> -----ABAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB
AS3-18.1	CB <sup>o</sup> Bo <sup>o</sup> -----ABAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS3-29.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAA-----BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB
AS3-32.2	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAAC-----AABAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB
AS3-24.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACAC-----ABCAAA <sup>o</sup> BBB

**More complex structures**

Progenitor	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
Progenitor	.....AAABCAAAA <sup>o</sup> BAAB <sup>o</sup> BAAB <sup>o</sup> BAAB <sup>o</sup> AAAA <sup>o</sup> AABoHBBB
AB1-45.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAABCAAAA <sup>o</sup> BAAB <sup>o</sup> BAAB <sup>o</sup> BAAB <sup>o</sup> AAAA <sup>o</sup> AABoHBBB
AS2-44.2	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACA A-----AABAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BB
AS3-59.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA CACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB

**Simple duplications**

AB1-58.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACA A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS1-53.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAA o <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS1-48.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BB BCAAA <sup>o</sup> BBB
AS1-45.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> CAAA <sup>o</sup> BBB
AS1-44.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAA A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABCAAA <sup>o</sup> BBB
AS2-51.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABABCAAA <sup>o</sup> ABABCAAA <sup>o</sup> BB
AS2-51.2	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABABCAAA <sup>o</sup> BABABCAAA <sup>o</sup> BB
AS2-44.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABABCAAA <sup>o</sup> AA <sup>o</sup> BB
AS2-60.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAB <sup>o</sup> ABABCAAA <sup>o</sup> BB
AS2-46.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> AAAA <sup>o</sup> BABABCAAA <sup>o</sup> BB
AS3-46.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABA BABCAAA <sup>o</sup> BBB
AS3-46.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABA BABCAAA <sup>o</sup> BBB
AS3-49.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCA <sup>o</sup> BABCAAA <sup>o</sup> BBB
AS3-49.2	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB
AS3-65.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BB AAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB
AS3-58.1	CB <sup>o</sup> Bo <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> A <sup>o</sup> AAACACAAAA <sup>o</sup> ACAAAA <sup>o</sup> BAAAA <sup>o</sup> BABABCAAA <sup>o</sup> BBB

alleles. Therefore, while the majority of true mutants in sperm displayed complex rearrangements, PCR artefacts had simple structures and were likely to have been generated by a slippage-like event arising during the first cycle of PCR.

## Discussion

Variant repeat analysis of allele diversity at the insulin minisatellite demonstrated that in Caucasians, virtually all alleles fall into three very distinct lineages of alignable alleles, classes I, IIIA, and IIIB. As with *D2S44* (Holmlund and Lindblom, 1998), *D19S20* and *MS51* (Chapter 7), there was a strong correlation between allele lineage and size with small class I alleles forming a single lineage and class III alleles dividing into two lineages. As will be discussed in Chapter 10, the correlation between allele size and lineage is even greater when class I alleles are divided by both MVR code and flanking haplotype. The restriction of the majority of alleles of the insulin minisatellite to just three highly diverged lineages is reminiscent of the lineage restriction observed at minisatellite *MS205* in Caucasian populations (Armour *et al.*, 1996a). A far greater allele lineage diversity was identified at this locus in African populations, consistent with a recent African origin for modern humans (Armour *et al.*, 1996a). Whilst it has been shown that haplotype diversity surrounding the insulin minisatellite is far higher in African populations compared with Caucasians (Mijovic *et al.*, 1997), almost no information exists on minisatellite allele diversity in non-Caucasians, other than a small number of alleles of Japanese origin. Sequence analysis of these Japanese alleles found them to be very similar to Caucasian class I alleles both in size and structure (Awata *et al.*, 1997). The lack of structural similarity of the previously published  $\lambda$ HI-4 allele of African-American origin (Rotwein *et al.*, 1986) to any Caucasian allele suggests that other populations may show different or additional allele lineages. Surveys of MVR code diversity in African populations to investigate levels of lineage diversity are planned.

While selection may maintain allele size within specific boundaries, the bimodal size distribution suggests that most mutation events result in small changes in overall repeat number. This was confirmed by analysis of patterns of allele diversity within each lineage, which revealed that the major mode of germline mutation involves insertions and deletion of 1-2 repeats, preferentially occurring within homogeneous arrays of A-type repeats. This



mutation process is reminiscent of microsatellite instability being facilitated by homogeneous repeat arrays (Chung *et al.*, 1993; Garza *et al.*, 1995; Weber, 1990), and suggests that polymerase slippage during replication may play an important role in mutation at the insulin minisatellite. This pattern of structural diversity within lineages is in striking contrast to that observed at minisatellite MS51 (Chapter 7) where the dominant form of allele variation involves repeat-type switches or larger insertions or deletions of repeat motifs. Polymerase slippage at the insulin minisatellite could be promoted by the formation of hairpin G-quartet structures mediated by the two (G)<sub>4</sub> blocks of residues present within most repeats, and which form *in vitro* most readily within homogeneous arrays of A-type repeats (Catasti *et al.*, 1996). (G)<sub>4</sub> blocks are not present within the sequence of MS51 repeats which display no obvious potential for the formation of secondary DNA conformations.

Whilst the major mode of mutation at the insulin minisatellite involves simple deletions and duplications, allele diversity studies also provided evidence for more complex mutation processes including complex gene conversion, in-register inter-allelic repeat transfers, and complex intra-allelic duplications. These complex mutation processes were demonstrated directly in sperm DNA by the isolation of mutant molecules derived from class I progenitor alleles. Highly complex reshuffling of repeats and repeat transfers between alleles were detected in sperm DNA but not in blood. The recombinational nature of inter-allelic transfers and their presence exclusively in the germline strongly suggests that these mutants have a meiotic origin. In contrast, simple duplications and deletions in the germline were frequently mosaic, indicative of a premeiotic origin for at least some mutants, as was supported by the detection of similar mutants in blood DNA. Variation at the insulin minisatellite is therefore driven by two distinct mutation processes; relatively frequent simple deletions and duplications most probably arising by replication slippage or unequal sister chromatid exchange and at least partially of premeiotic origin, and a low frequency of highly complex rearrangements similar to those seen at previously characterised hypermutable GC-rich minisatellites, and which are almost certainly of meiotic origin. The detection of mild polar instability, particularly for inter-allelic transfers, implies that a weak promotor of recombination may be located in the flanking DNA 3' of the insulin minisatellite towards the insulin gene.

Mutation rate was estimated from allele diversity at about  $9 \times 10^{-4}$  per gamete, and is surprisingly high given the bimodal allele size distribution and the apparent absence of any mutants detected in pedigrees (Prof. John Todd, pers. commun.) despite intensive analysis of allele transmissions due to the association of the minisatellite with disease (e.g. Bennett *et al.*, 1995; Bennett *et al.*, 1997). However, *de novo* mutations present in pedigrees may have been discarded as incidences of either non-paternity or human error resulting in sample mix up. Furthermore, previous determination of class III allele size has only been performed using genomic Southern blot hybridisation (Dr. M. McCarthy, pers. commun.) which would not provide the resolution of allele sizes required to detect small changes in allele length. This estimation of mutation rate should be treated with caution as selection almost certainly acts at (or near) the locus, and there is likely to be departure from mutation-drift equilibrium, particularly if there has been a population bottleneck in Caucasians. In addition, parallel changes will lead to an underestimation of mutation rate. Such parallel changes were detected, for example mutant S1-38.2 is identical to allele ID38.3, as are S2-40.1/ID40.1 and S3-34.1/ID42.1 (Appendices 7 and 8).

Mutation rate of class I alleles was also determined directly at  $0.14\text{--}0.3 \times 10^{-4}$  (Table 8.4) by the detection of *de novo* germline mutations. This is clearly much lower than the rate estimated from allele diversity with the discrepancy almost certainly due to an inability to detect mutations which change allele length by  $\pm 1$  repeat, and the major loss of efficiency in the detection of mutants changing allele length by  $\pm 2$  repeats. In addition, nothing is known of mutations which arise in the female germline and which may contribute substantially to allele diversity.

The high frequency of specific alleles in the Caucasian population (Appendix 7) allowed the effects of homozygosity on minisatellite mutation rate and mechanism to be investigated for the first time. Under a double-strand break (DSB) model of recombination, Holliday junction branch migration between homologues which display high levels of sequence divergence would result in the extensive formation of heteroduplex DNA. These mismatches could trigger abortion of the recombination complex without crossover and may account for the low frequency of crossover associated with gene conversion at human minisatellites (Jeffreys *et al.*, 1994). Under the synthesis-dependent strand annealing model

(SDSA) of recombination where each end of a double strand break invades the template independently without the formation of Holliday junctions, is extended by *de novo* DNA synthesis, then the strands are displaced and anneal to each other (Nassif *et al.*, 1994; Paques *et al.*, 1998), complex rearrangements may result either from differences in allele lengths or the formation of heteroduplex DNA upon strand invasion between alleles. Both models predict that mutants would be more complex in heterozygotes, and (at least for the DSB model) accompanied by an elevated frequency of crossover. While it was not possible to determine the rate of crossover or the frequency of inter-allelic mutations in the individual homozygous at the insulin minisatellite, germline mutations detected in sperm DNA were of at least equal complexity to those detected in the sperm of the two class I/III heterozygotes. Furthermore, mutation rate was higher in the homozygote than either heterozygote. It is therefore apparent that, at least at the insulin minisatellite and in this one homozygote, homozygosity does not reduce the frequency or complexity of mutational rearrangements.

## Chapter 9

# The insulin minisatellite and type 1 diabetes: a literature review

### Introduction to diabetes.

Diabetes results from the failure to properly regulate glucose homeostasis and affects ~6% of the general population (Todd and Farrall, 1996). Diabetes insipidus is characterised by copious micturation, whilst in diabetes mellitus large amounts of glucose are excreted in the urine due to an elevated fasting blood glucose concentration. Although multiple factors are involved in glucose homeostasis, insulin plays a central role. Insulin is secreted from  $\beta$ -cells of the pancreatic islets of Langerhans in response to an elevated concentration in the blood of glucose and amino acids and promotes the storage of nutrients in target tissues (mainly liver, muscle, and adipose tissue).

There are multiple forms of diabetes mellitus. Type 2 diabetes, also known as non-insulin-dependent diabetes mellitus (NIDDM) or maturity onset diabetes, has a diverse aetiology and can result from impaired insulin secretion due to the relative failure of pancreatic  $\beta$ -cells, resistance to insulin of target tissues which often results in elevated levels of insulin in the blood, or non-suppressible glucose production by the liver (Hager *et al.*, 1995; Kulkarni *et al.*, 1999). It is the most common form of diabetes and affects ~5% of the world population (Hager *et al.*, 1995). Clinically similar to type 2 diabetes is maturity-onset diabetes of the young (MODY).

Type 1 diabetes, also called insulin-dependent diabetes mellitus (IDDM) or juvenile-onset diabetes mellitus, is characterised by a total lack of insulin in the blood. With a median age-of-onset of 12 years, it is the most severe form of diabetes resulting in severe imbalance of glucose homeostasis with associated complications such as blindness, kidney failure, and neuropathy (Todd, 1995). While insulin injection can prolong life, it does not

completely prevent the associated pathologies. The lack of insulin results from T cell-mediated autoimmune destruction of the insulin-secreting  $\beta$ -cells of the pancreas.

## **Type 1 diabetes is caused by genetic and environmental factors**

Although type 1 diabetes generally displays polygenic inheritance, one exception has been reported. Wolfram syndrome (WFS) was first clinically described in 1938 as a combination of type 1 diabetes with optic atrophy and is believed to account for 1/150 type 1 diabetes patients (Fraser and Gunn, 1977; Wolfram and Wagener, 1938). WFS patients generally display clinical features diagnostic of multiple diseases as evidenced by the alternative name for WFS, DIDMOAD (diabetes insipidus, diabetes mellitus, optic atrophy, and deafness). Additional common symptoms include urinary tract atony, ataxia, peripheral neuropathy, mental retardation, and psychiatric illness (see Strom *et al.* (1998)). The syndrome displays monogenic autosomal recessive inheritance and linkage analysis within pedigrees mapped the disease locus to chromosome 4p16 (Collier *et al.*, 1996; Polymeropoulos *et al.*, 1994). Mutation screening of candidate loci identified the aetiological gene, *wolframin*, as a predicted transmembrane protein (Strom *et al.*, 1998). Expression of *wolframin* is ubiquitous but is at its highest levels in pancreatic islets. Mutations in the gene result in premature death of pancreatic  $\beta$ -cells resulting in type 1 diabetes (Inoue *et al.*, 1998).

Evidence for the combined effects of genetic and environmental factors underlying the vast majority of cases of type 1 diabetes susceptibility can be divided into three categories; familial clustering, differences in incidence of disease between populations, and evidence from animal experimentation.

### ***Familial clustering***

Concordancy for type 1 diabetes between monozygotic twins is estimated at 35-50% while concordance between dizygotic twins is 11% (Barnett *et al.*, 1981; Kyvik *et al.*, 1995; Olmos *et al.*, 1988). This compares to concordance between sibs of 6% (Risch, 1987). This demonstrates that both multifactorial genetic and environmental components (including *in utero* environment) affect disease susceptibility. Familial clustering is often defined by

the  $\lambda_s$  statistic; the relative risk to sibs of type 1 diabetes patients compared to the mean population risk. For type 1 diabetes,  $\lambda_s=15$  (6% lifetime risk to sibs compared to 0.4% in the general population (Risch, 1987)). It should be noted that  $\lambda_s$  does not directly express the degree of heritability as it also reflects the effects of environmental susceptibility factors shared between family members. Furthermore, the contribution of environmental factors to susceptibility cannot be precisely derived from concordancy levels between monozygotic twins as in some tissues they are not genetically identical. Type 1 diabetes is an autoimmune disease. V(D)J recombination will generate different populations of both T cell receptors (TCRs) and antibodies between monozygotic twins, and the composition of these lymphocyte populations may affect susceptibility. For example, a study of monozygotic twins suffering from the autoimmune disease multiple sclerosis (MS) found differences between expressed TCR molecules in MS discordant twins, but not MS concordant twins (Utz *et al.*, 1993).

### ***Variation in the incidence of type 1 diabetes between populations.***

In general, type 1 diabetes incidence rates in different populations can be divided into three categories; high in Nordic, Anglo Saxon, and Sardinian populations, medium in the rest of Europe, and low in developing countries (Buzzetti *et al.*, 1998). While this difference could be entirely due to genetic differences between populations, the reported inter-population variation in disease frequency is larger than for any other chronic non-communicable disease with a genetic basis (Diabetes Epidemiology Research International Group, 1988). Further support for a significant environmental component comes from temporal studies within a population. For example, incidence of type 1 diabetes in Finnish children below the age of 15 increased by an average of almost 3% per year between 1965 and 1992 (Tuomilehto *et al.*, 1995).

If environmental factors do have significant effects on susceptibility, immigration of a population with a low endemic incidence of type 1 diabetes to an area with a high incidence should result in elevated risk in the immigrant population. Two studies of Asian immigrants living in the UK found an increase in type 1 diabetes incidence in the first generation of UK-born Asians (reviewed by Akerblom and Knip (1998)). In contrast, immigration studies on Sardinians in Italy highlight the genetic component of susceptibility. Immigrants

maintained their high endemic levels of disease, whilst children of Sardinian/Italian couples had half the incidence (Buzzetti *et al.*, 1998).

### ***Animal experimentation***

The nonobese diabetic (NOD) strain of mice is a spontaneous model of autoimmune type 1 diabetes (Castano and Eisenbarth, 1990). Despite the high levels of homozygosity in inbred strains, the frequency of disease is less than 100%, with about 50-80% of females and 10-50% of males developing diabetes by 220 days (Wicker *et al.*, 1995). In a germfree environment, the frequency of disease increases significantly towards 100%. By contrast, disease frequency decreases if NOD mice are deliberately or accidentally infected with bacterial, viral, or parasitic organisms (reviewed by Todd (1999a)). This suggests that the elevation of type 1 diabetes incidence in several populations is due to improvements in standards of health and sanitation. Indeed, Gibbon *et al.* (1997) found that multiple infections during the first few years of life correlated with a reduction in diabetes susceptibility.

This chapter will focus on the genetic basis of type 1 diabetes with particular emphasis on *IDDM2*, and is divided into three sections which will address three questions: i) What loci contribute to type 1 diabetes susceptibility?; ii) How do loci contribute to type 1 diabetes susceptibility?, and ; iii) What other phenotypic effects associate with *IDDM2*?

### **i) The identification of type 1 diabetes susceptibility loci**

#### **Principles for genetic mapping of type 1 diabetes susceptibility loci**

There are three factors which make the identification of type 1 diabetes susceptibility loci more complex than loci which cause diseases showing strict Mendelian inheritance: i) Type 1 diabetes is multigenic and each gene may account for only a small degree of the observed familial clustering; ii) Alleles which predispose to type 1 diabetes are often found at high frequencies in the general population, and; iii) Type 1 diabetes is caused by interactions between multiple genetic and environmental factors. Despite these difficulties,

familial clustering and the knowledge of functional candidate loci have allowed the identification of a number of susceptibility loci.

### ***Linkage and association analysis***

There have been two general approaches to the identification of type 1 diabetes associated genes; linkage analysis and association studies. Linkage analysis involves the screening of cohorts of affected sib pairs (ASPs) for loci in which alleles show elevated identity by descent (IBD) in sibs. Under random segregation, two sibs would be expected to share 0, 1, or 2 copies of a given allele at frequencies of 25%, 50%, and 25% respectively. If a region shows IBD (scored as 0 and 2) at a frequency significantly in excess of 50% in affected sibs, that region is a candidate locus for a susceptibility gene. There are however considerable limitations to this approach. For example, the relative risk of many susceptibility loci may be low. It is estimated that in order to be 95% confident of detecting loci with  $\lambda_s=1.25$ , a cohort of 1500 ASPs would have to be analysed (Todd and Farrall, 1996).

Once a candidate locus has been identified either by linkage analysis or by a functional candidate gene approach, it can be analysed further by association studies. While linkage analysis considers co-inheritance of a locus, association studies analyse the frequency of specific alleles at a locus. Association studies take two forms: population-based, and family-based association studies. Population-based association studies compare allelic and genotypic frequencies at specific loci between a case population (disease sufferers) and a control population (non-sufferers). Significant differences between the populations indicate association of specific alleles with disease. A common criticism of this approach is the potential for population mismatching where genetic differences between case and control populations result from sampling effects as opposed to any disease association. Family-based association studies avoid many of the problems of population mismatching. One example of this approach is the transmission disequilibrium test (TDT) (Spielman *et al.*, 1993). If a parent is heterozygous for disease-associated and non-associated alleles at a susceptibility locus, the disease-associated allele will be transmitted to affected offspring at a frequency of >50%. The relative effects on disease susceptibility of the two parental



alleles can therefore be determined by measuring transmission deviation from the expected frequency of 50% across multiple families using a standard  $\chi^2$  test.

### ***Definition of the aetiological variant***

The above approaches all use polymorphic markers to test for disease associations. The detection of an association means either that the marker analysed is the aetiological variant causing the association, or that it is in linkage disequilibrium (LD) with the true aetiological variant. The identification of the aetiological variant requires two stages of analysis:

- i) The identification of all polymorphisms in LD with the associated locus, and;
- ii) The exclusion of all candidate polymorphisms other than the aetiological variant.

Whilst the first stage seems relatively straightforward, there are certain pitfalls to be avoided. One approach is to screen for multiple polymorphic markers surrounding the region of interest and individually test loci for disease association. The most proximal markers which do not associate, or associate poorly with disease susceptibility define the boundaries of LD and therefore the boundaries of the susceptibility locus. However, the use of single markers to identify disease-associated regions depends on the relative frequencies of alleles on disease-associated and non-associated haplotypes. Identical alleles could be present on both protective and predisposing haplotypes due to mutation (resulting in identity by state as opposed to descent) or gene conversion. Analysis of such a marker in isolation could therefore result in the false demarcation of the boundary of LD. Combinatorial analysis of multiple markers over a wider area is a more robust approach (Merriman *et al.*, 1998).

The identification of the aetiological variant within a region of LD relies on two main approaches; functional studies to demonstrate the effects of different variants on gene function, and cross-match haplotype analysis (Bennett *et al.*, 1995). Pedigree analysis or multiple-DNA-variant-association analysis (Julier *et al.*, 1994) allows single locus allelic associations to be grouped into haplotypic associations. If an allele is common both on haplotypes associated with disease predisposition and protection, that locus can be excluded as an aetiological candidate. This is the basis of cross-match haplotyping. The strategy will

fail if the region containing the aetiological variant is not included, and will be complicated if there is more than one common aetiological variant in the region (Todd, 1999b).

Within an ethnically homogeneous population, extensive regions of LD can exist as a result of population bottlenecks and genetic drift. If a region is affected by strong selective forces as may be the case for disease-associated loci, the extent of this LD would also be amplified. Consequently, there may be insufficient haplotypic diversity for the elimination of all candidate polymorphisms from a region of disease association by cross-match haplotyping. One solution is to compare predisposing haplotypes between populations of different ethnic origin thus elevating haplotype diversity. This relies upon the assumption that the aetiological variant is identical for different populations, which may not always be valid. Indeed, all association studies which analyse mean allele transmission or allele sharing must assume that the effects of a locus on disease susceptibility are uniform either between families or populations.

## **Examples of the genetic mapping of type 1 diabetes susceptibility loci**

### ***Genome-wide scans for type 1 diabetes susceptibility loci***

Two animal models have been used to investigate type 1 diabetes susceptibility loci; the nonobese diabetic (NOD) mouse (Castano and Eisenbarth, 1990), and the BioBreeding (BB) rat (Jacob *et al.*, 1992). The number of susceptibility genes in humans is probably higher than in animal models due to the elevated genetic heterogeneity in humans (Davies *et al.*, 1994; Todd *et al.*, 1987). (A locus which predisposes to disease will only be identified if it has variant alleles which do not predispose to, or protect against disease.) Studies on NOD mice provided the first evidence that polygenes could be mapped by linkage analysis and found evidence for at least 13 putative susceptibility loci (Ghosh *et al.*, 1993; Risch *et al.*, 1993; Todd *et al.*, 1991).

The first genome-wide scan for type 1 diabetes susceptibility loci in humans analysed IBD at 289 microsatellites in 96 UK ASPs and found evidence for linkage at *IDDM1*, *IDDM2*, and an additional 18 loci (Davies *et al.*, 1994). However, these results should be treated

with caution as is demonstrated by two examples. In this screen, the greatest IBD distortion was observed at locus *D4S430* ( $p < 10^{-5}$ ) (Davies *et al.*, 1994), but as opposed to an increase in allele sharing, the result reflected 'negative sharing' where significantly more ASPs inherited different alleles than the same alleles which is biologically implausible. Secondly, many of the regions identified as putative susceptibility loci failed to show significant linkage with disease in later studies. For example, a more recent screen of 679 ASPs failed to detect any linkage of *IDDM2* with disease (Concannon *et al.*, 1998). This is however not surprising due to the low level of familial clustering attributed to *IDDM2* ( $\lambda_s = 1.25$ ) (Todd and Farrall, 1996). While genome-wide scans using ASPs clearly have limitations, they have generated numerous candidate regions which are the focus of further analysis by the application of TDT and cross-match haplotyping. Evidence for type 1 diabetes association of three loci, *IDDM4* (11q13), *IDDM5* (6q25) and *IDDM8* (6q27) has been extended to genome-wide levels of significance (Davies *et al.*, 1994; Luo *et al.*, 1996), despite the relative risk associated with the *IDDM4* region being low ( $\lambda_s = 1.09$ ) (Nakagawa *et al.*, 1998).

### ***The functional candidate gene approach to the identification of type 1 diabetes susceptibility loci***

The functional candidate gene approach has resulted in the identification of three type 1 diabetes susceptibility loci; *IDDM1*, *IDDM2*, and *IDDM12*. While this chapter focuses on *IDDM2*, epistatic interactions have been described between *IDDM1* and *IDDM2*, so both loci will be discussed.

#### ***IDDM1 maps to the HLA complex***

HLA (Human Leukocyte Antigen) serotyping first identified a correlation between the presence of specific products of the HLA region and incidence of type 1 diabetes (Janeway *et al.*, 1999). Genetic analysis of the HLA showed that the two most common predisposing haplotypes *HLA-DR3* and *HLA-DR4* were found in >95% of European type 1 diabetes patients compared to 55% in non-diabetics, giving an attributable risk of ~90% (if these haplotypes were removed, incidence of type 1 diabetes would fall by ~90%; reviewed by Todd (1999b)). The effects of *IDDM1* on type 1 diabetes susceptibility are greater than for

any other locus with  $\lambda_s=2.6$ , accounting for ~40% of familial clustering (Davies *et al.*, 1994; Todd, 1995).

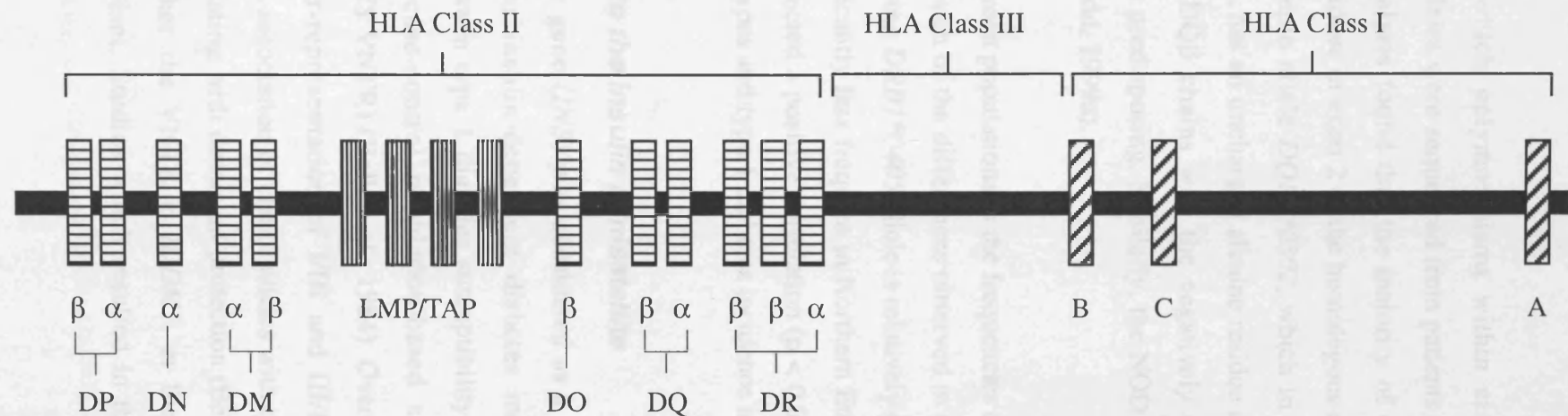
*IDDM1* refers to multiple tightly linked susceptibility loci (Nepom and Kwok, 1998). The HLA region spans ~3500 kb on chromosome 6p21 and contains over 200 genes, many of which are involved in the immune system (Janeway *et al.*, 1999). The region can be divided into three regions (Figure 9.1). The HLA class I region contains the three genes which encode  $\alpha$  chains of the HLA class I molecules HLA-A, -B, and -C. The class II region includes genes for both  $\alpha$  and  $\beta$  chains of HLA class II molecules called HLA-DR, -DP, and -DQ, as well as the *TAP* and *LMP* genes. Class III genes include many components of the complement, and cytokines such as tumour necrosis factor  $\alpha$  and  $\beta$  (*TNF- $\alpha$ , $\beta$* ), lymphotoxins (*LTA*, *LTB*), and the 21-hydroxylase genes *CYP21p* and *CYP21*.

Many genes within the HLA are highly polymorphic and the whole region is in strong linkage disequilibrium (LD) due to the strong selective forces acting on it from host-pathogen interactions. The first association of the HLA with type 1 diabetes was detected at the *HLA-B* class I locus (Nerup *et al.*, 1974; Singal and Blajchman, 1973). However, cross-match haplotype analysis identified the primary aetiological locus to be within the class II region 1.45 Mb telomeric of, but in strong LD with, *HLA-B* (Todd *et al.*, 1988). The majority of susceptibility and resistance associated with *IDDM1* has been mapped to *HLA-DQA1*, *-DQB1*, *-DRB1*, and probably *-DPA1/DPB1* (Todd, 1999b; Todd *et al.*, 1987) which code for the peptide chains HLA-DQ $\alpha$ , -DQ $\beta$ , -DR $\beta$ , and -DP $\alpha$ / $\beta$  respectively, although other loci within the HLA complex are thought to contribute to the overall effects of *IDDM1* (Lie *et al.*, 1999).

A primary role for the *HLA-DRB1* locus was established by analysing haplotypes which shared DQ alleles, but had different *DRB1* alleles (e.g. Noble *et al.* (1996)). These studies demonstrated both the effects of specific alleles, and epistatic interactions between alleles. For example, alleles *DRB1\*0405*, *DQB1\*0302* and *DQB1\*0201* result in type 1 diabetes susceptibility. However, a single protective allele on the haplotype such as *DRB1\*0403* or *DQB1\*0301* is sufficient to confer protection against type 1 diabetes. The relative risks

G

**Figure 9.1**



### **Genetic organisation of the HLA complex**

The Human Leukocyte Antigen (HLA) region is divided into three sections. The class I region contains three main genes; *HLA-A*, *-B*, and *-C*, which encode  $\alpha$ -chains of HLA class I molecules. The class II region includes the genes for the  $\alpha$  and  $\beta$  chains of the HLA class II molecules *HLA-DM*, *-DO*, *-DP*, *-DQ*, and *-DR*. *HLA-DO*  $\alpha$  and  $\beta$  chains are encoded by the *HLA-DN* and *HLA-DO* genes respectively. Alleles are named by identification numbers, for example allele *DQB1\*0302* is allele 0302 of the gene encoding the  $\beta$  chain of the DQ molecule (*DQB1*). The number '1' distinguishes the gene from pseudogenes located within the HLA region. The class II region also contains genes for the TAP-1:TAP-2 peptide transporter, and the *LMP* genes that encode proteasome subunits. The class III region contains genes encoding a variety of proteins that function in immunity. Figure 9.1 was adapted from Janeway *et al.* (1999).

associated with a number of haplotypes containing both susceptible and protective alleles are presented in Figure 9.2 (reviewed by Buzzetti *et al.* (1998)).

To determine which polymorphisms within class II genes affect type 1 diabetes susceptibility, alleles were sequenced from patients and controls in different ethnic groups. Cross-match analysis found that the majority of disease susceptibility associated with polymorphic residues in exon 2 of the homologous class II genes (Todd, 1990; Todd *et al.*, 1989). For example allele *DQB1\*0302*, which in many populations is the most highly associated allele, has an uncharged alanine residue at position 57 (reviewed by Nepom and Kwok (1998)). DQ $\beta$  chains with the negatively charged aspartic acid at this position (Asp57) are less predisposing. Similarly, the NOD mouse is the only mouse strain which lacks Asp57 (Todd, 1999a).

Differences between populations in the frequencies of *HLA* alleles and haplotypes are likely to account for much of the differences observed in disease susceptibility. For example, the highly predisposing *DRB1\*0405* allele is relatively common in Sardinia (8%) (Cucca *et al.*, 1995) but significantly less frequent in Northern Europe (Undlien *et al.*, 1997). Ronningen *et al.* (1994) detected a positive correlation ( $p < 0.003$ ) between the frequency of high risk *HLA-DQ* haplotypes and type 1 diabetes incidence in different populations.

### ***IDDM2 maps to the insulin minisatellite***

That the insulin gene (*INS*) was considered as a functional candidate gene involved in susceptibility to insulin-dependent diabetes mellitus requires no explanation. An association between type 1 diabetes susceptibility and the insulin gene region was first detected in a case-control population-based association study typing the insulin minisatellite (*INS* VNTR) (Bell *et al.*, 1984). Over-representation of class I homozygotes (I/I) with under-representation of I/III and III/III genotypes in the case population demonstrated an association of class I alleles with type 1 diabetes susceptibility, with class III alleles associating with dominant protection (Bell *et al.*, 1984). This led to two possible hypotheses; either the VNTR is *IDDM2*, or the VNTR is in LD with the *IDDM2* aetiological variant. Studies which resulted in the identification of the *INS* VNTR as

## Figure 9.2

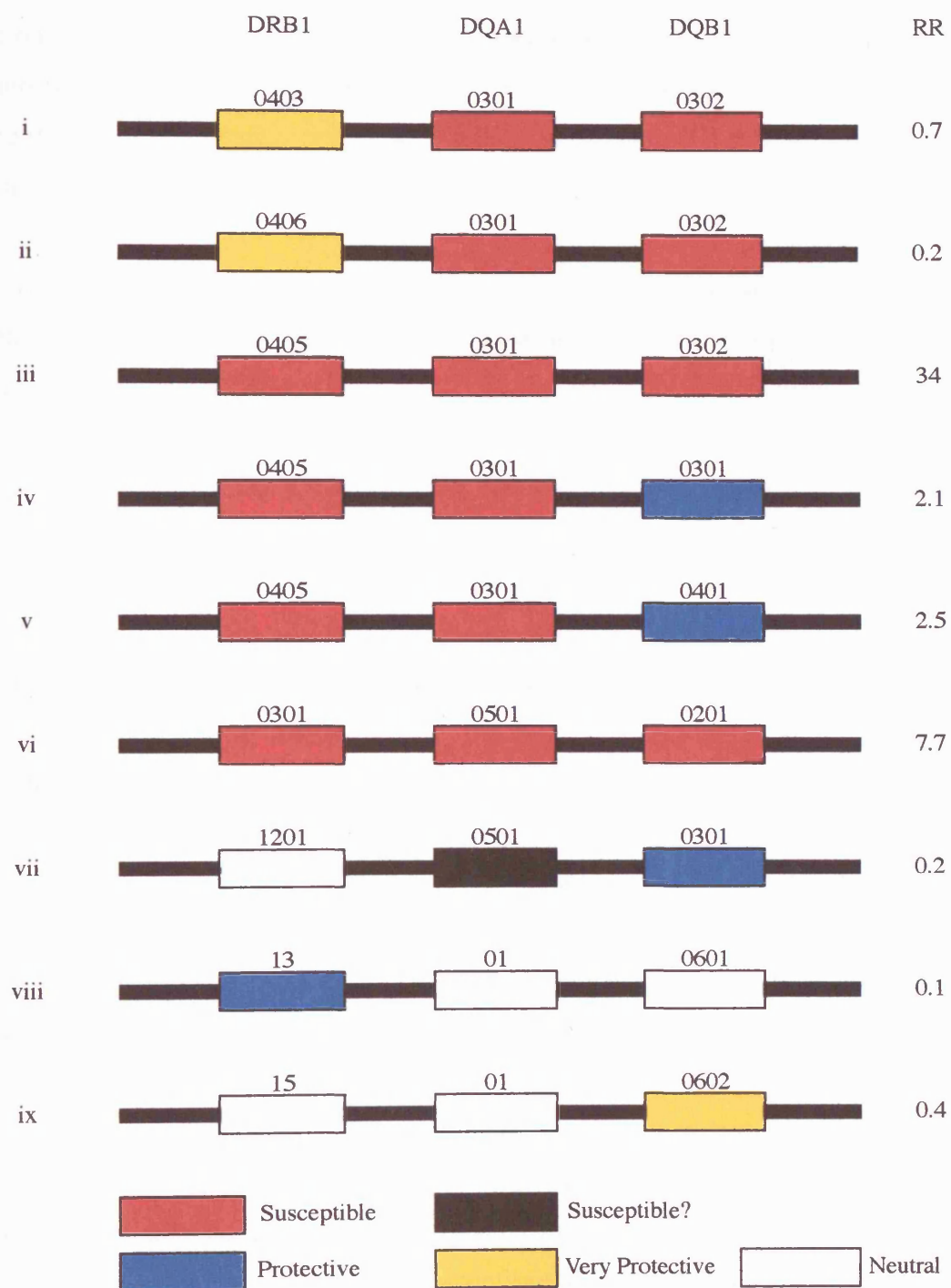
### ***Epistatic interactions between alleles of HLA class II haplotypes***

Among different combinations of HLA class II alleles, a single protective allele is sufficient to confer protection against type 1 diabetes, whereas susceptibility is due to a combination of susceptibility alleles. For example, haplotype iii has three alleles conferring type 1 diabetes susceptibility, resulting in an associated relative risk of 34. Haplotype ii is identical to haplotype iii with the exception of a single protective allele, *DRB1\*0406*. The associated relative risk falls to 0.2, indicative of a strongly protective haplotype. Figure 9.2 was adapted from Buzzetti *et al.* (1998).



C

**Figure 9.2**



*IDDM2* are of central importance to the results presented in Chapter 10, so will be considered in some detail.

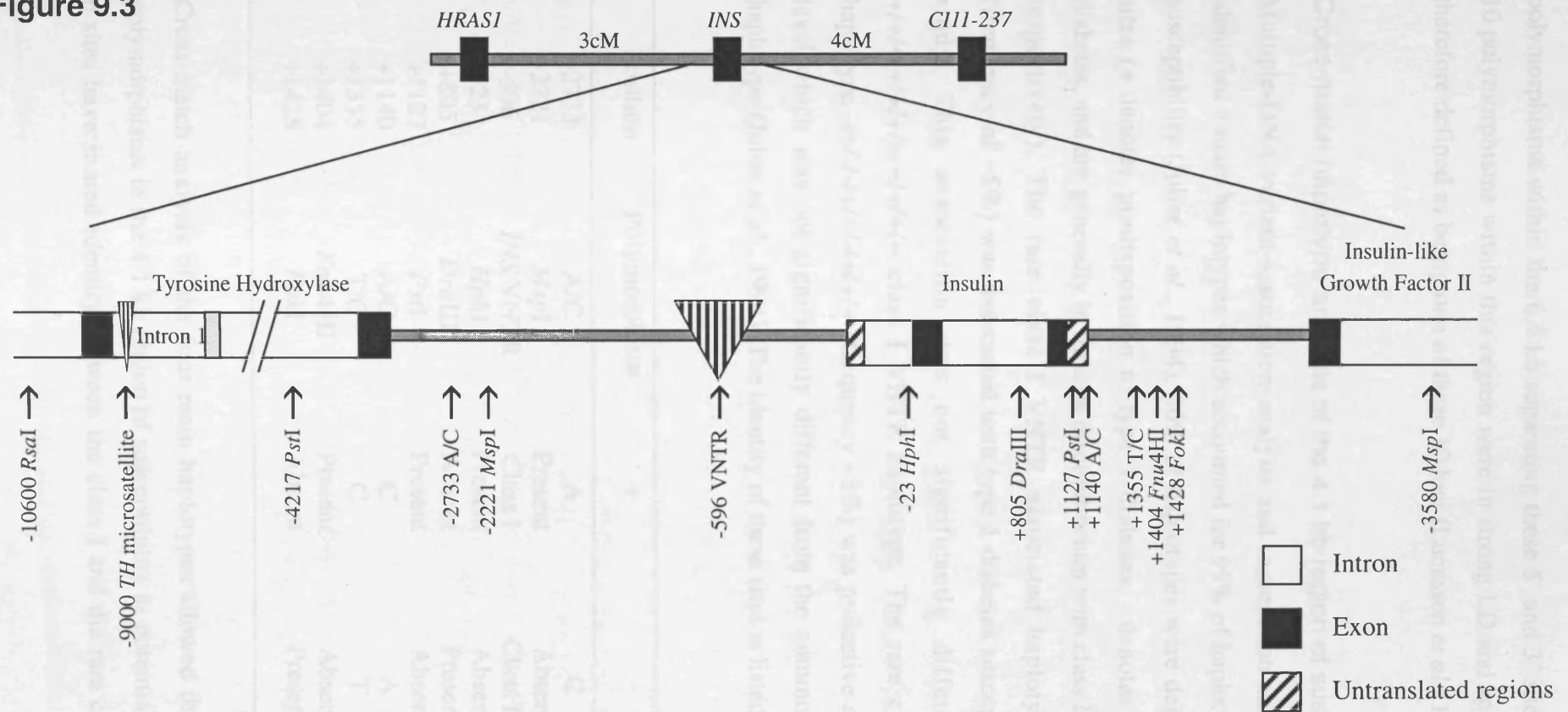
### Defining the boundaries of *IDDM2*

The definition by association studies of the aetiological basis of *IDDM2* began with two questions: i) How far from the VNTR does LD extend, and; ii) Were any other polymorphisms near the VNTR equally or more associated with type 1 diabetes susceptibility. In this chapter and Chapter 10, I will suggest that neither question has been fully answered.

To define the boundaries of association between the *INS* region and type 1 diabetes susceptibility, polymorphisms were selected for association studies at sites between the Harvey-ras-1 gene (*HRAS1*) 3 cM 5' of *INS* and the *C111-237* polymorphism at locus *D11S454* 4 cM 3' of *INS* (Julier *et al.*, 1991) (Figure 9.3). In addition to single polymorphisms at these loci, 2 sites in the gene for insulin-like growth factor II (*IGF2*), 4 sites in the tyrosine hydroxylase (*TH*) gene, and 8 sites within a 2 kb region including the *INS* gene were analysed. The higher density of polymorphisms surrounding *INS* reflects the strategy employed for the detection of polymorphisms which focused on *INS* as the major candidate gene. *INS* was resequenced in 4 haplotypes, with the 5' untranslated region between *INS* and the *INS* VNTR resequenced in 28 haplotypes. All other polymorphisms were identified by comparisons between previously published sequences (Julier *et al.*, 1991). Each site was typed for type 1 diabetes association in French case-control populations and multiplex families from France, N. Africa, and the USA. Only polymorphisms within 2 kb of *INS* were found to associate with susceptibility. However, the two sites distal to the 2 kb of association were separated by 19 kb, so this was initially defined as the region containing *IDDM2* (Julier *et al.*, 1991).

A region of 12.5 kb surrounding *INS* was resequenced in 4 individuals with the identification of a further 9 polymorphisms (Lucassen *et al.*, 1993). Association of each site was analysed independently in a larger French Caucasian cohort. Sites at -4217 and +2331 did not associate with type 1 diabetes susceptibility. [+1 is defined as the first base of the initiating codon on *INS* with the VNTR being 490 bp in length.] It was assumed that all

**Figure 9.3**



### Genetic organisation of the insulin-linked region on 11p15.5

Early investigations to define the aetiological basis of *IDDM2* analysed polymorphisms between the Harvey-ras-1 (*HRAS1*) genes and the *CIII-237* locus (Julier *et al.*, 1991). Further analysis of the insulin gene (*INS*) region identified multiple polymorphisms within a region including the tyrosine hydroxylase (*TH*) gene and the gene for insulin-like growth factor II (*IGF2*). Association analysis of each polymorphic site identified a 4.1 kb region of tight linkage disequilibrium between the polymorphic sites -2733A/C and +1428*Fok*I, within which *IDDM2* was identified (Lucassen *et al.*, 1993). +1 is defined as the first base of the ATG translation initiation codon of the *INS* gene. In the reference sequence, the VNTR is defined as being 490 bp long. Figure 9.3 was adapted from Julier *et al.* (1991) and Lucassen *et al.* (1993).

polymorphisms within the 6.5 kb separating these 5' and 3' sites had been identified. The 10 polymorphisms within this region were in strong LD and spanned 4.1 kb. *IDDM2* was therefore defined as being one of these 10 loci (Lucassen *et al.*, 1993).

#### Cross-match haplotype analysis of the 4.1 kb region of susceptibility.

Multiple-DNA-variant-association-analysis and segregation analysis of the French cohort identified 4 main haplotypes which accounted for 99% of haplotypes in the 4.1 kb region of susceptibility (Julier *et al.*, 1994). 86% of haplotypes were designated either + or - at all sites (+ denotes predisposition to type 1 diabetes, - denotes protection against type 1 diabetes, and are generally in linkage disequilibrium with class I and class III VNTR alleles respectively). The rare class I VNTR associated haplotype -/+ / + / + / + / + / + / + / + / + (frequency of ~5%) was associated with type 1 diabetes susceptibility in the case-control study. This association was not significantly different from the common + / + / + / + / + / + / + / + / + / + class I VNTR haplotype. The rare class III VNTR associated haplotype - / + / - / - / + / + / - / + / + / + (frequency ~8%) was protective against type 1 diabetes at a level which was not significantly different from the common - / - / - / - / - / - / - / - / - / - class III haplotype (Julier *et al.*, 1994). The identity of these sites is listed below.

Position	Polymorphism	+	-
-2733	A/C	A	C
-2221	<i>MspI</i>	Present	Absent
-596	<i>INS</i> VNTR	Class I	Class III
-23	<i>HphI</i>	Present	Absent
+805	<i>DraIII</i>	Absent	Present
+1127	<i>PstI</i>	Present	Absent
+1140	A/C	C	A
+1355	T/C	C	T
+1404	<i>Fnu4HI</i>	Present	Absent
+1428	<i>FokI</i>	Absent	Present

Cross-match analysis of the four main haplotypes allowed the exclusion of 6 of the 10 polymorphisms in the 4.1 kb region of susceptibility as potential *IDDM2* candidates. These 6 sites have shared identity between the class I and the rare class III haplotypes, yet the

associations of the haplotypes with diabetes are very different. The 4 sites which remain are -2733A/C, the *INS* VNTR, -23HphI, and +1140A/C. Each could be a functional candidate for *IDDM2*.

Support for the further exclusion of the -23HphI site came from association studies in Afro-Caribbean case-control populations (Mijovic *et al.*, 1997). The composition of haplotypes in this population was very different from Caucasians with the +/+ +/+ +/+ +/+ +/+ +/+ +/+ haplotype rare, and the -/- -/- -/- -/- -/- -/- haplotype absent. The + allele at -23HphI was at very low frequency in both case and control populations (4% and 1% respectively). By assuming that *IDDM2* has a common aetiology between Caucasian and Afro-Caribbean populations, the -23HphI site was excluded as a candidate (Mijovic *et al.*, 1997).

The rare class III haplotype was also identified in a UK cohort (Bennett *et al.*, 1995). The relative risk associated with the rare class III haplotype was lower than that of the common haplotype (RR=0.28 and 0.5 respectively). Analysis by TDT from I/III heterozygous parents in 425 multiplex families found this difference to be significant (p=0.048). The common haplotype was therefore designated the protective haplotype (PH) with the rare haplotype the very protective haplotype (VPH) (Bennett *et al.*, 1995). As described above, cross-match haplotype analysis between class I and VPH haplotypes allowed the exclusion of all but 4 polymorphic sites. Of the 4 remaining sites, 3 are biallelic and shared by both PH and VPH. The multi-allelic *INS* VNTR was therefore the only site to which the difference between PH and VPH could be attributed. By assuming that the effects of *IDDM2* were due to a single aetiological variant, it was concluded that the insulin minisatellite was *IDDM2* (Bennett *et al.*, 1995).

To test for variation in the levels of susceptibility associated with different class I alleles, alleles were divided into 21 classes by size. 15 size classes were present at sufficient frequencies for analysis by TDT from I/III heterozygous parents (30-44 repeats, designated alleles 641-843 respectively). All class I alleles were transmitted to diabetic offspring at a frequency of >50% with the exception of allele 698 (34 repeats), although the apparent protective effect of allele 698 was not significant. Heterogeneity in the level of

predisposition between size classes was detected ( $p=0.014$ ). 9/15 subclasses showed significant predisposition to type 1 diabetes. However the most common allele (allele 814, 42 repeats) did not show significant transmission deviation from 50% (transmission=58%,  $p=0.058$ ). The vast majority of haplotypes associated with class I alleles within the 4.1 kb region of susceptibility are identical. The transmission heterogeneity could therefore be attributed directly to variation at the VNTR providing further support for its identity as *IDDM2* (Bennett *et al.*, 1995).

To further investigate the anomalous behaviour of allele 814, transmission was analysed from 814/III parents from 1316 Caucasian families from Canada, Denmark, Italy, Norway, Sardinia, UK and USA (Bennett *et al.*, 1997). As in the previous study, transmission of allele 814 was lower than for all class I alleles combined (60.2% compared to 63.8%) but the difference between groups was not significant. The data were reanalysed after separation of maternal and paternal transmissions. Significant transmission distortion of allele 814 from 814/III mothers was detected ( $t_{814}=69.1\%$ ). However, 814 was not preferentially transmitted from fathers ( $t_{814}=51.6\%$ ). This difference between paternal and maternal transmission was significant ( $p=0.0017$ ). In addition, the lack of preferential paternal transmission was specific to allele 814 as the other class I alleles showed a 63.4% transmission bias from I/III heterozygous fathers, significantly different from the 814 transmission frequency ( $p=0.0093$ ). This anomalous behaviour was specific to the I/III genotype as 814 transmission from 814/I fathers did not differ from 50% (Bennett *et al.*, 1997). From this data, two interpretations are possible. Either both alleles from 814/III fathers are equally predisposing to, or equally protective against, type 1 diabetes. In a case-control study, the odds ratio for the 814/814 genotype was significantly lower than the I/I genotype, indicating the 814 allele was protective (Bennett *et al.*, 1997).

### *IDDM2* in different populations

Included in the above summary were details of the origin of the cohorts analysed as considerable heterogeneity has been observed between populations (Bennett *et al.*, 1996b). Class I allele predisposition to type 1 diabetes is greater when transmitted from I/III fathers in French, Canadian, and USA cohorts. However, maternally transmitted class I alleles were more predisposing in samples from Denmark, Sardinia and UK (Bennett *et al.*,

1996b). This difference between UK and USA populations was significant ( $p < 0.025$ ). This heterogeneity is difficult to explain without evoking significant genetic differences between populations, with a possible basis in differential imprinting patterns.

Type 1 diabetes is rare in Japanese populations and genetic factors play a major role in the low incidence ( $\lambda_s = 271$  in the Japanese, compared to  $\lambda_s = 15$  in Caucasians (Ikegami and Ogihara, 1996)). Although the frequency of class III alleles in the Japanese is low ( $< 10\%$ ) and there are considerable differences from Caucasian populations between class I allele size distribution (Kawaguchi *et al.*, 1997), three class I alleles with published sequences from Japanese populations were very similar to those from Caucasians. In addition, trimodal size distributions are seen for both populations with peaks at 32-33, 40-41 and 44 repeats in the Japanese compared to 31, 40, and 42 repeats in Caucasians (Awata *et al.*, 1997; Bennett *et al.*, 1995; McGinnis and Spielman, 1995). Subsequent subdivision of Japanese class I alleles into three groups (1S/1M/1L) according to this size distribution found that the small 1S/1S genotype was more predisposing to type 1 diabetes than other class I genotypes (Awata *et al.*, 1997). A study of UK and USA Caucasians divided class I alleles into 3 size groups apparently analogous to the Japanese subdivisions (McGinnis and Spielman, 1995). No significant heterogeneity between groups was observed. However, in the larger study by Bennett *et al.* (1995), 7/9 class I size subclasses of below 39 repeats were transmitted at frequencies significantly above 50% to diabetic offspring, compared to 2/6 classes of 39 repeats and larger. This trend is therefore consistent with that observed in the Japanese.

Allele 814 is at low frequency in the Japanese (Awata *et al.*, 1997). However, it was the most frequent allele in a non-diabetic population of Basques (Urrutia *et al.*, 1998). Comparison of allele frequencies with a case population found it to be the only class I allele which did not have an elevated frequency in the diabetic population, supporting the anomalous behaviour observed in other Caucasian populations (Urrutia *et al.*, 1998).

The *INS* region in Afro-Caribbeans shows considerably greater diversity than in Caucasians, both in the identity of haplotypes and VNTR size distribution (Mijovic *et al.*, 1997; Rotwein *et al.*, 1986) (Figure 9.4). The bimodal VNTR size distribution is not



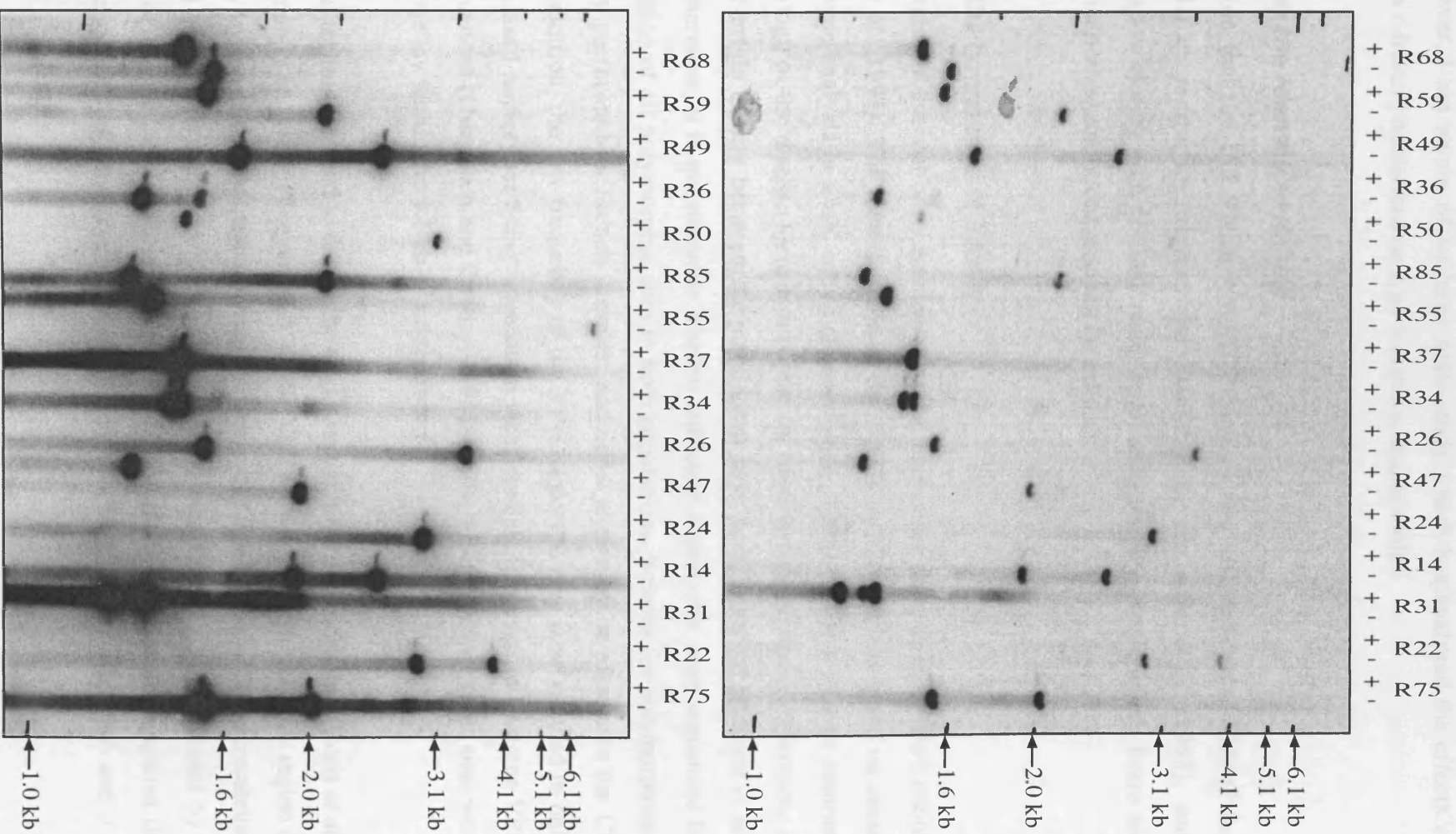
## Figure 9.4

### ***INS VNTR size distribution in an African sample population***

Two autoradiographs with different exposure times are presented showing allele-specific PCR products from amplification of the *INS* VNTR in a sample of 16 DNAs of African origin (laboratory reference codes are listed). PCR conditions were as described in Figure 8.2. The bimodal size distribution observed in Caucasian populations is not apparent, and whilst large alleles are generally associated with the lack of an *HphI* site at position -23 (as with Caucasian class III alleles), smaller alleles are detected in association with both the presence (+) and absence (-) of the -23*HphI* site.

Allele sizes range from ~1.2 kb (approximately 30 repeats) to >6 kb (approximately 350 repeats), substantially larger than any allele detected within Caucasian populations. The large allele size distribution and apparent haplotype diversity compared to Caucasian populations (Mijovic *et al.*, 1997) suggests a recent population bottleneck within the Caucasian ancestry. Autoradiographs were produced by K. Robbe.

**Figure 9.4**



apparent, and in one study analysing VNTR length polymorphisms in multiple races including African-Americans, an allele of over 540 repeats was detected (Rotwein *et al.*, 1986). However, I am aware of no large scale study which has analysed the effects of alleles size in African populations on type 1 diabetes susceptibility.

### ***Doubts over the identity of IDDM2***

The conclusion that the *INS* VNTR is *IDDM2* requires two assumptions: i) *IDDM2* lies within the 4.1 kb region of susceptibility identified by Lucassen *et al.* (1993), and; ii) All effects associating with *IDDM2* result from variation at a single locus. There are indications that both assumptions may not be valid.

### ***Redefining the boundaries of IDDM2***

The 4.1 kb region of susceptibility was identified by association analysis in a French cohort (Lucassen *et al.*, 1993). The identification of the VNTR as *IDDM2* was based on cross-match haplotype analysis in a UK cohort (Bennett *et al.*, 1995). While it may be assumed that the aetiology of the disease locus is identical for these two populations, differences in the parent-of-origin effects between the populations have been observed (Bennett *et al.*, 1996b). Furthermore, it is questionable whether sufficient haplotypes were sequenced for the identification of all polymorphic sites in both populations. Screens for polymorphisms were initially performed on French samples: additional sites may be present in the UK samples. In addition, the low frequency of the VPH haplotype may have resulted in only the PH and class I haplotypes being screened in these studies. However, at least one VPH has been sequenced (Owerbach and Gabbay, 1993). Whilst 3 novel polymorphic sites were identified, none lay within the 4.1 kb of susceptibility.

The statistical techniques used to define the 4.1 kb region have been criticised (Doria *et al.*, 1996; Merriman *et al.*, 1998). Sequences 5' of the VNTR were excluded from the region of susceptibility by the analysis of two polymorphic loci, -4127*Pst*I, and the *TH* microsatellite (*HUMTH01*) (Lucassen *et al.*, 1993). For each locus, relative risk was calculated by a comparison of the frequency of homozygotes for one allele (designated +) against the combined frequencies of heterozygotes and homozygotes for the other allele (+/- and -/-).

A low relative risk at both -4127*Pst*I and *HUMTH01* resulted in their exclusion as candidate susceptibility loci (Lucassen *et al.*, 1993). However, the allele designated + was defined as being either the most frequent allele, or the allele giving the greatest  $\chi^2$  value for association. *HUMTH01* was treated in their analysis as a biallelic marker, despite having 5 alleles (Brinkmann *et al.*, 1996; Polymeropoulos *et al.*, 1991; Puers *et al.*, 1993). Allele Z (also referred to as 122) was designated + as it gave a  $\chi^2$  of 5.6 with a relative risk of 0.35, resulting in its exclusion as a candidate. If allele Z-4 (118) had been designated +, the  $\chi^2$  value would be lower ( $\chi^2=5.0$ ), but the relative risk would be 4.7, higher than that associated with the *INS* VNTR (RR=4.5) (Lucassen *et al.*, 1993). Furthermore, if the *TH* and -4127*Pst*I sites were not associated with type 1 diabetes, the relative risk should approximate to 1. Values obtained for these loci (RR=0.35 and 0.7 respectively (Lucassen *et al.*, 1993)) show a substantial deviation from 1 indicating that if the alternative allele had been designated as +, a positive RR would have been obtained for each locus. The relative risk at -4127*Pst*I can be approximated as the reciprocal of the above value giving RR=1.4 (Doria *et al.*, 1996). This is still substantially lower than values determined at 5' and 3' sites. While this can exclude the *Pst*I site as a candidate for *IDDM2*, the possibility of gene conversion in this region prevents the use of a single site as a marker to define the boundary of LD.

Doria *et al.* (1996) typed the +1140A/C marker within the 4.1 kb region of susceptibility which is in linkage disequilibrium with class I and class III alleles, and an *Rsa*I polymorphic site located immediately 5' of the *TH* locus 10-15 kb upstream of the VNTR. Association with diabetes predisposition was analysed in a population of American Caucasian diabetics and controls. Association of the *TH* marker with disease was not significantly different from that of the *INS* marker (RR=2.6 and 3.3 respectively). It was concluded that the region associated with *IDDM2* could not be restricted to a 4.1 kb segment. This conclusion was confirmed by Merriman *et al.* (1998). Their re-analysis of the Lucassen *et al.* (1993) data using the *HUMTH01* allele which is in strongest LD with class I VNTR alleles identified a sufficiently significant association with type 1 diabetes to warrant an extension of the 4.1 kb region of susceptibility to a size of at least 9.1 kb. This re-analysis examined transmission of combinatorial haplotypes of markers, as opposed to

the association of individual markers with disease, an approach which has now been generally recommended (Merriman *et al.*, 1998).

### *IDDM2* may have a multi-locus aetiological basis

To postulate that the effects of *IDDM2* may have a multi-locus basis seriously affects the conclusion that the *INS* VNTR is *IDDM2*. If the majority of the effects associating with *IDDM2* are the result of a primary aetiological variant within the 4.1 kb of susceptibility, but secondary sites lying outside of this region can modify the effects of the primary site, then the variation seen between different class I and class III alleles may be caused by linkage of specific VNTR allele subclasses to the secondary variant as opposed to variation between different VNTR allele subclasses. Cross-match haplotype analysis within the 4.1 kb region identified the *INS* VNTR as the only locus which could account for the transmission heterogeneity of class I and class III alleles due to the multi-allelic nature of the minisatellite (Bennett *et al.*, 1995). A secondary site could equally account for this variation. Therefore, assuming that the primary locus of *IDDM2* is within the 4.1 kb region, the aetiological variant could be any one of the four sites which differ between class I and both class III PH and VPH haplotypes.

There are indications for the existence of such a secondary site. Five alleles at *HUMTH01* have been identified designated Z-16 (106), Z-12 (110), Z-8 (114), Z-4 (118) and Z (122). The majority of class III PH alleles are linked to the Z allele, whilst the majority of VPH alleles are linked to Z-8 (Bennett *et al.*, 1995). The difference between the effects of the PH and VPH was identified by dividing alleles by haplotype within the 4.1 kb region. Haplotypes were then extended to incorporate *HUMTH01* alleles. Any haplotype which differed from the Z/PH or Z-8/VPH modes was excluded. The difference between PH and VPH was greatest when the haplotype was extended to incorporate *HUMTH01* (%transmission PH=37%, VPH=26% compared to Z/PH=42%, Z-8/VPH=24%). This was interpreted as providing further support for the identity of the VNTR as *IDDM2* (Bennett *et al.*, 1995) as the extended haplotypes most closely resemble the ancestral PH/VPH haplotypes and exclusion of all non-ancestral haplotypes would increase VNTR homogeneity within each haplotypic group. An alternative interpretation is that the differences observed between PH and VPH haplotypes result from the effects of a

secondary locus such as *HUMTH01*. A similar result comes from the study of Doria *et al.* (1996) analysing LD between a site within the 4.1 kb region and a locus 5' of the *TH* gene. Multi-locus analysis found that individuals homozygous for both predisposing *INS* VNTR alleles (+/+) and alleles 5' of the *TH* gene (+/+) had a significantly higher RR than individuals with the same genotype at the minisatellite, but either +/- or -/- *TH* genotypes ( $p < 0.025$ ).

