DNA Diversity and Meiotic Crossover Distribution in the Xp/Yp Pseudoautosomal Region

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

by Michael Timothy Slingsby
Department of Genetics
University of Leicester
November 2003

University *of* Leicester

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

Contents

Chapter 1 Introduction

Chapter 2 Materials and Methods

Chapter 3 Sequencing the *PGPL* Region

Chapter 4 Extending and Characterising the Known Sequence in the PAR1 Telomere–PGPL Interval

Chapter 5 Analysis of Genes in the N0434.1 Interval

Chapter 6 DNA Diversity in the N0434.1 Interval

Chapter 7 Linkage Disequilibrium in the N0434.1 Interval

Chapter 8 Sperm Crossover Analysis in the N0434.1 Interval

Chapter 9 Discussion

Abstract

High resolution analyses indicate that meiotic crossovers in human autosomes tend to cluster into 1–2 kb hotspots separated by blocks of high LD tens to hundreds of kilobases long. In contrast, low resolution data suggest only modest regional variation in recombination efficiency across the 2.6 Mb Xp/Yp pseudoautosomal region (PAR1), a male-specific recombination hot domain with a recombination rate about twenty times higher than the genome average.

Recent data suggest a more complex picture of PAR1 recombination. Around the *SHOX* gene, 500 kb from the telomere, LD decays extremely rapidly with physical distance, but nearly all crossovers cluster into a highly localised hotspot about 2 kb wide. In contrast, SNPs in a 1.5 kb region immediately adjacent to the PAR1 telomere are in intense LD, implying that this region is recombinationally inert and that male crossover activity terminates at a currently unidentified boundary in the distal region of PAR1.

To further investigate PAR1 recombination, the *PGPL* gene, 80 kb from the telomere, was targeted for analysis. This region had to be sequenced prior to SNP discovery and recombination analysis, revealing a novel gene that is potentially the most telomeric gene in PAR1.

SNP analysis of a 33 kb *PGPL* interval showed that this region is in free association with the telomere, suggesting recombinational activity in the intervening region, which this study proved to be rich in tandem repeats. Within the *PGPL* region, LD decays slowly with physical distance at a rate consistent with randomly-distributed crossovers occurring at close to the genome average rate. However, sperm crossover analysis revealed it to be the most recombinationally active region of DNA yet identified. Moreover, the novel distribution of crossovers in the region, suggests that there is not a unified set of hotspot-based rules that govern meiotic recombination in the human genome.

Acknowledgements

First and foremost I would like to thank Alec. It has been a genuine privilege to work with him and I will always be grateful for his excellent insight and for the time he (somehow!) always managed to make for me.

Thanks to ALL those who have made life in the lab so much fun, but particularly big hugs and kisses or manly handshakes to the following: Maria, whose existence I have been aware of for some time now…; John Yauk (what? WHAT? What's all this shouting? We'll have no trouble here…); Carole Stead, "I'm gonna live forever…;" Hilda (do you want any food with your salt?); Yuri (who's a scientist and nobody loves him, but I wash my hands, it's all b*******, so will you kindly please….?); Liisa, whose performance in The Loaded Dog will be talked about for years to come; Mark - thanks for all the terrible cocktails and, no matter what the others say, you are not a fat, gay, minger with a bony back; Kim -you've just nearly fallen off your chair as I write this, so I'll mention that here for posterity, Celia, thank you for all your help, you have been superb; Rita - we got there in the end! Foxy, tea and cake are two of life's important treasures – thank you for keeping me stocked up, Ila (who won't be going on holiday this year), Richard (a regular contender for geek of the week but an awful lot of fun too – how did you get in that dress?) and I dare not try and get away without mentioning Zoe – huge thanks to you for helping me remind everyone around Leicester that there is more than one kind of rugby! You have all been fantastic and I am lucky to know each one of you.

To all those who manage to get on the 5-a-side pitch, particularly Alistair (are you SURE it was just whisky?), Colin (who has to run around in a shower to get wet), Marcus (I didn't know so many four-letter words could be put to so many uses before I met you), Ben (sorry I didn't get you an Ah-Ha ticket) and Gianni (I assume that you won't be coming to training on Tuesday….).

Thanks to all of the Accies – Sunday football's finest bunch. Particularly thanks to Jit (is it your turn to buy lunch again?), and to Amin (come to the good side of the force!). You have both become really good mates and I thank you for all your generosity and friendship over the past four years.

To Mum and Dad. Thanks for all of the opportunities you have given me, and all of the help you are always willing to give – even though I am often too proud to take it. You are by far and away the best parents anyone could wish for. To Maff and Diana, I am lucky to be your older brother and very proud of all that you do. Thank you for always being there.

Most of all, thanks to Nina. I know that it hasn't been easy for you, and goodness knows how you did it, but you have been my rock Neen. It's difficult to express just how much your support and encouragement has meant to me, but I cannot imagine how I would have got through this without you. So, thank you from the bottom of my heart. This thesis is for you.

Abbreviations

ASO	allele-specific oligonucleotide
AS-PCR	allele-specific PCR
CEPH	Centre d'Etude du Polymorphisme Humain
dNTP	deoxynucleotide triphosphate
DSB	double-strand break
(d)HJ	(double) Holliday Junction
LD	linkage disequilibrium
LINE	long interspersed nuclear element
mAF	minor allele frequency
Mb, kb, bp	mega-, kilo- base pair
MHC	major histocompatibility complex
MVR-PCR	minisatellite variant repeat mapping by PCR
µg	micro-gram
µl, ml, l	micro-, milli- litre
µM, mM, M	micro-, milli- molar
PAR1	pseudoautosomal region at the ends of Xp/Yp
PAR2	pseudoautosomal region at the ends of Xq/Yq
PCR	polymerase chain reaction
PI-PLC	phosphatidylinositol-specific phospholipase C
RN	recombination nodule
SC	synaptonemal complex
SDSA	synthesis-dependent strand annealing model
SNP	single nucleotide polymorphism
ssDNA	single stranded DNA
STIR	subtelomeric interspersed repeat
TF	transcription factor
UTR	untranslated region

Chapter 1
Introduction

Ultimately an individual is genetically unique as a consequence of three processes that operate on germline DNA. The first of these stems from Mendel's law of independent assortment and reflects the fact that each chromosome of any pair is free to combine with any chromosome from each of the remaining pairs during meiosis. Hence, as the human genome consists of 23 pairs of chromosomes, their separate distribution alone ensures that there are 2^{23} possible chromosomal combinations for the formation of each gamete. The second process, mutation, creates new variations in DNA that can be passed on to future generations. Most mutations are neutral in that they have no obvious effect on the individual but occasionally mutations are beneficial and lead to a selectively favoured characteristic or detrimental and lead to inherited disease. The third process is meiotic recombination, which increases diversity by shuffling pre-existing variation between chromosome pairs from one generation to the next. Obviously these processes must be strictly controlled if they are not to cause any highly deleterious effects. However, the level of control must be balanced against allowing a species to adapt and proliferate through the manipulation of new variants by population processes such as selection, migration and genetic drift. Throughout evolutionary time recombination ensures that chromosomes are merely temporary associations of particular alleles and thereby provides the population processes with new assortments of variants to be tested by selection. In addition, recombination is essential in the short-term as it is required for the proper segregation of homologous chromosomes during meiosis. Hence, recombination is a process that is vital to both the well being of an individual organism and to the survival of a species.

1.1: Meiotic Recombination

Essentially there are two kinds of meiotic recombination events; crossovers, which involve the reciprocal exchange of information between homologous chromosomes, and gene conversions, which are strictly non-reciprocal. Regions of both high and low meiotic recombination have been observed in several organisms (Lichten and Goldman, 1995; Wahls, 1998), so it is clear that there is no simple linear relationship between genetic and physical map distances. Understanding the rules that govern the distribution of recombination events will rely on mapping crossovers and gene conversions along the chromosomes, the identification of the processes operating during recombination and an understanding of how the patterns and processes of recombination both influence sequence and haplotype diversity and ensure proper chromosome segregation during meiosis. This understanding will be of great value to those using genetic association analysis to identify susceptibility loci for complex disease (Jorde, 2000) and it will aid the study of aberrant recombination events such as unequal crossover, which are a major source of genome rearrangements, generating pathological variants in the human genome by exchanges in directly repeated genes (e.g. α-thalassaemia) (Higgs *et al.*, 1989) or between distal dispersed repeats (e.g. Charcot-Marie Tooth 1A, CMT1A) (Pentao *et al.*, 1992). It is further relevant to understanding the origins of human populations and the dynamics of human DNA evolution as the distribution of exchanges influences the probability of assembling new configurations of physically linked genes (Pääbo, 2003). Finally, an understanding of meiotic recombination could be relevant to studies of somatic mutation, particularly in cancer genetics (Feunteun, 1998) and it is likely that an understanding of areas of high and low recombination will be relevant to comprehending other DNA-related processes affected by chromosome context such as transcription and replication (Petes, 2001).

Currently, meiotic recombination is best understood in yeast as these lower eukaryotes allow the recovery and analysis of all of the products of individual meiotic recombination events. Consequently, the bulk of this first chapter will concentrate on what has been learned from the budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*). I will then discuss the types of analysis used to characterise human meiotic recombination and consider their relevance to this thesis.

1.1.1: The Prominent Features of Meiotic Recombination

A high level of recombination is usually a prominent feature of meiosis. Consequently, a review of meiotic recombination would not be complete without an overview of the early stages of the meiotic process itself. Meiosis is a specialised form of cell division that reduces chromosome number through a single round of DNA replication followed by two rounds of chromosome segregation (figure 1.1). It is through meiosis that haploid gametes are produced from diploid parental cells and it is during the lengthy prophase of meiosis I that double-strand breaks (DSBs) are formed to initiate recombination. The prophase of meiosis I can be divided into several stages based on chromosome morphology (table 1.1).

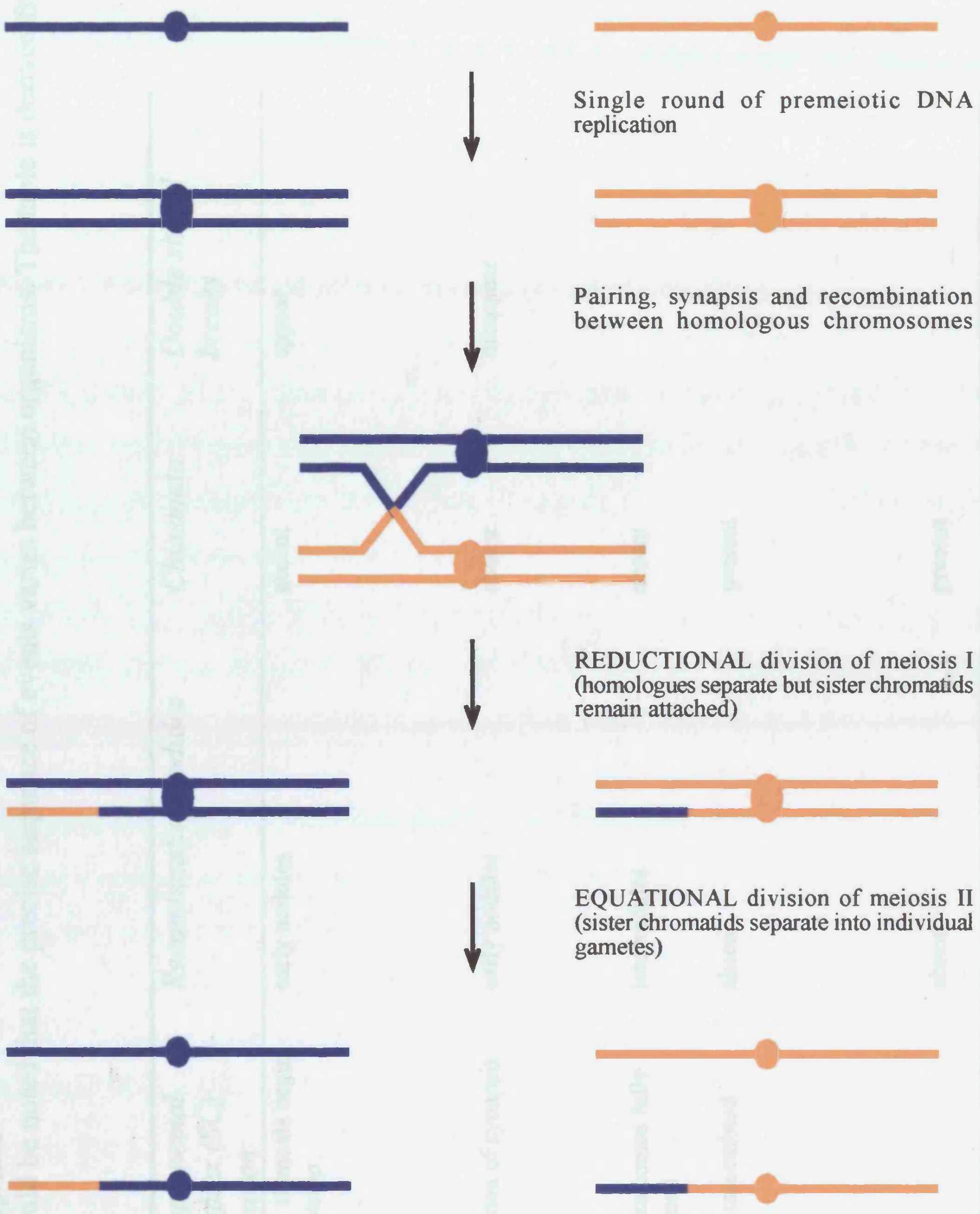

Single round of premeiotic DNA replication
Pairing, synapsis and recombination between homologous chromosomes
REDUCTIONAL division of meiosis I (homologues separate but sister chromatids remain attached)
EQUATIONAL division of meiosis II (sister chromatids separate into individual gametes)

Figure 1.1: Segregation of homologous chromosomes and sister chromatids during meiosis. The process is represented here by a single pair of homologous chromosomes, one of which is orange and the other blue. The simple exchange shown in this diagram has involved a reciprocal exchange between nonsister chromatids and so is a *crossover*.

Table 1.1. The five stages of meiosis I prophase
This is a generalisation of events and it should be noted that the precise sequence of events varies between organisms. The table is derived from Roeder (1997).

Stage of MI prophase	*Chromosome morphology*	*Synaptonemal Complex (SC) formation*	*Recombination nodules*	*Chiasmata*	*Double strand breaks*
Leptotene	Chromosomes are generally unpaired but telomeres begin to cluster. Sister chromatids are tightly bound	Axial elements begin to develop	early nodules	absent	appear
Zygotene	Homologous chromosomes are paired and telomeres are tightly clustered	Initiation of synapsis	early nodules	absent	disappear
Pachytene	Telomeres disperse	Chromosomes fully synapsed	late nodules	absent	
Diplotene	Chromosomes condense as homologues separate but remain held together by chiasmata	SC disassembled	absent	present	
Diakinesis	Further compaction of chromosomes	-	absent	present	

1.1.1.a: The Synaptonemal Complex and Cohesin

The prominent structural change during the prophase of meiosis I is the formation of the synaptonemal complex (SC) (for review see Roeder, 1997; Paques and Haber, 1999; Zickler and Kleckner, 1999 and references therein). It is strictly meiotic and mediates the intimate connection of homologous chromosomes along their lengths. Its ribbon-like tripartite structure is built sequentially between homologues following premeiotic replication (figure 1.2). Assembly begins during the leptotene stage with the organisation of the two sister chromatids of each chromosome along a common proteinaceous core called the axial element. Synapsis between homologous chromosomes begins at zygotene, during which the axial elements become the lateral elements of the SC and are connected by transverse filaments that extend between them. Upon completion of synapsis in pachytene there is a central element within the central region that lies parallel to and equidistant between the lateral elements. The central region contains very little DNA (Vazquez Nin *et al.*, 1993). The SC is dissociated during diplotene in order to allow diakinesis.

A different protein structure, the cohesin complex, is a major effector of sister chromatid cohesion in mitotic cells (Nasmyth *et al.*, 2000). Recently, cohesin was shown to be required for chromosome pairing and segregation and for recombination in yeast meiotic cells (Watanabe and Nurse, 1999; Klein *et al.*, 1999). Cohesion must be maintained in coordination with the deconstruction of the SC in order to allow the correct segregation of homologous chromosomes (Lee and Orr-Weaver, 2001).

1.1.1.a.i: Synaptonemal Complex and Cohesin Proteins

A number of components of the lateral SC elements have been identified. One of these is the mammalian Corl/SCP3, which is a primary determinant of axial element assembly and is required for the assembly of a second axial element protein, SCP2, which shapes the *in vivo* structure of the axial elements (Pelttari *et al.*, 2001). In *S. cerevisiae* the Redl and Hopl proteins are part of the lateral elements of the SC (Hollingsworth *et al.*, 1990; Smith and Roeder, 1997). Redlp localises discontinuously along the lateral elements and has been proposed to act as a 'catalyst' of lateral element formation - promoting the assembly of other proteins responsible for building the elements themselves (Smith and Roeder, 1997). For example, the product of the HOPl gene requires Redlp for its assembly onto chromosomes (Smith and Roeder, 1997). However, whether Hoplp then contributes to lateral element assembly or not is unclear, as experiments have shown that, although both Redlp and Hoplp are required for synapsis, Hoplp is not required for axial element formation (Hollingsworth

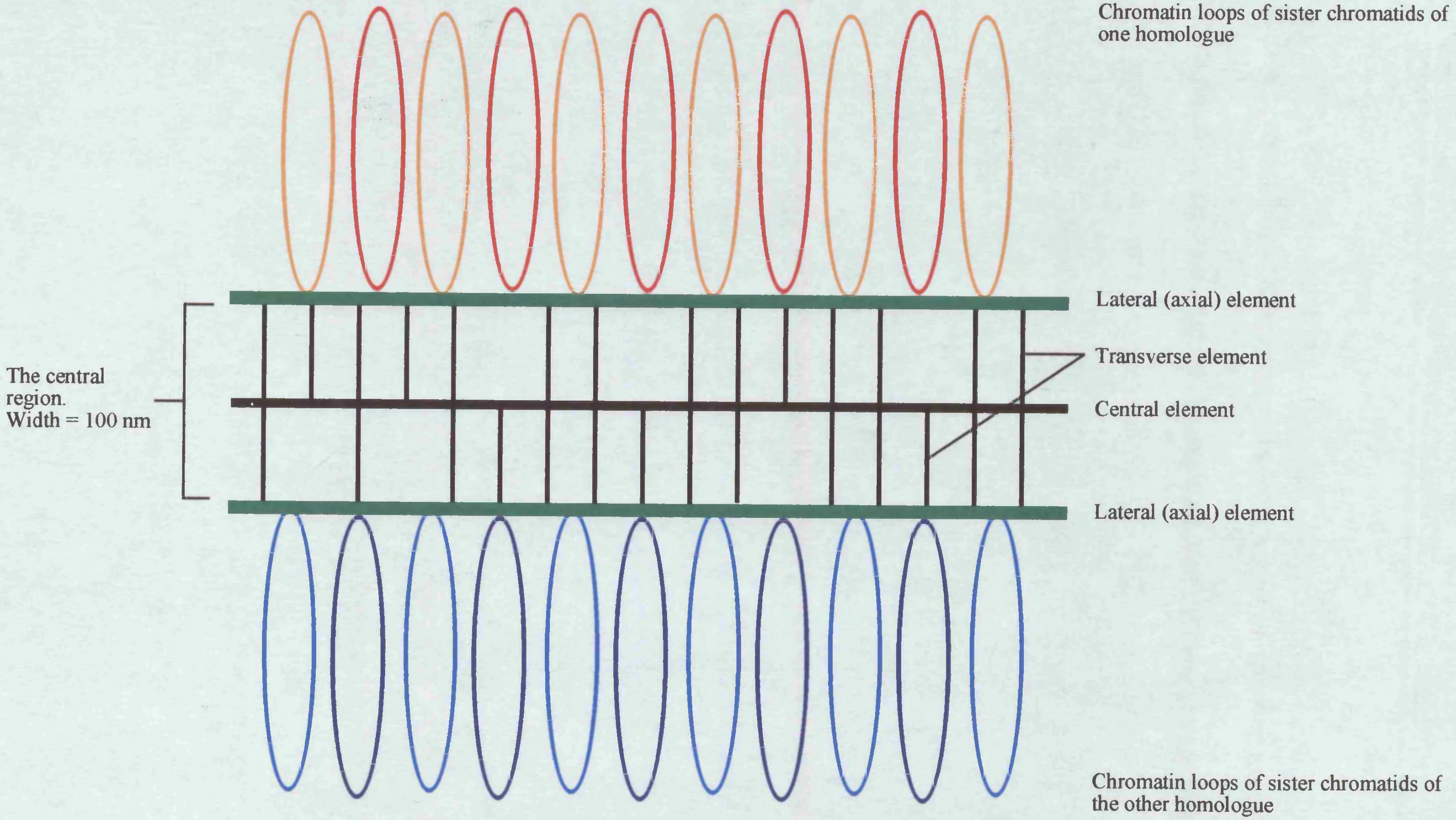

Chromatin loops of sister chromatids of one homologue
Lateral (axial) element
Transverse element
Central element
Lateral (axial) element
Chromatin loops of sister chromatids of the other homologue
The central region.
Width = 100 nm

Figure 1.2: The Synaptonemal Complex (SC)

and Byers, 1989; Roeder, 1997). In turn, the *S. cerevisiae* MEK1 gene encodes a kinase that is needed for Red1p localisation and hence correct SC assembly.

The *S. cerevisiae* protein Zip1 is a component of the central element and appears to stitch the lateral elements together along the whole central element (Sym *et al.*, 1993; Sym and Roeder, 1995). Zip2p may have a recruitment role in the central element similar to the lateral element function proposed for Red1p as it is required for the initiation of chromosome synapsis, has a punctate distribution along the SC and is required for normal Zip1p distribution (Chua and Roeder, 1998).

A number of cohesin proteins have also been identified. In *S. cerevisiae* these include Smc1, Smc3, Scc3/Irr1, Scc1/Mcd1/Rad21 (Guacci *et al.*, 1997; Michaelis *et al.*, 1997) and Rec8, which is meiosis-specific and acts in place of Rad21 (Klein *et al.*, 1999). These proteins have been highly conserved through evolution, which has enabled identification of their homologues in other organisms (Lee and Orr-Weaver, 2001).

1.1.1.a.ii: Synaptonemal Complex / Cohesin Interaction

There is considerable evidence that the SC and cohesin act together to ensure proper meiotic chromosome segregation. For example, *rec8* mutants of *S. cerevisiae* do not contain intact SCs and Red1p (part of the SC lateral elements, see above) cannot be detected (Klein *et al.*, 1999). In addition, work with mammalian cells has shown that cohesin proteins colocalise with the SC (Eijpe *et al.*, 2000). It has been suggested that the proper formation of axial elements depends on cohesin but that the organisation of cohesin complexes in mammalian meiotic cells is not affected by the SC (Lee and Orr-Weaver, 2000; Pelttari *et al.*, 2001). However, it has been suggested that a mature SC is required for proper sister chromatid cohesion in maize (Maguire *et al.*, 1991). This further suggests that the same meiotic structures and processes have variable importance in different organisms.

1.1.1.b: Chiasmata

Before detailed molecular analysis was possible, recombination was detected in many species by studies of chiasmata, which are the cytologically visible structures of meiosis I that correspond to a site of crossover between two non-sister chromatids. Chiasmata have been observed in many higher eukaryotes (e.g. Hultén, 1974; Herickhoff *et al.*, 1993, Latos-Bielenska and Vogel, 1990, Morton *et al.*, 1982; Nilsson and Pelger, 1991) and are clearly important structures; failure of homologous chromosomes to undergo at least one crossover results in elevated rates of nondisjunction and unviable gametes. As a result, each chromosome pair in meiotic cells generally contains at least one obligate chiasma (Lawrie *et al.*, 1995). Other than covalent bonds between the non-sister chromatids there is nothing particularly stable about the chiasma structure. Most evidence suggests that distal sister chromatid cohesion is the additional factor that locks chiasmata into place. This can be extrapolated to predict that terminal chiasmata will be relatively unstable. Therefore, studies that have shown that crossovers near the ends of chromosomes do not ensure proper meiosis I disjunction as effectively as exchanges close to the centromere provide very strong support for this model (Lamb *et al.*, 1996; Ross *et al.*, 1996). However, in male meiosis of both mice and humans, chiasmata show strong telomeric or subtelomeric localisation (e.g. Laurie and Hultén, 1985; reviewed in Lichten and Goldman, 1995; below, section 1.2.2). Furthermore, chiasmata are *obliged* to develop in the telomere-adjacent pseudoautosomal regions of the X and Y chromosomes during male meiosis (Burgoyne, 1982), suggesting that there is at least one other system of chiasma stabilisation.

Assuming that cohesion is required for chiasma stability, chromosomes that have undergone a reciprocal exchange would not be expected to disjoin properly if they were subjected to mutations that destabilise cohesin. In *S. cerevisiae* mutations to the RED1 and MEK1 genes, which are important for SC assembly and strongly suggested to be required for cohesion, result in premature separation of sister chromatids and the consequent random segregation of chromosomes at both meiotic divisions (Rockmill and Roeder, 1988, Bailis and Roeder, 1998). The same result is seen when maize, which needs a mature SC for correct sister chromatid cohesion, is subjected to mutations in which synapsis fails (Maguire *et al.*, 1991). The strongest evidence that cohesion is vital to chiasma stability is derived from experiments in which the ORD gene of *Drosophila* was mutated; *ord* mutations lead to the visible separation of sister chromatids during prophase I and result in nondisjunction of exchange chromosomes (Miyazaki and Orr-Weaver, 1992; Bickel *et al.*, 1997). The overall suggestion is that a stable chiasma, and hence crossover, is the feature of recombination required for proper chromosome segregation at meiosis I. The stability of chiasmata is reliant upon correct sister chromatid cohesion, which in turn is dependent on the SC and cohesin.

1.1.1.c: Recombination Nodules

Recombination nodules (RN) are multicomponent structures that are found in association with SCs from leptotene through to pachytene (Carpenter, 1975; for review see Zickler and Kleckner, 1999). They can be visualised through electron microscopy and have been identified in a number of organisms. Two classes of RN have been defined, based on their time of appearance, size, shape, relative number and distribution.

1.1.1.c.i: Early Nodules

Early nodules are present during leptotene or zygotene and it has been suggested that they have a significant role in deciding where recombination is or is not to occur (Anderson *et al*, 2001). Support for the hypothesis that early nodules are initiators of recombination is derived from the discovery that, in the lily, early nodules contain the Dmc1 and Rad51 proteins (Anderson *et al.*, 1997). Both are homologues of RecA, a protein that promotes strand exchange and is required for nearly all homologous recombination events in *E. coli* (Kowalczykowski, 2000). However, a study of early nodule distribution in six different plant species has suggested a more complex model for early nodule activity; Anderson *et al.* (2001) proposed that early nodule distribution was random in each species tested but that a minority (10–40%) of early nodules attached to axial elements prior to synapsis, whilst the majority (60–90%) attached at synaptic forks (the interface between the formation of the tripartite SC, as seen in figure 1.2, and two separated axial elements). In turn this suggests one of two possibilities. First, early nodules can be classified into two types, those that initially associate with axial elements and function in the initiation of synapsis and those that assemble at synaptic forks and might function in recombination. Alternatively early nodules may mature over time such that the youngest nodules are found at synaptic forks and the more mature nodules lie closer to the synaptic initiation sites.

1.1.1.c.ii: Late Nodules

Certain lines of evidence suggest that late nodules are derived from a subset of early nodules (Sherman *et al.*, 1992; Plug *et al.*, 1998; Zickler and Kleckner, 1999; Agarwal and Roeder, 2000). Late nodules are present during pachytene, are structurally distinct and 2 to 20 times less numerous than early nodules (Zickler and Kleckner 1999). The distribution pattern of late nodules correlates very strongly with the distribution and number of chiasmata, which has led to the proposal that late nodules are multi-enzyme complexes that resolve recombination intermediates as crossovers (Carpenter, 1987, Zickler *et al.*, 1992; Zickler and Kleckner, 1999; Allers and Lichten, 2001). Components of late nodules in *S. cerevisiae* include two homologues of the *E. coli* MutS mismatch repair protein, Msh4 and Msh5, and Mlh1, a homologue of MutL, a second *E. coli* mismatch repair protein (Ross-Macdonald and Roeder 1994; Hollingsworth *et al.*, 1995; Hunter and Borts, 1997).

1.1.1.d: Premeiotic Pairing and Homologue Recognition

How homologous chromosomes find each other prior to recombination is still not very well understood. In many organisms there is evidence to suggest that homologue pairing (the detectable coming-together of homologues, as opposed to synapsis, which is the intimate association of chromosomes) in nonmeiotic cells contributes to pairing in meiosis. For example, observations in yeast cells suggest that nuclear architecture contributes to the apparent premeiotic pairing of homologous chromosomes; Scherthan *et al.* (1994) demonstrated that different homologue pairs occupy separate nuclear territories of *Schizosaccharomyces pombe* (*S. pombe*) diploid cells and that centromeric regions show a high level of pairing. In vegetative cells of *S. cerevisiae* centromeres are tightly clustered together and telomeres are distributed elsewhere in the nucleus (J. Loidl, cited in Roeder, 1997). The studies of Weiner and Kleckner (1994) indicated that homologues are paired via multiple interstitial interactions before meiosis, which must limit searches for homology once a DSB has formed and so prevent an excessive number of ectopic recombination events.

In many organisms, such as mice, humans and maize, chromosomes are not paired prior to meiosis (Scherthan *et al.*, 1996; Bass *et al.*, 1997), which indicates that there must be mechanisms for homologue searching in the very early stages of meiosis. The products of RAD17, RAD24 and MEC1 are essential for the maintenance of cell cycle arrest in *dmc1* mutants (see 'the recombination checkpoint,' below) (Lydall *et al.*, 1996) but also appear to function in the correct choice of recombination partner, as their mutants result in a redirection of interhomologue events into pathways that favour ectopic recombination (Grushcow *et al.*, 1999). Perhaps this is because *rad17, rad24* and *mec1* strains fail to check that recombination is complete and allow the premeiotic separation of homologous chromosomes, thereby increasing the probability of intersister and ectopic events. Alternatively, fully functional RAD17, RAD24 and MEC1 could ensure recombination by maintaining the SC, which, through linking homologues together, promotes allelic over ectopic partner choice. (Goldman and Lichten, cited as unpublished results in Grushcow *et al.*, 1999). This latter mechanism is unlikely as RAD17, RAD24 and MEC1 are also required to arrest mitotic progression, and so they are likely to have an indirect role common to both meiosis and mitosis, rather than tethering recombination events to the meiosis-specific SC.

Interhomologue interactions must restrict ectopic recombination events to some degree during *S. cerevisiae* meiosis, as allelic recombination is reduced by at least 100-fold between homoeologous chromosomes in hybrid *S. cerevisiae–carlsbergensis* strains, whilst ectopic recombination is increased. Experiments demonstrate that this is due to the diverged chromosomes not undergoing end-to-end alignment during MI prophase (Goldman and Lichten, 2000). Therefore, in normal *S. cerevisiae* meiosis, homologous pairing must serve to direct repair towards allelic sequences and thus limit a potential genome-wide homology search that could otherwise lead to deleterious rearrangements (Goldman and Lichten, 2000).

As a loss of cohesion leads to defects in recombination (Lee and Orr-Weaver, 2001) models have been proposed in which the drive for meiotic repair between homologous chromosomes is mediated by a link between cohesin and recombination proteins (Pelttari *et al.*, 2001; van Heemst and Heyting, 2000), and not by the SC as suggested by Goldman and Lichten.

1.1.2: Models for the Mechanisms of Recombination

It is clear from the data presented above that meiotic recombination is not a simple, linear process and that continual interaction between components and systems within the meiotic recombination machinery is vital for the initiation of recombination, the physical interaction of non-sister chromatids and the conclusion of the process via the correct disjunction of chromosomes. A number of models have been proposed to explain how DNA recombination occurs and current thinking favours the two that are discussed at length below.

1.1.2.a: The Double Strand Break Repair Model

The event initiating most, if not all, meiotic recombination events in *S. cerevisiae* has been unambiguously identified as a double strand break (DSB) (Sun *et al.*, 1989). Paradoxically DSBs are also DNA lesions that, unless repaired, will trigger a cell's DNA-damage response systems to arrest the cell cycle or to induce apoptosis. Observations in mitotic cells led to the proposal of the DSB repair model of meiotic recombination (figure 1.3) (Szostak *et al.*, 1983; Sun *et al.*, 1991), in which double-strand cleavage is followed by the 5′ to 3′ resection of the ends to generate overhanging 3′ single-stranded termini. The single-stranded DNA tails then invade an intact homologous duplex, and prime repair synthesis. The accompanying branch migration results in long stretches of heteroduplex DNA and two Holliday junctions. Resolving the Holliday junctions in the same or opposite direction will then result in gene conversion or reciprocal crossover (between markers that flank the region of strand exchange) respectively. If mismatches contained within heteroduplex DNA are not repaired during meiosis, postmeiotic segregation (PMS) of gene conversions will be observed.

1.1.1.a.i: Initiation

In *S. cerevisiae*, premeiotic DNA replication is necessary for the formation of DSBs (Borde *et al.*, 2000; Smith *et al.*, 2001). It has been suggested that sites for DSB formation are marked by the addition of cohesin proteins, such as Rec8, as the replication fork passes through each potential DSB site (Mizuno *et al.*, 2001; Watanabe *et al.*, 2001). Various groups have demonstrated that at least 11 genes are then essential for the DSB formation step. These are

Figure 1.3: The double-strand break repair model of meiotic recombination (Szostak *et al.*, 1983; Sun *et al.*, 1991).

Shown are two double-stranded non-sister chromatids (one blue and one orange). The other two chromatids (sisters to the pair shown here) are not included in the diagram. Gene products are indicated at specific steps where they have been demonstrated to be important for proper meiotic recombination (see text).

(a) Recombination is initiated by a DSB in one duplex. (b) Each side of the break is subjected to extensive 5′ to 3′ degradation, which leaves overhanging 3′ single-stranded ends. (c) One 3′ end invades the intact homologous duplex and displaces one of the strands to form a D-loop and heteroduplex DNA. (d) The 3′ end of the invading strand then acts as a primer for the initiation of new DNA synthesis, which enlarges the D-loop and leads to annealing between the D-loop and the homologous sequences of the second 3′ single-stranded end. Repair synthesis from the second 3′ end then takes place. (e) Repair synthesis and branch migration results in two Holliday junctions. (f) Independent resolution of the Holliday junctions by cutting either inner (open triangles) or outer (filled triangles) strands then leads to noncrossover or (g) crossover configurations. In the diagram the crossover molecule was produced by resolution of the left Holliday junction inner strands and the right Holliday junction outer strands.

Note that the site at which recombination initiates via DSB formation is the *recipient* of genetic information.

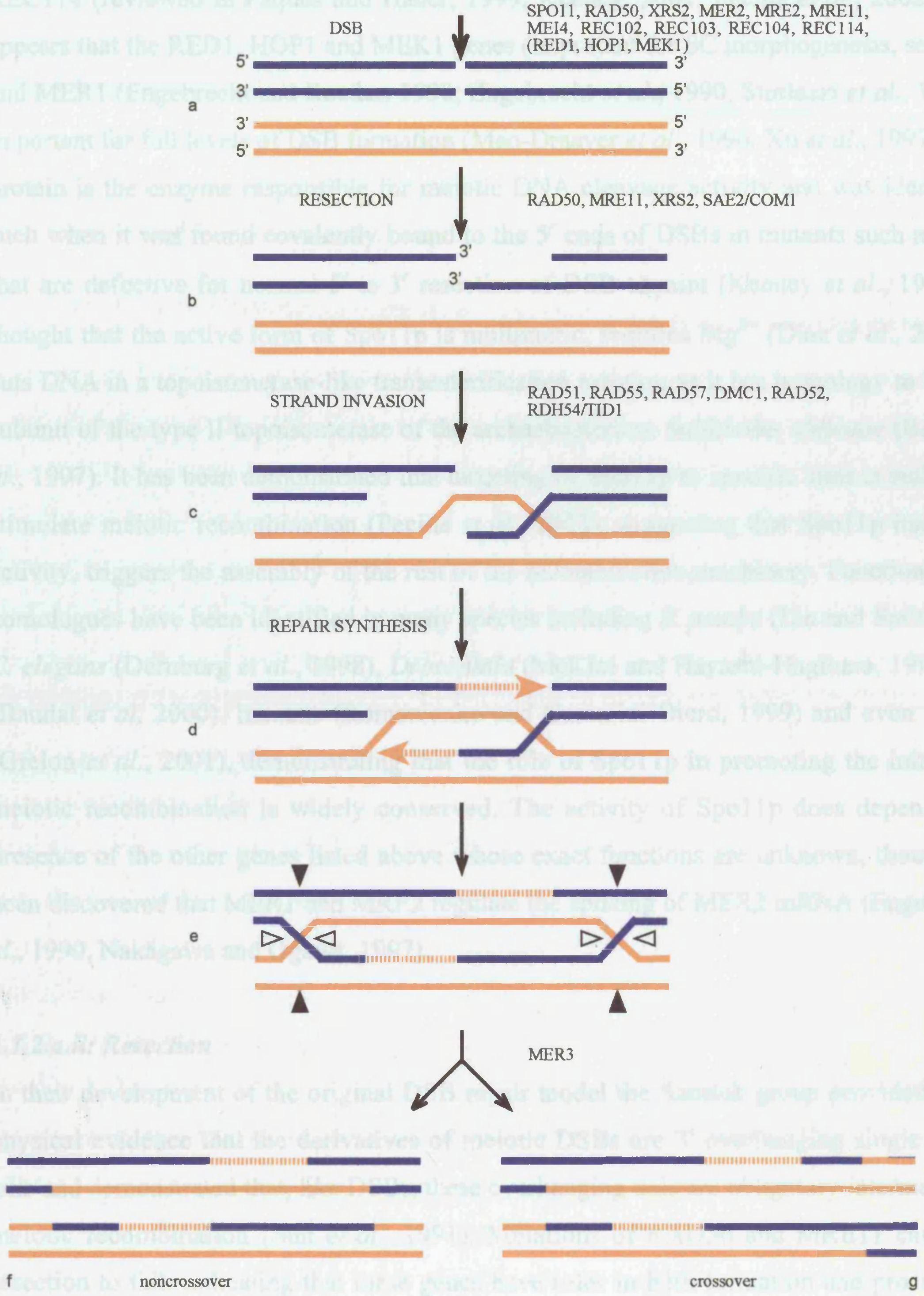

DSB
SPO11, RAD50, XRS2, MER2, MRE2, MRE11, MEI4, REC102, REC103, REC104, REC114, (RED1, HOP1, MEK1)
5'
3'
3'
5'
a
3'
5'
5'
3'
RESECTION
RAD50, MRE11, XRS2, SAE2/COM1
3'
3'
b
STRAND INVASION
RAD51, RAD55, RAD57, DMC1, RAD52, RDH54/TID1
c
REPAIR SYNTHESIS
d
e
MER3
f
noncrossover
crossover
g

Figure 1.3: The double-strand break repair model of meiotic recombination.