However, the insulin minisatellite is a highly polymorphic site located in a region which may affect expression of the insulin gene. Despite a lack of absolute proof, the VNTR does therefore remain a very strong candidate for the *IDDM2* aetiological variant.

## **ii) The development of type 1 diabetes**

In this section, the question will be addressed of how genetic and environmental factors act and interact to result in the development of type 1 diabetes. Type 1 diabetes is due to the autoimmune destruction of insulin-secreting  $\beta$ -cells of the pancreatic islets of Langerhans. This autoimmune response is dominated by T cells, specifically helper 1 T cells (Th1) (Todd, 1999b). A role has also been described for cytotoxic T cells (Augstein *et al.*, 1998). To provide a context within which the effects of various susceptibility factors can be understood, a general model of T cell development and activation will be described. [Unless otherwise stated, the reference source for the following section is Janeway *et al.* (1999).]

### ***The cell-mediated immune response***

#### **Antigen recognition**

Many pathogens display common antigens which are recognised as foreign by components of the innate immune response. This generates strong selective pressure on pathogens for antigenic variation. The immune system must therefore be capable of recognising a wide and ever-changing diversity of foreign antigens whilst distinguishing foreign from self antigens. The cell-mediated immune response targets intracellular pathogens which are not recognised by the mediators of the innate immune response. The binding of antigens by

T cell receptors (TCRs) is central to both T cell development and activation, and can only occur when antigens are presented by HLA molecules. To maximise the diversity of antigens which can be recognised in the context of the TCR-peptide-HLA complex, both TCR and HLA molecules are highly variable.

There are two classes of HLA molecule. Class I molecules display antigens from within the cytosolic compartment of almost every nucleated cell. Antigens from pathogens in the cell cytosol (such as viruses) will also be presented. Antigen-HLA class I complexes are recognised by CD8+ cytotoxic T cells which induce apoptosis in infected cells. Class II HLA molecules display antigens derived from intracellular vesicles and are found on the surfaces of only a few cell types, collectively termed antigen presenting cells (APCs). Antigens from pathogens (such as some bacteria) which are able to grow within intracellular vesicles are recognised by CD4+ helper T cells, whose primary function is the activation and recruitment of various components of both innate and adaptive immune systems.

Class I and class II molecules are composed of two peptides,  $\alpha$  and  $\beta$ . HLA class I and class II genes (class II genes are the major component of *IDDM1* (Bennett and Todd, 1996a)) encode the  $\alpha$  chain of class I molecules and the  $\alpha$  and  $\beta$  chains of class II molecules respectively. The gene for the class I  $\beta$ -microglobulin is located on a different chromosome. To elevate the diversity of antigens to which HLA molecules can bind, the HLA is polygenic and the genes are highly polymorphic. Both types of HLA molecule are unstable unless a peptide is bound. In class II molecules, this binding is mediated by interactions between four sites on the antigen (named anchor residues 1,4,6,9) and four residues on the HLA molecule (anchor pockets 1,4,6,9).

TCRs are also composed of two polypeptide chains ( $\alpha$  and  $\beta$ ) joined by a disulphide bond. The genes encoding both chains are composed of discrete segments that are joined by V(D)J somatic recombination during T cell development in the thymus generating a massive diversity of TCRs. Despite this diversity (an estimated  $10^{18}$  different TCRs could potentially form) the receptors have a relatively invariant shape as interaction with HLA molecules is required for their function. To produce an appropriate T cell response, TCRs

must have two properties: i) The ability to bind self HLA molecules, and; ii) The ability to recognise only foreign antigens. T cells which fail these two criteria are largely removed during development in the foetal thymus by positive and negative selection respectively.

### T cell selection in the thymus

Epithelial cells within the thymus express HLA class I and class II molecules on their surfaces. Immature thymocytes are in close apposition with these cells. If they fail to bind to an HLA-peptide complex, the thymocytes undergo apoptosis. Binding indicates that the TCR is able to recognise self-HLA molecules, so the immature T cell is rescued. This is the basis of positive selection. During the process of positive selection, each thymocyte expresses both CD4 and CD8 co-receptors. If the TCR binds to a class I HLA molecule, only CD8 will continue to be expressed. Binding to class II HLA molecules, restricts expression to CD4 co-receptors. It is this restriction that determines whether the T cell develops into a cytotoxic CD8+ T cell, or a helper CD4+ T cell.

From the developing T cell population, it is necessary to remove any T cells which might recognise, therefore react against, self antigens. This process is called negative selection. Antigen presenting cells (APCs) activate mature T cells in the periphery by displaying antigens complexed to both class I and class II HLA molecules, and releasing secondary activation signals. There are three main cell types which can function as APCs; bone marrow-derived dendritic cells, macrophages, and B cells. Bone marrow-derived dendritic cells are also central to the process of antigen presentation and recognition that leads to negative selection in the thymus. These APCs express a range of self epitopes presented on HLA molecules. If a developing thymocyte encounters an antigen-HLA complex to which it can bind strongly, the cell undergoes apoptosis, thus removing autoreactive T cells from the population.

There is a clear paradox within the above model. Positive selection means that a T cell which fails to bind an HLA-peptide complex through its TCR will not develop. Negative selection means that any cell able to bind and HLA-peptide complex will undergo apoptosis. The resolution of this paradox is thought to relate to binding affinity. If a TCR binds weakly to an HLA-peptide complex, this indicates a recognition of self HLA



molecules, so is sufficient for positive selection to rescue the T cell from cell death. However, if the TCR binds with high affinity, this indicates a recognition both of HLA molecule and a self peptide, which results in negative selection by apoptosis. Binding affinity may be measured both by the direct interaction within the TCR-peptide-HLA complex (possibly measured by the longevity of the binding) and the density of complex formation on the T cell membrane. It has been shown that, at least for some T cell-APC interactions, full T cell response requires a density of HLA-peptide complexes of 60 molecules  $\mu\text{m}^{-2}$ , which represents 100-200 molecules per APC (Grakoui *et al.*, 1999).

### Induction of T cell anergy

Following clonal selection, mature T cells leave the thymus as naive (inactive) T cells, enter the blood stream and continually circulate between blood and peripheral lymphoid tissues such as the lymph nodes, spleen, and mucosal lymphoid tissues, until they encounter antigen. However, not all self proteins expressed throughout the body are expressed in the thymus, so the mature T cell population will contain many self reactive clones unaffected by negative selection. To prevent an autoimmune response, T cell activation requires not only effective binding of the TCR-peptide-antigen complex, but also detection of secondary activating signals. TCR-peptide-HLA binding in the absence of these signals not only fails to activate the cell, but reduces its potential for future activation even in the presence of secondary signals, inducing a state called anergy. This self tolerance which develops in the periphery complements the negative selection in the thymus.

### T cell activation

APCs are the only cell types which can display both class I and class II HLA molecules, and release the secondary signals required for T cell activation. Dendritic cells are the most effective T cell stimulators and are present in an immature form under most surface epithelia and in most solid organs. Immature cells are incapable of T cell stimulation but are very active in antigen uptake both by receptor-mediated phagocytosis and non-specific macropinocytosis. The detection of infection inhibits antigen uptake and stimulates their migration to local lymphoid tissues where they express high levels of HLA class I and class II molecules and adhesion molecules which result in transient T cell binding. Antigen

recognition through the TCR-peptide-HLA complex elevates the binding affinity of T cell-expressed adhesion molecules stabilising the APC-T cell interaction. A greater density of peptide-HLA complexes increases the capacity for T cell antigen recognition. This depends in part on the number of different peptide-HLA complexes expressed on the APC surface. Insulin (which associates with *IDDM2* (Bennett and Todd, 1996a)) is a very small protein (31 amino acids) so degradation within an APC will generate a low diversity of epitopes. This may increase the density of specific insulin-derived antigens thereby facilitating T cell recognition and elevating the probability of an autoimmune response.

Co-stimulatory signals from the APC are required for T cell activation. The best characterised signals are the B7 glycoproteins (CD80 and CD86) which interact with the CD28 receptor of the T cell. Activated T cells express the cytokine interleukin-2 (IL-2) and its receptor. This self-stimulation results in the proliferation and differentiation of T cells. The proliferation must be regulated, so activated T cells express the CTLA-4 receptor which binds B7 molecules with ~20 times the avidity of CD28, and inhibits T cell proliferation by competing with the stimulatory CD28 and limiting IL-2 production. (*CD28* and *CTLA-4* genes are both candidates for *IDDM12*.) It is the inability to generate IL-2 that characterises anergic T cells, resulting in self tolerance.

### T cell response

After 4-5 days of proliferation and differentiation, the activated T cells leave the peripheral lymphatic tissues and migrate to sites of infection and inflammation. Once activated, further encounters with its specific HLA-bound antigen result in immune attack without the need for co-stimulatory signals. Following antigen recognition, CD8+ cytotoxic T cells release cytotoxins (which cross target cell membranes inducing apoptosis) and cytokines. The main cytokine released is interferon- $\gamma$  (IFN- $\gamma$ ) which blocks viral replication and activates macrophages. CD4+ Th1 cells promote cell-mediated immune responses (as opposed to Th2 cells which promote humoral immune responses) primarily by activating anti-microbial mechanisms in macrophages. Pathogens within intracellular vesicles survive by either inhibiting the fusion of lysosomes with endocytosed vesicles, or preventing the acidification of vesicles which is required for the activation of lysosomal

proteases. Activated macrophages also release free oxygen radicals (FORs) and nitric oxide (NO) which destroy extracellular pathogens. This also results in some tissue damage.

Once initiated, the immune response can be amplified by the recruitment of further components of the immune system. For example, activated macrophages become more potent APCs recruiting further naive T cells, whilst activated Th1 cells release factors which attract additional macrophages. Local inflammation activates more dendritic cells which again activate T cells. Pancreatic  $\beta$ -cells may display HLA class II molecules (this can be induced *in vitro* by the application of cytokines IFN- $\gamma$  and TNF- $\alpha$  (Pujol-Borrell *et al.*, 1987)). Th1 cells release the co-stimulatory signals required for T cell activation so more T cells can be recruited, and so the immune response may continue until the target antigens are removed.

From the above model, it is possible to define 7 stages which may lead to the development of type 1 diabetes.

- 1 Autoreactive T cells escape negative selection.
- 2 Anergy fails to develop.
- 3 Pancreatic  $\beta$ -cell damage results in endocytosis of  $\beta$ -cell antigen by APCs.
- 4 Pancreatic inflammation results in APC activation.
- 5 Activated APCs display  $\beta$ -cell antigen-HLA class II molecules complexes resulting in Th1 cell activation.
- 6 The activated T cells initiate an autoimmune response.
- 7 The autoimmune response is sustained, resulting in  $\beta$ -cell destruction and type 1 diabetes. (Transient autoimmune responses to tissue damage are common but are usually regulated and cause no further tissue damage. A sustained response may result from prolonged inflammation, or an unusual susceptibility to tissue damage.)

Each genetic and environmental type 1 diabetes susceptibility factor could act at one or multiple stages through this progression towards disease. A number of potential determinants of susceptibility will now be considered.

## ***The Pathology of environmental determinants of type 1 diabetes susceptibility***

Environmental factors are likely to elevate type 1 diabetes susceptibility by two main routes: i)  $\beta$ -cell damage resulting in the release of  $\beta$ -cell antigens, inflammation, and APC activation, and; ii) T cell activation.

### **$\beta$ -cell damage**

$\beta$ -cells have low scavenging capacity for free oxygen radicals (FORs) so are unusually sensitive to FOR damage (Malaisse, 1982). Exposure to *N*-nitroso compounds has been correlated with type 1 diabetes and may act by the oxidation of  $\beta$ -cell components. High levels of zinc in groundwater have been correlated with a decrease in diabetes incidence, possibly due to the role of zinc in Zn/Cu superoxide dismutase. Activation of macrophages will also result in higher levels of tissue damage due to the release of FORs (reviewed by Akerblom and Knip (1998)).

A number of viruses are capable of inducing diabetes in experimental animals. In humans, both case reports and epidemiological studies have described an elevated incidence of type 1 diabetes following infection with viruses such as the mumps virus, rubella, and a number of RNA enteroviruses. In some cases, this could be due to viral infection of  $\beta$ -cells which would induce a cytotoxic T cell response both destroying  $\beta$ -cells and sensitising potentially autoreactive T cells to  $\beta$ -cell components (reviewed by Akerblom and Knip (1998)).

### **T cell activation**

Superantigens are produced by many different pathogens including bacteria and viruses, and bind directly to the outer surfaces of both the HLA class II molecules and a subset of the TCR population. This can result in the non-specific stimulation of 2-20% of all Th cells increasing the probability of an autoimmune response (Janeway *et al.*, 1999).

Autoimmunity may also be induced by molecular mimicry in which T cells that are activated in response to a foreign antigen cross-react with self antigens with similar properties (reviewed by Akerblom and Knip (1998)). Many putative mimics have been

described, for example antibodies against a retroviral p73 antigen are thought to cross-react with insulin. Many components of cow's milk may be molecular mimics and numerous studies have found an inverse correlation between the frequency and duration of breast feeding and type 1 diabetes susceptibility. This correlation is greatest where infants were exposed to cow's milk during the first 2-3 months when the gut is permeable to proteins. Three components of cow's milk have been identified as possible mimics. Bovine serum albumin (BSA) differs from human serum albumin in a small region from residues 152-168. T cells which recognise this region cross-react with a p69 antigen transiently expressed by  $\beta$ -cells in response to viral infection, so could result in an autoimmune response.  $\beta$ -casein concentration in cow's milk is much higher than in human milk. A putative region of mimicry has been identified between  $\beta$ -casein and the  $\beta$ -cell glucose transporter molecule Glut-2. Finally, bovine insulin which differs from human insulin by just 3 amino acids is present in cow's milk. A humoral response to bovine insulin has been detected and the resulting antibodies were found to cross-react with human insulin.

### ***The pathology of genetic determinants of type 1 diabetes susceptibility***

#### ***IDDM1***

Class II HLA molecules are a central component of antigen presentation to helper T cells resulting in positive and negative selection in the thymus, the development of anergy in the periphery, and APC-mediated T cell activation. Most of the effects associated with *IDDM1* were mapped to polymorphisms within exon 2 of class II HLA genes (Todd, 1990; Todd *et al.*, 1989). Variation in this region significantly effects HLA-peptide binding affinity and therefore the capacity of an individual to mount an immune response against both foreign and self antigens (Tisch and McDevitt, 1996). HLA class II molecules with low antigen binding affinities could theoretically predispose to type 1 diabetes in many ways.

Positive and negative T cell selection depends both on binding affinity and density. An HLA class II molecule which binds a self-peptide weakly could positively select for self-reactive T cells but prevent negative selection. A subset of HLA molecules with low binding affinities would also reduce the overall diversity and density of antigens displayed. This would have a general effect of reducing negative selection whilst maintaining positive

selection. More autoreactive T cells would therefore leave the thymus, elevating autoimmune predisposition. Low HLA-peptide binding affinity would also reduce the probability of autoimmune T cell contact with self antigens in the periphery in the absence of infection, thereby preventing the induction of anergy.

Not all self proteins are expressed in the thymus. A self antigen which binds HLA molecules weakly in the thymus would encourage maturation of autoreactive T cells. If a similar self antigen was present in the periphery which could bind the same HLA molecule with high avidity, an autoimmune response would result. Expression of the insulin gene in the thymus results in the production primarily of proinsulin (Pugliese *et al.*, 1997), while insulin is produced in the periphery. Insulin is the only known  $\beta$ -cell-specific autoantigen (Vafiadis *et al.*, 1997). T cells which bind proinsulin-derived antigen-HLA complexes weakly in the thymus could cross-react strongly with insulin-derived antigen-HLA complexes in the periphery, resulting in  $\beta$ -cell autoimmune destruction. Finally, a higher density of HLA-peptide complexes in the periphery (perhaps resulting from the release of large amounts of stored insulin from within damaged  $\beta$ -cells) could compensate for low binding affinity resulting in T cell activation.

An *IDDM1* susceptibility haplotype requires all *DQB1* and *DRB1* alleles to be predisposing (Buzzetti *et al.*, 1998) (Figure 9.2). HLA molecules with high antigen avidity would sequester autoantigens both in the thymus and periphery. This would reduce positive selection for T cells which could bind HLA-antigen complexes weakly in the thymus. Autoreactive T cells which do escape thymic selection due to poor affinity of HLA-peptide binding would be less likely to encounter the same complex in the periphery, again due to antigen sequestration. The overall effect of HLA molecules with high binding affinities would therefore be one of dominant protection.

The *DQB1*\*0302 allele is the most common allele associated with type 1 diabetes in many populations (Nepom and Kwok, 1998). Its product has an alanine residue at position 57 of the DQ $\beta$  chain as opposed to aspartic acid (Todd *et al.*, 1987). Asp57 normally forms an ionic bond with Arg79 of the DQ $\alpha$  chain stabilising the heterodimer. DQ molecules such as DQ3.2 without Asp57 therefore display elevated instability, the consequences of which

have been discussed. If DQ3.2 binds an antigen with a negative charge at anchor residue 9, an ionic bond can form between this site and Arg79 (reviewed by Nepom and Kwok (1998)). Intrinsically unstable HLA molecules can therefore be stabilised by specific antigens. If a self antigen expressed in the periphery differed from a self antigen in the thymus by a single residue, binding affinity could therefore be greatly altered. If a T cell was able to recognise both antigens, it would escape thymic negative selection, but be activated in the periphery. To support this, Hemmer *et al.* (1997) have shown that TCRs are able to cross-react against many antigens bound to self-HLA molecules.

## **IDDM2**

### **Levels of gene expression correlate with *INS* VNTR identity**

The insulin minisatellite is thought to affect susceptibility to type 1 diabetes by acting as a transcriptional enhancer element (Paquette *et al.*, 1998). Transcriptional variation at three genes in the *INS* region could theoretically alter susceptibility; the tyrosine hydroxylase gene (*TH*), the insulin gene (*INS*) and the gene for insulin-like growth factor II (*IGF2*). The *TH* gene produces the rate limiting enzyme in the synthesis of catecholamines (such as adrenaline) which are involved in the control of insulin secretion by pancreatic  $\beta$ -cells (Meloni *et al.*, 1998). IGF-II promotes T cell survival in the thymus by the inhibition of apoptosis. It may also promote T cell proliferation in the periphery. To support this, mice which overexpress transgenic *IGF2* have an unusually large thymus and T cell population, whilst circulating T cells display IGF-II receptors and respond to IGF-II by elevated cell division (Kooijman *et al.*, 1994). Therefore, increased expression of *IGF2* could elevate the number of T cells which escape thymic negative selection and promote their proliferation in the periphery. *IGF2* is expressed monoallelically in most tissue (this is relaxed to varying degrees in lymphocytes (Polychronakos *et al.*, 1995)) so this could explain the parent-of-origin effects observed at *IDDM2*. A number of studies analysing *IGF2* mRNA expression failed to detect any difference which associated with the identity of the VNTR in either the thymus, pancreas or leukocytes (Vafiadis *et al.*, 1996; Vafiadis *et al.*, 1998a; Vafiadis *et al.*, 1998b). In contrast, *in vitro* expression studies and analysis of steady state *IGF2* mRNA levels in human placentae found higher expression to associate with class I

VNTR alleles than class III alleles, indicating that the VNTR may affect *IGF2* expression at least in some tissues (Paquette *et al.*, 1998).

Elevated insulin gene expression in the thymus could increase the potency of negative selection against any T cells which recognise insulin-derived epitopes, whilst decreased expression in the pancreas may reduce the density of insulin antigen presentation (Pugliese *et al.*, 1997; Vafiadis *et al.*, 1997). Immunological data from both humans and mice indicate a primary role for both insulin and its precursors as autoantigens (Wegmann, 1996). Alternatively  $\beta$ -cell damage could relate to their activity. For example if elevated *INS* expression resulted in a substantial increase in cellular metabolic rate, it could increase the rate of synthesis of FORs thereby increasing tissue damage.

Analysis of *INS* transcription in transiently transfected  $\beta$ -cell lines detected an increase (Kennedy *et al.*, 1995) and a decrease (Lucassen *et al.*, 1995) in transcription associated with class III alleles. Allele-specific RT-PCR analysis of adult and foetal pancreata have also generated mixed results (Bennett *et al.*, 1995; Bui *et al.*, 1996; Owerbach *et al.*, 1982; Permutt *et al.*, 1985; Weaver *et al.*, 1992). Analysis of 10 foetal pancreata found a small but significant reduction in *INS* transcription associated with class III alleles (Vafiadis *et al.*, 1996) and reduced *INS* expression in adult pancreas also associated with class III alleles in most, but not all samples (e.g. Bennett *et al.* (1995)). However, it is difficult to resolve a class III-associated reduction in transcription with the dominant protective effect of class III alleles. In addition, these studies failed to account for the pulsatile nature of insulin secretion which is central to insulin signalling, the effects of which on type 1 diabetes susceptibility are unknown (McCarthy, 1998).

Analysis of foetal thymus in two studies found a 2- to 3-fold higher transcription level associated with class III alleles. However, in 2/12 (Vafiadis *et al.*, 1997) and 3/10 (Pugliese *et al.*, 1997) I/III heterozygous thymi analysed, *INS* expression from class III-associated alleles was completely suppressed. If susceptibility results from reduced thymic insulin transcription, this result suggests that while most class III alleles protect against type 1 diabetes, some may associate with elevated susceptibility. Insulin expression in the thymus is very low (3-4 orders of magnitude lower than the pancreas (Vafiadis *et al.*, 1997)).



However, variation in thymic antigen expression of as little as 10-15% can have a major effect on T cell selection (Pugliese *et al.*, 1997) so the presence of a class III allele may be sufficient to induce self-tolerance.

### Putative mechanisms for *INS* VNTR enhancer activity

The insulin minisatellite may act by either directly or indirectly affecting the binding or action of transcription factors (TFs). Pur-1 is a ubiquitously expressed Zn-finger TF which has been shown to bind to G1 *cis*-acting elements in the 5' regulatory regions of both human and mice insulin promoters, and has the capacity to induce *INS* expression in cells that do not normally express the gene (reviewed by Bennett and Todd (1996a); Dumonteil and Philippe (1996)). Alternatively, Pur-1-G1 binding may result in transcriptional termination of *TH* gene expression. The close proximity of *TH* to the *INS* promoter could cause transcriptional interference unless *TH* transcription was efficiently terminated (Ashfield *et al.*, 1994; Bennett and Todd, 1996a). The *INS* VNTR contains several high affinity binding sites for Pur-1 so the VNTR may supplement the *cis*-acting effects of the G1 element.

To investigate a role for the VNTR in Pur-1 mediated transcriptional regulation, 11 synthetic oligonucleotides were generated, each composed of 4 identical tandem repeats of the VNTR variants a-k (Kennedy *et al.*, 1995). Significant variation in binding affinity was detected, with the a-type repeat binding most strongly followed by the h-type repeat. Transcriptional activity of each variant was then analysed and found to generally correlate with binding affinity, with two notable exceptions. The k-type repeat stimulated moderate transcription despite a low binding affinity, whilst the h-type repeat showed ~2.5 times greater transcriptional activity than the a-type repeat, again despite a lower binding affinity (Kennedy *et al.*, 1995). If the VNTR results in protection against disease by elevating thymic insulin gene expression, this could in part be mediated by the presence of h-type repeats. However, it is unknown whether Pur-1 binds to the VNTR *in vivo*. In addition, while transcriptional activity was assayed with four identical tandem copies of each repeat, only the a-type repeat has ever been detected in an allele in four tandem copies (Chapter 8).

Alternatively, the VNTR could mediate transcription of *INS* by altering the distance between regulatory elements 5' of the VNTR and the insulin gene, similar to the EPM1 minisatellite (Lalioti *et al.*, 1999). The Ink region (insulin kilobase upstream) contains three putative binding sites for a subfamily of steroid hormone receptor TFs including the retinoic acid receptor (RAR), thyroid hormone receptor (TR) and vitamin D receptor (VDR) (see Clark *et al.* (1995)). Electrophoretic mobility-shift assays and reporter gene assays in cell lines exposed to retinoic acid and expressing the retinoic acid receptor indicated that RAR is able to bind to Ink resulting in suppression of reporter gene expression. In addition, the incubation in retinoic acid of human islets of Langerhans led to increased mRNA expression. It is conceivable that allele length at the VNTR, by affecting the proximity of the Ink sequence to the transcriptional start site, affects *INS* transcription (Clark *et al.*, 1995).

S1 nuclease treatment of supercoiled plasmids containing the *INS* VNTR (Hammond-Kosack *et al.*, 1992), and high resolution nuclear magnetic resonance (NMR) analysis (Catasti *et al.*, 1997; Catasti *et al.*, 1996) identified the *in vitro* formation of unusual DNA conformations at the minisatellite. The G-rich strand forms hairpin G-quartet structures (Catasti *et al.*, 1996), whilst the C-rich strand forms hairpin dimers stabilised by intercalated C<sup>+</sup>-C pairs (Catasti *et al.*, 1997). In both cases, the probability for the formation of each structure increases with repeat copy number. In addition, many sequence deviations from the a-type repeat destabilise the G-quartet structure (Catasti *et al.*, 1996). Formation of these structures would create negative supercoiling at the VNTR which could be compensated for *in vivo* by positive supercoiling within the *INS* promotor (Catasti *et al.*, 1996; Felsenfeld, 1993). This would facilitate denaturation of the duplex and therefore increase transcription. The high level of transcriptional activity associated with the a-type repeat (Kennedy *et al.*, 1995) was interpreted as support for this model (Catasti *et al.*, 1996).

A functional role for tandemly repetitive DNA has been proposed in the context of genetic imprinting (Neumann *et al.*, 1995). Foreign DNA such as retroviral elements, inserted into genomic sequences, are often associated with tandem repeats. It was suggested that methylation of tandemly repeated regions may have been a function of host defence

systems which silence expression of the foreign element (Barlow, 1993). Many if not all imprinted genes are associated with tandem repeats (Constancia *et al.*, 1998). No obvious homology between these repeats other than a propensity for GC-richness has been detected and a so common secondary structure associated with repetitive DNA may act to attract methylation of these regions (Constancia *et al.*, 1998). Putative methylation sites are contained within d, f, m, l, and n repeats (of which f and possibly d are the only common repeats in Caucasians; Chapter 8; Bennett and Todd (1996a)). The VNTR is located in a region known to be affected by imprinting, and parent-of-origin effects associate with *IDDM2* in many studies (Bennett and Todd, 1996a; Bennett *et al.*, 1997). Ghazi *et al.* (1990) analysed methylation states at sites within the VNTR and flanking sequences and detected extensive methylation in sperm with reduced methylation in other tissues. It was suggested that the number of methylation sites within the VNTR correlate with levels of protection (Awata *et al.*, 1997). In the study of the effects of VNTR size on susceptibility to diabetes in a Japanese population, the 1S alleles were significantly more predisposing to disease. Approximately 300 1S and 1L alleles from this study were analysed by restriction digestion and with only 7 exceptions, 1L alleles contained 2 additional methylation sites compared to 1S alleles (Awata *et al.*, 1997). Inclusion of previously published class III alleles supported the observation that the number of d/f repeats increases with allele protection: 1S alleles have 3 d/f repeats, 1L alleles have 5-6 d/f repeats, PH alleles have 8 d/f repeats, and VPH alleles have 17 d/f repeats (Awata *et al.*, 1997).

Finally, a protective role for the e-type repeat has been suggested based on its absence from class I alleles (except the  $\lambda$ H1.1 clone (Bell *et al.*, 1982)) and an elevated frequency in VPH- compared to PH-associated alleles (Bennett *et al.*, 1995). No mechanistic basis for the effect of e-type repeats has been proposed.

Results from *in vitro* studies should be interpreted with caution for two reasons: i) If a result is obtained *in vitro*, it does not mean that the same process operates *in vivo*, and; ii) Many *in vitro* studies use the cloned class I allele  $\lambda$ H1.1 which has an atypical class I structure apparently resulting from a deletion within a class III PH-associated allele. Further elucidation of *INS* VNTR action is unlikely to be facilitated by the use of mouse

models as the VNTR is only found in primates, and mice have two copies of the insulin gene.

### ***The effects of other type 1 diabetes susceptibility loci***

Candidate genes have been identified for many type 1 diabetes susceptibility loci (reviewed by Buzzetti *et al.* (1998); Todd and Farrall (1996)). For example *IDDM4* maps near *CD3* which represents one of the most important membrane-bound complexes involved in T cell activation (Buzzetti *et al.*, 1998). *IDDM5* may be the Mn-superoxide dismutase gene, mutation of which may elevate  $\beta$ -cell susceptibility for FOR damage (Buzzetti *et al.*, 1998). The peak of linkage of *IDDM10* is at the *TCF8* locus which encodes Nil-2-a, a negative regulator of interleukin 2 (IL2) expression (Mein *et al.*, 1998; Schwartz, 1997). The effects of IL2 include its action as a T cell growth factor and a role in the termination of T cell activity (Todd, 1999a). *SEL-1L* is near *IDDM11* and is expressed at its highest levels in the pancreas (Donoviel and Bernstein, 1999). Its homologues in *C. elegans* (SEL-1) and *S. cerevisiae* (Hrd3) have putative protein processing/degradation functions (Grant and Greenwald, 1997; Hampton *et al.*, 1996). It has therefore been suggested that a defective *SEL-1L* gene could result in a failure to correctly process a  $\beta$ -cell antigen which could result in their targeting by the immune system (Donoviel and Bernstein, 1999). *IDDM12* on chromosome 2q33 was identified as a type 1 diabetes susceptibility locus by the functional candidate gene approach (Donner *et al.*, 1997; Marron *et al.*, 1997; Nistico *et al.*, 1996; Van der Auwera *et al.*, 1997). *CD28* and *CTLA-4* both map to this location and are involved in the respective proliferation and inhibition of proliferation of T cells during APC-mediated activation. Disruption of this interaction between *CD28* and *CTLA-4* could result in T cell overproliferation and an elevated autoimmune susceptibility (Marron *et al.*, 1997).

### **Interaction between *IDDM1* and *IDDM2***

It has been proposed that *IDDM1* may elevate susceptibility by low avidity binding of thymic proinsulin, whilst *IDDM2* predisposition results from low expression levels of thymic proinsulin (Pugliese *et al.*, 1997; Vafiadis *et al.*, 1997). A prediction of this model is that a high affinity of HLA binding may compensate for lower *INS* expression levels, whilst

high *INS* expression may compensate for poor HLA binding. Epistasis would therefore be expected between *IDDM1* and *IDDM2*. Whilst epistasis between these loci has been demonstrated using a multiplicative genetic model (Cordell *et al.*, 1995) it remains a controversial issue. For example, one study found that the identity of *IDDM2* had no effect on susceptibility in the high risk genotype *DQA1\*0301-DQB1\*0302/DQA1\*0501-DQB1\*0201*, but in intermediate risk *HLA-DQ* genotypes, class III alleles of the *INS* VNTR were strongly protective (reviewed by Bennett *et al.* (1995)).

## **Type 1 diabetes and other autoimmune diseases**

It is apparent from the model presented for the aetiology of type 1 diabetes that some susceptibility factors predispose to autoimmune disease, whilst others predispose specifically to type 1 diabetes. Type 1 diabetes is just one of many clinically characterised autoimmune diseases, which together affect approximately 5% of the population (Becker *et al.*, 1998). Genome-wide scans have been performed for loci implicated in other autoimmune diseases such as multiple sclerosis, Crohn's disease, familial psoriasis, and asthma (Daniels *et al.*, 1996; Kuokkanen *et al.*, 1996; Satsangi *et al.*, 1996; Tomfohrde *et al.*, 1994). A comparison of 23 genome-wide scans for autoimmune or immune-mediated disease susceptibility loci identified 18 loci to which the majority (~65%) of positive linkages mapped (Becker *et al.*, 1998). It is therefore likely that a common set of genes predispose to autoimmune disease. The progression to a specific disease would result from additional genetic and environmental factors. This is supported by reports of familial clustering of different autoimmune diseases (such as rheumatoid arthritis, coeliac disease, thyroiditis, and multiple sclerosis), and the co-association of multiple autoimmune diseases in the same individual (Bias *et al.*, 1986; Vyse and Todd, 1996).

### **iii) Further investigations of *IDDM2***

While the high frequency of predisposing class I alleles in Caucasian populations could simply reflect the effects of genetic drift on a locus with a low associated relative risk ( $\lambda_s=1.25$ ), it is perhaps surprising that disease-associated alleles of the *INS* VNTR are at such a high frequency. One explanation comes from a recent study analysing transmission

of class I and class III alleles from I/III heterozygous parents to offspring unaffected by *INS* VNTR-associated diseases. Significant evidence for transmission distortion was detected, with class I alleles transmitted to unaffected offspring at a frequency of 54% (95% C.I. limits of 0.51- 0.56) (Eaves *et al.*, 1999). This is the first example of meiotic drive operating in humans. A consequence of this result is that previously reported levels of type 1 diabetes predisposition associated with class I alleles would have been overestimated, whilst levels of protection which associated both with class III, and specific subclasses of class I allele were underestimated. Alternatively, it is possible that the selective pressure on the VNTR due to type 1 diabetes risk is counterbalanced by opposing selective forces.

Polycystic ovary syndrome (PCOS) is a common endocrine disorder affecting up to 10% of women of reproductive age. Women with anovulatory PCOS have hyperinsulinaemia, insulin resistance, dyslipidaemia, and an elevated risk of developing type 2 diabetes. The *INS* VNTR III/III genotype was found to significantly associate with PCOS susceptibility (Waterworth *et al.*, 1997). The *INS* VNTR has also been associated with type 2 diabetes in which class III alleles associate with elevated susceptibility (Bennett and Todd, 1996a; Ong *et al.*, 1999). In addition, the minisatellite has recently been shown to associate with adult obesity (O'Dell *et al.*, 1999). An *ApaI* polymorphism located within *IGF2* is in linkage disequilibrium with a subset of smaller class I alleles and associates with an elevated body mass index (BMI). It was proposed that either the minisatellite, *ApaI*, or a third site may be aetiological (O'Dell *et al.*, 1999).

Birth size is an important determinant of perinatal survival (Alberman, 1991) and has been associated with the *INS* VNTR (Dunger *et al.*, 1998). Analysis of birth size in 758 infants from the ALSPAC cohort (Avon Longitudinal Study of Pregnancy and Childhood) revealed the III/III genotype to associate with significantly larger babies (as defined by head circumference measurements) than both the I/III and I/I genotypes. A subset of infants were identified for whom variation in birth size was due primarily to genetic factors as opposed to maternal uterine factors. In these children, VNTR genotype also associated significantly with both birth length and weight (Dunger *et al.*, 1998).

The first large scale detailed survey of allele diversity at the insulin minisatellite defined by internal structure as opposed to allele length was presented in Chapter 8. These alleles were derived from the parents of type 1 diabetes affected sib pair families, allowing the association of alleles of the insulin minisatellite characterised by structure as opposed to size or haplotype to be re-evaluated. In Chapter 10, the relationship between allele structure at the minisatellite and associated predisposition to type 1 diabetes will be discussed.

## Chapter 10

### Evidence that *IDDM2* has a multi-locus aetiology

#### Summary

MVR-PCR analysis at the insulin minisatellite divided almost all alleles from a Caucasian cohort into four lineages, IC, ID, IIIA, and IIIB (Chapter 8). Haplotype analysis demonstrated that lineages IIIA and IIIB correspond perfectly to the Protective (PH) and Very Protective (VPH) haplotypes respectively. Analysis of an *MspI* polymorphic site (+3580*MspI*<sup>+/−</sup>) located in the first intron of *IGF2* (Lucassen *et al.*, 1993) found the restriction site to be present on all haplotypes with the exception of a subset of larger class ID alleles. A combination of MVR analysis and the presence (+) or absence (−) of +3580*MspI*<sup>+/−</sup> thus divides class I alleles into three newly defined ancestral lineages, IC+, ID+, and ID−. Different alleles at the tyrosine hydroxylase microsatellite (*HUMTH01*) each associate with a different lineage of the insulin minisatellite, confirming the identity of these lineages and demonstrating linkage disequilibrium to extend over at least 10 kb. Transmissions of each allele class to type 1 diabetic offspring were analysed in 219 affected sib pair families. All class I alleles were equally predisposing to type 1 diabetes except for ID− alleles which were protective when transmitted from ID−/III fathers. Similar results had been previously reported for class I alleles of 42 repeats in length (allele 814) (Bennett *et al.*, 1997). Division of ID− alleles into those of 42 repeats and those of other sizes demonstrated that the protective effect associated with all ID− alleles, irrespective of size. ID− alleles are only clearly distinguished from all other alleles by the absence of the +3580*MspI* polymorphic site. All observed heterogeneity in the effects of class I alleles detected in this study may therefore be due to variation in the flanking haplotype, and not a property of the insulin minisatellite. It is proposed that *IDDM2* has a multi-locus aetiological basis. A hypothetical model to account for *IDDM2*-associated susceptibility to type 1 diabetes is presented.



## Introduction

A brief summary of the salient points introduced in Chapter 9 will be presented. The insulin minisatellite is the best known candidate for the type 1 diabetes susceptibility locus *IDDM2* (Bennett and Todd, 1996a). Class I alleles associate with predisposition to disease, while class III alleles are dominantly protective (Bell *et al.*, 1984). Three flanking haplotypes, the class I haplotype, class III Protective Haplotype (PH) and class III Very Protective Haplotype (VPH) (Bennett *et al.*, 1995) were defined across 10 polymorphic sites (9 SNPs and the minisatellite) within a 4.1 kb region of tight linkage disequilibrium, within which *IDDM2* had been localised (Lucassen *et al.*, 1993). Subdivisions within both class I and class III alleles revealed heterogeneity between the effects of different allele subclasses (Awata *et al.*, 1997; Bennett *et al.*, 1995). For example, division of class I alleles by size demonstrated that whilst most class I alleles associate with predisposition to type 1 diabetes, allele 814 (42 repeats) is protective when transmitted from class I/III heterozygous fathers (Bennett *et al.*, 1995; Bennett *et al.*, 1997). The minisatellite was the only single polymorphic site within the 4.1 kb region of susceptibility to which the observed heterogeneity could be attributed (Bennett *et al.*, 1995).

The insulin minisatellite may act as a transcriptional regulator of the insulin gene and *IGF2* (Paquette *et al.*, 1998). Regulation could be mediated by variant repeat composition and distribution which could alter DNA conformation (Catasti *et al.*, 1997; Catasti *et al.*, 1996), levels of methylation within the locus (Awata *et al.*, 1997), or the binding and activity of transcription factors (Kennedy *et al.*, 1995). However, previous studies which identified the insulin minisatellite as *IDDM2* classified alleles by either size or flanking haplotype, and not by variant repeat distribution (Bennett and Todd, 1996a).

In Chapter 8, patterns of allele diversity defined by MVR-PCR at the insulin minisatellite were described and four main allele lineages (IC, ID, IIIA and IIIB) were identified. The 876 alleles analysed in this study represented all alleles from within 219 families of type 1 diabetes affected sib pair families. Inheritance patterns of alleles within families could in most cases be unequivocally determined simply from analysing allele length. Where parents were homozygous for allele length but heterozygous for MVR code, inheritance patterns were determined by MVR-PCR analysis of both parents and offspring. Patterns of

inheritance could therefore be established within each family and are presented in Appendix 9. In this chapter, a combination of flanking haplotype and allele lineage defined by MVR-PCR are used to re-investigate associations of the minisatellite with type 1 diabetes.

## Results

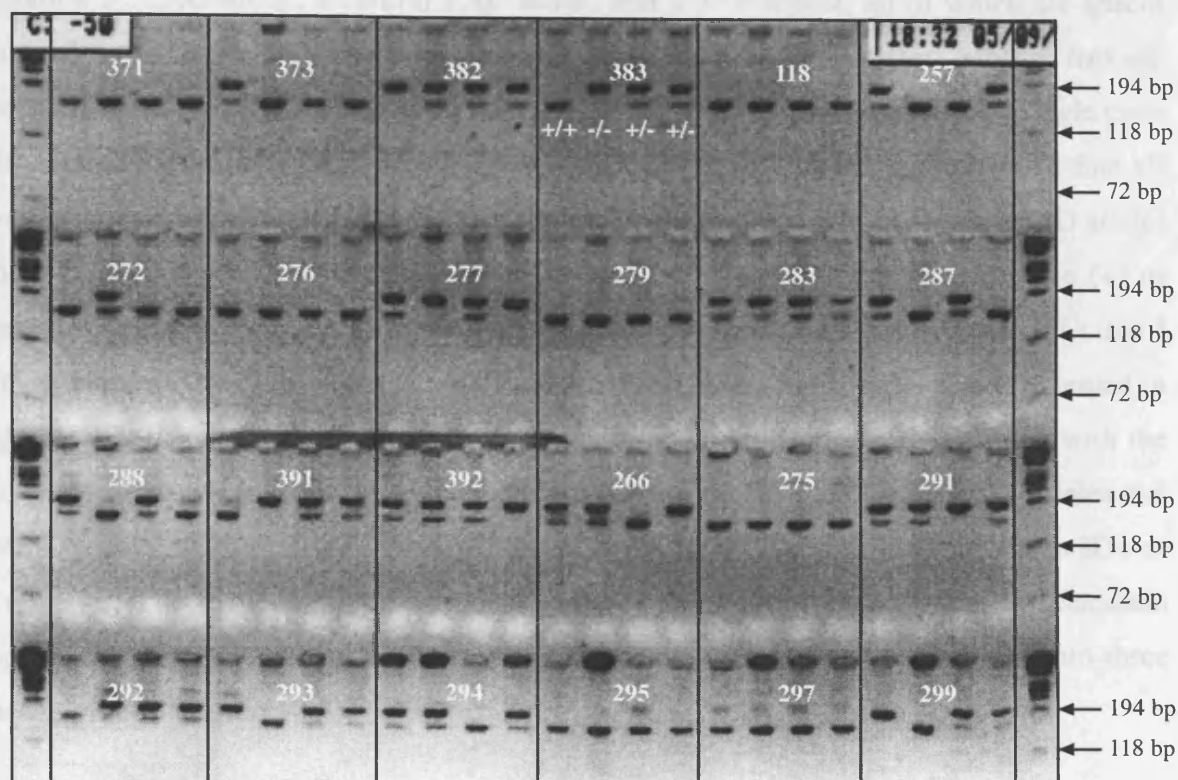
### Analysis of flanking haplotypes

Polymorphisms at sites -2733A/C, -2221*Msp*I, -23*Hph*I, +805*Dra*III, +1127*Pst*I, +1428*Fok*I and the tyrosine hydroxylase (*TH*) microsatellite (*HUMTH01*) 9 kb upstream of the insulin gene (Figure 9.3) had been previously typed in 128 of the 219 families analysed in this study (Bennett *et al.*, 1995). Genotype data for each polymorphism were kindly supplied by Prof. John Todd. The -23*Hph*I polymorphism was typed in all families as described in Chapter 8. Integration of minisatellite variant repeat codes with genotypic data allowed the conversion of genotypes to haplotypes using patterns of allele transmissions within these families. An additional *Msp*I polymorphic site located 3580 bp downstream of the insulin gene translation initiation site (+3580*Msp*I<sup>+/−</sup>) and within intron 1 of *IGF2* (Lucassen *et al.*, 1993) had also been typed in ~25% of families. For reasons described later, this site was re-typed in all 219 families (Figure 10.1).

### Class III allele haplotypes

Variant repeat analysis of 171 class III alleles defined two distinct lineages, class IIIA (126 alleles) and class IIIB (43 alleles) (Appendix 7; <http://www.le.ac.uk/genetics/ajj/insulin>). Lineages IIIA and IIIB defined by MVR analysis correspond exactly to the PH and VPH respectively. With only one exception, all 169 IIIA and IIIB haplotypes were +3580*Msp*I<sup>+</sup>. Two rare alleles of class III size had been characterised. The R213.1 allele haplotype was identical to most IIIA (VPH) alleles both 5' and 3' of the locus as described in Chapter 8. MVR typing of allele R188.1 demonstrated that it belonged to a different lineage from any other allele analysed. No haplotype data were available for most flanking sites, but analysis of +3580*Msp*I demonstrated that in contrast to the other class III lineages, the *Msp*I site was absent from the haplotype.

**Figure 10.1**



### ***Analysis of the +3580MspI<sup>+</sup> polymorphism***

Analysis of the +3580MspI<sup>+</sup> polymorphic site in four individuals from each of 24 families of affected sib pairs is presented. Samples were loaded from left to right in the order father, then mother, then the two children. Numbers correspond to family reference codes, a full list of which is presented in Appendix 9. To type the +3580MspI polymorphic site, a 188 bp region containing the polymorphic site was amplified using primers 3580-A and 3580-B (Table 2.1) in 10 µl PCR reactions with 10xPCR buffer (Advanced Biotechnologies) diluted to 75 mM Tris-HCl pH 8.8, 20 mM (NH<sub>4</sub>)<sub>2</sub>SO<sub>4</sub>, 0.1% (v/v) Tween 20, and supplemented with 2 mM MgCl<sub>2</sub>, 0.2 mM of each dNTP, 0.4 µM of each primer, and 0.05 U/µl *Taq* polymerase. 10 ng of genomic DNA was amplified at 96°C 40 sec, 58°C 30 sec, 70 °C 1 min for 31 cycles and digested with 2 units of *MspI* in REact 1<sup>TM</sup> buffer for 2 hours at 37°C. 10xPCR buffer was selected as, in contrast to the standard 11.1xPCR buffer, its pH and salt concentrations are similar to REact 1<sup>TM</sup> digestion buffer. PCR samples could therefore be digested with no prior precipitation, without substantially compromising digestion efficiency or specificity. Samples were electrophoresed for 2 hours at 5 V/cm through 4% Metaphor® agarose gels (FMC Bioproducts). *MspI* digestion cleaves the 188 bp amplicon into fragments of 35 bp and 153 bp. In family 383, the father is a +/+ homozygote, the mother a -/- homozygote, and both offspring clearly heterozygous as indicated.

### ***Class I allele haplotypes***

Class I alleles divide by MVR code into two lineages, IC and ID. Typically, ID alleles contain a 5' CAC motif, a central FAC motif, and a 3' F repeat, all of which are absent from class IC alleles (Appendix 7). Previous studies identified strong linkage disequilibrium between polymorphic sites -2733A/C and +1428*FokI* defining a single class I associated haplotype. Analysis of the +3580*MspI* polymorphism demonstrated that all class I alleles were also +3580*MspI*<sup>+</sup> with the exception of a subset of the larger ID alleles which lacked the restriction site. A combination of MVR code and the presence (+) or absence (-) of the +3580*MspI* site therefore defines three ancestral lineages of class I alleles, classes IC+, ID+, and ID-. MVR codes of all ID+ and ID- alleles are presented in Figure 10.2. ID+ and ID- alleles show very similar variant repeat organisation, with the exception of an ABA motif present 10 repeats upstream of the end of most ID- alleles and absent from most ID+ alleles. As a result, most ID- alleles are slightly larger than ID+ or IC+ alleles. The trimodal allele size distribution observed for class I alleles in Caucasian populations (Bennett *et al.*, 1995) clearly reflects this new division of alleles into three lineage groups (Figure 10.3).

### ***Allele distribution at HUMTH01***

To determine whether the five classes of minisatellite defined by MVR and haplotype (IC+, ID+, ID-, IIIA and IIIB) correspond to true allele lineages, linkage disequilibrium was analysed between the minisatellite and the tyrosine hydroxylase microsatellite (*HUMTH01*) 9 kb upstream of the insulin gene. *HUMTH01* has 5 common alleles composed of 6 (allele Z-16) to 10 (allele Z) tandem repeats of a TCAT tetramer (O'Malley and Rotwein, 1988; Polymeropoulos *et al.*, 1991). Haplotypes of the five insulin minisatellite subclasses were each dominated by a different allele of *HUMTH01* (Figure 10.4), showing that each subclass represents a distinct lineage of closely related alleles that have retained linkage disequilibrium over a region of over 10 kb surrounding the minisatellite. The greatest breakdown of linkage disequilibrium between the minisatellite and *HUMTH01* was within the ID- lineage, where ID- haplotypes were associated with both allele Z (33% of haplotypes) and Z-16 (59% of haplotypes) at *HUMTH01*. 89% of all Z/ID- haplotypes share identical alleles at the insulin minisatellite; allele ID42.4 (Figure 10.2). Surprisingly, ID42.4 alleles are linked to both Z-16 and

## Figure 10.2

### ***+3580MspI<sup>+/+</sup> divides class ID into two groups***

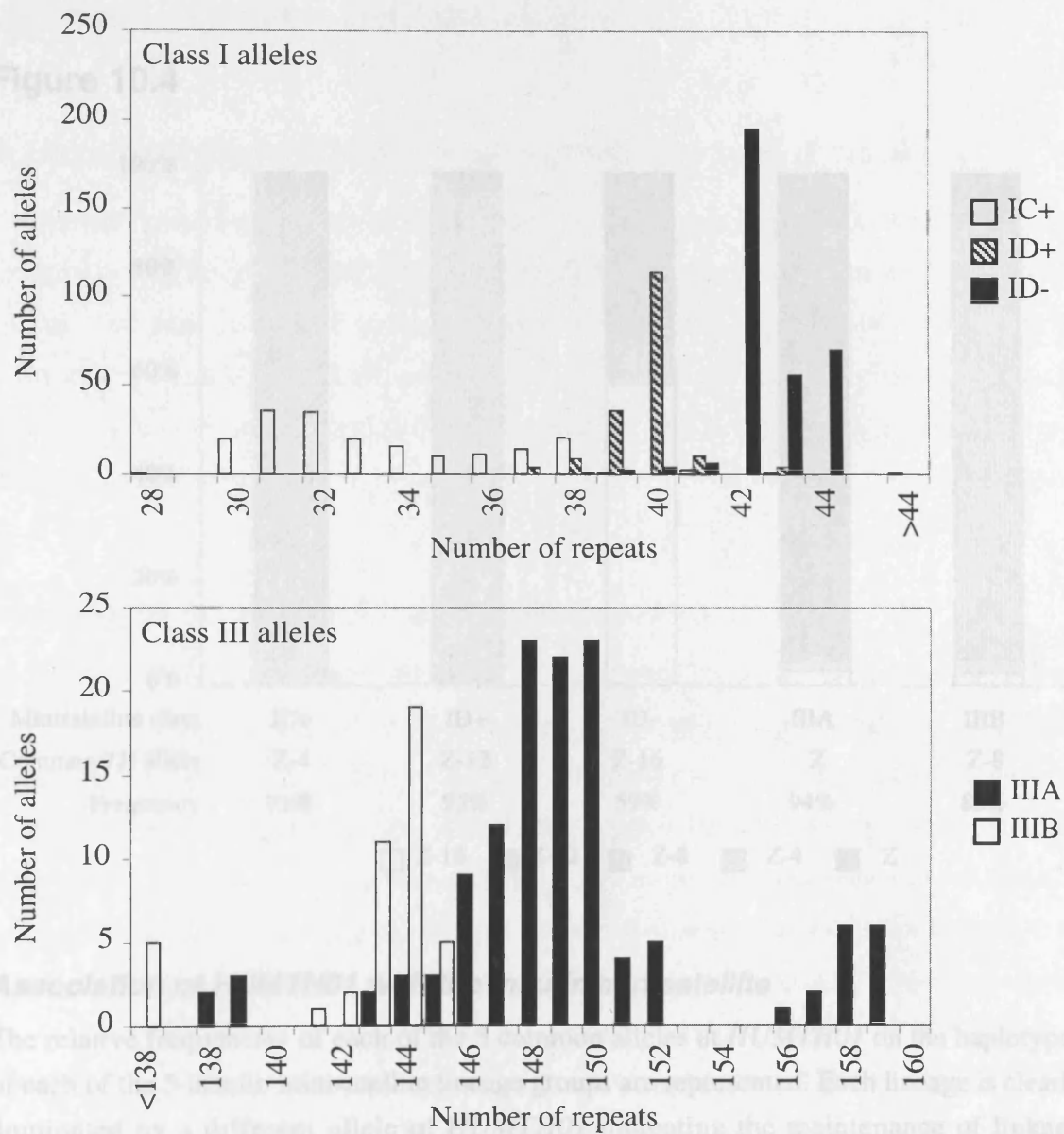
MVR typing of class I alleles divided alleles into two groups, IC and ID (Figure 8.7). Analysis of the +3580MspI<sup>+/+</sup> polymorphism further divides class ID alleles into ID+ and ID-. All ID+ and ID- alleles are presented, with a selection of IC+ alleles for structural comparison. The number of copies of each ID+/- allele is presented. In 11 of the 219 families in this study, each family member was heterozygous at the MspI site (Appendix 9). Linkage of the MspI variant to each allele of the VNTR was assumed for these families, based on patterns of MspI-VNTR linkage disequilibrium observed at all other alleles. Haplotypes of two class ID alleles of the insulin minisatellite (ID40.3 and ID43.9) were found to be associated with both the presence and absence of the +3580MspI site and are indicated by \*.

**Figure 10.2**

Allele	Copies	MVR code
<b>IC+</b>		
IC28.1		CBoBoFAFAAAAAA-----BABABC--AAABBB
IC31.1		CBoBoFAFAAAAAA-----BABABC--AAABBB
IC31.2		CBoBoFAFAAAAAA-----BAAA-BABABC-AAAABBB
IC33.1		CBoBoFAFAAAAAA-----BAAAABABABC-AAAABBB
IC34.1		CBoBoFAFAAAAAA-----BAAAABABABC-AAAABBB
IC30.6		CBoBoFAFAAAAA-----CAAAAA-----BABABC--AAABBB
IC32.6		CBoBoFAFAAAAA-----CAAAAA-----BABABC--AAABBB
IC32.7		CBoBoFAFAAAAA-----CAAAAA-----BABABCC-AAAB-B
IC35.4		CBoBoFAFAAAAA-----CAAAAA--BAAAABABABC-AAAABBB
IC37.1		CBoBoFAFAAAAA-----CAAAAA--BAAAABABABCC-AAABBB
IC38.3		CBoBoFAFAAAAACAAA-CAAAA--BAAAABABABC---AABBB
IC41.2		CBoBoFAFAAAAACAAAA-----BAAAABABABC-AAAAB-B
<b>ID+</b>		
ID37.1	4	CBoBoFAFAAAC--AAA---FACAAAABAAA---BABCAAAFB
ID38.1	1	CBoBoFAFAAACACAAAA--FACAAA-BAAA---BABCAAA-B
ID38.2	2	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAA--B
ID38.3	6	CBoBoFAFAAAC--AAAA---FACAAAABAAA---BABCAAAFB
ID39.1	6	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAA-B
ID39.2	1	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAAFB
ID39.4	21	CBoBoFAFAAACACAAAA--FACAAA-BAAA---BABCAAAFB
ID39.6	8	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAAF--B
ID40.2	112	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAAFB
ID40.3*	1	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAAF--B
ID40.4	1	CBoBoFAFAAACACAAAA--FA-AAAAABAAA---BABCAAAFB
ID41.5	10	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAAFB
ID43.9*	4	CBoBoFAFAAACACAAAA--FACAAAABAAAABABABCAAAF--B
<b>ID-</b>		
ID38.4	1	CBoBoF--AAACACAAAA--FACAAAABAAA---BABCAAAF--B
ID39.3	1	CBoBoFAFAAACACAAAA--FACAAA-BAA---BABABCAAAF--B
ID39.5	1	CBoBoFAFAAACACAAAA--FACAAA-BAAA---BABCAAAF--B
ID40.1	2	CBoBoF--AAACACAAAA--FACAAAABAAAABABABCAAAF--B
ID40.3*	2	CBoBoFAFAAACACAAAA--FACAAAABAAA---BABCAAAF--B
ID41.1	1	CBoBoFAFAAAC--AAAA---FAFAAAAABAAAABABABCAAAFB
ID41.2	2	CBoBoFAFAAACACAAAA--FACAAA-BAAAABABABCAAAF--B
ID41.3	2	CBoBoFAFAAACACAAAA--FACAAAABAAA-BABABCAAAF--B
ID41.4	1	CBoBoFAFAAACACAAAA--FACAAAABAAAABABABCAA-F--B
ID42.1	2	CBoBoFAFAAAC--AAAA---FACAAAABAAAABABABCAAAFB
ID42.2	6	CBoBoFAFAAAC-CAAAA--FACAAAABAAAABABABCAAAFB
ID42.3	1	CBoBoFAFAAACACAAAA--FACAAA-BAAAABABABCAAAFB
ID42.4	184	CBoBoFAFAAACACAAAA--FACAAAABAAAABABABCAAAF--B
ID42.5	2	CBoBoFAFAAACACAAAA--FACAAA-BAAAABABABCAAAF--B
ID43.1	1	CBoBoFAFAA-CACAAAA--FACAAAABAAAABABABCAAAFB
ID43.2	2	CBoBoFAFAAAC--AAAA---FACAAAABAAAABABABCAAAFB
ID43.3	5	CBoBoFAFAAAC-CAAAA--FACAAAABAAAABABABCAAAFB
ID43.4	1	CBoBoFAFAAAC-CAAAA--FAFAAAAABAAAABABABCAAAFB
ID43.5	10	CBoBoFAFAAACACAAAA--FACAAAABAAAABABABCAAAFB
ID43.6	2	CBoBoFAFAAACACAAAA--FACAAAABAAA-BABABCAAAFB
ID43.7	1	CBoBoFAFAAACACAAAA--FACAAAABAAAABABABCAA-FBB
ID43.8	1	CBoBoFAFAAACACAAAA--FACAAAABAAAABABAB-AAAFBB
ID43.9*	33	CBoBoFAFAAACACAAAA--FACAAAABAAAABABABCAAAF--B
ID44.1	69	CBoBoFAFAAACACAAAA--FACAAAABAAAABABABCAAAFB
ID44.2	1	CBoBoFAFAAACCCAAAA--FAAAAAABAAAABABABCAAAFB



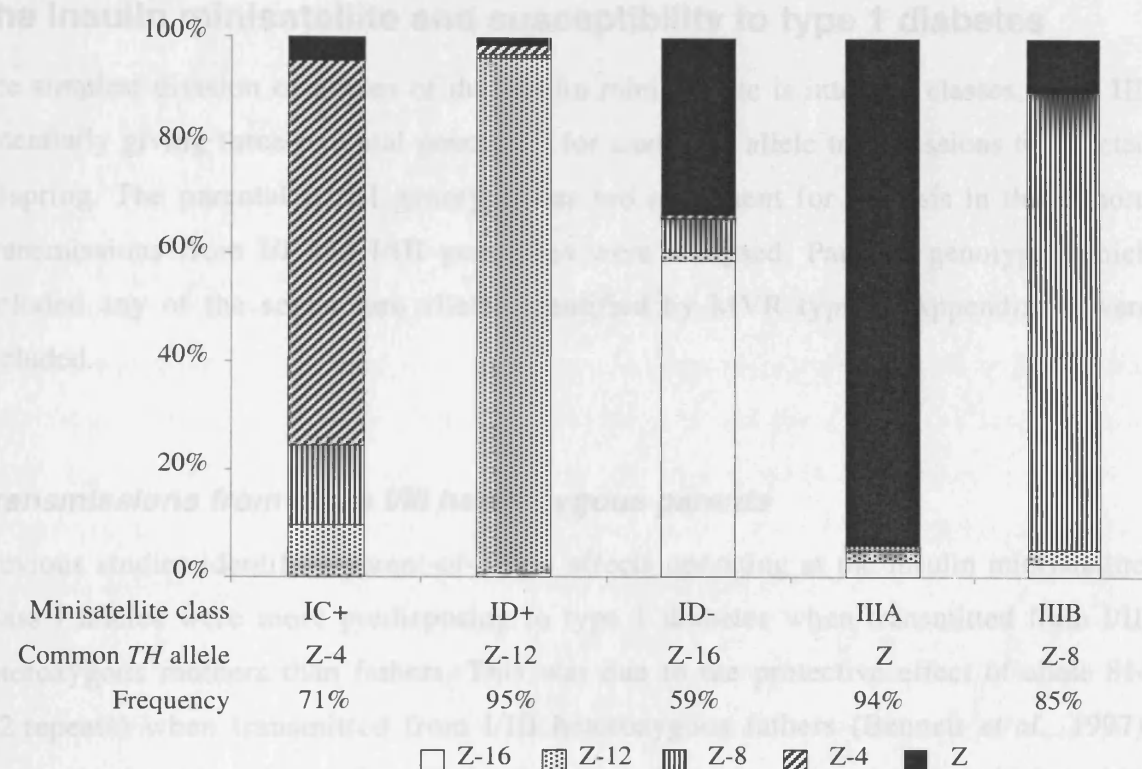
**Figure 10.3**



***Size distribution of allele subgroups***

Eight alleles lay outside of the size ranges shown, of which 6 had lengths of class II alleles and were structurally related to either IC+ or IIIB alleles (Chapter 8).

**Figure 10.4**



#### **Association of *HUMTH01* with the insulin minisatellite**

The relative frequencies of each of the 5 common alleles at *HUMTH01* on the haplotypes of each of the 5 insulin minisatellite lineage groups are represented. Each lineage is clearly dominated by a different allele of *HUMTH01* indicating the maintenance of linkage disequilibrium across at least 10 kb surrounding the minisatellite.



Z alleles of *HUMTH01* (40% and 56% of ID42.4 alleles respectively), indicating either a recent mutation at the microsatellite or (more plausibly given the stepwise mode of microsatellite mutation) a recent recombination event between *HUMTH01* and the insulin minisatellite, most likely between a Z-16/ID42.4 haplotype and a Z/IIIA haplotype. The apparent breakdown of linkage disequilibrium between the ID- haplotype and *HUMTH01* can therefore be attributed to a single recombination event.

## **The insulin minisatellite and susceptibility to type 1 diabetes**

The simplest division of alleles of the insulin minisatellite is into two classes, I and III, potentially giving three parental genotypes for analysing allele transmissions to affected offspring. The parental III/III genotype was too infrequent for analysis in this cohort. Transmissions from I/I and I/III genotypes were analysed. Parental genotypes which included any of the seven rare alleles identified by MVR typing (Appendix 7) were excluded.

### ***Transmissions from class I/III heterozygous parents***

Previous studies identified parent-of-origin effects operating at the insulin minisatellite. Class I alleles were more predisposing to type 1 diabetes when transmitted from I/III heterozygous mothers than fathers. This was due to the protective effect of allele 814 (42 repeats) when transmitted from I/III heterozygous fathers (Bennett *et al.*, 1997). Consequently, transmission frequencies in this study were analysed separately from I/III heterozygous mothers and fathers. Analysis using the transmission disequilibrium test (TDT) (McGinnis and Spielman, 1995) similarly found that while class I alleles associated with predisposition to disease when transmitted from I/III mothers ( $\chi^2=19.7$ , 1 df,  $p<0.00001$ ), they were not significantly predisposing when transmitted from fathers ( $\chi^2=2.35$ , 1 df,  $p=0.13$ ) (Table 10.1). The difference between maternal and paternal transmissions was significant (G test for heterogeneity,  $\chi^2=4.17$ , 1 df,  $p<0.05$ ). This result was fully expected as there is considerable overlap between the families analysed in this study and those used by Bennett *et al.* (1997).

**Table 10.1*****Transmissions from I/III heterozygous parents***

Test		Maternal	Paternal	Heterogeneity
TDT	t(I)	101	78	$\chi^2=4.17$ $p<0.05$
	t(III)	47	60	
	$\chi^2$	19.7	2.35	
	p	<0.00001†		
IBD	Shared	51	45	$\chi^2=0.22$
	Not shared	23	24	
	$\chi^2$	10.6	6.39	
	p	<0.002	<0.02	

TDT analyses the association of specific allele classes with predisposition to disease. Transmission (t) of class I and class III alleles to type 1 diabetes affected offspring was compared against an expected ratio of 1:1 using a standard  $\chi^2$  test. Class I alleles transmitted from I/III heterozygous mothers are significantly predisposing to type 1 diabetes ( $p<0.0001$ ), and significantly more predisposing than class I alleles transmitted from I/III heterozygous fathers (G test for heterogeneity,  $p<0.05$ ).

IBD analyses linkage of the minisatellite locus with predisposition to disease. The numbers of affected sibs which shared the same parental allele and the numbers of affected sibs which did not share the same parental allele was tested against an expected ratio of 1:1 using a standard  $\chi^2$  test. The minisatellite was significantly linked with disease from both maternal and paternal transmissions. All levels of significance below  $p<0.05$  are presented.

Transmission ratio distortion (TRD) has been reported at the insulin-linked region with class I alleles transmitted to unaffected offspring from I/III heterozygous parents at a frequency of 54% ( $TRD=0.54$ , 95% confidence limits= $0.51<TRD<0.56$ ). To correct for the effects of TRD, TDT results were reanalysed against an expected frequency for class I allele transmissions of 54%. IBD was reanalysed against an expected frequency of shared alleles of 50.3% ( $0.54^2+0.46^2$ ). The only significance level notably affected by this correction is indicated as †, where  $p<0.001$ .

In addition to analysing allele associations with disease by TDT, the same transmissions were analysed for linkage of the locus to type 1 diabetes by determining the frequency at which affected sib pairs (ASPs) inherit the same alleles identical by descent (IBD) from a given parent. Surprisingly, significant linkage of the minisatellite with disease susceptibility was identified for both maternal and paternal transmissions from I/III heterozygotes (Table 10.1), despite no significant association of either paternally inherited class I or class III alleles with disease. The difference between linkage and association is clarified in Figure 10.5. Significant linkage without association of the minisatellite with type 1 diabetes has two possible interpretations. The first is that (at least for transmissions from I/III heterozygous fathers) the insulin minisatellite is not the aetiological locus but instead is linked to, but not in linkage disequilibrium with, *IDDM2* (Figure 10.5). It was for this reason that analysis was extended to incorporate the +3580*MspI*<sup>+/−</sup> polymorphism. The second interpretation is that the insulin minisatellite may be *IDDM2* but the division of alleles into classes I and III does not reflect the differences between the aetiological effects of different alleles, so that some class I alleles protect against type 1 diabetes whilst some class III alleles are predisposing. It can be concluded that either the identity of *IDDM2*, or the identity of predisposing alleles at *IDDM2*, differs between transmissions from I/III heterozygous mothers and fathers. To clarify the nature of this difference between maternal and paternal transmissions, the same transmissions were re-analysed following subdivision of either class I or class III alleles into the subgroups defined by a combination of minisatellite allele structure and flanking haplotype.

### Subdivision of class III alleles

Variant repeat analysis divides class III alleles into two lineages, IIIA and IIIB, which are identical to the protective haplotype (PH) and very protective haplotype (VPH) respectively, with the latter being more protective against type 1 diabetes (Bennett *et al.*, 1995). In this study, division of class III alleles into IIIA and IIIB identified no heterogeneity between transmission from I/IIIA and I/IIIB parental genotypes (Table 10.2). Since there is significant overlap between the families used in this study and those by Bennett *et al.* (1995), this failure to detect differences between IIIA and IIIB alleles was presumably due to our smaller sample size. Pairwise tests for heterogeneity between maternal and paternal genotypes identified significant heterogeneity only for transmission

## Figure 10.5

### ***A comparison of linkage and association analysis***

Two scenarios are presented. In both cases, at locus 1 allele A is the aetiological variant which predisposes to disease. Loci 1 and 2 are physically linked. Hypothetical transmissions of the chromosomal region containing loci 1 and 2 from a sample of heterozygous parents to affected sib pairs (ASPs) are considered, and the number of transmissions (t) of each allele at locus 2 to affected offspring are presented on the right. Linkage and association of locus 2 with disease is analysed. In scenario **a**, alleles at locus 1 are in complete linkage disequilibrium with alleles at locus 2. In scenario **b**, alleles at locus 1 are in free association with alleles at locus 2. Linkage analysis by IBD (identical by descent) analyses co-transmission of alleles to affected sib pairs. The identity of the allele is not considered. Association analysis by the transmission disequilibrium test (TDT) analyses transmission of specific alleles to affected offspring.

For scenario **a**, allele B at locus 2 is transmitted to both affected offspring from every parent. It is therefore shared by 100% of ASPs so locus 2 is linked to disease. Analysis of the same transmissions for association with disease by TDT compares the number of transmissions of allele B to affected offspring with allele b. Allele B is transmitted to every offspring so allele B is associated with predisposition to disease.

In scenario **b**, testing locus 2 for linkage by IBD again finds that both affected offspring in each ASP family inherit the same allele at locus 2 from every parent. This is due to physical linkage to locus 1 at which allele A predisposes to disease. Locus 2 is therefore linked with predisposition to disease. However, analysis of association by TDT reveals that only half of the offspring inherit allele B, whilst half inherit allele b as loci 1 and 2 are in free association. Neither B or b therefore associates with predisposition to disease.

Linkage is therefore a property of a locus. Significant linkage can be obtained by physical linkage of the marker being analysed to the aetiological site. In contrast, association is a property of specific alleles. Significant association can be obtained only if a marker is both linked to, and in linkage disequilibrium with, the predisposing allele at the aetiological site.

Table 10.2

Figure 10.5

a Allele A predisposes to disease	b Allele A predisposes to disease
Allele B is physically linked to and in linkage disequilibrium with allele A	Allele B is physically linked to but in free association with allele A
Allele B is analysed for linkage (by IBD) and association (by TDT) to disease	Allele B is analysed for linkage (by IBD) and association (by TDT) to disease
Parental Genotype	Parental Genotype
<div> Locus 1                      Locus 2 </div> <div> </div> <div> t=2                      t=0 </div>	<div> Locus 1                      Locus 2 </div> <div> </div> <div> t=2                      t=0 </div>
<div> </div> <div> t=2                      t=0 </div>	<div> </div> <div> t=2                      t=0 </div>
<div> </div> <div> t=2                      t=0 </div>	<div> </div> <div> t=2                      t=0 </div>
<div> </div> <div> t=2                      t=0 </div>	<div> </div> <div> t=2                      t=0 </div>
<div> </div> <div> t=2                      t=0 </div>	<div> </div> <div> t=2                      t=0 </div>
<div> </div> <div> t=2                      t=0 </div>	<div> </div> <div> t=2                      t=0 </div>
<p><b>IBD (Linkage):</b></p> <p>Shared    100%</p> <p>Not shared   0%</p> <p>Locus B shows linkage to disease</p> <p><b>TDT (Association):</b></p> <p>t(B)        100%</p> <p>t(b)        0%</p> <p>Locus B shows association with disease</p>	<p><b>IBD (Linkage):</b></p> <p>Shared    100%</p> <p>Not shared   0</p> <p>Locus B shows linkage to disease</p> <p><b>TDT (Association):</b></p> <p>t(B)        50%</p> <p>t(b)        50%</p> <p>Locus B shows no association with disease</p>

**Table 10.2*****Class III alleles are equally protective against type 1 diabetes***

Genotype		Maternal	Paternal
I/IIIA	t(I)	81	52
	t(IIIA)	35	44
I/IIIB	t(I)	20	26
	t(IIIB)	12	16
Heterogeneity		$\chi^2=0.6$ p=0.4	$\chi^2=0.71$ p=0.4

Subdivision of class III alleles by MVR code and flanking haplotype into lineages IIIA (PH) and IIIB (VPH) identified no difference between transmission frequencies of the two allele classes upon transmission from I/III heterozygous parents to affected offspring (G test for heterogeneity).

Transmissions from each genotype to affected offspring were tested by the TDT against an expected 1:1 ratio. However, significance levels obtained by TDT are highly dependant on sample size. Differences between significance levels obtained from analysis of each genotype were therefore found to be misleading so are not presented.

from I/IIIA parents (data not shown) which may indicate that the very protective effect of IIIB (VPH) alleles is specific to transmissions from I/IIIB heterozygous fathers. However, this is no more than a trend and would require further investigation using a much larger sample size. Due to the absence of any detectable heterogeneity between class III allele subgroups, in following analyses class III alleles will be considered as a single homogeneous group.

### Subdivision of class I alleles

Class I alleles divide by a combination of variant repeat distribution and haplotype into three newly defined ancestral lineages, IC+, ID+, and ID- (Figure 10.2). Transmission frequencies of each of the three lineages were analysed from I/III heterozygous parents (Table 10.3). There was significant heterogeneity between transmissions from all I/III heterozygous genotypes (G test for heterogeneity,  $\chi^2=11.48$ , 5 df,  $p<0.05$ ). This heterogeneity was due specifically to transmissions from the paternal ID-/III genotype where ID- alleles appear to be at least as protective against type 1 diabetes as class III alleles. Class I alleles were equally predisposing from all other I/III heterozygous parents (G test for heterogeneity,  $\chi^2=1.00$ , 4 df,  $p=0.9$ ) with class I alleles transmitted to affected offspring at a frequency of 67%. The anomalous behaviour of ID- alleles transmitted from ID-/III fathers was highly significant when tested against an expected transmission frequency of 67% ( $\chi^2=14.1$ , 1 df,  $p<0.0002$ ).

If ID- alleles are more protective against type 1 diabetes than class III alleles inherited from ID-/III heterozygous fathers, this could account for the significant linkage of the insulin minisatellite to disease but the lack of significant association of class I alleles with diabetes susceptibility (Table 10.1). The minisatellite would therefore be expected to show evidence of linkage to disease from every parental genotype following division of class I alleles into three lineages. Linkage was therefore analysed by IBD for each I/III heterozygous parent (Table 10.4). Consistent with this hypothesis, no evidence of heterogeneity in linkage of the minisatellite to type 1 diabetes was observed between all six parental genotypes (G test for heterogeneity,  $\chi^2=6.75$ , 5 df,  $p=0.2$ ). Furthermore, pairwise tests for heterogeneity between transmissions from maternal and paternal ID-/III genotypes found no evidence for a difference in the levels of linkage of the locus with disease (G test for heterogeneity,

**Table 10.3*****ID- alleles from ID-/III fathers are protective against type 1 diabetes***

Genotype		Maternal	Paternal
IC+/III	t(IC+)	23	28
	t(III)	9	16
ID+/III	t(ID+)	28	29
	t(III)	16	15
ID-/III	t(ID-)	50	21
	t(III)	22	29

The numbers of class I and class III alleles transmitted from I/III heterozygous parents are presented. There is significant heterogeneity between transmission frequencies from the six parental genotypes (G test for heterogeneity,  $\chi^2=11.48$ , 5 df,  $p<0.05$ ), due exclusively to transmissions from the paternal ID-/III genotype (G test for heterogeneity excluding the paternal ID-/III genotype,  $\chi^2=1.00$ , 4 df,  $p=0.9$ ) from which class I alleles are transmitted at a frequency  $<50\%$ . Significance levels obtained by TDT analysis are not presented for the reasons described in Table 10.2.



**Table 10.4**

***Levels of linkage of the insulin minisatellite to type 1 diabetes transmitted from I/III parents are indistinguishable***

Genotype		Maternal	Paternal
IC+/III	Shared	13	18
	Not shared	3	4
ID+/III	Shared	14	11
	Not shared	8	11
ID-/III	Shared	24	16
	Not shared	12	9

The numbers of affected sib pairs (ASPs) inheriting the same allele (shared) or different alleles (not shared) from a given parent were determined for each I/III heterozygous parental genotype. In contrast to association analysis (Table 10.4), no heterogeneity in the levels of linkage of the minisatellite transmitted from the six genotypes was identified (G test for heterogeneity,  $\chi^2=6.75$ , 5 df,  $p=0.2$ ).

$\chi^2=0.05$ , 1 df,  $p=0.8$ ) despite substantial heterogeneity in the associations of the ID- allele with disease (G test for heterogeneity,  $\chi^2=9.05$ , 1 df,  $p<0.005$ ). (It should be noted that heterogeneity tests are not directly comparable between TDT and IBD data as there is a 2-fold difference in sample size which will affect significance levels.)

With the present sample size, it cannot be concluded that ID- alleles transmitted from ID-/III fathers are more protective against type 1 diabetes than class III alleles, as TDT analysis of this genotype found no significant transmission deviation from 50% ( $\chi^2=1.28$ , 1 df,  $p=0.26$ , or correcting for the effects of transmission ratio distortion ( $\chi^2=2.9$ , 1 df,  $p=0.09$ ; (Table 10.1); (Eaves *et al.*, 1999)). However, the data are consistent with ID- protection (or class III predisposition) and can account for the lack of association of paternally transmitted class I alleles despite significant linkage.

#### Further analysis of the ID-/III paternal genotype

Previous studies have demonstrated that protection against type 1 diabetes associates with class I alleles of 42 repeats in length (allele 814) when transmitted from 814/III heterozygous fathers (Bennett *et al.*, 1995; Bennett *et al.*, 1997; Urrutia *et al.*, 1998), similar to that seen for the ID- allele lineage. In this study, 195/196 alleles of 42 repeats in length were within the ID- lineage. The single exception was allele R42.1 which bore no structural similarity to any other allele (and was associated with the +3580MspI<sup>+</sup> variant). The protective effect associated with the ID- lineage may therefore be due specifically to the effects of allele 814. All ID- alleles were therefore divided into those of 42 repeats (814) and those of other sizes (non-814) and transmissions analysed from all ID-/III fathers (Table 10.5). Nineteen fathers with 814/III genotypes were identified in this study (resulting in a possible 38 transmissions of each allele). Of the 19 alleles of 42 repeats in ID-/III paternal genotypes, 18 had identical internal structures (ID42.4) with the single exception (allele ID42.2) transmitted once from the heterozygous father.

No heterogeneity between the transmission frequencies from 814/III and non-814/III fathers was observed (G test for heterogeneity,  $\chi^2=0.47$ , 1 df,  $p=0.5$ ). Furthermore, class I alleles from both genotypes were transmitted at frequencies <50%, significantly different from the 67% transmission frequency seen for class I alleles from other I/III heterozygous

**Table 10.5*****ID- alleles are protective against type 1 diabetes irrespective of size***

Genotype		Observed	Expected
814/III	t(814)	17	25.4
	t(III)	21	12.6
			$\chi^2=8.47$ $p<0.005$
non-814/III	t(non-814)	4	8.0
	t(III)	8	4.0
			$\chi^2=6.13$ $p<0.02$

Class ID- alleles were divided into alleles of 42 repeats (814) and alleles of other sizes (non-814) to investigate whether the protective effects associated with class ID- alleles transmitted from ID-/III fathers were specific to alleles of a certain size (Bennett *et al.*, 1997), or a general property of the ID- class. Class I alleles from all other I/III parental genotypes were equally predisposing to type 1 diabetes and were transmitted to diabetic offspring at a frequency of 67% (Table 10.3). Class I allele transmissions from 814/III and non-814/III fathers were tested for deviation from an expected transmission frequency of 67% using a standard  $\chi^2$  test.

parents ( $p < 0.005$  and  $p < 0.02$  for 814 and non-814 alleles respectively). This demonstrates that the protective effect previously associated with allele 814 is not a property of a specific size class of alleles, but instead of the newly defined class I allele lineage, class ID-. Finally, the protective effects of ID- alleles were the same when transmitted from ID-/IIIA and ID-/IIIB fathers, indicating that ID- protection is not dependent on the nature of the untransmitted class III allele (data not shown).

### ***Transmissions from class I/I heterozygous parents***

Class I alleles were again divided into the three subgroups and transmissions analysed by TDT from I/I heterozygous mothers and fathers (Table 10.6). There was no significant heterogeneity across the six parental genotypes (G test for heterogeneity,  $\chi^2 = 3.46$ , 5 df,  $p = 0.6$ ). Transmissions from only one parental genotype showed any evidence for transmission deviation from the expected 50%. This paternal IC+/ID- genotype showed over-transmission of ID- alleles to affected offspring. Therefore in contrast to transmissions from ID-/III fathers, ID- alleles not only fail to protect against type 1 diabetes, but may even be relatively predisposing when transmitted from IC+/ID- fathers. However, this transmission deviation was of borderline significance ( $p = 0.035$ ) and the lack of significant linkage of the minisatellite to type 1 diabetes when these transmissions are analysed by IBD (data not shown) suggests that all class I alleles are equally predisposing when transmitted from I/I parents.

## **Discussion**

A combination of variant repeat distribution at the insulin minisatellite and the analysis of flanking haplotype divided the insulin-linked region into five newly defined ancestral lineages, IC+, ID+, ID-, IIIA and IIIB. The three lineages of class I alleles differ in their associated predisposition to type 1 diabetes. Class ID- alleles are protective against type 1 diabetes, but only when transmitted from ID-/III heterozygous fathers. Class I alleles transmitted from all other parental genotypes appear to be equally predisposing. ID- alleles are generally larger than either IC+ or ID+ (Figure 10.3). A size threshold may exist above which class I alleles transmitted from I/III fathers are protective. Alternatively, protection could in theory be due to variant repeat distribution at the insulin minisatellite. Typically,

**Table 10.6*****Class I allele transmissions from I/I parents***

Genotype		Maternal	Paternal
IC+/ID+	t(IC+)	17	15
	t(ID+)	15	15
		$\chi^2=0.13$	$\chi^2=0.00$
IC+/ID-	t(IC+)	33	35
	t(ID-)	33	55
		$\chi^2=0.00$	$\chi^2=4.44$ $p<0.05$
ID+/ID-	t(ID+)	37	25
	t(ID-)	39	25
		$\chi^2=0.06$	$\chi^2=0.00$

Transmissions of class I alleles were tested for deviation from an expected 50% transmission frequency by TDT analysis as described in Table 10.1. Significant deviation was only detected for transmissions from the paternal IC+/ID- genotype, though this was of borderline significance. No heterogeneity was observed between transmissions from all six genotypes.

the protective ID- alleles differ from ID+ alleles by the presence of an ABA motif 10 repeats upstream of the end of ID- alleles (Figure 10.2). However, this motif is also present in all class IC+ alleles which are predisposing to type 1 diabetes. Whilst ID- can be distinguished from IC+ by the presence of a 5' CAC motif, a central FAC motif, and a 3' F repeat, these motifs are also present in most ID+ alleles (Figure 10.2). There was therefore no consistent difference between repeat-type distribution within ID- alleles and both ID+ and IC+ alleles. It is possible that the protective effects of ID- alleles are mediated by a combination of multiple variant repeats (the 5' CAC motif, the central FAC motif, the 3' F repeat, *and* the additional ABA motif) which can distinguish ID- alleles from other class I alleles. However, it is unlikely that such minor differences in repeat composition would have any major effect on levels of associated susceptibility, when more substantial differences such as those between IC+ and ID+ alleles have no apparent effect. This could be tested by analysing transmissions of the small number of ID+ alleles which do not share the ABA deletion from ID+/III fathers (Figure 10.2). However in this study, of the five ID+ alleles identified without the ABA deletion, only one was found in ID+/III heterozygous fathers.

The obvious aetiological candidate for the protective effects of ID- alleles is within the flanking haplotype. All ID- alleles lack an *MspI* site at +3580*MspI* within the first intron of *IGF2* over 4 kb downstream of the insulin minisatellite, which is present on 99.6% of all other haplotypes. All observed heterogeneity in the effects associated with class I allele susceptibility to type 1 diabetes observed in this study can therefore be attributed to this single polymorphic site, or to any other polymorphism in linkage disequilibrium with the ID- haplotype either 5' or 3' of the minisatellite. The effects of this secondary site are specific to transmissions from the paternal ID-/III genotype. ID- alleles are not protective when transmitted from any other parental genotype and may even be more predisposing than IC+ alleles transmitted from IC+/ID- fathers. The overall effects of *IDDM2* cannot therefore be readily explained by either the insulin minisatellite or a secondary polymorphism when the loci are considered independently. It is therefore concluded that *IDDM2* has a multi-locus aetiological basis, with the functional significance of different aetiological loci determined both by parental genotype and gender.

The insulin minisatellite was previously identified as *IDDM2* by consideration of polymorphisms within a 4.1 kb region of susceptibility which excludes the +3580*MspI* site (Bennett *et al.*, 1995; Lucassen *et al.*, 1993). Protection against type 1 diabetes would not have been associated with the *MspI* polymorphism when considered in isolation as the *MspI* variant would associate with predisposition to disease when transmitted from any parental genotype other than *I/III* heterozygous fathers. To postulate that *IDDM2* has a multi-locus aetiology raises doubts as to whether the insulin minisatellite has any causative role in susceptibility to type 1 diabetes. The minisatellite was the only single locus polymorphism within the 4.1 kb region of susceptibility to which heterogeneity in the effects of class I alleles could be attributed as all other polymorphisms were identical. However, if all transmission heterogeneity can be correlated with a site outside of this 4.1 kb region, then the primary aetiological variant could be any of the 4 polymorphisms (-2733A/C, *INS* VNTR, -23*HphI*, and +1140A/C) the identity of which differs between the class I haplotype and both the PH and VPH.

Three of the four polymorphisms within the 4.1 kb region which were shared by PH and VPH were biallelic markers so could not account for the differences in the levels of protection associated with the two haplotypes (Bennett *et al.*, 1995). Only the multi-allelic minisatellite could explain the observed heterogeneity between PH and VPH. However, a multi-locus model for *IDDM2* aetiology again raises doubts over this conclusion. The VPH shares 6 polymorphic variants with class I alleles. If one of these variants resulted in mild protection against type 1 diabetes, it could account for the observed difference between PH and VPH. For the class I haplotype, this mild protection would be masked by the predisposition conferred by the primary locus. It could therefore be argued that there is no definitive evidence to demonstrate that the insulin minisatellite has any aetiological role in type 1 diabetes susceptibility. Furthermore, it is unknown whether all polymorphisms in disequilibrium with the five minisatellite lineages have been identified as previous sequencing strategies (Chapter 9) which screened for polymorphisms generally assumed that there were just two main haplotypes, class I and class III. Sequence analysis of the five newly defined haplotypes could therefore identify additional candidates for *IDDM2* aetiological variants.

Nevertheless, the insulin minisatellite is a highly polymorphic locus located in a region which could affect transcription of both the insulin gene and *IGF2*, so must remain a strong candidate for the primary aetiological locus of *IDDM2*. Differences in repeat-type distribution (Appendix 7) and composition (Table 10.7) may account for the different associations of class I, IIIA, and IIIB alleles with type 1 diabetes susceptibility. For example, the H-type repeat may protect against disease by increasing insulin gene transcription through the binding of the Pur-1 transcription factor (Kennedy *et al.*, 1995), although its presence in IIIA (PH) but not IIIB (VPH) alleles excludes it as the sole determinant of protection. A-type repeats encourage *in vitro* formation of hairpin G-quartet structures which may elevate *in vivo* insulin gene transcription by facilitating denaturation at the insulin gene promoter (Catasti *et al.*, 1996). The higher frequency and copy number of A-type repeats in IIIA alleles compared with IIIB alleles again argues against a functional role. However in IIIB alleles, 5/6 repeats closest to the insulin gene were A-type compared with 0/6 in IIIA alleles (Appendix 7), and so repeat distribution may be a more important factor than absolute composition. Methylation within the insulin minisatellite may result in protection against type 1 diabetes (Awata *et al.*, 1997), and imprinting is an obvious mechanism to account for parent-of-origin effects observed at *IDDM2*. In this study, known methylation sites were present only within F-type repeats. Fewer F-type repeats were detected within class IIIA alleles than within either ID+ or ID- alleles arguing against any correlation of protection with methylation state. However, comparison of IIIA alleles with the previously published III-G allele (Owerbach and Gabbay, 1993) (data not shown) indicates that many IIIA null (o-type) repeats may contain CpG sites. The absence of any detectable difference between the effects of IC+ alleles (2 F-type repeats) and ID+ alleles (4 F-type repeats) does not favour a role for methylation within the minisatellite. Nevertheless, without detailed sequence analysis of all null repeats and investigation of methylation states within the minisatellite, a general correlation between levels of methylation and levels of protection against type 1 diabetes cannot be excluded.

### ***A model for the aetiology of IDDM2***

The protective effects of ID- alleles transmitted exclusively from ID-/III fathers raises some baffling questions about possible aetiological mechanisms. It is generally assumed that the effects of *IDDM2* are mediated by alterations in levels of gene expression, most probably at



**Table 10.7*****Variant repeat composition of allele subclasses***

Variant Repeat	IC+		ID+		ID-		IIIA		IIIB	
A	18.1	(54.0%)	20.9	(52.6%)	23.2	(54.7%)	91.7	(61.5%)	78.0	(55.6%)
B	8.7	(25.8%)	7.9	(20.0%)	8.3	(19.5%)	22.8	(15.3%)	9.0	(6.4%)
C	2.8	(8.4%)	4.9	(12.4%)	5.0	(11.7%)	17.6	(11.8%)	19.3	(13.8%)
E	0.0	(0.0%)	0.0	(0.0%)	0.0	(0.0%)	5.6	(3.8%)	9.6	(6.8%)
F	2.0	(5.9%)	3.9	(9.9%)	4.0	(9.4%)	2.0	(1.3%)	15.5	(11.0%)
H	0.0	(0.0%)	0	(0.0%)	0.0	(0.0%)	1.0	(0.7%)	0.0	(0.0%)
o	2.0	(6.0%)	2.0	(5.0%)	2.0	(4.7%)	8.3	(5.6%)	8.9	(6.4%)
Total	33.6		39.8		42.5		149.1		140.2	

Mean variant repeat number and frequency (presented in parentheses) were determined for each allele class, with mean allele sizes (in number of repeats) presented below.

either the insulin gene or *IGF2* (Paquette *et al.*, 1998; Pugliese *et al.*, 1997; Vafiadis *et al.*, 1997). For an untransmitted class III allele to affect expression levels associated with the transmitted ID- allele is an example of paramutation (Bennett *et al.*, 1997); where one allele or locus alters the activity of another allele or locus, in a way which persists after segregation of the two alleles into the offspring (Hollick *et al.*, 1997). Paramutation is poorly understood, but is thought to involve physical interactions between alleles or loci (Wolffe and Matzke, 1999). While some cases of paramutation involve changes in methylation states, others do not suggesting a role for altered and heritable chromatin conformation (Wolffe and Matzke, 1999).

The mechanistic basis of many epigenetic effects is thought to involve the recognition of nucleic acid sequence homologies both at the DNA and RNA levels, although the precise mechanism by which this recognition occurs is unknown (Wolffe and Matzke, 1999). In *Ascombolus*, DNA methylation can be transferred inter-chromosomally between paired alleles by a mechanism related to homologous recombination in which DNA-DNA pairing serves as a signal for *de novo* methylation (Selker, 1997). Alternatively, RNA molecules have been implicated in the methylation of homologous DNA sequences (Mette *et al.*, 1999; Pelissier *et al.*, 1999). Both DNA-DNA and RNA-DNA mediated gene silencing are dependent on nucleotide homology. This dependence does provide an indication for how the ID-/III associated paramutation effects may operate. If the insulin minisatellite is considered as a biallelic marker (all alleles divided into either class I or class III or alternatively any biallelic marker such as -23*Hph*I which is in linkage disequilibrium with these two minisatellite classes is considered), then the only genotype which is heterozygous at *both* the insulin minisatellite and at +3580*Msp*I is ID-/III.

I propose the following hypothetical model to explain the paramutation effects observed at the insulin minisatellite. In this model, I assume that susceptibility to type 1 diabetes is due to reduced levels of insulin gene expression in the foetal thymus. This assumption is not an intrinsic aspect of the model but is simply used to illustrate how changes in gene expression could affect disease susceptibility.

By default, insulin gene expression in the foetal thymus is assumed to occur at a high level. Gene expression is reduced by (for example) DNA methylation or alterations in chromatin conformation in response to inter-allelic DNA-DNA interactions in the parent, which are dependent on homology between parental haplotypes.

Class III alleles possess an intrinsic property which prevents or reduces their inactivation. Class I alleles lack this property. This difference could be due to a functional single nucleotide polymorphism which differs between class I and class III haplotypes, or could be a property of the minisatellite. For example, the large tandem repeat array of class III alleles may prevent chromatin condensation and gene inactivation. Alternatively, the formation of extensive G-quartet structures within class III alleles may result in denaturation of the insulin gene promotor (Catasti *et al.*, 1996), either preventing or antagonising gene silencing. The smaller tandem repeat arrays in class I alleles may be insufficient to prevent gene repression. Under this model, allele size might therefore be the only characteristic of the insulin minisatellite which affects disease susceptibility.

Therefore, DNA-DNA interactions in the parent result in reduced expression of class I allele-associated genes, but an intrinsic property of class III alleles prevents or reduces the level to which they are inactivated. This would then correlate with differences in insulin gene expression observed in the foetal thymus and susceptibility to disease. In all I/I parental genotypes, high levels of homology between interacting alleles would result in reciprocal gene inactivation. In general, there would also be sufficient allele homology in I/III heterozygous parents for the inactivation of class I alleles. However, ID- alleles transmitted specifically from ID-/III heterozygous fathers would have to escape inactivation for this model to be feasible. If there was a threshold level of homology required for gene inactivation, the heterozygosity at the +3580MspI site may result in this threshold being exceeded, thereby preventing ID- inactivation. Homology searches could either be postulated to extend over a region of up to 4 kb, or could involve multiple homology searches in close proximity, each of which covers a small region. Heterozygosity at the +3580MspI site in for example the ID+/ID- parental genotype would not prevent class I allele inactivation due to the higher homozygosity at other polymorphic sites at this genotype (assuming that either no, or few unidentified polymorphisms exist between the

ID+ and ID- lineages). The fact that ID- alleles escape inactivation specifically when transmitted paternally can be easily explained by invoking differences in the regions involved in homology searches between the male and female germlines, so that homozygosity at +3580*Msp*I is not necessary for class I allele inactivation in females.

This model could even be extended to account for the differences between the effects associated with PH and VPH. The class I/PH heterozygous genotype is heterozygous at more sites than the class I/VPH genotype. The greater homology between class I and class III VPH haplotypes could elevate class I allele inactivation. This would result in a redefinition of PH and VPH where both classes of allele were intrinsically equally protective against type 1 diabetes, but the untransmitted VPH conferred greater predisposition upon the transmitted class I allele than the untransmitted PH as a result of elevated class I inactivation. Increased class III protection and increased class I susceptibility are indistinguishable in TDT analyses.

The apparent protective effect of ID- alleles transmitted from ID-/VPH fathers does raise concerns about this model. It would be expected that the greater sequence identity between ID- and class III VPH haplotypes would result in ID- inactivation. Nevertheless, the region or regions involved in the proposed homology searches which lead to inactivation are open to speculation, and the effects of heterozygosity at different polymorphisms may vary. The model is therefore not necessarily invalidated by the protective effects of ID- alleles from ID-/VPH fathers.

If a threshold of heterozygosity does block gene inactivation, and if a single 3' SNP such as +3580*Msp*I was sufficient to surpass this threshold, homology searches at least in males would have to include a region 3' of the minisatellite. It is therefore interesting that *de novo* mutation studies at the insulin minisatellite (described in Chapter 8) provide evidence for inter-allelic recombination processes in the male germline in which the flanking DNA 3' of the repeat array is paired.

Finally, one prediction of this model is that true homozygosity in parents at the insulin-linked region would result in increased inactivation of class I-associated alleles and

therefore elevated susceptibility to type 1 diabetes in the offspring. This prediction is testable using the data presented in this chapter.

By analysing parents of affected sib pairs, there was a preselection for alleles which associated with susceptibility to type 1 diabetes. As a result, 80% of alleles were class I compared with 70% in the general population (Bell *et al.*, 1984). If specific parental genotypes elevate susceptibility to disease in the offspring, similar enrichment for these genotypes would be expected. By the above model, the most predisposing genotypes would be true homozygotes.

In Chapter 8, the mutation rate at the insulin minisatellite was estimated from levels of parental homozygosity (9%) by  $H = \frac{1}{1 + 4N_e\mu}$  at  $10^{-4}$ . This was approximately 10-fold lower than the mutation rate estimated in the cohort by Ewens distribution (Ewens, 1972). A mutation rate of  $10^{-3}$  would correspond to a homozygosity of 1.2%. One interpretation of this result is therefore that the frequency of true homozygotes in the parents of type 1 diabetes affected offspring was greater than expected.

A more direct test for this hypothesis is possible by comparing the number of homozygous parents and offspring. If parental homozygosity has no effect on disease susceptibility in offspring, the number of homozygous offspring would be higher than the number of homozygous parents for two reasons. The first reason is that the probability of the sib of a homozygote also being homozygous (given 50% transmission of alleles) is 0.25. This is obviously elevated if one or both parents are also homozygous to 0.5 and 1 respectively. The second reason is that due to the dominant protection against type 1 diabetes conferred by class III alleles (Bell *et al.*, 1984), more affected offspring would have the I/I genotype than parents (>90% of whom were unaffected, John Todd, pers. commun.), and therefore more offspring would be true homozygotes. Transmission ratio distortion (Eaves *et al.*, 1999) would have similar implications.

To correct for the lack of independence of the genotypes of two offspring, the number of *families* in which one or both parents, or one or both offspring were true homozygotes was determined (Appendix 9). 36/219 families had at least one homozygous parent

(38 homozygous parents in total) compared with 21/219 families in which at least one child was homozygous (29 homozygous children in total). Testing this difference against an expected 1:1 ratio of homozygous parents to offspring underestimates the expected frequency of homozygous children for the reasons described above. Nevertheless, even testing against an expected ratio of 1:1, there were significantly more families with homozygous parents than families with homozygous offspring ( $\chi^2=3.95$ , 1 df,  $p<0.05$ ).

There are two interpretations to this result; either parental homozygosity increases susceptibility in the offspring to type 1 diabetes (as predicted by the above model), or homozygosity in offspring reduces their susceptibility to type 1 diabetes. With the current data set, it is not possible to distinguish between these interpretations. However, class III alleles are dominantly protective against disease (Bell *et al.*, 1984). If heterozygosity in offspring increased susceptibility to disease, I/III heterozygotes would be more predisposed to disease and not protected against disease, thus favouring the parental homozygosity model.

Elevated susceptibility to type 1 diabetes in children due to parental homozygosity has some intriguing implications. Selection will act on the parental genotype by manifesting a phenotype in the offspring, thereby reducing the probability of grandchildren. Furthermore, selection would favour parental heterozygosity and it would be predicted that populations with lower levels of genetic diversity or elevated inbreeding would display increased incidence of type 1 diabetes. This selection may also impose an upper limit on allele frequency within a population, and could in fact result in (albeit weak) positive selection for *de novo* mutation at the insulin minisatellite. However, extensive work would be required to confirm this association of parental homozygosity with type 1 diabetes in the offspring, as the level of significance identified in this cohort was of borderline significance ( $p=0.047$ ). Further studies would ideally compare levels of homozygosity between parents of affected and unaffected offspring. This comparison would have to account for the expected difference in class I allele homozygosity in parents between the two cohorts due to enrichment for predisposing class I alleles in the parents of affected offspring (predicted frequency of I/I genotypes in parents of affected offspring is  $0.8^2=0.64$ , compared with  $0.7^2=0.49$  in parents of unaffected offspring).

Another approach to test this model for *IDDM2* aetiology would involve RT-PCR analysis of insulin gene expression in the foetal thymus. The model would predict that all parental alleles are protective against type 1 diabetes and would be transcribed at higher frequency in the foetal thymus unless levels of sequence identity between the transmitted and untransmitted parental alleles were above a certain threshold. If samples could be obtained from non-Caucasian (especially African) parents and foetuses, then sequence analysis of parental haplotypes and RT-PCR analysis of foetal thymic expression, could be used to test for correlations between parental homozygosity and foetal gene expression.

## Future Directions

It is ironic that after extensive analysis of the insulin minisatellite by MVR-PCR the main result of this chapter, that ID- alleles from ID-/III fathers are protective against type 1 diabetes, could have been determined by the analysis of just two single nucleotide polymorphisms; the -23*Hph*I polymorphism (which is a surrogate marker for class I and class III alleles), and the +3580*Msp*I polymorphism which defines the ID- haplotype. Analysis could be extended to incorporate the division of class III alleles into lineages IIIA and IIIB by inclusion of the +1428*Fok*I site which is generally used to distinguish the PH and VPH haplotypes. It would therefore be relatively easy to extend this study to analyse additional type 1 diabetic affected sib pair families to increase the robustness of the results, and to investigate whether the nature of the untransmitted class III allele from ID-/III fathers does affect levels of protection associated with ID- fathers. No differences between the IIIA and IIIB lineages were detected in this study, but this may have been due to a small sample size.

There is strong evidence (Doria *et al.*, 1996; Merriman *et al.*, 1998) that linkage disequilibrium extends 5' of the insulin minisatellite further than the 4.1 kb region of susceptibility identified by (Lucassen *et al.*, 1993). In this chapter, linkage disequilibrium has also been shown to be maintained over at least 4 kb 3' of the minisatellite. Further analysis of *IDDM2* should therefore include extensive sequence analysis of perhaps 200 kb each side of the minisatellite (Todd, 1999b) to identify further polymorphisms which may be associated with disease. The definition in this chapter of three new class I allele haplotypes may be used to target appropriate haplotypes for sequence analysis.

High levels of linkage disequilibrium across the insulin-linked region make the functional dissection of variant sites by cross-match haplotype analysis difficult. Greater haplotype diversity may therefore be essential for the genetic identification of aetiological loci. Studies are planned for the characterisation of allele lineages in African populations where lineage diversity defined by haplotype is far higher than in Caucasians (Mijovic *et al.*, 1997). These studies would require characterisation of both the insulin minisatellite and flanking polymorphisms. The main problem facing this work is that increased lineage diversity would require very large sample sizes for the detection of any association between a specific allele lineage and susceptibility to type 1 diabetes. Sufficiently large cohorts of individuals affected by type 1 diabetes are not currently available. In addition, full characterisation of the insulin minisatellite is very labour intensive so would be prohibitive in such a large cohort. In Caucasians, low levels of allele variation within each lineage made it apparent that the characterisation of a small number of alleles would have been sufficient to define allele structure for each lineage. If levels of diversity within each lineage identified in non-Caucasians is also low, it may be possible to define minisatellite lineages from MVR analysis of a small number of alleles, after which flanking polymorphisms could be used as surrogate markers for allele lineage.

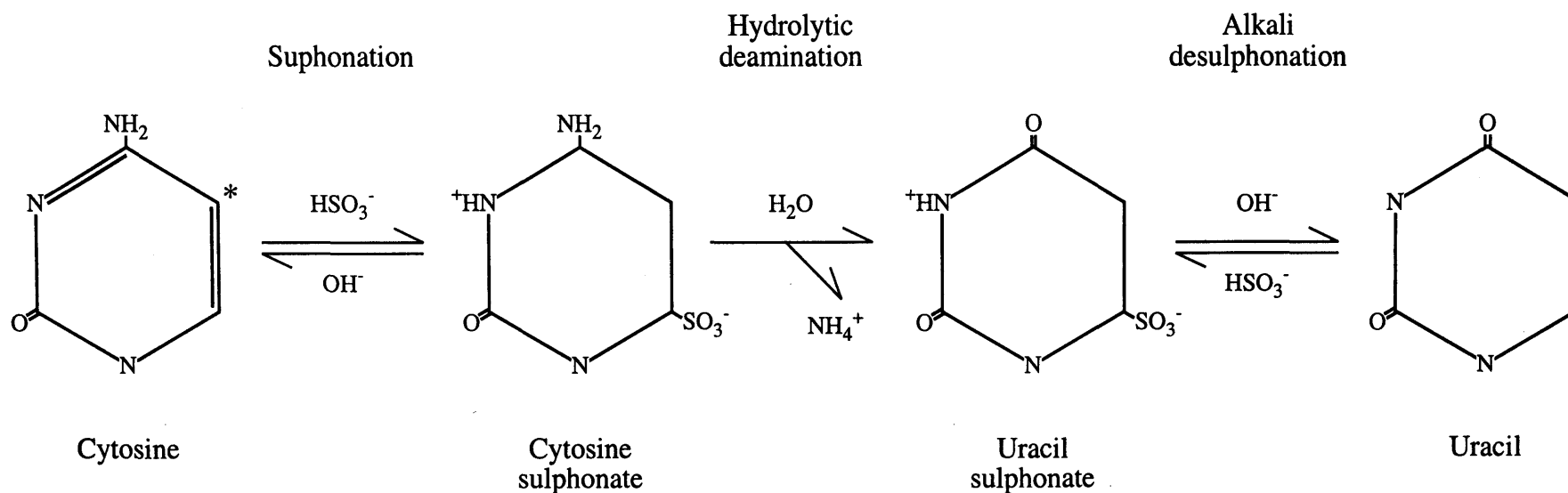
To complement the genetic identification of aetiological variants by linkage and association studies, further functional studies are required at *IDDM2* candidate polymorphisms. Despite the lack of an apparent correlation between the number of methylation sites within the insulin minisatellite and the association of minisatellite alleles with disease (Table 10.7), the parent-of-origin effects which associate with the insulin-linked region do implicate a role for imprinting. The detailed characterisation of methylation state across the insulin-linked region in a range of tissues including the pancreas, placenta, and foetal thymus would therefore be a very interesting study. Perhaps the most obvious approach for methylation detection would be to use sequence analysis of bisulphite-treated DNA (Frommer *et al.*, 1992), due to the sensitivity of the technique, and the absence of any bias towards specific target sequences which is a feature of methylation detection using methylation-sensitive restriction endonucleases.



Bisulphite treatment of DNA converts unmethylated cytosine residues to uracil, whilst methylated cytosines remain unchanged (Grigg and Clark, 1994). The chemistry of bisulphite treatment is summarised in Figure 10.6. Conversion of cytosines to uracil creates non base-complementary strands (i.e. uracils opposite guanines). DNA amplification using strand-specific primer pairs converts the single stranded products of bisulphite treatment to double strands, and converts uracils to thymines. The double-strand product can be sequenced to screen for methylated, and therefore unconverted, cytosine residues (Grigg and Clark, 1994). While this technique has been successfully applied to the analysis of non-repetitive DNA (Clark *et al.*, 1995; Clark *et al.*, 1994; Gonzalgo and Jones, 1997; Guldborg *et al.*, 1998; Olek *et al.*, 1996; Paulin *et al.*, 1998; Warnecke *et al.*, 1998; Warnecke *et al.*, 1997), the difficulties commonly associated with the amplification of minisatellites and, more seriously, complete sequence analysis of large tandem repeat arrays using flanking primers, would have prevented sequence analysis of at least class III alleles of the insulin minisatellite from bisulphite-treated DNA.

From the work presented in this chapter and in Chapter 8, it is apparent that a system of MVR-PCR, modified for specificity to bisulphite-treated DNA, could be established at the insulin minisatellite and potentially used in two ways. The first would be to create a series of deletion amplicons using an MVR primer analogous to INS-MER as described in Figure 8.5. This would divide the locus into small sections which could be directly sequenced. However, such sequence analysis of the locus in many tissues and from many subjects would be relatively labour intensive. A more efficient alternative would be to develop a system of MVR-PCR which discriminates between methylated and unmethylated sites. For example, following bisulphite treatment of an unmethylated F-type repeat, the CCCGGGGACAGGGGT sequence would be converted to TTTGGGGATAGGGGT whilst a methylated F-type repeat would be converted to TTCGGGGATAGGGGT. The resulting repeat variants could be readily distinguished throughout the entire allele in a single step by MVR-PCR using just two MVR primers for each allele. My primary concern for such a technique is whether it would allow the quantitative description of levels of methylation in partially methylated samples. Quantitation may be possible using densitometry analysis of band intensity for corresponding methylated and unmethylated repeats on autoradiographs. It should be possible to determine whether the technique is

**Figure 10.6**



***The chemistry of bisulphite-based methylation analysis***

Cytosine is capable of forming adducts with a number of reagents including bisulphite across the 5-6 bond. Hydrolytic deamination of cytosine- $\text{SO}_3$  converts it to uracil- $\text{SO}_3$  which can be desulphonated to form uracil. Although methylated cytosine can also react with bisulphite, the reaction is extremely slow and the equilibrium favours 5-MeC rather than the deaminated product, thymine (Clark *et al.*, 1994). The C residue which is methylated in 5-MeC is indicated by \*. Figure 10.6 was adapted from Grigg and Clark (1994) .

quantitative by mixing different ratios of fully methylated DNA (e.g. sperm DNA) with unmethylated DNA (e.g. preamplified DNA or DNA from cell lines) prior to bisulphite treatment. Alternatively, bisulphite-treated genomic DNA could be diluted to single molecule levels prior to amplification so that in most PCR reactions a maximum of one input molecule of DNA would be included. MVR mapping of amplified molecules would effectively allow methylation state to be determined individually in large number of molecules of genomic DNA.

The insulin minisatellite has been most intensively studied in relation to its association with susceptibility to type 1 diabetes (Bennett and Todd, 1996a). However, it has also been implicated in type 2 diabetes (Ong *et al.*, 1999), polycystic ovary syndrome (Waterworth *et al.*, 1997), obesity (O'Dell *et al.*, 1999), and infant birth size (Dunger *et al.*, 1998). Levels of variation within allele lineages in Caucasian populations have now been extensively characterised. Full MVR-PCR analysis is therefore unnecessary to determine minisatellite lineage. It is possible to develop a 'short-hand' system of MVR-PCR by which class III alleles can be divided into the IIIA and IIIB lineages due to differences in their interspersed patterns of E-type repeats (Figure 8.5c). Lineage analysis can therefore be performed in a single PCR reaction with visualisation of PCR products on an ethidium bromide-stained gel (Figure 8.5c). This approach would not be recommended for subclassification of class III alleles; a more efficient approach would be typing of the +1428*FokI* polymorphic site. However, an analogous system detecting F-type repeats could be readily established for subtyping class I alleles into lineages IC and ID, with the further division into ID+ and ID- determined by the +3580*MspI* polymorphism. I would estimate that this combination of 'short-hand' MVR with analysis of flanking polymorphisms would allow ~5 times as many families to be analysed in a given period of time compared with the full MVR analysis presented in this thesis.

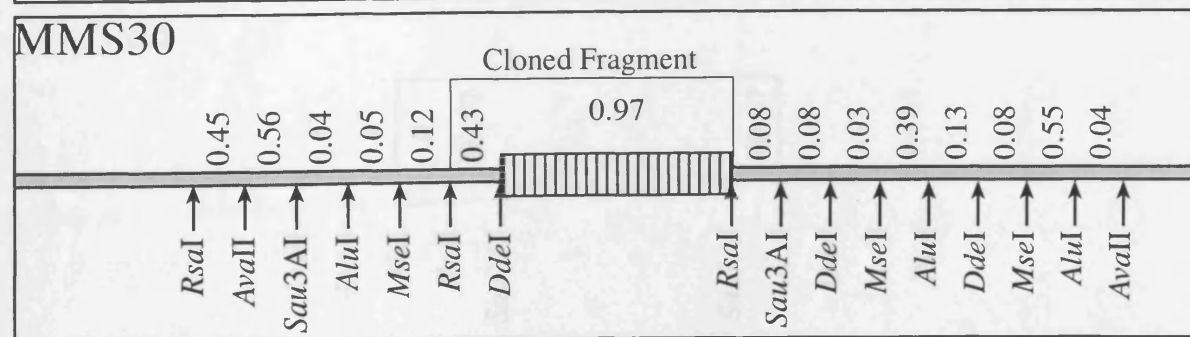
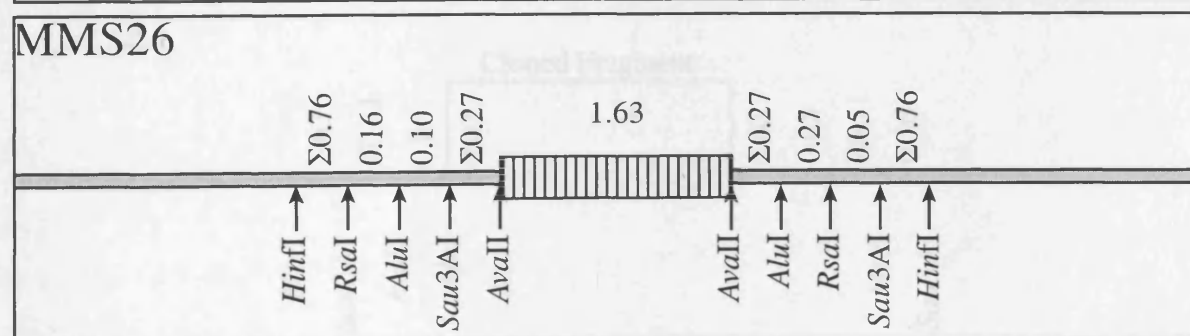
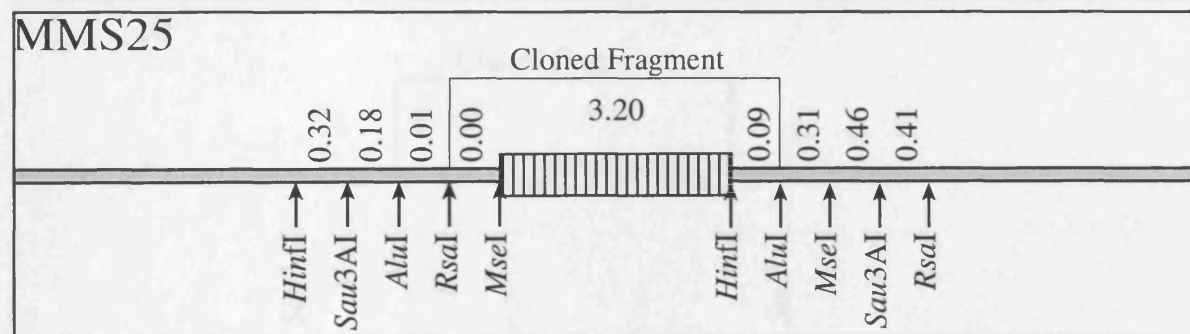
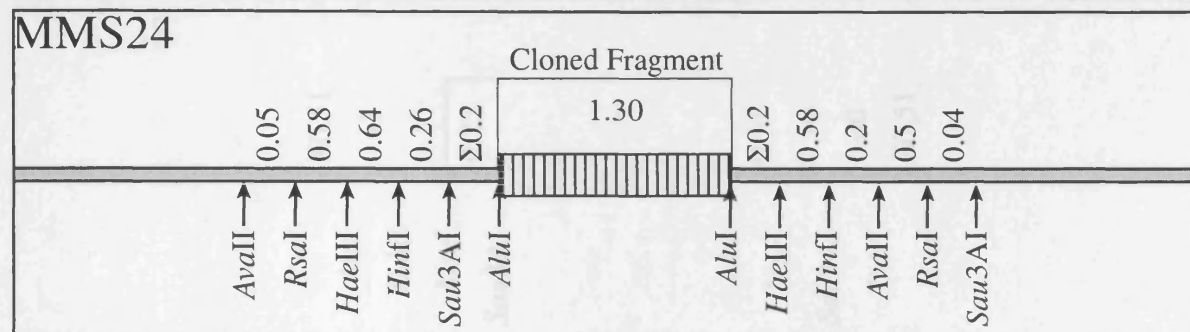
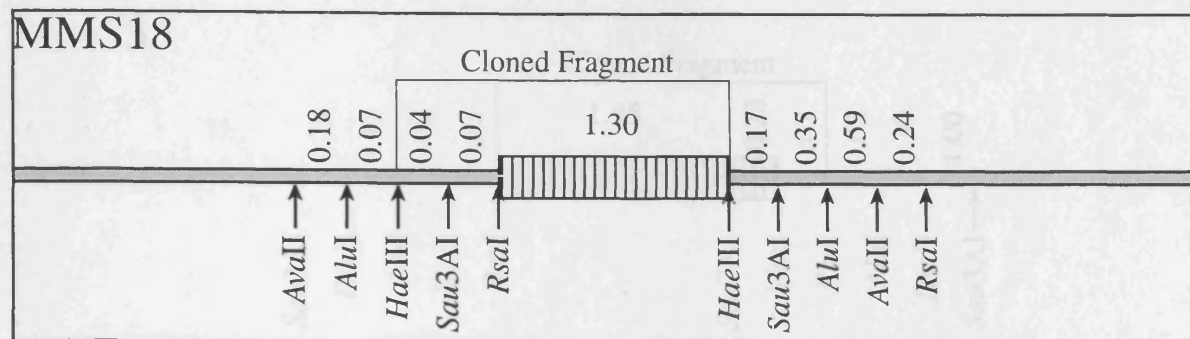
Current plans are to apply 'short-hand' MVR-PCR for the subclassification of class I alleles, to investigate the effects of minisatellite lineage on both type 2 diabetes and polycystic ovary syndrome (this work is planned in collaboration with Dr. Mark McCarthy). In contrast to type 1 diabetes, class III alleles are predisposing to both these diseases. In addition, subdivision of class III alleles into PH and VPH lineages revealed that

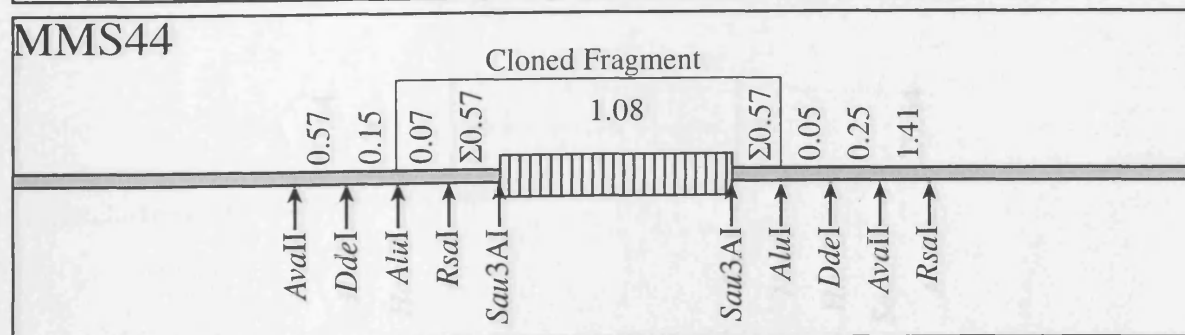
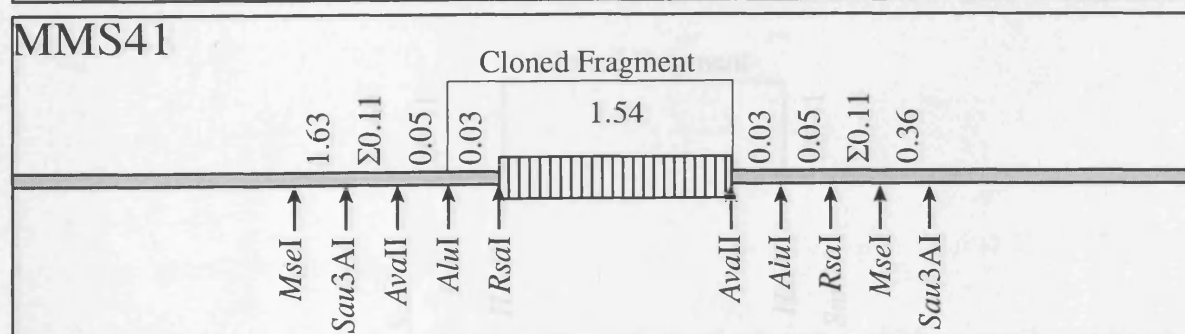
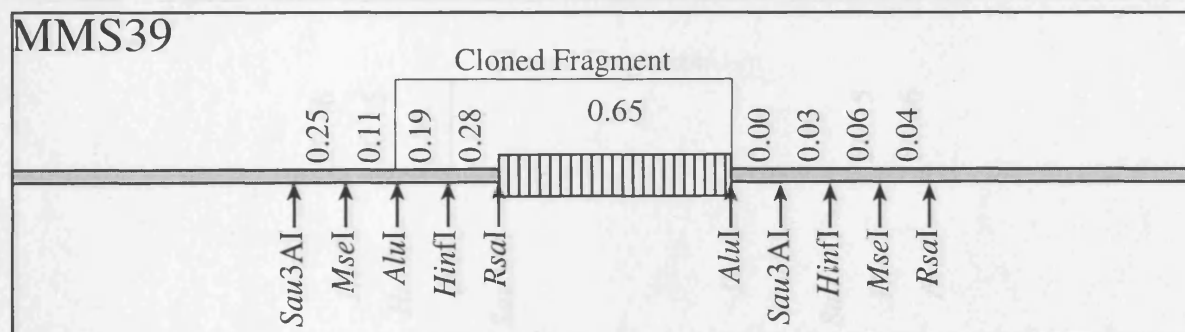
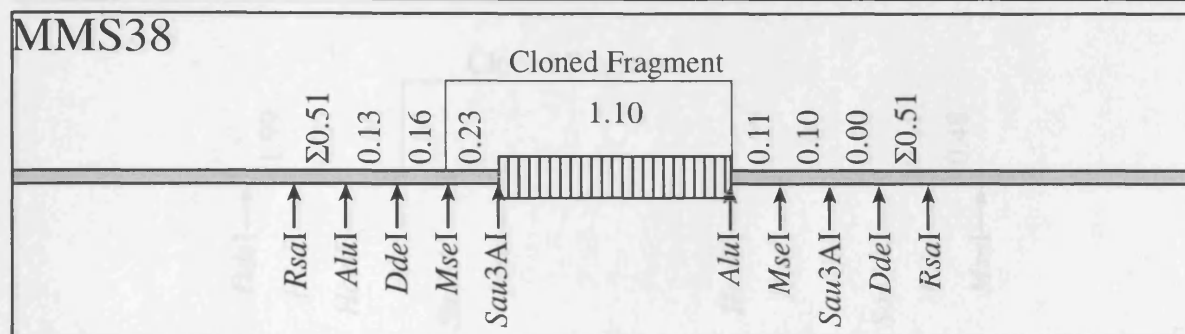
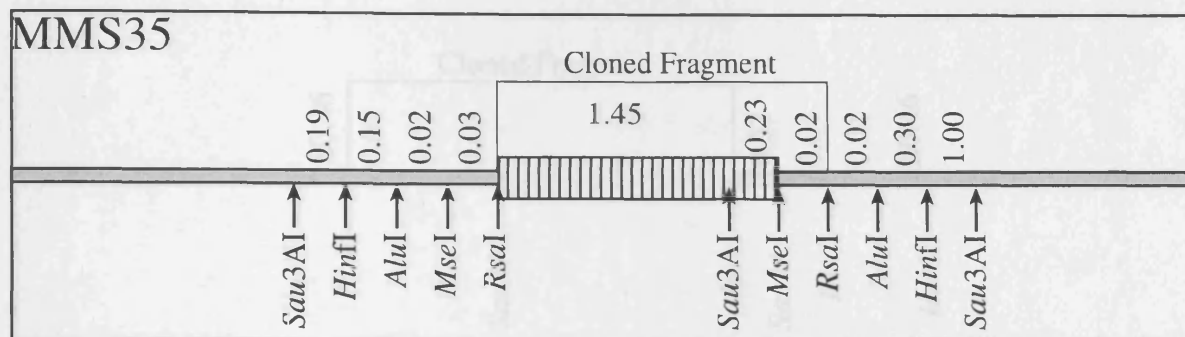
VPH alleles associated with greater susceptibility than PH alleles when transmitted from I/III parents; the reverse of type 1 diabetes (Mark McCarthy, pers. commun.). No subdivision of class I alleles has yet been attempted in these studies and it would be interesting to determine whether ID- alleles transmitted from ID-/III heterozygous fathers associated with elevated susceptibility to type 2 diabetes and polycystic ovary syndrome. If elevated susceptibility is detected, it would support the hypothesis that the effects of the insulin-linked region on type 1 and type 2 diabetes share a common aetiology. Furthermore, with a multi-locus model for *IDDM2*, ID- associated susceptibility to type 2 diabetes would suggest that both the primary and secondary locus (or loci) of *IDDM2* act within a common aetiological pathway.