SPO11, RAD50, XRS2, MER2, MRE2, MRE11, MEI4, REC102, REC103, REC104 and REC114 (reviewed in Paques and Haber, 1999; Keeney, 2001; Peciña *et al.*, 2002). It also appears that the RED1, HOP1 and MEK1 genes (important for SC morphogenesis, see above) and MER1 (Engebrecht and Roeder, 1990; Engebrecht *et al.*, 1990, Storlazzi *et al.*, 1995) are important for full levels of DSB formation (Mao-Draayer *et al.*, 1996; Xu *et al.*, 1997). Spo11 protein is the enzyme responsible for meiotic DNA cleavage activity and was identified as such when it was found covalently bound to the 5′ ends of DSBs in mutants such as *rad50S* that are defective for normal 5′ to 3′ resection of DSB termini (Keeney *et al.*, 1997). It is thought that the active form of Spo11p is multimeric, requires Mg^{2+} (Diaz *et al.*, 2002) and cuts DNA in a topoisomerase-like transesterification reaction as it has homology to the small subunit of the type II topoisomerase of the archaebacterium *Sulfobolus shibatae* (Bergerat *et al.*, 1997). It has been demonstrated that targeting of Spo11p to specific sites is sufficient to stimulate meiotic recombination (Pecina *et al.*, 2002), suggesting that Spo11p itself, or its activity, triggers the assembly of the rest of the recombination machinery. Functional Spo11 homologues have been identified in many species including *S. pombe* (Lin and Smith, 1994), *C. elegans* (Dernburg *et al.*, 1998), *Drosophila* (McKim and Hayashi-Hagihara, 1998), mice (Baudat *et al*, 2000), humans (Romanienko and Camerini-Otero, 1999) and even in plants (Grelon *et al.*, 2001), demonstrating that the role of Spo11p in promoting the initiation of meiotic recombination is widely conserved. The activity of Spo11p does depend on the presence of the other genes listed above whose exact functions are unknown, though it has been discovered that MER1 and MRE2 regulate the splicing of MER2 mRNA (Engebrecht *et al.*, 1990, Nakagawa and Ogawa, 1997).

1.1.2.a.ii: Resection

In their development of the original DSB repair model the Szostak group provided the first physical evidence that the derivatives of meiotic DSBs are 3' overhanging single stranded tails and demonstrated that, like DSBs, these overhanging tails are obligatory intermediates in meiotic recombination (Sun *et al.*, 1991). Mutations of RAD50 and MRE11 cause DSB resection to fail, indicating that these genes have roles in both formation and processing of DSBs (Alani *et al.*, 1992; Nairz and Klein, 1997). Like *rad50S*, a mutation of MRE11 causes Spo11p to remain associated with DSB ends (Tsubouchi and Ogawa, 1998). It is not then unreasonable to assume that Rad50p and Mre11p remove Spo11p from DSB ends in order to stimulate resection. XRS2 also has a probable role in resection (Paques and Haber, 1999), as does the meiosis-specific SAE2/COM1 (McKee and Kleckner, 1997; Prinz *et al.*, 1997).

1.1.2.a.iii Single strand invasion

In a detailed analysis of the transition from DSBs to double Holliday junctions (dHJs), Hunter and Kleckner (2001) identified single-end invasions (SEIs) as the primary products of strand exchange between one DSB end and its homologue. Consequently, SEIs have an asymmetric structure, which supports the idea that the two 3' ends of a resected DSB interact with a homologous partner via temporally distinct mechanisms; i.e. the first interaction (SEI formation) is a strand invasion step and, whilst the second end may also interact via strand invasion, it is probable that enlargement of the initial D-loop, caused by invasion of the first end, permits interaction of the second end via annealing (figure 1.3). RAD51, RAD55, RAD57 and DMC1 are required for the strand invasion step in *S. cerevisiae* (Bishop *et al.*, 1992; Shinohara *et al*, 1992; Schwacha and Kleckner, 1997). All are homologues of the bacterial RecA strand exchange enzyme but DMC1 is the only gene to encode a meiosis-specific product. It is thus not inconceivable that other genes involved in the mitotic repair of DSBs also function in the DSB repair method of meiotic recombination. These genes include RAD52 and RAD54. RAD51 is conserved from yeast to humans with very high sequence similarity (Shinohara and Ogawa, 1999), providing clear evidence that its role in meiotic recombination is also widely conserved. The clear importance of Dmc1p in meiotic strand exchange is demonstrated by *dmc1* mutants, which arrest in prophase with hyper-resected, unrepaired DSBs (Bishop *et al.*, 1992).

Biochemical evidence suggests that a Rad55/Rad57 heterodimer, in the presence of RPA (the yeast single stranded DNA (ssDNA) binding protein) stimulates the Rad51 enzyme to promote ATP-dependent strand invasion (Sung, 1994; Sung, 1997). The situation *in vivo* is likely to be more complex; if RPA is pre-incubated with ssDNA, it prevents the binding of Rad51p to DNA and so the *in vitro* strand exchange reaction is inhibited. This effect can be overcome by the addition of Rad52p (Benson *et al.*, 1998; New *et al.*, 1998; Shinohara and Ogawa, 1998), which interacts with a subunit of RPA (Shinohara *et al.*, 1998). This suggests that RPA and Rad51p are ordinarily in competition to bind with the overhanging ssDNA and that Rad52p limits the binding action of RPA, thereby allowing Rad51p to promote strand invasion. However, if RPA is added after Rad51p has bound DNA, it stimulates the strand exchange reaction. Therefore RPA is believed to remove secondary structures of ssDNA and thereby allows Rad51p and DNA to form a nucleoprotein filament, which is an essential intermediate of the strand exchange step (Shinohara and Ogawa, 1999). The inhibitory action of RPA is also alleviated by the Rad55/Rad57 heterodimer (Sung and Robberson, 1995; Sugiyama *et al.*, 1997). Rad52p is required for nearly all recombination events but Rad55p and Rad57p are sometimes dispensable (Paques and Haber, 1999). Therefore, Rad52p is likely to be the catalytic subunit of the strand exchange reaction in *S. cerevisiae* and the role of the Rad55/Rad57 heterodimer may not be to stimulate the activity of Rad51p but rather to direct it to the ssDNA tails in a manner similar to that of the RecO and RecR proteins, which direct RecAp to ssDNA in bacteria (Umezu and Kolodner, 1994).

1.1.2.a.iv: Meiotic recombination partner choice

Assuming that the homologous pairing discussed above greatly reduces the probability of ectopic interactions, at the point of strand exchange there has to be a decision with regard to using a sister or non-sister chromatid as the template for repair of the DSB. In mitotic cells most DSBs are likely to occur during replication, which means that the most conservative choice for a repair template is the sister chromatid (Kadyk and Hartwell, 1992). During meiosis, if repair of DSBs was to occur between sister chromatids there would be no interhomologue crossing over, a feature of the process that is essential for proper chromosome segregation. Therefore, meiotic cells direct repair to ensure that most recombination events occur between allelic sequences of non-sister chromatids (Schwacha and Kleckner, 1997). There is considerable evidence that the meiosis-specific Dmc1p is heavily involved in this step, as are Rad54p and its homologue Tid1p/Rdh54p. The indications are that Dmc1p and Tid1p act together to direct repair towards non-sisters and that Rad51p and Rad54p normally mediate exchange between sisters (Haber, 2000). During meiosis this latter type of exchange is presumably further suppressed by SC formation, as it has been shown that *red1*, *mek1* and *hop1* mutations that prevent normal formation of axial elements all increase sister chromatid recombination (Thompson and Stahl, 1999). As SEIs occur at the same time as full length SCs (Hunter and Kleckner, 2001), it is possible that events leading to SC formation concurrently direct DSBs to enter a DMC1-dependent interhomologue-only recombination pathway. In support of this, mutation of RED1 is actually thought to eliminate a highly specific interhomologue-only recombination pathway (Schwacha and Kleckner 1997; Xu *et al.*, 1997).

1.1.2.a.v: Double Holliday Junctions

The DSB repair model predicts the formation of an intermediate containing two Holliday junctions (HJs) (figure 1.3). Schwacha and Kleckner (1994) showed that branched DNA molecules detected after DSB formation are indeed dHJs that can be resolved *in vitro*, by the *E. coli* HJ cleaving enzyme RuvC, into noncrossover and crossover products (Schwacha and Kleckner, 1995). The ratio of these products is approximately 1:1, as is seen is wild-type strains of *S. cerevisiae*, and as would be predicted from random resolution of dHJs.

1.1.2.b: The Synthesis-Dependent Strand Annealing Model

All of the evidence reviewed above, and particularly the demonstration of dHJs, appears to provide unequivocal support for the DSB repair model. However, the DSB repair model also predicts the formation of two regions of heteroduplex DNA, one on each side of the initiating DSB and each on different chromatids (figure 1.3). In the studies of Gilbertson and Stahl (1996) and of Porter *et al.* (1993) these expectations were not met; most events were one-sided and, when two heteroduplex regions were detected, both were on the same chromatid. These results support a second type of recombination model; that of synthesis-dependent strand annealing (SDSA, figure 1.4) (reviewed in Paques and Haber, 1999). Inevitably the initiating event is once again the formation of DSBs and extensive 5′ to 3′ degradation takes place to leave 3′ overhanging single-stranded ends. The strand invasion and repair synthesis then takes place on only one side of the DSB. The newly synthesised DNA strand is displaced and extensive heteroduplex DNA is formed only at the initiating (recipient) locus by the annealing of the new strand to the other 3′ overhanging end. Heteroduplex DNA arises only at the time of resolution and only noncrossover products are formed, though Paques *et al.* (1998) do suggest a modification of the SDSA model containing a dHJ that can be resolved with or without crossover (figure 1.4). However, only one experiment, in which crossovers seemed to account for about 5% of DSB repair events, has suggested that SDSA associated with crossing over occurs in *S. cerevisiae* (Paques *et al.*, 1998).

1.1.2.c: Crossover and and gene conversion during meiotic recombination

In all current models of meiotic recombination, and particularly the DSB repair model, there is an assumption that conversion and crossover products arise at the same time as a result of alternative resolution of a recombination intermediate. However, more recent experiments have indicated that there are distinct pathways of crossover and noncrossover meiotic recombination. In *S. cerevisiae* there are a number of genes that are required for normal frequencies of crossover that are not required for gene conversion. These include ZIP1 and ZIP2 (indicating a possible role for the SC in this choice), ZIP3, MSH4, MSH5, MLH1, MLH3 and MER3 (see Nakagawa and Kolodner, 2002). Mer3p is a meiosis-specific DNA helicase (Nakagawa and Kolodner, 2002). As MER3 is not required for gene conversion it is conceivable that its helicase activity is important for branch migration and/or the resolution of HJs (figure 1.3) once a decision between a noncrossover or crossover pathway has been made. It is worth noting, however, that *mer3* mutants accumulate DSBs, indicating that Mer3p has an earlier function in meiosis involving the transition of DSBs to later intermediates (Nakagawa and Kolodner, 2002). As gene conversion remains unaffected in *mer3* mutants, this suggests that the discrimination between a pathway of crossover and one

Figure 1.4: The synthesis-dependent strand annealing model of meiotic recombination (reviewed in Paques and Haber, 1999).

The basic feature of the SDSA model is that newly synthesised strands are displaced from the template and returned to the molecule that suffered the DSB. Furthermore, in contrast to the DSB repair model, heteroduplex DNA is found only at the recipient locus and the donor template remains unchanged.

(a) Recombination is initiated by a DSB in one duplex. (b) Each side of the break is subjected to extensive 5′ to 3′ degradation, which leaves overhanging 3′ single-stranded ends. (c) One 3′ end invades the intact homologous duplex, initiates DNA synthesis and forms heteroduplex DNA. (d) The newly synthesised strand is displaced and (e) anneals with the other DSB end. Repair of the break is completed by DNA synthesis and ligation. Note that in this model only noncrossover products are formed. However, as suggested by Paques *et al.* (1998), invasion of the template by the second 3′ end (f) may sometimes stabilise the strand displaced by the first 3′ end. DNA synthesis would become semiconservative (g) and two Holliday junctions would be formed (h), and subsequently cut, as shown in figure 1.3.

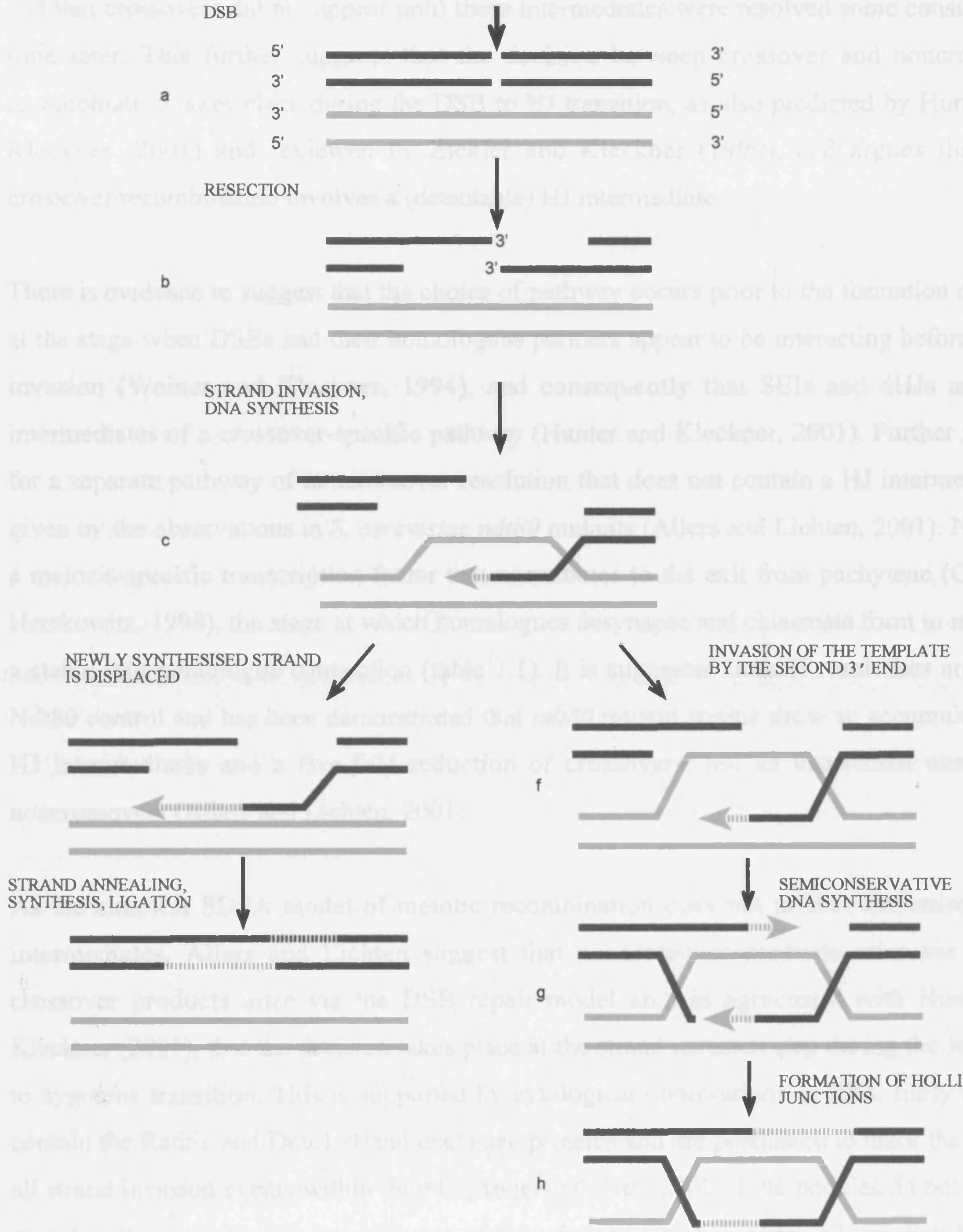


Figure 1.4: The synthesis-dependent strand annealing (SDSA) model of meiotic recombination

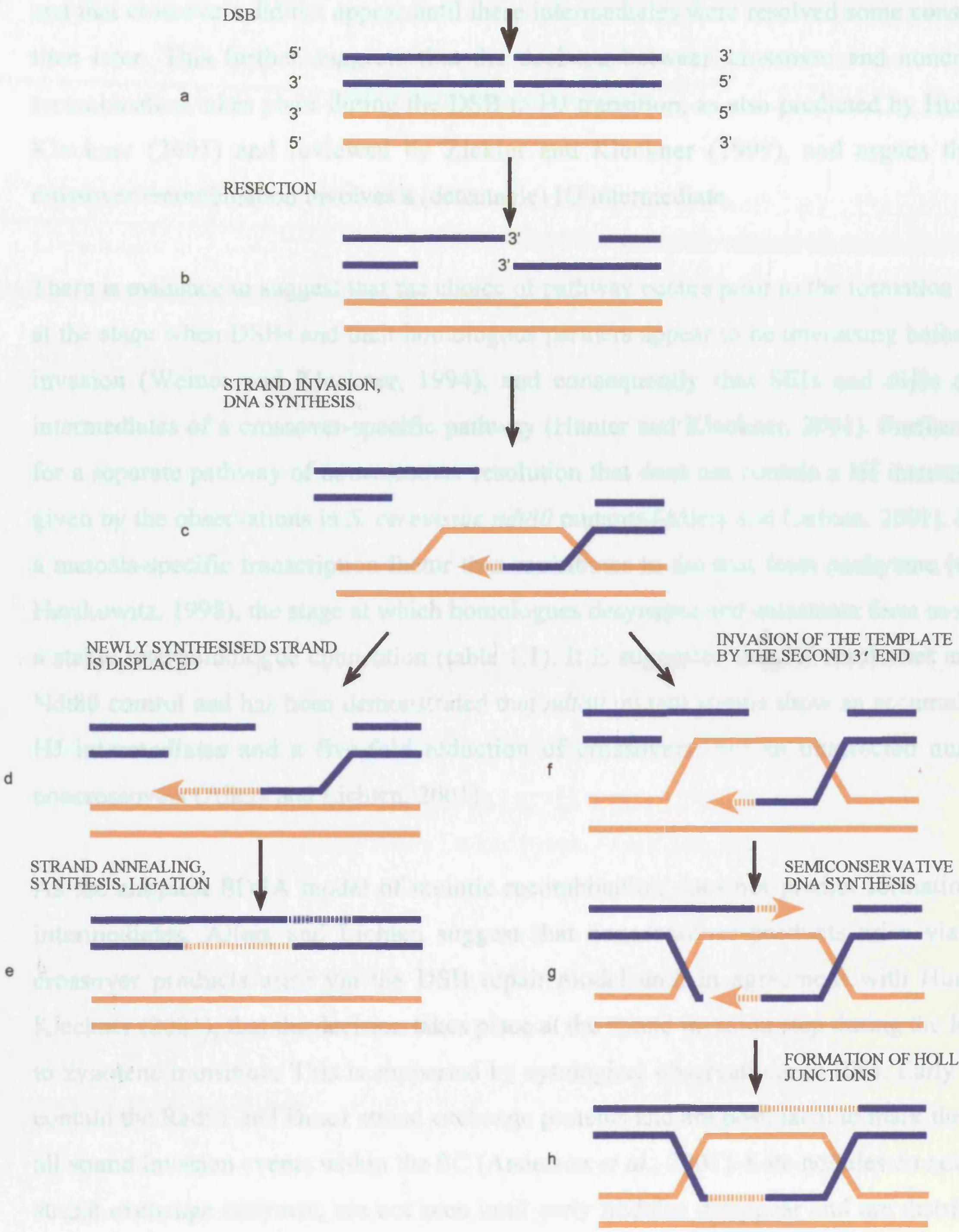


Figure 1.4: The synthesis-dependent strand annealing (SDSA) model of meiotic recombination

of conversion takes place at a relatively early stage of recombination. Allers and Lichten (2001) demonstrated that noncrossover products occur at the same time as HJ intermediates and that crossovers did not appear until these intermediates were resolved some considerable time later. This further suggests that the decision between crossover and noncrossover recombination takes place during the DSB to HJ transition, as also predicted by Hunter and Kleckner (2001) and reviewed by Zickler and Kleckner (1999), and argues that only crossover recombination involves a (detectable) HJ intermediate.

There is evidence to suggest that the choice of pathway occurs prior to the formation of SEIs, at the stage when DSBs and their homologous partners appear to be interacting before strand invasion (Weiner and Kleckner, 1994), and consequently that SEIs and dHJs are both intermediates of a crossover-specific pathway (Hunter and Kleckner, 2001). Further support for a separate pathway of noncrossover resolution that does not contain a HJ intermediate is given by the observations in *S. cerevisiae ndt80* mutants (Allers and Lichten, 2001). Ndt80 is a meiosis-specific transcription factor that contributes to the exit from pachytene (Chu and Herskowitz, 1998), the stage at which homologues desynapse and chiasmata form to maintain a stable interhomologue connection (table 1.1). It is suggested that HJ resolvases are under Ndt80 control and has been demonstrated that *ndt80* mutant strains show an accumulation of HJ intermediates and a five-fold reduction of crossovers, but an unaffected number of noncrossovers (Allers and Lichten, 2001).

As the simplest SDSA model of meiotic recombination does not predict formation of HJ intermediates, Allers and Lichten suggest that noncrossover products arise via SDSA, crossover products arise via the DSB repair model and, in agreement with Hunter and Kleckner (2001), that the decision takes place at the strand invasion step during the leptotene to zygotene transition. This is supported by cytological observations of RNs. Early nodules contain the Rad51 and Dmc1 strand exchange proteins and are postulated to mark the sites of all strand invasion events within the SC (Anderson *et al.*, 2001). Late nodules do not contain strand exchange enzymes, are not seen until early nodules disappear and are distributed in patterns that correlate with those of crossovers (Carpenter, 1987, Zickler *et al.*, 1992; Zickler and Kleckner, 1999; Allers and Lichten, 2001).

1.1.3: Distribution and Regulation of Meiotic Recombination Events

As part of recombination crossovers are required for proper chromosome segregation and the generation of diversity. Too few crossovers are therefore undesirable both in the long and the short-terms and too many crossovers will almost inevitably lead to deleterious effects such as inherited disease. To compensate, the distribution and frequency of recombination events appears to be regulated by features of chromatin and by higher order chromosome structure. As a result, there are areas across chromosomes in which recombination is intense and highly localised. Such areas are known as recombination hotspots.

1.1.3.a: Meiotic Recombination Hotspots in Yeast

DSBs occur non-randomly along *S. cerevisiae* chromosomes; almost all occur in intergenic regions corresponding to gene promoters (Cao *et al.*, 1990; Baudat and Nicolas 1997). As DSBs are initiators of recombination it is a natural consequence that yeast meiotic recombination hotspots correspond to these same regions. Each *S. cerevisiae* chromosome has at least four hotspots, as demonstrated by global mapping of meiotic DSBs for all 16 *S. cerevisiae* chromosomes (Gerton *et al.*, 2000) and several have been identified, including those near the ARG4 (Nicolas *et al.*, 1989; de Massy and Nicolas, 1993; Sun *et al.*, 1991), HIS4 (White *et al.*, 1991,1993; Detloff *et al.*, 1992) and HIS2 (Malone *et al.*, 1994) loci. Furthermore, the ectopic insertion of cloned yeast and bacterial sequences immediately downstream of the HIS4 gene created the HIS4LEU2 hotspot (Cao *et al.*, 1990) and there exists a well-characterised hotspot in *S. pombe* that is caused by the *ade6-M26* G to T transversion mutation (Schuchert *et al.*, 1991; Fox *et al.*, 1997). DSBs do not occur at a specific sequence but are dispersed through a region of 50-500 bp (de Massy *et al.*, 1995; Liu *et al.*, 1995, Xu and Kleckner, 1995; Xu and Petes, 1996). The formation of DSBs is therefore not sequence-specific but is site-specific and, more often than not, these sites display increased meiosis-specific DNaseI- or micrococcal nuclease– (MNase) sensitivity (Fan and Petes, 1996; Ohta *et al.*, 1994, Wu and Lichten, 1994), indicating that hotspots reflect the presence of open chromatin domains that allow the Spo11 endonuclease access to the DNA. Of course, the change in chromatin configuration may actually be provoked by a Spo11p-containing recombination complex as the open chromatin domains are observed prior to the formation of DSBs (Ohta *et al*, 1994; 1998) and the extent of nuclease-sensitivity seems to require functional MRE11, RAD50, XRS2 and MRE2 (Ohta *et al.*, 1998). In the presence of nucleotide heterology at DSB hotspots, the frequency of meiotic DSBs is reduced (Goldman and Lichten, 2000), suggesting that interhomologue recognition is also important for DSB formation. RAD50, XRS2 and MRE11 have been implicated in this role (Ohta *et al.*, 1998).

1.1.3.a.i: α-Hotspots

The HIS4 hotspot region contains binding sites for the transcription factors (TFs) Rap1, Gcn4, Bas1 and Bas2 (Devlin *et al.*, 1991; White *et al.*, 1991; 1993). Loss of any, other than Gcn4, leads to the loss of hotspot activity through the elimination of DSBs and the demise of any significant changes in chromatin structure (White *et al.*, 1991; 1993). However, deletion of the upstream TATAA sequence, which substantially reduces transcription, has no effect on hotspot activity (White *et al.*, 1992). Therefore, transcription factors, but not transcription, are required for HIS4 hotspot activity.

The activity of Rap1p at the HIS4 hotspot has been studied most extensively demonstrating that, in order to fully stimulate hotspot activity, transcription factors require both intact DNA binding and activation domains (Kirkpatrick *et al.*, 1999a). This suggests that the TF binds to DNA and opens chromatin thus allowing Spo11p access to the DNA. Alternatively, once the TF has opened the chromatin, its activation domain then recruits the Spo11 endonuclease. This contact may be indirect; the Rad50, Mre11 and Xrs2 proteins have once again been implicated and are suggested to form a complex that binds to open chromatin and interacts with Spo11p (Johzuka *et al.*, 1995; Raymond and Kleckner, 1993). This complex may be acting as a link between Spo11p and the TFs that open the chromatin (Fan and Petes, 1996). This also provides a reasonable explanation as to why RAD50, MRE11 and XRS2 are vital to DSB formation (figure 1.3).

Variation in hotspot activity could be explained by a varying ability of different TFs to efficiently recruit the Spo11 endonuclease and so stimulate recombination. This may also explain why all TF-binding regions in yeast do not represent hotspots (Baudat and Nicolas, 1997).

The G→T transversion at the *S. pombe* ade6-M26 hotspot creates a binding site for the Atf1/Pcr1 (Mts1/Mts2) transcription factor, which must attach before recombination activity begins (Kon *et al.*, 1997). It has been suggested that hotspots whose activity depends on TFs may be common in eukaryotes. These have been termed α-hotspots (Kirkpatrick *et al.*, 1999a).

1.1.3.a.ii: β-Hotspots

Ectopic insertion of DNA sequences into yeast chromosomes can stimulate both hotspots of meiotic recombination and gene expression without the association of TFs (Cao *et al.*, 1990; Wu and Lichten, 1995; Xu and Kleckner, 1995, Kirkpatrick *et al.*, 1999a). It is possible that these discontinuities in chromosomal context create open chromatin domains that are accessible to the recombination machinery. Open chromatin domains are free of nucleosomes, the basic unit of DNA compaction. The presence of nucleosomes may thereby inhibit the initiation of meiotic recombination. As the formation and positioning of nucleosomes is sensitive to DNA sequence (Kirkpatrick *et al.*, 1999b) the ectopic DNA insertions must consist of sequences that exclude nucleosomes. This has clearly been demonstrated for tandem arrays of the 5′ CCGNN 3′ repeat (Wang and Griffith, 1996; Kirkpatrick *et al.*, 1999b). As there is no TF known to bind to this sequence, and yet it stimulates transcription and meiotic recombination, it is clear that nucleosome-excluding sequences behave in *cis* to create hotspots of meiotic recombination, in contrast to the *trans*-acting TFs of α-hotspots. This second kind of hotspot has been termed a β-hotspot (Kirkpatrick *et al.*, 1999a).

A naturally occurring β-hotspot may be located between LEU2 and CEN3 on chromosome III of *S. cerevisiae*. The majority of recombination events at this hotspot are dependent upon the presence of ARS307, a replication origin. The stimulation is independent of DNA replication and is likely to be caused by another feature of the ARS307, such as the relatively nucleosome free environment of ARS elements (Rattray and Symington, 1993). As subsequent studies have failed to find any significant association between the location of replication origins and hotspots (Gerton *et al.*, 2000) the ARS307 hotspot may be due to a coincidence of factors; perhaps divergence from the ARS core consensus sequence has created a sequence at ARS307 that is particularly inhospitable to nucleosome formation or predisposed to creating an open chromatin domain.

1.1.3.a.iii: Identification of Yeast Hotspots

Blumenthal-Perry *et al* (2000) identified a motif, which they called the CoHR profile, that seemed to associate with *S. cerevisiae* recombination hotspots. The profile is 50 bp long with a poly(A) tract in its centre but it is also flexible, in that it can have several gaps of unrelated sequence that stretch its length out to 250 bp. The global study of Gerton *et al* (2000) could not identify a significant association between the CoHR profile and meiotic recombination hotspots suggesting that, at best, the CoHR profile tentatively identifies only a subset of hotspots. So, beyond examining every putative transcription factor binding sequence in every putative gene promoter region or testing every putative nucleosome-excluding sequence for hotspot activity, there is no simple primary sequence determinant that allows the quick and easy identification of meiotic recombination hotspots in *S. cerevisiae*. However, significant associations of hotspots with regions of high G + C base composition have been observed (Gerton *et al.*, 2000), which suggests that DSBs (and recombination activity) are regulated in part by global features of chromosome structure.

1.1.3.b: Global Regulation of Meiotic Recombination

It is clear that the distribution of recombination events is regulated to some extent at a local level through the requirement of an open chromatin domain, which in turn requires the activity of certain TFs or the exclusion of nucleosomes. However, it is clear that other factors, reflecting an imposition of more global controls upon local determinants, also affect hotspot activity. Experiments have shown that hotspots can be influenced by chromosomal location; Wu and Lichten (1995) showed that the frequency of DSBs and associated recombination at the ARG4 hotspot varied by a factor of between 3 and 10, depending upon the location of an ARG4-containing insert. Further evidence was provided by a study that used a recombination-proficient reporter to map meiotic recombination domains on chromosome III of *S. cerevisiae* (Borde *et al.*, 1999).

The molecular basis of the global regulation of crossover events has not been elucidated but it has been demonstrated that targeting of Spo11p to specific sites is sufficient to stimulate meiotic recombination even in naturally 'cold' regions (Pecina *et al.*, 2002). Spo11p was in a construct also containing the DNA binding domain of the Gal4 TF and the initiation of recombination required all other DSB gene functions, suggesting that the differential recruitment, assembly and/or maturation of Spo11p-containing recombination machinery is a part of the genome-wide level of recombination regulation (Pecina *et al.*, 2002).

1.1.3.b.i: Chromatin Organisation

The DNA within a SC is organised into a series of chromatin loops (figure 1.2) and the rate of meiotic recombination appears to be influenced by the relative density of this chromatin packaging. The average rate of recombination per unit length of DNA in yeast is nearly 300 times that in humans but DNA is compacted about 25 times more in the SCs of human cells than it is in the SCs of yeast cells (Loidl *et al.*, 1995); i.e. the tighter the packaging, the lower the rate of recombination. Evidence that this is regulated is seen when human DNA is introduced into yeast, where it adopts both the packaging and the rate of recombination typical of yeast DNA (Sears *et al.*, 1992; Loidl *et al.*, 1995). Therefore, meiotic chromatin organisation is not an inherent chromosomal property but appears to be evidence for a (host-specific) nuclear level of recombination regulation.

1.1.3.b.ii: Competitive Inactivation

Insertion of a strong DSB site near a pre-existing DSB site reduces the hotspot activity of both sites without significantly altering chromatin structure, as assayed with DNaseI (Wu and Lichten, 1995; Fan *et al.*, 1997; Xu and Kleckner, 1995; Ohta *et al.*, 1999). This is called competitive inactivation (Ohta *et al.*, 1999) and its affect is stronger in *cis* than in *trans* (Fan *et al.*, 1997), a feature that is perhaps most strongly demonstrated by observations at LEU2, where meiotic recombination was suppressed by the insertion of a strong DSB site 17 kb away (Wu and Lichten, 1995). Although chromatin structure is unaffected, there is a suppression of MNase hypersensitivity, suggesting that MNase hypersensitivity is a critical step in DSB formation (Ohta *et al.*, 1999). If MNase hypersensitivity is a reflection of the binding of a multiprotein complex in preparation for DSB formation, competitive inactivation may then be a reflection of competition between adjacent hotspots for proteins contributing to this complex, and their distribution being controlled at the chromosomal or nuclear domain.

One interesting idea for the mechanics of this competition suggests that a critical density of these proteins must assemble non-cooperatively within a restricted region of DNA (presumably open chromatin domains) in order to catalyse a recombination event (Kirkpatrick *et al.*, 1999b). Twelve copies of the 5′ CCGNN 3′ repeat act as a β-hotspot, but a $(CCGNN)_{48}$ tract is a coldspot. As the $(CCGNN)_{48}$ tract consists of four tandem copies of $(CCGNN)_{12}$ it is possible that the $(CCGNN)_{48}$ tract behaves as four adjacent and competing hotspots. The rate-limiting protein(s) required for the initiation of recombination still bind to the longer tract but they do not reach the critical density required for recombination to proceed (Kirkpatrick *et al.*, 1999b). Together with the observation that competitive inactivation is stronger in *cis* than in *trans* this implies that the diffusion of the recombination-initiating proteins is limited and cannot be recruited from other hotspots in the genome.

1.1.3.b.iii: Crossover Interference

The regulation of crossover distribution along a chromosome is further inferred from the observation of crossover interference (also known as chiasma interference), in which a crossover at one locus decreases the probability of another crossover in the vicinity. This suggests that there is an inhibitory signal from a site of crossover to nearby potential sites. The SC has been implicated in this role as mutations in the ZIP1 gene abolish interference (Sym and Roeder, 1995). In addition, organisms that do not form the SC, such as *S. pombe* and *Aspergillus nidulans*, do not exhibit interference (Egel-Mitani *et al.*, 1982; Kohli and Bahler, 1994). Perhaps a signal is transmitted along the SC that causes the release of any late recombination nodules not already involved in a crossover event (Zickler *et al.*, 1992 and references therein), an idea supported by observations of the putative late nodule component Msh4p. Mutant *msh4* strains form the SC but interference is virtually eliminated (Ross-Mcdonald and Roeder, 1994; Roeder, 1997). If late nodules do indeed represent those sites that are resolved as crossovers, then Msh4p may well be acting as an acceptor or initiator of an inhibitory signal.

Recently, a simple reaction-diffusion model for the mechanism of interference has been proposed (Fujitani *et al.*, 2002) in which "randomly-walking precursors" become immobilised and mature into crossover points. Interference could then be caused by the collision and subsequent destruction of two random walkers and by the collision of one random walker with an immobilised point. This is obviously not a genetic model and only serves to predict what might be happening physically. The randomly walking precursor could be anything from a premeiotic contact point, as identified by Weiner and Kleckner (1994), to an early RN or even a component of the putative Spo11p-containing recombination machinery.

It is tempting to suggest that competitive inactivation and crossover interference amount to one and the same thing. However, there are several reasons for caution. First, if the two processes were, at the very least, related, then all events initiated by DSBs ought to exhibit interference. This does not appear to be the case as gene conversion events look as though they are distributed without interference (Mortimer and Fogel, 1974). Second, crossover interference has been observed over much larger distances than the 17 kb representing the longest distance of competitive inactivation so far observed (Wu and Lichten, 1995). Third, *zip1* mutants, which abolish interference, have normal levels of gene conversion (Sym and Roeder, 1995). This leads to the conclusion that competitive inactivation occurs prior to DSB formation and that crossover interference operates at some stage after DSB formation, probably along the crossover-only recombination pathway. As suggested by Fan *et al.* (1997) a simple test of any link between inactivation and interference would be to determine if there is normal competitive inactivation in *zip1* mutant strains that are defective for interference.

1.1.3.b.iv: Obligate chiasma

In order to ensure proper chromosome segregation there has to be at least one crossover per chromosome, a phenomenon referred to as obligate chiasma. This is likely to be related to the observation that crossover frequency is regulated by chromosome size. Of the 300 DSBs that appear at every *S. cerevisiae* meiosis about 100 are resolved as crossovers distributed along 16 chromosomes of unequal size (Paques and Haber, 1999). As is the case with most organisms, the number of crossovers per kilobase increases with decreasing chromosome size, rather than remaining constant as would be expected if crossovers occurred randomly (Jones, 1984; Kaback *et al.*, 1989). Furthermore, when a large chromosome of *S. cerevisiae* was bisected the two resulting smaller chromosomes underwent an increased number of crossovers per kilobase (Kaback *et al.*, 1992). It is likely that meiotic recombination hotspots contribute to these observations as it has also been noted that both the density and the recombinational activity of hotspots on smaller chromosomes is significantly greater than on the larger chromosomes (Gerton *et al.*, 2000). Mutations that reduce interference also randomise the distribution of crossovers among chromosomes such that some homologous pairs fail to crossover and so nondisjoin (Egel, 1995; Sym and Roeder, 1995; Chua and Roeder, 1997). Global regulatory mechanisms, as reflected by interference, might therefore exist to ensure that large chromosomes do not undergo an excessive number of exchanges and that at least one crossover occurs per homologue pair.

1.1.3.b.v: The Recombination Checkpoint

As recombination is such an essential process it is inevitable that checkpoint machinery acts during meiosis to ensure that events are properly coordinated. Many mutants in which recombination is defective, including *rad50S*, *rad51*, *dmc1* and *zip1* are arrested at the pachytene stage of M1 prophase (Bishop *et al.*, 1992; Xu *et al.*, 1997; Shinohara *et al.*, 1992; Storlazzi *et al.*, 1996). The intermediates that activate checkpoint-mediated arrest (presumably unrepaired DSBs and/or unresolved HJs) can be bypassed by mutations to RAD17, RAD24 or MEC1 (Lydall *et al.*, 1996), providing clear evidence that these are meiotic recombination checkpoint genes. The SC proteins RED1 and MEK1 also appear to be important to this checkpoint as deletions of either also alleviate the arrest of *rad50S*, *rad51*, *dmc1* and *zip1* mutants (Xu *et al.*, 1997). If loss of Red1p does indeed result in the loss of a specific interhomologue-only pathway of meiotic recombination (Schwacha and Kleckner 1997; Xu *et al.*, 1997) it is possible that any alternative pathway does not result in intermediates that activate the checkpoint. Recognition of these intermediates may involve the Mer3 helicase, which is suggested to have an important role in crossover resolution (Nakagawa and Kolodner, 2002). Mer3p has been suggested to trigger the checkpoint if it is not bound to DNA, perhaps at the stage of HJ resolution (Nakagawa and Kolodner, 2002).