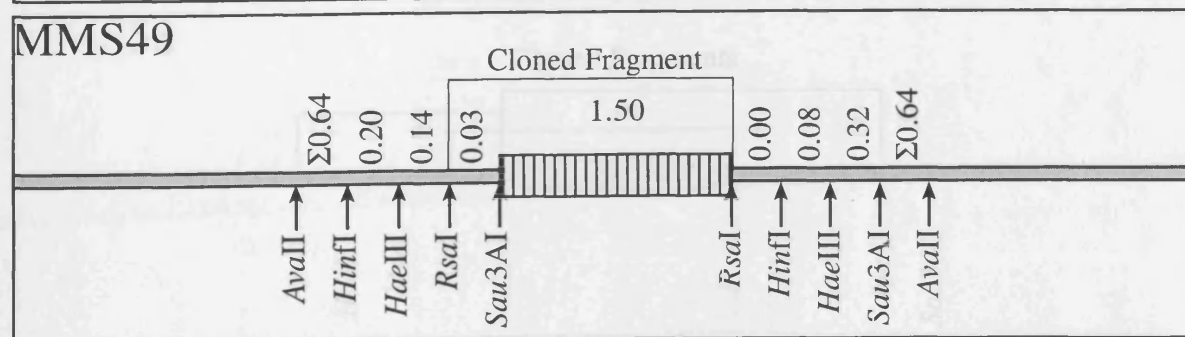
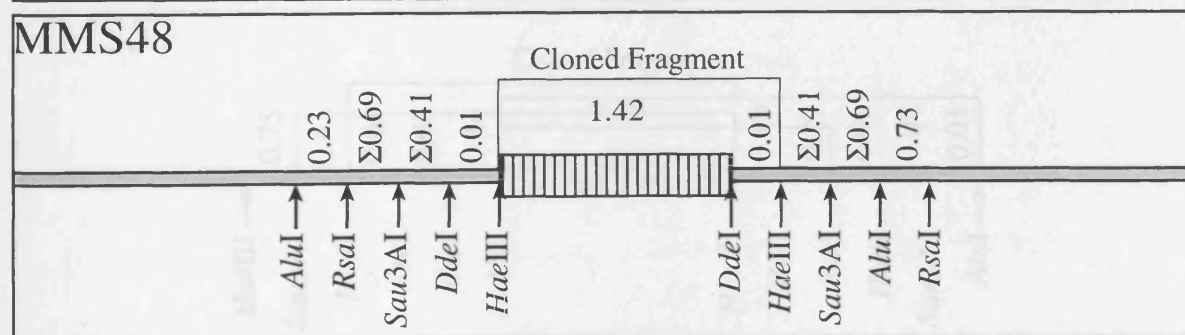
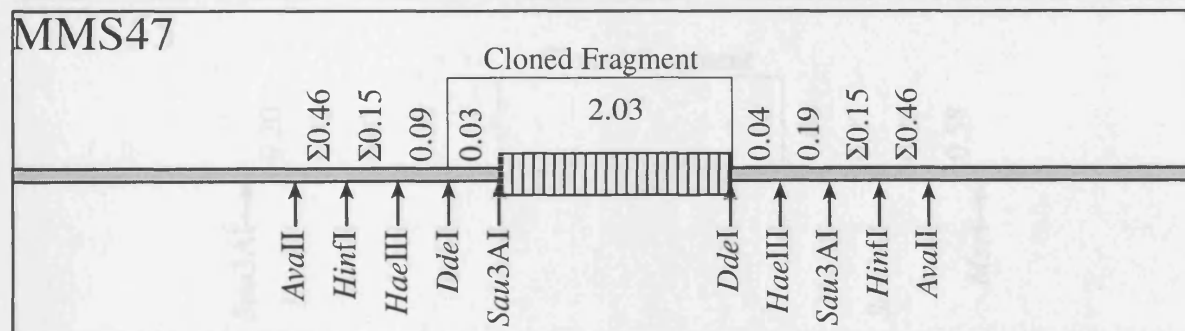
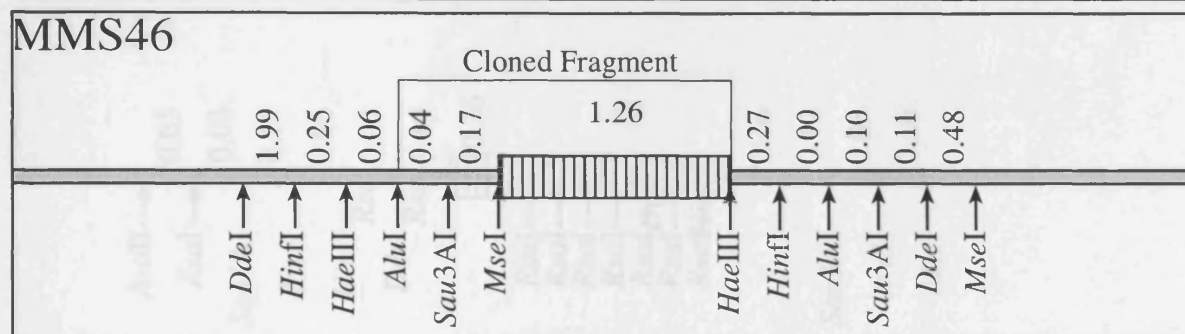
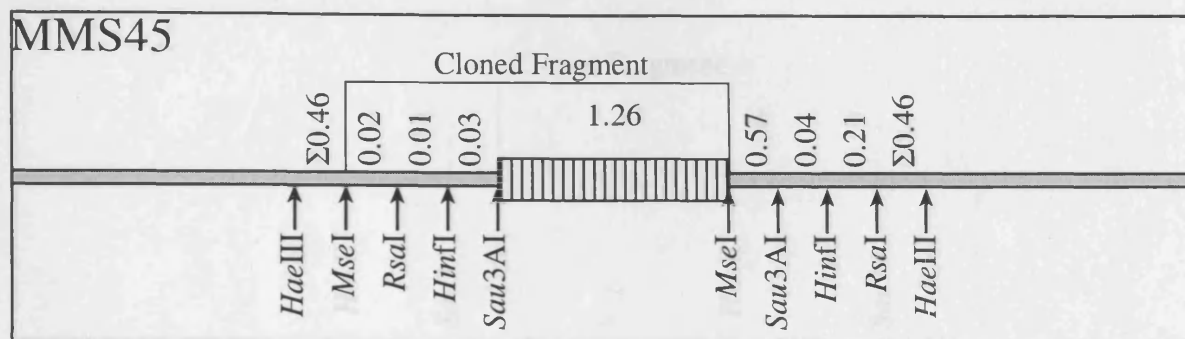
## Appendix 1

### ***Proximal restriction maps of mouse minisatellites***

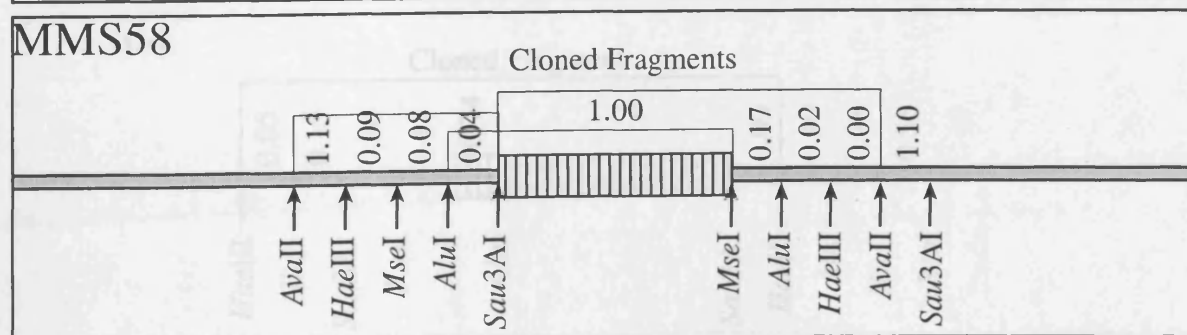
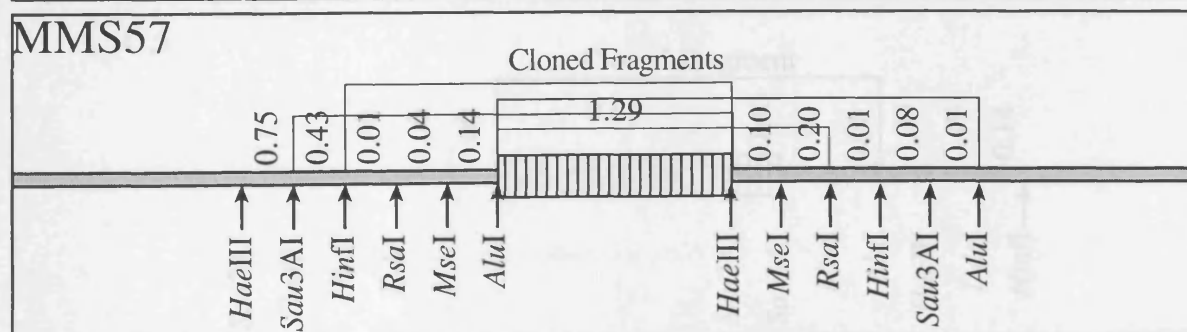
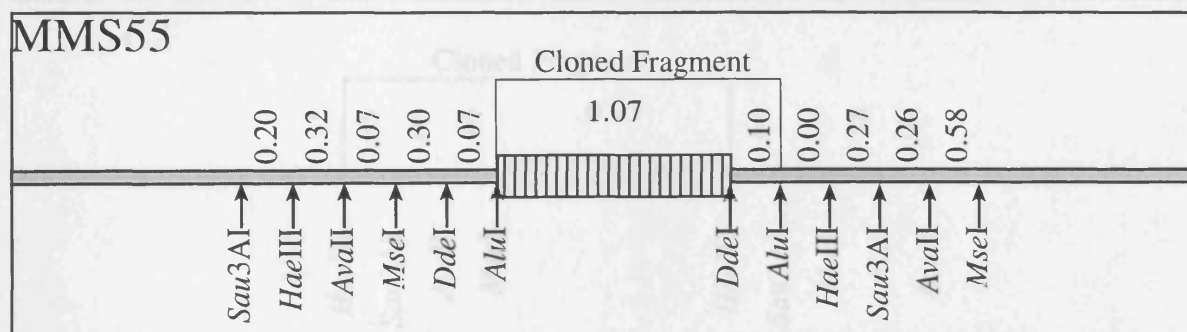
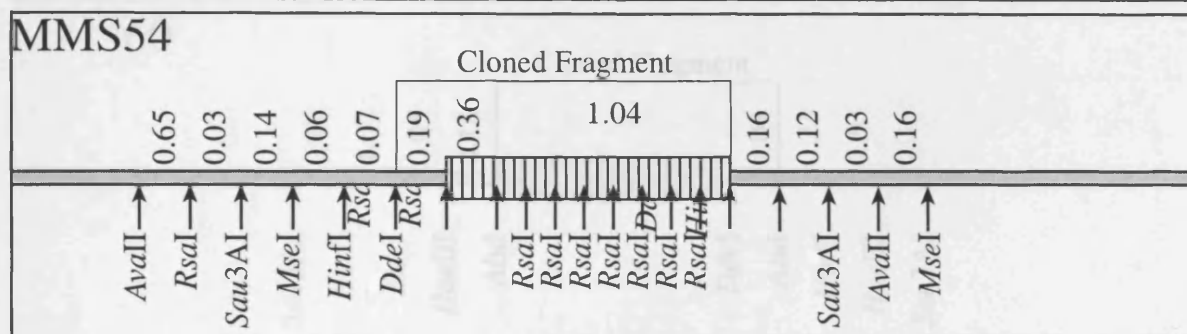
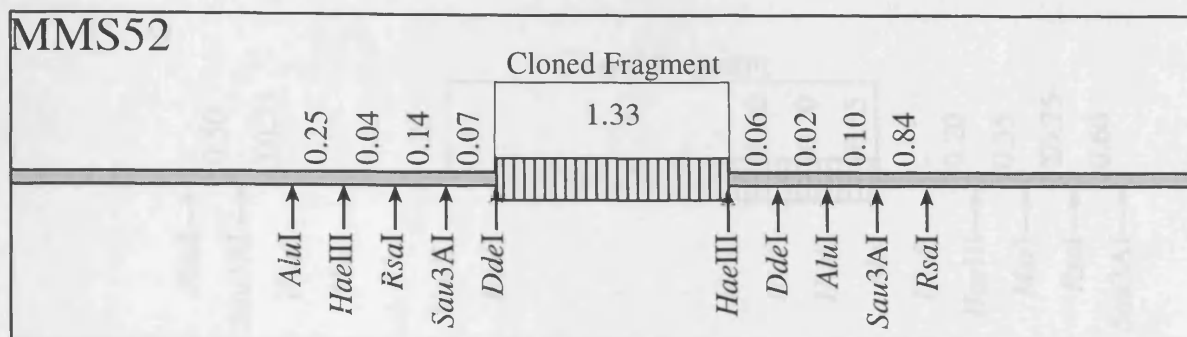
Schematic proximal restriction maps of 30 mouse minisatellites were generated as described in Chapter 3. Arrows indicate the most proximal restriction sites for each enzyme, with numbers referring to the distance between restriction sites in kilobases. The size of the repeat array (estimated as the distance between the most proximal restriction sites) is provided in kilobases. The fragments subcloned for sequence analysis are identified. As described in Figure 3.8, the precise order of restriction sites could not always be determined and whilst the sum (denoted  $\Sigma$ ) of the 5' and 3' distances between outer and inner nested restriction sites can be estimated, each specific 5' and 3' distance could not.

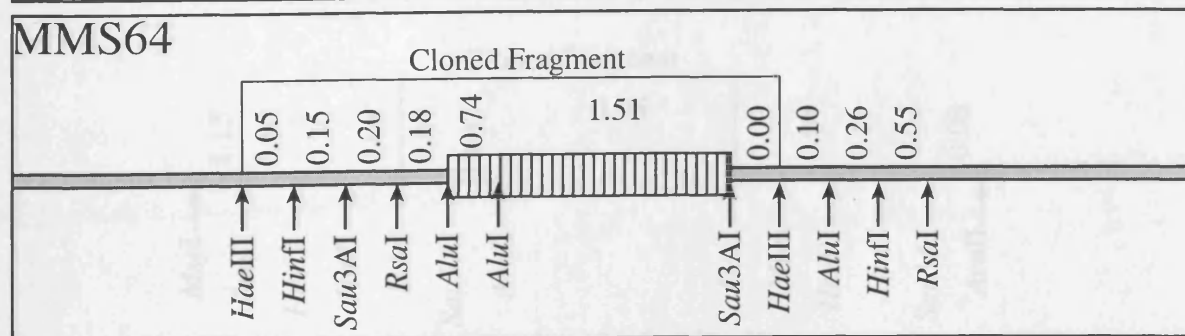
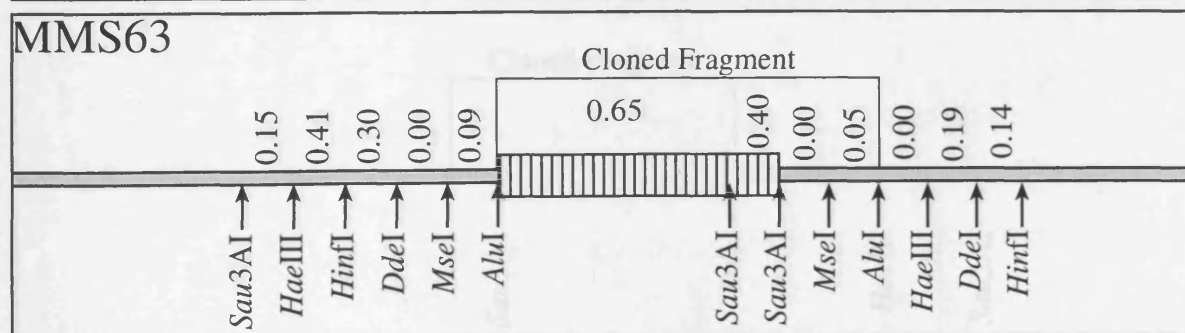
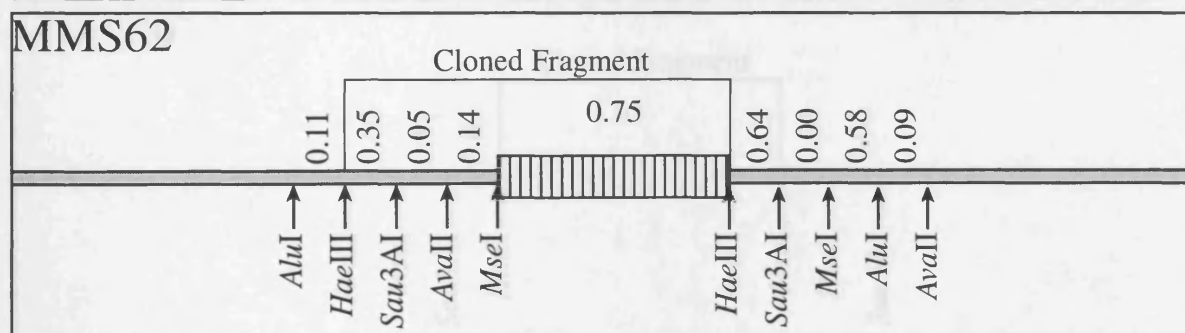
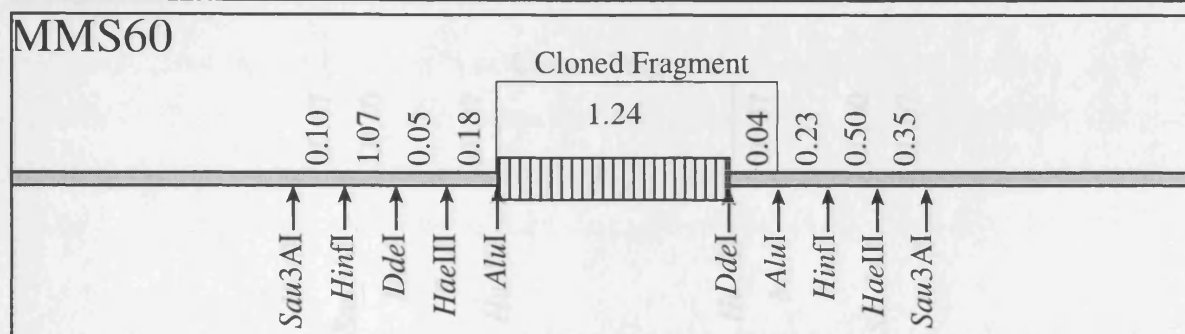
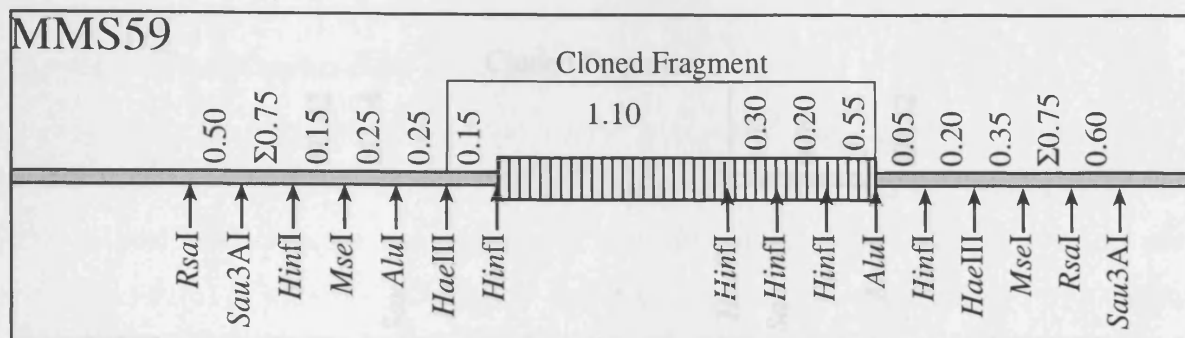


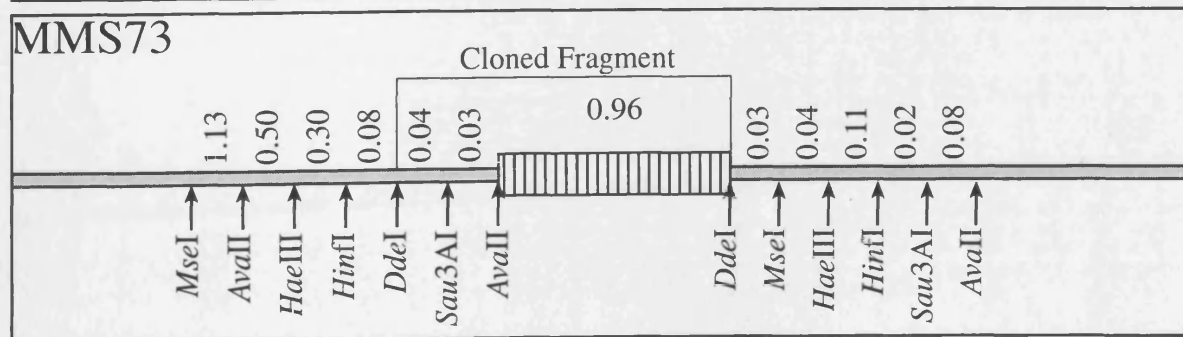
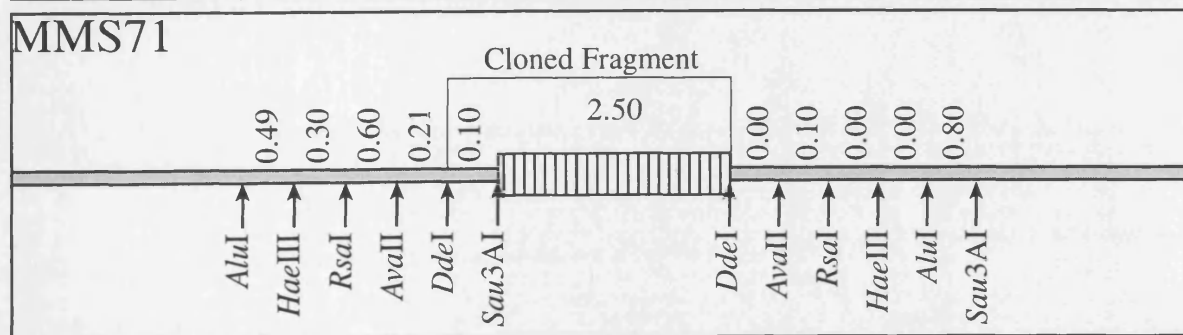
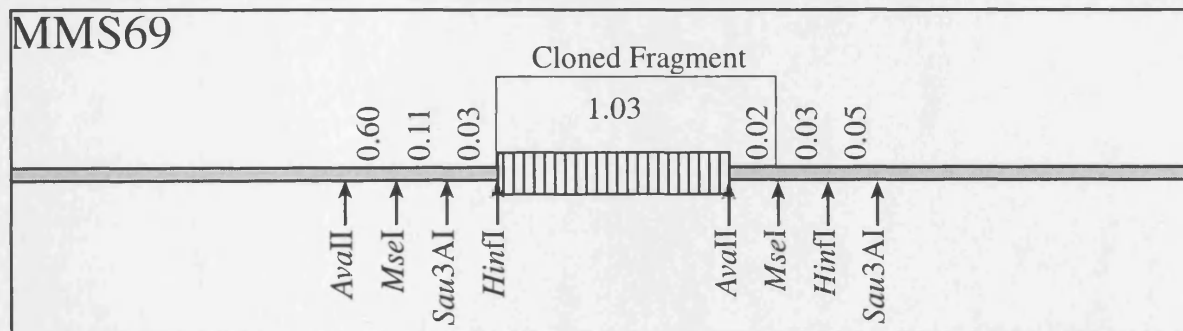
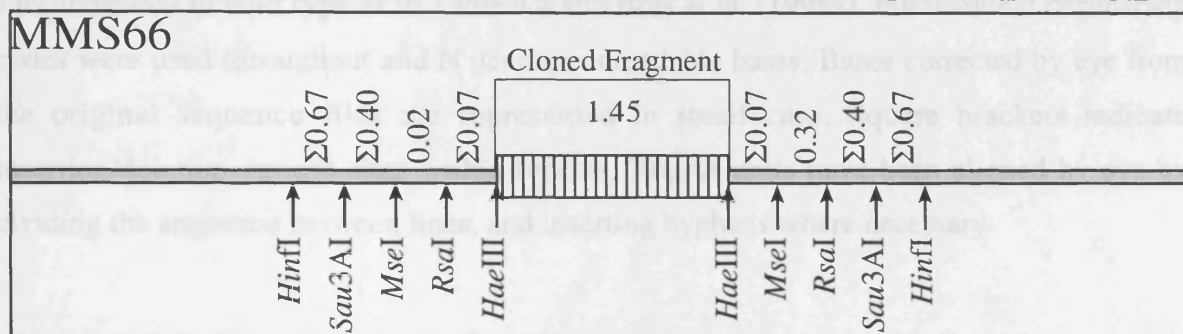
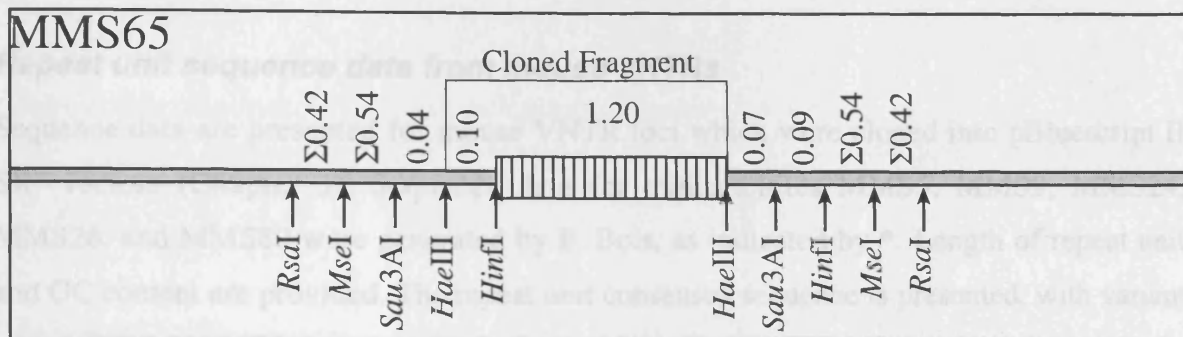












## Appendix 2

### ***Repeat unit sequence data from mouse VNTRs***

Sequence data are presented for mouse VNTR loci which were cloned into pBluescript II SK<sup>+</sup> vectors (Chapter 3). Sequence data for minisatellites MMS5, MMS9, MMS24, MMS26, and MMS80 were generated by P. Bois, as indicated by \*. Length of repeat unit and GC content are provided. The repeat unit consensus sequence is presented, with variant sites indicated in bold type as in Table 3.4 and Bois *et al.* (1998a). International degeneracy codes were used throughout and N denotes unreadable bases. Bases corrected by eye from the original sequence files are represented in small case. Square brackets indicate insertion/deletion variant sites within repeats. Repeat units have been aligned by eye by dividing the sequence between lines, and inserting hyphens where necessary.

## MMS5\*

Repeat unit length 24 bp

%GC of repeat 63%

Repeat unit consensus: CTTARGRTCYGTGGGCAGGCTCAV

AGGCCACACACCTAGAGACTGTATTCAAACCTCCATTTCTGCCACTATACCNTTAGCTTCAGATCAAGAGGTTTGAGAAAGGGCA  
ACSAGTGTCAATTATGAGGAACTGTGCAGCTCTTCGGTACTATAAACACTGTGAGGCTCAA  
CTTAGGGTCTGTGGGCAGGCTCAG  
CTTAAGATCCGTGGGCAGGCTCAC  
CTTAAGATCTGTGGGCATGCTCAG  
CTTAGGATCTGTGGGCAGGCTCCA

## MMS9\*

Repeat unit length 20 bp

%GC of repeat 65%

Repeat unit consensus: GGRGYAGGGTASGAGAGTGA

GAATGAATGGCAGAGCTGAACTTGAACCCCAAGTTCAGAGCTGTGAAGCATCACACTAGGCAGGATTGAGAATTTCTGTGCTGT  
TTATATGTGTTCAGGGAGAACTAT  
--GGCAGGGTACGAGAGTGA  
GGGGCAGAGTACGAGAGTAA  
GGGGCAGGGTGGGACACTGA  
GGGGCACGGTACGACAATGA  
GGGGCAGGGTAyGAYaATGA  
CGGGCAGGGTAwGACAmTsA  
CGGGCAGGGTTTCGACACTCA  
GGGGCACGGTAyGACACTsAC  
ggAGcAsGGTAyGAGAGTGA  
GGGGCAGGGTACGAgAGTGA  
GGAGTAGGGTAGGAgAgTGA  
GGGGCAGGGTAGGAsAsTGA  
GGAsTAsGGTACGAsAGTGA  
gGGGCAGGGTACGAgAsTsA  
NNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNN  
GGAsTAgGGTaggAsAgTgA  
GGAGTAGGGtACGAsAGTGA  
GGAsTAGGGtAgGAgAgTGA  
GGGGCagGGTAgGAgAgTGA  
GGAgTAsGGTAGGAgAgTgA  
GGGGCggGGTACGAgAgTGA  
GGGGCAsGGTACGAgAgTGA  
CGAGTACGGTNCGAgAgTGA  
GGAGTAGGGTAGGAGAGTGA  
GGAGTAGGGTAGGAGAGTGA  
GGAGTAGGGTAGG  
GGAGCGAGGGCAGGGTAGGGAGAGGTGAGGAGAGAAGACATAGACCAGCACCTTTTTCTGTGTTTCCACAAAGCAGTAGCAATG  
AGGCAGGATTAACACAGCCTAAAGAATGGCTAGTTCGATGGGCAGTTGTGGCACATGCCTTTAATCCCAGCACTTGGGAGGTAGAG  
TAGGTGGATTTTTGAGTTTCGAGGCCAGCCTGGGCTACAGAGTGAAGTTCCAGGACAGCCAGGGCTACACAGAGAAACCTGTCTC  
GAAAAACAAACAAACAAAAAACCATAACAAACAAACAAAAAGAATGGTAGTTCAAATAGTTTACCTGACTCTGAGGAA  
CAGGGTTACCCCATTTGCCTAGTACTTGACCCTGATGTGCATAGGACAGATGGAAAGCCCAAGCCCAAGGGGAGACTGGCCT  
GCAGAATTTGCATTGATTGGTCCGCTTGGTCCACATATTAAGGTCCTTGTACAGGCATATTCTTTACTAGCCCCATGAATTGA  
CTGGTCTTCAGAAGGGCATCTCTCAGGGTCAAGCAAGGCCCAAAGTGCACAGCCTTTGGGACACAAATACCATGCAGTCAGAG  
AAGTCCCAGGAAGTTTGTGCAAGGCTACTTTCAAAGAACAGGAGTCAACAGGGTCAAAGATGCTGATTTGTCCAsGCTAATCAG  
GACTGAACTCArCCACTTTATTTAGGAACACAGTAAATAGGCTTGCAGTCAACACAAGGAAGTTTAAAGTGGAGAAAGCCTCCT  
CTCACACAGGAAAGACTTCATCTCCTCAAAGTGGCCCATGCTAGCTACCCACCACATCCACAGGGCACTCTTGAAAGAGCAG  
GGAAAGAAACAGATACGTATTGAATAACTTTACTGTATTAGACCAAAGGTAGGAACCTATAAGCTACCCAAGAGGATGAATAT  
TCATGTGGATAACACACACCAGGTCCCCTTAACAAAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAAATAA  
TTGCATGTCTCTGTATCCCAGCATTTGCAATCCTTAACAAGCCTTTAATAGCATGGATGTTACCTGCTTGTAACTGGAAGGC  
TAGCGTTTTATGTATTCAAGGGTCAAGCACAGACTGGAGCCAGCCTCTTCAAAGTGTGAGGAGAGAAGCAACATACGTAGCT  
GAGCCAGGCTTGGCTTTGAAAAGCATCGACTCTCAATCAGGTTCTTATCAGGAAGAACATCATAAGTCAAGGATGGAATTA  
GGTTGTCTTTCTTGTACTGTAAATTTCTCAAAGTTTACTATTCTCTCTCCTGTCCATAGGTTTTTTTTTTTTTTTTTTTT  
TAAAGAGTGTCAAGGAAAGTCCATTCTGGATAACTTAATTTGGGCTTCTGTTnGAATACGGTACTACAAGTGGCCTCTTCCCT  
CGGTACCCCTGACAACCAACTCAATCCTTTCAAATTATCTCTTTGGGGGGAATAAAGACTATTTACAGAGAGTTGGACATTTG  
TTTCTCAAACAGGGCAAGAAGCTTGCACAGCACCACCTCTcTtTcTGCTGGGGACACTGTGGACAACATTCTTTTTTTAAATTA  
TAAATCCCTTTATTATTTTTCCTCTTTAAAAATACATTGTAGAGAACAGATTGTGTTTATACAGGGACATTAAGTGGT  
TTTCAATATCAAAAAGGTTATTAAGACACCACAGGATGGAGTTTTAGCTCATAAATAGAAAACAAGAACTCAAAGAGCTTCAGA  
GGACATTTAGTTAGAAAGTCATAAGAGTGAAGGACATAAGGcTAAGAGAGGCACAGAAGAAGCGGTTTTCTTTTAAATTAAC  
AACAGTAATAGCAGTGATTTACCAACCACAAGCCTAAGACAGTCCAAGGACATCAGAGAACTTCATAGGCAAGGTAGGATGTGG  
GCTCCCCCTGGAGGGCTGGGCAGAACTTACAGGGGAAAAACAGGCAAAACAGGAAGCAAAGCAGAG

## MMS18

**Repeat unit length**                      14-18 bp

%GC of repeat                      60%

**Repeat unit consensus:** GGGTGACA[V][H]G<sub>1-4</sub>A[C]DGRT

**Sequenced with primer KS**

AGACAGCATGGACATCTACTGAGAGGGTGTCCAGGGATCTGCAGAAGCAAAACTAGGGAGGTGGCATGCGTGATATGGAGAGGTCA  
gGAGGGCTGGGGGAGGGTCACAgAGTGGAGTAAGTACTGATGATAGAGAGGGT  
GGGTAAACAC--GGG-A-GGAT  
GGGTGACAGT--GG--A-GGGA  
GGGTGACAGCAG---AT-GGT  
GTGTGATA---GGG-A-gGGT  
GGGTGACA---GGGA-GGAT  
GGGTGACAAA-G---A-gGGT  
GGGTGACAgT--GG--ACgGAT  
GGGTGACA---gGG-ACAGAT  
G--TGATA---gGG-A-gGGT  
GGGTAACA---GGGA-GGAT

**MMS24\***

Repeat unit length 28 bp

%GC of repeat 64%

**Repeat unit consensus:** TGTGAGCAC**R**TGMCTGCAGTGTCTGC**YV**

[illegible]

TGTGAGCACATGACTTCAATGTTTGCTA  
 GGTGAGCACATG-----  
 TGTGCGCTCACAAACCTCCGTCCGCACTGTGCCCCACAGCCCGCTTACGTTCTTGCCCCACGTCATCCTTGCTGCTCAGCAGTGCT  
 TCCAGCTCAGCCCTCAGAGCACGGTTCTGCCTCTCCAGCTCCTCCCGGGCCTCCTGCTCcTCTTCCAGGGCCCGGGTCAGCGACA  
 GGGCCC

## **MMS25**

Repeat unit length            17 bp

%GC of repeat                68%

Repeat unit consensus:      GGGTCCCT**M**CTCC**Y**CAT

## **Sequenced with primer KS**

GGTGGCgGCCGCTCTAGAACTAGTGGATCCCCGGGCTGCAGGAATTCGATCTCATTCTTCTcAGTAGTAAAGGTTGGGAGAGGAC  
 TGAGA  
 -----TGGCCAT  
 GGGTCCCTCCTCCCCAT  
 GAGTCCCTCCTCCCCAT  
 GGGTCCCTCCTCCTCAT  
 GGGTCCCTCCTCCCCAT  
 GGGTCCCTCCTCCCCAT  
 GGGTCCCTCCTCCCCAT  
 GGGTCCCTACTCCCCAT  
 GGGTCCCTCCTCCCCAT  
 GGGTCCCTCCTCCTCAT  
 GGGTCCCTcCTCCCCAT  
 GGGTCCCTCCTCCCCAT  
 GGGTCCCTCCTcCTcAT  
 GGGTCCCTCCTCCTCAT  
 GGGTCCCTCCTCCCCAT  
 GGGTcCCTCCTCCCCAT  
 GGGTCCCTcCTC-----





## MMS30

Repeat unit length            39 bp

%GC of repeat 44%

Repeat unit consensus: AGGAGATTC**MS**TTTCACTATACAGAAGATGGTGTCTCAGC

**Sequenced with primer SK**

NATAGGGCCAATGAGCGAGGCAGAATTAGAAGATTGGACATCTGTTAGAAAGACTGACATTCTGGGATAGATTTCATGCACAGCAG  
AAGATTCAATTGAAACTCTG---

-----CAGAAGATGGTGTCAAC  
AGAAGATTCCCTTCACACTATACAGAAGATGGTGTTCAGC  
AGGAGATTCACTTCACACTGTACAGAAGATGGTGTTCAGC  
AGGAGATTCACTTCACACTATACAGAAGATGGTGTTCAGC  
AGGAGATTCACTTCACACTATACAGAAGATGGTGTTCAGC  
AGGAGATTCACTTCACACTATACAGAAGATGGTGTTCAGC  
AGGAGATTCACTTCACACTATACAGAAGATGGTGTTCAGC  
AGGAAATTCCTTCACACTATACAGAAGATGGTGTTCAGC  
AGGAGATTCACTTCAGACTATACA

**MMS35**

Repeat unit length 28 bp

%GC of repeat 56%

**Repeat unit consensus:** GGCCATGCCAGTGGTCCTTTCACWCTCA

**Sequenced with primer SK**

GTCATGTGTAACCTGTAGGGGCAAAATACATTTATTCC

-----TTCACACTCA  
 GGCCATGCCAGTGGTCCCTTTCACACTCA  
 GGCCATGCCAGTGGTCCCTTTCACTCTCA  
 GGCCATGCCAGTGGTCCCTTTCACACTCA  
 GGCCATGCCAGTGGTCCCTTTCACACTCA  
 GGCCATGCCAGTGGTCCCTTTCACTCTCA  
 GGCCATGCCAGTGGTCCCTTTCACACTCA  
 GGCCATGCCAGTGGTCCCTTTCACTCTCA  
 GgCCATGCCAGTGGTCCCTTTCACTCTCA  
 GGCCATGCCAGTGGTCCCTTTCACTCTCA  
 GGCCATGCCAGTGGTCCCTTTCACACTCA  
 GGCCATGCCAGTGGTCCCTTTCACTCTCA  
 GGCCATGCCAGTGGTCCCTTTCACTCTCA  
 GGCCATGC

**MMS38**

Repeat unit length 47 bp

%GC of repeat 62%

**Repeat unit consensus:** GGGGATTCCACAGGG**K**GCCTGTGGTCCAGCACCTGGACAACATGGCT

**Sequenced with primer KS**

AAAGCCTGCCTGCCTCTGTAAAAACACTCATATTTACAGCATAATCCCTCGGAGGCCCTTCTGTCTCTCCAGGCCAGTCCCCATC  
TGTATGCGACATCGTGAGAGAACCACCGAGGAGGAGTGGAGGTGGAACAGAGCACATCGGGCTGCCCTCGTGCAGGAGGAAAGCAA  
AGCGAGGGCGCGCGTAGCCCTGCACGTCTACCGAGCATCTGGAGGGCCATGGGCAGCGGCCTGGCCGCGCCCATCTCTGTCTCTC  
TGCTCTCATGCAAGTGTAGGTGATCACCAGAAAGTTCCTTACGCTACCGACCTGCACC-----

-----GAAT  
CCCCAGCCATGTTGTCCAGGTGCTGGACCACAGGCACCTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGCTG-----

**Antisense of sequence from primer SK**

-----GGCCCCNTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGCTGGACCACAGGCACCCTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGCTGGACCACAGGCACCCTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGcTGGACCACAGGCACCCTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGCTGGACCACAGGCACCCTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGCTGGACCACAGGCCCCCCTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGCTGGACCACAGGCACCCTGTGGAAT  
CCCCAGCCATGTTGTCCAGGTGCTGGACCACAGGCACCCTGTGGAAT  
CCCCA-----  
CAGGGCCCTGAACCTTCCCAAGCT

## MMS39

Repeat unit length 30 bp

%GC of repeat 58%

Repeat unit consensus: GATGTYSCWGTGYGTGCTCCACCTCCTGT

### Sequenced with primer KS

CTAATTTACTTCAGGATTAGGGTAAGCAGCGGGACCACAGGAGATGGA  
---AGCATAACACAGGGACATCACAGGAGGT  
GGGAGCACACACAGGGACATCACAGGAGGT  
GGGAGCACACACAGGGACATCACAGGAGGT  
GGGAGCACACACAGGGACATCACAGGAGGT  
GGGAGCACACACTGCAACATCACAGGAGAT  
GGGAGAACACACAGAGACATCACAGGAGGT  
AGGAGCACGCACAGCAACATCACAGAAGGT  
GGGAGCACACACAGGGACATCACAGGAGGT  
GGGAGCACACACAGGGACATCACAGGAGGT  
GGGAGCACACACAGAGACATCACAGGAGGT  
GGGAGCACACACAGGGACATCACAGGAGGT  
GGGAGCACACACAGGGACATC-----

### Antisense of sequence from primer SK

-----ACACACTGCAACATCACAGGAGGT  
GGGAGCACACACTGCAACATCACAGGAGGT  
GGGAGCACACACAGGGACATCACAGGAGGT  
GGGAGCATGCACAGCGACATCACAGGAGGT  
GGGAGCACGTACAGGGACATCACAGGAGGC  
GGGAGCACACACAG-----  
CAATGGGCGAGGGACAAGAATAGGCTCTGGTCCCTCCCTCCTCACCATCTGGAGAAAGCGACGGATGCGCTGGGCATCCTGGGCC  
CGGAACACATGCCACACAGTGCTGGTCTGGTCCCTGGAGACCAGAGTCTTCCCATCCAGGCCTGAGAGGAAATCTGCAGGAG  
AAAAATGGCTCAGGTGAGCCCAGGGTAGGCAGAGGCAGACATCAGCGACTGCTTGACTGGCCAAGAACCACCTTTCTGTGCTCG  
ATACCA

## MMS41

Repeat unit length 42 bp

%GC of repeat 71%

Repeat unit consensus: GGCCACACCCAGGGGCTGACTCCSAGGAGCAGGCTGGGAGCA

### Sequenced with primer SK

TGTTAGCCTGCGCCACACCACCACCAGCCAGCCAGAACAGCAGCCACCTACCTCTGAGGGGCCAGAACTCTCCATCCGGGGCA  
GCAACTCTTGATAAGACAGGGGAGAGAGAACGAATAAACAAGCGGGGCGAG  
-----CGAGGAGCAGGCTGGGAGCA  
GGCCACACACCAGGGGCTGACTCCCAGGAGCAGGCTGGGAGCA  
GGCCACACACCAGGGGCTGACTCCCAGGAGCAGGCTGGGAGCA  
GGCCACACACCAGGGGCTGACTCCCAGGAGCAGGCTGGGAGCA  
GGCCACACACCAGGGGCTGACTCCCAGGAGCAGGCTGGGAGCA  
GGCCACACACCAGGGGCTGACTCCCAGGAGCAGGCTGGGAGCA  
GGCCACACACCAGGGGCTGACTCCCAGGAGCAGGCTGGGAGCA

## MMS44

Repeat unit length 39 bp

%GC of repeat 43%

Repeat unit consensus: CCTGCTGASASCATCTTCTGTATAGTGTGAASKGAATCT

### Sequenced with primer SK

GTTTATTATATTGTAAGCCCTGCCACATGGTTAGTTACCTCTCCTTAGTCTCCTAAG  
-----ACCATCTTCTGTATAGTGTGAAGGGAATCT  
CCTGCTGAAACCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGAGAGCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGACACCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGACACCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGACACCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGACANCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGACACCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGAGAGCATCTTCTGTATAGTGTGAAGTGAATCT  
CCTGCTGACACCATCTTCTGTATAGTGTGAANGGAATCT

## MMS45

Repeat unit length 40 bp

%GC of repeat 44%

Repeat unit consensus: GGGTAGGGT**R**GAGATACTCAGTTGTTACACTGTCATCTAA

### Sequenced with primer SK

-----TCAA  
GGGTANGGTACAGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAGGGTAGAGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAGAGTGGAGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAgGGTA**g**AGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAGGGTGGAGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAGGGTGGAGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAGGGTGGAGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAGGGTGGAGATACTCAGTTGTTACACTGTCATCTAA  
GGGTAGGGTGGAGATACTCAGTTGTTACACTGTCATCTAA

## MMS46

Repeat unit length 18-19 bp

%GC of repeat 53%

Repeat unit consensus: GG[G]AT**R**AGAACATAGGTGT

### Sequenced with primer KS

CCAAGGACTGGGGAGCCCAGTGCCATTCCAGGGCAGGACACAGTCCTGTTGTGTTTCATGCCTGACCTCTG**c**ACTCTGTTCCACAT  
GTGACACTGAGTGGACTGACTGGTCTTG**C**ATAAAGACCCA  
-----GTGT  
GGGATGAGAACATAGGTGT  
GG-ATGAGAGCACAGGTGT  
GGGATGAGAGCACAGGTGT  
GG-ATGAGAGCACAGGTGT  
GGGATAAGAGCACAgGTGT  
GGGATAAGAGCACAGGT**g**T  
GGGATGAGAGCACAGGTGT  
GGGATAAGAGCACAGGTGT  
GGGATAAGAGCACAGGTGT  
GG-ATGATAGCACAgGTAT  
GGGATA-----

## MMS47

Repeat unit length 38 bp

%GC of repeat 46%

Repeat unit consensus: TTTY**C**TGACCTAGCTTACCTTTGGTGTTAGAGCGTGTG

### Sequenced with primer SK

CAGTTATAATCATAACATACTGTTTTTTCTTACTCCACACAGGCATAGAGTGTCTGCTATTAATAACTATGCTCAAAAATTGTGT  
ACCTTTTAGTCAGAGATGCCTCTGCTGCCTGAATGCTGGGATCAAAGGCACGTGCCACCACACCTGGCTATTTTCACTTTATAAGT  
GGTTACCCATAATGACACAGACCACTTAAGTCAGGAAACGCCTGTTATGTTTCGCTTCTCACCTCCATAGACAAATGCTGTTGGAC  
ACAGCTGTCTTTTACTTCCTGGTATTTGACTTGGCTCTGCCCTCCTA**A**CCCCACAAAGTACTGCAAATGTTGCACTATTTCTTGA  
GGAGAACAGTAACACTAACCACC-----  
-----TGACCTAGCTTACCTTTGGTGTTAGAGCGTGTG  
TTTTCTGACCTAGCTTACCTTTGGTGTTAGAGCGTGTG  
TTTTCTGACCTAGCTTACCTTTGGTGTTAGAGCGTGTG  
TTTCCTGNCCTANCTTACCTTTGGTGTTAGAGCGTGTG

**MMS48**

**Repeat unit length**                      19 bp

%GC of repeat 68%

**Repeat unit consensus:**      **GGC**S**AGGG**A**NG**R**SAGCAG**

**Sequenced with primer KS**

ACTACAGA-----GCCAGa  
 -----GCCAGa  
 GGGAGGAGAgCagGGCACA  
 GGGAGGACAcCagGGCAGa  
 gGGATGaCAGCAgGGCACA  
 GGGAAgGgCAGcAGGCAGa  
 GGgAcGgcAGCAGGgCAGa  
 gGGAcGgGAGCAgG-----

**MMS49**

Repeat unit length      Degenerate

%GC of repeat 58%

Repeat unit consensus: C1-4 (TS) 1-15

**Sequenced with primer SK**

[illegible]

## MMS52

Repeat unit length            Degenerate  
%GC of repeat                57%  
Repeat unit consensus:      C<sub>1-4</sub>T<sub>1-2</sub>CA[T][YT]

### Sequenced with primer SK

GATTACAGGCGTGAACCTACCACACCTGGTTTATGGAGCACTGAGGACCAAGCCCAGGGTTTGCCACATGCACGCTACCAAGTGAG  
CCACAGCCCCAGAACTCTTGATACTTCACGGGATGTAAATGAAGGAGTCATGGTGGATCAGAAAGGGCTTAAGCCT  
CC--T-C-A  
CCCCT-C-A  
CCC-T-C-A  
CCC-TTC-AT  
CCC-T-C-AT  
CCC-T-C-A  
CC--TTC-AT  
CCC-TT--A  
CCC-T-C-A-G  
CCC-T-C-ATCT  
C---T-C-A-CT  
C---T-C-A  
CCC-T-CCA  
CCC-T-----CT  
CC--TTC-AACT  
C---T-C-A  
CC--T-CCAT-T  
CC--T-C-A  
CC--TTC-AT  
CCC-T-C-A  
CCC-T-C-AT  
CCC-T-C-AT  
C---TTC-ATCT  
C---T-C-A  
CCC-T-C-A  
CCCCT-C-AT  
CCC-T-C-AT  
CCC-T-C-A  
CCC-T-C-A-CT  
CC--T-C-A  
CC--TTC-AT  
CCC-T-C-A  
CCC-T-C-A-CT  
CC--T-C-AN  
CC--T-C-AT  
CCC-T-C-A  
CCC-T-C-AT  
CCG-T-C-AT

## MMS54

Repeat unit length            15 + 20 bp  
%GC of repeat                56%  
Repeat unit consensus:      ARGATGCTGTGTGAGYACAGC

### Sequenced with primer KS

-----C  
AGGATGCTGGTGAGCACAGC  
AGGA-----GTGAGCACAGC  
AGGA-----GTGAGCACAGC  
AGGATGCTGGTGAGTACAGC  
AGGATGCTGGTGAGTACAGC  
AGGATGCTGGTGAGTACAGC  
AAGA-----GTGAGCACAGC  
AnGATGCTGGTGAGTACAgC  
AGGATGCTGGTGAGCACAAAC  
AAGA-----GTGAGCACAGC  
AGGATGCTGGTgAGCACAGC  
AAGA-----GTGAGCACAgC  
Ag-----

**MMS55**

**Repeat unit length**                      24-44 bp

**%GC of repeat**                      **46%**

**Repeat unit consensus:** GGKAGGGGCASATKCTGAG (TG) 2-4 [T] ACATG (T) 3-5GCATGAT

**Sequenced with primer KS**

TATGCCCCGTGGAATCTCACTCATGACAACACC-----  
-----TGTGGTTG-TATGT---CTTCA-TGT---ACACT-----  
-----TGTGT---ACACA-TGTGCTTA---CATGAT  
GGTGAGGGCAGATGCTGAG-TGTGTGT---ACA-TGTG-TTTT---GCATGAT  
GGGGAGGGCACATTC--AG-TGTGTGTGT--ACG-TGT--TT-----  
-----G-TGTGTGTGT--ACA-TGTG---CAATGCATGAT  
GGGGAGGGCaCATGTCTATA-TGTATGTG-----

**MMS57**

Repeat unit length 24 bp

%GC of repeat 50%

**Repeat unit consensus:** GCTGTGTAGACAGAGCAGTAGAGT

**Sequenced with primer KS**

[illegible]

**Antisense of sequence from primer SK**

[illegible]



## **MMS60**

Repeat unit length            18-22 bp

%GC of repeat                60%

Repeat unit consensus:       CCTCC [TCC] [A] TGTGCTCCY [Y] TGT [TCT]

### **Sequenced with primer SK**

```
AAGTAAACCTAGGAAAGAGTTGCTGCACACTCTTCTGTCCT
CCTCC---ATGTGCTCCC--TGT
CCTCC---TGTGCTCCCC-TGTTCT
CCTCC---ATGTGCTCCCT-TGT
CCTCCTCCATGTGCTCCCC-TGT
CCTCC---ATGTGCTCCC--TGT
CCTCC---ATGTGCTCCTT-TGT
CCTCC---TGTGCTCCCT-TGT
CCTCC---ATGTGCTCCCT-TGT
CCTCCTCCATGTGCTCCCC-TGT
CCTCC---ATGTGCTCCCT-TGT
CCTCCTCCATGTGCTCCCT-TGT
CCTCC---ATGTGCTCCCC-TGTTCT
CCTCC---ATGTGCTCCC--TGT
CCTCC---ATGTGCTCCc--TGT
CCTCC---ATGTGCTCCTT-TGT
CCTCC---TGTGCTCCCT-TGT
```

## **MMS63**

Repeat unit length            31 bp

%GC of repeat                52%

Repeat unit consensus:       TCCCCAGTCTGACCTC**R**TAGTCTATCTGTCC

### **Sequenced with primer SK**

```
-----TAGTCTATCTGTCC
TCCCCAGTCTGACCTCATAGTCTATCTGTCC
TCCCCAGTCTGACCTCATAGTCTATCTGTCC
TCCCCAGTCTGACCTCATAGTCTATCTGTCC
TCCCCAGTCTGATCTnATAGTCTATCTGTCC
TCCCCAGTCTGACCTCGTAGTCTATCTGTCC
TCCCCAGTCTGACCTCGTAGTCTATCTGTCC
TCCCCAGTCTGACCTCGTAGTCTATCTGTCC
TCCCCAGTCTGACCTCGTAGTCTATCTGTCC
TCCCCAGTCTGACCTCGTAGTCTATCTGTCC
TCCCCAGTCTGACCTCATAGTCTATC-----
```

### **Antisense of sequence from primer KS**

```
-----TCTATCTGTCC
TCcCCAGTCTGACCTCATAGtCTAtcTGTCC
TCcCCAGWMTGACCTCATAGTCTAtcTGTCS
TCcCCAGTCTGACCTCATRgTcgATcTGTCC
TCcCCAGTcTGACcTcaTaGacTATCTGTCC
TCCCCAGTaTGACCTCATAGTCTATCTGTCC
TCCCCAGTgWGACCTCATAGTCTATCTGTCC
TCCCCAGTCTGACCTCATAGTCTATCTGTGC
TCC-----
ATAGTCTGACCCYWTWKTCTATCTGTcSTCCSTAGTCTG
```



## **MMS65**

Repeat unit length            Degenerate  
%GC of repeat                62%  
Repeat unit consensus:      GC-rich STR

### **Sequenced with primer KS**

```
GACAGATCAAAGAAACACCAGAGTTCCAAATTCTCTGGGGCTTACT
CCG
CCCCG
CC---T
CG
CCCCCT
CC---T
CC---T
C---T
CCCCCT
CCCC-TGACTGAGCGGCACCAGCTTT
CC---TGA
-----GTCT
-----GT
-----GTGCTTCTCAGGCTCTACAGaCATGTAA
CCCC-T
CCCCCATCACACACA
CCGC-TGCATT
CC---TACACAG
CC---TCCATTGGTCCATT
CC---TCAGGAGAA
CCC--TGGCTAATGCAGAAGGACGGTCCATGCAG
CCCC-TGAAA
-----CTG
-----CAGG
-----CAG
-----CAGCAACTGCACATGCACTCGCCAT
CC--TCCAG
CC--TCCGCCGCTCA
-----GGGGCT
-----GGGAGACA
-----GG---AGA
----TGT
----GT
----GTGATGCGTGGAG
CC--TG
-----GT
-----GT
-----GTGCAG
-----GT
-----GTG
CC--TG
-----GTGCT
-----GTGCACAT
-----GT
-----GT
-----GTG
-----GTGCTG
```

## **MMS69**

Repeat unit length            41 bp  
%GC of repeat                53%  
Repeat unit consensus:      GGGTRCCAGCATYCCCAGCTCTATCTGAGCACACTCTCTAT

### **Sequenced with primer SK**

```
CACTCTGGGGATGGCTATCCTTTAGTGGACCTGAGTGT---
-----TCTATCTGAGCACACTCTCTAT
GGGTACCAGCATCCCCAGCTCTATCTGAGCACACTCTCTAT
GGGTACCAGCATTCCCAGCTCTATCTGAGCACACTCTCTAT
GGGTGCCAGCATCCCCAGCTCTATCTGAGCACACTCTCTAT
GGGTACCAGCATCCCCAGCTCTATCTGAGCACACTCTCTAT
GGGTACCAGCATTCCCAGCTCTATCTGAGCACACTCTCTAT
GGGTGCCAGCATCCCCAGCTCTATCTGAGCACACTCTCTAT
GGGTACCAGCATCCCCAGCTCTATCTGAGCACACTCTCTAT
GGGTACCAGCATCCTCAGCTCTATCTGAGCACACTCTCTAT
GGGTACCAGCATTCCCAGCTCTATCTGAGCAC-----
```

### **MMS71**

Repeat unit length            20 bp

%GC of repeat                44%

Repeat unit consensus:        **CYKGC<sup>T</sup>RTAGATGWTGACTT**

#### **Sequenced with primer KS**

```
CATATAGTCTCCTCATCAAAATGCCCGAC
CCTACTGTAAA-----TGATGACTT
-----TGATGACTT
CCTGCTATAGATGATGACTT
CCTGCTATAGATGATGACTT
CCTGCTGTAGATGATGACTT
CCTGCTGTAGATGTTGACTT
CTGGCTGTAGATGTTGACTT
CTGGCTGTAGATGATGACTT
CCTGCTGTAGATGATGACTT
CCTGCTGTAGATGTTGACTT
CTGGCTGTAGATGTTGACTT
CTGGCTGTAGA-----
```

### **MMS73**

Repeat unit length            29-39 bp

%GC of repeat                47%

Repeat unit consensus:        **(TG)<sub>3</sub>TA(TG)<sub>4-8</sub>[C][A][T][G]CACTATASCCY**

#### **Sequenced with primer SK**

```
TAGAACTATACCTATTTATGTATGCATGTGTGTGTACTATAGCTACATGTA
TGTATG-CA-TGTGTGTG-----CATGTACTATAGCCC
TGTGTG-TA-TGTGTGTGTGTG-----CACTATAGCCC
TGTGTG-TA-TGTGTGTGCATG-----CACTATAGCCC
TGTGTG-TA-TGTGTGTGTG-----CACTATAGCCT
TTTGTG-TA-TGTGTGTG-----CATCCACTATAGCCC
TGTGTG-TA-TGTGTGTGTGTG-----CATGCACTATAGCCC
TGTGTG-TA-TGTGTGTGTGTGTG-----CATGCACTGTAGCCC
TGTGTA-TA-TGTGTGTGTGTGT-----ATGnACTAtAGCCC
TGTGTG-TA-TGTGTGTGTGTGTGTG-CATGCacTATACCCt
TGTGTG-TA-TGTGTGTGTGTGTGTG-----CAcTANACCCN
TGTGTG-TA-TGTGTGNG-----CACGCAATATANCCC-
TGTGTG-TA-TGTGTGTGTGTG-----CACT-----
```

Repeat unit length	40 bp
%GC of repeat	53%

Repeat unit consensus: CCAGCCCATGGGACAGACTGTA **[TA]** RCTAGGTCAGTCCT

GATCATACAAAACATTCTTTCTGATATACTCAATAGCAGAAAATATTTAGTGTTTACAATAGTTTTGGCACTCCGTATTTGAACC  
AGATTTTTTAAATTTGTTTGTGATCTGGAAGTAGACAATTGTAGAGAAATGTAGATGTTTAAAGCAATGGTAGTACTCAGACTGGCA  
CGATTTAGAGTCAAGATCTGTTTAAATCTATTGGGAATGTAGCCAGGAATGTTTCCAGGAATGTTTCCAGCTCAGTTTATCCCTACACTGAAGAC  
GGAAAGACTGAGAGTAACATGTTAAGGCGAGCATAGCCACGAGCACCCTGCTGTTTCATAAGTAGGTTGTGCTGCTTACTCTACCAAAA  
CAAATCTCTAGGGGCTGGAAGTCAAGGAAGTATGCTGTGTCTAGCAGATAGGCTGGTTCTGAAGGAGTTTTTAAAAACAGAGAC  
AGGTTTAAAGCTGTAACACAGGAGCTGTGTAGACACTGAGTACTGTAAATTAGGACCTTAATGTGCTTTTAAAGTTCATAAGTACAAG  
GAAAAATTTAAATTTGCAAGAAGAAATGAATGGAAATGTACACAAGTTAAAGGCAACATACACAGGGTTCAGGAGACCCAGTT  
TGTTAAGAGCACTGCAGTTACTTTTAGAAACAGAATAAATCAAGTTTAGCCTTACCTCTTTGAGTTAGAATTTTTCATCTGCCCTCT  
ATTCTATCTGTCTAGACATCTATCAGCTGCCACCTGCCCTCTGCTTTGCAGCGCTGGGAAGTGTCAACTGTGACCTCATGTTGCTAG  
GTACATGTTCTGTCACTAAGCAGTTTCCAGTCCCATCTTCTCTTAGAAAATGCATTGTGTTCATTGACTTGTGTAAAGTAGAG  
CTTCT

-----TGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTAG--CACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGTCTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTAGCNNNNNNNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN  
NNNNNNNNNNNNNNNNNNNNNTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTAG--CACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGTGACACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGTCTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGTCTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTAG--CACTAGGTCAGTCCT  
CCAGCCCATGGGATAGACTGTAG--CACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCCATGGGACAGACTGTATAACACTAGGTCAGTCCT  
CCAGCCC-----  
GTCCTGTGTTTGCCTTCCTCACTCCAGGGCTTGTTGGCGATAAGGCTGCTGCTGCTCTGGACCAGACTCTTAATTCTCTGTAC  
CTTGAGACGTTGCTTTGATGATTAGGTTTTGTGTGCTTTTGTGTTTAGAAAAATAAAGCAGAAGATTGAGTTTCTACATCCCACTC  
CTCCAGAAAAGATC

## Appendix 3

### ***Sequences flanking two mouse minisatellites***

MMS57 and MMS58 were identified from cosmid clones derived from a BALB/c mouse. Various regions of each cosmid containing 5' or 3' DNA flanking and including the repeat array were subcloned into a pBluescript II SK<sup>+</sup> vector and sequenced. Known sequences surrounding and including each minisatellite are presented for MMS57 (Appendix 3a) and MMS58 (Appendix 3b). Flanking primers are underlined. The repeat region for each locus is represented in grey with sequences of each repeat to which MVR primers anneal underlined. Repeat unit sequences polymorphic between repeats and detected by MVR-PCR are double underlined, and represent the 3' base for each MVR primer.

## a: Sequence of MMS57 clone

```

1  ATCTGAGACTTCAGTTGCCGTACCTTCTTACCCAGCACATGCGCGGTAGAAGGGTGGGA
   -----+-----+-----+-----+-----+-----+-----+
61 TAGACTCTGAAGTCAACGGCATGGAAGAATGGGGTCGTGTACGCGCCATCTTCCCACCTT
   -----+-----+-----+-----+-----+-----+-----+
121 TGGCTCTTCAGCAATGCAGCTGCCCTCCCTCCCAGGTGCTTGTGTTCCTTAACACACCC
   -----+-----+-----+-----+-----+-----+-----+
181 ACCGAGAAGTCGTTACGTCGACGGAGGGAGGGTCCACGAACACAAGGGGATTGAGTGGGG
   -----+-----+-----+-----+-----+-----+-----+
241 AGGAAACTCCCTGGCTCACCAAGCCACCTTGAGTGGGGTCTTCCCTCTGGTTTGTCAAGG
   -----+-----+-----+-----+-----+-----+-----+
301 TCCTTTGAGGGACCGAGTGGTTCGGTGAACTCACCCAGAAGGGAGACCAACAGTTCC
   -----+-----+-----+-----+-----+-----+-----+
361 GTGTGTCTGCATCTGGGATGAATATGTTTCCCTCCACCAACTTATGGACAGACAACATGA
   -----+-----+-----+-----+-----+-----+-----+
421 CACACAGACGTAGACCTTACTTATACAAAGGAAGGTGGTTGAATACCTGTCTGTGTACT
   -----+-----+-----+-----+-----+-----+-----+
481 TTCTTACTTAGAAAAATCAAAACAAGAATCCCTCCCTGGAAGTTGTGTGTCCCGTTGACGT
   -----+-----+-----+-----+-----+-----+-----+
541 AAGGATGAATCTTTTAGTTTGTCTTAGGGAGGGACCTTCAACACACAGGGCAACTGCA
   -----+-----+-----+-----+-----+-----+-----+
601 GCCTCATTCATCACCAGAAAGGTCATGTTTCATCTTCACTGAGCTAAAGCATGGGTGCTTG
   -----+-----+-----+-----+-----+-----+-----+
661 CGGAGTAAGTAGTGGTCTTTCCAGTACAAGTAGAAGTGACTCGATTTCGTACCCACGAAC
   -----+-----+-----+-----+-----+-----+-----+
721 GATGCTGTCTATGTTATTCAAGAGTACAGCCAAGGCACCAGATTACCAACCCCCATTACAA
   -----+-----+-----+-----+-----+-----+-----+
781 CTACGACAGTACAATAAGTTCTCATGTCCGTTCCGTGGTCTAAGTGGTGGGGTAATGTT
   -----+-----+-----+-----+-----+-----+-----+
841 CTGTGTGTTACTGAGACCACTGGATAAGTCACCAGCACCTCCAGAGAGGAATGTGTTCC
   -----+-----+-----+-----+-----+-----+-----+
901 GACACACAATGACTCTGGTGACCTATTCAGTGGTCGTGGGAGGTCTCTCCTTACACAAGG
   -----+-----+-----+-----+-----+-----+-----+
961 CGGGCTTGCTGACCACAGGCAGTGAGACAATTCAAGAACTAGAGATGAAACTTGAGGTG
   -----+-----+-----+-----+-----+-----+-----+
1021 GCGCGAACGACTGGTGTCCGTCACTCTGTTAAGTCTTTTGATCTCTACTTTGAACTCCAG
   -----+-----+-----+-----+-----+-----+-----+
1081 TGAACCTTGGAGATGTGTGATATATTACTCTGATAACTGGCTAAGAAGAATTAAGGACTG
   -----+-----+-----+-----+-----+-----+-----+
1141 ACTTGAAACCTCTACACACTATATAATGAGACTATTGACCGATTCTTCTTAATTCCTGAC
   -----+-----+-----+-----+-----+-----+-----+
1201 CAGAGGGGCTGGGGAGACTGCTGGTGGGTAACATCACGTTTCAGAGTTTGAATTCCCAGC
   -----+-----+-----+-----+-----+-----+-----+
1261 GTCTCCCCGACCCCTCTGACAGACCACCCATTGTAGTGCAAGTCTCAAACCTAAGGGTCG
   -----+-----+-----+-----+-----+-----+-----+
1321 ACTACATCAAAGCTAAACACTGACTGCACGCCTATAACCCAGCCTTGTGTTGGGGGAGCTGA
   -----+-----+-----+-----+-----+-----+-----+
1381 TGATGTAGTTTCGATTGTGTGACTGACGTGCGGATATTGGGTGCGGAACAACCCCTCGACT
   -----+-----+-----+-----+-----+-----+-----+
1441 AATAGGCAGGTCCAGGAGCTTGCTAAAGAAGGAATGAAGTAAAAGAAATCCGTTTAAAAA
   -----+-----+-----+-----+-----+-----+-----+
1501 TTATCCGTCCAGGTCTCGAACGATTCTTCTTACTTCATTTCCTTAGGCAAAATTTT
   -----+-----+-----+-----+-----+-----+-----+
1561 TACAGCAGCAAGTAAGGCATCCGTGCACTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGA
   -----+-----+-----+-----+-----+-----+-----+
1621 ATGTCGTCGTTTCAATTCGGTAGGCACGTGACACATCTGTCTCGTCACTCACGACACATCT
   -----+-----+-----+-----+-----+-----+-----+
1681 CAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGA
   -----+-----+-----+-----+-----+-----+-----+
1741 GTCTCGTCACTCACGACACATCTGTCTCGTCACTCACGACACATCTGTCTCGTCACTCT
   -----+-----+-----+-----+-----+-----+-----+
1801 GTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGA
   -----+-----+-----+-----+-----+-----+-----+
1861 CACGACACATCTGTCTCGTCACTCACGACACATCTGTCTCGTCACTCACGACACATCT
   -----+-----+-----+-----+-----+-----+-----+

```

Primer 57-C

CAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGA  
 961 -----+-----<-----+-----<-----+-----+----- 1020  
GTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCT  
 GTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGA  
 1021 ---<-----+-----+-----<-----+-----+-----<-----+ 1080  
CACGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCT  
 CAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGA  
 1091 -----+-----<-----+-----<-----+-----+-----+ 1140  
GTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCT  
 GTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGA  
 1141 ---<-----+-----+-----<-----+-----+-----+<-----+ 1200  
CACGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCT  
 CAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGA  
 1201 -----+-----<-----+-----<-----+-----+-----+ 1260  
GTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCT  
 GTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGA  
 1261 ---<-----+-----+-----<-----+-----+-----+<-----+ 1320  
CACGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCT  
 CAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGA  
 1321 -----+-----<-----+-----<-----+-----+-----+ 1380  
GTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCT  
 GTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGACAGAGCAGTAGAGTGCTGTGTAGA  
 1391 ---<-----+-----+-----<-----+-----+-----+<-----+ 1440  
CACGACACATCTGTCTCGTCATCTCAGACACATCTGTCTCGTCATCTCAGACACATCT  
 CAGAACAGTAGAGTTGTCTTTGGACGCTGACTGGCATTGGCAGACATGGTCCCAAGTG  
 1441 -----+-----+-----+-----+-----+-----+ 1500  
GTCTTGTCTATCTCAAAACAGAAACCTGCGACTGACCGTAAACCGTCTGTACCAGGGTTAC  
 GTACAAACAGAGCAAGATGGAGTTGATGTGAGAGAGGAGGCTCCTGGGGCTAGAGCAGAT  
 1501 -----+-----+-----+-----+-----+-----+ 1560  
 CATGTTTGTCTCGTTCTACCTCAACTACACTCTCTCCTCCGAGGACCCCGATCTCGTCTA  
 TGCCATCTCCTGGAGAAGTCACAGGTTCTATAAGGCCCTCTCTTGGGCTTGTGTATGGGA  
 1561 -----+-----+-----+-----+-----+-----+ 1620  
 ACGGTAGAGGACCTCTTCAGTGTCCAAGATATTCCGGAGAGAACCCGAACACAGTACCCT  
 CCTGTGTCATGTTTGGGGTCTTGTGTCCCCGTCAGAAATGCATACCTTGCCCATGCCTGTC  
 1621 -----+-----+-----+-----+-----+-----+ 1680  
 GGAACAGTACAAACCCAGAACGACAGGGGACGTCTTACGTATGGAACGGGTACGGACAG  
 ATCCCCTACTTAAACATAAGCAGGAGGTGAAGCGCAGCCTCAGTGCCTGCCTGGGCAGC  
 1681 -----+-----+-----+-----+-----+-----+ 1740  
 TAGGGGATGAATTTTGTATTCTGCTCCACTTCGCGTCCGAGTCACGGACGGACCCGTCG  
 AGGAGGAGATCCTGACTCAAAAAGCGCAGGGCTTGGATCCCATATCTTGTACCAGAAGAA  
 1741 -----+-----+-----+-----+-----+-----+ 1800  
 TCCTCCTCTAGGACTGAGTTTTCGCGTCCCGAACCTAGGGTATAGAACATGGTCTTCTT  
 ACAGAACTCTCCCTAGGAAACTCAGACACAGTTCCCCAGT  
 1801 -----+-----+-----+-----+ 1840  
 TGTCTTGAGAGGGATCCTTTGAGTCTGTGTCAAGGGGTCA

Primer 57-D

## b: Sequence of MMS58 clone

```

1  TCTGGGGCTTCCTGGCCCCAGATAATAAAGTGGAGGGTGACTGTGGTAGACACTGACTAT
   -----+-----+-----+-----+-----+-----+ 60
   AGACCCCGAAGGACCGGGGTCTATTATTTACCTCCCACTGACACCATCTGTGACTGATA

   CAACCTCTGGCCTCTAGGAGTGTATGGATGGACACACCTTCTCACAAGCACATACACACC
61 -----+-----+-----+-----+-----+-----+ 120
   GTTGGAGACCGGAGATCCTCACATACCTACCTGTGTGGAAGAGTGTTCGTGTATGTGTGG

   CATGCATCAAACCTACACACGCACATGTACACACATATGTACACACATACTCTCTCTCAC
121 -----+-----+-----+-----+-----+-----+ 180
   GTACGTAGTTTGGATGTGTGCGGTGTACATGTGTGTATACATGTGTGTATGAGAGAGAGTG

   ACACACATATACACACACACACACATACACACACTGTCTCTGTCTCTCTCTCTTTGTCTC
181 -----+-----+-----+-----+-----+-----+ 240
   TGTGTGTATATGTGTGTGTGTGTGTATGTGTGTGACAGAGACAGAGAGAGAGAAACAGAG

   TGTCTCTCTCTCACACATATACACATAGACATATTTCTTGGTCTCTGTTTCTCTCTCTC
241 -----+-----+-----+-----+-----+-----+ 300
   ACAGAGAGAGAGTGTGTATATGTGTATCTGTATAAGAGAACCAGAGACAAAGAGAGAGAG

   TCTCACACACATGGGATGTAGAATGAACATCTATGGACACTATTTTGTACTTACTATTTCT
301 -----+-----+-----+-----+-----+-----+ 360
   AGAGTGTGTGTACCTTACATCTTACTTGTAGATACCTGTGATAAAACATGAATGATAAGA

   CTAGACAAAAATCACAGCTGGTACTGGTTATTGTAGATAGAGATGCTTTTAACTAGGTG
361 -----+-----+-----+-----+-----+-----+ 420
   GATCTGTTTATTAGTGTGACCATGACCAATAACATCTATCTCTACGAAAAATTATGACCAC

   GGAAGAAATTTGGTAAATTTCTATGTAAGCATTCATCATTTAATGTAAAGAGCCTGAGCTT
421 -----+-----+-----+-----+-----+-----+ 480
   CCTTCTTTAACCATTTAAGATACATTCTGTAAGGTAGTAAATTACATTTCTCGGACTCGAA

   GGTGGATTTTATTATCCAAAGAAGACGGGGCAGTCTCCTGAGCAGTGCTCCCTGGCAGTG
481 -----+-----+-----+-----+-----+-----+ 540
   CCACCTAAATAATAGGTTTCTTCTGCCCCGTCAGAGGACTCGTCACGAGGGACCGTCAC

   CTCTGGAATATCCCTTCTGGAACACCAATGAGTGACTATGGTTTGGACTTGGTAAAGG
541 -----+-----+-----+-----+-----+-----+ 600
   GAGGACCTGATAGGGAAGACCTTGTGGTTACTACTGATACCAAAACCTGAACCATTTCC

   CAGAGATGGGTAGAGGGAAGAATCTTTCTGCCTGCCTGCCTGCCTGCCTTGTAGAGC
601 -----+-----+-----+-----+-----+-----+ 660
   GTCTCTACCCATCTCCCTTCTTAGAAAGACGGACGGACGGACGGACGGAACTCTG

   AAGGTTTCTATGTTGCCCTTGGCTGTCTGTAACTTACTTTGCAGACCAGGTTGGCCTC
661 -----+-----+-----+-----+-----+-----+ 720
   TTCCAAAAAGATACAACGGAACCGACAGGACATTGAATGAAACGCTCTGGTCCAACCGGAG
   Primer 58-C
   AAACCTACAGAAATATATCTGCCTCTGTCTCCAGAGTGCTGAGATTAAAGTCATATGCTC
721 -----+-----+-----+-----+-----+-----+ 780
   TTTGAGTGTCTTTATATAGACGGAGACAGAGGTCTCACGACTCTAATTTCACTATACGAG

   CCACATTCCTGTAGTGTCTTTCAAAGTGACTTTATCTTTAAAAATTTTTTTATTTGTTCT
781 -----+-----+-----+-----+-----+-----+ 840
   GGTGTAAGACATCACAAGAAAGTTTCACTGAAATAGAAATTTTATAAAAAATAAACAAGA

   TTGGCAATTTTATACTTGTATACAATGTATCCACCACCAGACATTCCTCGCAAACCTGACTT
841 -----+-----+-----+-----+-----+-----+ 900
   AACCGTTAAATATGAACATATGTTACATAGGTGGTGGTCTGTAAGGGCGTTTGACTGAA

```

```

TCTTCAAGCTATGGATGATGGAGAGGGAATAGTGATCAAACGAAAAAGTTACCAGATA
901 -----+-----+-----+-----+-----+-----+ 960
AGAAGTTCGATACCTACTACCTCTCCCTTATCACTAGTTTGACTTTTTTCAATGGTCTAT

CATGCTCACTTGGTTAGATAGTTGATACATGCTCACTTGGTTAGGTACTTTGATACATGCT
961 -----+----->+-----+-----+----->+-----+ 1020
GTACGAGTGAACCAATCTATCAACTATGTACGAGTGAACCAATCCATCAACTATGTACGA

CACTTGGTTAGATAGTTGATACATGCTCACTTGGTTAGATAGTTGATACATGCTCACTTG
1021 -----+-----+-----+-----+-----+-----+ 1080
GTGAACCAATCTATCAACTATGTACGAGTGAACCAATCTATCAACTATGTACGAGTGAAC

GTTAGATAGTTGATACATGCTCACTTGGTTAGGTAGTTGATACATGCTCACTTGGTTAGA
1081 ----->+-----+-----+-----+-----+-----> 1140
CAATCTATCAACTATGTACGAGTGAACCAATCTATCAACTATGTACGAGTGAACCAATCT

TAGTTGATACATGCTCACTTGGTTAGGTAGTTGATACATGCTCACTTGGTTAGATAGTTG
1141 -----+-----+-----+-----+-----+-----+ 1200
ATCAACTATGTACGAGTGAACCAATCTATCAACTATGTACGAGTGAACCAATCTATCAAC

ATACATGCTCACTTGGTTAGGTAGTTGATACATGCTCACTTGGTTAGGTAGTTGATACAT
1201 -----+-----+-----+-----+-----+-----+ 1260
TATGTACGAGTGAACCAATCTATCAACTATGTACGAGTGAACCAATCTATCAACTATGTA

GCTCACTTGGTTAGATAGTTGATACATGCTCACTTGGTTAGATAGTTGATACATGCTCAC
1261 -----+-----+-----+-----+-----+-----+ 1320
CGAGTGAACCAATCTATCAACTATGTACGAGTGAACCAATCTATCAACTATGTACGAGTG

TTGGTTAGATAGTTGATACATGCTCACTTGGTTAGGTAGTTGATACATGCTCACTTGGTT
1321 -----+-----+-----+-----+-----+-----+ 1380
AACCAATCTATCAACTATGTACGAGTGAACCAATCTATCAACTATGTACGAGTGAACCAA

AGGTAGTTGATACATGCTCACTTGGTTAGGTAGTTGATAGGGAAGTAGTGAGTTTAAATA
1381 -----+-----+-----+-----+-----+-----+ 1440
TCCATCAACTATGTACGAGTGAACCAATCCATCAACGTCCCTTCATCACTCAAAATTTAT

ACTTACAGTTATGTAGGAAGTCTGCTTGAGTTTAAATAGCGACTTTTCTTCTCCTTTGAG
1441 -----+-----+-----+-----+-----+-----+ 1500
TGAATGTCAATACATCCTTCAGACGAACATCAAATTATCGCTGAAAAGAAGAGGAAACTC

TTTCAGTTTATTTGGCGTTAGTGGCTGGGGGACGGGGTGGGGGACAGCCCTAGCACGTT
1501 -----+-----+-----+-----+-----+-----+ 1560
AAAGTCAAATAACCGCAATCACCGACCCCTGCCCCACCCCTGTGCGGGATCGTGCAA

TTAACCCCTTCCTCTGCCAAGTAGCCACGGTAGCTTTCCACCC
1561 -----+-----+-----+-----+-----+ 1602
AATTGGGAAGGAGACGGTTCATCGGTGCCATCGAAAGGTGGG

```

Primer 58-D



## Appendix 4

### ***Sequences flanking the Hm-1 repeat array***

Sequence data surrounding the *Hm-1* locus had been previously generated by R. Kelly using a cloned fragment derived from a C57BL/6J mouse (Kelly *et al.*, 1991). The region was resequenced using genomic DNA from a BALB/c strain (Chapter 6). Codes were aligned using the bestfit program from the GCG molecular biology software package (Devereux *et al.*, 1984). Resequenced code is presented on the top and with previous sequence on the bottom. Each primer site is underlined. Many mismatches between the sequences were due to the insertion or deletion of bases within arrays of identical nucleotides, which were most likely compression artefacts generated during the original manual sequencing of the region.

## Hm1p1

[illegible]



Appendix 4 Page 4

2249 CTCTTCTGAAAATCTTCTGAAATTTCTGCTTCTGTCAATTCCTGCTTGGGT 2298  
 |||  
 2222 CTCTTCTGAAAATCTTCTGAAATTTCTGCTTCTGTCAATTCCTGCTTGGGT 2271  
 |||  
 2299 TTCTGCTCTCACTTCCCTCAGCAATGAACTATGACCTGGAAGTGTGAGTC 2348  
 |||  
 2272 TTCTGCTCTCACTTCCCTCAGCAATGAACTATGACCT GAAGTGTGAGTC 2320  
 Hm1p24  
 2349 GAATAAGCCCTTCCCTCTCCTAAGCTGTTTTTGGTCATGATGTTTACCAC 2398  
 |  
 2321 G ATAAGCCC TCCCTCTCCTAAGCTG TTTTGGTCATGATGTTTACCAC 2367  
 |||  
 2399 AGCAACAGAAAGCAAACAAGGACAAATACTTTCTTCAACTACTAAAAAA 2448  
 |||  
 2368 AGCAACAGAAAGCAAACAAGGACAAATACTTTCTTCAACTACTAAAAAA 2417  
 Hm1p8  
 2449 TCAACAGAGGGATGATGcTAAGAGTATCTGCACAGGGAGATTTGGGTG T 2497  
 |||  
 2418 TCAACAGAGGGATGATGCTAAGAGTATCTGCACAAGGAGATTTGGGTGTT 2467  
 |||  
 2498 TTGTAACATGCGTTACAAGTCAGGCTTTCACAAGAAGAGAACTGAGATTA 2547  
 |||  
 2468 TTG AACATGCGTTACAAGTCAGGCTTTCACAAGAAGAGAACTGAGATTA 2516  
 Hm1p7  
 2548 CGTGGAACCTTTGAATTGGAAGAAGAAACATAAAATATGCTGTTGGAGAC 2597  
 |||  
 2517 CGTGGAACCTTTGAATTGGAAGAAGAAACATAAAATATGCTGTTGGAGAC 2566  
 |||  
 2598 CAGGGCCATAGCTTACCAGCTAGAAAGATCCTGATAGGTATACAGGCCAA 2647  
 |||  
 2567 CAGGGCCATAGCTTACCAGCTAGAAAGATCCTGATAGGTATACAGGCCAA 2616  
 |||  
 2648 TATGGACAGAATACAACCAGGTTGTATTCTGTATTCTAAAGGCTAGGTTA 2697  
 |||  
 2617 TATGGACAGAATACAACCAGGTTGTATTCTGTATTCTAAAGGCTAGGTTA 2666  
 |||  
 2698 TTACCTGACAAAATTGTGTATTTTGTGCCAAGAAGACCAATTCTATCTTC 2747  
 |||  
 2667 TTACCTGACAAAATTGTGTATTTTGTGCCAAGAAGACCAATTCTATCTTC 2716  
 |||  
 2748 TCACTGCAGGG TCTACACTTATGTAATAACTACTTAGCTTGCAGAATGT 2796  
 |||  
 2717 TCACTGCAGGGTTCTACACTTATGTAATAACTACTTAGCTTGCAGAATGT 2766  
 |||  
 2797 GCGAGTCAAGTAACTTCCAGTATCCCTTCTGCAAGTAAGTAAGTGACTAG 2846  
 |||  
 2767 GCGAGTCAAGTAACTTCCAGTATCCCTTCTGCAAGTAAGTAAGTGACTAG 2816  
 |||  
 2847 TAACAATCACGTTTTCTAATAGGGTCTCAAGGAGAAAACAAACACTCTAA 2896  
 |||  
 2817 TAACAATCACGTTTTCTAATAGGGTCTCAAGGAGAAAACAAACACTCTAA 2866  
 |||  
 2897 TAGGGCGTGGCCAAGGCAGAACAAACAAGAAGAAGAGGAGTAGGA 2946  
 |||  
 2867 TAGGGCGTGG CAAGGCAGAACAAACAAGAAGAAGAGGAGTAGGA 2915  
 |||  
 2947 GGCAGAAAGAAAAACAGACAAACAACAAAAaCAAAACAAAACAAaAAAGC 2996  
 |||  
 2916 GGCAGAAAGAAAAACAGACAAACAACAAAAACAAAACAAAACAAAAGC 2965  
 |||

2997 CCTCAGAACATAGAGAGGTAAGAACCTTAAATCTGCTTTAGACTCAAGG 3046  
 |||||  
 2966 CCTCAGAACATAGAGAGGTAAGAACCTTAAATCTGCTTTAGACTCAAGG 3015  
 Hmlp10  
 3047 GCACTTACTGACCACAGCAATAATGCTCTCTTCTCTcCTCTTCTTTCCCC 3096  
 |||||  
 3016 GCACTTACTGACCACAGCAATAATGCTCTCTTCTCTCTCTTCTTTCCCC 3065  
 Hmlp25  
 3097 GTGGAAAAGAAGGCTCTTCTACA TTTTTTTGTGGGCTAGCAACTCATCA 3145  
 |||||  
 3066 GTGGAAAAGAAGGCTCTTCTACATTTTTTTTGTGGGCTAGCAACTCATCA 3115  
 3146 TATCTTTGTACCCCTAAnaATAGATTCCAGTTTATCTGTAACTAATGTC 3195  
 ||||| : |||||  
 3116 TATC TTGTA CCCT AAATAGATTCCAG TTATCTGTAACTAATGTC 3160  
 Hmlp9  
 3196 ATTTTGGTTGCTGCTAATACTTAGCAATTACTTACAGAGACCAGCAATTG 3245  
 | |||||  
 3161 A TTTGGTTGCTGCTAATACTTAGC ATTACTTACAGAGACCAGCAATTG 3208  
 3246 TTAGACTGTCAAGCCATCTTCCAAAGCAGATGTGGGTTCTGTGTTCCCAT 3295  
 :::  
 3209 nnn 3258  
 3296 CAGTGGTGAGTGCAACTCATATACTCAGATCTGATGTTCACTGGGGTGTA 3345  
 :::  
 3259 nnn 3308  
 3346 GTTATTCTAACAGGCAGTAACTGGCTGTTGCTCTTACCAACTGCCCAATG 3395  
 :::  
 3309 nnn 3358  
 3396 ATGCGTGATGCTCCCATACACTCAAATAGGAGGCGATATGAGAAGTTACT 3445  
 :::  
 3359 nnn 3408  
 Hmp26  
 3446 GGAGATCACTCTTAGATAACGACCAGTACAGAACCTTTCTTCTCTCTGtA 3495  
 ::::::::::::::::::::::::::::::::::| ||| |||||  
 3409 nnnnnnnnnnnnnnnnnnnnnnnnnnnnnAG ACC TTCTTCTCTCTGTA 3456  
 3496 TTAATTTCTGCTGTGACAATGAATGTCCTAAAATACATTTAAGACCATG 3545  
 |||||  
 3457 TTAA TTCCTGCTGTGACAATGAATGTCCTAAAATACATTTAAGACCATG 3505  
 3546 TAAATTCATTGtTCTGGAGCTCACACGTaCAAAAAATAGATCTTACTGGG 3595  
 |||||  
 3506 TAAATTCATTGTTCTGGAGCTCACACGTACAAAAATAGATCTTACT GG 3554  
 3596 GGCTGGTGAGATGGCTCAGTGGGTAAGAGCACCCGACTGCTCTTCTGAAG 3645  
 |||||  
 3555 GGCTGGTGAGATGGCTCAGT GGTAAGAGCA CCGACTGCTCTTCTGAAG 3602  
 3646 GTCCGGAGTTCAAATCCCAGCAACCACATGGTGGCTCACAACCATCCGTA 3695  
 |||||  
 3603 GTCCGGAGGTC AATCCCAGCAA CACATGGTGGCTCACAACCATCCGTA 3650  
 3696 ATGTGATCTGACTCCCTCTTCTGGAGTGCTGAAGACAGCTACAGTGTA 3745  
 ||| : |||||  
 3651 ATGNNATCTGACTCCCTCTTCTGGAGTGCTGAAGACAGCTACAGTGTA 3700

3746 TTACATAAAATAAAATAAAATAAAATaAAATaAAATaAAATAAATCT GTTTGTTT 3794  
 |||||  
 3701 TTACATAAAATAAAATAAAATAAAATAAAATAAATAAATCTGGTTTG TT 3749  
  
 3795 TAAATAAAAAAAAAAATAGATCTTACTTTGCTGCAAATCCCTGCATAGA 3844  
 |||||  
 3750 TAAATAAAAAAAAAAATAGATCTTACTTTGCTGCAAATCCCTGCATAGA 3799  
  
 3845 CAGCACACATTCCCTTTTGGAGACTCTGGGAATAGTTATTTTTTATATTTT 3894  
 |||||  
 3800 CAGCACACATTCCCTTTTGGAGACTCTGGGAATAGTTATTTTTTATATTTT 3849  
  
 Hmlp12  
 3895 AGAGTCTATACATGCTCTTCTGTCCGCnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn 3944  
 |||||:.....  
 3850 AGAGTCTATACATGCTCTTCTGTCCGCCCCACCCACCTTTCCTACATTT 3899  
  
 3945 nnn 3994  
 :.....  
 3900 TAAACCTAGCTATATTGGGCTAAATCTGCTTCATATTATTCTCTCTGAT 3949  
  
 Hmlp11  
 3995 nnn 4022  
 :.....  
 3950 TGTCTTCCTGTCCACGCTCTAAGGATCC 3977

## Appendix 5

### ***Sequences flanking three human minisatellites***

Sequences flanking minisatellites *D17S74* (Appendix 5a) were generated from subclones of the pCMM86 clone (Nakamura *et al.*, 1987), whilst sequences flanking *D19S20* (Appendix 5b) were derived from subclones of both pJCZ3 and cMCOB19 (Nakamura *et al.*, 1988; Nakamura *et al.*, 1987), all of which were supplied by the Human Genome Mapping Project (HGMP). The sequence of minisatellite MS51 (Appendix 5c) is as published in Jeffreys *et al.* (1988). Flanking primers are underlined. The repeat region of each locus is represented in grey with sequences of each repeat to which MVR primers anneal underlined. Repeat unit sequences polymorphic between repeats and detected by MVR-PCR are double underlined. Sequences are presented in a 5' to 3' orientation from top to bottom. Designation of orientation was arbitrary.



## a: Sequence surrounding *D17S74*

```

      GTCTCCCCGCTTCnCTCCCCCACCCCCACCCCGACCCAGTAGTTAGAACAGCCGT
1  -----+-----+-----+-----+-----+-----+ 60
      CAGGAGGGGCGGAAGnGGAGGGGGTGGGGGTGGGGCGTGGGTCAATCTTGTTCGGCA

      CCTCCAAGCGATCTTATTATTAGAACTTAATAGAAACGCACATTCTCAGGTCCTCCCCCA
61 -----+-----+-----+-----+-----+-----+ 120
      GGAGGTTTCGCTAGAATAATAATCTTGAATTATCTTTGCGTGAAGAGTCCAGGAGGGGGT
      426-C
      GACCTCCTGAATCCAACTCTTTTAACACACTGTCCTGATGCACACTTAAGTTTGAAAAT
121 -----+-----+-----+-----+-----+-----+ 180
      CTGGAGGACTTAGGTTTGAGAAAATTGTGTGACAGGACTACGTGTGAATTCAAACTTTTA

      CACGGTCCTAGAATCAGCCTCTGGGTAACCTGGGAAGAAGTGAGCAGGAAGCGGTGAAAT
181 -----+-----+-----+-----+-----+-----+ 240
      GTGCCAGGATCTTAGTCGGAGACCCATTGAACCCCTTCTTCACTCGTCCTTCGCCACTTTA

      ATGAGGGTGGGTGTGTGGAGGGGGTGAGGGTGGGTGTGTGGAGGGGGTGAGGGTGGGT
241 -----+-----+-----+-----+-----+-----+ 300
      TACTCCACCCACACAACCTCCCCACTCCACCCACACAACCTCCCCACTCCACCCA

      GTGTTGGAGGGGGTGAGGGTGGGTGTGCTGGAGGGGGTGAGGGTGGGTGTGTTGGAGGGG
301 -----+-----+-----+-----+-----+-----+ 360
      CACAACCTCCCCACTCCACCCACACGACCTCCCCACTCCACCCACACAACCTCCCC

      ATGAGGGTGGGTGTGTGGAGGGGATGAGGGTGGGTGTGTGGAGGGGGTGAGGGTGGGT
361 -->-----+-----+-----+-----+-----+-----+ 420
      TACTCCACCCACACAACCTCCCCACTCCACCCACACAACCTCCCCACTCCACCCA

      GTGCTGGAGGGGGTGAGGGTGGGTGTGTTGGAGGGGGTGAGGGAGGGTGACTGAGGAAG
421 -----+-----+-----+-----+-----+-----+ 480
      CACGACCTCCCCACTCCACCCACACAACCTCCCCACTCCCTCCCACCTGACTCCTTC

      GGCTGCTGAGTGAAGGTCACAGTGCAAGACAATTCCACAGCGCTTGTCCCAATCAGGGA
481 -----+-----+-----+-----+-----+-----+ 540
      CCGACGACTCACTTCCAGTGTCACGTTCTTGTAAAGGTGTCGGAACAGGGTTAGTCCCT

      ACCTTTAAGGAACAAGATATCGGATGGCAATTTTGTATTTTCCTTTTGTAGGTAAGACATA
541 -----+-----+-----+-----+-----+-----+ 600
      TGGAAATTCCTTGTCTATAGCCTACCGTTAAACATAAAAGGAAAAATCCATTCTGTAT

      TCCAATGCTGTCAGCTGGATAGAGGGAAAGGGGGGTTACTAGGGGACCCACAACCTTAGG
601 -----+-----+-----+-----+-----+-----+ 660
      AGGTTACGACAGTGCACCTATCTCCCTTTCCCCCAATGATCCCCTGGGTGTTGAATCC
      426-D
      GGCAAGGTCTTGGACCAGGAGCAGAAACACAGCGGCTGGCTGAGTGTGCAGGGAGCCCT
661 -----+-----+-----+-----+-----+-----+ 720
      CCGTTCAGAACCTGGTCCTCGTCTTTGTGTGCGCGACCGACTCACGACGTCCCTCGGGA

      CAGCCAACCTGTGTCCCCACTCTGGCTGCTTGTCTCTGCCTGTTGAGATTTACCAAACCT
721 -----+-----+-----+-----+-----+-----+ 780
      GTCGGTTGGACACAGGGGTGAGACCGACGAACAGAGACGGACAACCTCTAAAGTGGTTTGA

      ACCACTGTGTTAATGGACTAGAATCTTGCAACCCCTCTCTAGATGACCCAGGAAAATGGAT
781 -----+-----+-----+-----+-----+-----+ 840
      TGGTGACACAATTACCTGATCTTAGAACGTTGGGAGAGATCTACTGGGTCCTTTTACCTA

      GCTCAGAAATCAAAAAGAAGCTTGAGACTGCATGGAAAAATCACTAGTGAATTAATCTT
841 -----+-----+-----+-----+-----+-----+ 900
      CGAGTCTTTAGTTTTTCTTGAACCTCTGACGTACCTTTTGTAGTATCACTTAATTTTGA

      TATTTCTGTTTTTTTTCAATCCTGTCTGTATCCTAAGAAAAAGCTGGTGCCGAATGAATG
901 -----+-----+-----+-----+-----+-----+ 960
      ATAAAGACAAAAAAGTTAGGACAGACATAGGATTCTTTTTCGACCACGGCCTTACTTAC

```

```

961  CGTATCTTTTTGTTGTTGTTGTTTGGACACTGACAAAAATTCCTTTTGCAGATGGCTACT
-----+-----+-----+-----+-----+-----+ 1020
GCATAGAAAAACAACAACAACAACTGTGACTGTTTTTAAGGAAAACGTCTACCGATGA

GAGGCTTCAAAGGGGCAATGTATCATGCAAATAATTTTTTCAGCTATGACGTGTACAGT
1021 -----+-----+-----+-----+-----+ 1080
CTCCGAAGTTTCCCCGTTACATAGTACGTTTATTAAAAAAGTCGATACTGCACATGTCA

TCAGTTTGATTTATTTTTGATGTCAGAAACAAATGGAAAGTTAGCGCTTTAGAAATGAG
1081 -----+-----+-----+-----+-----+ 1140
AGTCAAACATAAATAAACTACAGTCTTTGTTTTACCTTTCAATCGCGAAATCTTTACTC

CGTGGnTCCTGTGAATCAGATAAAGAACC
1141 -----+-----+----- 1169
GCACChAGGACACTTAGTCTATTTCTTGG

```

**b: Sequence surrounding *D19S20***

CAACCGTCACCACATATCCATAGTCTCATACCTTTCTCATTCCCCCGGACAGAAACCCGCACC  
1 -----+-----+-----+-----+-----+-----+-----+ 60  
GTTGGCAGTGGTGATAGGTATCAGGTATGGAAGAGTAAGGGGCCCTGTCTTTGGCGTGG

CCGCACATGCCACTTCTCACCCCTCCAACCCCCGGCCCCCGGCACCCATGCATCCCCTTC  
61 -----+-----+-----+-----+-----+-----+-----+ 120  
GGCGTGACGGTGAAGAGTGGGGAGGTTGGGGGCCGGGGGCCGTGGGTACGTAGGGGAAG

CTGTCTCTGGATTGGCCTGTCCTGGACATTTCTGTAGAAATGGGCTCACACGGCCGGGCGC  
121 -----+-----+-----+-----+-----+-----+-----+ 180  
GACAGAGACCTAACCGGACAGGACCTGTAAAGCATCTTTACCCGAGTGTGCCGGCCCGCG

AGTGGCTCAGGCCGGAATCCCCACACTTTGCGAGGCCTAGGCAGAAGGATCATGAGGTC  
181 -----+-----+-----+-----+-----+-----+-----+ 240  
TCACCGAGTCCGGCCATTAGGGGTGTGAAACGCTCCGGATCCGCTCTCTTAGTACTCCAG

AGGGTTTCGAGACCAGCCTGACCAACATGGTGAAACCTGTCTCTACTAAAAATACAAA  
241 -----+-----+-----+-----+-----+-----+-----+ 300  
TCCCAAAGCTCTGGTCGGACTGGTTGTACCATTGTTGGGACAGAGATGATTTTTATGTTTT

ATGAGCCAGGAGTGGTGGCTCATGCCTGTAATCCAGCTACTCAGGAGGCTGAGGCAGGA  
301 -----+-----+-----+-----+-----+-----+-----+ 360  
TACTCGGTCTCACCACCGAGTACGGACATTAGGGTCGATGAGTCTCCGACTCCGTCCT

GAATCGCTTGAAC TTGGGAGGCGGAGGTGGCAGTGAGCCGAGATGGTGCCACTGCACTCC  
361 -----+-----+-----+-----+-----+-----+-----+ 420  
CTTAGCGAACTTGAACCTCCGCCTCCACCGTCACTCGGCTCTACCACGGTGACGTGAGG

AGCCTGGGCGACAGGGTGAGACTCCGTCTCGGAAAAAAAAAAGAAATGGCTTCTCTCACTG  
421 -----+-----+-----+-----+-----+-----+-----+ 480  
TCGGACCCGCTGTCCCACTCTGAGGCAGAGCCTTTTTTTTTCTTTACCGAAGAGAGTGAC

AACGTGACGTCTCAAGGGGCATCTGCGCCGTGGCCTGGGTGAGAGCCTCACTCCTTTTC  
481 -----+-----+-----+-----+-----+-----+-----+ 540  
TTGCACTGCAGGAGTTCCCCGTAGACGCGGCACCGGACCCAGTCTCGGAGTGAGGAAAAG

GTGGCTGAGTCGTGTTCCATGGTGGACGGGTCGCGCCGTGTTTGTCCTCCGTTTGGTGA  
541 -----+-----+-----+-----+-----+-----+-----+ 600  
CACCGACTCAGCACAAAGGTACCACCTGCCCAGCGCGGCACAAACAGGGAGGCAAACCACT

TGGGCACCTGGGCTGCCTCTGCCTTTTTTGCTACTGnnnnnnnnnnnnnnnnnnnnnnnnn  
601 -----+-----+-----+-----+-----+-----+-----+ 660  
ACCCGTGGACCCGACGGAGACGGAAAAACGATGACnnnnnnnnnnnnnnnnnnnnnnnnn

Gap of ~600 bp

[illegible]

1381 TCGGCCTCCCCTGTGGCGCCAGCCCCGTCTCTGCTTGGCTTCCCTCCCTCCTGTGTGGAC  
 -----+-----+-----+-----+-----+-----+-----+ 1440  
 AGCCGGAGGGGACACCGCGGTGCGGGCAGAGACGAACCGAAGGGAGGGAGGACACACCTG  
  
 1441 GCCCCACTTCCTCCCTCTCTCTCTGCGTGGACGCCCCACTCCCTCCTCTCTCTCTCTG  
 -----+-----+-----+-----+-----+-----+-----+ 1500  
 CCGGGTGAAGGAGGGGAGAGGAGGACGCACCTGCGGGGTGAGGGAGGAGAGAGGAGGACG  
  
 1501 GTGGACGCCCCACTTcccTCCTCTCTCTCTCTGCGTGGACGCCCCACTCCCTCCTCTCTCTC  
 -----+-----+-----+-----+-----+-----+-----+ 1560  
 CACCTGCGGGGTGAgggAGGAGaGAGgAGGACGCACCTGCGGGGTGAGGGAGGAGAGAGG  
  
 1561 TCCTGCGTGGACGCCCCACTcccTCCTCTCTCTCTCTGCGTGGACGCCCCACTCCCTCTCT  
 -----+-----+-----+-----+-----+-----+-----+ 1620  
 AGGACGCACCTGCGGGGTGAgggAGGAGaGAGgAGGACGCACCTGCGGGGTGAGGGAGGA  
  
 1621 CTCTCTCTCTGCGTGGACGCCCCACTTcCTCTCTCTCTCTCTGCGTGGACGCCCCACTT  
 -----+-----+-----+-----+-----+-----+-----+ 1680  
 GAGAgGAGGACGCACCTGCGGGGTGArGGAGGAGAGAgGAGGACGCACCTGCGGGGTGA  
  
 1681 CCTCTGGAATGACGCGGTGCTGCCCTGGTTTCACTCTCTCTGTGGCCCCCTGTGGAGGCT  
 -----+-----+-----+-----+-----+-----+-----+ 1740  
 GGAGACCTTTACGTGCGCCACGACGGGACCAAGTCAGAAGGACACCGGGGACACCTCCGA  
  
 1741 CCCGCTTCTCAGTAAGGCATCTGGGGCCCTCACCTCCCGGCGTGAGCTGAGCACGGTTCA  
 -----+-----+-----+-----+-----+-----+-----+ 1800  
 GGGCGAAGAGTCATTCCGTAGACCCCGGAGTGAGGGCCGCACTCGACTCGTGCCAAGT  
  
 1801 TTTCCACCCCTGGCTTCCCGCCTTCGCTCTCCCTGTGCGCGCCGCCCTGCCTCGCCCTGC  
 -----+-----+-----+-----+-----+-----+-----+ 1860  
 AAAGGGTGGGACCGAAGGGCGGAAGCGAGAGGGACACGCGCGGCGGACGGAGCGGGACG  
  
 1861 TGCTTCCCAGTGCCCCCTGACCCCCGGCTCCTCTCTGCTCCTCTGCTGTCTGTCTCTGT  
 -----+-----+-----+-----+-----+-----+-----+ 1920  
 ACGAAGGGTCACGGGGAGACGTGGGGCCGAGGAGAGACGAGGAGACGACAGGACAGGACA  
  
 1921 TATTTTTTTAAACGCCAACCTTGTGCCAGGCACATTGTCCCTCTCTGGTCATGTGAAGCAC  
 -----+-----+-----+-----+-----+-----+-----+ 1980  
 ATAAAAAATTGCGGTTGGAACACGGTCCGTGTAACAGGGAGAGACCAGTACACTTCGTG  
  
 1981 TTAATnTATTTTGAGCCACCTCTTGGGCCTTGTGAGGTCTCCTGCATCTTCAAGGACACG  
 -----+-----+-----+-----+-----+-----+-----+ 2040  
 AATTAnATAAACTCGGTGGAGAACCCGGAACTCCAGAGGACGTAGAAGTTCCTGTGC  
 118-D  
 2041 GGGTCAAAGTTGGGCCTGAGAAAGGGACCTGGGCTAGGGCAGCnCAGTGGGCGTTGGGCT  
 -----+-----+-----+-----+-----+-----+-----+ 2100  
 CCCAGTTTCAACCCGACTCTTTCCCTGGACCCGATCCCGTCGnGTCACCCGCAACCCGA  
  
 2101 CCCTCGGGGCTGGGTGGCCTTCCTGGGGTGGGAGCCTGGTCCCGAGGGAGGAGTGCCCAG  
 -----+-----+-----+-----+-----+-----+-----+ 2160  
 GGGAGCCCCGACCCACCGGAAGGACCCACCTCGGACCAGGGCTCCCTCCTCACGGGTC  
  
 2161 GGCTTGTCCTGCAGGCGCCTGGGGGGAAGGCACCGGCCTGAGGTGTGGGCACCCCTCGCCC  
 -----+-----+-----+-----+-----+-----+-----+ 2220  
 CCGAACAGGACGTCCGCGGACCCCCCTTCCGTGGCCGACTCCACACCCGTGGGAGCGGG  
  
 2221 CCCA  
 ---- 2224  
 GGGT

## c: Sequence surrounding MS51

51-A  
 1 GATCAGCGAACTTCCTCTCGGCTCCCGATATCCTCCTCGATACGCACTCTGCCACAACGG 60  
 CTAGTCGCTTGAAGGAGAGCCGAGGGCTATAGGAGGAGCTATGCGTGAGACGGTGTGGC  
 51-C  
 61 GCAGGGTCCCTTTCAGCGTCTCATCCACAGTGAACGGGAGTTGAGGCTTTCTTAGCGGAG 120  
 CGTCCCAGGGAAAGTCGCAGAGTAGGTGTCACTTGCCCTCAACTCCGAAAGAATCGCCTC  
 GGGCTGGAGGGACACAGCAGGAGGGCAGGAGGGCAAGTGGAGGGACATGGCAGGAGGGCA  
 121 -----+-----+-----+-----+-----+ 180  
 CCCGACCTCCCTGTGTCTGTCCTCCGTCCTCCCGTTCACCTCCCTGTACCGTCCTCCCGT  
 GGTGGAGGGACATGGCAGGAGGGCAGGTGGAGGGACATGGCAGGAGGGCAGGTGAAGGGGA  
 181 -----+-----+-----+-----+-----+ 240  
 CCACCTCCCTGTACCGTCCTCCCGTCCACCTCCCTGTACCGTCCTCCCGTCCACTCCCT  
 GGGCAGGAGGGCAAGTGGAGGGACATGGCATGATGGCAGGAGGGCAGGTGGAGGGACATG  
 241 <-----+-----+-----+-----+-----+ 300  
 CCGTCCTCCCGTTCACCTCCCTGTACCGTACTACCGTCCTCCCGTCCACCTCCCTGTAC  
 GCAGGAGGGCAGGAGGGTAGGTGGAGGGACACGGCAGGAGGGTAGGTGGAGGGACACGGC  
 301 -----+-----+-----+-----+-----+ 360  
 CGTCCTCCCGTCCTCCCATCCACCTCCCTGTGCGTCCTCCCATCCACCTCCCTGTGCGG  
 AGGAGGGCAGGTGGAGGGACACGGCAGGAGGGCAGGTGGAGGGACACGGCAGGAGGGCAG  
 361 -----+-----+-----+-----+-----+ 420  
 TCCTCCCGTCCACCTCCCTGTGCGTCCTCCCGTCCACCTCCCTGTGCGTCCTCCCGTC  
 GTGGAGGGACACCGCAGGAGGGCAGGTGGAGGGACACGGCAGGAGGGCAGGTGGAGGGAC  
 421 -----+-----+-----+-----+-----+ 480  
 CACCTCCCTGTGCGTCCTCCCGTCCACCTCCCTGTGCGTCCTCCCGTCCACCTCCCTG  
 AGAGCAGGAGGGCAGGCCTCCCTGCGGTTTCCGGATGCTACGGGGTGGATCGGAGTGTGG  
 481 -----+-----+-----+-----+-----+ 540  
 TCTCGTCCTCCCGTCCGAGGGACGCCAAAGGCCTACGATGCCCCACCTAGCCTCACACC  
 TGTTAAGCACATCTGGACACGCTCTGTCCGAGACACATAGTCCCCAGGCGACCTACAGCC  
 541 -----+-----+-----+-----+-----+ 600  
 ACAATTTCGTGTAGACCTGTGCGAGACAGGCTCTGTGTATCAGGGGTCCGCTGGATGTGG  
 51-D  
 601 ACAGCCTGACCTCCTGAAAATTTCCAGCTTCCCACAGTCCTCAATGTGGAAACCAAGTGC 660  
 TGTCCGACTGGAGGACTTTTAAAGGGTCCGAAGGGTGTGAGGAGTTACACCTTTGGTCACG  
 51-B  
 661 CCAAAGGCCACCTGCCCCACACTGGCACC GAATTC 694  
 GGTTTCCGGTGGACGGGTGTGACCGTGGCTTAAG

## Appendix 6

### ***Sequences flanking the insulin minisatellite***

~3 kb of sequence surrounding the insulin minisatellite from the GENBANK data base (Accession No. L15440) is presented. The minisatellite is indicated in grey, and primer sites underlined. Overlapping primers are indicated by wavy lines. The two most proximal *Hinf*I sites both 5' and 3' of the repeat array are indicated. *Hinf*I was used to digest genomic DNA prior to size enrichment for mutant alleles (Chapter 8).

HinfI

TCTCCCCAGCGAGGCAGGATGGGGGCTGGATTTCAGACTCTGTAAGATGCCCCCTGGCTTA  
 1 -----+-----+-----+-----+-----+-----+-----+ 60  
 AGAGGGGTGCTCCGTCCCTACCCCCGACCTAAAGTCTGAGACATTCTACGGGGACCGAAT  
  
 CTCGAGGGGCCCTAGACATTGCCCTCCAGAGAGAGCACCCAACACCTCCAGGCTTGACCG  
 61 -----+-----+-----+-----+-----+-----+-----+ 120  
 GAGCTCCCCGGATCTGTAACGGGAGGTCTCTCTCGTGGGTGTGGGAGGTCCGAACTGGC  
  
 GCCAGGGTGTCCCCCTTCCTACCTTGGAGAGAGCAGCCCCAGGGCATCCTGCAGGGGGTGC  
 121 -----+-----+-----+-----+-----+-----+-----+ 180  
 CGGTCCCACAGGGGAAGGATGGAACCTCTCTCGTGGGGTCCCGTAGGACGTCCCCACG  
  
 TGGGACACCAGCTGGCCTTCAAGGTCTCTGCCCTCCCTCCAGCCACCCCACTACACGCTGC  
 181 -----+-----+-----+-----+-----+-----+-----+ 240  
 ACCCTGTGGTTCGACCGAAGTTCCAGAGACGGAGGGAGGTGGGTGGGTGATGTGCGACG  
  
 TGGGATCCTGGATCTCAGCTCCCTGGCCGACAACACTGGCAAACCTCTACTCATCCACGA  
 241 -----+-----+-----+-----+-----+-----+-----+ 300  
 ACCCTAGGACCTAGAGTCGAGGGACCGGTGTTGTGACCGTTTGAGGATGAGTAGGTGCT  
  
 AGGCCCTCCTGGGCATGGTGGTCCTTCCCAGCCTGGCAGTCTGTTCTCACACACCTTGT  
 301 -----+-----+-----+-----+-----+-----+-----+ 360  
 TCCGGGAGGACCCGTACCACCAGGAAGGGTCGGACCGTCAGACAAGGAGTGTGTGGAACA  
  
 TAGTGCCAGCCCCCTGAGGTTGCAGCTGGGGGTGTCTCTGAAGGGCTGTGAGCCCCCAGG  
 361 -----+-----+-----+-----+-----+-----+-----+ 420  
 ATCACGGGTTCGGGACTCCAACGTTCAGCCCCACAGAGACTTCCCGACACTCGGGGGTCC  
  
 AAGCCCTGGGGAAGTGCCTGCCTTGCTTCCCCCGGCCCTGCCAGCGCTGGCTCTGCCC  
 421 -----+-----+-----+-----+-----+-----+-----+ 480  
 TTCGGGACCCCTTCACGGACGGAACGGAGGGGGCCGGGACGGTCGCGGACCGAGACGGG  
  
 TCCTACCTGGGCTCCCCCATCCAGCCTCCCTCCCTACACACTCCTCTCAAGGAGGCACC  
 481 -----+-----+-----+-----+-----+-----+-----+ 540  
 AGGATGGACCCGAGGGGGGTAGGTTCGGAGGGAGGGATGTGTGAGGAGAGTTCTCCGTGG  
  
 CATGTCCTCTCCAGCTGCCGGGCCTCAGAGCACTGTGGCGTCTGGGGCAGCCACCGCAT  
 541 -----+-----+-----+-----+-----+-----+-----+ 600  
 GTACAGGAGAGGTTCGACGGCCCGGAGTCTCGTGACACCGCAGGACCCCGTCGGTGGCGTA  
  
 GTCCTGCTGTGGCATGGCTCAGGGTGGAAAGGGCGGAAGGGAGGGTCTGCAGATAGCT  
 601 -----+-----+-----+-----+-----+-----+-----+ 660  
 CAGGACGACACCGTACCGAGTCCACCTTTCCCGCCTTCCCTCCCCAGGACGTCTATCGA  
  
 GGTGCCCCACTACCAAACCCGCTCGGGGCAGGAGGCCAAAGGCTGGGTGTGTGCAGAGCG  
 661 -----+-----+-----+-----+-----+-----+-----+ 720  
 CCACGGGTGATGGTTTGGGCGAGCCCGTCCTCTCGGTTTCCGACCCACACACGTCTCGC  
  
 GCCCCGAGAGGTTCCGAGGCTGAGGCCAGGGTGGGACATAGGGATGCGAGGGGGCCGGGGC  
 721 -----+-----+-----+-----+-----+-----+-----+ 780  
 CGGGGCTCTCCAAGGCTCCGACTCCGGTCCCAACCTGTATCCCTACGCTCCCCGGCCCCG  
  
 ACAGGATACTCCAACCTGCCTGCCCCCATGGTCTCATCCTCCTGCTTCTGGGACCTCTG  
 781 -----+-----+-----+-----+-----+-----+-----+ 840  
 TGTCTTATGAGGTTGGACGGACGGGGGTACCAGAGTAGGAGGACGAAGACCTGGAGGAC  
  
 ATCCTGCCCCTGGTGCTAAGAGGCAGGTAAGGGGCTGCAGGCAGCAGGGCTCGGAGCCCA  
 841 -----+-----+-----+-----+-----+-----+-----+ 900  
 TAGGACGGGGACACGATTCCTCCGTCCATTCCTCCGACGTCCGTCTCCGAGCCTCGGGT  
  
 TGCCCCCTCACCATGGGTCAGGCTGGACCTCCAGGTGCCTGTTCTGGGGAGCTGGGAGGG  
 901 -----+-----+-----+-----+-----+-----+-----+ 960  
 ACGGGGGAGTGGTACCCAGTCCGACCTGGAGGTCCACGGACAAGACCCCTCGACCTCCC

```

          CCGGAGGGGTGTACCCAGGGGCTCAGCCCAGATGACACTATGGGGGTGATGGTGTCTATG
961  -----+-----+-----+-----+-----+-----+-----+ 1020
          GGCCTCCCCACATGGGGTCCCCGAGTCGGGTCTACTGTGATACCCCCACTACCACAGTAC

          GGACCTGGCCAGGAGAGGGGAGATGGGCTCCCAGAAGAGGAGTGGGGGCTGAGAGGGTGC
1021 -----+-----+-----+-----+-----+-----+-----+ 1080
          CCTGGACCGGTCTCTCCCTCTACCCGAGGGTCTTCTCCTCACCCCCGACTCTCCACG

          HinfI
          |
          CTGGGGGGCCAGGACGGAGCTGGGCCAGTGCACAGCTTCCCACACCTGCCCACCCCCAGA
1081 -----+-----+-----+-----+-----+-----+-----+ 1140
          GACCCCCCGGTCTCTGCCTCGACCCGGTCACGTGTCTGAAGGGTGTGGACGGGTGGGGGTCT

          MINS-A          INS-5          INS-1296
          GTCCTGCCGCCACCCCAGATCACACGGAAGAATGAGGTCCGAGTGGCCTGCTGAGGACT
1141 -----+-----+-----+-----+-----+-----+-----+ 1200
          CAGGACGGCGGTGGGGGTCTAGTGTGCCTTCTTACTCCAGGCTCACCGGACGACTCCTGA

          TGCTGCTTGCTCCCCAGGTCCCCAGGTCATGCCCTCCTTCTGCCACCCTGGGGAGCTGAGG
1201 -----+-----+-----+-----+-----+-----+-----+ 1260
          ACGACGAACAGGGGTCCAGGGGTCCAGTACGGGAGGAAGACGGTGGGACCCCTCGACTCC

          GCCTCAGCTGGGGCTGCTGTCTTAAGGCAGGGTGGGAAGTAGGCAGCCAGCAGGGAGGGG
1261 -----+-----+-----+-----+-----+-----+-----+ 1320
          CGGAGTCGACCCCGACGACAGGATTCCGTCCCACCCTTGATCCGTGCGTCGTCCCTCCCC

          ACCCCTCCCTCACTCCCACCTCTCCCACCCCCACCACCTTGGCCCATCCATGGCGGCATCT
1321 -----+-----+-----+-----+-----+-----+-----+ 1380
          TGGGGAGGGAGTGAGGGGTGAGAGGGTGGGGGTGGTGAACCGGGTAGGTACCGCCGTAGA

          TGGGCCATCCGGGACTGGGGACAGGGGTCTGGGGACAGGGGTCCGGGGACAGGGTCTCTG
1381 -----+-----+-----+-----+-----+-----+-----+ 1440
          ACCCGGTAGGCCCTGACCCCTGTCCCCAGGACCCCTGTCCCCAGGCCCTGTCCCAGGAC

          GGGACAGGGGTGTGGGGACAGGGGTCTGGGGACAGGGGTGTGGGGACAGGGGTGTGGGGGA
1441 -----+-----+-----+-----+-----+-----+-----+ 1500
          CCCTGTCCCCACACCCCTGTCCCCAGACCCCTGTCCCCACACCCCTGTCCCCACACCCCT

          CAGGGGTCTGGGGACAGGGGTGTGGGGACAGGGGTCCGGGGACAGGGGTGTGGGGACAGG
1501 -----+-----+-----+-----+-----+-----+-----+ 1560
          GTCCCCAGACCCCTGTCCCCACACCCCTGTCCCCAGGCCCTGTCCCCACACCCCTGTCC

          GGTCTGGGGACAGGGGTGTGGGGACAGGGGTGTGGGGACAGGGGTCTGGGGACAGGGGTG
1561 -----+-----+-----+-----+-----+-----+-----+ 1620
          CCAGACCCCTGTCCCCACACCCCTGTCCCCACACCCCTGTCCCCAGACCCCTGTCCCCAC

          TGGGGACAGGGGTCTGGGGACAGGGGTGTGGGGACAGGGGTGTGGGGACAGGGGTGTGG
1621 -----+-----+-----+-----+-----+-----+-----+ 1680
          ACCCCTGTCCCCAGGACCCCTGTCCCCACACCCCTGTCCCCACACCCCTGTCCCCACACC

          GGACAGGGGTGTGGGGACAGGGGTCTGGGGATAGGGGTGTGGGGACAGGGGTGTGGGGGA
1681 -----+-----+-----+-----+-----+-----+-----+ 1740
          CCTGTCCCCACACCCCTGTCCCCAGGACCCCTATCCCCACACCCCTGTCCCCACACCCCT

          CAGGGGTCCCGGGACAGGGGTGTGGGGACAGGGGTGTGGGGACAGGGGTCTGGGGACA
1741 -----+-----+-----+-----+-----+-----+-----+ 1800
          GTCCCCAGGGCCCTGTCCCCACACCCCTGTCCCCACACCCCTGTCCCCAGGACCCCTGT

          GGGGTCTGAGGACAGGGGTGTGGGGACAGGGGTCTGGGGACAGGGGTCTGGGGACAGG
1801 -----+-----+-----+-----+-----+-----+-----+ 1860
          CCCCAGACTCCTGTCCCCACACCCGTGTCCCCAGGACCCCTGTCCCCAGGACCCCTGTCC

          GGTCTGGGGACAGGGGTCTGGGGACAGCGCGAAAGAGCCCCGCCCTGCAGCCTCCAG
1861 -----+-----+-----+-----+-----+-----+-----+ 1920
          CCAGGACCCCTGTCCCCAGACCCCTGTCGTGCGGTTTCTCGGGGCGGGACGTGGGAGGTC

```



CTCTCCTGGTCTAATGTGGAAAGTGGCCCAGGTGAGGGCTTTGCTCTCCTGGAGACATTT  
 1921 -----+-----+-----+-----+-----+-----+ 1980  
 GAGAGGACCAGATTACACCTTTCACCGGGTCCACTCCCAGAAACGAGAGGACCTCTGTAAA  
 MINS-C  
 HinfI  
 GCCCCAGCTGTGAGCAGGGACAGGTCTGGCCACCGGGCCCCCTGGTTAAGACTCTAATGA  
 1981 -----+-----+-----+-----+-----+-----+ 2040  
 CGGGGGTCGACACTCGTCCCTGTCCAGACCGGTGGCCCCGGGGACCAATTCTGAGATTACT  
 MINS-B  
 CCCGCTGGTCTCTGAGGAAGAGGTGCTGACGACCAAGGAGATCTTCCACAGACCCAGCAC  
 2041 -----+-----+-----+-----+-----+-----+ 2100  
 GGGCGACCAGGACTCCTTCTCCACGACTGCTGGTTCTCTAGAAGGGTGTCTGGGTCTGTG  
 CAGGGAAATGGTCCGAAATTGCAGCCTCAGCCCCAGCCATCTGCCGACCCCCCACCC  
 2101 -----+-----+-----+-----+-----+-----+ 2160  
 GTCCCTTTACAGGCCTTTAACGTCCGAGTCCGGGGTCCGGTAGACGGCTGGGGGGTGGG  
 CGCCCTAATGGGCCAGGCGCAGGGGTTGACAGGTAGGGGAGATGGGCTCTGAGACTATA  
 2161 -----+-----+-----+-----+-----+-----+ 2220  
 GCGGGATTACCCGGTCCGCGTCCCCAACTGTCCATCCCCTCTACCCGAGACTCTGATAT  
 AAGCCAGCGGGGGCCCAGCAGCCCTCAGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCA  
 2221 -----+-----+-----+-----+-----+-----+ 2280  
 TTCGGTCGCCCCCGGGTCGTCGGGAGTCGGGAGGTCCTGTCCGACGTAGTCTTCTCCGGT  
 TCAAGCAGGTCTGTTC AAGGGCCTTTGCGTCAGGTGGGCTCAGGGTTCAGGGTGGCTG  
 2281 -----+-----+-----+-----+-----+-----+ 2340  
 AGTTCGTCCAGACAAGGTTC CCGAAACGCAGTCCACCCGAGTCCCAAGGTCCCACCGAC  
 GACCCAGGCCCCAGCTCTGCAGCAGGGAGGACGTGGCTGGGCTCGTGAAGCATGTGGGG  
 2341 -----+-----+-----+-----+-----+-----+ 2400  
 CTGGGGTCCGGGGTCGAGACGTGCTCCCTCCTGCACCGACCCGAGCACTTCGTACACCC  
 GTGAGCCCAGGGGGCCCCAAGGCAGGGCACCTGGCCTTCAGCCTGCCTCAGCCCTGCCTGT  
 2401 -----+-----+-----+-----+-----+-----+ 2460  
 CACTCGGGTCCCCGGGGTTCCGTCCCGTGGACCGGAAGTCGGACGGAGTCGGGACGGACA  
 CTCCCAGATCACTGTCTTCTGCCATGGCCCTGTGGATGCGCCTCCTGCCCCCTGCTGGCG  
 2461 -----+-----+-----+-----+-----+-----+ 2520  
 GAGGGTCTAGTGACAGGAAGACGGTACCGGGACACCTACGCGGAGGACGGGGACGACCGC  
 INS-23+/- INS-3  
 CTGCTGGCCCTCTGGGGACCTGACCCAGCCGAGCCTTTGTGAACCAACACCTGTGCGGC  
 2521 -----+-----+-----+-----+-----+-----+ 2580  
 GACGACCGGGAGACCCCTGGACTGGGTCCGGCTCGGAAACACTTGGTTGTGGACACGCCG  
 TCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACC  
 2581 -----+-----+-----+-----+-----+-----+ 2640  
 AGTGTGGACCACCTTCGAGAGATGGATCACACGCCCTTGCTCCGAAGAAGATGTGTGGG  
 AAGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCATTTGCTGCCCCCTGG  
 2641 -----+-----+-----+-----+-----+-----+ 2700  
 TTCTGGGCGGCCCTCCGTCTCCTGGACGTCCCACTCGGTGGCGGGTAACGACGGGGACC  
 CCGCCCCCAGCCACCCCTGCTCCTGGCGCTCCCAACCAGCATGGGCAGAAGGGGGCAGG  
 2701 -----+-----+-----+-----+-----+-----+ 2760  
 GCGGGGGTTCGGTGGGGGACGAGGACCGCGAGGGTGGGTTCGTACCCGTCTTCCCCGTCC  
 AGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAAGTTCTCTTGGTCACG  
 2761 -----+-----+-----+-----+-----+-----+ 2820  
 TCCGACGGTGGGTCTGTCCTCCAGTCCACGTGAAAAAATTTTCTTCAAGAGAACCAGTGC  
 HinfI  
 TCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGAGGACGGTGTGG  
 2821 -----+-----+-----+-----+-----+-----+ 2880  
 AGGATTTTCACTGGTCGAGGGACACCGGGTCAGTCTTAGAGTCGGAATCCTGCCACAACC

CTTCGGCAGCCCCGAGATACATCAGAGGGTGGGCACGCTCCTCCCTCCACTCGCCCCTCA  
2881 -----+-----+-----+-----+-----+-----+ 2940  
GAAGCCGTCGGGGCTCTATGTAGTCTCCCACCCGTGCGAGGAGGGAGGTGAGCGGGGAGT

## Appendix 7

### ***Allele diversity at the insulin minisatellites***

All 189 different MVR codes identified from within a Caucasian cohort are presented in 5' to 3' orientation with the insulin gene to the right. Allele names reflect lineage, repeat number, and a further discriminator. Hyphens were inserted to improve alignments. 'o' denotes unamplifiable variant repeats. The number of copies of each allele detected in this study of 876 alleles is presented. Alleles R188.1 and R213.1 were divided over two lines due to their size. Variant repeats are as defined in Table 8.2. A-type repeats are presented in green, B in red, C in blue, E in cyan, F in yellow, H in pink, and null repeats in black.