1.2: Human Meiotic Recombination

Studies in yeast have demonstrated that meiotic recombination is a complex and highly regulated process involving a wide array of proteins and complexes. Determination of the functions that these proteins and complexes perform, together with the identification and study of structural intermediates, is likely to provide major insights into the process of meiotic recombination in humans, which have much larger and more repetitive genomes.

1.2.1: Pedigree Analysis

The detailed analysis of meiotic recombination in the human genome is very difficult because of the small size of human families and the low frequency of crossover events per unit of physical distance in the germline, which equates to a mean rate of only 1 cM/Mb. Until recently this imposed a resolution limit of 0.1–1.0 cM (typically 0.1–1.0 Mb) on crossovers studied through the traditional method of pedigree analysis. Nevertheless linkage maps so constructed have detected considerable large-scale variation in recombination event distribution across human chromosomes and between the sexes. (NIH/CEPH Collaborative Mapping Group, 1992; Weissenbach *et al.*, 1992; Gyapay *et al.*, 1994; Broman *et al*, 1998; Mohrenweiser *et al.*, 1998).

The human genome sequencing project (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001) has since allowed the comparison of the genetic linkage maps with a detailed physical map, thereby providing a comparison of rates of crossover, rather than the absolute number of recombination events (Yu *et al.*, 2001; Kong *et al* 2002). The Yu *et al.* study, based on 188 meioses within about 58% of the genome, estimated that recombination rate across the human genome varied from 0 to 8.8 cM/Mb and identified regions of up to 6 Mb in length with particularly low or high recombination rates, which they termed recombination deserts and jungles respectively. They also suggested that the average male recombination rate was 0.92 cM/Mb, slightly more than half the female average rate of 1.68 cM/Mb. This sex-specific variation is actually rather complex; there are areas of the genome, such as sub-telomeric regions, where the rate of recombination is particularly high in men but not in women, and other regions, near centromeres for example, where the reverse is true (Broman *et al*, 1998; Mohrenweiser *et al.*, 1998; Kong *et al.*, 2002).

More recently, Kong *et al.*, (2002) were able to improve on the resolution of the Yu *et al.* map by approximately five times through looking at 1,257 meioses in 146 Icelandic families. There are some minor differences between the results of the studies, e.g. the Iceland map revealed a slightly lower sex averaged recombination rate of 1.13 cM/Mb, but these can be attributed to the higher resolution of the Iceland map and the fact that the groups used drafts of the human genome sequence at different stages of development. Overall, the major conclusions of the studies are the same, but reveal nothing about fine-scale crossover activity across the chromosomes.

1.2.2: Cytological Analysis

Classically, meiotic recombination has been investigated directly in human spermatocytes by making use of chiasmata to determine the numbers and distributions of crossovers at diakinesis or metaphase I. More recently, an alternative approach has been developed to exploit the behaviour of the mismatch repair protein MLH1, which, as a component of late RNs (above, section 1.1.1.c.ii), forms discrete foci along the axes of homologous chromosomes prior to meiosis (Baker *et al.*, 1996).

Cytological analyses have shown that, although chiasmata can arise at any position along an autosomal bivalent, they do show significant preferences. The preferential localisation of chiasmata must result, in part, from crossover interference (above, section 1.1.3.b.iii) but morphological features also appear to play a role. For example, both heterochromatin and the relative position of the centromere appear to effect the positions favoured for chiasma formation (Laurie *et al.*, 1981; Saadallah and Hultén, 1983; Goldman and Hultén, 1993) and, in male meiosis, chiasma formation is favoured in telomeric and subtelomeric regions (Hultén, 1974; Laurie and Hultén, 1985). This distribution pattern is mirrored by MLH1 foci in pachytene spermatocytes (Barlow and Hultén, 1998). In contrast, it has been shown that chiasmata are more interstitial in oocytes (Wallace and Hultén, 1985; Hultén *et al.*, 1990; Tease *et al.*, 2002). Therefore, in agreement with pedigree analysis (above) this considerable intersex difference must be a reflection of a system that controls the localisation of crossovers. Unfortunately both pedigree analysis and detection of chiasmata suffer from low resolution and an inability to detect gene conversion events. Therefore, they do not provide any clues as to how this control might function.

1.2.3: Single Sperm Typing

As there is an almost unlimited number of sperm or meioses available from any male, sperm typing presents an opportunity to study recombination in individuals (Li *et al.*, 1988). In addition, in regions with accurate physical maps, sperm-typing experiments are able to provide evidence for variation in recombination rates at a higher resolution than that obtained with either cytological or pedigree analyses, and have led to the identification of a putative recombination hotspot (in a DNA segment that is 280 kb long) near the locus for Huntington disease (Hubert *et al.*, 1994). More recently, sperm typing has identified modest variation in recombination rate across the pseudoautosomal region at the ends of the short arms of the X and Y chromosomes (Lien *et al.*, 2000) and has been used to produce a map of recombination across the major histocompatibility complex (MHC) (Cullen *et al.*, 2002). In addition, by suggesting that reduced recombination in the pseudoautosomal region is a significant cause of XY nondisjunction, and thus Klinefelter syndrome (Shi *et al.*, 2001), single sperm typing provided the first direct evidence that reduced recombination has an effect on nondisjunction in human gametes (above, section 1.1.1.b). However, single sperm studies are very technically challenging and, unfortunately, their resolution remains limited.

1.2.4: Linkage Disequilibrium Studies

The family-based approaches that have been successfully employed to identify the variants responsible for monogenic diseases are largely ineffective when faced with the complexity of common multifactorial diseases. Hence, the main method used to identify common low penetrance susceptibility alleles is a case-control association study, in which genetic and phenotypic variation is compared in large population samples. The ability to detect association between marker alleles and disease critically depends on the extent of linkage disequilibrium (LD) between disease-causing alleles and surrounding marker alleles. LD is the non-random association of alleles at closely linked loci and, in an ideal high LD scenario, an allele found at one locus would predict which allele would be found at the other, making one of the loci redundant for association mapping purposes.

The number of markers that have to be applied in association studies depends on the distance over which useful LD can be detected. A theoretical simulation, which assumed that crossovers were randomly distributed, predicted that useful LD extends to an average of only 3 kb (Kruglyak, 1999), implying that 500,000 SNPs would be required for a whole-genome association study. In general, LD decays as distance increases and the main force that breaks down LD is recombination, so a genomic region in intense LD is expected to have been recombinationally inactive through human evolutionary history. LD studies thereby allow a higher resolution analysis of crossover through the use of haplotype diversity to infer the recombination events that have accumulated over thousands of generations in the history of a contemporary population. This has led to the identification of a number of putative human meiotic recombination hotspots, including at the β-globin gene cluster (Chakravarti *et al.*, 1984) and at the *PGM1* gene (Yip *et al.*, 1999). Furthermore, the physical mapping of occasional recombination breakpoints detected in family studies revealed heterogeneity in recombination rates across the human major histocompatibility complex class II region (MHC II) (Cullen *et al.*, 1995; 1997). This is not entirely unexpected given the important immunological role of the MHC; high frequencies of recombination may indicate selective pressures in favour of haplotype diversification, which may eventually prove useful against an invading pathogen. Conversely, low frequencies of recombination could indicate selective pressure favouring specific allele associations.

Patterns of LD can be disrupted by factors that are specific to the genomic region, including recombination rate, recurrent mutation and demographic processes such as natural selection, genetic drift, population bottlenecks and admixture (Hedrick, 1987; Ardlie *et al.*, 2002). Hence, a number of studies have shown that there is considerable variation in the extent of LD from one genomic region to another (Taillon-Miller *et al.*, 2000; Abecasis *et al.*, 2001; Reich *et al.*, 2001; 2002). Variation in LD patterns has also been observed between populations (Laan and Pääbo, 1997; Frisse *et al.*, 2001; Reich *et al.*, 2001), as exemplified by the consistent observation that LD in non-African populations extends over longer distances than in Africans, a relatively ancient population that has had time to accumulate many recombination events. Therefore, the feasibility of association studies depends on an understanding of the rules that govern patterns of LD in the human genome.

1.2.4.a: Haplotype Blocks

The recent evidence suggesting that recombination events localise into hotspots (below; Jeffreys *et al.*, 2001; Reich *et al.*, 2002) indicates that human LD patterns may not be so complicated as they first appear; recombination hotspots might break the genome into a series of discrete blocks of high LD, 3–100 kb or more in length, in which there are just a few common haplotypes (Collins, 2000; Daly *et al.*, 2001; Patil *et al.*, 2001; Dawson *et al.*, 2002; Gabriel *et al.*, 2002; Reich *et al.*, 2002). These "haplotype blocks" would reduce genotyping requirements for large-scale association studies because their high LD would render many variant sites redundant (Johnson *et al.*, 2001). However, it is also plausible that block patterns could arise solely as a result of random crossover and demographic processes (Subrahmanyan *et al.*, 2001), resulting in inconsistent block boundaries within and between populations, and making haplotype blocks less useful for association studies. A recent study in the MHC II region suggested that it had been subjected to a high rate of haplotype turnover, but also demonstrated that three populations with different demographic histories (UK north Europeans, Saami and Zimbabweans) have very similar LD patterns (Kauppi *et al.*, 2003). This suggested that the recombination hotspots in the region (Jeffreys *et al.*, 2001), and not population history, do indeed direct the formation of haplotype blocks (Kauppi *et al.*, 2003). However, as the MHC II region is under selective pressure to diversify, it cannot be regarded as a "typical" area of the genome. Therefore, analysis of LD patterns in other areas of the genome is required before it can be determined if the structuring of LD blocks in the MHC II region is reflective of processes operating in the genome as a whole.

1.2.5: High Resolution Analysis of Meiotic Recombination

An understanding of human meiotic recombination in molecular terms will require very high resolution data so that hotspots can be defined in terms of DNA sequence and chromatin structure, as they have been in yeast. Unfortunately, because of the many factors that disrupt LD patterns (above), LD studies are limited, and thus the conclusions drawn from them are contentious. Recently, an alternative method has been developed that is capable of analysing recombination events at a resolution of 0.0001 cM or less (figure 1.5). This is essentially a two-tier strategy in which single nucleotide polymorphism (SNP) genotypes are first subjected to LD analysis in order to screen the target region for evidence of localised historical recombination. This analysis can lead to one of several results: first, all markers may be in intense LD, suggesting that the whole region is recombinationally inert. Second, if LD declines slowly and uniformly with physical distance this suggests that recombination events have occurred randomly across the region. Third, there may be regions of intense LD separated by intervals of free association; this sudden breakdown of LD implies the possible existence of a localised recombination hotspot. Fourth, if there is no LD even between markers that are physically very close, there is an implication of extreme recombination activity over the whole region. It is worth re-emphasising that LD studies merely allow inference of historical recombination events. For instance, LD blocks can be generated, without the need to invoke hotspot activity, through chance clustering of a few historical crossovers (Wang *et al.*, 2002). A picture of what is truly happening can only be gained

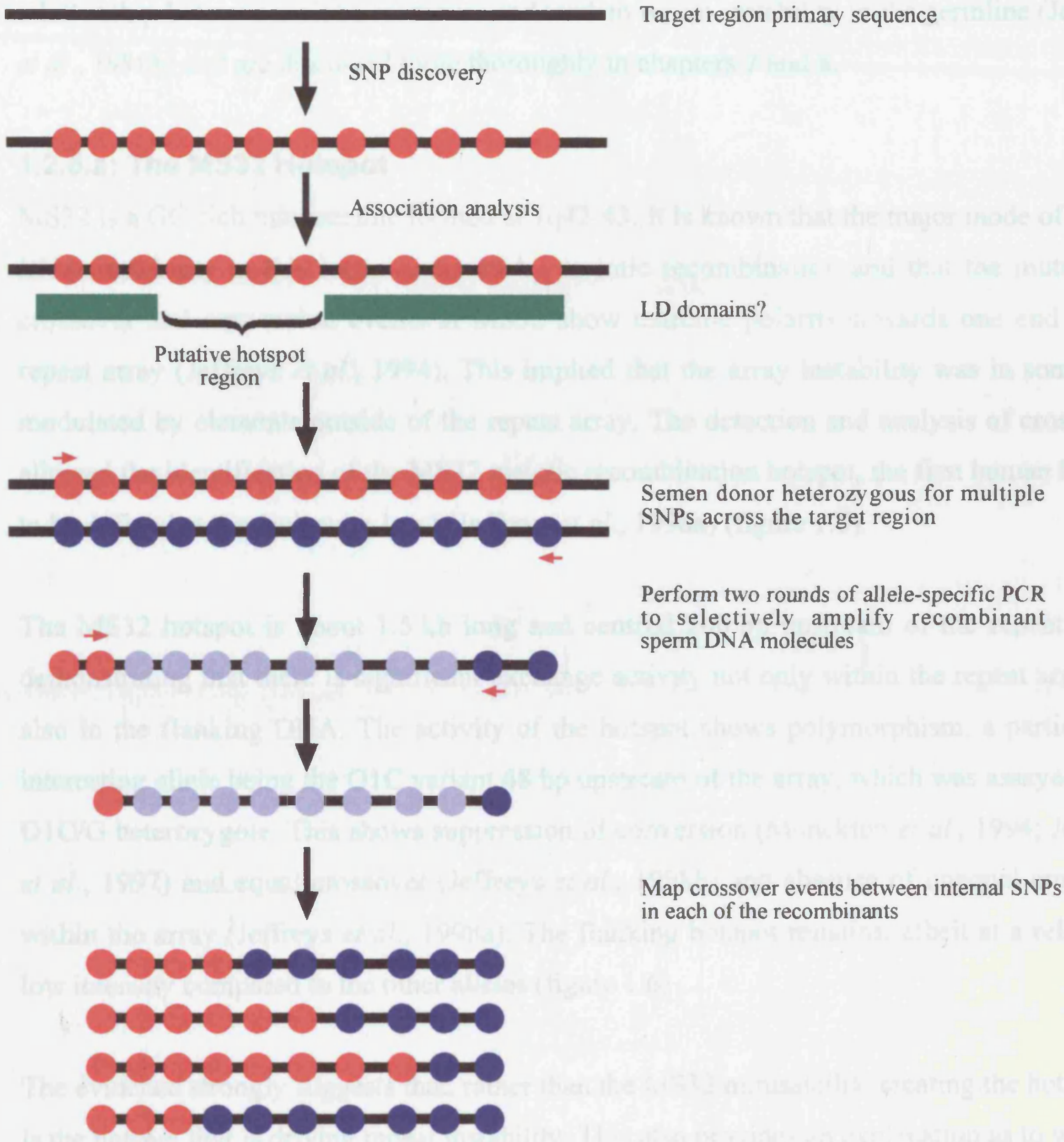

Target region primary sequence
SNP discovery
Association analysis
LD domains?
Putative hotspot region
Semen donor heterozygous for multiple SNPs across the target region
Perform two rounds of allele-specific PCR to selectively amplify recombinant sperm DNA molecules
Map crossover events between internal SNPs in each of the recombinants

Figure 1.5: The two-tier strategy for high-resolution analysis of meiotic crossovers in human DNA. A chosen target region is subjected to high density SNP discovery, followed by genotyping of all SNPs in a reference human population. The haplotypes so revealed can be used to infer the recombination events that have occurred in the target region during the history of the population. The second stage begins with the identification of a man with multiple heterozygous SNPs across the target region. Two rounds of repulsion-phase allele-specific PCR are used to selectively amplify recombinant molecules from batches of sperm DNA and crossover events are mapped by ASO hybridisation.

through the second step of this strategy, in which the evidence gained from the LD analysis is used to direct allele-specific PCR to heterozygous SNPs in order to recover crossover molecules directly from sperm DNA. These methods were initially developed to explore the relationship between meiotic crossover and tandem repeat instability in the germline (Jeffreys *et al.*, 1998b) and are discussed more thoroughly in chapters 7 and 8.

1.2.5.a: The MS32 Hotspot

MS32 is a GC-rich minisatellite located at 1q42-43. It is known that the major mode of repeat DNA instability at this locus is driven by meiotic recombination and that the mutational crossover and conversion events at MS32 show extreme polarity towards one end of the repeat array (Jeffreys *et al.*, 1994). This implied that the array instability was in some way modulated by elements outside of the repeat array. The detection and analysis of crossovers allowed the identification of the MS32 meiotic recombination hotspot, the first human hotspot to be defined at the molecular level (Jeffreys *et al.*, 1998a) (figure 1.6).

The MS32 hotspot is about 1.5 kb long and centred 200 bp upstream of the repeat array, demonstrating that there is significant exchange activity not only within the repeat array but also in the flanking DNA. The activity of the hotspot shows polymorphism, a particularly interesting allele being the O1C variant 48 bp upstream of the array, which was assayed in an O1C/G heterozygote. This shows suppression of conversion (Monckton *et al.*, 1994; Jeffreys *et al.*, 1997) and equal crossover (Jeffreys *et al.*, 1998b) and absence of unequal crossover within the array (Jeffreys *et al.*, 1998a). The flanking hotspot remains, albeit at a relatively low intensity compared to the other alleles (figure 1.6)

The evidence strongly suggests that, rather than the MS32 minisatellite creating the hotspot, it is the hotspot that is driving repeat instability. This also provides an explanation as to why the conversion and crossover events in the repeat array show such extreme polarity. A similar hotspot has been provisionally identified upstream of the MS31 minisatellite (C. Hollies, M. Panayi and A.J. Jeffreys, unpublished data) and significant levels of exchange activity have been detected within the repeat array and flanking DNA of the extremely unstable CEB1 minisatellite (Buard *et al.*, 2000). It thus appears that minisatellites are occasionally generated as by-products of recombination hotspot activity.

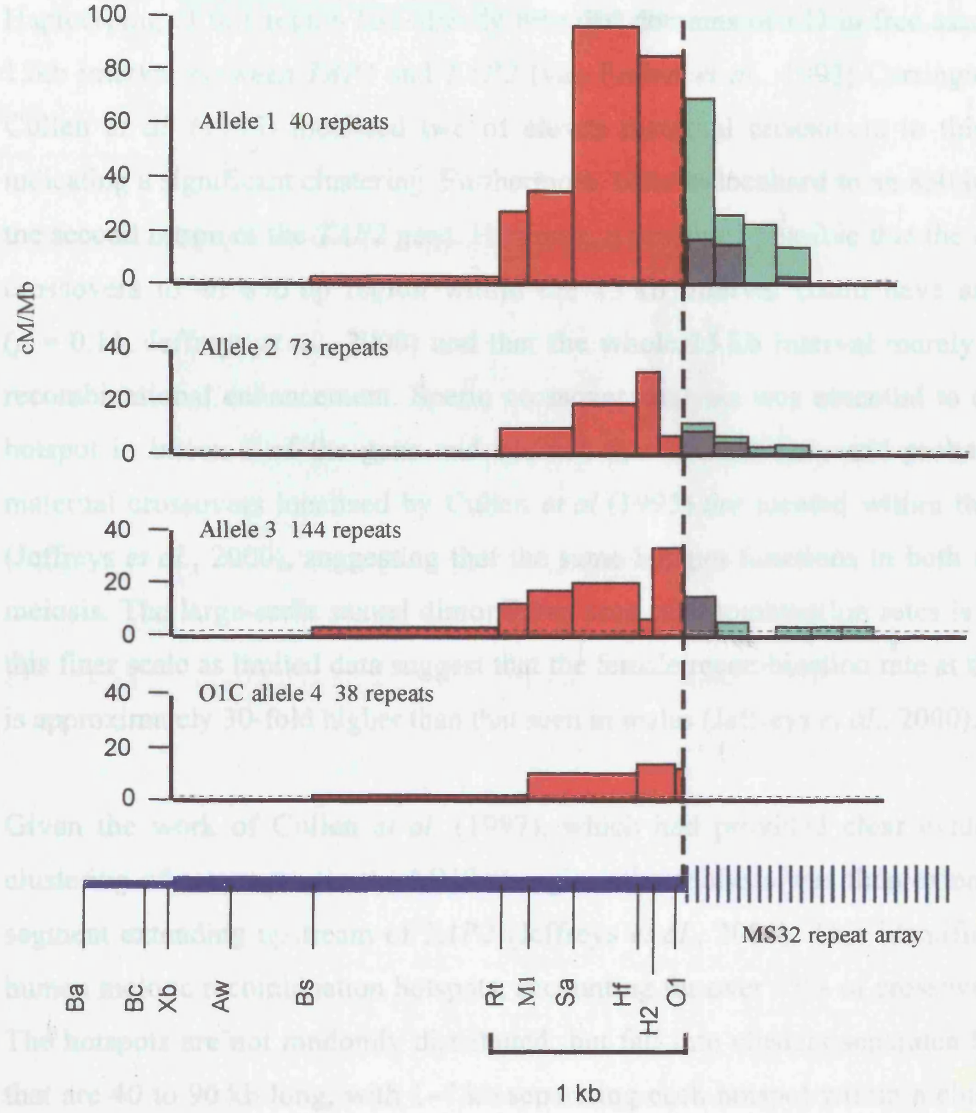


Figure 1.6: The polymorphic activity of the MS32 meiotic recombination hotspot. Data taken from Jeffreys *et al.* (1998*a*). Flanking crossovers are shaded in red, equal and unequal crossovers within the repeat array are shaded in grey and green respectively. Positions of SNPs used to locate crossover breakpoints are marked at the bottom of the diagram.

1.2.5.b: Recombination within the MHC class II region

The discovery of the MS32 hotspot raised the possibility that human meiotic crossovers may in general concentrate into local hotspots interspersed with recombinationally inert regions. To test this a putative hotspot in the MHC II region was analysed (Jeffreys *et al.*, 2000). Haplotyping of this region had already revealed domains of LD in free association across a 15kb interval between *TAP1* and *TAP2* (van Endert *et al.*, 1992; Carrington *et al.*, 1994). Cullen *et al.* (1995) localised two of eleven maternal crossovers to this 15 kb interval indicating a significant clustering. Furthermore, both co-localised to an 850 bp interval within the second intron of the *TAP2* gene. However, it remained possible that the clustering of two crossovers to an 850 bp region within the 15 kb interval could have arisen by chance (p = 0.11, Jeffreys *et al.*, 2000) and that the whole 15 kb interval merely showed modest recombinational enhancement. Sperm crossover analysis was essential to define the *TAP2* hotspot in intron 2 of the gene and showed that at least one, and probably both, of the maternal crossovers localised by Cullen *et al* (1995) are located within the sperm hotspot (Jeffreys *et al.*, 2000), suggesting that the same hotspot functions in both male and female meiosis. The large-scale sexual dimorphism seen in recombination rates is also apparent at this finer scale as limited data suggest that the female recombination rate at the *TAP2* hotspot is approximately 30-fold higher than that seen in males (Jeffreys *et al.*, 2000).

Given the work of Cullen *et al.* (1997), which had provided clear evidence for further clustering of crossovers in the MHC II region, the analysis was then extended to a 240 kb segment extending upstream of *TAP2* (Jeffreys *et al.*, 2001). This identified a further five human meiotic recombination hotspots, accounting for over 95% of crossovers in the region. The hotspots are not randomly distributed, but fall into clusters separated by blocks of LD that are 40 to 90 kb long, with 1–7 kb separating each hotspot within a cluster (figure 1.7). This study and others (above, section 1.2.4.a) provide growing evidence that recombination hotspots strongly influence patterns of LD and that structuring of diversity into LD blocks is common in the human genome.

1.2.5.c: Distribution of Human Meiotic Recombination Hotspots

Whereas nearly all yeast recombination hotspots are associated with transcriptional promoter regions (Cao *et al.*, 1990; Baudat *et al.*, 1997), this situation has only been seen for the weakest human hotspot so far identified (Jeffreys *et al.*, 2001). The others are found in a wide variety of genomic locations. Furthermore, the hotspots do not share any obvious primary sequence similarity (Jeffreys *et al.*, 2000; 2001), so it is currently impossible to predict hotspot location from human DNA sequences. However, in a low resolution analysis along

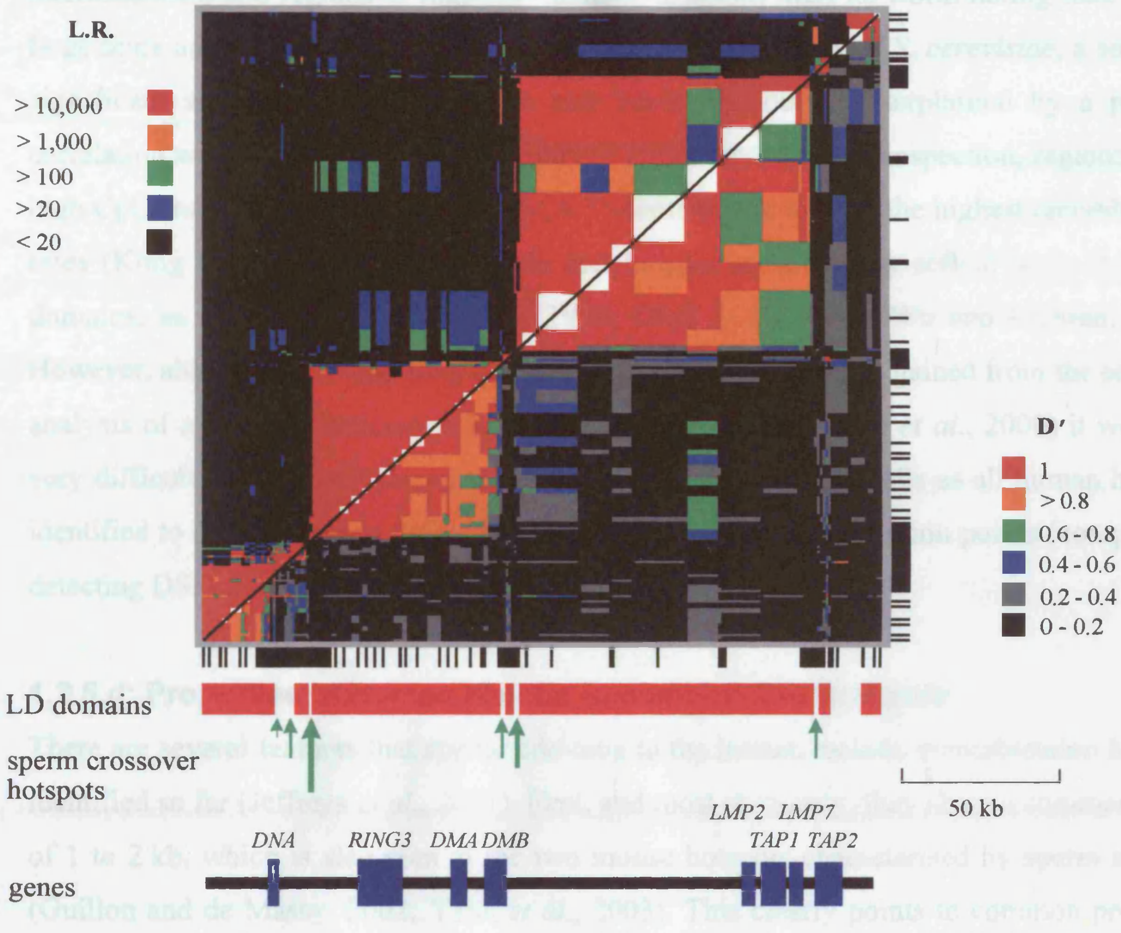


Figure 1.7: The LD pattern and location of meiotic recombination hotspots across the MHC II region (adapted from Jeffreys *et al.*, 2001). $|D'|$ measures (Lewontin, 1988) of complete LD between all pairs of markers with minor allele frequencies of at least 0.15 are shown in the bottom right triangle of the plot. $|D'| = 1$ for marker pairs showing only two or three haplotypes (shown in red), values of less than 1 indicate pairs with all four haplotypes and values of 0 (in black) indicate pairs in free association. The top left triangle shows the likelihood ratio in favour of significant association. These measures were determined from unphased diploid genotype data on 179 markers typed in a panel of 50 unrelated UK semen donors. Points are plotted as rectangles centred on each SNP (shown below and to the right of the plot) and extending half way to each marker. The very clear pattern of strong LD domains (indicated below the plot in red) that are abruptly separated by small intervals of free association implies the existence of local meiotic recombination hotspots. Sperm crossover analysis across the regions of complete or partial breakdown of LD identified a total of six hotspots across the MHC II region. Their positions and relative intensities are shown by the green arrows.

chromosome 22, Majewski and Ott (2000) did find a positive correlation between GT microsatellites and regions of high recombination. It may well be worth noting that, on this large scale and in agreement with the Gerton *et al.* (2000) study of *S. cerevisiae*, a small but significant amount of recombination rate variation could be explained by a positive correlation with GC content (Yu *et al.*, 2001). However, on closer inspection, regions with a high CpG fraction, but low GC and poly(A/T) content tend to have the highest recombination rates (Kong *et al.*, 2002). It is possible that hotspot location may reflect open chromatin domains, as in yeast (Fan and Petes, 1996; Ohta *et al.*, 1994, Wu and Lichten, 1994). However, although this suggestion is tenuously supported by results gained from the sequence analysis of a putative male-specific sub-telomeric hotspot (Badge *et al.*, 2000) it would be very difficult to confirm it through direct observation in meiotic cells as all human hotspots identified to date have been defined by mapping of crossover resolution points (compared to detecting DSBs in *S. cerevisiae*) and extend for only 1–2 kb.

1.2.5.d: Properties of Human Meiotic Recombination Hotspots

There are several features that appear common to the human meiotic recombination hotspots identified so far (Jeffreys *et al.*, 2001). First, and most obviously, they share a common width of 1 to 2 kb, which is also seen at the two mouse hotspots characterised by sperm analysis (Guillon and de Massy, 2002; Yauk *et al.*, 2003). This clearly points to common processes operating within these hotspots. Second, the vast majority of crossovers in hotspots are simple with a only a few (about 1%) showing a patchwork of DNA from both haplotypes at the site of crossover. These rare, complex exchanges presumably result from patchy repair of heteroduplex DNA generated during recombination. Third, most hotspots show fully reciprocal exchange, with reciprocal products arising in sperm at the same rate and with the same distribution across the hotspot (see figure 1.5, reciprocal exchanges in this example would be recovered by using a forward primer directed to the "blue" haplotype and a reverse primer specific to the red haplotype). Fourth, the centres of hotspots can be mapped quite precisely (within ± 30 bp) due to the symmetrical distribution of exchanges across them. However, the peak intensity over different hotspots varies quite considerably, from 0.4 cM/Mb for the *DNA1* hotspot in the MHC II region to 300 cM/Mb for the *SHOX* hotspot found within the pseudoautosomal region at the ends of the short arms of the X and Y chromosomes (Jeffreys *et al.*, 2001, May *et al.*, 2002). Furthermore, some hotspots show significant variation in crossover rates between men (Jeffreys *et al.*, 1998*a*; May *et al.*, 2002). High resolution analysis only identifies sites of crossover resolution, and so it remains possible that the initiating sites of recombination (presumably DSBs) occur remote from the hotspots, perhaps tens of kilobases away, as seen in *S. pombe* (Cervantes *et al.*, 2000; Young *et al.*, 2002). However this is unlikely as preliminary data show that human hotspots are also very active in gene conversion and that this activity is focused on the centre of the hotspot as defined by resolution of crossovers (Jeffreys and Neumann, 2002; A.J. Jeffreys, C.A. May, manuscript in preparation). Since conversions and crossovers in yeast result from alternative pathways initiated by the same DSB, it is most probable that human hot spots contain the sites or zones of initiation of recombination that results in either crossover or conversion products.

1.2.6: Mechanisms of Human Meiotic Recombination

It is difficult to make generalisations as so few human hotspots have been mapped. However, hotspots do appear to dominate human recombination and it is probable that the mechanisms are very similar to those identified in yeast as numerous homologues of *S. cerevisiae* genes involved in DSB repair have been identified (Shinohara *et al.*, 1993; Petrini *et al.*, 1995; Dolganov *et al.*, 1996; Romanienko and Camerini-Otero, 1999). It would also appear that human meiotic recombination is regulated at two levels; a global regulation, as evidenced by the male-specific enhanced recombination in sub-telomeric regions (Broman *et al*, 1998; Mohrenweiser *et al.*, 1998; Kong *et al.*, 2002) and chiasma interference (above, section 1.1.3.b.iii) and at a local level through properties of local sequence and chromatin structure.

Although the concordance of LD breakdown and the locations of sperm crossover hotspots strongly indicates that the same hotspots exist in females, little can actually be observed of the female recombination rates as the high-resolution analysis of recombination is limited to sperm DNA. This impedes quantitative analysis of the relationship between crossover distribution and the levels of LD. Therefore, analysing regions known to be proficient in male recombination ought to facilitate the analysis of recombination processes. Consequently, the high resolution analysis of human meiotic recombination has been extended to the pseudoautosomal region at the tips of the short arms of the X and Y chromosomes, and it is this that marks the basis of the work in presented in this thesis

1.3: The Pseudoautosomal Regions

The human X and Y chromosomes are morphologically and genetically distinct but they do share two terminal regions of homology whose sequences can crossover and be inherited as if they were autosomal (Cooke *et al.*, 1985; Simmler *et al.*, 1985; Freije *et al.*, 1992). These areas are known as pseudoautosomal regions (PARs).

1.3.1: A Brief Overview of PAR2

PAR2 is only 320 kb in length and is at the ends of the long arms of the X and Y chromosomes (Freije *et al.*, 1992). Recently the entire region has been sequenced revealing two distinct subregions, as defined by their GC content (Ciccodicola *et al.*, 2000). The proximal PAR2 subregion (zone 1) stretches over 295 kb, has an average GC content of only 34.5% and is very rich in Alu and LINE sequences, which constitute 67% of the nucleotides. Its two known genes, HSPRY3 and SYBL1 are inactivated on both the Y and inactive X chromosomes (D'Esposito *et al.*, 1996; Ciccodicola *et al.*, 2000). In contrast, the distal 35 kb (zone 2) has a GC content of 51% and only 29% of its sequence consists of Alus and LINEs, though the density of Alu sequences in zone 2 is much greater than it is in zone 1. Furthermore, the two genes in the distal subregion, IL9R and CXYorf1, show the biallelic expression that is characteristic of autosomal genes (Vermeesch *et al.*, 1997; Ciccodicola *et al.*, 2000). The rate of recombination across the whole of PAR2 is six-fold higher than genome average (Li and Hamer, 1995), but much less than the rate across the Xp/Yp PAR (PAR1), where recombination occurs at an average rate of 20 cM/Mb, approximately twenty times faster than the genome average.

1.3.2: PAR1

1.3.2.a: The Structure of PAR1

PAR1 is at the tips of the short arms, is 2.6 Mb long (Brown, 1988; Petit *et al.*, 1988) and is bounded at its proximal end by a Y-specific Alu sequence, followed by a 220 bp region of 78% homology (Ellis *et al.*, 1989). PAR1 contains thirteen known genes; *PGPL*, *PPP2R3B*, *SHOX*, *XE7*, *CSF2RA*, *IL3RA*, *CRLF2*, *SLC25A6* (*ANT3*), *ASMTL*, *DHRSXY*, *ASMT*, *ALTE* (*TRAMP*) and *CD99* (*MIC2*) plus the 5′ region of a fourteenth, *XG* (*PBDX*) (Graves *et al.*, 1998, Ried *et al.*, 1998; Esposito *et al.*, 1999; Schiebel *et al.*, 2000; Gianfrancesco *et al.*, 2001, Tonozuka *et al.*, 2001) (figure 1.8a). There is a great deal of evidence to suggest that

Figure 1.8: PAR1 shown with the results of LD and sperm crossover analysis in the *SHOX* region (May *et al.*, 2002).

(a) PAR1, at the tips of the short arms of the X and Y chromosomes, is 2.6 Mb long and contains thirteen known genes plus the 5′ region of a fourteenth. (b) High resolution anlysis of PAR1 has initially been targeted to three regions; first, the telomere-adjacent 1.5 kb, which is SNP-dense and appears to be in very high LD. Second, the region around the *SHOX* gene, approximately 500 kb from the telomere and third, the *PGPL* region, approximately 80 kb from the telomere. (c) Analysis of a 43 kb region within the *SHOX* region initially revealed an extremely rapid decay of LD with physical distance. Sperm crossover analysis was still able to show that crossovers were not randomly distributed but clustered into the most active hotspot yet defined by sperm typing.

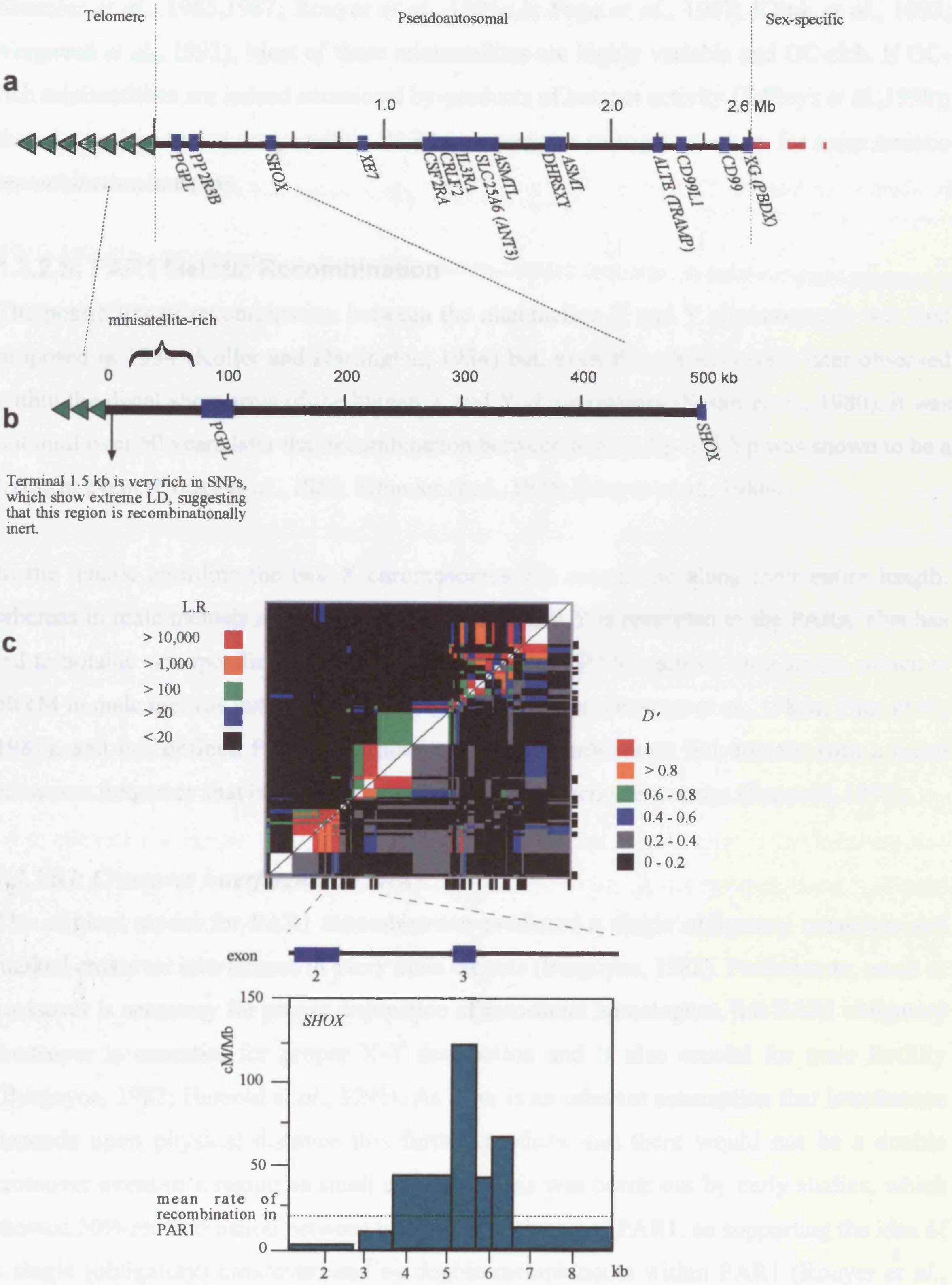


Figure 1.8: PAR1 shown with the results of LD and sperm crossover analysis in the *SHOX* region (May *et al.*, 2002).