# IC 189 alleles, 59 different codes

Name	Copies	
IC28.1	1	CBoBoAFAAAAAA-----BABABC-AAA-BBB
IC30.1	1	CBoBoAFAAAAAA-----BABABC-AAA-BBB
IC30.2	7	CBoBoAFAAAAA-----BAAA-BABABC-AAAABBB
IC30.3	6	CBoBoAFAAAAA-----BAAAABABABC-AAA-BBB
IC30.4	1	CBoBoAFAAAAA-----BAAA-BABABC-AAA-BBB
IC30.5	1	CBoBoFAFAAAAA-----BAAAABABABC-AA--BBB
IC30.6	3	CBoBoAFAAAAA--CAAAAA-----BABABC-AAA-BBB
IC30.7	1	CBoBoAFAAAAA--CAAAAA-----BABABC-AAA-B-B
IC31.1	1	CBoBoAFAAAAA-----AAAABABABC-AAA-BBB
IC31.2	16	CBoBoAFAAAAA-----BAAA-BABABC-AAAABBB
IC31.3	1	CBoBoAFAAAAA-----BAAAABABABC-AAA-BBB
IC31.4	1	CBoBoAFAAAAA-----BAAA-BABABC-AAA-BBB
IC31.5	14	CBoBoAFAAAAA--CAAAAA-----BABABC-AAA-BBB
IC31.6	2	CBoBoAFAAAAA--CAAAAA-----BABABC-AA--BBB
IC31.7	1	CBoBoAFAAAAA--CAAAAA-----BABABCCAAA-B-B
IC32.1	7	CBoBoAFAAAAA-----BAAAABABABC-AAAABBB
IC32.2	1	CBoBoAFAAAAA-----BAAAABABABCCAAA-BBB
IC32.3	2	CBoBoAFAAAAA-----BAAA-BABABC-AAAABBB
IC32.4	4	CBoBoAFAAAAA-----BAAAABABABC-AAA-BBB
IC32.5	1	CBoBoAFAAAA--CAAAAAA-----BABABC-AAA-BBB
IC32.6	12	CBoBoAFAAAA--CAAAAAA-----BABABC-AAA-BBB
IC32.7	8	CBoBoAFAAAA--CAAAAA-----BABABCCAAA-B-B
IC32.8	1	CBoBoAFAAAAA-----BAAA-BABABC-AAAABBB
IC33.1	8	CBoBoAFAAAAA-----BAAAABABABC-AAAABBB
IC33.2	1	CBoBoAFAAAAA-----BAAAABABABCCAAA-BBB
IC33.3	2	CBoBoAFAAAAA-----BAAAABABABC-AAA-BBB
IC33.4	6	CBoBoAFAAAA--CAAAAAA-----BABABC-AAA-BBB
IC33.5	2	CBoBoAFAAAAA--CAAAAA-----BABABCCAAA-BBB
IC34.1	10	CBoBoAFAAAAAA-----BAAAABABABC-AAAABBB
IC34.2	3	CBoBoAFAAAAAA-----BAAAABABABCCAAA-BBB
IC34.3	1	CBoBoFAFAAAAA--CAAAAA-----BAAAABABABC-AAAAB-B
IC34.4	1	CBoBoAFAAAAA--CAAAAA-----BAAA-BABABC-AAA-BBB
IC34.5	1	CBoBoAFAAAAA--CAAA-----BAAAABABABC-AAA-BBB
IC35.1	1	CBoBoAFAAAAAA-AA-----BAAAABABABC-AAAABBB
IC35.2	1	CBoBoAFAAAAAA-AA-----BAAAABABABCCAAA-BBB
IC35.3	1	CBoBoAFAAAAAA-AAA-----BAAAABABABC-AAA-BBB
IC35.4	5	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AAA-BBB
IC35.5	1	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AA--BBB
IC35.6	1	CBoBoAFAAAAA--CAAAA-----BAAAABABABC-AAAAB-B
IC36.1	2	CBoBoAFAAAAAA-AAA-----BAAAABABABC-AAAABBB
IC36.2	2	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AAAABBB
IC36.3	1	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AAAABoB
IC36.4	2	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AAA-BBB
IC36.5	1	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AAAAB-B
IC36.6	1	CBoBoAFAAAAA--CAAAAA-----BAAAABABABCCAAA-B-B
IC36.7	1	CBoBoAFAAAAA--CAAAA-----BAAAABABABC-AAA-BBB
IC36.8	1	CBoBoAFAAAAAA-CAAAA-----BAAAABABABC-AA--BBB
IC37.1	7	CBoBoAFAAAAA--CAAAAA-----BAAAABABABCCAAA-BBB
IC37.2	2	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AAA-BBB
IC37.3	4	CBoBoAFAAAAA--CAAAA-----BAAAABABABC-AAAABBB
IC37.4	1	CBoBoAFAAAAA--CAA--CAAAA-BAAAABABABC-AA--BBB
IC38.1	2	CBoBoAFAAAAA--CAAAAA-----BAAAABABABC-AAAABBB
IC38.2	6	CBoBoAFAAAAAA-CAAAA-----BAAAABABABC-AAAABBB
IC38.3	13	CBoBoAFAAAAA--CAAA--CAAAA-BAAAABABABC-AA--BBB
IC39.1	1	CBoBoAFAAAAA--CAAA--CAAAA-BAAAABABABC-AAAAB-B
IC41.1	1	CBoBoAFAAAAA--CAAAoCAAAA-BAAAABABABC-AAAAB-B
IC41.2	1	CBoBoAFAAAAA--CAAAA-CAAAA-BAAAABABABC-AAAAB-B
IC43.1	1	CBoBoAFAAAA--CAAAA-CAAAA-BAAABAAAABABABCAAAABBB
IC60.1	1	CBoBoAFAAAA--CAAAAAAABABABCACAAAAABABABACAAAAABABABCAAAABBB

## ID 511 alleles, 36 different codes

Name	Copies	
ID37.1	4	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC--AAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID38.1	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAA-BAAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID38.2	2	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> B-B
ID38.3	6	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC--AAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID38.4	1	CB <sub>o</sub> Bo <sub>o</sub> F--AAACACAAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> B-B
ID39.1	6	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID39.2	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID39.3	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAA-BAA--BABABC <sub>A</sub> AA <sub>F</sub> B-B
ID39.4	21	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAA-BAAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID39.5	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAA-BAAAA--BABC <sub>A</sub> AA <sub>F</sub> B-B
ID39.6	8	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> B-B
ID40.1	2	CB <sub>o</sub> Bo <sub>o</sub> F--AAACACAAAA--FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> B-B
ID40.2	112	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID40.3	3	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> B-B
ID40.4	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FA-AAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID41.1	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC--AAAA--FA-AAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID41.2	2	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAA-BAAAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> B-B
ID41.3	2	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA-BABABC <sub>A</sub> AA <sub>F</sub> B-B
ID41.4	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> B-B
ID41.5	10	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FACAAAA <sub>B</sub> AAA--BABC <sub>A</sub> AA <sub>F</sub> BBB
ID42.1	2	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC--AAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID42.2	6	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC-CAAAA--FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID42.3	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAA-BAAAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID42.4	184	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> B-B
ID42.5	2	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FACAAA-BAAAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> B-B
ID43.1	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAA-CACAAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID43.2	2	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC--AAAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID43.3	5	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC-CAAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID43.4	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAAC-CAAAAA-FA-AAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID43.5	10	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAA--FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID43.6	2	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FACAAAA <sub>B</sub> AAA-BABABC <sub>A</sub> AA <sub>F</sub> BBB
ID43.7	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID43.8	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABAB-AAA <sub>F</sub> BBB
ID43.9	37	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> B-B
ID44.1	69	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACACAAAAA-FACAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB
ID44.2	1	CB <sub>o</sub> Bo <sub>o</sub> FAFAAACCCAAAAA-FAAAAA <sub>B</sub> AAA <sub>B</sub> ABABC <sub>A</sub> AA <sub>F</sub> BBB





IIIB 43 alleles, 14 different codes

R 7 alleles, 7 different codes

[illegible]

## Appendix 8

### ***De novo mutation at the insulin minisatellite***

All mutants isolated from sperm DNA by SESP-PCR (Jeffreys and Neumann, 1997) of three donors and blood DNA of one donor were analysed by MVR-PCR. Mutant alleles are aligned with the progenitor molecule using hyphens to denote deletions, and by dividing alleles over multiple lines to clarify the nature of the duplication. For putative inter-allelic repeat transfers, repeats gained by the class I progenitor from the class III progenitor alleles are underlined. Mosaic mutants are indicated by \* and where necessary grouped by vertical lines. Mutant names reflect the donor (1-3) and tissue (blood (B) or sperm (S)) from which the mutants was isolated, repeat number, and a further discriminator. Variant repeats are as defined in Table 8.2.



## Donor 1

Donor 2

Donor 3



# Class I germine deletion mutants

## Donor 1

Progenitor CBoBoFAFAAACACAAAAFACAAAABAAABABCAAAFBBB

S1-10.1 C-----BCAAAFBBB  
 S1-24.1 CB-----FACAAAABAAABABCAAAFBBB  
 S1-38.1 CBoB--AFAAACACAAAAFACAAAABAAABABCAAAFBBB  
 S1-37.1\* CBoBoF---AACACAAAAFACAAAABAAABABCAAAFBBB  
 S1-37.2\* CBoBoF---AACACAAAAFACAAAABAAABABCAAAFBBB  
 S1-38.2\* CBoBoFAFAAAC--AAAAFACAAAABAAABABCAAAFBBB  
 S1-38.3\* CBoBoFAFAAAC--AAAAFACAAAABAAABABCAAAFBBB  
 S1-38.4\* CBoBoFAFAAAC--AAAAFACAAAABAAABABCAAAFBBB  
 S1-38.5\* CBoBoFAFAAAC--AAAAFACAAAABAAABABCAAAFBBB  
 S1-38.6\* CBoBoFAFAAAC--AAAAFAFAAAABAAABABCAAAFBBB  
 S1-38.7\* CBoBoFAFAAAC--AAAAFAFAAAABAAABABCAAAFBBB  
 S1-38.8\* CBoBoFAFAAAC--AAAAFAFAAAABAAABABCAAAFBBB  
 S1-38.9 CBoBoFAFAAAC--AAAAFAAAAAABAAABABCAAAFBBB  
 S1-38.10 CBoBoFAFAAACACAAAA--CAAAABAAABABCAAAFBBB  
 S1-34.1 CBoBoFAFAAACACAAAAFACA-----oBABCAAAFBBB  
 S1-36.1 CBoBoFAFAAACACAAAAFACAAA-----ABABCAAAFBBB  
 S1-38.11 CBoBoFAFAAACACAAAAFACAAAA--AABABCAAAFBBB  
 S1-33.1 CBoBoFAFAAACACAAAAFACAAAABA-----AAFBBB  
 S1-34.2 CBoBoFAFAAACACAAAAFACAAAABAA-----AAFBBB  
 S1-35.1 CBoBoFAFAAACACAAAAFACAAAABAAA-----AAFBBB  
 S1-32.1\* CBoBoFAFAAACACAAAAFACAAAABAAABA-----B  
 S1-32.2\* CBoBoFAFAAACACAAAAFACAAAABAAABA-----B

## Donor 2

Progenitor CBoBoFAFAAACACAAAAFACAAAABAAABABABCAAAFB

S2-40.1 CBoBoF--AAACACAAAAFACAAAABAAABABABCAAAFB  
 S2-39.2 CBoBoFAFAA---CAAAAFACAAAABAAABABABCAAAFB  
 S2-40.2\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S2-40.3\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S2-40.4 CBoBoFAFAAACACAA--FACAAAABAAABABABCAAAFB  
 S2-40.5\* CBoBoFAFAAAC--AAAAFAFAAAABAAABABABCAAAFB  
 S2-40.6\* CBoBoFAFAAAC--AAAAFAFAAAABAAABABABCAAAFB  
 S2-40.7 CBoBoFAFAAACAC--AAFAFAAAABAAABABABCAAAFB  
 S2-40.8 CBoBoFAFAAACACAAAA--CAAAABAAABABABCAAAFB  
 S2-35.1\* CBoBoFAFAAACACAAAA-----BAAABABABCAAAFB  
 S2-35.2\* CBoBoFAFAAACACAAAA-----BAAABABABCAAAFB  
 S2-37.1\* CBoBoFAFAAACACAAAAFACAAAA-----BABABCAAAFB  
 S2-37.2\* CBoBoFAFAAACACAAAAFACAAAA-----BABABCAAAFB  
 S2-37.3\* CBoBoFAFAAACACAAAAFACAAAA-----BABABCAAAFB  
 S2-33.1 CBoBoFAFAAACACAAAAFACAAAA-----BCAAAFBB  
 S2-40.9 CBoBoFAFAAACACAAAAFACAAAABAA--BABABCAAAFB  
 S2-40.10\* CBoBoFAFAAACACAAAAFACAAAABAAA--BABCAAAFB  
 S2-40.11\* CBoBoFAFAAACACAAAAFACAAAABAAA--BABCAAAFB

## Donor 3

Progenitor CBoBoFAFAAACACAAAAFACAAAABAAABABABCAAAFB

S3-36.1 CBoBoFAFAA-----FACAAAABAAABABABCAAAFB  
 S3-24.1 CBoBoFAFAA-----BABABCAAAFB  
 S3-42.1\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S3-42.2\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S3-42.3\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S3-42.4\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S3-42.5\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S3-42.6\* CBoBoFAFAAAC--AAAAFACAAAABAAABABABCAAAFB  
 S3-30.1 CBoBoFAFAACA-----AAAABABABCAAAFB  
 S3-41.1 CBoBoFAFAACA?A---AFACAAAABAAABABABCAAAFB  
 S3-42.7 CBoBoFAFAAACACAA--AFACAAAABAAABABABCAAAFB  
 S3-23.1 CBoBoFAFAAACACAAAAF-----BBB  
 S3-36.2 CBoBoFAFAAACACAAAAFACAAA-----BABCAAAFB  
 S3-35.1 CBoBoFAFAAACACAAAAFACAAAA-----BCAAAFBB  
 S3-32.1 CBoBoFAFAAACACAAAAFACAAAA-----AAFBB  
 S3-42.8 CBoBoFAFAAACACAAAAFACAAAABA--ABABABCAAAFB  
 S3-35.2 CBoBoFAFAAACACAAAAFACAAAABAAA-----BBB

### Class III germline deletion mutations

## Donor 1

[illegible]

### Donor 2

Progenitor CBoBoBAo AEAACCEAAACA AAAA AC EAACACA ACECCA AoBoBAo ACEAAACAAA ACAACACAAAA CAAA ACEoEA AEA ACAA ACEFAEABAAAAA AAA AAAABCAA AAAAA AABoCAoAAAAAB  
S2-70.1 CBoBoBAo AEAACCEAAACA AAAA AC EAACACA ACEC -----AAABCAA AAAAA AABoCAoAAAAAB  
S2-58.1 CBoBoBAo AEAACCEAAACA AAAA A-----AAAABCAA AAAAA AABoCAoAAAAAB

## Donor 1

Progenitor CB<sup>o</sup>Bo<sup>o</sup>A<sup>o</sup>AAACACAAAA<sup>o</sup>ACAAAA<sup>o</sup>BAAAB<sup>o</sup>ABC<sup>o</sup>AAA<sup>o</sup>BBB<sup>o</sup>  
S1-44.1 CB<sup>o</sup>Bo<sup>o</sup>A<sup>o</sup>AAACACAAAA<sup>o</sup>ACAAAA<sup>o</sup>  
AAAA<sup>o</sup>BAAAB<sup>o</sup>ABC<sup>o</sup>AAA<sup>o</sup>BBB<sup>o</sup>

Progenitor CBoBoAFAAACACAAAAACAAAAABAAABABCAAA BBB  
S1-43.1 CBoBoAFAAACACAAAAACAAAAABAAA  
AAABABCAAA BBB

Progenitor CBoBoAFAAACACAAAAACAAAAABAAABABCAAAABBB  
S1-42.1 CBoBoAFAAACAC  
ACAAAAACAAAAABAAABABCAAAABBB

## Donor 2

Progenitor CBBoAFAAACACAAAAACAAAABAAAABABABCAAAFB

S2-54.1\* CBBoAFAAACACAAAAACAA  
CACAAAAACAAAABAAAABABABCAAAFB

Progenitor CB<sub>0</sub>Bo<sub>0</sub>F<sub>0</sub>AAACACAAAA<sub>0</sub>ACAAAA<sub>0</sub>BAAAA<sub>0</sub>BABABCAAA<sub>0</sub>BB  
S2-54.2\* CB<sub>0</sub>Bo<sub>0</sub>F<sub>0</sub>AAACACAAAA<sub>0</sub>ACAA  
CACAAAA<sub>0</sub>ACAAAA<sub>0</sub>BAAAA<sub>0</sub>BABABCAAA<sub>0</sub>BB

Progenitor CBBoA AAACACAAAA ACAAAABAAAABABABCAAA BB  
S2-49.1 CBBoA AAACACAAAA ACAAAABAAAABAB  
AAAAABABABCAAA BB

Progenitor CB<sup>o</sup>Bo<sup>o</sup> A<sup>o</sup> AAACACAAAA<sup>o</sup> ACAAAA<sup>o</sup>BAAAA<sup>o</sup>BABABCAAA<sup>o</sup> BB<sup>o</sup>  
S2-48.1 CB<sup>o</sup>Bo<sup>o</sup> A<sup>o</sup> AAACACAAAA<sup>o</sup> ACAAAA<sup>o</sup>BAAAA<sup>o</sup>  
ABAAAA<sup>o</sup>BABABCAAA<sup>o</sup> BB<sup>o</sup>

Progenitor CBoBo-AFAAACACAAAAACAAAABAAAABABABCAAABB  
S2-47.1 CBoBo-AFAAACACAAAAACAAAABAAAABABABCAAACAAAB

Progenitor CBoBoFAFAAACACAAAAACAAAABAAAABABABCAAAFB

S2-46.1 CBoBoFAFAAACACAAAAACAAAABAAAABABAB  
ABABCAAAFB

Progenitor CBoBoFAAAACACAAAACAAAABAAAABABABCAAAABB  
S2-45.1 CBoBoFAAAACACAAAACAAAAB  
AABAAAABABABCAAAABB

Progenitor CBBoBoFAFAAACACAAAAACAAAABAAAABABABCAAAABB  
S2-45.2 CBBoBoFAFAAACACAAAAACAAAABAAAABABA  
ABABCAAAABB



# Simple germline duplications (continued)

## Donor 3

Progenitor S3-65.1	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAA FAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB
Progenitor S3-65.2	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFAC BoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB
Progenitor S3-62.1	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAA CAAAAFAAAAABAAAABABABCAAAFBBB
Progenitor S3-59.1	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAABABAB CAAAAABAAAABABABCAAAFBBB
Progenitor S3-54.1*	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAAB AAAAB AAAABABABCAAAFBBB
Progenitor S3-54.2*	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAAB AAAAB AAAABABABCAAAFBBB
Progenitor S3-54.3	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAA FACAAAABAAAABABABCAAAFBBB
Progenitor S3-53.1	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAA AAFAAAAABAAAABABABCAAAFBBB
Progenitor S3-52.1	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAABABABC BABABCAAAFBBB
Progenitor S3-51.1*	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAABA BAAAABABABCAAAFBBB
Progenitor S3-51.2*	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAABA BAAAABABABCAAAFBBB
Progenitor S3-51.3*	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAABA BAAAABABABCAAAFBBB
Progenitor S3-51.4	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAAB ACAAAABAAAABABABCAAAFBBB
Progenitor S3-50.1	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACAC AAACACAAAAAFACAAAABAAAABABABCAAAFBBB
Progenitor S3-49.1	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAAB AAAABABABCAAAFBBB
Progenitor S3-49.2	CBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAAFBBB CBoBoFAFAAACACAAAAAFACAAAABAAAABAB AABABABCAAAFBBB

# Complex intra-allelic germline duplications

## Donor 1

Progenitor S1-68.1  
 CBoBoFAAAACACAAAAACAAAABAAABABCAAA BBB  
 CBoBoFAAAACACAAAAACAAAABAAABABCAB  
 FAAACACAAAAACAAAABAAABABCAAA BBB

Progenitor S1-67.1  
 CBoBoFAAAACACAAAAACAAAABAAABABCAAA BBB  
 CBoBoFAAAACACAAAAACAAAABAAA--BCAAA B  
 AB--BABCAA  
 BAAABCAA BABCAA BBB

Progenitor S1-60.1  
 CBoBoFAAAACACAAAAACAAAABAAABABCAAA BBB  
 CBoBoFAAAACACAAAA  
 AAAAAACAAAA  
 AAAAAAAACAAAABAAABABCAAA BBB

Progenitor S1-52.1  
 CBoBoFAAAACACAAAAACAAAABAAABABCAAA BBB  
 CBoBoFAAAACACAAAAACAAAABA--BAB  
 ABA--BAB-AA  
 AABABCAA BBB

Progenitor S1-47.1  
 CBoBoFAAAACACAAAAACAAAABAAABABCAAA BBB  
 CBoBoFAAAACACAAAAACAAAA--ABAB  
 AABABABABCAAA BBB

Progenitor S1-44.2  
 CBoBoFAAAACACAAAAACAAAABAAABABCAAA BBB  
 CBoBoFAAAACACAAAAACAAAABAAABABCAAA BB  
 BCA---BB

## Donor 2

Progenitor S2-80.1  
 CBoBoFAAAACACAAAAACAAAABAAABABA--BCAAA BB  
 CBoBoFAAAACACAAAAACAAAABAAABABABABC  
 -AAAABABABABC  
 AAAAACAAAABAAABABABABCAAA BB

Progenitor S2-57.1  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACACAAAAACAA  
 -AA AA  
 CACAAAABAAABABABCAA FBB

Progenitor S2-57.2  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACACAAAAACAAAA  
 AAFAA-C-CAAAA---AABAAAABABABCAA FBB

Progenitor S2-54.3  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACACAAAAACAA  
 F-----CAAAA AAAAABAAABABABCAA FBB

Progenitor S2-54.4  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACACAAAAACAAAABAAABABA  
 AAAAAACAAAA-----CAAA FBB

Progenitor S2-49.2  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACACAAAAACAAAABAAABABAB  
 AAAABA--BCAAA FBB

Progenitor S2-49.3  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACACAAAAAC-AAA BAA-B  
 AABAAAABABABCAA FBB

Progenitor S2-48.2  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACAC  
 CA-AAAA  
 C-AAAACAAAABAAABABABCAA FBB

Progenitor S2-48.3  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACA-AAAA  
 CACCAAAACAAAABAAABABABCAA FBB

Progenitor S2-48.4  
 CBoBoFAAAACACAAAAACAAAABAAABABABCAA FBB  
 CBoBoFAAAACACAAAAACAAAA BAA  
 CAAAA BAA--BABABCAA FBB

# Complex intra-allelic germline duplications (continued)

## Donor 3

Progenitor S3-109.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>B  
 AA<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>B  
 A-AAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-100.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA  
 CA-----C<sup>A</sup>AAAA<sup>B</sup>AAAA  
 CA-----C<sup>A</sup>AAAA<sup>B</sup>AAAA  
 CA-----C<sup>A</sup>AAAA<sup>B</sup>AAAA  
 CACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-93.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA  
 AAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA  
 CAAAA  
 CAAAAAAAC  
 A<sup>F</sup>  
 AAAA<sup>F</sup>AAAA  
 CAAAAAA  
<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-90.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB  
 AAAA<sup>B</sup>AAA-B  
 CAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>  
 AABAB  
 AAAB  
 CAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-90.2 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>  
 -AAAA<sup>B</sup>BAB<sup>B</sup>C  
 AA<sup>F</sup>ACAAA  
 AA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>  
 AAAABAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-89.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BBAAA<sup>F</sup>AABA  
 AA<sup>F</sup>  
 AA<sup>F</sup>AABA  
 oBo<sup>o</sup>FA<sup>F</sup>AAA-----<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-88.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>AAAA<sup>B</sup>ABAA  
 CAAAA-AAAA-----CAA  
 CAAAA<sup>B</sup>AAAA  
 AABAAA-BA  
 AAAABAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-86.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>  
 AAA<sup>F</sup>ACAAAA<sup>B</sup>  
 AAA<sup>F</sup>ACAAAA<sup>B</sup>-----ABAB<sup>B</sup>CAAA<sup>F</sup>  
 AAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-86.2 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>AB  
 BAAAAA  
 AAAAA<sup>F</sup>A  
 ACAA--<sup>F</sup>  
 CACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-83.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>ABABA  
<sup>F</sup>ACAA--BAAAA  
<sup>F</sup>ACAAAA-AAAA<sup>B</sup>AB  
 ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB

Progenitor S3-78.1 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB  
 CB<sup>o</sup>Bo<sup>o</sup>FA<sup>F</sup>AAACACAAAA<sup>F</sup>ACAAAA<sup>B</sup>AAAA<sup>B</sup>BAB<sup>B</sup>BCAA  
 BAAAA  
 CAAAA<sup>B</sup>  
 ACAA--<sup>B</sup>  
 AAAABAAA<sup>B</sup>BAB<sup>B</sup>BCAAA<sup>F</sup>BBB



Progenitor CBBoBoFAFAAACACAAAAAACAAAABAAAABABABCAAAFBBB  
S3-77.1 CBBoBoFAFAAACACAAAAAACAAAABAAAABABA  
F----AABA  
AAACACAAAAAACAAAABAAAABABABCAAAFBBB

Progenitor CBBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAA BBB  
S3-73.1 CBBoBoFAFAAACACAAAAAFACAAAABAAAABABABCAAA  
BABABCAAA  
BA---BABABCAAA  
BABABCAAA BBB

Progenitor S3-72.1

CB<sup>o</sup>Bo<sup>+</sup>F<sup>+</sup>AACAC<sup>+</sup>CAAAAA<sup>+</sup>ACAAAA<sup>+</sup>BAAAA<sup>+</sup>BABA<sup>+</sup>BCAAA<sup>+</sup>BBB<sup>+</sup>

CB<sup>o</sup>Bo<sup>+</sup>F<sup>+</sup>AFA<sup>+</sup>----CAAA<sup>+</sup>

CACAA<sup>+</sup>

CAACAA<sup>+</sup>

FAAACAA<sup>+</sup>CAA<sup>+</sup>

FAAACACAAAAA<sup>+</sup>ACAAAA<sup>+</sup>BAAAA<sup>+</sup>BABA<sup>+</sup>BCAAA<sup>+</sup>BBB<sup>+</sup>

Progenitor CBoBo A AAACACAAAAA ACAAAA BAAAA BABA - BCAAA BBB  
S3-69.1 CBoBo A AAACACAAAAA ACAAAA BAAAA BABAB CAAA  
ACA AAA BAAAA BABA B CAAA BBB

Progenitor CBoBoFAFAAACACAAAAAACAAAAAABABABCAAAFBBB  
S3-67.1 CBoBoFAFAAACACAAAAAACAAAAAABABABCAAA  
ABA  
ACAAAAAABABABCAAAFBBB

Progenitor CB<sup>o</sup>Bo<sup>o</sup>F<sup>o</sup>FAAACACAAAAA<sup>o</sup>ACAAAA<sup>o</sup>BAAAA<sup>o</sup>BABABCAA<sup>o</sup>BBB<sup>o</sup>  
S3-67.2 CB<sup>o</sup>Bo<sup>o</sup>F<sup>o</sup>FAAACACAAAAA<sup>o</sup>ACAAAA<sup>o</sup>B  
A<sup>o</sup>ACAA<sup>o</sup>  
A<sup>o</sup>ACAA<sup>o</sup>  
A<sup>o</sup>ACAAAA<sup>o</sup>BAAAA<sup>o</sup>BABABCAA<sup>o</sup>BBB<sup>o</sup>

Progenitor CBBoFAAAACACAAAAACAAAABAAAABABABCAAABBB  
S3-61.1 CBBoFAAAACACAAAAACAAAABAAAABABABCAA  
BCAA  
BAAAABABABCAAABBB

Progenitor CBoBoAFAAACACAAAAACAAAAAABABABCAAA BBB  
S3-60.1 CBoBoAFAAACACAAAAACAAAAAABABA  
AAAAAABAAAAABABCAAA BBB

Progenitor CBoBoFAFAAACACAAAAAACAABABABCAAAABBE  
S3-60.2 CBoBoFAFAAACACAAAAAACAABABAB  
ACAABABABCAAAABBE

Progenitor CBoBoAFAAACACAAAAAFAAAAAABAAAAABABABCAAA BBB  
S3-55.1 CBoBoAFAAACACAAAAA  
BoBoAFAAACACA-----FAAAAAABAAAAABABABCAAA BBB

Progenitor CB<sub>0</sub>Bo<sub>0</sub>FAFAAACACAAAAAACAACAAABAAAAABABABCAAA<sub>0</sub>BBB<sub>0</sub>  
 S3-54.4 CB<sub>0</sub>Bo<sub>0</sub>FAFAAACACAAAAA  
 A<sub>0</sub>-----AB  
 CAAAAA<sub>0</sub>ACAACAAABAAAAABABABCAAA<sub>0</sub>BBB<sub>0</sub>

Progenitor CB<sub>0</sub>Bo<sub>0</sub>TA<sub>0</sub>AAACACAAAAA<sub>0</sub>ACAAAA<sub>0</sub>BAAAA<sub>0</sub>BABABCAAA<sub>0</sub>BBB<sub>0</sub>  
 S3-54.5 CB<sub>0</sub>Bo<sub>0</sub>TA<sub>0</sub>AAACACAAAAA<sub>0</sub>ACAAAA<sub>0</sub>BAAA<sub>0</sub>  
 BAAA<sub>0</sub>  
 AA<sub>0</sub>BAAAA<sub>0</sub>BABABCAAA<sub>0</sub>BBB<sub>0</sub>

Progenitor CBoBoFAAAACACAAAAACAAAAABAAAABABABCAAA BBB  
S3-52.2 CBoBoFAAAACACAAAAA  
ACAAAAACAAAAABAAAABABABCAAA BBB

Progenitor CBoBoFAAAACACAAAAACAAAAABAAAABABABCAAABBB  
S3-52.3 CBoBoFAAAACACAAAAACAAAAABAAAABABABC  
BA---BABABCAAABBB

Progenitor CB<sup>o</sup>Bo<sup>+</sup>FA<sup>+</sup>AAACACAAAAA<sup>+</sup>ACAAAA<sup>+</sup>BAAAA<sup>+</sup>BABABC<sup>+</sup>AAA<sup>+</sup>BBB<sup>+</sup>  
S3-50.2 CB<sup>o</sup>Bo<sup>+</sup>FA<sup>+</sup>AAACACAAAAA<sup>+</sup>ACAAAA<sup>+</sup>BAAAA<sup>+</sup>BABABC<sup>+</sup>AAA<sup>+</sup>  
BABABC<sup>+</sup>AAA<sup>+</sup>---

# Inter-allelic germline mutations

## Donor 1

Progenitor .....CEAAABCEAAAAACAAAABAAAEEAABAAABAAFAAAFAABoHBBB  
 Progenitor CBoBoFAFAAACACAAAAFA-----CAAAAABAAABABCAAAFB  
 S1-51.1 CBoBoFAFAAACACAAAAFAAAEAAACEAAACAAAABAAABABCAAAFB

Progenitor .....BCEAAAAACAAAABAAAEEAABAAABAAFAAAFAABoHBBB  
 Progenitor CBoBoFAFAAACACAAAAFACAAAABAAAAB-----BCAAFB  
 S1-46.1 CBoBoFAFAAACACAAAAFACAAAABAAAABAAFAAAABCAAAFB

## Donor 2

Progenitor .....EoEAFAEAFAACAAACEAEAAABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAAABAAAAB-----ABABCAAAFB  
 S2-66.1 CBoBoFAFAAACACAAAAFACAAAABAAAABAAFACAAAAABAAAABAAFAABABABCAAAFB

Progenitor .....FACAAACEAEAAABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAAABAAAAB-----BABABCAAAFB  
 S2-58.2 CBoBoFAFAAACACAAAAFACAAAABAAAFAACAAAABAAAFAABABABCAAAFB

Progenitor .....CAAACEAEAAABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAA-----BAAAABABABCAAAFB  
 S2-56.1 CBoBoFAFAAACACAAAAFACAAAAAAAABCAAAAABAAAABABABCAAAFB

Progenitor .....AAACEAEAAABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAAABAAAAB-----ABABCAAAFB  
 S2-55.1 CBoBoFAFAAACACAAAAFACAAAABAAAABAACAAAAFAABABABCAAAFB

Progenitor .....AACEAEAAABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAA-----BAAAABABABCAAAFB  
 S2-54.5 CBoBoFAFAAACACAAAAFACAAAFACAAAAFAAABAAAABABABCAAAFB

Progenitor .....CEAEAAABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAAABAAAABABABCAAAFB-----B  
 S2-51.1 CBoBoFAFAAACACAAAAFACAAAABAAAABABABCAAAFAACaOAAAAAB

Progenitor CBoBoBAoFAEAAACEAAACAAFAAAFACEAAACACAFACECC.....  
 Progenitor CBoBo-----FAFAAACACAAAAFACAAAABAAAABABABCAAAFB  
 S2-48.5 CBoBoBAAAAAFAFAAACACAAAAFACAAAABAAAABABABCAAAFB

Progenitor .....AABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAAABAAAAB-----BABABCAAAFB  
 S2-47.2 CBoBoFAFAAACACAAAAFACAAAABAAAAAAFAABABABCAAAFB

Progenitor .....ABAAAAFAAAFAAABCAAFAAAFAABoACaOAAAAAB  
 Progenitor CBoBoFAFAAACACAAAAFACAAAABAAAABABAB-----CAAAFB  
 S2-46.2 CBoBoFAFAAACACAAAAFACAAAABAAAABABABoACAAAAFB

## Appendix 9

### ***Genotypes within each type 1 diabetes affected sib pair family***

Genotype data at the +3580MspI<sup>+</sup> polymorphism and the insulin minisatellite from each of the 219 type 1 diabetes affected sib pair families is presented. Allele names at the insulin minisatellite are as described in Appendix 7. At the minisatellite, for both fathers (reference .1) and mothers (reference .2) alleles are presented with the smaller allele to the left. For both children (references .3 and .4) alleles are presented with the paternally derived allele to the left. The number of times that each parental allele is transmitted to affected offspring is indicated on the right. H denotes parental genotypes homozygous both for allele length and MVR code at the minisatellite, preventing the number of transmissions of specific copies of the allele to affected offspring from being determined. Offspring which were homozygous at the minisatellite are indicated by \*. The -23HphI site was also analysed in all samples. With the single exception of allele R42.1, all class I alleles were -23HphI<sup>+</sup> and all class III alleles were -23HphI<sup>-</sup>.

DNA	Msp I	MVR		
3.1	+/+	ID40.2	IIIA148.5	2 0
3.2	+/-	ID43.9	IIIA147.2	2 0
3.3	+/-	ID40.2	ID43.9	
3.4	+/-	ID40.2	ID43.9	
13.1	+/-	IC36.1	ID42.4	1 1
13.2	+/+	IC34.1	IIIA150.5	0 2
13.3	+/-	ID42.4	IIIA150.5	
13.4	+/+	IC36.1	IIIA150.5	
7.1	+/-	IC34.2	ID42.2	2 0
7.2	+/-	ID40.2	ID44.1	2 0
7.3	+/+	IC34.2	ID40.2	
7.4	+/+	IC34.2	ID40.2	
9.1	+/+	IIIB144.1	IIIA150.6	2 0
9.2	-/-	ID43.5	ID44.1	1 1
9.3	+/-	IIIB144.1	ID44.1	
9.4	+/-	IIIB144.1	ID43.5	
14.1	+/+	IC34.2	IIIA150.8	0 2
14.2	+/-	IC32.4	ID42.4	2 0
14.3	+/+	IIIA150.8	IC32.4	
14.4	+/+	IIIA150.8	IC32.4	
23.1	-/-	ID42.4	ID44.1	1 1
23.2	+/-	ID39.6	ID42.1	1 1
23.3	-/-	ID42.4	ID44.1	
23.4	+/-	ID44.1	ID39.6	
26.1	+/-	ID42.4	IIIA145.2	2 0
26.2	+/-	ID40.2	ID42.4	2 0
26.3	+/-	ID42.4	ID40.2	
26.4	+/-	ID42.4	ID40.2	
44.1	+/-	ID40.2	ID42.4	1 1
44.2	+/+	IC32.1	ID40.2	0 2
44.3	+/-	ID42.4	ID40.2	
44.4	+/+	ID40.2	ID40.2	*
46.1	-/-	ID42.4	ID42.4	H H
46.2	+/+	IC30.3	ID41.5	2 0
46.3	+/-	ID42.4	IC30.3	
46.4	+/-	ID42.4	IC30.3	
49.1	+/+	ID38.1	ID40.2	2 0
49.2	+/-	IC32.3	ID44.1	1 1
49.3	+/-	ID38.1	ID44.1	
49.4	+/+	ID38.1	IC32.3	
146.1	+/-	IC31.2	ID42.4	2 0
146.2	+/-	IC33.1	ID42.4	1 1
146.3	+/-	IC31.2	ID42.4	
146.4	+/+	IC31.2	IC33.1	
51.1	+/-	ID40.2	ID43.9	1 1
51.2	+/+	ID39.6	ID39.2	2 0
51.3	+/-	ID43.9	ID39.6	
51.4	+/+	ID40.2	ID39.6	
53.1	+/-	IC38.3	ID44.2	1 1
53.2	+/-	ID42.4	IIIA146.2	2 0
53.3	+/-	IC38.3	ID42.4	
53.4	-/-	ID44.2	ID42.4	
55.1	+/-	IC33.4	ID44.1	1 1
55.2	+/-	ID40.2	ID43.9	0 2
55.3	+/-	IC33.4	ID43.9	
55.4	-/-	ID44.1	ID43.9	
79.1	+/+	IC37.1	ID40.2	1 1
79.2	+/-	ID40.2	ID44.1	0 2
79.3	+/-	ID40.2	ID44.1	
79.4	+/-	IC37.1	ID44.1	

DNA	Msp I	MVR		
80.1	+/+	IIIB145.1	IIIA158.4	1 1
80.2	+/-	ID42.4	IIIB144.1	0 2
80.3	+/+	IIIA158.4	IIIB144.1	
80.4	+/+	IIIB145.1	IIIB144.1	
81.1	-/-	ID38.4	ID42.4	2 0
81.2	+/-	ID42.4	IIIA159.1	2 0
81.3	-/-	ID38.4	ID42.4	
81.4	-/-	ID38.4	ID42.4	
93.1	+/-	IC33.1	ID42.4	1 1
93.2	-/-	ID42.4	ID42.4	H H
93.3	-/-	ID42.4	ID42.4	*
93.4	+/-	IC33.1	ID42.4	
98.1	+/+	IC33.4	IC38.2	2 0
98.2	+/+	IC38.2	IIIA150.4	2 0
98.3	+/+	IC33.4	IC38.2	
98.4	+/+	IC33.4	IC38.2	
102.1	+/-	ID39.4	ID40.3	1 1
102.2	-/-	ID41.1	ID44.1	1 1
102.3	+/-	ID39.4	ID41.1	
102.4	-/-	ID40.3	ID44.1	
103.1	+/+	IC34.2	ID40.2	2 0
103.2	+/+	IC31.1	IIIB144.1	1 1
103.3	+/+	IC34.2	IC31.1	
103.4	+/+	IC34.2	IIIB144.1	
109.1	+/+	IC32.4	R213.1	1 1
109.2	-/-	ID42.4	ID42.4	H H
109.3	+/-	IC32.4	ID42.4	
109.4	+/-	R213.1	ID42.4	
112.1	+/+	IC32.6	ID40.2	1 1
112.2	+/-	IC31.5	ID42.4	1 1
112.3	+/+	IC32.6	IC31.5	
112.4	+/-	ID40.2	ID42.4	
115.1	+/+	IC30.3	ID37.1	1 1
115.2	+/-	IC31.7	ID44.1	0 2
115.3	+/-	ID37.1	ID44.1	
115.4	+/-	IC30.3	ID44.1	
116.1	+/-	ID40.2	ID42.4	0 2
116.2	+/+	IC31.6	IC31.2	2 0
116.3	+/-	ID42.4	IC31.6	
116.4	+/-	ID42.4	IC31.6	
122.1	+/+	IC34.3	IC39.1	1 1
122.2	+/+	IC35.4	IC38.2	1 1
122.3	+/+	IC34.3	IC35.4	
122.4	+/+	IC39.1	IC38.2	
126.1	+/+	ID40.2	IIIA150.3	1 1
126.2	+/-	ID39.4	ID43.9	2 0
126.3	+/+	ID40.2	ID39.4	
126.4	+/+	IIIA150.3	ID39.4	
128.1	+/-	IC32.7	ID44.1	0 2
128.2	+/-	IC38.3	ID43.5	2 0
128.3	+/-	ID44.1	IC38.3	
128.4	+/-	ID44.1	IC38.3	
130.1	+/-	ID42.4	IIIA146.4	1 1
130.2	+/+	ID40.2	IIIA149.3	2 0
130.3	+/+	IIIA146.4	ID40.2	
130.4	+/-	ID42.4	ID40.2	
35.1	+/-	IC31.5	ID42.4	0 2
35.2	+/+	ID40.2	IIIA144.3	1 1
35.3	+/-	ID42.4	ID40.2	
35.4	+/-	ID42.4	IIIA144.3	

DNA	Msp I	MVR		
139.1	+/+	ID40.2	IIIB143.4	2 0
139.2	-/-	ID43.9	ID43.9	H H
139.3	+/+	ID40.2	ID43.9	
139.4	+/+	ID40.2	ID43.9	
140.1	+/+	ID44.1	IIIB142.1	0 2
140.2	+/+	IIIA148.9	IIIA148.9	H H
140.3	+/+	IIIB142.1	IIIA148.9	
140.4	+/+	IIIB142.1	IIIA148.9	
141.1	-/-	ID42.4	ID43.5	2 0
141.2	+/+	ID39.4	IIIB145.1	0 2
141.3	+/+	ID42.4	IIIB145.1	
141.4	+/+	ID42.4	IIIB145.1	
144.1	+/+	ID40.2	IIIB144.3	2 0
144.2	+/+	ID44.1	IIIB144.2	1 1
144.3	+/+	ID40.2	IIIB144.2	
144.4	+/+	ID40.2	ID44.1	
145.1	+/+	IC37.2	ID40.2	1 1
145.2	-/-	ID42.4	ID44.1	0 2
145.3	+/+	ID40.2	ID44.1	
145.4	+/+	IC37.2	ID44.1	
52.1	-/-	ID42.4	ID42.4	H H
52.2	+/+	IC33.5	ID42.4	2 0
52.3	+/+	ID42.4	IC33.5	
52.4	+/+	ID42.4	IC33.5	
133.1	+/+	IIIB144.1	IIIA150.3	0 2
133.2	+/+	IC36.3	ID39.4	1 1
133.3	+/+	IIIA150.3	IC36.3	
133.4	+/+	IIIA150.3	ID39.4	
165.1	+/+	IC32.6	ID42.4	1 1
165.2	+/+	IIIB144.1	IIIA149.7	1 1
165.3	+/+	IC32.6	IIIB144.1	
165.4	+/+	ID42.4	IIIA149.7	
167.1	-/-	ID42.4	ID42.4	H H
167.2	+/+	ID44.1	IIIA152.1	0 2
167.3	+/+	ID42.4	IIIA152.1	
167.4	+/+	ID42.4	IIIA152.1	
169.1	+/+	IC31.2	IIIA151.2	2 0
169.2	+/+	IC32.7	IIIA150.6	2 0
169.3	+/+	IC31.2	IC32.7	
169.4	+/+	IC31.2	IC32.7	
170.1	+/+	ID40.2	ID42.4	2 0
170.2	+/+	ID41.5	ID42.4	0 2
170.3	+/+	ID40.2	ID42.4	
170.4	+/+	ID40.2	ID42.4	
173.1	+/+	ID42.4	IC60.1	1 1
173.2	+/+	ID43.9	IIIA147.5	2 0
173.3	-/-	ID42.4	ID43.9	
173.4	+/+	IC60.1	ID43.9	
174.1	+/+	ID39.4	ID40.2	2 0
174.2	+/+	ID42.4	IIIA149.2	1 1
174.3	+/+	ID39.4	ID42.4	
174.4	+/+	ID39.4	IIIA149.2	
177.1	+/+	IC34.1	ID42.4	2 0
177.2	+/+	IIIA138.1	IIIB144.1	0 2
177.3	+/+	IC34.1	IIIB144.1	
177.4	+/+	IC34.1	IIIB144.1	
179.1	+/+	ID42.4	IIIA159.2	2 0
179.2	-/-	ID42.4	ID42.5	2 0
179.3	-/-	ID42.4	ID42.4	*
179.4	-/-	ID42.4	ID42.4	*

DNA	Msp I	MVR		
180.1	+/+	IC38.3	ID40.2	1 1
180.2	+/+	ID40.2	ID42.4	2 0
180.3	+/+	ID40.2	ID40.2	*
180.4	+/+	IC38.3	ID40.2	
18.1	+/+	IC34.1	ID43.3	1 1
18.2	+/+	IC30.6	ID44.1	1 1
18.3	+/+	IC34.1	IC30.6	
18.4	-/-	ID43.3	ID44.1	
31.1	-/-	ID44.1	ID44.1	H H
31.2	+/+	IC30.2	ID42.4	1 1
31.3	+/+	ID44.1	IC30.2	
31.4	-/-	ID44.1	ID42.4	
58.1	+/+	IC32.6	R34.1	1 1
58.2	+/+	ID40.2	IIIA148.9	2 0
58.3	+/+	IC32.6	ID40.2	
58.4	+/+	R34.1	ID40.2	
62.1	+/+	ID40.2	ID42.4	2 0
62.2	+/+	ID40.2	ID40.2	H H
62.3	+/+	ID40.2	ID40.2	*
62.4	+/+	ID40.2	ID40.2	*
96.1	-/-	ID42.4	ID43.5	1 1
96.2	-/-	ID43.9	ID43.9	H H
96.3	-/-	ID43.5	ID43.9	
96.4	-/-	ID42.4	ID43.9	
97.1	+/+	IC30.3	ID44.1	1 1
97.2	+/+	IC32.6	ID38.3	2 0
97.3	+/+	IC30.3	IC32.6	
97.4	+/+	ID44.1	IC32.6	
166.1	-/-	ID42.4	ID44.1	0 2
166.2	+/+	ID40.2	ID44.1	0 2
166.3	-/-	ID44.1	ID44.1	*
166.4	-/-	ID44.1	ID44.1	*
200.1	+/+	IC28.1	ID44.1	1 1
200.2	+/+	IC30.5	IIIA149.7	1 1
200.3	+/+	ID44.1	IIIA149.7	
200.4	+/+	IC28.1	IC30.5	
207.1	+/+	ID40.4	IIIA158.3	1 1
207.2	-/-	ID42.4	ID42.4	H H
207.3	+/+	IIIA158.3	ID42.4	
207.4	+/+	ID40.4	ID42.4	
212.1	+/+	IC31.2	ID42.4	0 2
212.2	-/-	ID42.4	ID42.4	H H
212.3	-/-	ID42.4	ID42.4	*
212.4	-/-	ID42.4	ID42.4	*
213.1	+/+	ID40.2	ID40.2	H H
213.2	+/+	ID39.4	ID42.4	1 1
213.3	+/+	ID40.2	ID42.4	
213.4	+/+	ID40.2	ID39.4	
214.1	-/-	ID42.4	ID44.1	1 1
214.2	+/+	ID40.2	ID42.4	1 1
214.3	+/+	ID44.1	ID40.2	
214.4	-/-	ID42.4	ID42.4	*
216.1	+/+	IC33.1	ID42.4	0 2
216.2	+/+	ID40.2	IIIA148.3	1 1
216.3	+/+	ID42.4	ID40.2	
216.4	+/+	ID42.4	IIIA148.3	
227.1	-/-	ID43.9	ID43.9	H H
227.2	+/+	IC32.8	ID42.4	1 1
227.3	-/-	ID43.9	ID42.4	
227.4	+/+	ID43.9	IC32.8	

DNA	Msp I	MVR			
229.1	+/+	ID40.2	IIIA146.2	1	1
229.2	+/+	ID40.2	IIIA144.1	1	1
229.3	+/+	IIIA146.2	ID40.2		
229.4	+/+	ID40.2	IIIA144.1		
232.1	+/+	ID40.2	ID40.2	H	H
232.2	-/-	ID42.4	ID43.9	1	1
232.3	+/-	ID40.2	ID42.4		
232.4	+/-	ID40.2	ID43.9		
301.1	+/+	IC31.2	IIIB142.1	2	0
301.2	+/+	IC33.1	IIIA149.5	2	0
301.3	+/+	IC31.2	IC33.1		
301.4	+/+	IC31.2	IC33.1		
302.1	+/-	IC32.3	ID42.4	1	1
302.2	+/-	ID39.4	ID44.1	1	1
302.3	-/-	ID42.4	ID44.1		
302.4	+/+	IC32.3	ID39.4		
303.1	+/-	ID42.4	IIIB143.3	0	2
303.2	+/+	IIIA149.6	IIIA151.3	1	1
303.3	+/+	IIIB143.3	IIIA151.3		
303.4	+/+	IIIB143.3	IIIA149.6		
304.1	+/+	IC31.5	IIIA147.2	0	2
304.2	+/-	ID44.1	IIIA151.1	0	2
304.3	+/+	IIIA147.2	IIIA151.1		
304.4	+/+	IIIA147.2	IIIA151.1		
305.1	+/-	IC32.7	ID42.4	1	1
305.2	+/-	ID41.3	IIIA139.1	2	0
305.3	-/-	ID42.4	ID41.3		
305.4	+/-	IC32.7	ID41.3		
306.1	+/-	IC33.4	ID42.4	1	1
306.2	+/-	ID42.4	IIIA147.5	1	1
306.3	+/-	ID42.4	IIIA147.5		
306.4	+/-	IC33.4	ID42.4		
307.1	+/-	ID42.4	IIIA158.3	2	0
307.2	+/+	IIIA147.4	IIIA147.2	2	0
307.3	+/-	ID42.4	IIIA147.4		
307.4	+/-	ID42.4	IIIA147.4		
308.1	+/-	IC32.7	ID43.9	1	1
308.2	+/-	ID40.2	ID43.9	2	0
308.3	+/-	ID43.9	ID40.2		
308.4	+/+	IC32.7	ID40.2		
309.1	+/-	ID42.4	IIIA146.1	1	1
309.2	+/+	IC32.6	IC37.1	2	0
309.3	+/-	ID42.4	IC32.6		
309.4	+/+	IIIA146.1	IC32.6		
310.1	+/-	IC36.6	ID40.1	0	2
310.2	+/-	ID40.2	ID42.4	0	2
310.3	-/-	ID40.1	ID42.4		
310.4	-/-	ID40.1	ID42.4		
311.1	+/+	IC30.1	IIIA159.2	2	0
311.2	+/+	ID39.1	IIIA144.2	1	1
311.3	+/+	IC30.1	IIIA144.2		
311.4	+/+	IC30.1	ID39.1		
313.1	+/+	IC31.2	IIIB110.1	2	0
313.2	+/-	ID44.1	IIIA150.5	2	0
313.3	+/-	IC31.2	ID44.1		
313.4	+/-	IC31.2	ID44.1		
314.1	+/-	ID41.5	ID42.4	2	0
314.2	-/-	ID42.4	ID42.4	H	H
314.3	+/-	ID41.5	ID42.4		
314.4	+/-	ID41.5	ID42.4		

DNA	Msp I	MVR			
317.1	+/-	ID43.9	IIIB144.1	2	0
317.2	+/-	ID40.2	ID42.4	0	2
317.3	-/-	ID43.9	ID42.4		
317.4	-/-	ID43.9	ID42.4		
318.1	+/-	IC34.1	IIIA149.13	0	2
318.2	+/-	IC41.1	ID42.5	1	1
318.3	-/-	IIIA149.13	ID42.5		
318.4	+/-	IIIA149.13	IC41.1		
321.1	+/-	IC41.2	ID42.4	1	1
321.2	+/-	ID43.3	IIIB145.2	1	1
321.3	-/-	ID42.4	ID43.3		
321.4	+/+	IC41.2	IIIB145.2		
327.1	+/-	IC34.5	ID41.3	1	1
327.2	+/-	ID40.2	ID42.4	2	0
327.3	+/+	IC34.5	ID40.2		
327.4	+/-	ID41.3	ID40.2		
333.1	+/-	ID41.4	IIIA148.2	1	1
333.2	+/-	ID42.4	IIIA150.3	2	0
333.3	+/-	IIIA148.2	ID42.4		
333.4	-/-	ID41.4	ID42.4		
163.1	-/-	ID42.4	ID42.4	H	H
163.2	+/-	IC31.2	ID43.2	2	0
163.3	+/-	ID42.4	IC31.2		
163.4	+/-	ID42.4	IC31.2		
203.1	+/-	ID40.2	ID44.1	1	1
203.2	+/+	IC35.4	IC37.2	1	1
203.3	+/-	ID44.1	IC35.4		
203.4	+/+	ID40.2	IC37.2		
226.1	-/-	ID43.9	ID44.1	0	2
226.2	+/+	IIIA148.10	IIIA158.4	1	1
226.3	+/-	ID44.1	IIIA158.4		
226.4	+/-	ID44.1	IIIA148.10		
230.1	+/-	IC31.5	ID44.1	1	1
230.2	+/+	IC33.3	ID40.2	1	1
230.3	+/-	ID44.1	ID40.2		
230.4	+/+	IC31.5	IC33.3		
233.1	+/-	ID39.4	ID43.6	0	2
233.2	+/+	IC32.7	ID39.1	1	1
233.3	+/-	ID43.6	ID39.1		
233.4	+/-	ID43.6	IC32.7		
234.1	+/+	IC31.5	IC32.6	1	1
234.2	-/-	ID43.6	ID44.1	1	1
234.3	+/-	IC32.6	ID43.6		
234.4	+/-	IC31.5	ID44.1		
239.1	+/+	ID40.2	IIIA149.8	1	1
239.2	+/-	ID42.4	IIIA148.4	0	2
239.3	+/+	IIIA149.8	IIIA148.4		
239.4	+/+	ID40.2	IIIA148.4		
241.1	+/-	IC38.3	ID42.4	0	2
241.2	+/+	ID39.1	ID40.2	0	2
241.3	+/-	ID42.4	ID40.2		
241.4	+/-	ID42.4	ID40.2		
320.1	+/+	IC36.2	IIIA147.4	2	0
320.2	+/-	ID44.1	IIIA150.3	2	0
320.3	+/-	IC36.2	ID44.1		
320.4	+/-	IC36.2	ID44.1		
322.1	+/+	IIIA149.2	IIIA149.10	2	0
322.2	+/-	ID40.2	ID44.1	0	2
322.3	+/-	IIIA149.2	ID44.1		
322.4	+/-	IIIA149.2	ID44.1		



DNA	Msp I	MVR		
323.1	+/-	ID42.4	IIIA148.4	0 2
323.2	-/-	ID42.4	ID43.3	1 1
323.3	+/-	IIIA148.4	ID43.3	
323.4	+/-	IIIA148.4	ID42.4	
324.1	+/+	IC32.6	IIIA158.2	0 2
324.2	+/+	ID39.1	IIIA148.1	0 2
324.3	+/+	IIIA158.2	IIIA148.1	
324.4	+/+	IIIA158.2	IIIA148.1	
326.1	+/+	IC30.4	ID40.2	1 1
326.2	+/+	IC37.4	ID40.2	1 1
326.3	+/+	ID40.2	IC37.4	
326.4	+/+	IC30.4	ID40.2	
332.1	+/+	IC36.1	IIIA139.1	2 0
332.2	+/+	IC31.2	IC32.1	1 1
332.3	+/+	IC36.1	IC32.1	
332.4	+/+	IC36.1	IC31.2	
337.1	-/-	ID42.4	ID42.4	H H
337.2	+/-	IC36.7	ID42.4	0 2
337.3	-/-	ID42.4	ID42.4	*
337.4	-/-	ID42.4	ID42.4	*
340.1	+/-	IC31.2	ID40.3	1 1
340.2	+/-	ID44.1	IIIA149.9	2 0
340.3	+/-	IC31.2	ID44.1	
340.4	-/-	ID40.3	ID44.1	
341.1	+/+	IC32.1	ID40.2	1 1
341.2	-/-	ID42.4	ID42.4	H H
341.3	+/-	ID40.2	ID42.4	
341.4	+/-	IC32.1	ID42.4	
155.1	+/-	ID42.4	IIIA150.3	0 2
155.2	+/-	ID40.2	ID42.4	0 2
155.3	+/-	IIIA150.3	ID42.4	
155.4	+/-	IIIA150.3	ID42.4	
161.1	+/-	IC37.3	ID43.3	0 2
161.2	+/-	ID42.4	IIIA147.1	1 1
161.3	+/-	ID43.3	IIIA147.1	
161.4	-/-	ID43.3	ID42.4	
195.1	+/-	ID42.4	IIIA152.2	0 2
195.2	+/+	ID43.9	IIIB144.1	2 0
195.3	+/+	IIIA152.2	ID43.9	
195.4	+/+	IIIA152.2	ID43.9	
208.1	+/+	IC37.1	IC38.3	1 1
208.2	+/-	ID40.2	ID42.4	1 1
208.3	+/-	IC37.1	ID42.4	
208.4	+/+	IC38.3	ID40.2	
220.1	+/-	ID43.9	R42.1	0 2
220.2	+/-	ID40.2	ID42.4	2 0
220.3	+/+	R42.1	ID40.2	
220.4	+/+	R42.1	ID40.2	
246.1	+/+	ID39.6	IIIB143.1	2 0
246.2	+/-	IC34.1	ID44.1	1 1
246.3	+/-	ID39.6	ID44.1	
246.4	+/+	ID39.6	IC34.1	
250.1	+/+	ID41.5	IIIA149.7	0 2
250.2	+/-	IC31.5	ID42.4	0 2
250.3	+/-	IIIA149.7	ID42.4	
250.4	+/-	IIIA149.7	ID42.4	
252.1	+/-	ID40.2	ID42.4	0 2
252.2	+/-	IC30.3	ID42.4	1 1
252.3	-/-	ID42.4	ID42.4	*
252.4	+/-	ID42.4	IC30.3	

DNA	Msp I	MVR		
325.1	-/-	ID42.4	ID44.1	0 2
325.2	+/+	IC31.2	IIIB143.1	2 0
325.3	+/-	ID44.1	IC31.2	
325.4	+/-	ID44.1	IC31.2	
335.1	+/-	IC37.1	ID42.4	1 1
335.2	-/-	ID42.4	ID42.4	H H
335.3	-/-	ID42.4	ID42.4	*
335.4	+/-	IC37.1	ID42.4	
343.1	+/-	ID42.4	IIIB144.1	1 1
343.2	+/+	ID40.3	IIIA150.2	2 0
343.3	+/-	ID42.4	ID40.3	
343.4	+/+	IIIB144.1	ID40.3	
344.1	+/+	IIIB144.1	IIIA147.2	1 1
344.2	+/-	ID44.1	IIIB144.1	1 1
344.3	+/+	IIIB144.1	IIIB144.1	*
344.4	+/-	IIIA147.2	ID44.1	
345.1	+/-	ID39.6	ID42.4	1 1
345.2	+/-	IC30.2	ID43.1	1 1
345.3	+/-	ID42.4	IC30.2	
345.4	+/-	ID39.6	ID43.1	
346.1	+/-	ID42.4	IIIA138.2	0 2
346.2	+/-	ID44.1	IIIA150.3	2 0
346.3	+/-	IIIA138.2	ID44.1	
346.4	+/-	IIIA138.2	ID44.1	
347.1	+/-	IC35.2	ID42.4	1 1
347.2	+/-	ID37.1	ID44.1	0 2
347.3	-/-	ID42.4	ID44.1	
347.4	+/-	IC35.2	ID44.1	
351.1	+/-	ID39.4	ID43.9	2 0
351.2	+/+	IC35.4	IC38.3	1 1
351.3	+/+	ID39.4	IC35.4	
351.4	+/+	ID39.4	IC38.3	
188.1	-/-	ID42.4	ID42.4	H H
188.2	+/-	ID38.3	ID42.4	1 1
188.3	-/-	ID42.4	ID42.4	*
188.4	+/-	ID42.4	ID38.3	
235.1	+/-	ID43.9	IIIA147.4	0 2
235.2	+/+	IC32.7	ID37.1	1 1
235.3	+/+	IIIA147.4	IC32.7	
235.4	+/+	IIIA147.4	ID37.1	
256.1	+/-	ID40.2	ID42.4	1 1
256.2	-/-	ID42.4	ID43.4	1 1
256.3	-/-	ID42.4	ID43.4	
256.4	+/-	ID40.2	ID42.4	
258.1	+/-	ID39.1	ID42.4	0 2
258.2	+/+	ID40.2	IIIA148.13	0 2
258.3	+/-	ID42.4	IIIA148.13	
258.4	+/-	ID42.4	IIIA148.13	
334.1	+/+	IC34.4	IIIB143.1	2 0
334.2	+/+	IC36.2	IC43.1	0 2
334.3	+/+	IC34.4	IC43.1	
334.4	+/+	IC34.4	IC43.1	
348.1	+/-	IC35.4	ID42.4	1 1
348.2	+/+	IC35.6	IIIA143.1	2 0
348.3	+/+	IC35.4	IC35.6	
348.4	+/-	ID42.4	IC35.6	
355.1	+/+	IC37.1	IIIA159.2	2 0
355.2	+/+	IC30.3	ID40.2	0 2
355.3	+/+	IC37.1	ID40.2	
355.4	+/+	IC37.1	ID40.2	

DNA	Msp I	MVR		
356.1	+/+	IC34.1	ID40.2	2 0
356.2	+/-	ID42.4	IIIA150.5	2 0
356.3	+/-	IC34.1	ID42.4	
356.4	+/-	IC34.1	ID42.4	
357.1	+/+	ID43.9	IIIB110.1	1 1
357.2	+/+	IC30.2	IC32.2	1 1
357.3	+/+	IIIB110.1	IC30.2	
357.4	+/+	ID43.9	IC32.2	
358.1	+/+	ID39.4	ID40.2	0 2
358.2	+/+	ID38.3	ID40.2	1 1
358.3	+/+	ID40.2	ID38.3	
358.4	+/+	ID40.2	ID40.2	*
361.1	-/-	ID42.4	ID44.1	0 2
361.2	+/+	IC30.6	IIIA157.1	2 0
361.3	+/-	ID44.1	IC30.6	
361.4	+/-	ID44.1	IC30.6	
362.1	+/+	ID40.2	IIIA152.2	2 0
362.2	+/-	IC38.3	ID42.4	0 2
362.3	+/-	ID40.2	ID42.4	
362.4	+/-	ID40.2	ID42.4	
366.1	-/-	ID42.4	ID44.1	1 1
366.2	+/+	ID40.2	IIIA149.1	1 1
366.3	+/-	ID44.1	ID40.2	
366.4	+/-	ID42.4	IIIA149.1	
368.1	+/+	ID40.2	IIIB143.4	2 0
368.2	+/+	IC32.1	ID40.2	2 0
368.3	+/+	ID40.2	IC32.1	
368.4	+/+	ID40.2	IC32.1	
247.1	+/+	R31.1	IIIA151.2	0 2
247.2	+/-	IC36.4	ID43.9	2 0
247.3	+/+	IIIA151.2	IC36.4	
247.4	+/+	IIIA151.2	IC36.4	
254.1	+/+	ID39.4	ID41.5	0 2
254.2	+/+	IC38.3	IIIB128.1	0 2
254.3	+/+	ID41.5	IIIB128.1	
254.4	+/+	ID41.5	IIIB128.1	
255.1	+/+	IC31.2	IC33.3	0 2
255.2	+/-	ID40.2	ID42.4	1 1
255.3	+/+	IC33.3	ID40.2	
255.4	+/-	IC33.3	ID42.4	
259.1	+/+	IC31.2	ID40.2	1 1
259.2	-/-	ID42.4	ID44.1	2 0
259.3	+/-	ID40.2	ID42.4	
259.4	+/-	IC31.2	ID42.4	
260.1	+/-	ID41.2	IIIA148.7	1 1
260.2	+/+	IC38.1	IC38.2	1 1
260.3	+/-	ID41.2	IC38.1	
260.4	+/+	IIIA148.7	IC38.2	
261.1	+/+	IC31.5	IIIA149.12	0 2
261.2	+/-	ID40.2	ID42.4	0 2
261.3	+/-	IIIA149.12	ID42.4	
261.4	+/-	IIIA149.12	ID42.4	
263.1	+/+	IC33.5	IIIB143.1	1 1
263.2	+/+	ID41.5	IIIA152.2	2 0
263.3	+/+	IIIB143.1	ID41.5	
263.4	+/+	IC33.5	ID41.5	
269.1	+/+	ID40.2	ID40.2	H H
269.2	+/-	ID44.1	IIIA149.2	1 1
269.3	+/+	ID40.2	IIIA149.2	
269.4	+/-	ID40.2	ID44.1	