PAR1 contains a particularly high density of minisatellite sequences (Cooke *et al.*, 1985; Simmler *et al.*, 1985,1987; Rouyer *et al.*, 1986*a*,*b*; Page *et al.*, 1987; Klink *et al.*, 1993; Vergnaud *et al.*, 1993). Most of these minisatellites are highly variable and GC-rich. If GC-rich minisatellites are indeed occasional by-products of hotspot activity (Jeffreys *et al.*,1998); then the tandem repeat arrays within PAR1 may serve as surrogate markers for some meiotic recombination hotspots.

1.3.2.b: PAR1 Meiotic Recombination

The possibility of recombination between the mammalian X and Y chromosomes was first proposed in 1934 (Koller and Darlington, 1934) but, even though RNs were later observed within the distal short arms of the human X and Y chromosomes (Solari *et al.*, 1980), it was not until over 50 years later that recombination between human Xp and Yp was shown to be a frequent event (Cooke *et al.*, 1985; Simmler *et al.*, 1985; Rouyer *et al.*, 1986*a*).

In the female germline the two X chromosomes can recombine along their entire length, whereas in male meiosis recombination between X and Y is restricted to the PARs. This has led to notable sex-specific differences in estimations of PAR1 genetic map length, which is 50 cM in male meiosis but only 4–18 cM in female meiosis (Rouyer *et al.*, 1986*a*; Page *et al.*, 1987), and has defined PAR1 as a male-specific recombination hot domain with a mean crossover frequency that is twenty times higher than the genome average (Rappold, 1993).

1.3.2.b.i: Crossover Interference in PAR1

The original model for PAR1 recombination predicted a single obligatory crossover and marked crossover interference in every male meiosis (Burgoyne, 1982). Furthermore, much as crossover is necessary for proper disjunction of autosomal homologues, this PAR1 obligatory crossover is essential for proper X-Y disjunction and is also crucial for male fertility (Burgoyne, 1982; Hassold *et al.*, 1991). As there is an inherent assumption that interference depends upon physical distance this further predicts that there would not be a double crossover event in a region as small as PAR1. This was borne out by early studies, which showed 50% recombination between markers at each end of PAR1, so supporting the idea of a single (obligatory) crossover, and no double recombinants within PAR1 (Rouyer *et al.*, 1986*a*, 1986*b*; Page *et al.*, 1987). However, double crossovers have since been detected in PAR1 (Rappold *et al.*, 1994; Lien *et al.*, 2001) raising the possibility that, at least in this region, interference is a function of genetic distance rather than physical distance.

1.3.2.c: Targeted High Resolution Analysis of Recombination in PAR1

Current low resolution data from families and from typing single sperm (Lien *et al.*, 2000) suggest that male crossovers are fairly randomly distributed across PAR1, with only modest regional variation in recombination efficiency. With very little known about the fine-scale distribution of crossovers within this region, coupled with a lack of primary sequence data, high resolution analysis was initially targeted to three regions (figure 1.8b) and has suggested a rather more complex picture of recombination.

1.3.2.c.i: SHOX Gene

To see whether hotspotting as seen in the MHC II region also occurs in PAR1, the *SHOX* gene, located 500 kb from the telomere (Rao *et al.*, 1997), was selected as it is one of the few PAR1 regions for which genomic sequence is available. Analysis of SNPs in a 43 kb interval around SHOX revealed an extremely rapid decay of LD with physical distance, with significant LD extending only a few kilobases at most, as expected for a region very active in (male) recombination (figure 1.8*c*). However, sperm crossover analysis in a 9.9 kb region that showed the strongest evidence for LD block structure accompanied by abrupt LD breakdown, demonstrated that crossovers were not randomly distributed. Instead, they clustered into a highly localised hotspot about 2 kb wide flanked by recombinationally much less active DNA (May *et al.*, 2002) (figure 1.8*c*). This hotspot shows a peak activity of 250 cM/Mb, making it the most active hotspot yet defined by sperm typing but, like the autosomal hotspots, nearly all crossovers are simple and exchanges are fully reciprocal. This suggests that hotspots also play a role in crossover distribution in PAR1 but, even outside the hotspot, there is a rapid decay of LD with distance preventing the use of LD breakdown as a method to localise putative hotspots, at least around the *SHOX* gene.

1.3.2.c.ii: Telomere-Adjacent Region

The second PAR1 interval analysed was a 1.5 kb region immediately adjacent to the telomere repeat array. This region has an extremely high density of SNPs but shows intense LD, with very few diverged haplotypes in human populations (Baird *et al.*, 1995). This implies that this region is recombinationally inert and that male crossover does not extend across the entire PAR1 but must instead terminate at a currently unidentified boundary proximal to the telomere.

1.3.2.c.iii: PGPL Region

To test whether recombination suppression extends further into PAR1, the PGPL gene, located approximately 80 kb from the telomere (Gianfrancesco *et al.*, 1998) was selected as the third target. The extensive work on the PGPL region is presented in this thesis.

1.4: Overview of this Thesis

In yeast, initiating DSBs can be readily detected biochemically and provide a surrogate physical measure of genetic recombination activity applicable across the entire genome (Gerton *et al.*, 2000). Unfortunately, an equivalent measure remains to be identified in the human genome and it is not possible to scale sperm typing (figure 1.5) to the genome level. Thus, despite there being a great deal of evidence to suggest that meiotic recombination hotspots are a common feature of the human genome, only those in the MHC II region, at *SHOX* and at the MS32 minisatellite have been identified at the molecular level. Therefore, the overall aim of this thesis is determine if the emerging pattern of hotspots that break the genome into a series of discrete haplotype blocks is also a feature of recombination around the *PGPL* gene. In addition, observations elsewhere in PAR1 provided at least three other questions that this work will aim to answer. First, analysis in the PAR1 telomere-adjacent region and in the *PPP2R3B* gene had suggested that PAR1 as a whole has a very high level of nucleotide variability (Baird *et al.*, 1995; Schiebel *et al.*, 2000). Consequently, analysing DNA diversity in the *PGPL* region will help to establish if the genome-average level of variability seen at *SHOX* (May *et al.*, 2002) is actually a more accurate reflection of diversity in PAR1. Second, LD and sperm crossover analysis in the *PGPL* region will ask if the observations at *SHOX*, where LD decays very rapidly with physical distance but crossovers still cluster into a hotspot, are applicable to the rest of PAR1. Third, the extreme LD in the telomere-adjacent region (Baird *et al.*, 1995) suggested that there is a distal region of PAR1 where recombination shuts down. Thus, identifying the rate of recombination in the *PGPL* region, located approximately 80 kb from the telomere (Gianfrancesco *et al.*, 1998), might narrow down the putative interval of recombinational inactivity.

1.4.1: Genomic Sequence in the *PGPL* Region

PAR1 was neglected by the Human Genome Consortium until very recently (June 2003). In fact, the limited amount of genomic sequence data that existed at the beginning of my work was restricted to the *SHOX* region. The *PGPL* gene was selected for analysis for two main reasons. First, its cDNA sequence had already been published (Gianfrancesco *et al.*, 1998) and second, there was already a handle on the region because it had been mapped to a cosmid that was part of a contig covering the distal 750 kb of PAR1 (Rao *et al.*, 1997; Gianfrancesco *et al.*, 1998). Hence, Chapter 3 describes the strategies that were employed to identify and analyse the sequence of the cosmid reported to contain *PGPL*. The long-awaited production of PAR1 sequence by the Human Genome Sequencing Consortium then allowed the genomic characterisation of a 100 kb interval between the *PGPL* gene and the PAR1 telomere. This work is described in Chapter 4 and includes the first accurate physical map of this region.

The genomic sequence around *PGPL* not only revealed a very high density of both interspersed and tandem repeats but also led to the identification of a novel gene. This provided an interesting side project to the main aims of my work, and has therefore been included in this thesis (Chapter 5).

1.4.2: DNA Diversity and Recombination in the *PGPL* Region

Identifying and analysing the genomic sequence around the *PGPL* gene provided the primary tool to answer the questions posed at the start of this section. Chapter 6 describes both the determination of nucleotide variability in the region of *PGPL* and the preliminary analysis of the instability of three novel tandem repeat arrays. This latter study was performed as it has been suggested that unstable minisatellites can serve as occasional markers of local meiotic recombination hotspots (Jeffreys *et al.*, 1998*a*).

Chapters 7 and 8 describe the high resolution analysis of recombination in the *PGPL* region. Not only does this analysis provide further insight into the putative distal PAR1 boundary of meiotic recombination, it also reveals a completely unexpected distribution of crossover events, which is considered at length both in Chapter 8 and in Chapter 9, the final discussion.

Chapter 2
Materials and Methods

2.1: Materials

2.1.1: Chemical reagents

Chemicals were supplied by Fisher Scientific (Loughborough, UK), Flowgen (Ashby de la Zouch, UK), FMC Bioproducts (Rockland, USA), and Sigma Biochemical Company (Poole, UK). Molecular biology reagents were obtained from ABgene (Epsom, UK), Ambion, Inc. (Austin, USA), Amersham Biosciences (Little Chalfont, UK), Applied Biosystems (Warrington, UK), Bio-Rad (Hemel Hemstead, UK), Invitrogen UK (Paisley, UK), Millipore (Watford, UK), NEN Life Sciences (Divison of PerkinElmer Life Sciences Ltd, Cambridge, UK), New England Biolabs (Hitchin, UK), Qiagen Ltd. (Crawley, UK), ResGen (Division of Invitrogen Ltd, Paisley, UK), Sigma-Aldrich Company Ltd (Poole, UK), Stratagene (Amsterdam, The Netherlands) and United States Biochemical Corp (USB) (Cleveland, USA)

2.1.2: Specialised Equipment

Specialised equipment was obtained from Bio-Rad (Hemel Hempstead, UK), Cecil Instruments (Cambridge, UK), Eppendorf (Hamburg, Germany), Fisher Scientific (Loughborough, UK), Genetic Research Instrumentation (GRI) (Braintree, UK), Helena Biosciences (Sunderland, UK), Heraeus Instruments (Hanau, Germany), Hybaid (Teddington, UK), MJ Research (Waltham, USA), Applied Biosystems (Warrington, UK), Thermo Shandon (Pittsburgh, USA), and Ultra Violet Products (UVP) Life Sciences (Cambridge, UK).

2.1.3: Oligonucleotides

Oligonucleotides for PCR amplification and ASO hybridisation were synthesised by the Protein and Nucleic Acid Chemistry Laboratory (PNACL), University of Leicester, UK. Hexadeoxyribonucleotides for random primed labelling were obtained from Pharmacia.

2.1.4: Enzymes

Restriction enzymes were supplied by Gibco-BRL, New England Biolabs and Boehringer Mannheim. T4 ligase, *Pfu* polymerase, calf intestinal alkaline phosphatase and React^{TM} buffers were obtained from Gibco-BRL. The Klenow fragment of DNA polymerase I of *E. coli* was supplied by Pharmacia. *Taq* polymerase was obtained from Advanced Biotechnologies. RNase and Proteinase K were obtained from Sigma and T4 polynucleotide kinase was supplied by New England Biolabs.

2.1.5: Molecular weight markers

1 kb ladder was supplied by Gibco-BRL. λ DNA digested with *Hin*dIII and ϕ X174 DNA cut with *Hae*III were supplied by ABgene.

2.1.6: Bacterial strains

Escherichia coli strain XL1-Blue MRF′ (supplied by Stratagene) was used in all cloning experiments. XL1-Blue is both recombination deficient (recA) and endonuclease deficient (endA1). These properties serve to increase the stability of inserts and improve the quality of purified plasmid DNA respectively. The hsdR mutation prevents the cleavage of cloned DNA by the EcoK (hsdR) endonuclease system and the lacI^{q}ZΔM15 gene on the F′ episome allows blue-white screening of recombinant vectors by α-complementation of β-galactosidase.

XL1-Blue MRF′ genotype: Δ(*mcrA*) *183* Δ(*mcrCB-hsdSMR-mrr*) *173 endA1 supE44 thi-1 recA1 gyrA96 relA1 lac* [F′ *proAB lacI* ^{q}ZΔ*M15*Tn*10*(Tet^{r})]

2.1.7: Cosmid DNA

Gudrun Rappold (Heidelberg University) provided three overlapping cosmids from the distal end of a cosmid contig that covers the terminal 750 kb of PAR1 (Rao *et al.*, 1997).

LLN0YCO3′M′29C1; LLN0YCO3′M′3F3; LLNLN0434

2.1.8: Cloning vector

DNA fragments were cloned into the *Eco*RV polylinker site of the pBluescript II SK$^+$ vector (supplied by Stratagene).

2.1.9: Human DNAs

Genomic DNA from EBV-transformed lymphoblastoid cell lines was supplied by the Centre d'Etude du Polymorphisme Humain (CEPH, Paris, France). Semen samples from anonymous north European donors of UK origin were supplied by J. Blower (Leicester Royal Infirmary, Leicester, UK). Additional semen samples were provided by members of the Department of Genetics (University of Leicester, UK). All studies were granted Local Ethical Committee approval.

2.1.10: Standard solutions

Southern blot solutions (depurinating solution, denaturing solution, and neutralising solution), 20 x Sodium 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 (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, Department of Genetics, University of Leicester, UK. SOC broth was as described by Sambrook *et al.* (1989).

2.1.11: Computers

This thesis was produced using a Power Macintosh G3 Minitower and an Epson Perfection 1250 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, Factura, Freehand, Microsoft Word, Microsoft Excel, and Microsoft Powerpoint all for Macintosh computers. Internet searches were performed using Microsoft Internet Explorer. All computer analyses were performed on Apple Macintosh computers.

2.2: Methods

The methods used during the course of this work are described below as general overviews of techniques. Where appropriate, detailed descriptions of novel or modified techniques are described, in context, in the Results chapters. Other techniques are very well described elsewhere and are referenced as such in the relevant Results chapters.

2.2.1: Selection and growth of bacterial cultures

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

2.2.1.a: Preparation of glycerol stocks

Each cosmid preparation provided by Gudrun Rappold was streaked on Luria-Bertani agar (LUA) containing kanamycin at 50 μg/ml and incubated for 16 hours at 37°C. Four individual clones of each cosmid were transferred to 3 ml of Luria-Bertani broth (LUB) containing kanamycin at 50 μg/ml and grown for 16 hours at 37°C with shaking at 200 rpm. A 0.5 ml aliquot of overnight LUB bacterial culture was mixed with 0.5 ml of sterile 60:40 (v/v) glycerol/LUB solution and stored at -80°C.

2.2.2: Agarose gel electrophoresis

2.2.2.a: Electrophoresis conditions

Unless stated otherwise, agarose gel electrophoresis was carried out using 1% (w/v) LE (SeaKem^{TM}) agarose gels in 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 the GeneGenius analysis system (Syngene). The Dark Reader System (Clare Chemical Research) was used for band excision.

2.2.2.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 Mitsubishi video copy processor (Genetic Research Instrumentation) with camera (UVP Life Sciences) or by use of the GeneGenius analysis system and a Sony digital graphic printer (Syngene).

2.2.3: Methods of DNA preparation

DNA was extracted from venous bloods and semen samples under PCR clean conditions using phenol/chloroform extraction followed by ethanol precipitation, as described elsewhere (Jeffreys *et al.*, 1990). DNA extractions from both sperm and blood were performed in a category II laminar flow hood under conditions designed to minimise contamination. Unless otherwise stated all centrifugation steps in this and subsequent sections were performed in either a Heraeus Septatech Biofuge 15 centrifuge or an Eppendorf 5415 D centrifuge.

2.2.3.a: DNA extraction from human sperm

Frozen semen samples were thawed on ice for 1 hour. 250 μl of semen was transferred to a 1.5ml screw top tube and diluted to 1ml with 1 x SSC. 20 μl 1% (w/v) SDS was added, mixed by flicking in order to lyse somatic cells including epithelial cells (sperm heads are not lysed by SDS alone) and centrifuged for 2 minutes. The supernatant was removed and lysis was repeated once by vigorous resuspension in 1 ml 1 x SSC, 0.2% (w/v) SDS. The pellet was washed in 1 ml 1 x SSC to remove any trace of SDS and non-sperm DNA and resuspended fully in a hypotonic solution of 450 μl 0.2 xSSC to aid lysis. Sperm heads were lysed by addition of SDS to a final concentration of 1% (w/v) and 2-mercaptoethanol (final concentration of 1 M) and incubated at room temperature for 5 minutes. Proteinase K was added to a final concentration of 200μg/ml and incubated for 1 hour at 37°C with occasional mixing. Proteins were removed by addition of 350 μl phenol/chloroform with gentle mixing to allow emulsification and centrifuged again for 2 minutes. The organic layer was re-extracted once with 1 x SSC and 0.2% (w/v) SDS. DNA was ethanol precipitated using 2 volumes of 100% ethanol and gentle swirling. The supernatant was removed and the pellet washed with 1 ml 80% (v/v) ethanol and gently mixed. The pellet was dissolved in 90 μl of distilled water; 10 μl 2 M sodium acetate (pH 7.0) was added followed by 200 μl 100% ethanol and centrifuged for 1 minute. The supernatant was removed, the pellet air dried and dissolved at 4°C in 50 μl 5 mM Tris-HCl (pH 7.5).

2.2.3.b: DNA extraction from human blood

Venous blood samples (delivered into equal volumes of 1 x SSC and stored at -80°C), were thawed at 37°C and 500 µl transferred into a 1.5ml screw top tube. 800 µl 1 x SSC was added and gently mixed before centrifugation for 2 minutes. The supernatant, including haemoglobin from lysed erythrocytes, was removed and the pellet washed twice in 1 ml 1 x SSC. The pellet was resuspended in 300 µl 0.2 x SSC and cells were lysed by adding 30 µl 10% (w/v) SDS and incubated at room temperature for 5 minutes. Proteinase K was added to a final concentration of 200 µg/ml and incubated for 1 hour at 37°C with occasional mixing. Trace proteins were removed by addition of phenol/chloroform with gentle mixing to allow emulsification and centrifuged for 3 minutes. The organic layer was re-extracted twice with 1 x SSC and 0.2% (w/v) SDS. DNA was ethanol precipitated as above and dissolved in 20 µl 5 mM Tris-HCl (pH 7.5).

2.2.3.c: Extraction of cosmid DNA

An isolated positive colony or 5–10 µl aliquot of glycerol stock, containing an appropriate recombinant cosmid, was cultured in 50 ml LUB with 50 µl/ml kanamycin at 37°C overnight. Next, 25 ml aliquots of the overnight culture were centifuged at 13000 rpm for five minutes at 4°C. The cells from each aliquot were resuspended in 2 ml cosmid lysis solution (25 mM Tris-HCl; 10 mM EDTA, pH 8.0; 50 mM glucose and freshly added 2 mg/ml lysozyme) and left for five minutes at room temperature. The cells were then lysed, and the DNA denatured, with 4 ml 0.2 M NaOH, 1% (w/v) SDS. The mixture was left on ice for 4 minutes before adding 3 ml 3 M potassium acetate, pH 4.8 to neutralise and precipitate the chromosomal DNA and most of the proteins. After 8 minutes on ice the precipitate was pelleted by centrifugation as above, the supernatant collected and residual protein was precipitated by addition of equal volumes of phenol: chloroform: isoamyl alcohol (25:24:1). The DNA was collected by ethanol precipitation and dissolved in 5 mM Tris-HCl, pH 7.5. To prepare DNA for shotgun library construction (see below), a large-scale version of this method was used in which a 40 µl aliquot of glycerol stock was cultured in 200 ml LUB with 50 µl/ml kanamycin at 37°C overnight, before the overnight culture was split into 25 ml aliquots and the extraction was completed as above. Where high quality cosmid DNA was required, it was prepared using a QIAGEN Plasmid Maxi Kit, which was employed according to the manufacturer's instructions.

2.2.4: Methods of DNA purification

2.2.4.a: Ethanol precipitation

Unless stated otherwise, double-stranded DNA was precipitated in microcentrifuge tubes by addition of 1/10 volume 2 M sodium acetate 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. The pellet was washed in 80% (v/v) ethanol and allowed to air dry. Single-stranded DNA was precipitated in microcentrifuge tubes by addition of 1/10 volume 2 M sodium acetate 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. The pellet was washed in 80% (v/v) ethanol and allowed to air dry. Oligonucleotides (section 2.1.3) were precipitated in microcentrifuge tubes by addition of 1/10 volumes 2 M Sodium acetate pH 7, and 3 volumes 100% ethanol followed by incubation at -80°C for 30 min and centrifugation for 30 min at 15000 rpm. The pellet was washed in 80% ethanol, allowed to air dry and redissolved in 50 μl of PCR-clean water.

2.2.4.b: Electroelution of DNA from agarose gels

Following electrophoresis under normal conditions (Sambrook *et al.*, 1989),the fragment of interest was excised from the agarose gel using the Dark Reader system (Clare Chemical Research) for visualisation. The Dark Reader emits light only in the blue spectrum so there is no UV light that would damage DNA samples. The agarose bands excised from gels were 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 Dark Reader 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. DNA was recovered from the eluate by ethanol precipitation.

2.2.5: DNA transfer to membrane

2.2.5.a: Southern blot

Following electrophoresis, the region of agarose gel required for DNA transfer was excised. The gel was partially depurinated in 0.25 M HCl for 2 x 5 minutes (depurinated DNA is cleaved more readily by NaOH), alkali-denatured in 0.5 M NaOH, 1 M NaCl for 8 minutes and 10 minutes (to cleave DNA into smaller fragments), and neutralised in 0.5 M Tris-HCl pH 7.5, 3 M NaCl for 8 minutes and 10 minutes. DNA was transferred to Hybond^{TM}-N^{fp} (Amersham International Plc.) nylon membrane (pre-soaked in 10xSSC) by the capillary transfer method using 20 x SSC as the transfer buffer (Southern blotting (Southern, 1975)). Blotting was continued for 1-8 hours, depending on the agarose gel concentration. The membrane was washed in 2 x SSC, dried at 80°C for 15 min, and the DNA covalently linked to the membrane by exposure to 7x10^4 J/cm^2 of UV light in the RPN 2500 ultraviolet crosslinker (Amersham).

2.2.5.b: Dot blot

PCR amplification was performed so that each spot would contain the equivalent of 3–100 ng DNA per 1 kb PCR product. Each reaction was then mixed with 0.25 volumes of dotblot loading mix (30% (v/v) glycerol, 0.5 x TBE and a trace of bromophenol blue – sufficient to give a blue spot on the dotblot for identification of spot locations during loading) followed by at least five volumes of denaturing mix (0.5 M NaOH, 2 M NaCl, 25 mM EDTA) in order to obtain at least 30 μl of solution per replica filter. A vacuum was applied to the assembled dotblot manifold harbouring one sheet of Hybond^{TM}-N^{fp} (Amersham International Plc.) nylon membrane, plus two additional backing sheets of 3MM Whatman chromatography paper (all pre-soaked in 2 x SSC). At leat 30 μl of denatured DNA was loaded into each well and, once all samples were loaded, each well was washed with 150 μl 2 x SSC to neutralise the DNA. The dotblots were dried for 10 minutes at 80°C and the DNA covalently linked to the membrane by UV exposure as above.

2.2.6: Hybridisation

2.2.6.a: Southern blot hybridisation

Double stranded DNA (10 ng) generated (unless stated otherwise) by PCR amplification of the locus of interest was added to 2 ng of of λ DNA x *Hin*dIII and ϕX174 RF DNA x *Hae*III ladders and labelled by the random primed labelling reaction (Feinberg and Vogelstein, 1983; Feinberg and Vogelstein, 1984) which involves the use of random-sequence hexamers and the *E. coli* DNA polymerase I Klenow fragment to incorporate α-^{32}P-dCTP (supplied by Amersham International Plc.) into the DNA. Labelling reactions were performed in 15 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% (w/v) SDS, 0.5 M sodium phosphate 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 4 changes of 40 mM sodium phosphate pH 7.2, 1 % (w/v) SDS and then in high stringency wash solution (0.1xSSC, 0.01% (w/v) SDS) at 65°C for fifteen minutes. The membranes were blot-dried and exposed to Fuji RX100 X-ray film, either at room temperature for strong signals, or at -80°C with an intensifying screen, in order to visualise the hybridisation pattern.

2.2.6.b: Dot blot hybridisation

This method uses a TMAC (tetramethylammonium chloride, Sigma Biochemical Company) hybridisation protocol which essentially eliminates the dependence of allele specific oligonucleotide (ASO) melting temperature on base composition, thereby allowing all ASOs of equal length to hybridise at the same temperature. ASOs were end-labelled for one hour at 37°C in 10 μl reaction volumes consisting of 8 ng ASO, 1 μl 10 x kinase buffer (700 mM Tris-HCl (pH 7.5), 100 mM $MgCl_2$, 50 mM spermidine trichloride, 20mM dithiothreitol), 3.5 units bacteriophage T4 polynucleotide kinase and 2 μCi γ-^{32}P-ATP. Following incubation, 20μl of kinase stop solution (25mM diNa EDTA, 0.1% (w/v) SDS, 10μM ATP) together with 20-fold excess (160ng) of unlabelled ASO for the other allele as competitor were added to the [γ -^{32}P]-labelled ASO and the probe was used immediately for hybridisation. Dotblot membranes were soaked briefly in 3 x SSC and pre-hybridised in 3 ml hybridisation solution (final concentrations: 3 M TMAC, 0.6% (w/v) SDS, 1 mM diNa EDTA (pH 8.0), 10 mM sodium phosphate (pH 6.8), 5 x Denhardt's solution (50 x Denhardt's solution: 1% (w/v) Ficoll 400, 1% (w/v) polyvinylpyrrolidone, 1% (w/v) BSA), 4 μg/ml yeast RNA) at 53°C for 5 minutes. The membranes were then incubated in 2.5 ml fresh hybridisation solution together with 7 μl 3 mg/ml single stranded (heat-denatured) herring sperm DNA competitor at 53°C for a further 2–10 minutes. The [γ-^{32}P]-labelled ASO/unlabelled competitor ASO mix was then added to the hybridisation solution and hybridised at 53°C for 1 hour. Following hybridisation, the membranes were washed in 3 changes of 2–3 ml TMAC wash solution (final concentrations: 3M TMAC, 0.6% (w/v) SDS, 1mM diNa EDTA pH 8.0, 10 mM sodium phosphate pH 6.8) at 56°C (5–20 minutes washing in total), followed by 4 ml TMAC wash solution at 56°C for a 15 minutes. The membranes were then rinsed twice in 3 x SSC at room temperature, blot-dried and exposed to Fuji RX100 X-ray film as above.

2.2.6.c: Removing probe for re-use of membranes

The membranes were washed in 2–4 changes of boiling 1 % (w/v) SDS (8–9 changes for dotblots) until monitoring of the membranes with a Geiger counter demonstrated complete probe removal.The membranes were then rinsed in 2 x SSC at room temperature and stored damp at 4°C or used directly for re-hybridisation as above.

2.2.7: 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.

2.2.8: DNA amplification

2.2.8.a: PCR Buffer

11.1 x PCR buffer (Jeffreys *et al.*, 1990) was prepared by R. Neumann (Department of Genetics, Leicester) as indicated below. 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
$MgCl_2$	1 M	33.5	4.5 mM
2-mercaptoethanol	100%	3.6	6.7 mM
EDTA pH 8.0	10 mM	3.4	4.4μM
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 μg/ml
Total Volume		676	

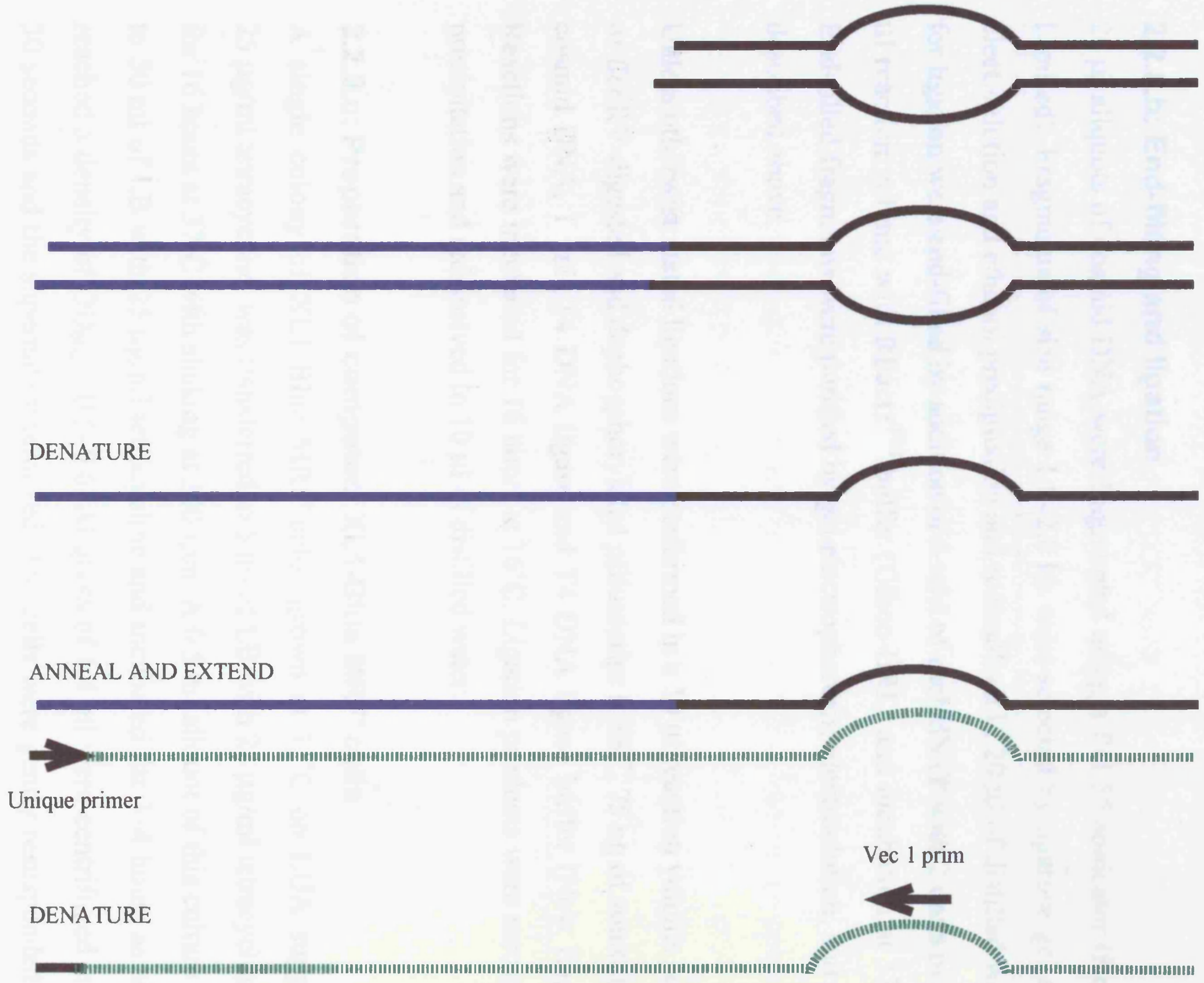
2.2.8.b: PCR conditions

2.2.8.b.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 amplifications were performed, unless stated otherwise, in 10 μl reactions containing 0.9 μl of 11.1 x PCR buffer as described above, 0.25 μM of each primer and 0.03 U/μl *Taq* polymerase under normal amplification conditions. For amplification of DNA longer than 2 kb normal PCR conditions were supplemented with 0.0035 U/µl *Pfu* polymerase. *Pfu* has 3′ to 5′ exonuclease activity which can edit regions of depurinated DNA by removing base mismatches that would otherwise cause *Taq* polymerase to stall, resulting in incomplete products and allowing jumping PCR to occur. Where indicated 1 µg/ml carrier herring sperm DNA was added to the reaction to preferentially coat the side of PCR tubes and prevent the target DNA being sequestered in this way. This was particularly important when using very low inputs (<1 ng) of target DNA. 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 and primer sequences are listed in the tables at the end of this chapter. Details of the templates used in the PCRs are included in the tables or appropriate chapters, where relevant.

2.2.8.b.ii: Vectorette PCR

The aim of vectorette PCR (vectorette walking) is to walk from a known sequence of DNA to an unknown, adjacent section using PCR. To do this vectorette linkers are ligated onto partially digested DNA, which is then amplified using a unique primer from the known sequence and a vectorette primer (figure 2.1). Vectorette PCR was used during the work presented in Chapter 3 of this thesis; N0434.1 cosmid DNA, in amounts of 1 µg, was digested with 1.0, 0.5, 0.25, 01.25 or 0.0625 units of *Hae*III or *Rsa*I restriction enzymes, for 30 minutes in 10 µl reaction volumes. The extent of the digests was then checked by agarose gel electrophoresis and the DNA from the tubes in which the digests were judged to have produced the best 'smear' (i.e. the digests were clearly partial) was purified by phenol/chloroform extraction and ethanol precipitation. Vectorette oligonucleotides were annealed to each other through the addition of 5 µg of each oligonucleotide to 1 µl of 10x annealing mix (1 M NaCl; 0.1 M $MgCl_2$; 0.1 M Tris-HCl, pH7.5) in a 10 µl reaction volume and incubation at 60°C for 30 minutes. The final concentration of the annealed oligonucleotides (vectorette linkers) was adjusted to 200 ng/µl and they were stored at –20°C. Partially digested DNA (0.2 µg) was ligated to the vectorette linkers (0.2 µg) at 15°C for 16 hours in a final volume of 20 µl using T4 ligase and its buffer as recommended by the supplier (New England Biolabs). The completed reactions were diluted with 80 µl of ddH_2O and stored at –20°C. This vectorette library was then used to seed PCR amplifications, which were subjected to electrophoresis and electroelution in order to isolate the amplicon that contained unknown sequence.



1: Anneal vectorette oligonucleotides; they have a region of non-homology resulting in a 20 bp single-stranded segment.

2: Ligation of vectorette linkers to digested DNA fragments, which creates a vectorette library.

3: PCR amplification using Vec 1 prim and a unique primer from known DNA sequence.

Vec 1 prim is identical to one of the single strands in the vectorette linker and can only prime once a copy of the vectorette sequence has been made. Thus the whole system relies on the specificity of the unique primer

Figure 2.1: Vectorette PCR

2.2.9: Preparation of shotgun library

2.2.9.a: Digestion and dephosphorylation of vector

20 μg of uncut pBluescript II SK$^+$ vector was digested with 50 units of *Eco*RV in REact^{TM} 2 buffer for 2 hours at 37°C, and dephosphorylated by the addition of 50 units of calf intestinal alkaline phosphatase (CIAP, Gibco) with incubation for 30 minutes at 37°C. The reaction was stopped by addition of SDS and EDTA to concentrations of 0.5% (w/v) and 5 mM respectively, followed by incubation at 65°C for 20 minutes for enzyme denaturation. The vector was collected by ethanol precipitation and redissolved in 50 μl TE (5 mM Tris-HCl, pH7.5; 0.1 mM EDTA). Concentration of the stock solution was assayed by measuring A_{260} using a Cecil Instruments 2040 UV spectrophotometer, and adjusted to 100 μg/ml.

2.2.9.b: End-filling and ligation

20 μl aliquots of cosmid DNA were fragmented using a Pul 55 sonicator (Kerry Ultrasonics Limited). Fragments of size range 1.0–2.0 kb were selected by agarose gel electrophoresis, electroelution and ethanol precipitation and redissolved in 20 μl of distilled water. 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^{TM} 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.

Unless otherwise stated ligations were performed in a 10 μl reaction volume containing 7.5 ng of *Eco*RV-digested and dephosphorylated pBluescript II SK$^+$, 75 ng of sonicated and repaired cosmid DNA, 1 unit T4 DNA ligase and T4 DNA ligase buffer (New England Biolabs). Reactions were incubated for 16 hours at 16°C. Ligation products were recovered by ethanol precipitation and redissolved in 10 μl of distilled water.

2.2.9.c: Preparation of competent XL1-Blue MRF' cells

A single colony of XL1-Blue MRF' 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 hours at 37°C with shaking at 300 rpm. A 0.5 ml aliquot of this culture was transferred to 50 ml of LB with 25 μg/ml tetracycline and incubated for 3–4 hours as above until cells reached a density of OD_{600} = 0.5–0.6. Aliquots of 1.4 ml were centrifuged at 13000 rpm for 30 seconds and the supernatant removed. The cells were gently resuspended in 0.5 ml fresh MR solution (10 mM MOPS pH 7.0, 10mM RbCl) and centrifuged for 30 seconds at

13000 rpm. The supernatant was removed, the cells were resuspended in 0.5 ml fresh MRC solution (100 mM MOPS pH6.5, 10 mM RbCl, 50 mM $CaCl_2$) and left on ice for 30 minutes. The cells were then centrifuged at 13000 rpm for 30 seconds before the supernatant was removed and the cells were reuspended in 0.15 ml MRC. The competent cells were then kept on ice prior to their use in transformation.

2.2.9.d: Transformation and selection of transformants

5 μl of ligation mix and 3 μl of DMSO were added to the 0.15 ml MRC suspensions of competent cells, which were then left on ice for one hour. The cells were heat-shocked at 55°C for exactly 35 seconds and cooled on ice for one minute. 1 ml of SOC broth was immediately added and the cell suspension was incubated at 37°C for 45 minutes. Centrifugation at 13000 rpm for 30 seconds was followed by removal of the supernatant and the resuspension of the cells in 30 μl of SOC broth. This culture was then plated onto LUA supplemented with 60 μg/ml ampicillin, 25 μg/ml tetracycline, 40 μg/ml X-gal and 0.5 mM IPTG. Plates were incubated for 16 hr at 37°C. For each ligation, 10 white colonies and 2 blue colonies were selected, cultured in 2.5 ml LB, and DNA extracted by alkali lysis. The DNA was digested with *Pst*I and *Hin*dIII and electrophoresed to check for insert presence.

2.2.9.e: Ligation reaction and transformation controls

For each newly prepared stock of either pBluescript II SK^+ vector or competent cells, the following controls were performed.

Plate	*Control Insert*	*Digested Dephosphorylated Vector*	*Insert*	*Ligase*	*Unmodified Vector*
1	✓	✓	✗	✓	✗
2	✗	✗	✓	✓	✗
3	✗	✓	✗	✓	✗
4	✗	✓	✗	✗	✗
5	✗	✗	✗	✗	✓
6	✗	✓	✓	✓	✗
7	✗	✓	✓	✓	✗

2.2.9.e.i: Expected results from controls:

- The 1.35 kb band from the ϕx174 x *Hae*III ladder was selected as the control insert fragment. Most colonies should be white due to disruption of the LacZ operon by the insert DNA. Tests the efficiency of the ligase.
- Ought to be no colonies due to the absence of the vector and hence the amp^R gene. Tests that insert is not contaminated with vector DNA.

- Dephosphorylated vector should not re-ligate. A small number of blue colonies would be seen if dephosphorylation or digestion was incomplete.
- Linearised vector so no colonies should be present. Any colonies would indicate that uncut vector remains.
- Should produce many blue colonies and is an indicator of transformation efficiency
- Controls 6 and 7 (and onwards) vary the insert:vector ratio and should produce mostly white colonies, representing recombinants.

2.2.9.f: Picking colonies into ordered arrays

10 ml of 10 x HMFM (36 mM $K_2HPO_4.3H_20$, 13 mM $KH_2PO_4.2H_2O$, 20 mM Na citrate dihydrate, 10 mM $MgSO_4.7H_2O$, 44% (w/v) glycerol) was sterilised and added to 100 ml of LUB with 55 μg/ml ampicillin and 27.5 μg/ml tetracycline. White colonies were picked from plates and added to 100 μl of HMFM:LUB mixture in the wells of a 96-well microtitre plate. Each plate also contained an aliquot of HMFM:LUB into which no colony was added and two wells into which a blue colony was picked. The plates were incubated at 37°C for 16 hours, frozen on a bed of dry ice and stored at -80°C.

2.2.9.g: Recovery of double stranded DNA directly from shotgun library clones

Shotgun library clones were thawed (on ice) and 20 μl aliquots were mixed with 500 μl of distilled water before centrifuging at 13000 rpm for 2 minutes. The supernatant was then discarded and the pellet was resuspended in 9 μl of distilled water. The mixture was then held at 100°C for 4 minutes, centrifuged at 13000 rpm for 4 minutes and stored at 4°C. Each clone was then ready for PCR amplification using the commercially-available KS and SK primers (Stratagene).

2.2.10: Sequencing

2.2.10.a: Automated DNA sequence analysis

Following PCR amplification and agarose gel electrophoresis, the amplified DNA was recovered by electroelution onto dialysis membrane as previously described. The sample was ethanol precipitated and dissolved in 10μl of distilled water. Sequencing was then carried out using a PE Applied Biosystems Model 377 DNA Sequencing System, with the ABI PRISM BigDye^{TM} Terminator Cycle Sequencing Ready Reaction Kit, in accordance with the manufacturer's instructions. 10μl sequencing reactions were cycled at 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes for 25 cycles using ~15 ng/kb of DNA template and 3.3pM sequencing primer. Sequencing products were purified by a modified ethanol precipitation protocol: 0.1 volumes 3M sodium acetate (pH 4.6) and 2.5 volumes 95% (v/v) ethanol were added to the finished reaction, followed by a 10 minute incubation on ice. The mixture was centrifuged for 20 minutes at 13000 rpm at room temperature, the pellet was washed with 70% (v/v) ethanol, air dried and dissolved in 2μl 83% de-ionised formamide, 8.3mM EDTA, prior to loading onto the sequencing gel. Gel electrophoresis and scanning was carried out in the Protein and Nucleic Acid Laboratory (PNACL), University of Leicester. Where appropriate sequence reads were assembled using ABI Autoassembler software.

Table 2.1: Oligonucleotides used in Chapter 3 (N0434.1 sequencing). All except those marked with an asterisk were used in sequencing reactions that took place under the conditions described in section ***10.b*** of this chapter. A subset of these oligonucleotides were also used in a number of PCRs (table 2.2*a*,*b*). Note that here and elsewhere oligonucleotides are listed in 5′ to 3′ orientation.

KS	TGC AGG TCG ACG GTA TC
SK	CGC TCT AGA ACT AGT GGA TC
N1.4F	TAT GGA AAG GGT CAC TTC CC
N1.7F	GTG TGT GTG TGT GTG TAG GC
N2.3R	TTA GAC CCG GCT TGG CAA CC
N2.4F	CTT CCA AGG ACC CCT TTT CC
N2.4R	GGA AAA GGG GTC CTT GGA AG
N2.5F	GAG TTT CAC TCT TGT TGC CC
N3.1R	ATG CAT GTC TGT GCA TGT CC
N4.7F	CCC AGT GTG ACA TCC TGA GG
N5.2R	GTT AGC TGA CAC ATG TAG CC
N7.3F	TGC TTT GCT CTC AGG GAG CC
N7.9F	CCG TGT TAG TCA GGA TGG
N8.2R	ACA GGA GAA TGG CGT GAA CC
N8.4R	TGT CTG TCT TGA GTG CAG GC
N8.5R	AAC TGC AGA CCC GCC TCA GG
N8.8R	CAA CAT GTT AGC CAG GAC GG
N9.1F	CAG CTG TTG TTA CGA CAT CC
N9.1R	CCA CGG GGA CCT GCA TTT GG
N9.9R	TTA GCT GGG CGT GGT GTC CG
N10.4F	GAT CTG CTT CCC TTG ACG G
N10.8F	GCG GAT CAT GAG GTC AGG AC
N10.9R	TTG TGT TAG TAG AGA CGG G
N11.3R	ATC CCG GCT AGT CAG AAG GC
N11.8F	TGG AGA CCA TGA AGA GCT GC
N12.5F	GAG GGG AAA TGA GGA CTG AC
N12.5R	TTC CAT GAA GCA GTC AGT CC
N12.6F	AAG GGG TCT CAC TGT GTT GG
N12.8R	CCT GAC CTC ATG ATC CGC CC
N13.1R	GAT GGG ATT ACA GGT GCC
N13.3R	CCT GGT TCA AGT GAT TCC CC
N13.5F	CTC ACG CCT GTA ATC CCA CC
N13.8R	GTG GTG AAA TTC ACG TAC
N14.2F	GAG ATG GAG TCT CAC TCT TG
N14.4F	TGA GCC ACG GTG CCA GGC GG
N14.7F	TTG GCC TCC CAA ACT GGT GG
N15.1R	GAT TCC ACG GCC AGA GAC GG
MVN15.3F*	GAT GGC ACA GAC TTA GTG CC
MVN17.6F*	CAC GTC GTA ACA CCG TGT GG
N18.4F	GCT CAG GCA GGA GAA CTG C
N19.1R	GGC TTG GTG TGA GAA CTA GC
N19.4F	GTG TGT ATG CGT GTA TAT ACA C
N19.7R	TCG GCG GCT GGG CAC AGT G
N20.2F*	TCA CTG TAA CCT TCA ACG CC
RN20.7R*	GAA CAC TTC AGT AGC ACT CC
N22.4F	GGG CCT GTA GTC CCA GCT AC
N22.8R	AGA ATC TCG CTC TGT CAC CC
N22.86R	CAC GCA CAT TAG CCG TCT CC
N23.3R	GCA GCA CAG AGA ACC CTT CC
N23.5F*	GTT ACG GAA ACA TTC CGA GGG A
RN23.8R*	TGA ACG GAC GCC CAC AGA GG
N24.1F	GTT CAA GTG ATT CTC CTG CC
N24.7R	TGG TGA GAC CCC ATC TCT AC
N26.3R	CAG TGG ACT CCT GGG GAC C
N30.6F	GGG GAG ATG GTG GTG TGG AC
N31.2R	CCC CTC CAT CTA CAC CAT CC
VEC1.TOP*	GAT CAG GCT GGA GAT GTA GCA GAT TGA GAT ATT CGT TAT AGT TTA CCT ATC CCG ACC GAG CAT G
VEC1.BOTTOM*	CAT GCT CGG TCG GGA TAG GCA CTG GTC TAG AGG GTT AGG TTC CTG CTA CAT CTC CAG CCT
VEC1.BOTTOM2*	CAT GCT CGG TCG GGA TAG GCA CTG GTC TAG AGG GTT AGG TTC CTG CTA CAT CTC CAG CCT GAT C
VEC1.PRIM	AGG CAC TGG TCT AGA GGG TTA GGT TC

Table 2.2*a*: The PCR conditions used to generate templates for the sequencing of cosmid N0434 (chapter 3)

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Other Comments*
SK x KS	variable	Each N0434 shotgun library clone	1 x (96°C, 1') 24 x (96°C, 20"; 56°C, 30"; 63°C, 3')	Variable amplicon size as these conditions were employed to isolate shotgun library inserts prior to sequencing. These PCR conditions were also employed to amplify cosmid 3F3 library clones (Chapter 4).
N12.5F x N15.1R	2.5	10 ng of isolated N0434.1 DNA	1 x (96°C, 1') 22 x (96°C, 20"; 56°C, 30"; 63°C, 3')	These PCRs were performed in order to generate amplicons across gaps in the sequence assembly of the N0434 cosmid. The amplicons were then sequenced and the gaps closed.
N30.6F x N31.2R	0.6			
N24.1F x N26.3R	2.2			
N18.4F x N19.1R	0.6			
N19.4F x N19.7R	0.4		1 x (96°C, 1') 5 x (96°C, 20"; 66°C, 30"; 72°C, 3') 5 x (96°C, 20"; 62°C, 30"; 68°C, 3') 12 x (96°C, 20"; 58°C, 30"; 63°C, 3')	
N11.8F x N12.5R	0.7			
N10.4F x N10.9R	0.6			
N9.1F x N9.9R	0.9			
N7.3F x N9.1R	1.7			
N4.7F x N5.2R	0.7			
N1.4F x N2.4R	1.2			
VEC1.PRIM x N23.3R	variable	1 µl of *Hae*III or *Rsa*I vectorette libraries (see this chapter, section *2.8.b.ii*, and chapter 3)	1 x (96°C, 1') 5 x (96°C, 20"; 66°C, 30"; 70°C, 3') 5 x (96°C, 20"; 63°C, 30"; 70°C, 3') 12 x (96°C, 20"; 60°C, 30"; 70°C, 3')	These were vectorette PCRs and were also employed as part of the strategy to close gaps in the sequence assembly of the N0434 cosmid.
VEC1.PRIM x N9.1R				
VEC1.PRIM x N9.1R				
SK x N22.8R	0.8 or 1.1	Shotgun library clone 1A5 or 1H2	1 x (96°C, 1') 5 x (96°C, 20"; 60°C, 30"; 63°C, 2') 5 x (96°C, 20"; 58°C, 30"; 63°C, 2') 12 x (96°C, 20"; 56°C, 30"; 63°C, 2')	In the final push to close gaps remaining in the N0434 cosmid sequence assembly these PCRs were used to generate relatively small amplicons in an effort to clear any secondary structure that might have obstructed previous sequencing reactions.
SK x N22.86R	0.9 or 1.1			
N12.5F x N12.8R	0.4	Shotgun library clone 2A9 or N12.5F x N15.1R amplicon		
N7.9F x N8.5F	0.6	Shotgun library clone 1C1 or 1H3		
N10.8F x N11.3R	0.5	Shotgun library clone 1B7 or 1C4		

Table 2.2*b*: The conditions employed for other Chapter 3 PCR amplifications

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Purpose of reaction*
N23.5F x RN23.8R	0.295	1 ng of isolated N0434.1-.4 DNA	1 x (96°C, 1') 4 x (96°C, 20"; 65°C, 30"; 70°C, 90") 4 x (96°C, 20"; 63°C, 30"; 70°C, 90") 9 x (96°C, 20"; 61°C, 30"; 70°C, 90") 1 x (60°C, 1'; 70°C, 5')	To confirm presence of *PGPL* within the N0434 cosmid
MVN15.3F x MVN17.6R		1 ng of isolated N0434.1 DNA	1 x (96°C, 1') 6 x (96°C, 20"; 65°C, 30"; 70°C, 6') 6 x (96°C, 20"; 64°C, 30"; 70°C, 6') 15 x (96°C, 20"; 63°C, 30"; 70°C, 6')	Estimation of the size of the PGMS2 minisatellite within the N0434.1 sequence

Table 2.3: The conditions employed for Chapter 5 PCR amplifications

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Purpose of reaction*
RN21.6F x RN22.9R (see table 2.4)	1.33	1 ng of isolated N0434.1 DNA	1 x (96°C, 1') 6 x (96°C, 20"; 59°C, 30"; 63°C, 90") 6 x (96°C, 20"; 58°C, 30"; 63°C, 90") 15 x (96°C, 20"; 56°C, 30"; 63°C, 90")	Production of novel gene probe for expression analysis.

Table 2.4: Primers and conditions used to investigate the variability of PGMS1, PGMS2 and the STIR element tandem repeat arrays (Chapter 5).

(a): Amplification of the tandem repeats

MVN15.3F	GAT GGC ACA GAC TTA GTG CC
MVN17.6R	CAC GTC GTA ACA CCG TGT GG
MVN24.8F	CTT ATT GGT AGG GAG ACC TG
MVN25.9R	GTA CAG ACG AGA TGA ATG GG
RN28.3F	ACA GAG AAG GCC ACG TGG AG
RN29.5R	CCA GAA AGC GGG AGC TGA CC

Primers Used	*Template*	*Tandem repeat amplified*	*Cycling Conditions*
MVN15.3F x MVN17.6R	10 ng of genomic DNA	PGMS2	1 x (96°C, 1') 6 x (96°C, 20"; 65°C, 30"; 70°C, 6') 6 x (96°C, 20"; 64°C, 30"; 70°C, 6') 15 x (96°C, 20"; 63°C, 30"; 70°C, 6')
MVN24.8F x MVN25.9R		PGMS1	1 x (96°C, 1') 5 x (96°C, 20"; 56°C, 30"; 63°C, 6') 5 x (96°C, 20"; 56°C, 30"; 61°C, 6') 15 x (96°C, 20"; 56°C, 30"; 59°C, 6')
RN28.3F x RN29.5R		STIR array	

(b): Amplification of the probes for hybridisation to the amplified and blotted tandem repeats

KS	TGC AGG TCG ACG GTA TC
SK	CGC TCT AGA ACT AGT GGA TC

Primers Used	*Template*	*Tandem repeat probed*	*Cycling Conditions*
KS x SK	Shotgun library clone 1H5	PGMS2	1 x (96°C, 1') 24 x (96°C, 20"; 56°C, 30"; 63°C, 3')
MVN24.8F x MVN25.9R	25 pg of isolated N0434.1 DNA	PGMS1	1 x (96°C, 1') 3 x (96°C, 20"; 56°C, 30"; 63°C, 2') 3 x (96°C, 20"; 56°C, 30"; 61°C, 2') 9 x (96°C, 20"; 56°C, 30"; 59°C, 2')
RN28.3F x RN29.5R		STIR array	

Table 2.5: Oligonucleotides used to amplify and resequence regions of the N0434.1 interval for the SNP discovery and genotyping strategies described in Chapters 6 and 7.

AMPLICON E (2.919KB)

RN0.2F	TTC ATT CTG GGC TCG AGA CC
RN0.7F	TCT AAG AGG AGG CCT CAG G
RN0.7R	TGT TGA TCT TGA GCC CAG
RN1.0F	CCG TAC ATA CAG CTG GGC AC
RN1.4R	GGG AAG TGA CCC TTT CCA TA
RN2.3F	CAA CGG TGG GAA TCG CCA TG
RN3.2R	CAT GTG TGC ATG CCT GTG TG

AMPLICON X (2.353KB)

RN3.3F	GCA TGT GCA TAC TGT GTT GG
RN4.3F	ACA CAC GTA GAC ACA CGT CC
N5.2R	GTT AGC TGA CAC ATG TAG CC
RN5.6R	CCC TTT AAA TCC CCT GAG GA

AMPLICON F (2.522KB)

RN6.5F	GTC AAG ATA TGG GCA GGA CC
RN7.0F	TCC TCT TGT GTC TTA GAC GG
N7.3F	TGC TTT GCT CTC AGG GAG CC
RN8.0F	GTG AGC ATC GCA GGC CTC CA
RN8.1R	GGT GTT GTC TCT GCT GGG AT
RN8.4F	GAC AAC CCC TTT CAT GCC TG
N8.5R	AAC TGC AGA CCC GCC TCA GG
RN8.6R	GCC TGT CTC CCT GTT TCC CA
N8.8R	CAA CAT GTT AGC CAG GAC GG
N9.1R	CCA CGG GGA CCT GCA TTT GG

AMPLICON Y (0.869KB)

RN9.9F	GGG GAT GAA ACA GCC ACA TG
N9.9R	TTA GCT GGG CGT GGT GTC CG
RN10.5R	CCG GCT GAC GCT GCA CAC AT

AMPLICON G (2.054KB)

N10.4F	GAT CTG CTT CCC TTG ACG G
N10.8F	GCG GAT CAT GAG GTC AGG AC
N11.3R	ATC CCG GCT AGT CAG AAG GC
N11.8F	TGG AGA CCA TGA AGA GCT GC
RN11.9R	GTG AAG GCA AGT CCA TGT GG
N12.5R	TTC CAT GAA GCA GTC AGT CC

AMPLICON H (2.638KB)

N12.5F	GAG GGG AAA TGA GGA CTG AC
RN12.8F	AAA TTA GCG GGC ACA GGC CG
N12.8R	CCT GAC CTC ATG ATC CGC CC
RN13.4F	CAA ACT TCT GGG GTC AAG CG
RN13.5R	GCA CAA CTC GGC TGG GAA TG
RN13.8F	GGG TGT GAG GTG CAG ACA GC
N14.4F	TGA GCC ACG GTG CCA GGC GG
RN14.4R	CCG CCT GGC ACC GTG GCT CA
N15.1R	GAT TCC ACG GCC AGA GAC GG

AMPLICON D (3.075KB)

RN17.8F	AGG AGG GTT GTT CGT GGC CG
RN18.2F	TGG TGA TCA TAG GAC CG
RN18.5F	CAG CCG AGT GTG CAG TGA C
RN19.0F	GAC AGC CCA CTT AGG AGC CG
N19.1R	GGC TTG GTG TGA GAA CTA GC
RN19.7F	CAC TGT GCC CAG CCG CCG A
N19.7R	TCG GCG GCT GGG CAC AGT G
RN20.2R	GGC GTT GAA GGT TAC AGT GA
RN20.7R	GAA CAC TTC AGT AGC ACT CC
RN20.9R	TTC CCG AGA GAA GAC GCT GC

AMPLICON C (2.946KB)

RN20.8F	ACC CAG GTA TCC GCT TGT GC
RN20.9F	CTT TCA GTG GCC TCT CTG TG
RN21.6F	TCG CTT GTT GGA GCT GCT GG
RN21.7R	CAC GAT TTG CTT GGG AAA GG
RN22.0F	GCG TAT CCA CAA AAT CAC CG
RN22.5F	GTT AAT TCA GGC CGG GCA CA
RN22.9F	GCC ACA CGC AGG TGA AAT TC
RN22.9R	TGA TGT GTC ACA GAC CAT CC
RN23.2F	ACG GTC TGG TTT GAA GCT GC
RN23.4R	GTT TCC CAT GGT GGT GTC GG
RN23.8R	TGA ACG GAC GCC CAC AGA GG

AMPLICON Z (1.313KB)

N23.5F	GTT ACG GAA ACA TTC CGA GGG A
N24.1F	GTT CAA GTG ATT CTC CTG CC
RN24.3R	CGC AGT GGC TCA TGC CTG TC
RN24.8R	CAG GTC TCC CTA CCA ATA AG

AMPLICON B (3.528KB)

RN26.0F	ACA TCC TAA AGC TCT CGG GG
RN26.2F	GGC TGT ACC TGC AAG GGT GG
N26.3R	CAG TGG ACT CCT GGG GAC C
RN26.5F	ACA GAA CGC TGC ATT TCT GG
RN27.0F	CTT ACA GGC AGG CTC CTT AC
RN27.8F	GTC GGG AGG AGT CGA AAG CG
RN27.8R	GCA AAT GTC CCA AAG AAC GG
RN28.3F	ACA GAG AAG GCC ACG TGG AG
RN28.3R	CTC CAC GTG GCC TTC TCT GT
RN29.5R	CCA GAA AGC GGG AGC TGA CC

AMPLICON A (2.730KB)

RN30.2F	AGA GGA AGC CGA TGG TGT CC
RN30.5F	TGG GAC AAA GCC ACC GTC GC
RN31.2F	GGG TGG TAA AGC TCT GGA TG
RN31.3R	TCC TGT GGT CTC TGT GTG GG
RN31.6F	CTT GCG AAG CCT GTC CAA GG
RN31.6R	TCG CAA GAA GAG GCA CCT GC
RN32.0F	GGG AAT GGG ACC TGA TTT GG
RN32.7R	GGA CAT AGG AGA CCC TCA GC
RN32.9R	CTT GAA AAG GGA CGT CGC CC

Table 2.6: The PCR conditions used to generate templates for the resequencing and genotyping strategies described in Chapter 6 and 7.

Primers Used	*Name of amplicon*	*Size of Amplicon (kb)*	*Cycling Conditions*	*Other Comments*
RN30.2F X RN32.9R	A	2.7	1 x (96°C, 1') 6 x (96°C, 20"; 58°C, 30"; 63°C, 4') 6 x (96°C, 20"; 57°C, 30"; 63°C, 4') 26 x (96°C, 20"; 56°C, 30"; 63°C, 4')	Templates for resequencing were amplified from 10 ng of the genomic DNA of 8 UK semen donors in reaction volumes of 10 µl. Amplicons that were spotted on to dot blots were generated from 20 ng of the genomic DNA of 50 UK semen donors in reaction volumes of 20 µl (see chapter 6).
RN26.0F X RN28.3R	B	3.5		
RN20.8F X RN23.8R	C	2.9		
RN17.8F X RN20.9R	D	3.1		
RN0.2F X RN1.4R	Ea	1.1	1 x (96°C, 1') 6 x (96°C, 20"; 58°C, 30"; 63°C, 2') 6 x (96°C, 20"; 57°C, 30"; 63°C, 2') 26 x (96°C, 20"; 56°C, 30"; 63°C, 2')	
RN2.3F X RN3.2R	Eb	0.8	1 x (96°C, 1') 6 x (96°C, 20"; 59°C, 30"; 63°C, 2') 6 x (96°C, 20"; 58°C, 30"; 63°C, 2') 26 x (96°C, 20"; 57°C, 30"; 63°C, 2')	
RN6.4F X N9.1R	F	2.6	1 x (96°C, 1') 6 x (96°C, 20"; 62°C, 30"; 66°C, 3') 6 x (96°C, 20"; 61°C, 30"; 66°C, 3') 26 x (96°C, 20"; 60°C, 30"; 66°C, 3')	
N10.4F X N12.5R	G	2.1		
N12.5F X N15.1R	H	2.6	1 x (96°C, 1') 6 x (96°C, 20"; 59°C, 30"; 63°C, 3') 6 x (96°C, 20"; 58°C, 30"; 63°C, 3') 26 x (96°C, 20"; 57°C, 30"; 63°C, 3')	
RN3.3F X RN5.6R	X	2.4		
RN9.9F X RN10.5R	Y	0.9	1 x (96°C, 1') 6 x (96°C, 20"; 59°C, 30"; 63°C, 2') 6 x (96°C, 20"; 58°C, 30"; 63°C, 2') 26 x (96°C, 20"; 57°C, 30"; 63°C, 2')	
N23.5F X RN24.8R	Z	1.3		
RN17.8F X N19.1R	Da	1.3	1 x (96°C, 1') 9 x (96°C, 20"; 60°C, 30"; 63°C, 1')	Due to the plethora of both tandem repeat and Alu sequences in the N0434.1 interval, these "sub-amplicons" had to be generated in order to obtain clean sequence and/or unambiguous genotypes. In each case they were amplified from larger amplicons that had already been generated from genomic DNA as above (see Chapter 6, figure 6.6*a* for more detail)
N12.5F X RN13.5R	Ha	1.0	1 x (96°C, 1') 2 x (96°C, 20"; 59°C, 30"; 63°C, 2') 2 x (96°C, 20"; 58°C, 30"; 63°C, 2') 8 x (96°C, 20"; 57°C, 30"; 63°C, 2')	
RN13.8F X N15.1R	Hc	1.3		
RN13.4F X N13.8R	Hb	0.4	1 x (96°C, 1') 9 x (96°C, 20"; 60°C, 30"; 63°C, 1')	
N14.4F X N15.1R	Hd	0.7	1 x (96°C, 1') 9 x (96°C, 20"; 57°C, 30"; 63°C, 1')	

Table 2.7: Allele-specific oligonucleotides for genotyping of the SNPs in the N0434.1 interval. The allele-specific sites are marked in blue. Degenerate nucleotide positions are marked in green. Nomenclature reflects the nucleotide position (in kb) of each SNP (and ASO) from the distal end of the N0434.1 interval. Hence, the ASOs for SNP 423C/T (table 6.1) are labelled NA0.42/T and NA0.42/C.

NA0.42/T	TGG TGT GAT TGT GGC TCA	NA14.89/G	TGG ATG TGT TTT TAC TAG
NA0.42/C	TGG TGT GAT CGT GGC TCA	NA14.89/A	TGG ATA TGT TTT TAC TAG
NA1.10/G	AAG ACT AGC CTG ACC AAG	NA18.53/C	CGC CTG TCA TCC CAG CAC
NA1.10/A	AAG ACT AAC CTG ACC AAG	NA18.53/A	CGC CTG TAA TCC CAG CAC
NA2.35/C	AGG CTC GCA TTT GCA TAT	NA18.61/C+	GTG AAA CCC CGT CTC TAT
NA2.35/T	AGG CTC GTA TTT GCA TAT	NA18.61/T-	GTG AAA CTC GTC TCT ATT
NA2.45/C	AGG ACC CCT TTT CCT CTC	NA18.63/T	CAA AAA TTA CCT GGG TGC
NA2.45/T	AGG ACC CTT TTT CCT CTC	NA18.63/C	CAA AAA CTA CCT GGG TGC
NA2.79/C	TGA TGG GCG GGC AGG TGA	NA18.66/C	CTG TAG TCC CAG CTA CTC
NA2.79/T	TGA TGG GTG GGC AGG TGA	NA18.66/G	CTG TAG TGC CAG CTA CTC
NA3.10/C	GTG CAT ACA CAT ATG TAC	NA18.97/A	CTT ACA GAC CCA GGG AGA
NA3.10/T	GTG CAT ACA TAT ATG TAC	NA18.97/G	CTT ACG GAC CCA GGG AGA
NA3.37/+	GGA CTC ACA GTT CCA CCT	NA19.96/C	TCC TGG GCT CCA GTG ATC
NA3.37/-	GGA CTC AGT TCC ACC TGG	NA19.96/G	TCC TGG GCT GCA GTG ATC
NA4.38/T	AAG GAA ATA AAC TTA GAC	NA20.72/G	CCG GAG TGC TGC TGA AGT
NA4.38/g	AAG GAA AGA AAC TTA GAC	NA20.72/C	CCG GAG TCC TGC TGA AGT
NA4.58/A	TGG CAT CAA GCA CTG ATC	NA21.30/G	ATG TTT TGA GAC GGA GTT
NA4.58/g	TGG CAT CGA GCA CTG ATC	NA21.30/C	ATG TTT TCA GAC GGA GTT
NA4.72/C	GGC TCC ACC TCA GCT CAT	NA22.03/G	AAT CAC CGA ATT CAT ACA
NA4.72/T	GGC TCC ATC TCA GCT CAT	NA22.03/A	AAT CAC CAA ATT CAT ACA
NA4.82/A	GCC TGG CTG ATT TTT GTA	NA22.14/A	GCA CGC CAG AAG CAT CTT
NA4.82/G	GCC TGG CTG GTT TTT GTA	NA22.14/G	GCA CGC CGG AAG CAT CTT
NA5.23/G	GTG GGG AAC GTG ACT GGG	NA22.16/C	ATG TTA ACT CAG GAG GCC
NA5.23/A	GTG GGG AAC ATG ACT GGG	NA22.16/T	ATG TTA ATT CAG GAG GCC
NA7.22/C	TGC TGT GGG CGG GCA GCA	NA23.13/G	AGC AGA GGT GGG GTT TCA
NA7.22/T	TGC TGT GGG TGG GCA GCA	NA23.13/A	AGC AGA GAT GGG GTT TCA
NA7.29/A	TGG TGA CAT GTC CGC GTG	NA23.48/A	GAA GTA AAA ACA AAG GTT
NA7.29/G	TGG TGA CGT GTC CGC GTG	NA23.48/T	GAA GTA AAA TCA AAG GTT
NA8.02/C	TGA GCA TCG CAG GCC TCC	NA23.75/G	AGG CAG CGA TGC CCC CAC
NA8.02/T	TGA GCA TTG CAG GCC TCC	NA23.75/A	AGG CAG CAA TGC CCC CAC
NA8.24/C	CTA CAG GCT CCC GCG ACC	NA24.15/C	GAT TCT CCT GCC TCA GCC
NA8.24/T	CTA CAG GTT CCC GCG ACC	NA24.15/T	GAT TCT CTT GCC TCA GCC
NA8.42/C	CTG CAC TCA AGA CAG ACA	NA26.64/T	TGG TCC CTG ACC CCA AGC
NA8.42/G	CTG CAC TGA AGA CAG ACA	NA26.64/C	TGG TCC CCG ACC CCA AGC
NA9.77/C	TCT TGC TCT GTC GCC CAG	NA26.90/G	CCC AGG GGG TCT TCA TTC
NA9.77/G	TCT TGC TGT GTC GCC CAG	NA26.90/A	CCC AGG GAG TCT TCA TTC
NA11.50/A	TGC GCC CAG CCA GAA TTT	NA27.68/C	GAT CCA CCC GCC TCA ACC
NA11.50/G	TGC GCC CGG CCA GAA TTT	NA27.68/G	GAT CCA CGC GCC TCA ACC
NA12.23/A	GGG GAA GGA TCA ACA AGG	NA27.91/G	GTT TTC CTG CCT TGT TGC
NA12.23/G	GGG GAG GGA TCA ACA AGG	NA27.91/A	GTT TTC CTA CCT TGT TGC
NA12.65/G	AGC CAG GTG CGG TGG CTC	NA30.38/T	CTG CAG AGA TCC CTG CGT
NA12.65/A	AGC CAG GTG CAG TGG CTC	NA30.38/C	CTG CAG AGA CCC CTG CGT
NA13.35/T	CCG AGA TTG AGC CAG TGT	NA31.67/C	TGA GCC ACG CCG GGA AAG
NA13.35/C	CCG AGA TCG AGC CAG TGT	NA31.67/T	TGA GCC ATG CCG GGA AAG
NA13.64/A	CAC CTG TAA TCC CAG CTC	NA31.69/T	GCA CAA GTG CGG GYG GTG
NA13.64/G	CAC CTG TGA TCC CAG CTC	NA31.69/A	GCA CAA GAG CGG GYG GTG
NA13.99/A	TCT TCA CAG CCG CGA GGG	NA31.70/C	GCG GGC GGT GCC GCG GAG
NA13.99/G	TCT TCA CGG CCG CGA GGG	NA31.70/T	GCG GGT GGT GCC GCG GAG
NA14.05/A	CGG GCA AAG GAG CAG GTT	NA31.71/G	GCG GAG GGT CTG CGG GGG
NA14.05/G	CGG GCA AGG GAG CAG GTT	NA31.71/C	GCG GAG GCT CTG CGG GGG
NA14.20/C	TCT CCC GGG TTC AAG CCA	NA32.25/G	ACA TCC TGA TTC CCA GAA
NA14.20/T	TCT CCT GGG TTC AAG CCA	NA32.25/A	ACA TCC TAA TTC CCA GAA
NA14.53/G	TCG AGC AGT TCT CCC GCC	NA32.45/T	GGA GTG ATG CGG CCA CAA
NA14.53/A	TCG AGC AAT TCT CCC GCC	NA32.45/C	GGA GTG ACG CGG CCA CAA
NA14.56/G	ACC ACG CCC GGC TAA TTT		
NA14.56/A	ACC ACG CCC AGC TAA TTT		

Table 2.8: Primers used in Chapter 8

Universal primers

N12.5F	GAG GGG AAA TGA GGA CTG AC
XOT13.0F	CCT GGG CGA CAG AGT GGA AC
XN13.5F	CAT TCC CAG CCG AGT TGT GC
RN13.5R	GCA CAA CTC GGC TGG GAA TG
RN13.8F	GGG TGT GAG GTG CAG ACA GC
N15.1R	GAT TCC ACG GCC AGA GAC GG
MVN15.3F	GAT GGC ACA GAC TTA GTG CC
MVN17.6R	CAC GTC GTA ACA CCG TGT GG
RN17.8F	AGG AGG GTT GTT CGT GGC CG
RN17.8R	TCT CCG TGA GGT TGA TGC CG
N18.4F	GCT CAG GCA GGA GAA CTG C
XOT18.5R	CTC ATG TTC CAC CCG CCT TC
N19.1R	GGC TTG GTG TGA GAA CTA GC
XN20.6R	CTG GCT CTA CTG GGC TGG GA
RN21.7R	CAC GAT TTG CTT GGG AAA GG

Allele-specific primers

Selector sites are at the 3′ end of each allele-specific primer and are shown in red. Synthetic 5′ extensions are shown in lower case.

12237FA	TAC ATT GGC GGG GGG AA
12237FG	TAC ATT GGC GGG GGG AG
12657FG	TGG TGT CAG CCA GGT GCG
12657FA	CTG GTG TCA GCC AGG TGC A
c18610/RT	ccc cGT GTT TTT AAT AGA GAC GA
18970/RA	AGT GTC TCC CTG GGT CT
18970/RG	AGT GTC TCC CTG GGT CC
c21309RG	ccc cCG AGA GCG AAA CTC CGT CTC
22145RA/Ad	gtc tac gta gtc agc tct ggC ATT TTC TAA AGA TGC TTC T

NB: Only the allele-specific primers that were used in assays that worked, and those used to identify linkage phase of particular semen donors are shown here. Many modifications of the primers directed towards these and other potential selector sites had to take place before a workable crossover molecule assay was achieved. Modifications included 5´ truncations and the addition of synthetic 5´ extensions including CCCC and:

Adap	gtc tac gta gtc agc tct gg
Tail	tgc aca tgc cga cca tac gc

Table 2.9: The PCR conditions used to determine the optimal annealing temperatures of the allele-specific primers (AS-primers) used in the crossover assays described in Chapter 8.

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Other Comments*
forward AS-primer x XN20.6R	6.6–9.8	10 ng genomic DNA from men known to be homozygous for one allele or the other.	1 x (96°C, 1') *then* 26 x (96°C, 20"; 56°C, 30"; 65°C, 10') *or* 26 x (96°C, 20"; 59°C, 30"; 65°C, 10') *or* 26 x (96°C, 20"; 62°C, 30"; 65°C, 10') *or* 26 x (96°C, 20"; 65°C, 30"; 65°C, 10')	These conditions were employed to optimise the primers used in the first crossover assays within the interval bounded by SNPs 12237A/G and 22145A/G (see figure 8.1).
XN13.5F x reverse AS-primer				
XN13.5F x XN20.6R				
forward AS-primer x RN17.8R	5.1–5.6		1 x (96°C, 1') *then* 26 x (96°C, 20"; 56°C, 30"; 65°C, 6') *or* 26 x (96°C, 20"; 59°C, 30"; 65°C, 6') *or* 26 x (96°C, 20"; 62°C, 30"; 65°C, 6') *or* 26 x (96°C, 20"; 65°C, 30"; 65°C, 6')	These conditions were employed to optimise the primers used to detect crossovers within the interval bounded by SNPs 12237A/G and 18970C/G (see figure 8.3).
XN13.5F x reverse AS-primer				
XN13.5F x RN17.8R				

Table 2.10: The PCR conditions used to establish linkage phase of semen panel donor 5 prior to the first crossover assay attempts described in Chapter 8.

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Other Comments*
12237/FA x RN13.5R	1.3	10 ng donor 5 genomic DNA	1 x (96°C, 1') 38 x (96°C, 20"; 59°C, 30"; 63°C, 2')	Dot blotted and ASO hybridised for 12657/A or /G
12237/FG x RN13.5R	1.3		1 x (96°C, 1') 38 x (96°C, 20"; 65°C, 30"; 65°C, 2')	
12657/FA x RN21.7R	8.6–8.9		1 x (96°C, 1') 19 x (96°C, 20"; 62°C, 30"; 65°C, 10') 19 x (96°C, 20"; 65°C, 30"; 65°C, 10')	Dot blotted and ASO hybridised for 21309/C or /G
RN17.8F x 22145/RA/Ad	4.3		1 x (96°C, 1') 38 x (96°C, 20"; 56°C, 30"; 63°C, 5')	
N12.5F x RN13.5R	1.0	10 ng genomic DNA of homozygote	1 x (96°C, 1') 38 x (96°C, 20"; 62°C, 30"; 63°C, 2')	Control amplifications used the genomic DNA of two men; one that was homozygous for one allele of the relevant SNP and one that was homozygous for the other allele.
RN17.8F x RN21.7R	3.9		1 x (96°C, 1') 38 x (96°C, 20"; 56°C, 30"; 63°C, 5')	

Table 2.11: The PCR conditions used to establish linkage phase of semen panel donor 7 prior to sperm crossover analysis within the interval defined by N0434.1 SNPs 12237A/G and 18970A/G (Chapter 8).