DNA	Msp I	MVR		
369.1	+/-	ID39.5	ID40.2	2 0
369.2	-/-	ID42.4	ID42.4	H H
369.3	-/-	ID39.5	ID42.4	
369.4	-/-	ID39.5	ID42.4	
370.1	+/-	IC32.5	ID40.1	1 1
370.2	+/+	IC31.2	IIIA148.13	2 0
370.3	+/+	IC32.5	IC31.2	
370.4	+/-	ID40.1	IC31.2	
371.1	+/+	IC32.6	IC38.2	2 0
371.2	+/+	IC34.1	ID40.2	2 0
371.3	+/+	IC32.6	IC34.1	
371.4	+/+	IC32.6	IC34.1	
373.1	+/-	ID42.4	IIIB143.4	0 2
373.2	+/+	IC30.6	IIIB143.4	2 0
373.3	+/+	IIIB143.4	IC30.6	
373.4	+/+	IIIB143.4	IC30.6	
382.1	-/-	ID43.9	ID43.5	1 1
382.2	+/-	IC38.3	ID43.9	2 0
382.3	+/-	ID43.9	IC38.3	
382.4	+/-	ID43.5	IC38.3	
383.1	+/+	IIIA150.1	IIIA150.1	H H
383.2	-/-	ID41.2	ID42.4	2 0
383.3	+/-	IIIA150.1	ID41.2	
383.4	+/-	IIIA150.1	ID41.2	
118.1	+/+	IC31.5	IIIB110.1	1 1
118.2	+/+	ID40.2	IIIB144.1	0 2
118.3	+/+	IIIB110.1	IIIB144.1	
118.4	+/+	IC31.5	IIIB144.1	
257.1	+/-	ID40.2	ID42.4	1 1
257.2	+/+	ID40.2	ID40.2	H H
257.3	+/+	ID40.2	ID40.2	*
257.4	+/-	ID42.4	ID40.2	
272.1	+/+	IC32.1	IIIB141.1	2 0
272.2	+/-	ID40.2	ID42.4	2 0
272.3	+/+	IC32.1	ID40.2	
272.4	+/+	IC32.1	ID40.2	
276.1	+/+	ID39.4	IIIA150.8	2 0
276.2	+/+	ID40.2	IIIA146.2	1 1
276.3	+/+	ID39.4	ID40.2	
276.4	+/+	ID39.4	IIIA146.2	
277.1	+/-	IC32.7	ID42.4	1 1
277.2	-/-	ID42.4	ID44.1	1 1
277.3	+/-	IC32.7	ID42.4	
277.4	-/-	ID42.4	ID44.1	
279.1	+/+	ID40.2	IIIA146.2	2 0
279.2	+/+	IC31.5	IIIA146.3	2 0
279.3	+/+	ID40.2	IC31.5	
279.4	+/+	ID40.2	IC31.5	
283.1	+/-	ID42.4	IIIB144.1	0 2
283.2	+/-	ID42.3	ID43.9	1 1
283.3	+/-	IIIB144.1	ID43.9	
283.4	+/-	IIIB144.1	ID42.3	
287.1	+/-	ID40.2	ID44.1	1 1
287.2	+/+	ID43.9	IIIA150.3	2 0
287.3	+/-	ID44.1	ID43.9	
287.4	+/+	ID40.2	ID43.9	
288.1	+/-	IC30.2	ID44.1	1 1
288.2	+/+	IC31.5	IIIA149.4	0 2
288.3	+/-	ID44.1	IIIA149.4	
288.4	+/+	IC30.2	IIIA149.4	



DNA	Msp I	MVR		
391.1	+/+	R41.1	IIIA146.5	1 1
391.2	-/-	ID42.4	ID42.4	H H
391.3	+/-	IIIA146.5	ID42.4	
391.4	+/-	R41.1	ID42.4	
392.1	+/-	IC31.5	ID43.3	0 2
392.2	+/-	ID40.2	ID42.4	1 1
392.3	+/-	ID43.3	ID40.2	
392.4	-/-	ID43.3	ID42.4	
266.1	+/-	ID42.4	ID43.9	1 1
266.2	+/-	ID40.2	ID44.1	1 1
266.3	+/+	ID43.9	ID40.2	
266.4	-/-	ID42.4	ID44.1	
275.1	+/+	ID40.2	IIIA149.13	1 1
275.2	+/+	IC34.1	ID40.2	1 1
275.3	+/+	ID40.2	IC34.1	
275.4	+/+	IIIA149.13	ID40.2	
291.1	+/-	IC38.3	ID42.4	1 1
291.2	+/-	ID42.4	IIIA148.14	2 0
291.3	-/-	ID42.4	ID42.4	*
291.4	+/-	IC38.3	ID42.4	
292.1	+/+	IC33.1	IIIB144.1	2 0
292.2	-/-	ID44.1	ID44.1	H H
292.3	+/-	IC33.1	ID44.1	
292.4	+/-	IC33.1	ID44.1	
293.1	-/-	ID42.4	ID42.4	H H
293.2	+/+	IC31.5	ID40.2	1 1
293.3	+/-	ID42.4	IC31.5	
293.4	+/-	ID42.4	ID40.2	
294.1	+/-	ID40.2	ID43.5	2 0
294.2	+/-	ID42.4	IIIA149.11	1 1
294.3	+/+	ID40.2	IIIA149.11	
294.4	+/-	ID40.2	ID42.4	
295.1	+/+	ID39.4	IIIA148.12	1 1
295.2	+/+	ID40.2	IIIA146.2	2 0
295.3	+/+	ID39.4	ID40.2	
295.4	+/+	IIIA148.12	ID40.2	
297.1	+/+	IC36.8	IIIA147.6	1 1
297.2	+/+	ID41.5	IIIB145.3	2 0
297.3	+/+	IC36.8	ID41.5	
297.4	+/+	IIIA147.6	ID41.5	
299.1	-/-	ID42.4	ID44.1	0 2
299.2	+/+	IC36.5	IIIA156.1	2 0
299.3	+/-	ID44.1	IC36.5	
299.4	+/-	ID44.1	IC36.5	
385.1	+/-	ID44.1	IIIB143.4	0 2
385.2	+/-	ID40.2	ID42.4	1 1
385.3	+/+	IIIB143.4	ID40.2	
385.4	+/-	IIIB143.4	ID42.4	
402.1	+/-	ID42.4	IIIB144.1	2 0
402.2	+/-	ID44.1	IIIB143.2	2 0
402.3	-/-	ID42.4	ID44.1	
402.4	-/-	ID42.4	ID44.1	
404.1	+/-	IC32.1	ID42.4	0 2
404.2	+/+	ID40.2	IIIB110.1	2 0
404.3	+/-	ID42.4	ID40.2	
404.4	+/-	ID42.4	ID40.2	
408.1	+/-	ID42.4	IIIA149.11	1 1
408.2	+/-	IC38.2	ID42.4	1 1
408.3	+/-	IIIA149.11	ID42.4	
408.4	+/-	ID42.4	IC38.2	

DNA	Msp I	MVR		
410.1	+/-	ID40.2	ID43.8	0 2
410.2	-/-	ID43.2	ID44.1	0 2
410.3	-/-	ID43.8	ID44.1	
410.4	-/-	ID43.8	ID44.1	
411.1	+/-	ID42.2	IIIA150.7	1 1
411.2	+/-	ID44.1	IIIA145.1	2 0
411.3	-/-	ID42.2	ID44.1	
411.4	+/-	IIIA150.7	ID44.1	
412.1	+/-	ID38.3	ID42.4	0 2
412.2	-/-	ID42.4	ID43.9	2 0
412.3	-/-	ID42.4	ID42.4	*
412.4	-/-	ID42.4	ID42.4	*
413.1	+/-	ID39.4	ID44.1	1 1
413.2	+/-	ID39.4	ID42.4	0 2
413.3	+/-	ID39.4	ID42.4	
413.4	-/-	ID44.1	ID42.4	
416.1	+/-	IC32.4	ID43.9	1 1
416.2	+/+	IC32.6	IIIA158.1	1 1
416.3	+/-	ID43.9	IIIA158.1	
416.4	+/+	IC32.4	IC32.6	
40.1	-/-	ID44.1	ID44.1	H H
40.2	+/-	ID40.2	ID44.1	1 1
40.3	+/-	ID44.1	ID40.2	
40.4	-/-	ID44.1	ID44.1	*
63.1	+/-	IC32.4	ID44.1	0 2
63.2	-/-	ID42.4	ID43.5	2 0
63.3	-/-	ID44.1	ID42.4	
63.4	-/-	ID44.1	ID42.4	
90.1	+/+	ID39.1	IIIA150.3	1 1
90.2	+/-	ID38.3	ID42.4	1 1
90.3	+/-	ID39.1	ID42.4	
90.4	+/+	IIIA150.3	ID38.3	
117.1	+/-	IC37.3	ID42.4	1 1
117.2	-/-	ID42.4	ID42.4	H H
117.3	-/-	ID42.4	ID42.4	*
117.4	+/-	IC37.3	ID42.4	
267.1	+/+	IC32.6	ID38.2	1 1
267.2	+/-	ID42.4	IIIB144.1	2 0
267.3	+/-	IC32.6	ID42.4	
267.4	+/-	ID38.2	ID42.4	
284.1	+/+	IC31.2	IC38.3	1 1
284.2	+/-	ID42.4	IIIA150.3	1 1
284.3	+/+	IC31.2	IIIA150.3	
284.4	+/-	IC38.3	ID42.4	
285.1	+/-	ID42.4	IIIA148.12	2 0
285.2	+/-	ID44.1	IIIA159.2	1 1
285.3	+/-	ID42.4	IIIA159.2	
285.4	-/-	ID42.4	ID44.1	
300.1	+/+	ID41.5	IIIB145.3	0 2
300.2	+/-	ID42.4	IIIA145.3	1 1
300.3	+/+	IIIB145.3	IIIA145.3	
300.4	+/-	IIIB145.3	ID42.4	
372.1	+/-	ID42.2	R44.1	2 0
372.2	+/-	ID42.4	IIIA149.14	2 0
372.3	-/-	ID42.2	ID42.4	
372.4	-/-	ID42.2	ID42.4	
376.1	-/-	ID42.4	ID44.1	1 1
376.2	+/+	IC34.1	ID40.2	0 2
376.3	+/-	ID44.1	ID40.2	
376.4	+/-	ID42.4	ID40.2	

DNA	Msp I	MVR		
421.1	+/+	IC31.6	IC32.6	0 2
421.2	+/-	ID44.1	IIIA150.3	2 0
421.3	+/-	IC32.6	ID44.1	
421.4	+/-	IC32.6	ID44.1	
425.1	-/-	ID39.3	ID42.4	0 2
425.2	-/-	ID42.4	ID42.2	2 0
425.3	-/-	ID42.4	ID42.4	*
425.4	-/-	ID42.4	ID42.4	*
431.1	+/+	IC30.2	ID40.2	1 1
431.2	+/-	IC30.3	ID42.4	1 1
431.3	+/-	IC30.2	ID42.4	
431.4	+/+	ID40.2	IC30.3	
436.1	+/+	ID39.6	IIIA148.8	1 1
436.2	+/-	IC37.1	ID42.4	1 1
436.3	+/-	IIIA148.8	ID42.4	
436.4	+/+	ID39.6	IC37.1	
124.1	+/+	IC33.1	IIIA143.2	1 1
124.2	+/-	ID42.4	IIIA148.11	1 1
124.3	+/+	IC33.1	IIIA148.11	
124.4	+/-	IIIA143.2	ID42.4	
147.1	+/+	IC36.4	ID39.4	0 2
147.2	+/-	IC33.1	ID42.4	0 2
147.3	+/-	ID39.4	ID42.4	
147.4	+/-	ID39.4	ID42.4	
150.1	+/-	ID38.3	ID43.5	2 0
150.2	+/-	IC38.1	ID42.4	1 1
150.3	+/+	ID38.3	IC38.1	
150.4	+/-	ID38.3	ID42.4	
405.1	+/+	ID40.2	ID40.2	H H
405.2	+/+	IC31.4	ID39.4	1 1
405.3	+/+	ID40.2	ID39.4	
405.4	+/+	ID40.2	IC31.4	
406.1	+/-	IC35.1	ID43.5	1 1
406.2	-/-	ID42.4	ID43.9	2 0
406.3	+/-	IC35.1	ID42.4	
406.4	-/-	ID43.5	ID42.4	
415.1	-/-	ID43.9	R188.1	0 2
415.2	-/-	ID42.4	ID43.9	1 1
415.3	-/-	R188.1	ID42.4	
415.4	-/-	R188.1	ID43.9	
418.1	+/-	IC37.3	ID43.5	1 1
418.2	+/-	ID42.4	IIIA157.2	2 0
418.3	-/-	ID43.5	ID42.4	
418.4	+/-	IC37.3	ID42.4	
428.1	+/+	IC31.2	IC38.3	0 2
428.2	+/+	IC32.1	IC33.4	1 1
428.3	+/+	IC38.3	IC33.4	
428.4	+/+	IC38.3	IC32.1	
439.1	+/-	IC31.5	ID42.4	0 2
439.2	+/+	ID40.2	IIIA149.4	2 0
439.3	+/-	ID42.4	ID40.2	
439.4	+/-	ID42.4	ID40.2	
446.1	+/-	ID37.1	ID42.4	2 0
446.2	+/-	ID44.1	IIIB144.3	2 0
446.3	+/-	ID37.1	ID44.1	
446.4	+/-	ID37.1	ID44.1	
192.1	+/+	ID40.2	IIIA148.1	1 1
192.2	+/+	ID41.5	IIIA148.6	1 1
192.3	+/+	ID40.2	IIIA148.6	
192.4	+/+	IIIA148.1	ID41.5	

DNA	Msp I	MVR		
414.1	+/+	ID39.6	ID41.5	0 2
414.2	+/+	ID40.2	ID40.2	H H
414.3	+/+	ID41.5	ID40.2	
414.4	+/+	ID41.5	ID40.2	
420.1	+/-	IC32.7	ID42.4	1 1
420.2	+/-	IC30.2	ID42.4	2 0
420.3	+/+	IC32.7	IC30.2	
420.4	+/-	ID42.4	IC30.2	
429.1	+/+	IC35.4	IIIA150.5	2 0
429.2	+/+	ID38.2	IIIA149.2	1 1
429.3	+/+	IC35.4	ID38.2	
429.4	+/+	IC35.4	IIIA149.2	
430.1	+/+	ID39.4	IIIA159.2	2 0
430.2	+/-	IC35.5	ID42.4	0 2
430.3	+/-	ID39.4	ID42.4	
430.4	+/-	ID39.4	ID42.4	
451.1	+/-	ID42.4	IIIA139.1	1 1
451.2	+/-	IC33.4	ID42.2	1 1
451.3	+/-	ID42.4	IC33.4	
451.4	+/-	IIIA139.1	ID42.2	
455.1	+/+	IC31.5	ID40.2	1 1
455.2	+/+	IC38.3	ID40.2	1 1
455.3	+/+	ID40.2	ID40.2	*
455.4	+/+	IC31.5	IC38.3	
456.1	-/-	ID42.4	ID44.1	2 0
456.2	+/-	ID39.6	ID42.4	2 0
456.3	+/-	ID42.4	ID39.6	
456.4	+/-	ID42.4	ID39.6	
5.1	-/-	ID42.4	ID42.4	H H
5.2	+/+	ID40.2	ID40.2	H H
5.3	+/-	ID42.4	ID40.2	
5.4	+/-	ID42.4	ID40.2	
186.1	+/-	ID39.4	ID43.7	1 1
186.2	-/-	ID42.4	ID44.1	0 2
186.3	-/-	ID43.7	ID44.1	
186.4	+/-	ID39.4	ID44.1	
204.1	-/-	ID42.4	ID43.9	1 1
204.2	+/-	IC33.2	ID42.4	1 1
204.3	+/-	ID42.4	IC33.2	
204.4	-/-	ID43.9	ID42.4	
379.1	+/+	ID40.2	ID40.2	H H
379.2	+/-	ID39.6	ID43.9	1 1
379.3	+/-	ID40.2	ID43.9	
379.4	+/+	ID40.2	ID39.6	
464.1	+/-	IC33.4	ID44.1	0 2
464.2	+/-	ID40.2	ID43.9	1 1
464.3	+/-	ID44.1	ID40.2	
464.4	-/-	ID44.1	ID43.9	
467.1	+/+	IC34.1	ID40.2	0 2
467.2	+/-	IC37.3	ID42.4	1 1
467.3	+/+	ID40.2	IC37.3	
467.4	+/-	ID40.2	ID42.4	
468.1	+/+	ID40.2	IIIA148.12	1 1
468.2	+/-	IC31.2	ID42.4	1 1
468.3	+/-	ID40.2	ID42.4	
468.4	+/+	IIIA148.12	IC31.2	
503.1	+/-	IC30.7	ID44.1	1 1
503.2	-/-	ID42.1	ID44.1	2 0
503.3	+/-	IC30.7	ID42.1	
503.4	-/-	ID44.1	ID42.1	

DNA	Msp I	MVR			
101.1	+/+	ID40.2	ID40.2	H	H
101.2	+/-	IC37.1	ID42.4	1	1
101.3	+/+	ID40.2	IC37.1		
101.4	+/-	ID40.2	ID42.4		
43.1	-/-	ID42.4	ID43.9	1	1
43.2	+/-	ID39.4	ID44.1	1	1
43.3	-/-	ID43.9	ID44.1		
43.4	+/-	ID42.4	ID39.4		
57.1	-/-	ID44.1	ID44.1	H	H
57.2	-/-	ID42.4	ID42.4	H	H
57.3	-/-	ID44.1	ID42.4		
57.4	-/-	ID44.1	ID42.4		
171.1	+/-	ID42.4	IIIA148.3	1	1
171.2	-/-	ID42.4	ID44.1	0	2
171.3	+/-	IIIA148.3	ID44.1		
171.4	-/-	ID42.4	ID44.1		
121.1	+/-	ID40.2	ID42.2	0	2
121.2	+/-	ID39.4	ID42.4	2	0
121.3	+/-	ID42.2	ID39.4		
121.4	+/-	ID42.2	ID39.4		

DNA	Msp I	MVR			
136.1	+/+	IC31.3	IIIA147.3	2	0
136.2	+/-	IC33.1	ID42.4	0	2
136.3	+/-	IC31.3	ID42.4		
136.4	+/-	IC31.3	ID42.4		
113.1	+/+	IC35.3	IIIA152.3	0	2
113.2	+/-	ID42.4	IIIA148.12	0	2
113.3	+/+	IIIA152.3	IIIA148.12		
113.4	+/+	IIIA152.3	IIIA148.12		
138.1	+/+	IC30.2	IC32.6	1	1
138.2	+/-	ID40.2	ID42.4	2	0
138.3	+/+	IC30.2	ID40.2		
138.4	+/+	IC32.6	ID40.2		
82.1	+/+	ID40.2	ID40.2	H	H
82.2	+/-	ID40.2	ID43.9	0	2
82.3	+/-	ID40.2	ID43.9		
82.4	+/-	ID40.2	ID43.9		

## References

- Ahmad, W., Faiyaz ul Haque, M., Brancolini, V., Tsou, H. C., ul Haque, S., Lam, H., Aita, V. M., Owen, J., deBlaquiere, M., Frank, J., Cserhalmi-Friedman, P. B., Leask, A., McGrath, J. A., Peacocke, M., Ahmad, M., Ott, J., and Christiano, A. M. (1998a). Alopecia universalis associated with a mutation in the human hairless gene. *Science* 279, 720-724.
- Ahmad, W., Panteleyev, A. A., Sundberg, J. P., and Christiano, A. M. (1998b). Molecular basis for the rhino (hrrh-8J) phenotype: a nonsense mutation in the mouse hairless gene. *Genomics* 53, 383-386.
- Ahmad, W., Zlotogorski, A., Panteleyev, A. A., Lam, H., Ahmad, M., ul Haque, M. F., Abdallah, H. M., Dragan, L., and Christiano, A. M. (1999). Genomic organization of the human hairless gene (*HR*) and identification of a mutation underlying congenital atrichia in an Arab Palestinian family. *Genomics* 56, 141-148.
- Aitman, T. J., Hearne, C. M., McAleer, M. A., and Todd, J. A. (1991). Mononucleotide repeats are an abundant source of length variants in mouse genomic DNA. *Mamm. Genome* 1, 206-210.
- Aker, M., and Huang, H. V. (1996). Extreme heterogeneity of minor satellite repeat arrays in inbred strains of mice. *Mamm. Genome* 7, 62-64.
- Akerblom, H. K., and Knip, M. (1998). Putative environmental factors in type 1 diabetes. *Diabetes Metab. Rev.* 14, 31-67.
- Alberman, E. (1991). Are our babies becoming bigger? *J. R. Soc. Med.* 84, 257-260.
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1989). *Molecular biology of the cell*, Second Edition (New York: Garland Publishing).
- Allawi, H. T., and SantaLucia, J., Jr. (1997). Thermodynamics and NMR of internal G. T mismatches in DNA. *Biochemistry* 36, 10581-10594.
- Allen, M. J., Jeffreys, A. J., Surani, M. A., Barton, S., Norris, M. L., and Collick, A. (1994). Tandemly repeated transgenes of the human minisatellite MS32 (*DIS8*), with novel mouse gamma satellite integration. *Nucleic Acids Res.* 22, 2976-2981.
- Allen, N. D., Barton, S. C., Surani, M. A. H., and Reik, W. (1987). *Mammalian development: a practical approach* (Oxford: IRL Press).
- Amarger, V., Gauguier, D., Yerle, M., Apiou, F., Pinton, P., Giraudeau, F., Monfouilloux, S., Lathrop, M., Dutrillaux, B., Buard, J., and Vergnaud, G. (1998). Analysis of distribution in the human, pig, and rat genomes points toward a general subtelomeric origin of minisatellite structures. *Genomics* 52, 62-71.
- Andreassen, R., and Olaisen, B. (1998). *De novo* mutations and allelic diversity at minisatellite locus *D7S22* investigated by allele-specific four-state MVR-PCR analysis. *Hum. Mol. Genet.* 7, 2113-2120.
- Armour, J. A., Anttinen, T., May, C. A., Vega, E. E., Sajantila, A., Kidd, J. R., Kidd, K. K., Bertranpetit, J., Paabo, S., and Jeffreys, A. J. (1996a). Minisatellite diversity supports a recent African origin for modern humans. *Nat. Genet.* 13, 154-160.
- Armour, J. A., Crosier, M., and Jeffreys, A. J. (1996b). Distribution of tandem repeat polymorphism within minisatellite MS621 (*D5S110*). *Ann. Hum. Genet.* 60, 11-20.
- Armour, J. A., Harris, P. C., and Jeffreys, A. J. (1993). Allelic diversity at minisatellite MS205 (*D16S309*): evidence for polarized variability. *Hum. Mol. Genet.* 2, 1137-1145.

- Armour, J. A., Neumann, R., Gobert, S., and Jeffreys, A. J. (1994). Isolation of human simple repeat loci by hybridization selection. *Hum. Mol. Genet.* 3, 599-605.
- Armour, J. A., Patel, I., Thein, S. L., Fey, M. F., and Jeffreys, A. J. (1989a). Analysis of somatic mutations at human minisatellite loci in tumors and cell lines. *Genomics* 4, 328-334.
- Armour, J. A., Povey, S., Jeremiah, S., and Jeffreys, A. J. (1990). Systematic cloning of human minisatellites from ordered array charomid libraries. *Genomics* 8, 501-512.
- Armour, J. A., Wong, Z., Wilson, V., Royle, N. J., and Jeffreys, A. J. (1989b). Sequences flanking the repeat arrays of human minisatellites: association with tandem and dispersed repeat elements. *Nucleic Acids Res.* 17, 4925-4935.
- Asghari, V., Schoots, O., van Kats, S., Ohara, K., Jovanovic, V., Guan, H. C., Bunzow, J. R., Petronis, A., and Van Tol, H. H. (1994). Dopamine D4 receptor repeat: analysis of different native and mutant forms of the human and rat genes. *Mol. Pharmacol.* 46, 364-373.
- Ashfield, R., Patel, A. J., Bossone, S. A., Brown, H., Campbell, R. D., Marcu, K. B., and Proudfoot, N. J. (1994). MAZ-dependent termination between closely spaced human complement genes. *Embo J.* 13, 5656-5667.
- Ashley, C. T., Jr., and Warren, S. T. (1995). Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.* 29, 703-728.
- Augstein, P., Stephens, L. A., Allison, J., Elefanty, A. G., Ekberg, M., Kay, T. W., and Harrison, L. C. (1998).  $\beta$ -cell apoptosis in an accelerated model of autoimmune diabetes. *Mol. Med.* 4, 495-501.
- Awata, T., Kurihara, S., Kikuchi, C., Takei, S., Inoue, I., Ishii, C., Takahashi, K., Negishi, K., Yoshida, Y., Hagura, R., Kanazawa, Y., and Katayama, S. (1997). Evidence for association between the class I subset of the insulin gene minisatellite (*IDDM2* locus) and IDDM in the Japanese population. *Diabetes* 46, 1637-1642.
- Bain, S. C., Prins, J. B., Hearne, C. M., Rodrigues, N. R., Rowe, B. R., Pritchard, L. E., Ritchie, R. J., Hall, J. R., Undlien, D. E., Ronningen, K. S., Dunger, D. B., Barnett, A. H., and Todd, J. A. (1992). Insulin gene region-encoded susceptibility to type 1 diabetes is not restricted to *HLA-DR4*-positive individuals. *Nat. Genet.* 2, 212-215.
- Baird, D. M., Jeffreys, A. J., and Royle, N. J. (1995). Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. *Embo J.* 14, 5433-5443.
- Baird, D. M., and Royle, N. J. (1997). Sequences from higher primates orthologous to the human Xp/Yp telomere junction region reveal gross rearrangements and high levels of divergence. *Hum. Mol. Genet.* 6, 2291-2299.
- Balazs, I., Purrello, M., Rubinstein, P., Alhadeff, B., and Siniscalco, M. (1982). Highly polymorphic DNA site *D14S1* maps to the region of Burkitt lymphoma translocation and is closely linked to the heavy chain  $\gamma$  1 immunoglobulin locus. *Proc. Natl. Acad. Sci. USA* 79, 7395-7399.
- Barlow, D. P. (1993). Methylation and imprinting: from host defense to gene regulation? *Science* 260, 309-310.
- Barnett, A. H., Eff, C., Leslie, R. D., and Pyke, D. A. (1981). Diabetes in identical twins. A study of 200 pairs. *Diabetologia* 20, 87-93.
- Bastien, L., and Bourgaux, P. (1987). The MT family of mouse DNA is made of short interspersed repeated elements. *Gene* 57, 81-88.

- Bates, G. P., Mangiarini, L., Mahal, A., and Davies, S. W. (1997). Transgenic models of Huntington's disease. *Hum. Mol. Genet.* 6, 1633-1637.
- Batzner, M. A., and Deininger, P. L. (1991). A human-specific subfamily of Alu sequences. *Genomics* 9, 481-487.
- Becker, K. G., Simon, R. M., Bailey-Wilson, J. E., Freidlin, B., Biddison, W. E., McFarland, H. F., and Trent, J. M. (1998). Clustering of non-major histocompatibility complex susceptibility candidate loci in human autoimmune diseases. *Proc. Natl. Acad. Sci. USA* 95, 9979-9984.
- Beckman, J. S., and Weber, J. L. (1992). Survey of human and rat microsatellites. *Genomics* 12, 627-631.
- Bell, G. I., Horita, S., and Karam, J. H. (1984). A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33, 176-183.
- Bell, G. I., Karam, J. H., and Rutter, W. J. (1981). Polymorphic DNA region adjacent to the 5' end of the human insulin gene. *Proc. Natl. Acad. Sci. USA* 78, 5759-5763.
- Bell, G. I., Selby, M. J., and Rutter, W. J. (1982). The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. *Nature* 295, 31-35.
- Bennett, K. L., and Hastie, N. D. (1984). Looking for relationships between the most repeated dispersed DNA sequences in the mouse: small R elements are found associated consistently with long MIF repeats. *Embo J.* 3, 467-472.
- Bennett, S. T., Lucassen, A. M., Gough, S. C., Powell, E. E., Undlien, D. E., Pritchard, L. E., Merriman, M. E., Kawaguchi, Y., Dronsfield, M. J., Pociot, F., Nerup, J., Bouzekri, N., Combon-Thomsen, A., Ronningen, K. S., Barnett, A. H., Bain, S. C., and Todd, J. A. (1995). Susceptibility to human type 1 diabetes at *IDDM2* is determined by tandem repeat variation at the insulin gene minisatellite locus. *Nat. Genet.* 9, 284-292.
- Bennett, S. T., and Todd, J. A. (1996a). Human type 1 diabetes and the insulin gene: principles of mapping polygenes. *Annu. Rev. Genet.* 30, 343-370.
- Bennett, S. T., Wilson, A. J., Cucca, F., Nerup, J., Pociot, F., McKinney, P. A., Barnett, A. H., Bain, S. C., and Todd, J. A. (1996b). *IDDM2*-VNTR-encoded susceptibility to type 1 diabetes: dominant protection and parental transmission of alleles of the insulin gene-linked minisatellite locus. *J. Autoimmun.* 9, 415-421.
- Bennett, S. T., Wilson, A. J., Esposito, L., Bouzekri, N., Undlien, D. E., Cucca, F., Nistico, L., Buzzetti, R., Bosi, E., Pociot, F., Nerup, J., Cambon-Thomsen, A., Pugliese, A., Shield, J. P., McKinney, P. A., Bain, S. C., Polychronakos, C., and Todd, J. A. The IMDIAB Group. (1997). Insulin VNTR allele-specific effect in type 1 diabetes depends on identity of untransmitted paternal allele. *Nat. Genet.* 17, 350-352.
- Berg, E. S., and Olaisen, B. (1993). Characterization of the *COL2A1* VNTR polymorphism. *Genomics* 16, 350-354.
- Bias, W. B., Reveille, J. D., Beaty, T. H., Meyers, D. A., and Arnett, F. C. (1986). Evidence that autoimmunity in man is a Mendelian dominant trait. *Am. J. Hum. Genet.* 39, 584-602.
- Bingham, P. M., Scott, M. O., Wang, S., McPhaul, M. J., Wilson, E. M., Garbern, J. Y., Merry, D. E., and Fischbeck, K. H. (1995). Stability of an expanded trinucleotide repeat in the androgen receptor gene in transgenic mice. *Nat. Genet.* 9, 191-196.
- Birnboim, H. C., and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* 7, 1513-1523.
- Bliskovskii, V. V. (1994). Human genes with variable number of tandem absolute tandem repeats in the coding region: possible role in pathogenesis. *Mol. Biol.* 28, 185-189.

- Boguski, M. S., Birkenmeier, E. H., Elshourbagy, N. A., Taylor, J. M., and Gordon, J. I. (1986). Evolution of the apolipoproteins. Structure of the rat apo-A-IV gene and its relationship to the human genes for apo-A-I, C-III, and E. *J. Biol. Chem.* 261, 6398-6407.
- Bois, P., Collick, A., Brown, J., and Jeffreys, A. J. (1997). Human minisatellite MS32 (*DIS8*) displays somatic but not germline instability in transgenic mice. *Hum. Mol. Genet.* 6, 1565-1571.
- Bois, P., and Jeffreys, A. J. (1999). Minisatellite instability and germline mutation. *Cell. Mol. Life Sci.* 55, 1636-1648.
- Bois, P., Stead, J. D., Bakshi, S., Williamson, J., Neumann, R., Moghadaszadeh, B., and Jeffreys, A. J. (1998a). Isolation and characterization of mouse minisatellites. *Genomics* 50, 317-330.
- Bois, P., Williamson, J., Brown, J., Dubrova, Y. E., and Jeffreys, A. J. (1998b). A novel unstable mouse VNTR family expanded from SINE B1 elements. *Genomics* 49, 122-128.
- Bonhomme, F., Catalan, J., Gerasimov, S., Orsini, P., and Thaler, L. (1983). Le complexe d'espece du genre *Mus* en Europe Centrale et Orientale. *Z. Saugetierkunde* 48, 78-85.
- Bonhomme, F., and Guénet, J. L. (1996). The laboratory mouse and its wild relatives. In *Genetics variants and strains of the laboratory mouse*, M. F. Lyon, S. Rastan and S. D. M. Brown, eds. (Oxford: Oxford University Press), pp. 1577-1596.
- Borts, R. H., and Haber, J. E. (1987). Meiotic recombination in yeast: alteration by multiple heterozygosities. *Science* 237, 1459-1465.
- Borts, R. H., Leung, W. Y., Kramer, W., Kramer, B., Williamson, M., Fogel, S., and Haber, J. E. (1990). Mismatch repair-induced meiotic recombination requires the pms1 gene product. *Genetics* 124, 573-584.
- Boursot, P., Din, W., Anand, R., Darviche, D., Dod, B., Vondeimling, F., Talwar, G. P., and Bonhomme, F. (1996). Origin and radiation of the house mouse: mitochondrial-DNA phylogeny. *J. Evol. Biol.* 9, 391-415.
- Bouzekri, N., Taylor, P. G., Hammer, M. F., and Jobling, M. A. (1998). Novel mutation processes in the evolution of a haploid minisatellite, MSY1: array homogenization without homogenization. *Hum. Mol. Genet.* 7, 655-659.
- Boyle, A. L., Ballard, S. G., and Ward, D. C. (1990). Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. USA* 87, 7757-7761.
- Breslauer, K. J., Frank, R., Blocker, H., and Marky, L. A. (1986). Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. USA* 83, 3746-3750.
- Brinkmann, B., Klitschar, M., Neuhuber, F., Huhne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* 62, 1408-1415.
- Brinkmann, B., Sajantila, A., Goedde, H. W., Matsumoto, H., Nishi, K., and Wiegand, P. (1996). Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci. *Eur. J. Hum. Genet.* 4, 175-182.
- Brock, G. J., Anderson, N. H., and Monckton, D. G. (1999). *Cis*-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. *Hum. Mol. Genet.* 8, 1061-1067.
- Brutlag, D., Fry, K., Nelson, T., and Hung, P. (1977). Synthesis of hybrid bacterial plasmids containing highly repeated satellite DNA. *Cell* 10, 509-519.

- Buard, J., Bourdet, A., Yardley, J., Dubrova, Y., and Jeffreys, A. J. (1998). Influences of array size and homogeneity on minisatellite mutation. *Embo J.* 17, 3495-3502.
- Buard, J., and Vergnaud, G. (1994). Complex recombination events at the hypermutable minisatellite CEB1 (*D2S90*). *Embo J.* 13, 3203-3210.
- Bui, M. M., Luo, D. F., She, J. Y., Maclaren, N. K., Muir, A., Thomson, G., and She, J. X. (1996). Paternally transmitted *IDDM2* influences diabetes susceptibility despite biallelic expression of the insulin gene in human pancreas. *J. Autoimmun.* 9, 97-103.
- Buresi, C., Desmarais, E., Vigneron, S., Lamarti, H., Smaoui, N., Cambien, F., and Roizes, G. (1996). Structural analysis of the minisatellite present at the 3' end of the human apolipoprotein B gene: new definition of the alleles and evolutionary implications. *Hum. Mol. Genet.* 5, 61-68.
- Burright, E. N., Clark, H. B., Servadio, A., Matilla, T., Feddersen, R. M., Yunis, W. S., Duvick, L. A., Zoghbi, H. Y., and Orr, H. T. (1995). *SCA1* transgenic mice: a model for neurodegeneration caused by an expanded CAG trinucleotide repeat. *Cell* 82, 937-948.
- Buzzetti, R., Quattrocchi, C. C., and Nistico, L. (1998). Dissecting the genetics of type 1 diabetes: relevance for familial clustering and differences in incidence. *Diabetes Metab. Rev.* 14, 111-128.
- Cachon-Gonzalez, M. B., Fenner, S., Coffin, J. M., Moran, C., Best, S., and Stoye, J. P. (1994). Structure and expression of the hairless gene of mice. *Proc. Natl. Acad. Sci. USA* 91, 7717-7721.
- Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizares, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., Trouillas, P., DeMichele, G., Filla, A., DeFrutos, R., Palau, F., Patel, P. I., DiDonato, S., Mandel, J. L., Coccozza, S., Koenig, M., and Pandolfo, M. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. *Science* 271, 1423-1427.
- Capon, D. J., Chen, E. Y., Levinson, A. D., Seeburg, P. H., and Goeddel, D. V. (1983). Complete nucleotide sequences of the T24 human bladder carcinoma oncogene and its normal homologue. *Nature* 302, 33-37.
- Carpenter, A. T. (1987). Gene conversion, recombination nodules, and the initiation of meiotic synapsis. *Bioessays* 6, 232-236.
- Casavant, N. C., Hardies, S. C., Funk, F. D., Comer, M. B., Edgell, M. H., and Hutchison, C. A. d. (1988). Extensive movement of LINES ONE sequences in  $\beta$ -globin loci of *Mus caroli* and *Mus domesticus*. *Mol. Cell. Biol.* 8, 4669-4674.
- Castano, L., and Eisenbarth, G. S. (1990). Type-1 diabetes: a chronic autoimmune disease of human, mouse, and rat. *Annu. Rev. Immunol.* 8, 647-679.
- Catasti, P., Chen, X., Deaven, L. L., Moyzis, R. K., Bradbury, E. M., and Gupta, G. (1997). Cytosine-rich strands of the insulin minisatellite adopt hairpins with intercalated cytosine<sup>+</sup>.cytosine pairs. *J. Mol. Biol.* 272, 369-382.
- Catasti, P., Chen, X., Moyzis, R. K., Bradbury, E. M., and Gupta, G. (1996). Structure-function correlations of the insulin-linked polymorphic region. *J. Mol. Biol.* 264, 534-545.
- Choo, K. H., Vissel, B., Nagy, A., Earle, E., and Kalitsis, P. (1991). A survey of the genomic distribution of  $\alpha$ -satellite DNA on all the human chromosomes, and derivation of a new consensus sequence. *Nucleic Acids Res.* 19, 1179-1182.
- Chung, M. Y., Ranum, L. P., Duvick, L. A., Servadio, A., Zoghbi, H. Y., and Orr, H. T. (1993). Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type 1. *Nat. Genet.* 5, 254-258.



- Church, G. M., and Gilbert, W. (1984). Genomic sequencing. *Proc. Natl. Acad. Sci. USA* *81*, 1991-1995.
- Clark, A. R., Wilson, M. E., London, N. J., James, R. F., and Docherty, K. (1995). Identification and characterization of a functional retinoic acid/thyroid hormone-response element upstream of the human insulin gene enhancer. *Biochem. J.* *309*, 863-870.
- Clark, S. J., Harrison, J., Paul, C. L., and Frommer, M. (1994). High sensitivity mapping of methylated cytosines. *Nucleic Acids Res.* *22*, 2990-2997.
- Collick, A., Drew, J., Penberth, J., Bois, P., Luckett, J., Scaerou, F., Jeffreys, A., and Reik, W. (1996). Instability of long inverted repeats within mouse transgenes. *Embo J.* *15*, 1163-1171.
- Collick, A., Norris, M. L., Allen, M. J., Bois, P., Barton, S. C., Surani, M. A., and Jeffreys, A. J. (1994). Variable germline and embryonic instability of the human minisatellite MS32 (*DIS8*) in transgenic mice. *Embo J.* *13*, 5745-5753.
- Collier, D. A., Barrett, T. G., Curtis, D., Macleod, A., Arranz, M. J., Maassen, J. A., and Bunday, S. (1996). Linkage of Wolfram syndrome to chromosome 4p16.1 and evidence for heterogeneity. *Am. J. Hum. Genet.* *59*, 855-863.
- Collins, K. (1996). Structure and function of telomerase. *Curr. Opin. Cell Biol.* *8*, 374-380.
- Concannon, P., Gogolin-Ewens, K. J., Hinds, D. A., Wapelhorst, B., Morrison, V. A., Stirling, B., Mitra, M., Farmer, J., Williams, S. R., Cox, N. J., Bell, G. I., Risch, N., and Spielman, R. S. (1998). A second-generation screen of the human genome for susceptibility to insulin-dependent diabetes mellitus. *Nat. Genet.* *19*, 292-296.
- Constancia, M., Pickard, B., Kelsey, G., and Reik, W. (1998). Imprinting mechanisms. *Genome Res.* *8*, 881-900.
- Copeland, N. G., Jenkins, N. A., Gilbert, D. J., Eppig, J. T., Maltais, L. J., Miller, J. C., Dietrich, W. F., Weaver, A., Lincoln, S. E., Steen, R. G., Stein, L. D., Nadeau, J. H., and Lander, E. S. (1993). A genetic linkage map of the mouse: current applications and future prospects. *Science* *262*, 57-66.
- Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y., and Farrall, M. (1995). Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of *IDDM2* and *IDDM4* with *IDDM1* in type 1 diabetes. *Am. J. Hum. Genet.* *57*, 920-934.
- Corpet, F. (1988). Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* *16*, 10881-90.
- Crothers, D. M., and Zimm, B. H. (1964). Theory of the melting transition of synthetic polynucleotides: evaluation of the stacking free energy. *J. Mol. Biol.* *9*, 1-9.
- Cucca, F., Lampis, R., Frau, F., Macis, D., Angius, E., Masile, P., Chessa, M., Frongia, P., Silveti, M., Cao, A., Devirgiliis, S., and Congia, M. (1995). The distribution of *DR4* haplotypes in Sardinia suggests a primary association of type 1 diabetes with *DRB1* and *DQB1* loci. *Hum. Immunol.* *43*, 301-308.
- Daniels, S. E., Bhattacharrya, S., James, A., Leaves, N. I., Young, A., Hill, M. R., Faux, J. A., Ryan, G. F., le Souef, P. N., Lathrop, G. M., Musk, A. W., and Cookson, W. O. (1996). A genome-wide search for quantitative trait loci underlying asthma. *Nature* *383*, 247-250.
- David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E., Drabkin, H., Gemmill, R., Giunti, P., Benomar, A., Wood, N., Ruberg, M., Agid, Y., Mandel, J. L., and Brice, A. (1997). Cloning of the *SCA7* gene reveals a highly unstable CAG repeat expansion. *Nat. Genet.* *17*, 65-70.

- Davies, J. L., Kawaguchi, Y., Bennett, S. T., Copeman, J. B., Cordell, H. J., Pritchard, L. E., Reed, P. W., Gough, S. C., Jenkins, S. C., Palmer, S. M., Balfour, K. M., Rowe, B. R., Farrall, M., Barnett, A. H., Bain, S. C., and Todd, J. A. (1994). A genome-wide search for human type 1 diabetes susceptibility genes. *Nature* 371, 130-136.
- DeBerardinis, R. J., and Kazazian, H. H., Jr. (1998). Full-length L1 elements have arisen recently in the same 1 kb region of the gorilla and human genomes. *J. Mol. Evol.* 47, 292-301.
- Deininger, P. L., Batzer, M. A., Hutchison, C. A., 3rd, and Edgell, M. H. (1992). Master genes in mammalian repetitive DNA amplification. *Trends Genet.* 8, 307-311.
- Delcourt, S. G., and Blake, R. D. (1991). Stacking energies in DNA. *J. Biol. Chem.* 266, 15160-15169.
- Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12, 387-395.
- Diabetes Epidemiology Research International Group. (1988). Geographic patterns of childhood insulin-dependent diabetes mellitus. *Diabetes* 37, 1113-1119.
- Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., and Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature* 380, 152-154.
- Dietrich, W., Katz, H., Lincoln, S. E., Shin, H. S., Friedman, J., Dracopoli, N. C., and Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* 131, 423-447.
- DiFiglia, M., Sapp, E., Chase, K. O., Davies, S. W., Bates, G. P., Vonsattel, J. P., and Aronin, N. (1997). Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science* 277, 1990-1993.
- Din, W., Anand, R., Boursot, P., Darviche, D., Dod, B., Jouvin-March, E., Orth, A., Talwar, G. P., Cazenave, P. A., and Bonhomme, F. (1996). Origin and radiation of the house mouse: clues from nuclear genes. *J. Evol. Biol.* 9, 519-539.
- Ding, S., Larson, G. P., Foldenauer, K., Zhang, G., and Krontiris, T. G. (1999). Distinct mutation patterns of breast cancer-associated alleles of the *HRAS1* minisatellite locus. *Hum. Mol. Genet.* 8, 515-521.
- Djian, P. (1998). Evolution of simple repeats in DNA and their relation to human disease. *Cell* 94, 155-160.
- Doktycz, M. J., Goldstein, R. F., Paner, T. M., Gallo, F. J., and Benight, A. S. (1992). Studies of DNA dumbbells. I. Melting curves of 17 DNA dumbbells with different duplex stem sequences linked by T4 endloops: evaluation of the nearest-neighbor stacking interactions in DNA. *Biopolymers* 32, 849-864.
- Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F., and Kazazian, H. H., Jr. (1991). Isolation of an active human transposable element. *Science* 254, 1805-1808.
- Donner, H., Rau, H., Walfish, P. G., Braun, J., Siegmund, T., Finke, R., Herwig, J., Usadel, K. H., and Badenhop, K. (1997). CTLA4 alanine-17 confers genetic susceptibility to Graves' disease and to type 1 diabetes mellitus. *J. Clin. Endocrinol. Metab.* 82, 143-146.
- Donoviel, D. B., and Bernstein, A. (1999). *SEL-1L* maps to human chromosome 14, near the insulin-dependent diabetes mellitus locus 11. *Genomics* 56, 232-233.
- Doria, A., Lee, J., Warram, J. H., and Krolewski, A. S. (1996). Diabetes susceptibility at *IDDM2* cannot be positively mapped to the VNTR locus of the insulin gene. *Diabetologia* 39, 594-599.
- Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. *Nature* 299, 111-117.

- Dover, G. A., and Flavell, R. B. (1984). Molecular coevolution: DNA divergence and the maintenance of function. *Cell* 38, 622-623.
- Dover, G. A., and Tautz, D. (1986). Conservation and divergence in multigene families: alternatives to selection and drift. *Phil. Trans. R. Soc. Lond. B. Biol. Sci.* 312, 275-289.
- Dubrova, Y. E., Jeffreys, A. J., and Malashenko, A. M. (1993). Mouse minisatellite mutations induced by ionizing radiation. *Nat. Genet.* 5, 92-94.
- Dubrova, Y. E., Nesterov, V. N., Krouchinsky, N. G., Ostapenko, V. A., Neumann, R., Neil, D. L., and Jeffreys, A. J. (1996). Human minisatellite mutation rate after the Chernobyl accident. *Nature* 380, 683-686.
- Dubrova, Y. E., Nesterov, V. N., Krouchinsky, N. G., Ostapenko, V. A., Vergnaud, G., Giraudeau, F., Buard, J., and Jeffreys, A. J. (1997). Further evidence for elevated human minisatellite mutation rate in Belarus eight years after the Chernobyl accident. *Mutat. Res.* 381, 267-278.
- Dubrova, Y. E., Plumb, M., Brown, J., Fennelly, J., Bois, P., Goodhead, D., and Jeffreys, A. J. (1998a). Stage specificity, dose response, and doubling dose for mouse minisatellite germline mutation induced by acute radiation. *Proc. Natl. Acad. Sci. USA* 95, 6251-6255.
- Dubrova, Y. E., Plumb, M., Brown, J., and Jeffreys, A. J. (1998b). Radiation-induced germline instability at minisatellite loci. *Int. J. Radiat. Biol.* 74, 689-696.
- Dumonteil, E., and Philippe, J. (1996). Insulin gene: organisation, expression and regulation. *Diabetes Metab.* 22, 164-173.
- Dunger, D. B., Ong, K. K., Huxtable, S. J., Sherriff, A., Woods, K. A., Ahmed, M. L., Golding, J., Pembrey, M. E., Ring, S., Bennett, S. T., and Todd, J. A. ALSPAC Study Team. Avon Longitudinal Study of Pregnancy and Childhood. (1998). Association of the *INS* VNTR with size at birth. *Nat. Genet.* 19, 98-100.
- Eaves, I. A., Bennett, S. T., Forster, P., Ferber, K. M., Ehrmann, D., Wilson, A. J., Bhattacharyya, S., Ziegler, A. G., Brinkmann, B., and Todd, J. A. (1999). Transmission ratio distortion at the *INS-IGF2* VNTR. *Nat. Genet.* 22, 324-325.
- Epstein, R. P., Novick, O., Umansky, R., Priel, B., Osher, Y., Blaine, D., Bennett, E. R., Nemanov, L., Katz, M., and Belmaker, R. H. (1996). Dopamine D4 receptor (*D4DR*) exon III polymorphism associated with the human personality trait of Novelty Seeking. *Nat. Genet.* 12, 78-80.
- Ellegren, H., Lindgren, G., Primmer, C. R., and Moller, A. P. (1997). Fitness loss and germline mutations in barn swallows breeding in Chernobyl. *Nature* 389, 593-596.
- Essers, J., Hendriks, R. W., Swagemakers, S. M., Troelstra, C., de Wit, J., Bootsma, D., Hoeijmakers, J. H., and Kanaar, R. (1997). Disruption of mouse RAD54 reduces ionizing radiation resistance and homologous recombination. *Cell* 89, 195-204.
- Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3, 87-112.
- Falconer, D. S. (1960). Introduction to quantitative genetics (Essex, England: Longman).
- Fang, G., and Cech, T. R. (1993). Characterization of a G-quartet formation reaction promoted by the  $\beta$ -subunit of the *Oxytricha* telomere-binding protein. *Biochemistry* 32, 11646-11657.
- Feinberg, A. P., and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* 132, 6-13.
- Feinberg, A. P., and Vogelstein, B. (1984). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Addendum. *Anal. Biochem.* 137, 266-267.

- Felsenfeld, G. (1993). Chromatin structure and the expression of globin-encoding genes. *Gene* 135, 119-124.
- Fennelly, J., Wright, E., and Plumb, M. (1997). Mini- and microsatellite mutations in radiation-induced acute myeloid leukaemia in the CBA/H mouse. *Leukemia* 11, 807-810.
- Festing, M. F. W. (1996). Origins and characteristics of inbred strains of mice. In *Genetics variants and strains of the laboratory mouse*, M. F. Lyon, S. Rastan and S. D. M. Brown, eds. (Oxford: Oxford University Press), pp. 1537-1576.
- Frantz, J. D., and Gilbert, W. (1995). A yeast gene product, G4p2, with a specific affinity for quadruplex nucleic acids. *J. Biol. Chem.* 270, 9413-9419.
- Fraser, F. C., and Gunn, T. (1977). Diabetes mellitus, diabetes insipidus, and optic atrophy. An autosomal recessive syndrome? *J. Med. Genet.* 14, 190-193.
- Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA* 89, 1827-1831.
- Garza, J. C., Slatkin, M., and Freimer, N. B. (1995). Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. *Mol. Biol. Evol.* 12, 594-603.
- Gelfand, D. H., and White, T. J. (1990). In: *PCR protocols; a guide to methods and applications*, M. A. Innis, D. H. Gelfand, J. J. Sninsky and T. J. White, eds. (San Diego: Academic Press).
- Gendler, S. J., Lancaster, C. A., Taylor-Papadimitriou, J., Duhig, T., Peat, N., Burchell, J., Pemberton, L., Lalani, E. N., and Wilson, D. (1990). Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. *J. Biol. Chem.* 265, 15286-15293.
- Ghazi, H., Magewu, A. N., Gonzales, F., and Jones, P. A. (1990). Changes in the allelic methylation patterns of c-H-ras-1, insulin and retinoblastoma genes in human development. *Dev. Suppl.* 115-123.
- Ghosh, S., Palmer, S. M., Rodrigues, N. R., Cordell, H. J., Hearne, C. M., Cornall, R. J., Prins, J. B., McShane, P., Lathrop, G. M., Peterson, L. B., Wicker, L. S., and Todd, J. A. (1993). Polygenic control of autoimmune diabetes in nonobese diabetic mice. *Nat. Genet.* 4, 404-409.
- Gibbon, C., Smith, T., Egger, P., Betts, P., and Phillips, D. (1997). Early infection and subsequent insulin dependent diabetes. *Arch. Dis. Child.* 77, 384-385.
- Gibbs, M., Collick, A., Kelly, R. G., and Jeffreys, A. J. (1993). A tetranucleotide repeat mouse minisatellite displaying substantial somatic instability during early preimplantation development. *Genomics* 17, 121-128.
- Gill, P., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E., and Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nat. Genet.* 6, 130-135.
- Giraldo, R., Suzuki, M., Chapman, L., and Rhodes, D. (1994). Promotion of parallel DNA quadruplexes by a yeast telomere binding protein: a circular dichroism study. *Proc. Natl. Acad. Sci. USA* 91, 7658-7662.
- Goldberg, Y. P., Kalchman, M. A., Metzler, M., Nasir, J., Zeisler, J., Graham, R., Koide, H. B., O'Kusky, J., Sharp, A. H., Ross, C. A., Jirik, F., and Hayden, M. R. (1996a). Absence of disease phenotype and intergenerational stability of the CAG repeat in transgenic mice expressing the human Huntington disease transcript. *Hum. Mol. Genet.* 5, 177-185.
- Goldberg, Y. P., Nicholson, D. W., Rasper, D. M., Kalchman, M. A., Koide, H. B., Graham, R. K., Bromm, M., Kazemi-Esfarjani, P., Thornberry, N. A., Vaillancourt, J. P., and Hayden, M. R. (1996b). Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. *Nat. Genet.* 13, 442-449.

- Gonzalzo, M. L., and Jones, P. A. (1997). Rapid quantitation of methylation differences at specific sites using methylation-sensitive single nucleotide primer extension (Ms-SNuPE). *Nucleic Acids Res.* 25, 2529-2531.
- Gotoh, O., and Tagashira, Y. (1981). Locations of frequently opening regions on natural DNAs and their relation to functional loci. *Biopolymers* 20, 1043-1058.
- Gourdon, G., Radvanyi, F., Lia, A. S., Duros, C., Blanche, M., Abitbol, M., Junien, C., and Hofmann-Radvanyi, H. (1997). Moderate intergenerational and somatic instability of a 55-CTG repeat in transgenic mice. *Nat. Genet.* 15, 190-192.
- 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.
- Grakoui, A., Bromley, S. K., Sumen, C., Davis, M. M., Shaw, A. S., Allen, P. M., and Dustin, M. L. (1999). The immunological synapse: a molecular machine controlling T cell activation. *Science* 285, 221-227.
- Grant, B., and Greenwald, I. (1997). Structure, function, and expression of SEL-1, a negative regulator of LIN-12 and GLP-1 in *C. elegans*. *Development* 124, 637-644.
- Gray, I. C., and Jeffreys, A. J. (1991). Evolutionary transience of hypervariable minisatellites in man and the primates. *Proc. R. Soc. Lond. B. Biol. Sci.* 243, 241-253.
- Green, H., and Djian, P. (1998). Amino acid repeats in proteins and the neurological diseases produced by poluglutamine. In *Genetic instabilities and hereditary neurological diseases*, R. D. Wells and S. T. Warren, eds. (New York: Academic Press), pp. 739-759.
- Grigg, G., and Clark, S. (1994). Sequencing 5-methylcytosine residues in genomic DNA. *Bioessays* 16, 431-436.
- Grimaldi, G., Skowronski, J., and Singer, M. F. (1984). Defining the beginning and end of KpnI family segments. *Embo J.* 3, 1753-1759.
- Gu, F., Hindkjaer, J., Gustavsson, I., and Bolund, L. (1996). A signal of telomeric sequences on porcine chromosome 6q21-q22 detected by primed in situ labelling. *Chromosome Res.* 4, 251-252.
- Guldborg, P., Gronbak, K., Aggerholm, A., Platz, A., Thor Straten, P., Ahrenkiel, V., Hokland, P., and Zeuthen, J. (1998). Detection of mutations in GC-rich DNA by bisulphite denaturing gradient gel electrophoresis. *Nucleic Acids Res.* 26, 1548-1549.
- Hager, J., Hansen, L., Vaisse, C., Vionnet, N., Philippi, A., Poller, W., Velho, G., Carcassi, C., Contu, L., Julier, C., Cambien, F., Passa, P., Lathrop, M., Kindsvogel, W., Demenais, F., Nishimura, E., and Forguet, P. (1995). A missense mutation in the glucagon receptor gene is associated with non-insulin-dependent diabetes mellitus. *Nat. Genet.* 9, 299-304.
- Hammond-Kosack, M. C., Dobrinski, B., Lurz, R., Docherty, K., and Kilpatrick, M. W. (1992). The human insulin gene linked polymorphic region exhibits an altered DNA structure. *Nucleic Acids Res.* 20, 231-236.
- Hampton, R. Y., Gardner, R. G., and Rine, J. (1996). Role of 26S proteasome and *HRD* genes in the degradation of 3-hydroxy-3-methylglutaryl-CoA reductase, an integral endoplasmic reticulum membrane protein. *Mol. Biol. Cell.* 7, 2029-2044.
- Harley, C. B., Futcher, A. B., and Greider, C. W. (1990). Telomeres shorten during ageing of human fibroblasts. *Nature* 345, 458-460.
- Harrington, J. J., and Lieber, M. R. (1994). The characterization of a mammalian DNA structure-specific endonuclease. *Embo J.* 13, 1235-1246.

- Harris, P. C., and Thomas, S. (1992). Length polymorphism of the subtelomeric regions of both the short (p) and long arms (q) of chromosome 16. *Cytogenet. Cell. Genet.* 60, 171-172.
- Hastie, N. D. (1996). Highly repeated DNA families in the genome of *Mus musculus*. In *Genetics variants and strains of the laboratory mouse*, M. F. Lyon, S. Rastan and S. D. M. Brown, eds. (Oxford: Oxford University Press), pp. 1425-1442.
- Hearne, C. M., Ghosh, S., and Todd, J. A. (1992). Microsatellites for linkage analysis of genetic traits. *Trends Genet.* 8, 288-294.
- Heath, S. K., Carne, S., Hoyle, C., Johnson, K. J., and Wells, D. J. (1997). Characterisation of expression of mDMAHP, a homeodomain-encoding gene at the murine DM locus. *Hum. Mol. Genet.* 6, 651-657.
- Heery, D. M., Gannon, F., and Powell, R. (1990). A simple method for subcloning DNA fragments from gel slices. *Trends Genet.* 6, 173.
- Heils, A., Teufel, A., Petri, S., Stober, G., Riederer, P., Bengel, D., and Lesch, K. P. (1996). Allelic variation of human serotonin transporter gene expression. *J. Neurochem.* 66, 2621-2624.
- Heiniger, H. J., Huebner, R. J., and Meier, H. (1976). Effect of allelic substitutions at the hairless locus on endogenous ecotropic murine leukemia virus titers and leukemogenesis. *J. Natl. Cancer Inst.* 56, 1073-1074.
- Hemmer, B., Fleckenstein, B. T., Vergelli, M., Jung, G., McFarland, H., Martin, R., and Wiesmuller, K. H. (1997). Identification of high potency microbial and self ligands for a human autoreactive class II-restricted T cell clone. *J. Exp. Med.* 185, 1651-1659.
- Herman, G. E., Nadeau, J. H., and Hardies, S. C. (1992). Dispersed repetitive elements in mouse genome analysis. *Mamm. Genome* 2, 207-214.
- Hewett, D. R., Handt, O., Hobson, L., Mangelsdorf, M., Eyre, H. J., Baker, E., Sutherland, G. R., Schuffenhauer, S., Mao, J. I., and Richards, R. I. (1998). FRA10B structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Mol. Cell* 1, 773-781.
- Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., and de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 6, 799-803.
- Hiai, H., Morrissey, P., Khirya, R., and Schwartz, R. S. (1977). Selective expression of xenotropic virus in congenic HRS/J (hairless) mice. *Nature* 270, 247-249.
- Hilkens, J., Ligtenberg, M. J., Vos, H. L., and Litvinov, S. V. (1992). Cell membrane-associated mucins and their adhesion-modulating property. *Trends Biochem. Sci.* 17, 359-363.
- Hollick, J. B., Dorweiler, J. E., and Chandler, V. L. (1997). Paramutation and related allelic interactions. *Trends Genet.* 13, 302-308.
- Holmlund, G., and Lindblom, B. (1998). Different ancestor alleles: a reason for the bimodal fragment size distribution in the minisatellite *D2S44* (YNH24). *Eur. J. Hum. Genet.* 6, 597-602.
- Horz, W., and Altenburger, W. (1981). Nucleotide sequence of mouse satellite DNA. *Nucleic Acids Res.* 9, 683-696.
- Houck, C. M., Rinehart, F. P., and Schmid, C. W. (1979). A ubiquitous family of repeated DNA sequences in the human genome. *J. Mol. Biol.* 132, 289-306.
- 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, 971-983.