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Other Comments*
12237/FA x RN13.5R	1.3	10 ng donor 7 genomic DNA	1 x (96°C, 1') 38 x (96°C, 20"; 59°C, 30"; 63°C, 2')	Dot blotted and ASO hybridised for 12657/A or /G
12237/FG x RN13.5R	1.3		1 x (96°C, 1') 38 x (96°C, 20"; 65°C, 30"; 65°C, 2')	
12657/FA x N19.1R	5.9–6.1		1 x (96°C, 1') 19 x (96°C, 20"; 62°C, 30"; 65°C, 7') 19 x (96°C, 20"; 65°C, 30"; 65°C, 7')	Dot blotted and ASO hybridised for 18610/C or /T, 18634/C or /T and 18970/A or /G
RN17.8F x 18970/RG	1.1		1 x (96°C, 1') 38 x (96°C, 20"; 62°C, 30"; 63°C, 2')	Dot blotted and ASO hybridised for 18610/C or /T and 18634/C or /T
MVN15.3F x 18970/RA	3.1 or 3.3		1 x (96°C, 1') 13 x (96°C, 20"; 59°C, 30"; 65°C, 4') 13 x (96°C, 20"; 59°C, 30"; 65°C, 4')	Agarose gel electrophoresis and comparison of product sizes allowed linkage of SNPs and PGMS2 allele sizes to be established.
MVN15.3F x 18970/RG			1 x (96°C, 1') 26 x (96°C, 20"; 62°C, 30"; 65°C, 4')	
N12.5F x RN13.5R	1.0	10 ng genomic DNA of homozygote	1 x (96°C, 1') 38 x (96°C, 20"; 62°C, 30"; 63°C, 2')	Control amplifications used the genomic DNA of two men; one that was homozygous for one allele of the relevant SNP and one that was homozygous for the other allele.
RN17.8F x N19.1R	1.3		1 x (96°C, 1') 38 x (96°C, 20"; 56°C, 30"; 63°C, 2')	

Table 2.12: The two rounds of repulsion phase allele-specific PCR used to detect recombinant molecules from batches of donor 7 sperm DNA within the interval defined by N0434.1 SNPs 12237A/G and 18970A/G (Chapter 8).

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Other Comments*
12237/FG x 18970/RG	6.1–6.3	Donor 7 blood or sperm DNA in various known amounts (see Chapter 8, section for details)	1 x (96°C, 1') 23 x (96°C, 20"; 65°C, 30"; 65°C, 7') 1 x (65°C, 8')	Primary PCR. Low extension temperatures reduced the incidence of inter-haplotype–jumping PCR artefacts, which occasionally occur with 70°C extension (Jeffreys *et al.*, 2000)
12657/FG x c18610/RT			1 x (96°C, 1') 6 x (96°C, 20"; 59°C, 30"; 65°C, 6') 23 x (96°C, 20"; 56°C, 30"; 65°C, 6') 1 x (56°C, 1'; 65°C, 6')	Secondary PCR, which took place post-S1 nuclease digestion of primary PCR products.

Table 2.13: The PCR conditions used to determine SNP haplotypes across the PGMS2 region (Chapter 8).

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Other Comments**
12657/FA x N15.1R	2.5	10 ng donor 7 genomic DNA	1 x (96°C, 1') 19 x (96°C, 20"; 62°C, 30"; 65°C, 3') 19 x (96°C, 20"; 65°C, 30"; 65°C, 3')	Dot blotted and ASO hybridised for for 13351/C or /T; 13642/A or /G and 14051/A or /G.
12657/FG x N15.1R				
RN13.8F x N15.1R	1.3	1 µl aliquots of completed 12657/FA or /FG x N15.1R	1 x (96°C, 1') 5 x (96°C, 20"; 58°C, 30"; 63°C, 3') 5 x (96°C, 20"; 57°C, 30"; 63°C, 3') 10 x (96°C, 20"; 56°C, 30"; 63°C, 3')	Dot blotted and ASO hybridised for 14209/C or /T and 14893/A or /G.
RN17.8F x 18970/RA	1.1	10 ng donor 7 genomic DNA	1 x (96°C, 1') 38 x (96°C, 20"; 59°C, 30"; 63°C, 2')	Dot blotted and ASO hybridised for 18530/A or /C
RN17.8F x 18970/RG			1 x (96°C, 1') 38 x (96°C, 20"; 62°C, 30"; 63°C, 2')	

Table 2.14: The PCR conditions used to reamplify positive secondary PCRs in order to map crossover points in the PGMS2 region (Chapter 8).

Primers Used	*Size of Amplicon (kb)*	*Template*	*Cycling Conditions*	*Other Comments**
XOT13.0F x N15.1R	2.1	1 µl aliquots of positive (or control) secondary PCR products.	1 x (96°C, 1') 5 x (96°C, 20"; 58°C, 30"; 63°C, 3') 5 x (96°C, 20"; 57°C, 30"; 63°C, 3') 10 x (96°C, 20"; 56°C, 30"; 63°C, 3')	Dot blotted and ASO hybridised for 13351/C or /T; 13642/A or /G and 14051/A or /G.
RN13.8F x N15.1R	1.3			Dot blotted and ASO hybridised for 14209/C or /T and 14893/A or /G.
N18.4F x XOT18.5R	0.1		1 x (96°C, 1') 5 x (96°C, 20"; 58°C, 30"; 63°C, 30") 5 x (96°C, 20"; 57°C, 30"; 63°C, 30") 10 x (96°C, 20"; 56°C, 30"; 63°C, 30")	Dot blotted and ASO hybridised for 18530/A or /C
MVN15.3F x MVN17.6R	1.67 or 1.9		1 x (96°C, 1') 6 x (96°C, 20"; 65°C, 30"; 63°C, 3') 6 x (96°C, 20"; 64°C, 30"; 63°C, 3') 11 x (96°C, 20"; 63°C, 30"; 63°C, 3')	Size of amplicon assumes crossover occurring outside PGMS2. PGMS2 allele sizes were visualised by agarose gel electrophoresis with ethidium bromide staining.

*ASO hybridisations to determine SNP haplotypes and to map points of crossover were performed concurrently. Thus, the haplotype blots served as internal controls when dotblots of recombinant molecules were being scored (Chapter 8).

Chapter 3
Sequencing the *PGPL* Region

3.1: Introduction

Despite its clear biological importance, and the advanced stage of the human genome sequencing project, it was surprising that publicly available PAR1 genomic sequence data was limited to an interval around the *SHOX* gene and the partial characterisation of a few minisatellites. As the limited amount of sequence data in and around the *PGPL* region was simply not adequate to begin a high resolution analysis of recombination, the generation of this data became an important first step.

3.1.1: Partial Characterisation of the Xp/Yp Telomere–*PGPL* Interval

A cosmid contig covering the distal 750 kb of PAR1 was constructed by Rao *et al.* (1997) (figure 3.1*a*) as part of their efforts to map a major locus involved in growth retardation. It is worth noting that the gene identified at the end of this study was the *SHOX* gene, which has already been subjected to a high resolution analysis of recombination as noted in Chapter 1 of this thesis and described more thoroughly by May *et al.* (2002). Gudrun Rappold (University of Heidelberg) was kind enough to provide three cosmids from the 750 kb contig that were purported to cover the region between *PGPL* and an interval adjacent to the PAR1 telomere. These cosmids were LLN0YCO3′M′29C1, LLN0YCO3′M′3F3 and LLNLN0434, hereafter referred to as 29C1, 3F3 and N0434 respectively. Both 29C1 and 3F3 were derived from the Lawrence Livermore Y-chromosome specific library (LLN0YCO3′M′), for which the cosmid vector is Lawrist 16 (GenBank accession number = L19898), while N0434, despite its prefix, appears to be derived from "a self-made library covering the entire genome" (Rao *et al.*, 1997). The interval immediately adjacent to the PAR1 telomere and the region covered by the 29C1, 3F3 and N0434 cosmids have been partially characterised and a number of structures have been identified (figure 3.1*b*).

3.1.1.a: Structures Adjacent to the PAR1 Telomere

Immediately adjacent to the telomere repeat array, in addition to the polymorphisms identified by Baird *et al.* (1995), there is a truncated copy of a short interspersed nuclear element (SINE)

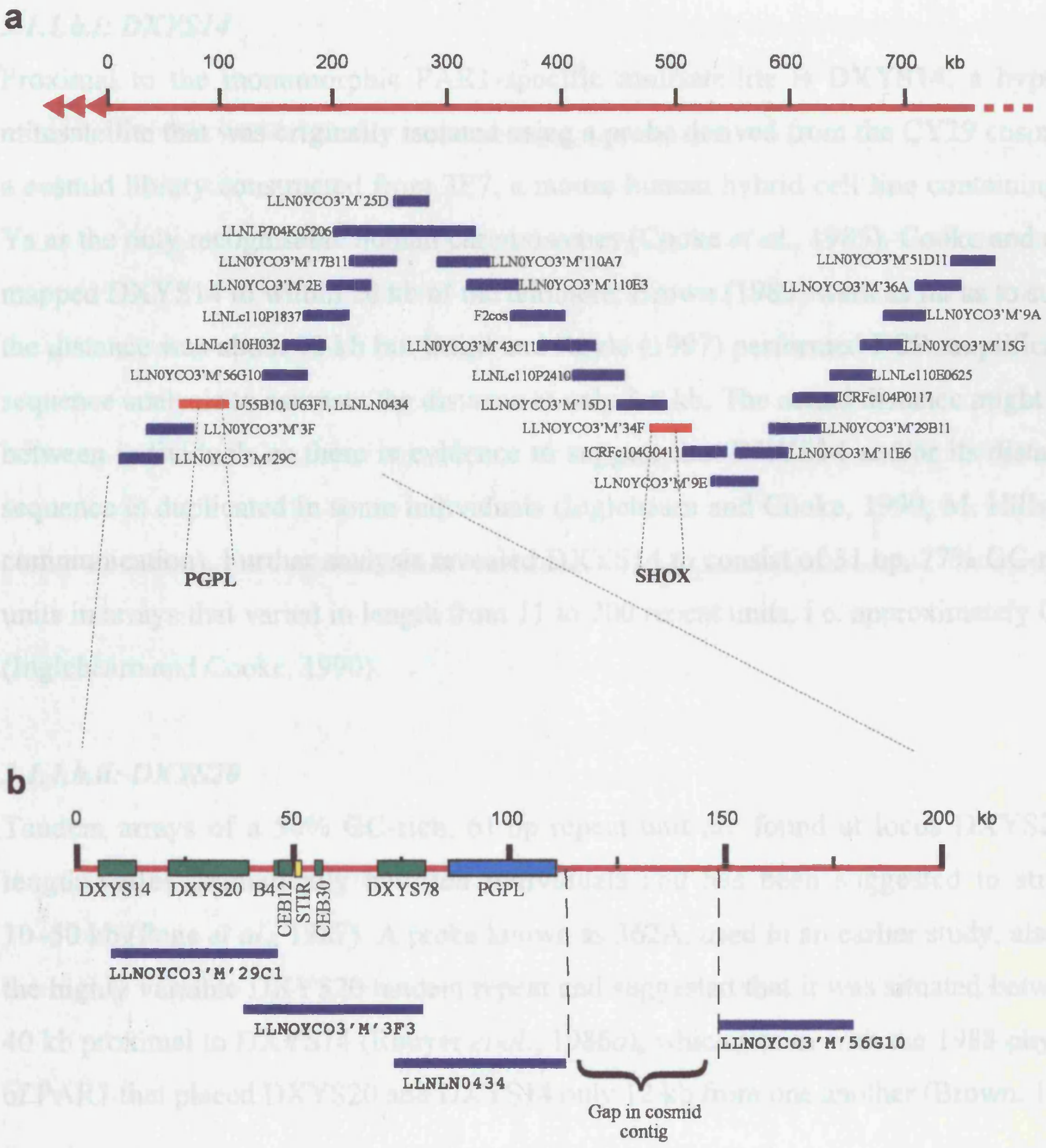


Figure 3.1: Cosmid contig covering the distal 750 kb of PAR1. (a) Adapted from Rao *et al*., 1997. All cosmids were derived from the Lawrence Livermore X- and Y-chromosome specific cosmid libraries and the ICRF X-chromosome specific library. *PGPL* was suggested to reside on cosmid LLNLN0434 and *SHOX* was mapped to LLN0YCO3'M'34F5, as shown by the red lines (Gianfrancesco *et al*., 1998). (b) The limited known loci in the region covered by the three cosmids provided to us by Gudrun Rappold (University of Heidelberg). Note that the lack of sequence data in this region has meant that each locus can only be shown at its estimated position relative to the telomere (position zero).

and a monomorphic PAR1-specific minisatellite consisting of four 63 bp repeat units (Brown, 1989; Royle *et al.*, 1992).

3.1.1.b: Tandem Repeats in the Region Covered by 29C1, 3F3 and N0434

3.1.1.b.i: DXYS14

Proximal to the monomorphic PAR1-specific minisatellite is DXYS14, a hypervariable minisatellite that was originally isolated using a probe derived from the CY29 cosmid, part of a cosmid library constructed from 3E7, a mouse-human hybrid cell line containing multiple Ys as the only recognisable human chromosomes (Cooke *et al.*, 1985). Cooke and colleagues mapped DXYS14 to within 20 kb of the telomere, Brown (1989) went as far as to suggest that the distance was about 13 kb but Baird and Royle (1997) performed PCR amplifications and sequence analysis to estimate the distance at only 3.6 kb. The actual distance might even vary between individuals as there is evidence to suggest that DXYS14 and/or its distal flanking sequence is duplicated in some individuals (Inglehearn and Cooke, 1990; M. Hills, personal communication). Further analysis revealed DXYS14 to consist of 31 bp, 77% GC-rich repeat units in arrays that varied in length from 11 to 200 repeat units, i.e. approximately 0.3–6.2 kb (Inglehearn and Cooke, 1990).

3.1.1.b.ii: DXYS20

Tandem arrays of a 50% GC-rich, 61 bp repeat unit are found at locus DXYS20, whose length varies dramatically between individuals and has been suggested to stretch over 10–50 kb (Page *et al.*, 1987). A probe known as 362A, used in an earlier study, also detected the highly variable DXYS20 tandem repeat and suggested that it was situated between 6 and 40 kb proximal to DXYS14 (Rouyer *et al.*, 1986*b*), which ties in with the 1988 physical map of PAR1 that placed DXYS20 and DXYS14 only 12 kb from one another (Brown, 1988).

3.1.1.b.iii: The B4, CEB12 and CEB30 Intervals

B4, CEB12 and CEB30 are subclones of cosmid 362, which is proximal and contiguous to CY29 (Vergnaud *et al.*, 1993). B4 contains a monomorphic minisatellite consisting of a 57% GC-rich, 28 bp tandem repeat unit that spans approximately 900 bp. In contrast, the CEB12 interval contains two minisatellites, at least one of which is variable. The first has a 26 bp repeat unit that is 72% GC-rich. An initial analysis suggested that this array varied in size from 1–4 kb. The second CEB12 tandem repeat array was discovered in the flanking sequence of the first. It has a 56% GC-rich 14 bp consensus unit but there appears to have been no investigation into its variability. CEB30 was found to contain a moderately variable minisatellite, with observed allele sizes of approximately 0.7–1.2 kb (Le Roux *et al.*, 1994) and a 19 bp, 59% GC-rich consensus repeat unit (Vergnaud *et al.*, 1993). The positions of the B4, CEB12 and CEB30 intervals as shown on figure 3.1*b* are approximate. Not only is a calculation of their exact distance from the telomere made difficult by the lack of sequence data outside of the DXYS14 and DXYS20 tandem repeats, but the highly variable nature of these repeat arrays further complicates matters. Nevertheless, CEB30 has been placed about 55 kb from the PAR1 telomere (Vergnaud *et al.*, 1993).

3.1.1.b.iv: DXYS78

A further minisatellite, DXYS78, of allele size range approximately 5–30 kb has been isolated and analysed by Armour *et al.* (1990). No sequence data is available for this locus but it is known that enzymes such as *Mbo*I and *Taq*I cut every few kilobases within the repeat array to give multi-band haplotypes for longer alleles and it is assumed that there is a *Hae*III site in most repeats (J.A.L. Armour, personal communication). Furthermore, it has been localised to an interval 70–80 kb from the telomere (Henke *et al.*, 1993). It has been suggested that DXYS78 may be one of the repeats within the CEB12 and CEB30 intervals (A.J. Jeffreys, personal communication). However, given the considerable differences in the reported array size of these loci and the respective differences in their apparent distance from the Xp/Yp telomere, this seems unlikely. Obviously this matter can only be resolved fully when sequence data for the DXYS78 minisatellite become available.

3.1.1.c: Subtelomeric Interspersed Repeats

A member of a SINE family known as subtelomeric interspersed repeat elements (STIRs, previously DXYZ2) has also been identified in the PAR1 telomere–*PGPL* interval (Rouyer *et al.*, 1990) (figure 3.1*b*). The first STIR elements were thought to be pseudoautosomal-specific (Simmler *et al.*, 1985; Rouyer *et al.*, 1986*b*) but, more recently, STIRs have been identified in the sex-specific parts of the X and Y chromosomes (Petit *et al.*, 1990) and at the distal ends of many human autosomes (Rouyer *et al.*, 1990). However, there are quite noticeable differences between the autosomal and the pseudoautosomal STIRs. In the loci analysed so far, the 350 bp pseudoautosomal STIRs are tandemly repeated two or four times, a situation that appears to have arisen through the duplication of individual elements. There are also two types of pseudoautosomal STIRs, which can be distinguished by differences in their first 80 bp. Both types then share a 55% GC-rich core of about 270 bp. In contrast, autosomal STIRs appear monomeric, are related to just one type of pseudoautosomal STIR and have a slightly longer homology that can extend up to 20 bp prior to the leader sequences defined by the pseudoautosomal elements. Furthermore, the sequence conservation between autosomal STIRs is much weaker than between the pseudoautosomal elements (Rouyer *et al.*, 1990).

STIRs have been found in most orders of mammals (W. Schempp and B. Weber, cited in Petit *et al.*, 1990), which argues in favour of a biological function. Considering that the pseudoautosomal STIRs appear to represent a more conserved and complex class of element, this raises two possibilities. Either the pseudoautosomal STIRs are involved in an additional or different function to the autosomal elements, or the correct mechanics and resolution of this putative function is much more important in the pseudoautosomal region than it is elsewhere. In this light, it is interesting to note that STIRs have been suggested to have a role in the recognition and initiation of pairing between homologues prior to meiotic recombination and that the STIR found at the distal end of Xq is an autosomal type element (Rouyer *et al.*, 1990). However, many more sequence and recombination analyses will have to be performed in order to test the above hypotheses, especially as Petit *et al.* (1990) suggested that STIRs might have some kind of role in escape from X inactivation.

3.1.1.d: The *PGPL* Gene

Only the cDNA sequence of the *PGPL* gene has been published (figure 3.2, GenBank accession number Y14391) and is believed to encode a putative GTP-binding protein of 442 amino acids (Gianfrancesco *et al.*, 1998). GTP-binding proteins have a number of functions ranging from cell proliferation and signal transduction to protein synthesis and protein targeting and have been identified in many organisms (Kjeldgaard *et al.*, 1996; Morimoto *et al.*, 2002). The exact function of *PGPL* is unclear, but the cDNA clone does hybridise to genomic DNA from a wide variety of species, from rodents to primates, and significant homology was observed with GTP-binding proteins of both *E. coli* and *C. elegans*, indicating that it has a conserved and important biological role (Gianfrancesco *et al.*, 1998). *PGPL* was localised to cosmid N0434 (figure 3.1) suggesting that it resides in an interval approximately 80–110 kb from the telomere and so proximal to the repeat arrays described above.

It thus appears that the distal 100 kb of PAR1 is largely composed of unrelated, variable, GC-rich tandem repeats. This is not in itself unusual, as it has been noted at the distal regions of many human autosomes (Royle *et al.*, 1988; Flint *et al.*, 1997*a*;*b*). However, considering the recombination hotspot-driven instability of the MS32 minisatellite (Jeffreys *et al.*, 1998a), it is interesting to note that the *PGPL* region, chosen for a high resolution analysis of recombination, lies just proximal to this distal PAR1 minisatellite-rich interval. Given the localisation of *PGPL* to N0434, this cosmid was chosen for initial analysis and sequencing.

```
1       gcgggccgccgtacgcccggggctgcggctctcccgcgtgggccgcggccgctcggctcc
61      gcgggcagccgcgccgtcctgccccgcgcgcgcgctagccgctgtcggccgcaggagccc
121     cgggaatctggagggggccgtggggcggagggagggggcctgcgggcggaaggcggacgaaa
181     acagaacggaagacgacaaggaggagccggaagatgcggacgagaacgccgaggaggagc
241     tgctgcggggagagcctctgctgccggcggggacccagcgcgtgtgtctggttcaccctg

301     acgtcaagtggggcccggggaagtcgcagatgactcgagccgagtggcaggtggcggagg
                                     M  T  R  A  E  W  Q  V  A  E
361     ccacagcgctggtgcacacgctggacggctggtccgtggtgcagacaatggtcgtgtcca
        A  T  A  L  V  H  T  L  D  G  W  S  V  V  Q  T  M  V  V  S
421     ccaaaacgccggacaggaagctcatctttggcaaagggaactttgagcacctgacagaaa
        T  K  T  P  D  R  K  L  I  F  G  K  G  N  F  E  H  L  T  E
481     agatccgagggtctccagacgtcacgtgcgtcttcctgaacgtggagaggatggctgccc
        K  I  R  G  S  P  D  V  T  C  V  F  L  N  V  E  R  M  A  A
541     cgaccaagaaagaactggaagccgcctggggcgtggaggtgtttgaccgcttcacggtcg
        P  T  K  K  E  L  E  A  A  W  G  V  E  V  F  D  R  F  T  V
601     tcctgcacatcttccgctgtaacgcccgcacgaaggaggcccggcttcaggtggccctgg
        V  L  H  I  F  R  C  N  A  R  T  K  E  A  R  L  Q  V  A  L
661     cggagatgccgctgcacaggtcgaacttgaaaagggacgtcgcccacctgtaccgaggag
        A  E  M  P  L  H  R  S  N  L  K  R  D  V  A  H  L  Y  R  G
721     tcggctcgcgctacatcatggggtcaggagaatccttcatgcagctgcagcagcgtctcc
        V  G  S  R  Y  I  M  G  S  G  E  S  F  M  Q  L  Q  Q  R  L
781     tgagagagaaggaggccaagatcaggaaggccttggacaggcttcgcaagaagaggcacc
        L  R  E  K  E  A  K  I  R  K  A  L  D  R  L  R  K  K  R  H
841     tgctccgccggcagcggacgaggcgggagttccccgtgatctccgtggtggggtacacca
        L  L  R  R  Q  R  T  R  R  E  F  P  V  I  S  V  V  G  Y  T
901     actgcggaaagaccacgctgatcaaggcactgacgggcgatgccgccatccagccacggg
        N  C  G  K  T  T  L  I  K  A  L  T  G  D  A  A  I  Q  P  R
961     accagctgtttgccacgctggacgtcacggcccacgcgggcacgctgccctcacgcatga
        D  Q  L  F  A  T  L  D  V  T  A  H  A  G  T  L  P  S  R  M
1021    ccgtcctgtacgtggacaccatcggcttcctctcccagctgccgcacggcctcatcgagt
        T  V  L  Y  V  D  T  I  G  F  L  S  Q  L  P  H  G  L  I  E
1081    ccttctccgccaccctggaagacgtggcccactcggatctcatcttgcacgtgagggacg
        S  F  S  A  T  L  E  D  V  A  H  S  D  L  I  L  H  V  R  D
1141    tcagccaccccgaggcggagctccagaaatgcagcgttctgtccacgctgcgtggcctgc
        V  S  H  P  E  A  E  L  Q  K  C  S  V  L  S  T  L  R  G  L
1201    agctgcccgccccgctcctggactccatggtggaggttcacaacaaggtggacctcgtgc
        Q  L  P  A  P  L  L  D  S  M  V  E  V  H  N  K  V  D  L  V
1261    ccgggtacagccccacggaaccgaacgtcgtgcccgtgtctgccctgcggggccacgggc
        P  G  Y  S  P  T  E  P  N  V  V  P  V  S  A  L  R  G  H  G
1321    tccaggagctgaaagctgagctcgatgcggcggttttgaaggcgacggggagacagatcc
        L  Q  E  L  K  A  E  L  D  A  A  V  L  K  A  T  G  R  Q  I
1381    tcactctccgtgtgaggctcgcaggggcgcagctcagctggctgtataaggaggccacag
        L  T  L  R  V  R  L  A  G  A  Q  L  S  W  L  Y  K  E  A  T
1441    ttcaggaggtggacgtgatccctgaggacggggcggccgacgtgagggtcatcatcagca
        V  Q  E  V  D  V  I  P  E  D  G  A  A  D  V  R  V  I  I  S
1501    actcagcctacggcaaattccggaagctctttccaggatgaacggacgcccacagaggcc
        N  S  A  Y  G  K  F  R  K  L  F  P  G  ●
1561    tgcggggtgggggcatcgctgcctggggagctgaggcgttaccgctgtgttgggggcagc
1621    ttggtgtcaggtgcagcagggtcctccttgtctggttctgcacccgtctcgctcccagcc
1681    atttgctgggatgaccgtgcaggccggtgacacggccgcacctgccccaaagcgggccgc
1741    ccgagcgtccactccaagcctgagcatccacacaattccagtgggccctcggtgcctgct
1801    gtgaactgcttccctcggaatgtttccgtaacaggacattaaacctttgattttaaaaa
1861    aaaaaaa
```

Figure 3.2: The cDNA sequence of *PGPL* (Gianfrancesco *et al.*, 1998) shown with the putative translation of the open reading frame. Primers used in the PCR to confirm the presence of *PGPL* in the N0434 cosmid are coloured pink. The position of the *Eag*I restriction enzyme site, which was used in the digest assay of the PCR products is denoted by the yellow box. The segment of *PGPL* cDNA that was subsequently shown not to be in cosmid N0434 is underlined.

3.2: Results

3.2.1: Assay for Insert Presence Within the Cosmids

The first stage was to confirm the presence of an insert within cosmids 29C1, 3F3 and N0434 and, particularly, to confirm the presence of *PGPL* within N0434. It remained possible that, for example, 90% of the N0434 glycerol stock provided by Gudrun Rappold was a contaminant. In this case, an assay for the presence of *PGPL* might have worked very well but any subsequent sequencing assembly would have been impeded by the inclusion of non-N0434 sequence. To this end, an aliquot of each cosmid glycerol stock was cultured overnight in LUB and plated out on LUA. Four single colonies of each cosmid were then selected, labelled as (cosmid name).1–.4 and stored as glycerol stocks at -80°C, as detailed in Chapter 2, section 2.1.a.

The N0434 vector was assumed to be Lawrist 16, as was known to be the case for 29C1 and 3F3. DNA was extracted from each cosmid "subclone" and digested with the *Sfi*I restriction enzyme, which cuts at each side of the multi-cloning site of Lawrist vectors. The digests were subjected to gel electrophoresis and were seen to have neatly separated the insert sequence from Lawrist 16, visible on the gel as a clean 5.2 kb band, in all cases apart from 29C1.1–.4 (data not shown). None of the four colonies picked from 29C1 contained an insert, suggesting that the highly variable DXYS14 and DXYS20 loci make this cosmid particularly unstable.

3.2.2: Assays to Confirm Presence of *PGPL* within N0434

3.2.2.a: PCR Assay

In order to confirm that the N0434.1–.4 inserts contained *PGPL*, the published cDNA sequence was used to design a simple PCR assay (table 2.2*b*) to amplify the putative 3′ untranslated region (3′ UTR), as this was expected to be intron-free. Cosmids 29C1 and 3F3 were used as negative controls in this amplification as these were not believed to contain *PGPL*. As can be seen from figure 3.2, the PCR was expected to generate an amplicon of 295 bp. The results are shown in figure 3.3*ai*. and clearly provided strong evidence that N0434 contains *PGPL*.

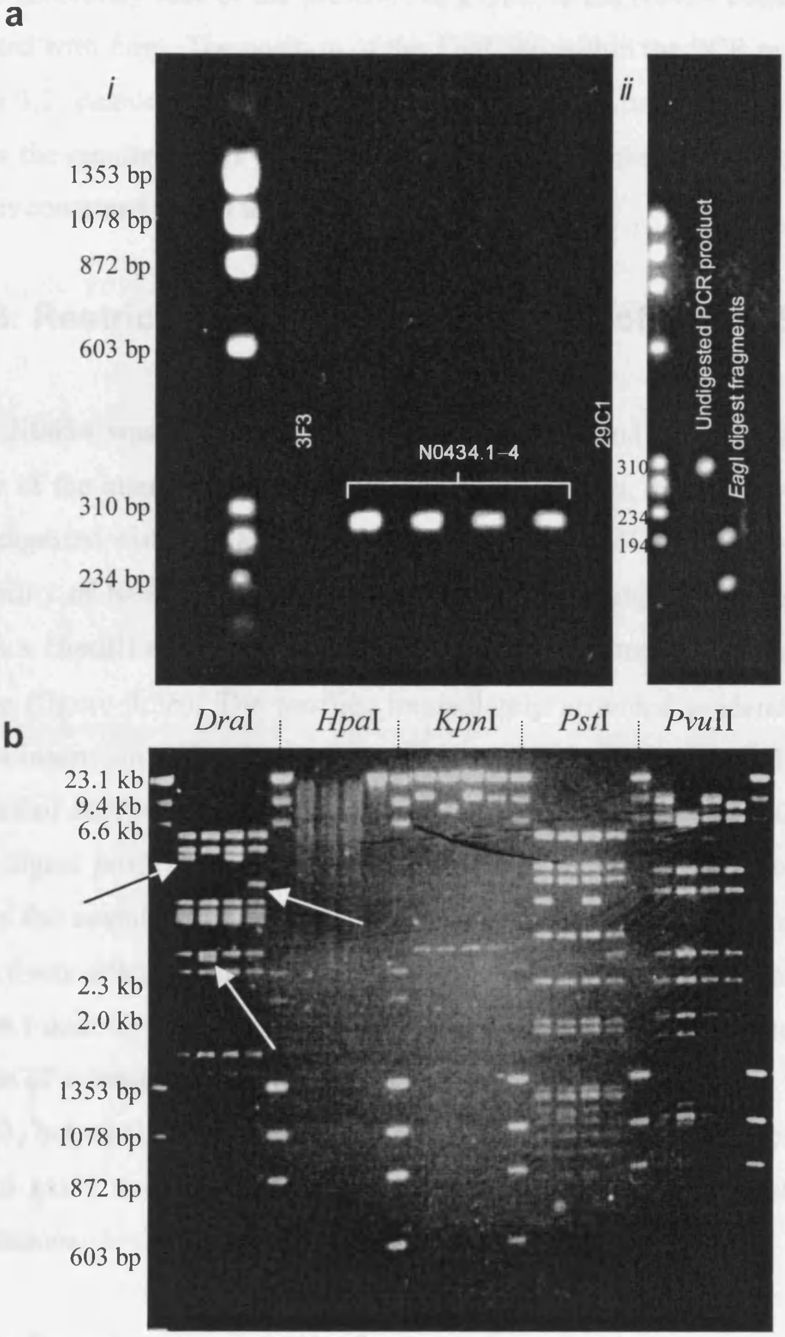


Figure 3.3: Assays to confirm the presence of *PGPL* in cosmid N0434 presented with the digest assays of N0434 sub-clones. (a) (i) PCR assay to confirm the presence of PGPL in N0434 sub-clones .1–.4. Cosmids 29C1 and 3F3 were also assayed as negative controls. (ii) To confirm this result the PCR product was digested with *Eag*I. The expected number and size of fragments are shown against undigested PCR product (b) Digestion of N0434. sub-clones .1–.4 using *Dra*I, *Hpa*I, *Kpn*I, *Pst*I and *Pvu*II restriction enzymes. In each case the profiles of .1 and .3 are identical but the profiles of .2 and .4 are different. This indicates that .2 and .4 have deletions of 2.2 kb and 0.8 kb respectively, relative to .1 and .3 and is denoted by the arrows over the *Dra*I profiles. This pattern suggests that the N0434 cosmid contains at least one tandem repeat array that is unstable in *E. coli*.

3.2.2.b: Digestion of PCR Product

To be absolutely sure of the presence of *PGPL* in the N0434 cosmid, the PCR product was digested with *Eag*I. The position of the *Eag*I site within the PCR product has been marked in figure 3.2, demonstrating that an *Eag*I digest should produce two fragments. Figure 3.3*aii* shows the results of this digest and clearly confirms that the putative 3' UTR of the *PGPL* gene is contained within the N0434 cosmid.

3.2.3: Restriction Enzyme Digest Assay of N0434 Sub-clones

Once N0434 was confirmed as containing *PGPL*, and knowing about the highly variable nature of the interval distally adjacent to *PGPL* region, the sub-clones of the N0434 cosmid were digested with the *Dra*I, *Hpa*I, *Kpn*I, *Pst*I and *Pvu*II restriction enzymes to determine the variability of N0434 inserts. The digests were then subjected to agarose gel electrophoresis with λ x *Hin*dIII and ϕx174 x *Hae*III standard weight molecular markers to produce a digest profile (figure 3.3*b*). The profiles immediately provided evidence of variability between N0434 inserts; in each case the digest profiles of N0434.1 and N0434.3 were identical but the profiles of N0434.2 and N0434.4 were different (but related to N0434.1 and N0434.3). As these digest profiles also enabled a reasonably accurate estimation of the length of DNA within the cosmid, this indicated that the rearrangements seen in N0434.2 and –.4 were respectively deletions of 2.2 kb and 0.8 kb relative to the approximately 35.5 kb length of the N0434.1 and –.3 sequences. The differing profiles can be explained by the enzymes cutting outside of at least one variable length tandem repeat array, which is the same size in N0434.1 and -.3, but smaller in both N0434.2 and –.4. These digest profiles also provided a quality control assay as they could be used for comparison with digests of subsequent N0434 preparations.

3.2.4: Preparation of N0434.1 Shotgun Library

Given the heterogeneity of the N0434 cosmid, it was decided that sequencing would focus exclusively on the N0434.1 subclone. First, 100 μg of N0434.1 DNA was prepared through the scaled-up version of the method described in Chapter 2, section 2.2.3.c. An aliquot of the DNA was then digested with the *Pst*I and *Dra*I restriction enzymes to confirm that the preparation had yielded the N0434.1 DNA, as defined by the restriction enzyme digest assay (above). A shotgun library of 192 colonies, containing 1–2 kb inserts was then prepared as Chapter 2, section 2.2.9. Before any colonies were picked into the ordered array that constituted the library, twelve recombinant colonies were selected at random to check for the presence of an inserted N0434.1 fragment by plasmid DNA preparation (Sambrook *et al.*, 1989) and restriction enzyme digestion with *Kpn*I and *Bam*HI to cut the insert out of the pBluescript SK$^+$ vector. The digests were then subjected to agarose gel electrophoresis to determine the size of the insert (figure 3.4). All twelve recombinant colonies were shown to contain an insert providing strong evidence that most, if not all, of the 186 recombinant colonies within the N0434.1 shotgun library contained an insert derived from the N0434.1 subclone (table 3.1).

3.2.5: Sequencing of N0434.1

3.2.5.a: Sizing the shotgun library inserts

Initial attempts to produce single-stranded phagemid DNA using VCS-M13 helper phage (Stratagene) were not successful. Consequently, double-stranded DNA was prepared from the library clones and the inserts were amplified using the SK and KS primers. The sizes of the amplified inserts were then determined by agarose gel electrophoresis against standard molecular weight markers (figure 3.4, table 3.1).

3.2.5.b: Sequencing strategy

All of the amplified inserts were purified through electroelution and ethanol precipitation and first sequenced with the KS primer. As the preparation of the shotgun library would have inevitably led to the inclusion of a small proportion of Lawrist 16 fragments, all of the retrieved sequences were aligned, *in silico*, against the full Lawrist 16 sequence to ensure that no Lawrist 16 sequences were included in the assembly of the N0434.1 sequence. If appropriate, the inserts were then sequenced with the SK primer. All sequence reads were assembled using ABI AutoAssembler software.

Unfortunately, the library was not quite large enough for full coverage of the N0434.1 sequence, which resulted in a small number of gaps in the sequence assembly. As the sizes of the shotgun library inserts had already been determined (table 3.1), it was possible to estimate the sizes of the gaps in the sequence assembly, which could then be filled by designing primers from the flanking sequence of a gap and then using these primers to PCR amplify the 'gap' from N0434.1 DNA (Chapter 2, table 2.2*a*). The gap primers were then used to sequence the gap amplicon. In a few cases vectorette PCR (Chapter 2, section 2.2.8.b.ii) had to be used to close particularly problematic gaps. In these cases vectorette libraries were amplified (Chapter 2, table 2.2*a*) and the reactions were subjected to agarose gel electrophoresis against standard molecular weight markers in order to separate and accurately size the amplified fragments. As the size of the problematic gaps could be estimated well, the fragments that were most likely to have one of their ends within the unknown sequence were separated by agarose gel electrophoresis, isolated by electroelution and sequenced with the vectorette primer, VEC1.PRIM (Chapter 2, table 2.1). A sequence interval was not considered to be 'complete' until it had been covered at least three times by sequence reads in the same direction but from different templates, or by two sequence reads in opposite directions.

3.2.5.c: Length and GC Composition of the Complete N0434.1 Sequence

Initial observations during sequence assembly had indicated the presence of at least three minisatellite arrays, which were provisionally named PGMS1, PGMS2 and PGMS3. It was not possible to sequence across PGMS2, the largest of the arrays, so its approximate size of 2.3 kb was estimated through a PCR amplification using primers designed within the PGMS2 flanking DNA (table 2.2*b*). This meant that the exact size of the insert within the N0434.1 cosmid could not be determined but, upon completion, the N0434.1 sequence assembly consisted of approximately 135 kb of sequence reads accumulated over a full length of about 33,400 bp (Appendix 1), in very good agreement with the 35.5 kb estimate gained through the restriction enzyme assay (above). The average GC content of the entire N0434.1 interval is 53.12%. As the 80–100 kb region that stretches from the *PGPL* gene to the PAR1 telomere appears to be composed almost entirely of 50–77% GC-rich tandem repeats (see above) this suggests that the N0434.1 interval could be part of a larger subtelomeric region forming a fraction of the H2 or H3 isochore families, representing the GC-richest class of human DNA as defined by Bernardi (1989).