- Igarashi, S., Takiyama, Y., Cancel, G., Rogaeva, E. A., Sasaki, H., Wakisaka, A., Zhou, Y. X., Takano, H., Endo, K., Sanpei, K., Oyake, M., Tanaka, H., Stevanin, G., Abbas, N., Durr, A., Rogaev, E. I., Sherrington, R., Tsuda, T., Ikeda, M., Cassa, E., Nishizawa, M., Benomar, A., Julien, J., Weissenbach, J., Wang, G. X., Agid, Y., St George-Hyslop, P. H., Brice, A., and Tsuji, S. (1996). Intergenerational instability of the CAG repeat of the gene for Machado-Joseph disease (*MJD1*) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat. *Hum. Mol. Genet.* 5, 923-932.
- Ikeda, H., Yamaguchi, M., Sugai, S., Aze, Y., Narumiya, S., and Kakizuka, A. (1996). Expanded polyglutamine in the Machado-Joseph disease protein induces cell death *in vitro* and *in vivo*. *Nat. Genet.* 13, 196-202.
- Ikegami, H., and Ogiwara, T. (1996). Genetics of insulin-dependent diabetes mellitus. *Endocr. J.* 43, 605-613.
- Imai, H., Nakagawa, H., Komatsu, K., Shiraishi, T., Fukuda, H., Sugimura, T., and Nagao, M. (1997). Minisatellite instability in severe combined immunodeficiency mouse cells. *Proc. Natl. Acad. Sci. USA* 94, 10817-10820.
- Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J. M., Weber, C., Mandel, J. L., Cancel, G., Abbas, N., Durr, A., Didierjean, O., Stevanin, G., Agid, Y., and Brice, A. (1996). Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. *Nat. Genet.* 14, 285-291.
- Inoue, H., Tanizawa, Y., Wasson, J., Behn, P., Kalidas, K., Bernal-Mizrachi, E., Mueckler, M., Marshall, H., Donis-Keller, H., Crock, P., Rogers, D., Mikuni, M., Kumashiro, H., Higashi, K., Sobue, G., Oka, Y., and Permutt, M. A. (1998). A gene encoding a transmembrane protein is mutated in patients with diabetes mellitus and optic atrophy (Wolfram syndrome). *Nat. Genet.* 20, 143-148.
- Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* 363, 558-561.
- Jacob, H. J., Pettersson, A., Wilson, D., Mao, Y., Lernmark, A., and Lander, E. S. (1992). Genetic dissection of autoimmune type 1 diabetes in the BB rat. *Nat. Genet.* 2, 56-60.
- Jacobson, D. P., Schmeling, P., and Sommer, S. S. (1993). Characterization of the patterns of polymorphism in a 'cryptic repeat' reveals a novel type of hypervariable sequence. *Am. J. Hum. Genet.* 53, 443-450.
- Janeway, C. A., Travers, P., Walport, M., and Capra, J. D. (1999). *Immunobiology: the immune system in health and disease*, Fourth Edition (New York: Elsevier Science Ltd/Garland Publishing).
- Jdo, J. W., Baldini, A., Ward, D. C., Reeders, S. T., and Wells, R. A. (1991). Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. USA* 88, 9051-9055.
- Jeffreys, A. J. (1987a). Highly variable minisatellites and DNA fingerprints. *Biochem. Soc. Trans.* 15, 309-317.
- Jeffreys, A. J., Barber, R., Bois, P., Buard, J., Dubrova, Y. E., Grant, G., Hollies, C. R., May, C. A., Neumann, R., Panayi, M., Ritchie, A. E., Shone, A. C., Signer, E., Stead, J. D., and Tamaki, K. (1999). Human minisatellites, repeat DNA instability and meiotic recombination. *Electrophoresis* 20, 1665-1675.
- Jeffreys, A. J., Bois, P., Buard, J., Collick, A., Dubrova, Y., Hollies, C. R., May, C. A., Murray, J., Neil, D. L., Neumann, R., Stead, J. D., Tamaki, K., and Yardley, J. (1997). Spontaneous and induced minisatellite instability. *Electrophoresis* 18, 1501-1511.
- Jeffreys, A. J., MacLeod, A., Tamaki, K., Neil, D. L., and Monckton, D. G. (1991a). Minisatellite repeat coding as a digital approach to DNA typing. *Nature* 354, 204-209.

- Jeffreys, A. J., Murray, J., and Neumann, R. (1998a). High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol. Cell* 2, 267-273.
- Jeffreys, A. J., Neil, D. L., and Neumann, R. (1998b). Repeat instability at human minisatellites arising from meiotic recombination. *Embo J.* 17, 4147-4157.
- Jeffreys, A. J., and Neumann, R. (1997). Somatic mutation processes at a human minisatellite. *Hum. Mol. Genet.* 6, 129-32; 134-136.
- Jeffreys, A. J., Neumann, R., and Wilson, V. (1990). Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60, 473-485.
- Jeffreys, A. J., Royle, N. J., Patel, I., Armour, J. A., MacLeod, A., Collick, A., Gray, I. C., Neumann, R., Gibbs, M., Crosier, M., Hill, M., Signer, E., and Monkton, D. (1991b). Principles and recent advances in human DNA fingerprinting. *Exs.* 58, 1-19.
- Jeffreys, A. J., Tamaki, K., MacLeod, A., Monkton, D. G., Neil, D. L., and Armour, J. A. (1994). Complex gene conversion events in germline mutation at human minisatellites. *Nat. Genet.* 6, 136-145.
- Jeffreys, A. J., Wilson, V., Kelly, R., Taylor, B. A., and Bulfield, G. (1987b). Mouse DNA 'fingerprints': analysis of chromosome localisation and germline stability of hypervariable loci in recombinant inbred strains. *Nucleic Acids Res.* 15, 2823-2836.
- Jeffreys, A. J., Wilson, V., Neumann, R., and Keyte, J. (1988). Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. *Nucleic Acids Res.* 16, 10953-10971.
- Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985a). Hypervariable 'minisatellite' regions in human DNA. *Nature* 314, 67-73.
- Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985b). Individual-specific 'fingerprints' of human DNA. *Nature* 316, 76-79.
- Jeffreys, A. J., Wilson, V., Thein, S. L., Weatherall, D. J., and Ponder, B. A. (1986). DNA 'fingerprints' and segregation analysis of multiple markers in human pedigrees. *Am. J. Hum. Genet.* 39, 11-24.
- Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L., and Schmid, C. W. (1980). Ubiquitous, interspersed repeated sequences in mammalian genomes. *Proc. Natl. Acad. Sci. USA* 77, 1398-1402.
- Jenkins, N. A., Copeland, N. G., Taylor, B. A., and Lee, B. K. (1981). Dilute (d) coat colour mutation of DBA/2J mice is associated with the site of integration of an ecotropic MuLV genome. *Nature* 293, 370-374.
- Jobling, M. A., Bouzekri, N., and Taylor, P. G. (1998). Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). *Hum. Mol. Genet.* 7, 643-653.
- Julier, C., de Gouyon, B., Georges, M., Guenet, J. L., Nakamura, Y., Avner, P., Lathrop, G. M. (1990). Minisatellite linkage maps in the mouse by cross-hybridization with human probes containing tandem repeats. *Proc. Natl. Acad. Sci. USA* 87, 4585-4589.
- Julier, C., Hyer, R. N., Davies, J., Merlin, F., Soularue, P., Briant, L., Cathelineau, G., Deschamps, I., Rotter, J. I., Froguel, P., Boitard, C., Bell, J. I., and Lathrop, G. M. (1991). Insulin-*IGF2* region on chromosome 11p encodes a gene implicated in *HLA-DR4*-dependent diabetes susceptibility. *Nature* 354, 155-159.
- Julier, C., Lucassen, A., Villedieu, P., Delepine, M., Levy-Marchal, C., Danze, P. M., Bianchi, F., Boitard, C., Froguel, P., Bell, J., and Lathrop, G. M. (1994). Multiple-DNA-variant-association-analysis: application to the insulin gene region in type 1 diabetes. *Am. J. Hum. Genet.* 55, 1247-1254.



- Kass, D. H., Kim, J., and Deininger, P. L. (1996). Sporadic amplification of ID elements in rodents. *J. Mol. Evol.* 42, 7-14.
- Kawaguchi, Y., Ikegami, H., Shen, G. Q., Nakagawa, Y., Fujisawa, T., Hamada, Y., Ueda, H., Fu, J., Uchigata, Y., Kitagawa, Y., Omori, Y., Shima, K., and Ogihara, T. (1997). Insulin gene region contributes to genetic susceptibility to, but may not to low incidence of, insulin-dependent diabetes mellitus in Japanese. *Biochem. Biophys. Res. Commun.* 233, 283-287.
- Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S., and Kakizuka, A. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.* 8, 221-228.
- Kazakov, V. S., Demidchik, E. P., and Astakhova, L. N. (1992). Thyroid cancer after Chernobyl. *Nature* 359, 21.
- Kelly, R. (1990). Hypervariable minisatellites in mouse DNA. (Ph. D. thesis). In Department of Genetics (Leicester: University of Leicester).
- Kelly, R., Bulfield, G., Collick, A., Gibbs, M., and Jeffreys, A. J. (1989). Characterization of a highly unstable mouse minisatellite locus: evidence for somatic mutation during early development. *Genomics* 5, 844-856.
- Kelly, R., Gibbs, M., Collick, A., and Jeffreys, A. J. (1991). Spontaneous mutation at the hypervariable mouse minisatellite locus *Ms6-hm*: flanking DNA sequence and analysis of germline and early somatic mutation events. *Proc. R. Soc. Lond. B. Biol. Sci.* 245, 235-245.
- Kelly, R. G. (1994). Similar origins of two mouse minisatellites within transposon-like LTRs. *Genomics* 24, 509-515.
- Kennedy, G. C., German, M. S., and Rutter, W. J. (1995). The minisatellite in the diabetes susceptibility locus *IDDM2* regulates insulin transcription. *Nat. Genet.* 9, 293-298.
- Kiaris, H., Ergazaki, M., and Spandidos, D. A. (1995). Instability at the *H-ras* minisatellite is associated with the spontaneous abortion of the embryo. *Biochem. Biophys. Res. Commun.* 214, 788-792.
- Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624-626.
- Kimura, M., and Crow, J. F. (1964). The number of alleles that can be contained in a finite population. *Genetics* 49, 725-738.
- Kipling, D., Ackford, H. E., Taylor, B. A., and Cooke, H. J. (1991). Mouse minor satellite DNA genetically maps to the centromere and is physically linked to the proximal telomere. *Genomics* 11, 235-241.
- Kipling, D., and Cooke, H. J. (1990). Hypervariable ultra-long telomeres in mice. *Nature* 347, 400-402.
- Kit, S. (1961). Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. *J. Mol. Biol.* 3, 711-716.
- Klesert, T. R., Otten, A. D., Bird, T. D., and Tapscott, S. J. (1997). Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of *DMAHP*. *Nat. Genet.* 16, 402-406.
- Kodaira, M., Satoh, C., Hiyama, K., and Toyama, K. (1995). Lack of effects of atomic bomb radiation on genetic instability of tandem-repetitive elements in human germ cells. *Am. J. Hum. Genet.* 57, 1275-1283.
- Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Naito, H., Tomoda, A., and Miike, T. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). *Nat. Genet.* 6, 9-13.

- Kooijman, R., van Buul-Offers, S. C., Scholtens, E. J., Rijkers, G. T., and Zegers, B. J. M. (1994). T cell development in IGF-II transgenic mice. *Growth Regul.* 4, 1-140.
- Koschinsky, M. L., Beisiegel, U., Henne-Bruns, D., Eaton, D. L., and Lawn, R. M. (1990). Apolipoprotein(a) size heterogeneity is related to variable number of repeat sequences in its mRNA. *Biochemistry* 29, 640-644.
- Kramerov, D. A., Grigoryan, A. A., Ryskov, A. P., and Georgiev, G. P. (1979). Long double-stranded sequences (dsRNA-B) of nuclear pre-mRNA consist of a few highly abundant classes of sequences: evidence from DNA cloning experiments. *Nucleic Acids Res.* 6, 697-713.
- Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A., and Georgiev, G. P. (1980). The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. *Nucleic Acids Res.* 8, 1201-1215.
- Krontiris, T. G. (1995). Minisatellites and human disease. *Science* 269, 1682-1683.
- Kulkarni, R. N., Bruning, J. C., Winnay, J. N., Postic, C., Magnuson, M. A., and Kahn, C. R. (1999). Tissue-specific knockout of the insulin receptor in pancreatic  $\beta$ -cells creates an insulin secretory defect similar to that in type 2 diabetes. *Cell* 96, 329-339.
- Kuokkanen, S., Sundvall, M., Terwilliger, J. D., Tienari, P. J., Wikstrom, J., Holmdahl, R., Pettersson, U., and Peltonen, L. (1996). A putative vulnerability locus to multiple sclerosis maps to 5p14-p12 in a region syntenic to the murine locus *Eae2*. *Nat. Genet.* 13, 477-480.
- Kyvik, K. O., Green, A., and Beck-Nielsen, H. (1995). Concordance rates of insulin dependent diabetes mellitus: a population based study of young Danish twins. *B. M. J.* 311, 913-917.
- La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E., and Fischbeck, K. H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. *Nature* 352, 77-79.
- Lalioti, M. D., Scott, H. S., and Antonarakis, S. E. (1999). Altered spacing of promoter elements due to the dodecamer repeat expansion contributes to reduced expression of the cystatin B gene in EPM1. *Hum. Mol. Genet.* 8, 1791-1798.
- Lalioti, M. D., Scott, H. S., Buresi, C., Rossier, C., Bottani, A., Morris, M. A., Malafosse, A., and Antonarakis, S. E. (1997). Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature* 386, 847-851.
- Lalioti, M. D., Scott, H. S., Genton, P., Grid, D., Ouazzani, R., M'Rabet, A., Ibrahim, S., Gouider, R., Dravet, C., Chkili, T., Bottani, A., Buresi, C., Malafosse, A., and Antonarakis, S. E. (1998). A PCR amplification method reveals instability of the dodecamer repeat in progressive myoclonus epilepsy (EPM1) and no correlation between the size of the repeat and age at onset. *Am. J. Hum. Genet.* 62, 842-847.
- Lancaster, C. A., Peat, N., Duhig, T., Wilson, D., Taylor-Papadimitriou, J., and Gendler, S. J. (1990). Structure and expression of the human polymorphic epithelial mucin gene: an expressed VNTR unit. *Biochem. Biophys. Res. Commun.* 173, 1019-1029.
- Larson, G. P., Ding, S., Lafreniere, R. G., Rouleau, G. A., and Krontiris, T. G. (1999). Instability of the EPM1 minisatellite. *Hum. Mol. Genet.* 8, 1985-1988.
- Leach, F. S., Nicolaides, N. C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomaki, P., Sistonen, P., Aaltonen, L. A., Nystrom-Lahti, M., Guan, X. Y., Zhang, J., Meltzer, P. S., Yu, J. W., Kao, F. T., Chen, D. J., Cerosaletti, K. M., Fournier, R. E. K., Todd, S., Lewis, T., Leach, R. J., Naylor, S. L., Weissenbach, J., Mecklin, J. P., Jarvinen, H., Petersen, G. M., Hamilton, S. R., Green, J., Jass, J., Watson, P., Lynch, H. T., Trent, J. M., Delachapelle, A., Kinzler, K. W., and Vogelstein, B. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* 75, 1215-1225.

- Leeflang, E. P., Liu, W. M., Chesnokov, I. N., and Schmid, C. W. (1993). Phylogenetic isolation of a human Alu founder gene: drift to new subfamily identity. *J. Mol. Evol.* 37, 559-565.
- Leeflang, E. P., Tavare, S., Marjoram, P., Neal, C. O., Srinidhi, J., MacFarlane, H., MacDonald, M. E., Gusella, J. F., de Young, M., Wexler, N. S., and Arnheim, N. (1999). Analysis of germline mutation spectra at the Huntington's disease locus supports a mitotic mutation mechanism. *Hum. Mol. Genet.* 8, 173-183.
- Lehrman, M. A., Russell, D. W., Goldstein, J. L., and Brown, M. S. (1987). Alu-Alu recombination deletes splice acceptor sites and produces secreted low density lipoprotein receptor in a subject with familial hypercholesterolemia. *J. Biol. Chem.* 262, 3354-3361.
- Levinson, G., and Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4, 203-221.
- Lewin, B. (1994). *Genes V* (New York: Oxford University Press).
- Lia, A. S., Seznec, H., Hofmann-Radvanyi, H., Radvanyi, F., Duros, C., Saquet, C., Blanche, M., Junien, C., and Gourdon, G. (1998). Somatic instability of the CTG repeat in mice transgenic for the myotonic dystrophy region is age dependent but not correlated to the relative inter-tissue transcription levels and proliferative capacities. *Hum. Mol. Genet.* 7, 1285-1291.
- Lichter, J. B., Barr, C. L., Kennedy, J. L., Van Tol, H. H., Kidd, K. K., and Livak, K. J. (1993). A hypervariable segment in the human dopamine receptor D4 (*DRD4*) gene. *Hum. Mol. Genet.* 2, 767-773.
- Lie, B. A., Todd, J. A., Pociot, F., Nerup, J., Akselsen, H. E., Joner, G., Dahl-Jorgensen, K., Ronningen, K. S., Thorsby, E., and Undlien, D. E. (1999). The predisposition to type 1 diabetes linked to the human leukocyte antigen complex includes at least one non-class II gene. *Am. J. Hum. Genet.* 64, 793-800.
- Lieber, M. R. (1997). The FEN-1 family of structure-specific nucleases in eukaryotic DNA replication, recombination and repair. *Bioessays* 19, 233-240.
- Likhtarev, I. A., Sobolev, B. G., Kairo, I. A., Tronko, N. D., Bogdanova, T. I., Oleinic, V. A., Epshtein, E. V., and Beral, V. (1995). Thyroid cancer in the Ukraine. *Nature* 375, 365.
- Lindstedt, B. A., Ryberg, D., and Haugen, A. (1997). Rare alleles at different VNTR loci among lung-cancer patients with microsatellite instability in tumours. *Int. J. Cancer* 70, 412-415.
- Liu, Z., Frantz, J. D., Gilbert, W., and Tye, B. K. (1993). Identification and characterization of a nuclease activity specific for G4 tetrastranded DNA. *Proc. Natl. Acad. Sci. USA* 90, 3157-3161.
- Liu, Z., and Gilbert, W. (1994). The yeast KEM1 gene encodes a nuclease specific for G4 tetraplex DNA: implication of *in vivo* functions for this novel DNA structure. *Cell* 77, 1083-1092.
- Livshits, L. A., Malyarchuk, S. G., Luk'yanova, E. M., Antipkin, Y. G., Arabskaya, L. P., Kravchenko, S. A., Matsuka, G. H., Petit, E., Giraudeau, F., Le Guen, B., and Vergnaud, G. (1999). Heritable mutations at some minisatellite loci; analysis in children of liquidators of Chernobyl accident consequences. *Int. J. Radiat. Med.* 1, 101-106.
- Loenen, W. A., and Brammar, W. J. (1980). A bacteriophage  $\lambda$  vector for cloning large DNA fragments made with several restriction enzymes. *Gene* 10, 249-259.
- Lucassen, A. M., Julier, C., Beressi, J. P., Boitard, C., Froguel, P., Lathrop, M., and Bell, J. I. (1993). Susceptibility to insulin dependent diabetes mellitus maps to a 4.1 kb segment of DNA spanning the insulin gene and associated VNTR. *Nat. Genet.* 4, 305-310.
- Lucassen, A. M., Sreaton, G. R., Julier, C., Elliott, T. J., Lathrop, M., and Bell, J. I. (1995). Regulation of insulin gene expression by the IDDM associated, insulin locus haplotype. *Hum. Mol. Genet.* 4, 501-506.

- Luo, D. F., Buzzetti, R., Rotter, J. I., Maclaren, N. K., Raffel, L. J., Nistico, L., Giovannini, C., Pozzilli, P., Thomson, G., and She, J. X. (1996). Confirmation of three susceptibility genes to insulin-dependent diabetes mellitus: *IDDM4*, *IDDM5* and *IDDM8*. *Hum. Mol. Genet.* 5, 693-698.
- Macina, R. A., Morii, K., Hu, X. L., Negorev, D. G., Spais, C., Ruthig, L. A., and Riethman, H. C. (1995). Molecular cloning and RARE cleavage mapping of human 2p, 6q, 8q, 12q, and 18q telomeres. *Genome Res.* 5, 225-232.
- Majerus, M. E. N., Amos, W., and Hurst, G. D. (1996). *Evolution; the four billion year war*: Addison Wesley Longman Limited.
- Malaisse, W. J. (1982). Alloxan toxicity to the pancreatic B-cell. A new hypothesis. *Biochem. Pharmacol.* 31, 3527-3534.
- Mangiarini, L., Sathasivam, K., Mahal, A., Mott, R., Seller, M., and Bates, G. P. (1997). Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. *Nat. Genet.* 15, 197-200.
- Manly, K. F. (1993). A Macintosh program for storage and analysis of experimental genetic mapping data. *Mamm. Genome* 4, 303-313.
- Mann, S. J. (1971). Hair loss and cyst formation in hairless and rhino mutant mice. *Anat. Rec.* 170, 485-499.
- Mariat, D., and Vergnaud, G. (1992). Detection of polymorphic loci in complex genomes with synthetic tandem repeats. *Genomics* 12, 454-458.
- Marron, M. P., Raffel, L. J., Garchon, H. J., Jacob, C. O., Serrano-Rios, M., Martinez Larrad, M. T., Teng, W. P., Park, Y., Zhang, Z. X., Goldstein, D. R., Tao, Y. W., Beaurain, G., Bach, J. F., Huang, H. S., Luo, D. F., Zeidler, A., Rotter, J. I., Yang, M. C., Modilevsky, T., Maclaren, N. K., and She, J. X. (1997). Insulin-dependent diabetes mellitus (IDDM) is associated with *CTLA4* polymorphisms in multiple ethnic groups. *Hum. Mol. Genet.* 6, 1275-1282.
- Masumoto, H., Masukata, H., Muro, Y., Nozaki, N., and Okazaki, T. (1989). A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *J. Cell Biol.* 109, 1963-1973.
- May, C. A., Jeffreys, A. J., and Armour, J. A. (1996). Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (*D16S309*). *Hum. Mol. Genet.* 5, 1823-1833.
- McCarthy, M. (1998). Weighing in on diabetes risk. *Nat. Genet.* 19, 209-210.
- McGinnis, R. E., and Spielman, R. S. (1995). Insulin gene 5' flanking polymorphism. Length of class I alleles in number of repeat units. *Diabetes* 44, 1296-1302.
- Mein, C. A., Esposito, L., Dunn, M. G., Johnson, G. C., Timms, A. E., Goy, J. V., Smith, A. N., Sebag-Montefiore, L., Merriman, M. E., Wilson, A. J., Pritchard, L. E., Cucca, F., Barnett, A. H., Bain, S. C., and Todd, J. A. (1998). A search for type 1 diabetes susceptibility genes in families from the United Kingdom. *Nat. Genet.* 19, 297-300.
- Meloni, R., Albanese, V., Ravassard, P., Treilhou, F., and Mallet, J. (1998). A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element *in vitro*. *Hum. Mol. Genet.* 7, 423-428.
- Merriman, T. R., Eaves, I. A., Twells, R. C., Merriman, M. E., Danoy, P. A., Muxworthy, C. E., Hunter, K. M., Cox, R. D., Cucca, F., McKinney, P. A., Shield, J. P., Baum, J. D., Tuomilehto, J., Tuomilehto-Wolf, E., Ionesco-Tirgoviste, C., Joner, G., Thorsby, E., Undlien, D. E., Pociot, F., Nerup, J., Ronningen, K. S., Bain, S. C., and Todd, J. A. (1998). Transmission of haplotypes of microsatellite markers rather than single marker alleles in the mapping of a putative type 1 diabetes susceptibility gene (*IDDM6*). *Hum. Mol. Genet.* 7, 517-524.

- Messier, W., Li, S. H., and Stewart, C. B. (1996). The birth of microsatellites. *Nature* 381, 483.
- Mette, M. F., van der Winden, J., Matzke, M. A., and Matzke, A. J. (1999). Production of aberrant promoter transcripts contributes to methylation and silencing of unlinked homologous promoters in *trans*. *Embo J.* 18, 241-248.
- Meyer, E., Wiegand, P., Rand, S. P., Kuhlmann, D., Brack, M., and Brinkmann, B. (1995). Microsatellite polymorphisms reveal phylogenetic relationships in primates. *J. Mol. Evol.* 41, 10-14.
- Mijovic, C. H., Penny, M. A., Jenkins, D., Jacobs, K., Heward, J., Knight, S. W., Lucassen, A., Morrison, E., and Barnett, A. H. (1997). The insulin gene region and susceptibility to insulin-dependent diabetes mellitus in four races; new insights from Afro-Caribbean race-specific haplotypes. *Autoimmunity* 26, 11-22.
- Mitas, M. (1997). Trinucleotide repeats associated with human disease. *Nucleic Acids Res.* 25, 2245-2254.
- Monckton, D. G., Coolbaugh, M. I., Ashizawa, K. T., Siciliano, M. J., and Caskey, C. T. (1997). Hypermutable myotonic dystrophy CTG repeats in transgenic mice. *Nat. Genet.* 15, 193-196.
- Monckton, D. G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A., and Jeffreys, A. J. (1994). Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nat. Genet.* 8, 162-170.
- Monckton, D. G., Tamaki, K., MacLeod, A., Neil, D. L., and Jeffreys, A. J. (1993). Allele-specific MVR-PCR analysis at minisatellite *D1S8*. *Hum. Mol. Genet.* 2, 513-519.
- Moriwaki, K. (1987). Genetic significance of laboratory mice in biomedical research. *Prog. Clin. Biol. Res.* 229, 53-72.
- Mulholland, J., and Botstein, D. (1986). Oligonucleotide repeats involved in the highly polymorphic locus *D14S1*. *Am. J. Hum. Genet.* 39, A226 abstract.
- Murray, J., Buard, J., Neil, D. L., Yeramian, E., Tamaki, K., Hollies, C., and Jeffreys, A. J. (1999). Comparative sequence analysis of human minisatellites showing meiotic repeat instability. *Genome Res.* 9, 130-136.
- Nachman, M. W., Boyer, S. N., Searle, J. B., and Aquadro, C. F. (1994). Mitochondrial DNA variation and the evolution of Robertsonian chromosomal races of house mice, *Mus domesticus*. *Genetics* 136, 1105-1120.
- Nakagama, H., Kaneko, S., Shima, H., Inamori, H., Fukuda, H., Kominami, R., Sugimura, T., and Nagao, M. (1997). Induction of minisatellite mutation in NIH 3T3 cells by treatment with the tumor promoter okadaic acid. *Proc. Natl. Acad. Sci. USA* 94, 10813-10816.
- Nakagawa, Y., Kawaguchi, Y., Twells, R. C., Muxworthy, C., Hunter, K. M., Wilson, A., Merriman, M. E., Cox, R. D., Merriman, T., Cucca, F., McKinney, P. A., Shield, J. P., Tuomilehto, J., Tuomilehto-Wolf, E., Ionesco-Tirgoviste, C., Nistico, L., Buzzetti, R., Pozzilli, P., Joner, G., Thorsby, E., Undlien, D. E., Pociot, F., Nerup, J., Ronningen, K. S., and Todd, J. A. (1998). Fine mapping of the diabetes-susceptibility locus, *IDDM4*, on chromosome 11q13. *Am. J. Hum. Genet.* 63, 547-556.
- Nakamura, Y., Cowen, J., Krapcho, K., O'Connell, P., Leppert, M., Lathrop, G. M., Lalouel, J. M., and White, R. (1988). Isolation and mapping of a polymorphic DNA sequence pJCZ3.1 on chromosome 19 [*D19S20*]. *Nucleic Acids Res.* 16, 1229.
- Nakamura, Y., Koyama, K., and Matsushima, M. (1998). VNTR (variable number of tandem repeat) sequences as transcriptional, translational, or functional regulators. *J. Hum. Genet.* 43, 149-152.
- Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E., and White, R. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* 235, 1616-1622.

- Nassif, N., Penney, J., Pal, S., Engels, W. R., and Gloor, G. B. (1994). Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. *Mol. Cell. Biol.* 14, 1613-1625.
- Neil, D. L., and Jeffreys, A. J. (1993). Digital DNA typing at a second hypervariable locus by minisatellite variant repeat mapping. *Hum. Mol. Genet.* 2, 1129-1135.
- Nepom, G. T., and Kwok, W. W. (1998). Molecular basis for *HLA-DQ* associations with IDDM. *Diabetes* 47, 1177-1184.
- Nerup, J., Platz, P., Andersen, O. O., Christy, M., Lyngsoe, J., Poulsen, J. E., Ryder, L. P., Nielsen, L. S., Thomsen, M., and Svejgaard, A. (1974). HLA antigens and diabetes mellitus. *Lancet* 2, 864-866.
- Neumann, B., Kubicka, P., and Barlow, D. P. (1995). Characteristics of imprinted genes. *Nat. Genet.* 9, 12-13.
- Nistico, L., Buzzetti, R., Pritchard, L. E., Van der Auwera, B., Giovannini, C., Bosi, E., Larrad, M. T., Rios, M. S., Chow, C. C., Cockram, C. S., Jacobs, K., Mijovic, C., Bain, S. C., Barnett, A. H., Vandewalle, C. L., Schuit, F., Gorus, F. K., Tosi, R., Pozzilli, P., and Todd, J. A. (1996). The *CTLA-4* gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Hum. Mol. Genet.* 5, 1075-1080.
- Noble, J. A., Valdes, A. M., Cook, M., Klitz, W., Thomson, G., and Erlich, H. A. (1996). The role of HLA class II genes in insulin-dependent diabetes mellitus: molecular analysis of 180 Caucasian, multiplex families. *Am. J. Hum. Genet.* 59, 1134-1148.
- Nothen, M. M., Cichon, S., Vogt, I. R., Hemmer, S., Kruse, R., Knapp, M., Holler, T., Faiyaz ul Haque, M., Haque, S., Propping, P., Ahmad, M., and Rietschel, M. (1998). A gene for universal congenital alopecia maps to chromosome 8p21-22. *Am. J. Hum. Genet.* 62, 386-390.
- O'Dell, S. D., Bujac, S. R., Miller, G. J., and Day, I. N. (1999). Associations of *IGF2* *ApaI* RFLP and *INS* VNTR class I allele size with obesity. *Eur. J. Hum. Genet.* 7, 821-827.
- O'Malley, K. L., and Rotwein, P. (1988). Human tyrosine hydroxylase and insulin genes are contiguous on chromosome 11. *Nucleic Acids Res.* 16, 4437-4446.
- Ogilvie, A. D., Battersby, S., Bubb, V. J., Fink, G., Harmar, A. J., Goodwim, G. M., and Smith, C. A. (1996). Polymorphism in serotonin transporter gene associated with susceptibility to major depression. *Lancet* 347, 731-733.
- Ohshima, K., Koishi, R., Matsuo, M., and Okada, N. (1993). Several short interspersed repetitive elements (SINES) in distant species may have originated from a common ancestral retrovirus: characterization of a squid SINE and a possible mechanism for generation of tRNA-derived retroposons. *Proc. Natl. Acad. Sci. USA* 90, 6260-6264.
- Okumura, K., Kiyama, R., and Oishi, M. (1987). Sequence analyses of extrachromosomal *Sau3A* and related family DNA: analysis of recombination in the excision event. *Nucleic Acids Res.* 15, 7477-7489.
- Olek, A., Oswald, J., and Walter, J. (1996). A modified and improved method for bisulphite based cytosine methylation analysis. *Nucleic Acids Res.* 24, 5064-5066.
- Olmos, P., A'Hern, R., Heaton, D. A., Millward, B. A., Risley, D., Pyke, D. A., and Leslie, R. D. (1988). The significance of the concordance rate for type 1 (insulin-dependent) diabetes in identical twins. *Diabetologia* 31, 747-750.
- Ong, K. K., Phillips, D. I., Fall, C., Poulton, J., Bennett, S. T., Golding, J., Todd, J. A., and Dunger, D. B. (1999). The insulin gene VNTR, type 2 diabetes and birth weight. *Nat. Genet.* 21, 262-263.
- Orr, H. T., Chung, M. Y., Banfi, S., Kwiatkowski, T. J., Jr., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P., and Zoghbi, H. Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. *Nat. Genet.* 4, 221-226.

- Osman, F., and Subramani, S. (1998). Double-strand break-induced recombination in eukaryotes. *Prog. Nucleic Acid Res. Mol. Biol.* 58, 263-299.
- Owerbach, D., and Gabbay, K. H. (1993). Localisation of a type 1 diabetes susceptibility locus to the variable tandem repeat region flanking the insulin gene. *Diabetes* 42, 1708-1714.
- Owerbach, D., and Gabbay, K. H. (1996). The search for IDDM susceptibility genes: the next generation. *Diabetes* 45, 544-551.
- Owerbach, D., Poulsen, S., Billesbolle, P., and Nerup, J. (1982). DNA insertion sequences near the insulin gene affect glucose regulation. *Lancet* 1, 880-883.
- Page, D. C., Mosher, R., Simpson, E. M., Fisher, E. M., Mardon, G., Pollack, J., McGillivray, B., de la Chapelle, A., and Brown, L. G. (1987). The sex-determining region of the human Y chromosome encodes a finger protein. *Cell* 51, 1091-1104.
- Paques, F., Leung, W. Y., and Haber, J. E. (1998). Expansions and contractions in a tandem repeat induced by double-strand break repair. *Mol. Cell. Biol.* 18, 2045-2054.
- Paquette, J., Giannoukakis, N., Polychronakos, C., Vafiadis, P., and Deal, C. (1998). The *INS* 5' variable number of tandem repeats is associated with *IGF2* expression in humans. *J. Biol. Chem.* 273, 14158-14164.
- Paulin, R., Grigg, G. W., Davey, M. W., and Piper, A. A. (1998). Urea improves efficiency of bisulphite-mediated sequencing of 5'-methylcytosine in genomic DNA. *Nucleic Acids Res.* 26, 5009-5010.
- Pelissier, T., Thalmeir, S., Kempe, D., Sanger, H. L., and Wassenegger, M. (1999). Heavy *de novo* methylation at symmetrical and non-symmetrical sites is a hallmark of RNA-directed DNA methylation. *Nucleic Acids Res.* 27, 1625-1634.
- Pennacchio, L. A., Lehesjoki, A. E., Stone, N. E., Willour, V. L., Virtaneva, K., Miao, J., D'Amato, E., Ramirez, L., Faham, M., Koskiniemi, M., Warrington, J. A., Norio, R., de la Chapelle, A., Cox, D. R., and Myers, R. M. (1996). Mutations in the gene encoding cystatin B in progressive myoclonus epilepsy (EPM1). *Science* 271, 1731-1734.
- Permutt, M. A., Rotwein, P., Andreone, T., Ward, W. K., and Porte, D., Jr. (1985). Islet  $\beta$ -cell function and polymorphism in the 5'-flanking region of the human insulin gene. *Diabetes* 34, 311-314.
- Petes, T. D., Greenwell, P. W., and Dominska, M. (1997). Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*. *Genetics* 146, 491-498.
- Phelan, C. M., Rebbeck, T. R., Weber, B. L., Devilee, P., Rutledge, M. H., Lynch, H. T., Lenoir, G. M., Stratton, M. R., Easton, D. F., Ponder, B. A., Cannon-Albright, L., Larsson, C., Goldgar, D. E., and Narod, S. A. (1996). Ovarian cancer risk in *BRCA1* carriers is modified by the *HRAS1* variable number of tandem repeat (VNTR) locus. *Nat. Genet.* 12, 309-311.
- Pietras, D. F., Bennett, K. L., Siracusa, L. D., Woodworth-Gutai, M., Chapman, V. M., Gross, K. W., Kane-Haas, C., and Hastie, N. D. (1983). Construction of a small *Mus musculus* repetitive DNA library: identification of a new satellite sequence in *Mus musculus*. *Nucleic Acids Res.* 11, 6965-6983.
- Polychronakos, C., Giannoukakis, N., and Deal, C. L. (1995). Imprinting of *IGF2*, insulin-dependent diabetes, immune function, and apoptosis: a hypothesis. *Dev. Genet.* 17, 253-262.
- Polymeropoulos, M. H., Swift, R. G., and Swift, M. (1994). Linkage of the gene for Wolfram syndrome to markers on the short arm of chromosome 4. *Nat. Genet.* 8, 95-97.
- Polymeropoulos, M. H., Xiao, H., Rath, D. S., and Merrill, C. R. (1991). Tetranucleotide repeat polymorphism at the human tyrosine hydroxylase gene (*TH*). *Nucleic Acids Res.* 19, 3753.

- Prosser, J., Frommer, M., Paul, C., and Vincent, P. C. (1986). Sequence relationships of three human satellite DNAs. *J. Mol. Biol.* 187, 145-155.
- Proudfoot, N. J., Gil, A., and Maniatis, T. (1982). The structure of the human  $\zeta$ -globin gene and a closely linked, nearly identical pseudogene. *Cell* 31, 553-563.
- Puers, C., Hammond, H. A., Jin, L., Caskey, C. T., and Schumm, J. W. (1993). Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus *HUMTH01*[AATG]<sub>n</sub> and reassignment of alleles in population analysis by using a locus-specific allelic ladder. *Am. J. Hum. Genet.* 53, 953-958.
- Pugliese, A., Zeller, M., Fernandez, A., Jr., Zalberg, L. J., Bartlett, R. J., Ricordi, C., Pietropaolo, M., Eisenbarth, G. S., Bennett, S. T., and Patel, D. D. (1997). The insulin gene is transcribed in the human thymus and transcription levels correlated with allelic variation at the *INS* VNTR-*IDDM2* susceptibility locus for type 1 diabetes. *Nat. Genet.* 15, 293-297.
- Pujol-Borrell, R., Todd, I., Doshi, M., Bottazzo, G. F., Sutton, R., Gray, D., Adolf, G. R., and Feldmann, M. (1987). *HLA* class II induction in human islet cells by interferon- $\gamma$  plus tumour necrosis factor or lymphotoxin. *Nature* 326, 304-306.
- Pulst, S. M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X. N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., DeJong, P., Rouleau, G. A., Auburger, G., Korenberg, J. R., Figueroa, C., and Sahba, S. (1996). Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat. Genet.* 14, 269-276.
- Pupko, T., and Graur, D. (1999). Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: role of length and number of repeated units. *J. Mol. Evol.* 48, 313-316.
- Richards, R. I., and Sutherland, G. R. (1997). Dynamic mutation: possible mechanisms and significance in human disease. *Trends Biochem. Sci.* 22, 432-436.
- Ricke, D. O., Liu, Q., Gostout, B., and Sommer, S. S. (1995). Nonrandom patterns of simple and cryptic triplet repeats in coding and noncoding sequences. *Genomics* 26, 510-520.
- Risch, N. (1987). Assessing the role of *HLA*-linked and unlinked determinants of disease. *Am. J. Hum. Genet.* 40, 1-14.
- Risch, N., Ghosh, S., and Todd, J. A. (1993). Statistical evaluation of multiple-locus linkage data in experimental species and its relevance to human studies: application to nonobese diabetic (NOD) mouse and human insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 53, 702-714.
- Robinson, M., Gautier, C., and Mouchiroud, D. (1997). Evolution of isochores in rodents. *Mol. Biol. Evol.* 14, 823-828.
- Ronningen, K. S., Halstein, T. S., Undlien, D. E., Ploski, R., Welsh, K., Todd, J. A., Kockun, I., de Vries, N., Kimura, A., Thorsby, E., and Green, A. (1994). Correlation between incidence of childhood onset diabetes and frequency of high risk *HLA*-markers in European populations. 15th International Federation Congress, 116.
- Rotwein, P., Yokoyama, S., Didier, D. K., and Chirgwin, J. M. (1986). Genetic analysis of the hypervariable region flanking the human insulin gene. *Am. J. Hum. Genet.* 39, 291-299.
- Royle, N. J., Armour, J. A., Webb, M., Thomas, A., and Jeffreys, A. J. (1992). A hypervariable locus *D16S309* located at the distal end of 16p. *Nucleic Acids Res.* 20, 1164.
- Royle, N. J., Baird, D. M., and Jeffreys, A. J. (1994). A subterminal satellite located adjacent to telomeres in chimpanzees is absent from the human genome. *Nat. Genet.* 6, 52-56.



- Royle, N. J., Clarkson, R., Wong, Z., and Jeffreys, A. J. (1988a). Human gene mapping 9: ninth international workshop on human gene mapping. *Cytogenet. Cell Genet.* 46, 685.
- Royle, N. J., Clarkson, R. E., Wong, Z., and Jeffreys, A. J. (1988b). Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* 3, 352-360.
- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239, 487-491.
- Saito, I., and Stark, G. R. (1986). Charomids: cosmid vectors for efficient cloning and mapping of large or small restriction fragments. *Proc. Natl. Acad. Sci. USA* 83, 8664-8668.
- Sakamoto, K., and Okada, N. (1985). Rodent type 2 Alu family, rat identifier sequence, rabbit C family, and bovine or goat 73 bp repeat may have evolved from tRNA genes. *J. Mol. Evol.* 22, 134-140.
- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular cloning; a laboratory manual*, Second Edition.
- Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., Wakisaka, A., Tashiro, K., Ishida, Y., Ikeuchi, T., Koide, R., Saito, M., Sato, A., Tanaka, T., Hanyu, S., Takiyama, Y., Nishizawa, M., Shimizu, N., Nomura, Y., Segawa, M., Iwabuchi, K., Eguchi, I., Tanaka, H., Takahashi, H., and Tsuji, S. (1996). Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. *Nat. Genet.* 14, 277-284.
- SantaLucia, J., Jr. (1998). A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA* 95, 1460-1465.
- SantaLucia, J., Allawi, H. T., and Seneviratne, P. A. (1996). Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35, 3555-3562.
- SantaLucia, J. J., Zuker, M., Bommarito, A., and Irani, R. J. (1999). (unpublished results).
- Sassaman, D. M., Dombroski, B. A., Moran, J. V., Kimberland, M. L., Naas, T. P., DeBerardinis, R. J., Gabriel, A., Swergold, G. D., and Kazazian, H. H., Jr. (1997). Many human L1 elements are capable of retrotransposition. *Nat. Genet.* 16, 37-43.
- Sato, T., Oyake, M., Nakamura, K., Nakao, K., Fukusima, Y., Onodera, O., Igarashi, S., Takano, H., Kikugawa, K., Ishida, Y., Shimohata, T., Koide, R., Ikeuchi, T., Tanaka, H., Futamura, N., Matsumura, R., Takayanagi, T., Tanaka, F., Sobue, G., Komure, O., Takahashi, M., Sano, A., Ichikawa, Y., Goto, J., Kanazawa, I., Katsuki, M., and Tsuji, S. (1999). Transgenic mice harboring a full-length human mutant *DRPLA* gene exhibit age-dependent intergenerational and somatic instabilities of CAG repeats comparable with those in *DRPLA* patients. *Hum. Mol. Genet.* 8, 99-106.
- Satoh, C., and Kodaira, M. (1996). Effects of radiation on children. *Nature* 383, 226.
- Satsangi, J., Parkes, M., Louis, E., Hashimoto, L., Kato, N., Welsh, K., Terwilliger, J. D., Lathrop, G. M., Bell, J. I., and Jewell, D. P. (1996). Two stage genome-wide search in inflammatory bowel disease provides evidence for susceptibility loci on chromosomes 3, 7 and 12. *Nat. Genet.* 14, 199-202.
- Schaffer, H. E., and Sederoff, R. R. (1981). Improved estimation of DNA fragment lengths from agarose gels. *Anal. Biochem.* 115, 113-122.
- Schichman, S. A., Severynse, D. M., Edgell, M. H., and Hutchison, C. A. d. (1992). Strand-specific LINE-1 transcription in mouse F9 cells originates from the youngest phylogenetic subgroup of LINE-1 elements. *J. Mol. Biol.* 224, 559-574.
- Schierer, T., and Henderson, E. (1994). A protein from *Tetrahymena thermophila* that specifically binds parallel-stranded G4-DNA. *Biochemistry* 33, 2240-2246.

- Schwartz, R. H. (1997). T cell clonal anergy. *Curr. Opin. Immunol.* 9, 351-357.
- Seino, S., Bell, G. I., and Li, W. H. (1992). Sequences of primate insulin genes support the hypothesis of a slower rate of molecular evolution in humans and apes than in monkeys. *Mol. Biol. Evol.* 9, 193-203.
- Selander, R. K., and Yang, S. Y. (1969). Protein polymorphism and genic heterozygosity in a wild population of the house mouse (*Mus musculus*). *Genetics* 63, 653-667.
- Selker, E. U. (1997). Epigenetic phenomena in filamentous fungi: useful paradigms or repeat-induced confusion? *Trends Genet.* 13, 296-301.
- Serdobova, I. M., and Kramerov, D. A. (1998). Short retroposons of the B2 superfamily: evolution and application for the study of rodent phylogeny. *J. Mol. Evol.* 46, 202-214.
- Shiels, P. G., Kind, A. J., Campbell, K. H., Waddington, D., Wilmut, I., Colman, A., and Schnieke, A. E. (1999). Analysis of telomere lengths in cloned sheep. *Nature* 399, 316-317.
- Signer, E. N., Dubrova, Y. E., Jeffreys, A. J., Wilde, C., Finch, L. M., Wells, M., and Peaker, M. (1998). DNA fingerprinting Dolly. *Nature* 394, 329-330.
- Singal, D. P., and Blajchman, M. A. (1973). Histocompatibility (HL-A) antigens, lymphocytotoxic antibodies and tissue antibodies in patients with diabetes mellitus. *Diabetes* 22, 429-432.
- Singer, M. F. (1982). Highly repeated sequences in mammalian genomes. *Int. Rev. Cytol.* 76, 67-112.
- Smit, A. F., and Riggs, A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* 23, 98-102.
- Smith, K. N., and Nicolas, A. (1998). Recombination at work for meiosis. *Curr. Opin. Genet. Dev.* 8, 200-211.
- Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* 98, 503-517.
- Southern, E. M. (1979). Measurement of DNA length by gel electrophoresis. *Anal. Biochem.* 100, 319-323.
- Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52, 506-516.
- Spurr, N. K., Bryant, S. P., Attwood, J., Nyberg, K., Cox, S. A., Mills, A., Bains, R., Warne, D., Cullin, L., Povey, S., Sebaoun, J. M., Weissenbach, J., Cann, H. M., Lathrop, M., Dausset, J., Marcadet-Troton, A. and Cohen, D. (1994). European Gene Mapping Project (EUROGEM): genetic maps based on the CEPH reference families. *Eur. J. Hum. Genet.* 2, 193-203.
- Stallings, R. L. (1994). Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. *Genomics* 21, 116-121.
- Starling, J. A., Maule, J., Hastie, N. D., and Allshire, R. C. (1990). Extensive telomere repeat arrays in mouse are hypervariable. *Nucleic Acids Res.* 18, 6881-6888.
- Stoye, J. P., Fenner, S., Greenoak, G. E., Moran, C., and Coffin, J. M. (1988). Role of endogenous retroviruses as mutagens: the hairless mutation of mice. *Cell* 54, 383-391.
- Strom, T. M., Hortnagel, K., Hofmann, S., Gekeler, F., Scharfe, C., Rabl, W., Gerbitz, K. D., and Meitinger, T. (1998). Diabetes insipidus, diabetes mellitus, optic atrophy and deafness (DIDMOAD) caused by mutations in a novel gene (wolframin) coding for a predicted transmembrane protein. *Hum. Mol. Genet.* 7, 2021-2028.

- Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K. (1996). Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.* 24, 4501-4505.
- Tamaki, K., Huang, X. L., Mizutani, M., Yamamoto, T., Katsumata, R., Uchihi, R., Katsumata, Y., and Jeffreys, A. J. (1999a). The potential contribution of MVR-PCR to paternity probabilities in a case lacking a mother. *J. Forensic Sci.* 44, 863-867.
- Tamaki, K., Huang, X. L., Yamamoto, T., Uchihi, R., Nozawa, H., and Katsumata, Y. (1995). Applications of minisatellite variant repeat (MVR) mapping for maternal identification from remains of an infant and placenta. *J. Forensic Sci.* 40, 695-700.
- Tamaki, K., May, C. A., Dubrova, Y. E., and Jeffreys, A. J. (1999b). Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7. *Hum. Mol. Genet.* 8, 879-888.
- Tamaki, K., Monckton, D. G., MacLeod, A., Allen, M., and Jeffreys, A. J. (1993). Four-state MVR-PCR: increased discrimination of digital DNA typing by simultaneous analysis of two polymorphic sites within minisatellite variant repeats at *DIS8*. *Hum. Mol. Genet.* 2, 1629-1632.
- Tamaki, K., Monckton, D. G., MacLeod, A., Neil, D. L., Allen, M., and Jeffreys, A. J. (1992). Minisatellite variant repeat (MVR) mapping: analysis of 'null' repeat units at *DIS8*. *Hum. Mol. Genet.* 1, 401-406.
- Tartof, K. D., and Hobbs, C. A. (1987). Improved media for growing plasmid and cosmid clones. *Bethesda Res. Lab. Focus* 9.
- Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* 17, 6463-6471.
- Taylor, B. A. (1978). Recombinant inbred strains: use in gene mapping. In *Origins of inbred mice*, H. C. Morse, ed. (New York: Academic Press), pp. 423-428.
- Thompson, C. C. (1996). Thyroid hormone-responsive genes in developing cerebellum include a novel synaptotagmin and a hairless homolog. *J. Neurosci.* 16, 7832-7840.
- Tisch, R., and McDevitt, H. (1996). Insulin-dependent diabetes mellitus. *Cell* 85, 291-297.
- Todd, J. A. (1999a). From genome to aetiology in a multifactorial disease, type 1 diabetes. *Bioessays* 21, 164-174.
- Todd, J. A. (1995). Genetic analysis of type 1 diabetes using whole genome approaches. *Proc. Natl. Acad. Sci. USA* 92, 8560-8565.
- Todd, J. A. (1999). Multifactorial diseases: ancient gene polymorphism at quantitative trait loci and a legacy of survival during our evolution. In *The metabolic and molecular bases of inherited disease*, C. R. Scriver, ed. (New York: McGraw-Hill, Health Professions Division).
- Todd, J. A. (1990). The role of MHC class II genes in susceptibility to insulin-dependent diabetes mellitus. *Curr. Top. Microbiol. Immunol.* 164, 17-40.
- Todd, J. A., Aitman, T. J., Cornall, R. J., Ghosh, S., Hall, J. R., Hearne, C. M., Knight, A. M., Love, J. M., McAleer, M. A., Prins, J. B., Rodrigues, N., Lathrop, M., Pressey, A., Delarato, N. H., Peterson, L. B., and Wicker, L. S. (1991). Genetic analysis of autoimmune type 1 diabetes mellitus in mice. *Nature* 351, 542-547.
- Todd, J. A., Bell, J. I., and McDevitt, H. O. (1987). *HLA-DQB* gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. *Nature* 329, 599-604.
- Todd, J. A., Bell, J. I., and McDevitt, H. O. (1988). A molecular basis for genetic susceptibility to insulin-dependent diabetes mellitus. *Trends Genet.* 4, 129-134.

- Todd, J. A., and Farrall, M. (1996). Panning for gold: genome-wide scanning for linkage in type 1 diabetes. *Hum. Mol. Genet.* 5, 1443-1448.
- Todd, J. A., Mijovic, C., Fletcher, J., Jenkins, D., Bradwell, A. R., and Barnett, A. H. (1989). Identification of susceptibility loci for insulin-dependent diabetes mellitus by trans-racial gene mapping. *Nature* 338, 587-589.
- Tomfohrde, J., Silverman, A., Barnes, R., Fernandez-Vina, M. A., Young, M., Lory, D., Morris, L., Wuepper, K. D., Stastny, P., Menter, A., and Bowcock, A. (1994). Gene for familial psoriasis susceptibility mapped to the distal end of human chromosome 17q. *Science* 264, 1141-1145.
- Trask, B. J., Friedman, C., Martin-Gallardo, A., Rowen, L., Akinbami, C., Blankenship, J., Collins, C., Giorgi, D., Iadonato, S., Johnson, F., Kuo, W. L., Massa, H., Morrish, T., Naylor, S., Nguyen, O. T., Rouquier, S., Smith, T., Wong, D. J., Youngblom, J., and van den Engh, G. (1998). Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum. Mol. Genet.* 7, 13-26.
- Trepicchio, W. L., and Krontiris, T. G. (1993). IGH minisatellite suppression of USF-binding-site and E-mu-mediated transcriptional activation of the adenovirus major late promoter. *Nucleic Acids Res.* 21, 977-985.
- Trepicchio, W. L., and Krontiris, T. G. (1992). Members of the rel/NF- $\kappa$ B family of transcriptional regulatory proteins bind the *HRAS1* minisatellite DNA sequence. *Nucleic Acids Res.* 20, 2427-2434.
- Tuomilehto, J., Virtala, E., Karvonen, M., Lounamaa, R., Pitkaniemi, J., Reunanen, A., Tuomilehto-Wolf, E., and Toivanen, L. (1995). Increase in incidence of insulin-dependent diabetes mellitus among children in Finland. *Int. J. Epidemiol.* 24, 984-992.
- Tyler-Smith, C., and Brown, W. R. (1987). Structure of the major block of alphoid satellite DNA on the human Y chromosome. *J. Mol. Biol.* 195, 457-470.
- Ullrich, A., Dull, T. J., Gray, A., Brosius, J., and Sures, I. (1980). Genetic variation in the human insulin gene. *Science* 209, 612-615.
- Umar, A., Boyer, J. C., and Kunkel, T. A. (1994). DNA loop repair by human cell extracts. *Science* 266, 814-816.
- Undlien, D. E., Bennett, S. T., Todd, J. A., Akselsen, H. E., Ikaheimo, I., Reijonen, H., Knip, M., Thorsby, E., and Ronningen, K. S. (1995). Insulin gene region-encoded susceptibility to IDDM maps upstream of the insulin gene. *Diabetes* 44, 620-625.
- Undlien, D. E., Friede, T., Rammensee, H. G., Joner, G., Dahl-Jorgensen, K., Sovik, O., Akselsen, H. E., Knutsen, I., Ronningen, K. S., and Thorsby, E. (1997). *HLA*-encoded genetic predisposition in IDDM: *DR4* subtypes may be associated with different degrees of protection. *Diabetes* 46, 143-149.
- Urrutia, I., Calvo, B., Bilbao, J. R., and Castano, L. (1998). Anomalous behaviour of the 5' insulin gene polymorphism allele 814: lack of association with type 1 diabetes in Basques. *Diabetologia* 41, 1121-1123.
- Utz, U., Biddison, W. E., McFarland, H. F., McFarlin, D. E., Flerlage, M., and Martin, R. (1993). Skewed T-cell receptor repertoire in genetically identical twins correlates with multiple sclerosis. *Nature* 364, 243-247.
- Vafiadis, P., Bennett, S. T., Colle, E., Grabs, R., Goodyer, C. G., and Polychronakos, C. (1996). Imprinted and genotype-specific expression of genes at the *IDDM2* locus in pancreas and leucocytes. *J. Autoimmun.* 9, 397-403.
- Vafiadis, P., Bennett, S. T., Todd, J. A., Grabs, R., and Polychronakos, C. (1998a). Divergence between genetic determinants of *IGF2* transcription levels in leukocytes and of *IDDM2*-encoded susceptibility to type 1 diabetes. *J. Clin. Endocrinol. Metab.* 83, 2933-2939.

- Vafiadis, P., Bennett, S. T., Todd, J. A., Nadeau, J., Grabs, R., Goodyer, C. G., Wickramasinghe, S., Colle, E., and Polychronakos, C. (1997). Insulin expression in human thymus is modulated by *INS* VNTR alleles at the *IDDM2* locus. *Nat. Genet.* *15*, 289-292.
- Vafiadis, P., Grabs, R., Goodyer, C. G., Colle, E., and Polychronakos, C. (1998b). A functional analysis of the role of *IGF2* in *IDDM2*-encoded susceptibility to type 1 diabetes. *Diabetes* *47*, 831-836.
- Van Arsdell, S. W., Denison, R. A., Bernstein, L. B., Weiner, A. M., Manser, T., and Gesteland, R. F. (1981). Direct repeats flank three small nuclear RNA pseudogenes in the human genome. *Cell* *26*, 11-17.
- Van der Auwera, B. J., Vandewalle, C. L., Schuit, F. C., Winnock, F., De Leeuw, I. H., Van Imschoot, S., Lamberigts, G., and Gorus, F. K. (1997). *CTLA-4* gene polymorphism confers susceptibility to insulin-dependent diabetes mellitus (IDDM) independently from age and from other genetic or immune disease markers. *Clin. Exp. Immunol.* *110*, 98-103.
- Vergnaud, G. (1989). Polymers of random short oligonucleotides detect polymorphic loci in the human genome. *Nucleic Acids Res.* *17*, 7623-7630.
- Vergnaud, G., Mariat, D., Apiou, F., Aurias, A., Lathrop, M., and Lauthier, V. (1991). The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. *Genomics* *11*, 135-144.
- Virtaneva, K., D'Amato, E., Miao, J., Koskiniemi, M., Norio, R., Avanzini, G., Franceschetti, S., Michelucci, R., Tassinari, C. A., Omer, S., Pennacchio, L. A., Myers, R. M., Dieguez-Lucena, J. L., Krahe, R., de la Chapelle, A., and Lehesjoki, A. E. (1997). Unstable minisatellite expansion causing recessively inherited myoclonus epilepsy, EPM1. *Nat. Genet.* *15*, 393-396.
- Vogel, F., and Motulsky, A. G. (1997). *Human genetics: problems and applications*, Third Edition (Berlin: Springer Verlag).
- Vologodskii, A. V., Amirikyan, B. R., Lyubchenko, Y. L., and Frank-Kamenetskii, M. D. (1984). Allowance for heterogeneous stacking in the DNA helix-coil transition theory. *J. Biomol. Struct. Dyn.* *2*, 131-148.
- Vyse, T. J., and Todd, J. A. (1996). Genetic analysis of autoimmune disease. *Cell* *85*, 311-318.
- Warnecke, P. M., Mann, J. R., Frommer, M., and Clark, S. J. (1998). Bisulfite sequencing in preimplantation embryos: DNA methylation profile of the upstream region of the mouse imprinted H19 gene. *Genomics* *51*, 182-190.
- Warnecke, P. M., Stirzaker, C., Melki, J. R., Millar, D. S., Paul, C. L., and Clark, S. J. (1997). Detection and measurement of PCR bias in quantitative methylation analysis of bisulphite-treated DNA. *Nucleic Acids Res.* *25*, 4422-4426.
- Waterworth, D. M., Bennett, S. T., Gharani, N., McCarthy, M. I., Hague, S., Batty, S., Conway, G. S., White, D., Todd, J. A., Franks, S., and Williamson, R. (1997). Linkage and association of insulin gene VNTR regulatory polymorphism with polycystic ovary syndrome. *Lancet* *349*, 986-990.
- Weaver, J. U., Kopelman, P. G., and Hitman, G. A. (1992). Central obesity and hyperinsulinaemia in women are associated with polymorphism in the 5' flanking region of the human insulin gene. *Eur. J. Clin. Invest.* *22*, 265-270.
- Weber, J. L. (1990). Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. *Genomics* *7*, 524-530.
- Weber, J. L., and Wong, C. (1993). Mutation of human short tandem repeats. *Hum. Mol. Genet.* *2*, 1123-1128.
- Wegmann, D. R. (1996). The immune response to islets in experimental diabetes and insulin-dependent diabetes mellitus. *Curr. Opin. Immunol.* *8*, 860-864.

- Weisman-Shomer, P., and Fry, M. (1993). QUAD, a protein from hepatocyte chromatin that binds selectively to guanine-rich quadruplex DNA. *J. Biol. Chem.* 268, 3306-3312.
- Weitzmann, M. N., Woodford, K. J., and Usdin, K. (1997). DNA secondary structures and the evolution of hypervariable tandem arrays. *J. Biol. Chem.* 272, 9517-9523.
- Weitzmann, M. N., Woodford, K. J., and Usdin, K. (1998). The mouse *Ms6-hm* hypervariable microsatellite forms a hairpin and two unusual tetraplexes. *J. Biol. Chem.* 273, 30742-30749.
- Wheeler, V. C., Auerbach, W., White, J. K., Srinidhi, J., Auerbach, A., Ryan, A., Duyao, M. P., Vrbanc, V., Weaver, M., Gusella, J. F., Joyner, A. L., and MacDonald, M. E. (1999). Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. *Hum. Mol. Genet.* 8, 115-122.
- Wicker, L. S., Todd, J. A., and Peterson, L. B. (1995). Genetic control of autoimmune diabetes in the NOD mouse. *Annu. Rev. Immunol.* 13, 179-200.
- Wierdl, M., Dominska, M., and Petes, T. D. (1997). Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 146, 769-779.
- Wilkie, A. O., Higgs, D. R., Rack, K. A., Buckle, V. J., Spurr, N. K., Fischel-Ghodsian, N., Ceccherini, I., Brown, W. R., and Harris, P. C. (1991). Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* 64, 595-606.
- Willard, H. F. (1990). Centromeres of mammalian chromosomes. *Trends Genet.* 6, 410-416.
- Wolff, R. K., Plaetke, R., Jeffreys, A. J., and White, R. (1989). Unequal crossingover between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. *Genomics* 5, 382-384.
- Wolffe, A. P., and Matzke, M. A. (1999). Epigenetics: regulation through repression. *Science* 286, 481-486.
- Wolfram, D. J., and Wagener, H. P. (1938). Diabetes mellitus and simple optic atrophy among siblings: report of four cases. *Mayo. Clin. Proc.* 1, 715-718.
- Wong, Z., Wilson, V., Jeffreys, A. J., and Thein, S. L. (1986). Cloning a selected fragment from a human DNA 'fingerprint': isolation of an extremely polymorphic minisatellite. *Nucleic Acids Res.* 14, 4605-4616.
- Wong, Z., Wilson, V., Patel, I., Povey, S., and Jeffreys, A. J. (1987). Characterization of a panel of highly variable minisatellites cloned from human DNA. *Ann. Hum. Genet.* 51, 269-288.
- Wyman, A. R., and White, R. (1980). A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* 77, 6754-6758.
- Wyman, A. R., Wolfe, L. B., and Botstein, D. (1985). Propagation of some human DNA sequences in bacteriophage  $\lambda$  vectors requires mutant *Escherichia coli* hosts. *Proc. Nat. Acad. Sci. USA* 82, 2880-2884.
- Yauk, C. (1998). Monitoring for induced heritable mutations in natural populations: application of minisatellite DNA screening. *Mutat. Res.* 411, 1-10.
- Yauk, C. L., and Quinn, J. S. (1996). Multilocus DNA fingerprinting reveals high rate of heritable genetic mutation in herring gulls nesting in an industrialized urban site. *Proc. Natl. Acad. Sci. USA* 93, 12137-12141.
- Yerle, M., Lahbib-Mansais, Y., Pinton, P., Robic, A., Goureau, A., Milan, D., and Gellin, J. (1997). The cytogenetic map of the domestic pig. *Mamm. Genome* 8, 592-607.
- Yonekawa, H., Moriwaki, K., Gotoh, O., Miyashita, N., Matsushima, Y., Shi, L. M., Cho, W. S., Zhen, X. L., and Tagashira, Y. (1988). Hybrid origin of Japanese mice *Mus musculus molossinus*: evidence from restriction analysis of mitochondrial DNA. *Mol. Biol. Evol.* 5, 63-78.

Yoshino, M., Sagai, T., Lindahl, K. F., Toyoda, Y., Moriwaki, K., and Shiroishi, T. (1995). Allele-dependent recombination frequency: homology requirement in meiotic recombination at the hot spot in the mouse major histocompatibility complex. *Genomics* 27, 298-305.

Yu, S., Mangelsdorf, M., Hewett, D., Hobson, L., Baker, E., Eyre, H. J., Lapsys, N., Le Paslier, D., Doggett, N. A., Sutherland, G. R., and Richards, R. I. (1997). Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell* 88, 367-374.

Zhong, N., Yang, W., Dobkin, C., and Brown, W. T. (1995). Fragile X gene instability: anchoring AGGs and linked microsatellites. *Am. J. Hum. Genet.* 57, 351-361.

Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D. W., Amos, C., Dobyns, W. B., Subramony, S. H., Zoghbi, H. Y., and Lee, C. C. (1997). Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the  $\alpha$  1A-voltage-dependent calcium channel. *Nat. Genet.* 15, 62-69.