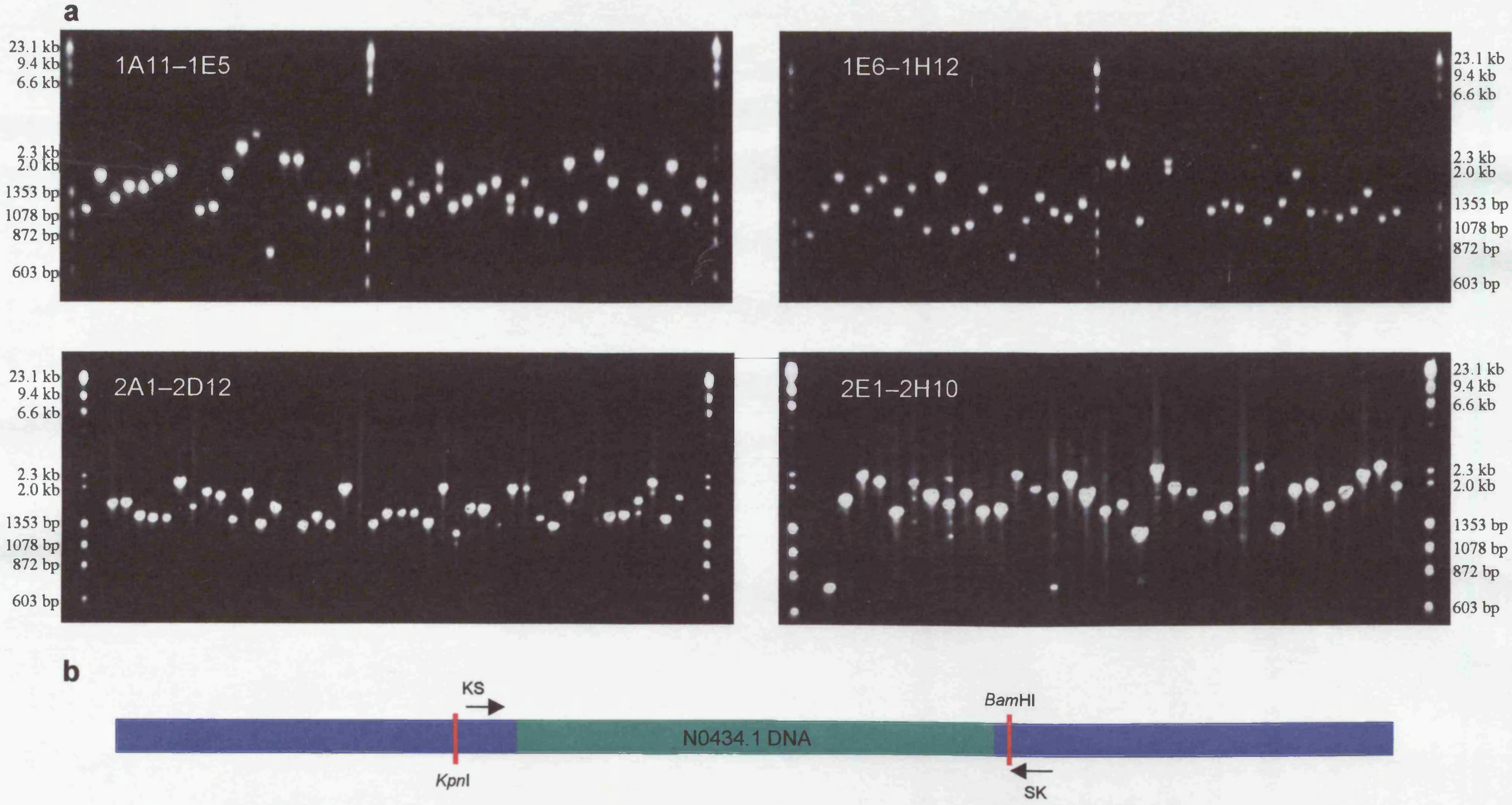


Figure 3.4: Determining the size of inserts in N0434.1 shotgun library clones (a) Amplifications of subsets of inserts from the shotgun library were subjected to gel electrophoresis against λ x *Hind*III and ϕx174 x *Hae*III standard weight molecular markers. (b) A representation of the plasmid DNA within a recombinant shotgun library colony. This shows how the inserted N0434.1 DNA (shown in green) can be separated from pBluescript SK$^+$ DNA (in blue) through digestion with the *Kpn*I and *Bam*HI restriction enzymes (whose sites are denoted by the red, vertical lines) or by amplification with the SK and KS primers.

Table 3.1: Sizes of inserts in the N0434.1 shotgun library. The bulk of the library is an ordered array of recombinant colonies picked into two 96-well plates. Plate one colonies were labelled 1A1–1H12 and plate two colonies were marked 2A1–2H12. Internal controls are marked within the table as 'non-recombinant,' which denoted those colonies that did not contain recombinant pBluescript SK$^+$ vectors, or 'no colony,' which indicated those wells that did not have a colony picked into them. PCR amplification with the KS and SK primers demonstrated that 86% of the colonies within the library contained a N0434.1 insert. A small proportion of the colonies produced two fragments after amplification which were most likely due to mixture of two recombinant colonies as they grew. In these cases the fragments were separated by agarose gel electrophoresis and purified separately through electroelution and ethanol precipitation. The separated fragments were sequenced as normal and included in the N0434.1 sequence assembly.

Table 3.1: Sizes of inserts in the N0434.1 shotgun library

clone	*insert size (kb)*	*comments*	*clone*	*insert size (kb)*	*comments*	*clone*	*insert size (kb)*	*comments*
T1a	1.4	Stored as DNA preps at -20ºC as these were the colonies selected at random to check for the presence of an inserted N0434.1 fragment (see text)	1C1	2.1		1F1	2.0	
T1b	1.4		1C2	2.1		1F2	1.3	
T2a	2.3		1C3	1.3		1F3	2.2	
T2b	1.9		1C4	1.2		1F4	1.3	
T3a	1.2		1C5	1.2		1F5	1.3	
T3b	1.3		1C6	1.9		1F6	2.0	
A1a	1.3		1C7	1.2		1F7	1.5	
A1b	1.4		1C8	1.4		1F8	0.9	
A2a	1.2		1C9	1.3		1F9	1.3	
A2b	1.0		1C10	1.4		1F10	1.7	
A3a	1.5		1C11	1.6/2.0	double	1F11	1.4	
A3b	2.1		1C12	1.3		1F12	1.3	
1A1		Non-recombinant	1D1	1.4		1G1	1.5	
1A2	—		1D2	1.5		1G2	2.5	
1A3	—		1D3	1.6		1G3	2.5	
1A4	2.2		1D4	1.2/1.4	double	1G4	1.2	
1A5	1.4		1D5	1.6		1G5	—	
1A6	1.2		1D6	1.2		1G6	2.3/2.6	double
1A7	1.3		1D7	1.1		1G7	—	
1A8	1.3		1D8	2.0		1G8	—	
1A9	1.5		1D9	1.3		1G9	1.4	
1A10	1.4		1D10	2.2		1G10	1.5	
1A11	1.2		1D11	1.6		1G11	1.4	
1A12	1.6		1D12	—		1G12	—	
1B1	1.3		1E1	1.5		1H1	1.2	
1B2	1.5		1E2	1.3		1H2	1.4	
1B3	1.5		1E3	1.9		1H3	2.0	
1B4	1.6		1E4	1.2		1H4	1.3	
1B5	1.7		1E5	1.6		1H5	1.3	
1B6	—		1E6	1.2		1H6	1.2	
1B7	1.2		1E7	1.6		1H7	1.3	
1B8	1.2		1E8	2.2		1H8	1.5	
1B9	1.8		1E9	1.6		1H9	1.2	
1B10	2.4		1E10	1.9		1H10	1.2	
1B11	—		1E11	2.2		1H11	—	Non-recombinant
1B12	0.8		1E12	1.5		1H12	—	No colony

clone	*insert size (kb)*	*comments*	*clone*	*insert size (kb)*	*comments*	*clone*	*insert size (kb)*	*comments*
2A1	—	Non-recombinant	2C9	1.4		2F5	—	
2A2	1.6		2C10	2		2F6	2.3	
2A3	1.6		2C11	2		2F7	—	
2A4	1.4		2C12	1.5		2F8	2.1	
2A5	1.4		2D1	1.3		2F9	—	
2A6	1.4		2D2	—		2F10	1.9	
2A7	2.2		2D3	1.8		2F11	2.3	
2A8	1.6		2D4	2.1		2F12	2.0	
2A9	1.9		2D5	—		2G1	1.6	
2A10	1.9		2D6	1.4		2G2	1.7	
2A11	1.4		2D7	1.4		2G3	1.3	
2A12	1.9		2D8	1.7		2G4	2.4	
2B1	—		2D9	—		2G5	1.9	
2B2	1.4		2D10	2.1		2G6	1.9	
2B3	1.6		2D11	1.4		2G7	—	
2B4	1.5		2D12	1.9		2G8	1.4	
2B5	1.3		2E1	0.8		2G9	1.6	
2B6	1.4		2E2	—		2G10	1.9	
2B7	1.3		2E3	1.9		2G11	2.4	
2B8	2.1		2E4	2.3		2G12	—	
2B9	—		2E5	2.2		2H1	1.3	
2B10	1.4		2E6	1.6		2H2	—	
2B11	1.4		2E7	—		2H3	2.7	
2B12	—		2E8	—		2H4	2.0	
2C1	1.4		2E9	2.2		2H5	2.0	
2C2	1.4		2E10	1.9		2H6	1.6	
2C3	—		2E11	1.8		2H7	1.8	
2C4	1.3		2E12	1.9		2H8	2.2	
2C5	2.1		2F1	1.6		2H9	2.4	
2C6	1.2		2F2	1.6		2H10	2.0	
2C7	1.6		2F3	—		2H11	—	Non-recombinant
2C8	1.6		2F4	—		2H12	—	No colony

3.2.6: Identification of the *PGPL* cDNA Sequences Within N0434.1

Although the presence of the *PGPL* 3' UTR had been confirmed within the N0434 cosmid, it was unclear whether all of the gene was contained within it. Therefore the completed N0434.1 sequence assembly was searched for the published *PGPL* cDNA sequence. This was achieved by including small sections of the cDNA sequence into the assembly and using the ABI AutoAssembler software to align them in their correct positions. For the first time this allowed some of the genomic structure of *PGPL* to be revealed but showed that a substantial proportion of the 5′ end of the published *PGPL* cDNA was not contained within the N0434 cosmid (figure 3.2). Incidentally, there were no sequence differences between the *PGPL* exons within N0434.1 and the published cDNA sequence.

3.2.7: Orientation of N0434.1 Within PAR1 and Overlap with 3F3

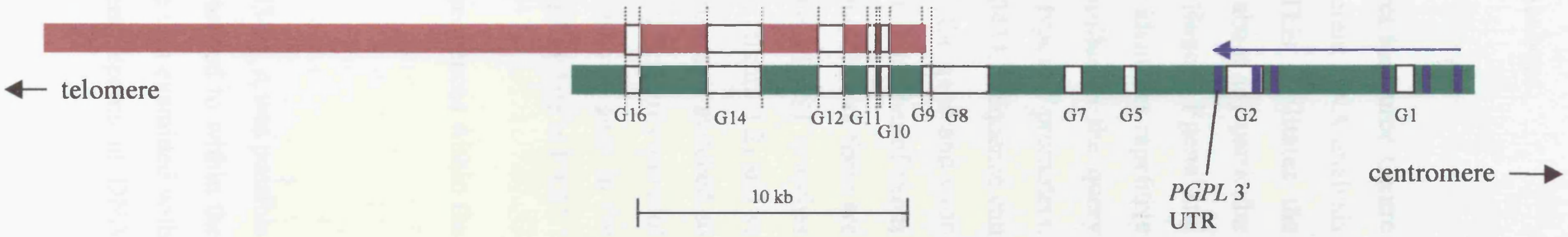
Given that the 3F3 cosmid was reported to partially overlap N0434 at its distal end (Rao *et al.*, 1997), and because the exons at the 3′ end of *PGPL* had been identified, the gap amplicons generated to close gaps in the N0434.1 sequence assembly were also used to determine the orientation of both the N0434 interval and *PGPL* within PAR1. This was achieved through attempting to amplify each of the gaps from both the N0434.1 and 3F3 cosmids whilst using DNA isolated from the 29C1 cosmid as a negative control. Clearly all of the gaps were expected to amplify from N0434.1 DNA, none were expected to amplify from 29C1 DNA and only those intervals contained within the N0434/3F3 overlap should have amplified from 3F3 DNA. The results of this set of reactions are shown in table 3.2 and figure 3.5 and clearly indicate that 3F3 extends approximately 12 kb into N0434. However, it is also clear that the 5′ part of *PGPL* that is missing from N0434 is centromeric to this interval and so resides in the only discernible gap in the cosmid contig covering the distal 750 kb of PAR1 (figure 3.1).

The products that arose from amplification of 3F3 using the G2 and G9 primer pairs are clearly not the same as those gained from N0434.1 DNA (table 3.2). This suggested two main alternatives; first, the N0434.1 G2 and G9 regions may be variable tandem repeat loci that are duplicated in the 3F3 interval. Second, the G2 and G9 primers may reside in a kind of interspersed repeat, which is present in the 3F3 interval and sufficiently similar to the repeat in N0434 to allow amplification.

Table 3.2: The PCR amplifications performed on isolated N0434, 3F3 and 29C1 cosmid DNA in order to determine the orientation of the N0434 and *PGPL* sequences within PAR1. All amplicons were successfully produced from N0434 DNA and not from 29C1, as expected. Those amplicons that were generated from 3F3 DNA corresponded to those from one end of N0434, thus enabling both the orientation of N0434 and an estimation of the degree of overlap between 3F3 and N0434., as shown in figure 3.6. The 3F3 amplicons generated with the G2 and G9 primers are discussed further in the main text. PCR cycling conditions were as described in table 2.2*a*,*b*.

Name of gap	*Primers used*	*Amplicon produced*			*Comments*
		N0434.1	*3F3*	*29C1*	
G1	N30.6F x N31.2R	✓	✗	✗	
G2	N24.1F x N26.3R	✓	(✓)	✗	N0434 amplicon was 2.2 kb (as expected) but amplicon generated from 3F3 cosmid was ~1.4 kb
G5	N20.2F x RN20.7R	✓	✗	✗	This amplicon was not part of the N0434.1 gap-closing strategy.
G7	N18.4F x N19.1R	✓	✗	✗	
G8	N12.5F x N15.1R	✓	✗	✗	
G9	N11.8F x N12.5R	✓	(✓)	✗	N0434 amplicon was 0.7 kb (as expected) but amplicon generated from 3F3 cosmid was ~5 kb
G10	N10.8F x N11.3R	✓	✓	✗	
G11	N10.4F x N10.9R	✓	✓	✗	
G12	N9.IF x N9.9R	✓	✓	✗	
G14	N5.2R x N7.3F	✓	✓	✗	
G16	N1.4F x N2.4R	✓	✓	✗	

Figure 3.5: The orientation of N0434 and *PGPL* within PAR1 and an estimation of the degree of overlap between the N0434 (green) and 3F3 (plum) cosmids. The positions of the *PGPL* exons identified within N0434.1 are shown in blue.



3.2.8: In silico analysis of the N0434.1 sequence

Before moving on to the next stage of a high resolution analysis of recombination, it was important to establish a comprehensive view of the genomic structure of the N0434.1 interval. Therefore the completed sequence was subjected to a number of *in silico* analyses.

3.2.8.a: NIX analysis

NIX, accessed through the MRC UK Human Genome Mapping Project Resource Centre (http://www.hgmp.mrc.ac.uk/), subjects a DNA sequence to many different DNA analysis programs and allows their results to be viewed simultaneously. This facilitates the identification of those instances when many programs have a consensus about any particular sequence feature. Programs used in a NIX analysis include Fex, Hexon, Fgene and HMMGene, which predict whole genes or exons; RepeatMasker, which identifies repetitive elements; BLAST, which searches many databases for sequences similar to the query sequence and GRAIL, which can predict protein coding genes, certain types of promoters, polyadenylation sites and CpG islands. The NIX analysis of the whole N0434.1 sequence can be seen in figure 3.6. There is a very good agreement between most of the gene and exon prediction programs (the blue and purple colours) in two major areas; first, a number of exons are predicted in the forward sense from approximately 3000–20,000 bp and second, there are suggested to be exons from 23,000–33,000 bp in the reverse sense. The BLAST searches showed that the latter exons were over 99% identical to the *PGPL* cDNA (figure 3.2) and so were part of the *PGPL* gene. The remaining putative exons in the forward sense matched an as yet uncharacterised cDNA reported to be slightly similar to a 1-phosphatidylinositol phosphodiesterase precursor. This clearly suggests the identification of a novel gene in the N0434 interval, distal to *PGPL* and thus the most telomeric gene discovered to date in PAR1.

In addition, the NIX analysis revealed a very high density of repetitive sequences within the N0434 interval, most of which are Alu elements.

3.2.8.b: Tandem Repeats Finder

As three minisatellite arrays had been provisionally identified within N0434.1, it was possible that the minisatellite-richness of the region distal to *PGPL* actually extended to within the N0434 interval. To investigate this further, the entire N0434.1 sequence was examined with Tandem Repeats Finder, a program that locates and displays tandem repeats in DNA sequences (http://tandem.biomath.mssm.edu/trf/trf.html; Benson, 1999).

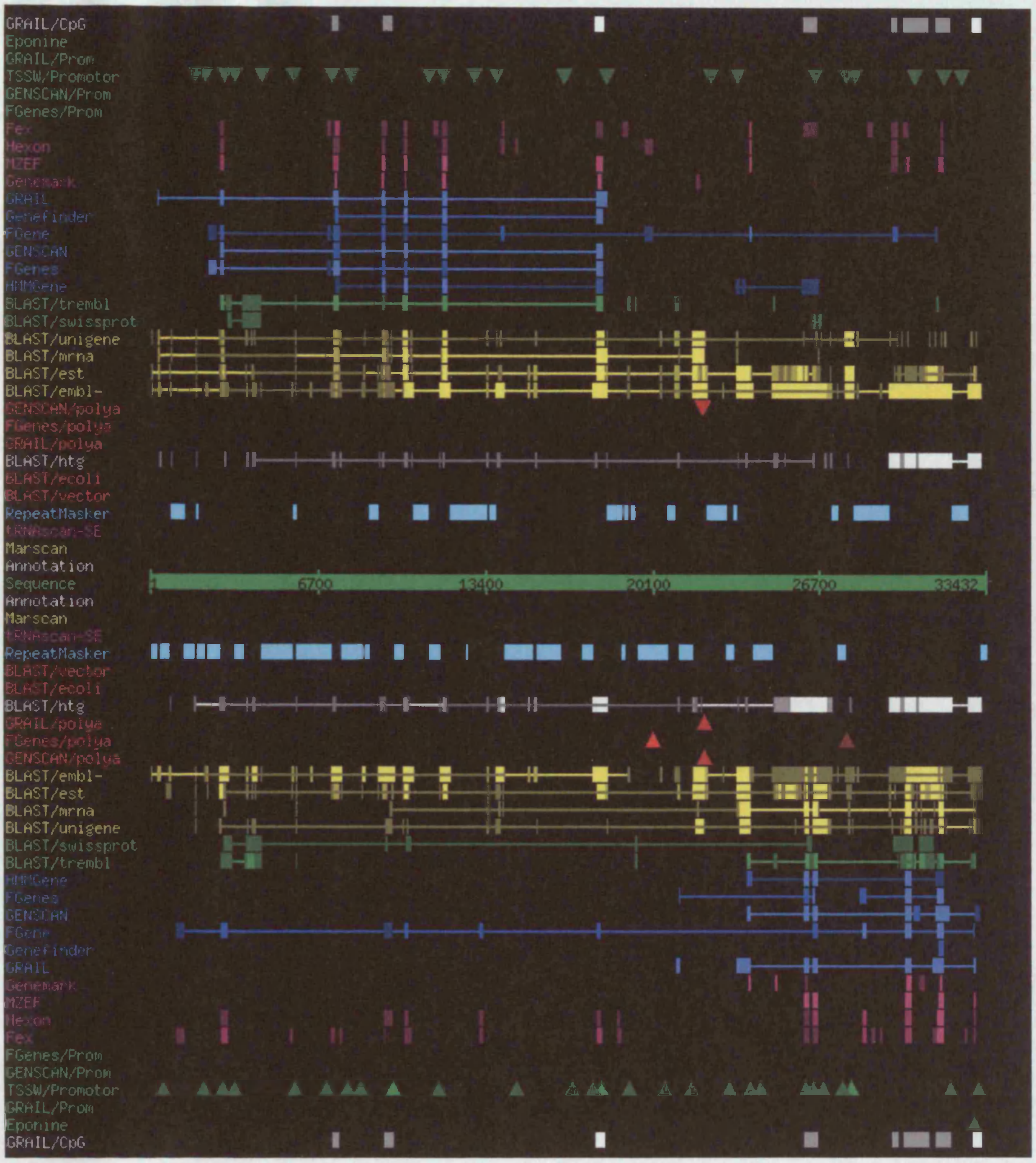


Figure 3.6: The NIX analysis of the entire N0434.1 sequence. N0434.1 is shown as the thick, central, green line. Those programs with similar purposes have generally been grouped together and given the same colour. Everything above the N0434.1 sequence line is a feature found in the forward sense and everything below is a feature in the reverse sense. Where colours are more intense, the quality or confidence of the prediction is stronger. Obviously, confirmation of these predictions requires a more focussed investigation of the individual DNA regions containing the putative features, and experimentation is usually essential. Also see appendix 1 and figure 3.7.

This analysis revealed at least 15 novel tandem repeat sequences, which together covered 22% of the N0434.1 interval (table 3.3, figure 3.7 and appendix 1); these included PGMS1, PGMS2, PGMS3 and a tandemly repeated array of pseudoautosomal STIR elements very similar to those described by Rouyer *et al.* in 1990 (figure 3.7). Given the supposed proximity of *PGPL* to the DXYS78 locus (above and figure 3.1*b*), it was conceivable that one of these novel tandem repeats might be DXYS78 itself. With no sequence data for DXYS78 available the simplest way to test this possibility was to screen the tandem repeats within the N0434.1 sequence for *Mbo*I, *Taq*I and *Hae*III restriction enzyme sites (see section 1.1.b.iv, above). This examination revealed that *Hae*III cut a number of the novel tandem repeats reasonably often, but the *Mbo*I and *Taq*I sites were present in only one or two of the tandem repeats and at no time were they found to occur together in the same repeat array (data not shown). This indicated that DXYS78 is not contained within the N0434.1 interval and, as there is about 23.5 kb of sequence between the distal end of *PGPL* and the distal end of the N0434.1 interval (Figure 3.7, appendix 1), it further suggested that the distance between *PGPL* and DXYS78 is greater than previously thought. The relative distances between the novel N0434.1 tandem repeats and the PAR1 telomere were thereby calculated with the assumption that the distal (3′) end of *PGPL* was 110 kb from the telomere (table 3.3).

3.2.8.c: RepeatMasker

RepeatMasker (Smit, A.F.A. and Green, P., http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) screens DNA sequences against a library of repetitive elements and is primarily a tool that can be used to remove interspersed repetitive elements from a query sequence prior to BLAST searches. As the NIX analysis had indicated that the concentration of repetitive elements within the N0434 interval was particularly high, it was decided to analyse the full N0434.1 sequence using RepeatMasker to more closely identify what the repetitive elements might be. The vast majority of the interspersed repetitive elements in the N0434.1 interval were shown to be Alu sequences, as suggested by the NIX analysis. Numerous studies have demonstrated that human Alu sequences are composed of a small number of distinct subfamilies (e.g. Slagel *et al.*, 1987; Willard *et al.*, 1987; Britten *et al.*, 1988; Batzer *et al.*, 1995) and RepeatMasker was able to determine that the following subfamilies are represented in N0434.1: AluSx, AluY (10 copies of each), AluSq (5 copies), AluSp, AluSg (3 copies of each), AluSc, AluJb, AluJo (2 copies of each) and AluYc (1 copy). In addition there are five Alu elements that RepeatMasker could not assign to a single family, plus three FLAM_C monomers and one FRAM monomer. Therefore in total there are 47 Alu elements, which constitute 33% of N0434.1. As a typical Alu element is a dimeric structure about 300 bp long, only those over 280 bp were considered to be "full-length." This defined

25 of the Alu elements within N0434.1. Most of the Alu sequences appear to be randomly distributed across the region. but their density does fall to only 10% in the proximal 8 kb of N0434.1. However, this decrease does not coincide with any significant alteration of the overall GC content in the region.

There is only one long interspersed nuclear element (LINE) within the whole N0434.1 region (figure 3.7). It is a member of the L1 family but is very truncated at its 5′ end, with only about 400 bp of sequence, derived from the 3' end of a full-length 6–7 kb L1, present in the N0434.1 sequence.

In addition there are a small number of medium reiteration frequency (MER) elements and one transposable human element (THE-1) within the N0434.1 sequence (not shown).

Table 3.3: The tandem repeat sequences identified within the N0434.1 sequence assembly, which corresponds to an interval of approximately 33 kb residing about 85–120 kb from the PAR1 telomere.

Name [a]	*Distance from telomere (kb)* [b, c]	*Length of repeat unit (bp)*	*Number of repeat units in N0434.1* [d]	*GC content (%)*
NMS1	87 (*i*)	14	2.8	16
N(TA)1	87 (*ii*)	2	116	10
N(TG)2	88 (*iii*)	2	15.5	45
NMS2	89 (*iv*)	28	6.3	72
N(CA)3	90 (*v*)	2	80.5	52
N(CA)4	90 (*vi*)	2	61.5	52
NMS3	92 (*vii*)	29	2.4	51
NMS4	93 (*viii*)	29	11.8	13
PGMS3	96 (*ix*)	62	3.9	74
PGMS2	102 (*x*)	49	47	45
NMS5	106 (*xi*)	13	28.5	30
PGMS1	111 (*xii*)	53	21	39
(STIRs)	115	339	3.6	53
NMS6	116 (*xiii*)	31	7.8	72
NMS7	117 (*xiv*)	98	6.6	51

a: Nomenclature system for the novel tandem repeats is as follows: PGMS1, PGMS2 and PGMS3 are the structures that were discovered during N0434.1 sequence assembly and are named after the *PGPL* region, MiniSatellite 1–3. NMS1–8 were discovered after N0434.1 assembly with Tandem Repeats Finder and are named N0434.1 MiniSatellite 1–7. in a distal to proximal order. The names of microsatellites are N0434.1 (repeat unit) 1–4. in a distal to proximal order.

b: Distances were calculated with the assumption that the distal (3′) end of *PGPL* is 110 kb from the telomere

c: The italicised Roman numerals in this column refer to the positions of the tandem repeats within the N0434.1 sequence as shown in figure 3.7.

d: The variable lengths of of PGMS1. PGMS2 and the STIR repeat array have been analysed in a panel of unrelated UK men of North European descent. These results are presented in Chapter 6.

Figure 3.7: The genomic structure of the N0434.1 sequence. This diagram is presented in a similar manner to Figure 3.6. In this case N0434.1 is shown as the thick, central, orange line. Everything above the orange line is a feature found in the forward sense and everything below is a feature in the reverse sense. Alu elements are denoted by the red triangles and tandem repeats by the green blocks. The Roman numerals refer to descriptions of the tandem repeats within Table 3.3. The only LINE in the N0434.1 interval is marked by the black box and the position of the array of pseudoautosomal STIR elements is shown by the yellow box. The exon positions of both the *PGPL* and novel genes are marked with vertical blue bars.

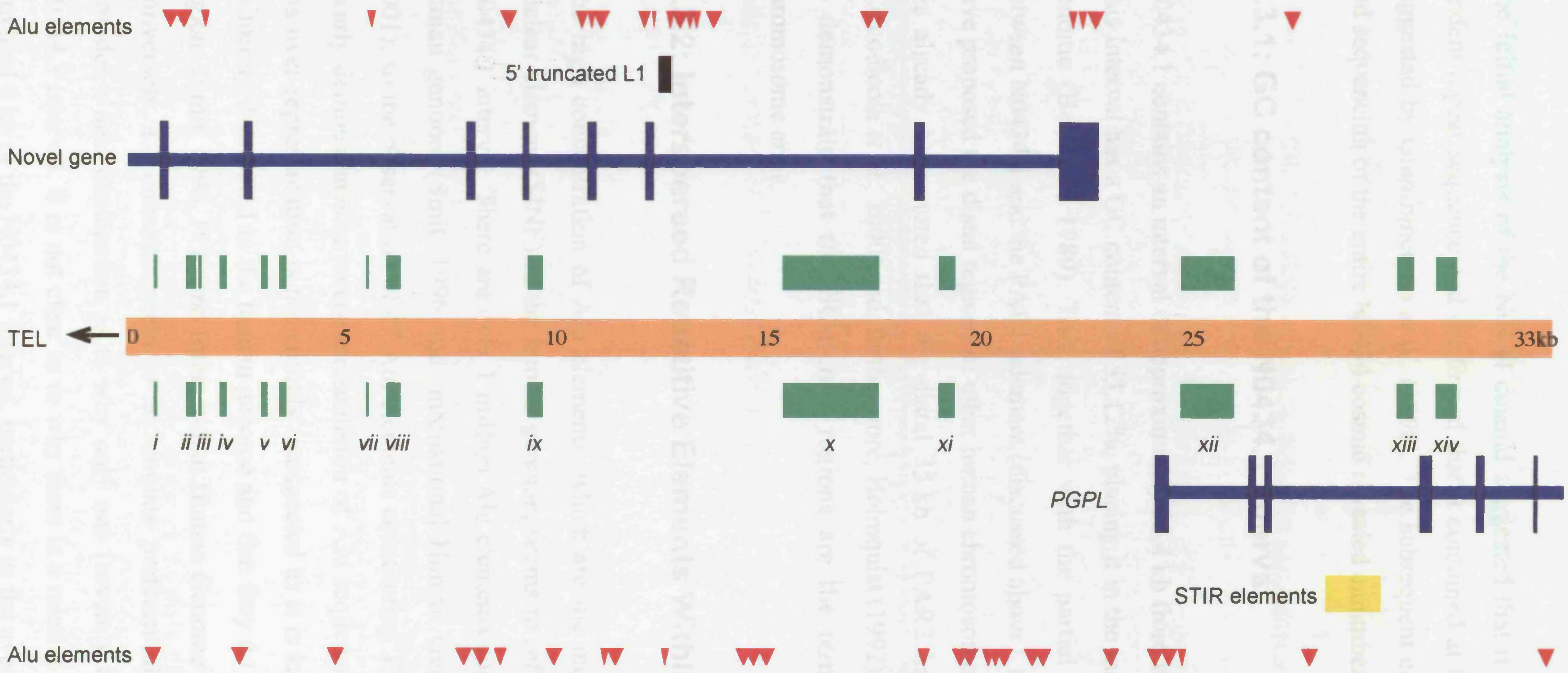


Figure 3.7: The genomic structure of the N0434.1 sequence

3.3: Discussion

The initial analysis of the N0434 cosmid suggested that it contained at least one variable tandem repeat sequence and confirmed that it contained at least part of the *PGPL* gene, as suggested by Gianfrancesco *et al.* (1998). The subsequent construction of a shotgun library and sequencing of the entire N0434 cosmid revealed a number of interesting factors.

3.3.1: GC content of the N0434.1 Interval

N0434.1 contains an interval of approximately 33.4 kb from within the distal region of PAR1. This interval has a GC content of 53.12%, placing it in the most GC-rich 3–5% of the human genome (Bernardi, 1989). This, together with the partial characterisation of the region between N0434.1 and the PAR1 telomere (discussed above), agrees with previous studies that have proposed the distal regions of other human chromosomes to be GC-rich. For example, it has already been noted that the distal 35 kb of PAR2 has a GC content of over 51% (Ciccodicola *et al.*, 2000) and, furthermore, Holmquist (1992) performed cytogenetic analyses to demonstrate that the GC-richest regions are the terminal regions of most human chromosome arms.

3.3.2: Interspersed Repetitive Elements Within the N0434.1 Interval

The high concentration of Alu elements, which are the most abundant short interspersed nuclear element (SINE) in the human genome, seems to reflect the high GC-content of the N0434.1 interval. There are over 1 million Alu elements comprising more than 10% of the human genome (Smit, 1996 and International Human Genome Sequencing Consortium, 2001), so the observation of 47 Alu elements constituting 33% of the DNA within N0434.1 clearly demonstrates an over-representation of Alu sequences within this region. However, this over-representation is not entirely unexpected as it is known that Alu elements are not uniformly distributed in the human genome and that they do preferentially locate in GC-rich regions (Smit, 1996, 1999 and International Human Genome Sequencing Consortium, 2001). Conversely, L1 elements seem to accumulate preferentially in GC-poor regions, which provides a neat explanation as to why only one (severely truncated) L1 was found in the N0434.1 interval. It is not clear as to why there is a relative scarcity of Alu elements in the proximal 8 kb of the N0434.1 interval, particularly as the average GC content of this region is actually higher than the N0434.1 interval as a whole. The region is far too small to be able to conclude that this represents a boundary between two PAR1 zones, as noted in PAR2 (Chapter 1, section 3.1). It remains quite possible that the noted difference in Alu density between the proximal 8 kb and distal 25 kb of the N0434.1 interval will not be quite so striking once sequence data for the surrounding region becomes available.

3.3.2.a: A Variety of Alu Subfamilies in the N0434.1 Interval

The variety of Alu subfamilies within the complete N0434.1 sequence is quite striking. The Alu family is primate-specific and is believed to have arisen from the tandem duplication of older 130 bp interspersed repeats called FLAM (free left Alu monomer) and FRAM (free right Alu monomer) (Jurka and Zuckerkandl, 1991; Quentin, 1992), both of which are represented in the N0434.1 interval. Dimeric Alu sequences are then classed as members of the AluJ or AluS subfamilies, based on nucleotide differences in at least 14 positions (Batzer *et al.*, 1996 and references therein). The AluJ subfamily includes the oldest Alu elements, with the AluS subfamily arising more recently and giving rise to AluY, the youngest Alu subfamily (Kapitonov and Jurka, 1996 and references therein). As representatives from each Alu subfamily are present in the N0434.1 interval, a closer examination may provide useful clues as to the evolution of the human PAR1.

3.3.3: Genes in the N0434.1 Interval

3.3.3.a: *PGPL*

The availability of the *PGPL* cDNA sequence allowed the identification of six *PGPL* exons, including the 3′ UTR. However, analysis also revealed that the 5′ 679 bp of the full-length 1867 bp cDNA sequence (as defined by Gianfrancesco *et al.*, 1998 and shown in figure 3.2) was not contained within the N0434 cosmid. Orientation of the N0434.1 sequences revealed that the missing portion of *PGPL* is situated proximal to the region covered by the 29C1, 3F3 and N0434 cosmids and thereby in the only identifiable gap in the cosmid contig that covers the distal 750 kb of PAR1 (figure 3.1*a,b*).

3.3.3.b: A Novel Gene

Perhaps the most interesting discovery made as a result of the *in silico* analyses was the identification of a novel gene, which reads in the opposite orientation to *PGPL* and ceases within 1 kb of the *PGPL* 3' UTR. A BLAST search (Altschul *et al.*, 1990) provided strong evidence that this gene is transcribed as it revealed that the exonic sequences of the novel gene were identical to a cDNA that is reported to be weakly similar to a 1-

phosphatidylinositol phosphodiesterase precursor. No genomic DNA matches could be found for this novel gene, which has been analysed in greater depth and is discussed further in Chapter 5.

3.3.4: Overlap Between the N0434 and 3F3 Cosmids

The experiments that were performed in order to determine the orientation of the N0434.1 sequences also verified that the N0434 and 3F3 cosmids overlap, and suggested that 3F3 extends approximately 12 kb into N0434.1. The 3F3 amplicons generated with the G2 and G9 primer pairs (table 3.2) suggested either that the G2 and G9 regions are variable tandem repeat loci that are duplicated in the 3F3 interval, or that the G2 and G9 primers reside in a kind of interspersed repeat, which is present in the 3F3 interval and sufficiently similar to the repeat in N0434 to allow amplification. Having identified the interspersed repeats within the N0434.1 interval, it was discovered that N24.1F, one of the G2 primers, resides in an Alu element and that N12.5R, one of the G9 primers, is located in the only L1 sequence in N0434.1. If the high density of interspersed repeats noted within N0434.1 continues into the interval covered by the 3F3 cosmid, this would provide ample opportunity for the N24.1F and N12.5R oligonucleotides to prime a PCR amplification within the 3F3 cosmid sequence, resulting in the amplicons noted in table 3.2. Clearly, this matter cannot be fully resolved without the complete sequence of the 3F3 cosmid itself.

3.3.5: Novel Tandem Repeats in the N0434.1 Interval

Fifteen novel tandem repeats have been discovered within the N0434.1 interval (table 3.3). Four are microsatellite sequences and consist of dinucleotide repeat units that stretch for no more than 232 bp; one is a tandemly repeated pseudoautosomal STIR element and the remaining ten tandem repeat arrays are minisatellites with repeat units of length 13–98 bp and GC contents ranging from 13–72%. It is interesting to note that the N0434.1 interval is an extension of the minisatellite-rich region that constitutes the distal 100 kb of PAR1, which suggests that the distal portion of PAR1, and perhaps PAR1 as a whole, is much more minisatellite-rich than previously thought. It also has further implications for a high resolution analysis of recombination within the N0434.1 interval, especially when considering the hotspot-driven instability of the MS32 minisatellite (Jeffreys *et al.*, 1998a). The length variability and repeat unit diversity of a subset of these novel minisatellites has been examined, and the results are presented in Chapter 6.

The identification of a novel pseudoautosomal STIR element that is similar, but not identical, to those described by Rouyer *et al.* (1990) raises further questions as to their putative role and supports a suggestion that the pseudoautosomal STIRs should be considered as a separate family to the autosomal STIRs. In light of this, the N0434.1 STIR has also been subjected to more analysis, the results of which are presented in Chapter 6.

3.4: Concluding Remarks

Sequencing of the N0434.1 cosmid has revealed some of the genomic structure of the *PGPL* gene and has putatively identified a novel gene that appears to be transcribed and is thus the most telomeric gene so far identified in PAR1. The N0434.1 interval is very GC-rich, which is reflected by the high density of Alu sequences in the region, and contains a number of novel tandem repeats. Most importantly, the N0434.1 sequence provides a tool for the discovery of SNPs, which can then be subjected to LD analysis as part of the high resolution analysis of recombination in the *PGPL* region.

Chapter 4

Extending and Characterising the Known Sequence in the PAR1 Telomere–*PGPL* Interval

4.1: Introduction

The work on the N0434.1 cosmid subclone, presented in Chapter 3, provided the first block of complete sequence information in the distal region of PAR1. This sequence provided the primary tool for an investigation of diversity in the *PGPL* region prior to a high resolution analysis of recombination, as detailed in Chapters 6, 7 and 8, but did not contain any of the tandemly repeated elements previously reported to exist in the PAR1 telomere–*PGPL* interval (Chapter 3, sections 3.1.1.a–c) (figure 4.1). Therefore, in hope of finding these elements and thus delineating more of the genomic structure of the PAR1 subtelomeric interval, an attempt was made to extend the known sequence towards the PAR1 telomere by sequencing the 3F3 cosmid. As the 3F3 cosmid had already been shown to overlap the N0434 cosmid by approximately 12 kb (table 3.2, figure 3.5), it was also possible that a comparison of the N0434.1 and 3F3 sequences could lead to the identification of a number of polymorphisms that might be of use in a subsequent recombination analysis.

4.1.1: A Brief Reminder of the Properties of the 3F3 cosmid

Cosmid 3F3 is derived from the Lawrence Livermore Y-chromosome specific library (LLN0YCO3′M′), for which the cosmid vector is Lawrist 16 (GenBank accession number = L19898), and forms part of a cosmid contig covering the distal 750 kb of PAR1 (Rao *et al.*, 1997). The 3F3 cosmid overlaps the distal part of N0434.1 (Rao *et al.*, 1997 and Chapter 3, section 2.7) and so purportedly covers a region that is approximately 40–80 kb from the PAR1 telomere. Therefore 3F3 is expected to contain several of the structures reported to exist within the distal 100 kb of PAR1 (Chapter 3, figure 3.1*b*).

Figure 4.1: The consensus sequences of the partially characterised minisatellites within the Xp/Yp telomere–*PGPL* interval. The loci are presented from the most distal to the most proximal, corresponding to the order provided by Page *et al.* (1987) and Vergnaud *et al.* (1993). The sequences read distal to proximal, as defined in the references that have been noted next to the relevant locus name.

DXYS14 (Ingleheam and Cooke, 1989)

CTCGGGACCACCCCAGACCCCCGCTCCTCCC

DXYS20 (Page *et al.*, 1987; Vergnaud *et al.*, 1993)

AAGGTTGCACAGTCTGCTCTCTATCTGTCCTCAATGAGACCTAGGCCCAATGCAGACTCTA

B4 (Vergnaud *et al.*, 1993)

TTGTCCCCACCAACATCCAGGGATGACC

CEB12 (repeat 2) (Vergnaud *et al.*, 1993)

GGAGGATGCACATG

CEB12 (repeat 1) (Vergnaud *et al.*, 1993)

CCCGAGACCCCCTCTTCCTGTCGCGG

CEB30 (Vergnaud *et al.*, 1993)

AGCTAGAGACAGTGGGGGT

4.2: Results

4.2.1: PCR Assay to Confirm Presence of Cosmid 3F3 DNA

As noted in Chapter 3, section 3.2.1, four single colony glycerol "subclones" of cosmid 3F3 (3F3.1–.4) had already been prepared. In addition, DNA had been extracted from each subclone and digested with the *Sfi*I restriction enzyme, a reaction that separates the Lawrist 16 vector from any insert sequence, thus demonstrating the presence of an insert within 3F3.1–.4. The PCR assays used to determine the extent of overlap between N0434.1 and 3F3 (Chapter 3, section 3.2.7) provided an assay to check that 3F3.1–.4 had not been contaminated with N0434.1 DNA, as the G2 and G9 primer pairs produced amplicons of a significantly different sizes when amplifying 3F3 DNA compared to when N0434.1 DNA was used as a template (Chapter 3, table 3.2). As expected, DNA prepared from 3F3.1–.4. produced G2 amplicons of 1.4 kb and G9 amplicons of 5 kb, the control amplifications from N0434.1 DNA produced amplicons of 2.2 kb and 0.7 kb respectively and negative control amplifications, which used 29C1, did not produce any amplicons (data not shown). The PCR conditions were as described in Chapter 2, tables 2.2*a*,*b* and Chapter 3, table 3.2. The assay demonstrated that 3F3 DNA had been successfully subcloned and was not contaminated with N0434 DNA

4.2.2: Restriction Enzyme Digest Assay of 3F3 Sub-clones

Given the number of tandem repeats that have been reported in the region (Cooke *et al.*, 1985; Page *et al.*, 1987; Armour *et al.*, 1990; Vergnaud *et al.*, 1993), and having already demonstrated the variability of N0434 inserts in *E. coli* (Chapter 3, section 3.2.3), the expected instability of the inserts in the 3F3 subclones was investigated by digestion of 3F3.1–.4 with the *Bgl*I, *Dra*I, *Pst*I and *Pvu*II restriction enzymes. The digests were then subjected to agarose gel electrophoresis in order to produce a digest profile (figure 4.2).

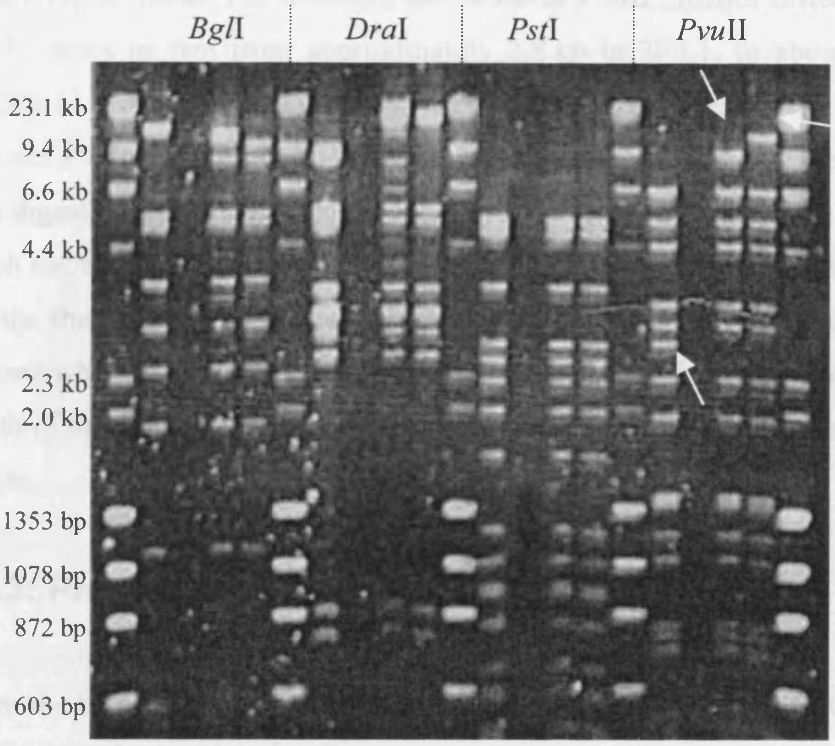


Figure 4.2: Digestion of 3F3 sub-clones .1–.4 using *Bgl*I, *Dra*I, *Pst*I and *Pvu*II restriction enzymes. The digests have been subjected to agarose gel electrophoresis with λ x *Hin*dIII and ϕx174 x *Hae*III standard weight molecular markers. The absence of bands in the 3F3.2 lanes was explained in a subsequent assay of the concentrations of each subclone DNA preparation, which demonstrated that the preparation of 3F3.2 was 25 times less concentrated than originally thought (data not shown). Nevertheless, the profiles of 3F3.1, .3 and .4 provide clear evidence for a level of variability between the 3F3 inserts. The *Pvu*II profiles differ by just one band of variable size, marked by the white arrows, which is consistent with the presence of an unstable tandem repeat array. Digestion with *Bgl*I and *Dra*I was only partial but these profiles are concordant with the presence of at least one variable tandem repeat locus. A suggestion to explain the identical profiles of the *Pst*I digests is provided in the main text.

4.2.2.a: Variability Between 3F3.1–.4

The digest profiles, whilst similar to each other, showed differences that indicated a level of variability between 3F3 inserts that was consistent with the presence of at least one variable tandem repeat locus. For instance, the 3F3.1–.4 *Pvu*II profiles differ by just one fragment, which varies in size from approximately 2.8 kb in 3F3.1, to about 9.4 kb in 3F3.3, and roughly 13 kb in 3F3.4 (figure 4.2). The *Bgl*I and *Dra*I digests were incomplete but are also consistent with the presence of tandem repeat loci. In addition, the identical patterns obtained after digestion with *Pst*I could be explained if 3F3 contained a large tandem repeat array in which each repeat unit contained a *Pst*I site, a possibility that appears to be supported by the fact the the length estimations of 3F3.1–.4 are significantly shorter using the *Pst*I data than they are when using the profiles of the *Pvu*II digests; e.g. for 3F3.1, the *Pst*I profile indicates a length of approximately 33 kb, compared to a length of 38 kb suggested by the *Pvu*II digest profile.

4.2.3: Preparation of the 3F3.1 Shotgun Library

Given the heterogeneity of the 3F3 cosmid in *E. coli*, it was decided that the simplest way to sequence 3F3 would be to focus exclusively on the shortest 3F3 subclone, 3F3.1. The construction of the 3F3.1 shotgun library followed much the same pattern as the assembly of the N0434.1 library (Chapter 3, section 3.2.4). As a result, only the differences between the two libraries, at specific stages of the construction process are included here. An aliquot of the 3F3.1 DNA preparation was digested with the *Pst*I and *Pvu*II restriction enzymes to confirm that the preparation had yielded 3F3.1 DNA. As the N0434.1 shotgun library of 192 colonies was insufficient for full coverage of the N0434.1 sequence, a 3F3.1 library of 238 colonies, consisting of 15 controls and 223 recombinant colonies expected to contain a 1–2 kb insert, was constructed and should have been adequate for full sequence coverage of the 3F3.1 sequence.

Before any 3F3.1 colonies were picked into the ordered array that constituted the library, twelve supposedly recombinant colonies were selected at random to check for the presence of an inserted 3F3.1 fragment by plasmid DNA preparation (Sambrook *et al.*, 1989) and digestion with *Kpn*I and *Bam*HI to cut the insert out of the pBluescript SK$^+$ vector. The digests were then subjected to agarose gel electrophoresis to determine the size of the insert. Only ten of the twelve colonies were shown to contain an insert, suggesting that only about 80% of the 223 putatively recombinant colonies within the 3F3.1 shotgun library actually contained an insert derived from the 3F3.1 subclone.

4.2.4: Sequencing of 3F3.1

4.2.4.a: Sizing the shotgun library inserts

Double-stranded DNA was prepared from the library clones and the inserts were amplified using the SK and KS primers (Chapter 2, table 2.2*a*). The sizes of the amplified inserts were then determined by agarose gel electrophoresis (table 4.1). Only 70% of the shotgun library recombinant colonies contained an amplifiable insert and four inserts (2A4, 2A7, 3C3 and 4A11) were only 0.2–0.3 kb in length. As the construction of the shotgun library involved selecting 1–2 kb DNA fragments of cosmid (Chapter 2, section 2.2.9.b) it was unclear as to what these short inserts represented. However, if 3F3.1 does contain variable tandem repeat arrays, it is quite possible that 0.2–0.3 kb inserts were fragments of the putative tandem repeat that were unstable in the system used to construct the shotgun library.

4.2.4.b: Sequencing strategy

All of the amplified inserts were purified by electroelution and ethanol precipitation and first sequenced with the KS primer. As for the sequencing of N0434.1, all of the retrieved 3F3.1 sequences were aligned, *in silico*, against the full Lawrist 16 sequence to ensure that no Lawrist 16 sequences were included in the assembly of the 3F3.1 sequence. In addition, novel 3F3.1 sequences were aligned against the complete N0434.1 sequence in order to more accurately establish the amount of overlap between the 3F3 and N0434 cosmids, and also to identify any differences that might exist between the two cosmids in their overlapping regions. If appropriate, the inserts were then sequenced with the SK primer. All sequence reads were assembled using ABI AutoAssembler software.

Assembling the 3F3.1 sequence was a difficult challenge, particularly as many inserts produced incomprehensible sequence reads. Evidently, repeat DNA was the source of this problem, as many sequence reads appeared to stutter and slip and superimpose on themselves. Up to this point all sequencing reactions had been carried out with an early version of the ABI PRISM BigDye^{TM} Terminator Cycle Sequencing Ready Reaction Kit (Chapter 2, section 2.2.10.a), which had been adequate for N0434.1 sequencing. However, in order to combat the problems that were arising during the assembly of 3F3.1, the most problematic library clones were resequenced using BigDye^{TM} version 3.1, which, according to the manufacturers, is able to read more effectively through regions of repetitive DNA and is usually able to extend a read beyond the upper limits of earlier BigDye^{TM} versions. In many cases improved sequence reads were obtained, but approximately 12% of the library clones that contained an amplifiable insert remained impossible to sequence.

4.2.4.c: The Final Assembly of the 3F3.1 Shotgun Library Sequences

Given these problems, the final assembly of 3F3.1 sequences within the library did not produce a contiguous sequence that covered the whole of the 3F3.1 interval. Instead, approximately 75 kb of sequence reads accumulated into 14 contigs of length 0.5–3.0 kb, and a further 10 kb of sequence reads did not align with any of the contigs. In some cases the KS and SK reads of one or more library inserts assembled at the ends of different contigs, which meant that the 14 contigs could be assembled into 9 non-contiguous reads that stretched over 0.5–6.5 kb, the longest of which included a novel minisatellite array, provisionally entitled 3FMS1, which appeared to stretch over 3.3 kb. The length of 3F3.1 had already been estimated at 38 kb (section 4.2.1, above), which meant that the sum total of the 3F3.1 shotgun library assembly covered only 60% of the 3F3 interval. Despite the poor coverage it was possible to accurately determine that the amount of N0434.1 sequence overlapped by the 3F3.1 cosmid was 12,238 bp (Appendix 1), in full agreement with the estimate of 12 kb drawn from the studies presented in Chapter 3, section 3.2.7. Therefore, with reference to Chapter 3, table 3.3, the 3F3.1 sequence includes the nine most distal tandem repeat sequences identified within N0434.1 as well as at least one major minisatellite locus in the 25–26 kb portion of 3F3.1 that lies distal to the N0434.1 interval. At this stage, it was decided to move on from primary sequencing in order to concentrate on the analysis of recombination in the *PGPL* region. In the meantime, it was hoped that the Human Genome Sequencing Consortium would finally address PAR1. Therefore, conatct was made with Dr Mark Ross, head of the X chromosome sequencing project at the Sanger Centre, Cambridge, UK, with the understanding that he would provide updates on the progress of any PAR1 sequencing project with particular reference to the PAR1 telomere–*PGPL* interval.

Table 4.1: Sizes of inserts in the 3F3.1 shotgun library. The library is an ordered array of recombinant colonies picked into five 96-well plates. The name of a colony reflects its position within the ordered array. Each plate contained three internal controls, marked within the table as 'non-recombinant,' which denoted those colonies that did not contain recombinant pBluescript SK$^+$ vectors, or 'no colony,' which indicated those wells that did not have a colony picked into them. PCR amplification with the KS and SK primers demonstrated that only 70% of the colonies within the library contained an amplifiable 3F3.1 insert. As was the case with the N0434.1 library, a small proportion of the colonies produced two fragments after amplification, which could have been caused through mixture of two recombinant colonies. In these cases the fragments were separated, sequenced as normal and included in the 3F3.1 sequence assembly.

clone	*insert size (kb)*	*comments*	*clone*	*insert size (kb)*	*comments*	*clone*	*insert size (kb)*	*comments*
1A1		Non-recombinant	3C7	1.4		4C5	1.2	
1A2	1.3		3C8	1.3		4C6	1.9	
1A3	1.2		3C9	1.4		4C7	1.4	
1A4	1.5		3C10	—		4C8	2.2	
1A5	1.2		3C11	—		4C9	1.8	
1A6	1.7		3C12	2.0		4C10	1.8	
1A7	1.2		3D1	1.1		4C11	1.7	
1A8	1.4		3D2	1.4		4C12	1.2	
1A9	—		3D3	1.1		4D1	1.5	
1A10	1.6		3D4	—		4D2	1.5	
1A11	—		3D5	—		4D3	1.9	
1A12	1.8		3D6	1.6		4D4	—	
1B1	1.4		3D7	1.6		4D5	1.5	
1B2	1.2		3D8	1.1		4D6	1.5	
1B3	1.6		3D9	1.4		4D7	1.1	
1B4	1.6		3D10	—		4D8	1.4	
1B5	1.6		3D11	0.8		4D9	—	
1B6	1.2		3D12	2.1		4D10	1.8	
1B7	1.5		3E1	—		4D11	1.7/2.0	double
1B8	1.7		3E2	1.9		4D12	1.7	
1H11	—	Non-recombinant	3E3	1.2		4E1	1.1	
1H12	—	No colony	3E4	1.2		4E2	1.6	
2A1	—	Non-recombinant	3E5	1.2		4E3	1.9	
2A2	1.5		3E6	1.1		4E4	1.0	
2A3	1.1		3E7	1.3		4E5	1.5	
2A4	0.6		3E8	2.2		4E6	1.8	
2A5	1.8		3E9	—		4E7	1.4	
2A6	1.1		3E10	—		4E8	1.9	
2A7	0.2		3E11	1.1		4E9	1.4	
2A8	0.9		3E12	3.4		4E10	—	
2A9	1.1		3F1	—		4E11	1.4	
2A10	1.3/1.3	double	3F2	—		4E12	1.8	
2A11	1.4		3F3	—		4F1	1.8	
2A12	1.7		3F4	—		4F2	1.4	
2B1	1.4		3F5	1.1		4F3	1.6	
2B2	1.5		3F6	1.1		4F4	1.3	
2B3	1.5		3F7	1.2		4H11	—	Non-recombinant
2B4	—		3F8	—		4H12	—	No colony
2B5	—		3F9	1.1/2.0	double	5A1	—	Non-recombinant
2B6	2.0		3F10	1.1		5A2	1.0	
2H11	—	Non-recombinant	3F11	1.0/2.2	double	5A3	—	
2H12	—	No colony	3F12	1.9		5A4	—	
3A1	—	Non-recombinant	3H11	—	Non-recombinant	5A5	1.7	
3A2	1.2		3H12	—	No colony	5A6	1.2	
3A3	—		4A1	—	Non-recombinant	5A7	1.0	
3A4	—		4A2	1.2		5A8	1.8	
3A5	1.2		4A3	—		5A9	1.4	
3A6	—		4A4	—		5A10	1.9	
3A7	—		4A5	—		5A11	2.0/2.3	double
3A8	1.4		4A6	1.0		5A12	1.6	
3A9	1.2		4A7	—		5B1	—	
3A10	—		4A8	—		5B2	1.2	
3A11	—		4A9	—		5B3	—	
3A12	—		4A10	—		5B4	1.8	
3B1	—		4A11	0.2		5B5	—	
3B2	—		4A12	1.3		5B6	1.1	
3B3	—		4B1	—		5B7	1.8	

Table 4.1 (*continued*): Sizes of inserts in the 3F3.1 shotgun library.

clone	***insert size (kb)***	***Comments***	***clone***	***insert size (kb)***	***comments***	***clone***	***Insert size (kb)***	***comments***
3B4	—		4B2	1.2		5B8	1.8	
3B5	1.3		4B3	1.6		5B9	—	
3B6	1.3		4B4	1.6		5B10	1.2	
3B7	1.8		4B5	—		5B11	—	
3B8	—		4B6	0.9/1.3	double	5B12	—	
3B9	1.3		4B7	1.2		5C1	—	
3B10	—		4B8	1.9		5C2	—	
3B11	—		4B9	1.8		5C3	1.3	
3B12	—		4B10	1.4		5C4	—	
3C1	1.3		4B11	1.3		5C5	1.2	
3C2	1.1		4B12	1.6		5C6	—	
3C3	0.3		4C1	—		5C7	—	
3C4	1.0		4C2	1.4		5C8	1.6	
3C5			4C3	1.3		5C9	1.5	
3C6	1.3		4C4	1.3		5C10	—	
5C11	—		5D7	—		5E3	—	
5C12	—		5D8	1.8		5E4	—	
5D1	1.3		5D9	—		5E5	1.1	
5D2	1.3		5D10	1.3		5E6	1.8	
5D3	2.1		5D11	2.2		5H11	—	Non-recombinant
5D4	1.3		5D12	1.2		5H12	—	No colony
5D5	1.3		5E1	2.0				
5D6	—		5E2	1.3				

4.2.5: Distal PAR1 Sequences Produced at the Sanger Centre

Very recently (June 2003), the Sanger Centre produced sequence data for almost the entire PAR1. However, they too found that sequencing the region was problematic, and it was thus rather pleasing for me to note that the Sanger Centre had found the process sufficiently difficult and frustrating for Dr Mark Ross to include the phrase, "What a bugger!" in his latest update on the progress in and around the *PGPL* region. The Sanger Centre sequences corresponding to the PAR1 telomere–*PGPL* interval were isolated, aligned with the N0434.1 sequence in order to determine their orientation, and subjected to *in silico* analyses, as detailed below, in order to identify the structures shown in figure 3.1*b*.

4.2.5.a: The 3F3 Region

Attempts to sequence the entire 3F3 cosmid at the Sanger Centre also appear to have been unsuccessful (M. Ross, personal communication). This led them to use a fosmid clone to sequence this interval, which they called G248P86101C3 (hereafter referred to as C3), the sequence of which is available under the GenBank accession number BX537334.

4.2.5.a.i: Length, Orientation and GC Composition of the C3 Sequence

The ABI AutoAssembler software was used to align the C3 sequence against the 3F3.1 sequence reads and a number of overlapping 2 kb fragments from the N0434.1 sequence. This established that the C3 sequence overlapped the distal end of N0434.1 by 19,518 bp and so included the PGMS2 and NMS6 minisatellites, in addition to the ten already shown to exist within 3F3.1 (this chapter, section 4.2.3.c). Given the fact that the orientation of the N0434.1 sequence had been determined (Chapter 3, section 3.2.7) it was also possible to ascertain that the database C3 sequence read in a proximal to distal orientation. The C3 sequence that did not overlap N0434.1 totalled 23,740 bp and was found to have an average GC content of 58.69%, providing ample evidence that the distal 100 kb of PAR1 is part of the GC-richest class of human DNA (Bernardi, 1989), as suggested in the previous chapter. Only the 23,740 bp of C3 DNA that did not overlap with N0434.1 was subjected to *in silico* analysis (this distal portion of C3 is hereafter referred to as C3A).

4.2.5.a.ii: NIX Analysis

The results of the NIX analysis, again accessed through the MRC UK Human Genome Mapping Project Resource Centre (http://www.hgmp.mrc.ac.uk/), can be seen in figure 4.3. There are no strong predictions for the presence of any exonic sequences, but a closer examination of the C3A sequence did reveal a short (66 bp) exon, which corresponded to the most 5′ end of the uncharacterised cDNA that matched the putative gene in the N0434.1 interval. This short segment had not been identified in the N0434.1 sequence. Therefore, the novel gene is 26 kb long and consists of nine exons. The NIX analysis also revealed that the relatively high density of Alu elements observed in the N0434.1 interval stretches into the C3A region.

4.2.5.a.iii: Tandem Repeats Finder

The heterogeneity of the different 3F3 cosmid subclones and the problematic 3F3.1 sequencing project had already suggested that the C3A sequence contained at least one other major minisatellite locus, in addition to those in the proximal part of C3 that overlapped with N0434.1. For that reason, and in an attempt to locate previously characterised tandem repeat arrays, the C3A sequence was analysed using Tandem Repeats Finder (http://tandem.biomath.mssm.edu/trf /trf.html, Benson, 1999). This indicated that C3A contained another nine novel tandem repeat sequences, plus the CEB30 tandem repeat already identified by Vergnaud *et al.* (1993) (figure 4.4, table 4.2). Hence, tandem repeat arrays cover 39% of the C3A sequence, showing the region to be more minisatellite-rich than the proximal N0434.1 interval.

4.2.5.a.iv: RepeatMasker

RepeatMasker (Smit, A.F.A. and Green, P., http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) identified 20 different Alu elements within the novel C3A sequence, consisting of five members of the AluSx subfamily, four copies of both the AluY and AluSq subfamilies, three AluSg elements, two AluJo elements plus one AluS and one Alu element that RepeatMasker could not assign to a single family. The density of Alu elements within C3A is 21%, still significantly greater than genome average of approximately 10% (International Human Genome Sequencing Consortium, 2001) but also much less than the density of Alu elements in the N0434.1 interval, as indicated by the NIX analysis (figure 4.3). Of the 20 Alu elements within C3A, 14 are "full-length," as defined by the criteria described in Chapter 3, section 2.8.c. The majority of the Alu sequences within C3A are clustered in the most distal 10 kb of the region (figure 4.4), where their density is actually 41%. This does not correspond to a significant shift in GC content but is due to the fact that the proximal part of

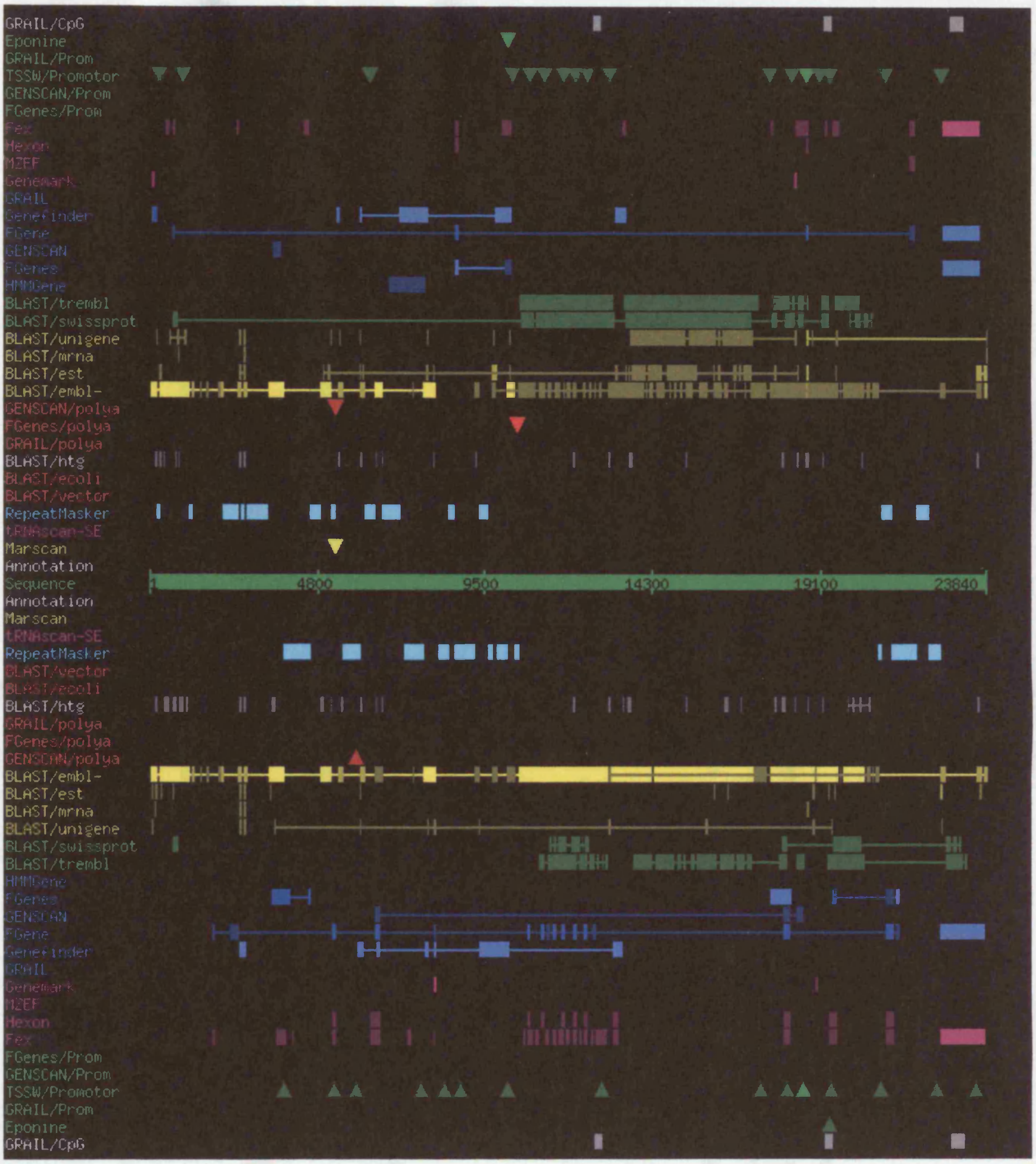


Figure 4.3: The NIX analysis of the portion of the Sanger Centre fosmid clone G248P86101C3 that lies distal to the N0434.1 sequence. The relatively high density of interspersed repeats in each orientation is denoted by the cyan blocks. Also see figure 4.5.

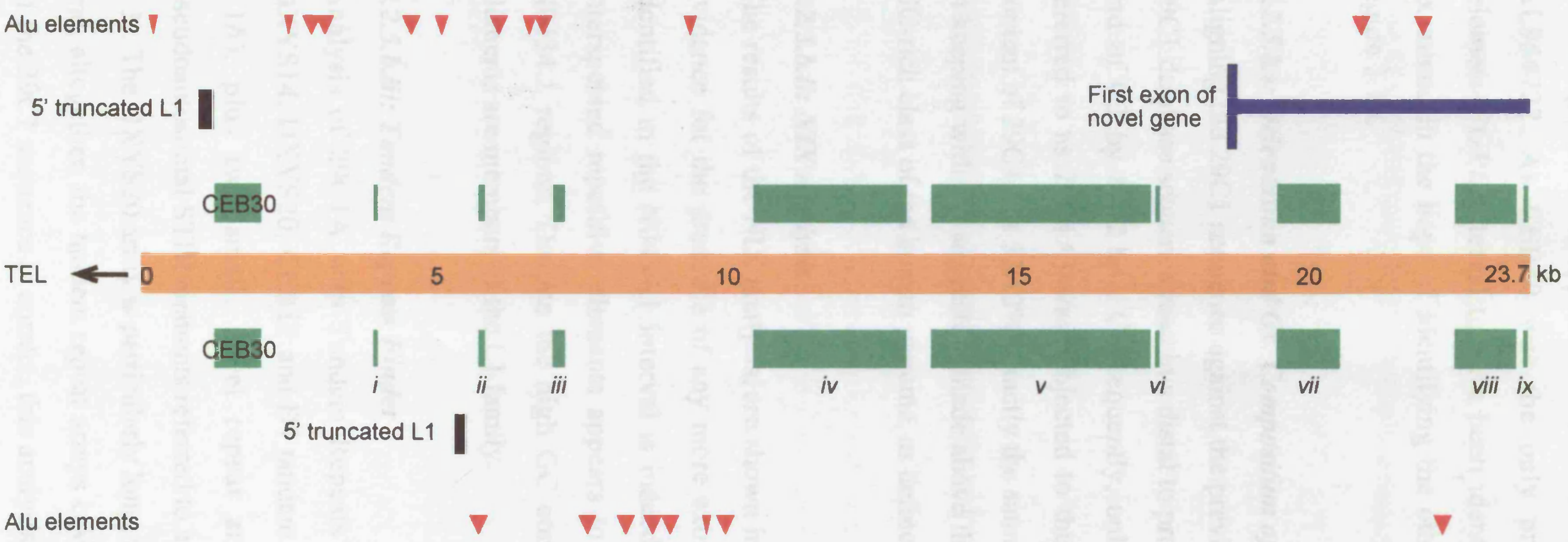


Figure 4.4: The genomic structure of the C3A sequence. Everything above the orange line is a feature found in the forward sense and everything below is a feature in the reverse sense. Alu elements are denoted by the red triangles and tandem repeats by the green blocks. CEB30 (Vergnaud *et al.*, 1993) is labelled and the Roman numerals refer to descriptions of the tandem repeats within table 4.2. The first exon of the novel gene identified during the sequencing of N0434.1 is marked by the vertical blue bar.

4.2.5.b: The 29C1 Region

Preliminary observations of the 29C1 cosmid suggested that it was highly unstable in *E. coli* (Chapter 3, section 2.1). Nevertheless, the Sanger Centre were able to fully sequence 29C1, producing a read of approximately 37.8 kb, available under GenBank accession number AL954722. As CEB30 was the only previously characterised locus in the PAR1 telomere–*PGPL* interval to have been identified within the C3 sequence, 29C1 was also examined in the hope of identifying the other previously characterised loci as laid out in figure 3.1*b*.

4.2.5.b.i: Orientation and GC Composition of the 29C1 Sequence

Aligning the 29C1 sequence against the previously oriented C3 sequence established that the 29C1 database sequence read in a distal to proximal direction and that it overlapped the distal end of C3 by 8252 bp . Consequently, only the proximal 29,522 bp of 29C1 (hereafter referred to as 29C1A) was subjected to the *in silico* analyses presented below. The GC content of 29C1A is 53.12%, exactly the same as the GC content of the N0434.1 interval, and in keeping with the suggestion made above that the distal 100 kb of PAR1 is part of the most GC-rich class of the human genome, as defined by Bernardi (1989).

4.2.5.b.ii: NIX Analysis

The results of the NIX analysis, are shown in figure 4.5. This analysis has not provided any evidence for the presence of any more exonic sequences, suggesting that the novel gene identified in the N0434.1 interval is indeed the most telomeric in PAR1. The density of interspersed repetitive elements appears to be relatively low compared to the C3A and N0434.1 regions. Despite the high GC content of 29C1A, a number of these repetitive elements are members of the L1 family.

4.2.5.b.iii: Tandem Repeats Finder

Analysis of 29C1A with Tandem Repeats Finder revealed the previously characterised DXYS14, DXYS20, CEB12 and B4 tandem repeat arrays (figure 4.1 and Chapter 3, figure 3.1*b*), plus two small, novel repeat arrays and the tandemly repeated array of pseudoautosomal STIR elements referred to as 362F by Rouyer *et al.* (1990) (figure 4.6, table 4.3). The DXYS20 array is particularly long, representing nearly 40% of the 29C1A sequence and, altogether, the tandem repeat arrays cover 53% of 29C1A. Assuming that the assembly of the 29C1 sequence is correct, this analysis has also revealed that the order and orientation of the CEB12 and B4 arrays, as defined by Vergnaud *et al.* (1993) and as presented in figures 3.1*b* and 4.1, is wrong. As shown in figure 4.6, the distal to proximal order of these arrays is actually CEB12 repeat 1, CEB12 repeat 2 and then B4. Similarly, the consensus sequences of these repeats, shown in figure 4.1 as reading distal to proximal (as defined by Vergnaud *et al.*, 1993), actually read in a proximal to distal direction. The original positioning of CEB12 and B4 took place via restriction enzyme mapping (Vergnaud *et al.*, 1993) and, as CEB12 and B4 are *Bam*HI fragments, it is quite conceivable that they were incorrectly oriented with respect to each other and the more centromeric 362F pseudoautosomal STIRs and CEB30.

4.2.5.b.iv: RepeatMasker

RepeatMasker was used to more closely identify the repetitive elements indicated by the NIX analysis. A total of 19 different Alu elements, covering just 15% of the 29C1A sequence was identified. As was the case for both the N0434.1 and C3A sequences, a wide range of Alu subfamilies were represented in the 29C1A sequence, including AluY (6 copies), AluSg, AluSq, AluSx (2 copies of each) and one copy each of the AluJb, AluJo, AluS and AluSp sequences. In addition there were two Alu elements that RepeatMasker could not assign to a single family and one FLAM_A monomer. Eleven of the 29C1A Alu elements are full-length. As was observed in the C3A sequence, the Alu elements are not randomly distributed in the 29C1A interval but, as is clear from figure 4.6, this is due to the large proportion of the 29C1A sequence that is taken up by tandem repeat arrays. In addition to the Alu elements and tandem repeat arrays, there are four 5′-truncated L1 elements (one of which is interrupted by an Alu sequence), which means that nearly 75% of 29C1A is composed of repetitive sequences (figure 4.6).

Table 4.3: The tandem repeats identified within the Sanger Centre 29C1A sequence.

Name [a]	*Length of repeat unit (bp)*	*Number of repeat units in 29C1*	*GC content (%)*
29(TYCC)1	4	14	61
DXYS14	31	44.5	72
29MS1	10	3.4	65
DXYS20	61	185	49
CEB12 (1)	26	49	67
CEB12 (2)	14	4	53
B4	28	34	51
(STIRs)	340	2	53

a: Nomenclature system for the novel tandem repeats is as follows: 29MS1 was named after the 29C1 region, MiniSatellite 1. The microsatellite is 29 (repeat unit) 1.

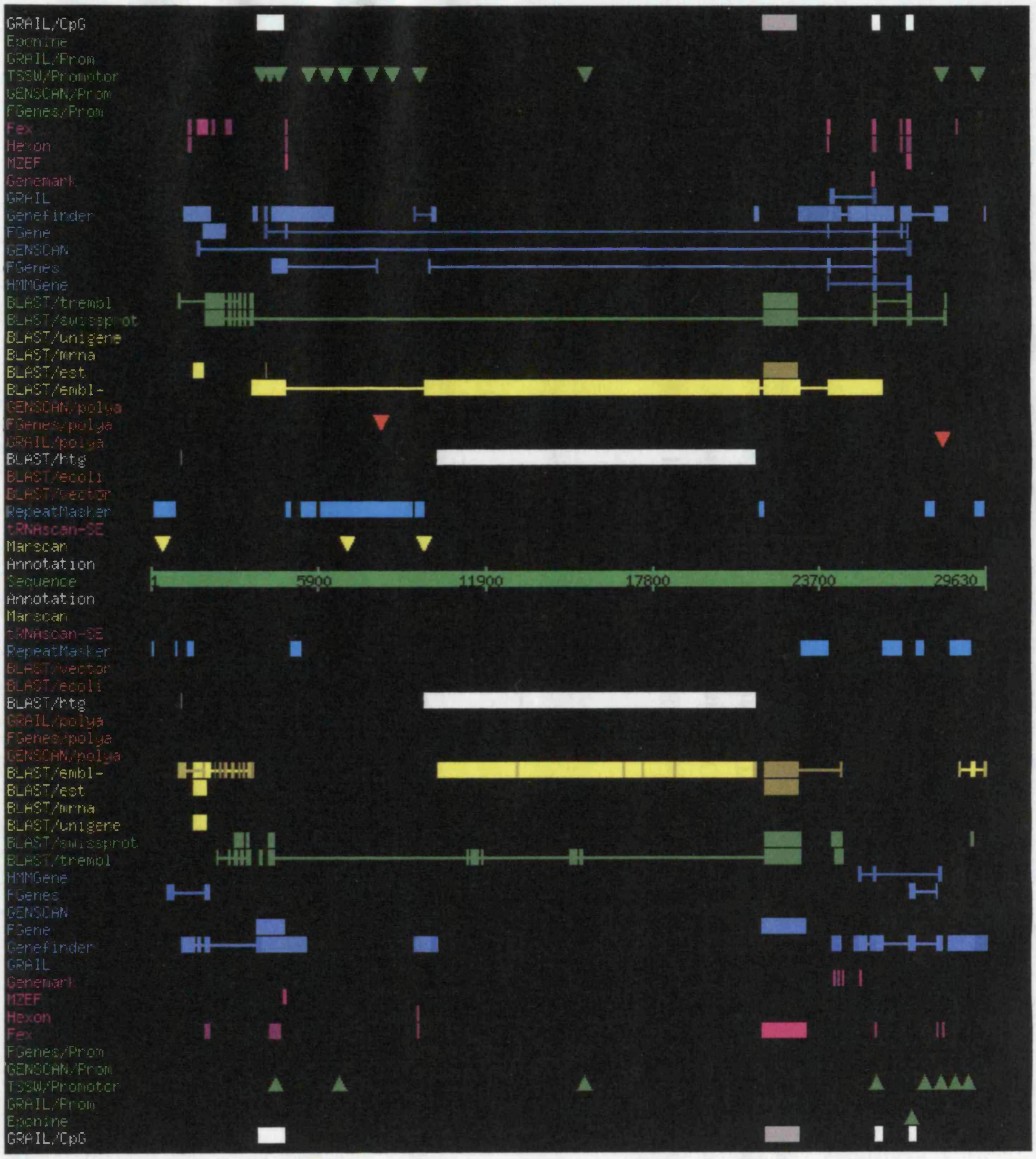


Figure 4.5: The NIX analysis of 29C1A, which lies distal to the Sanger Centre fosmid clone G248P86101C3.

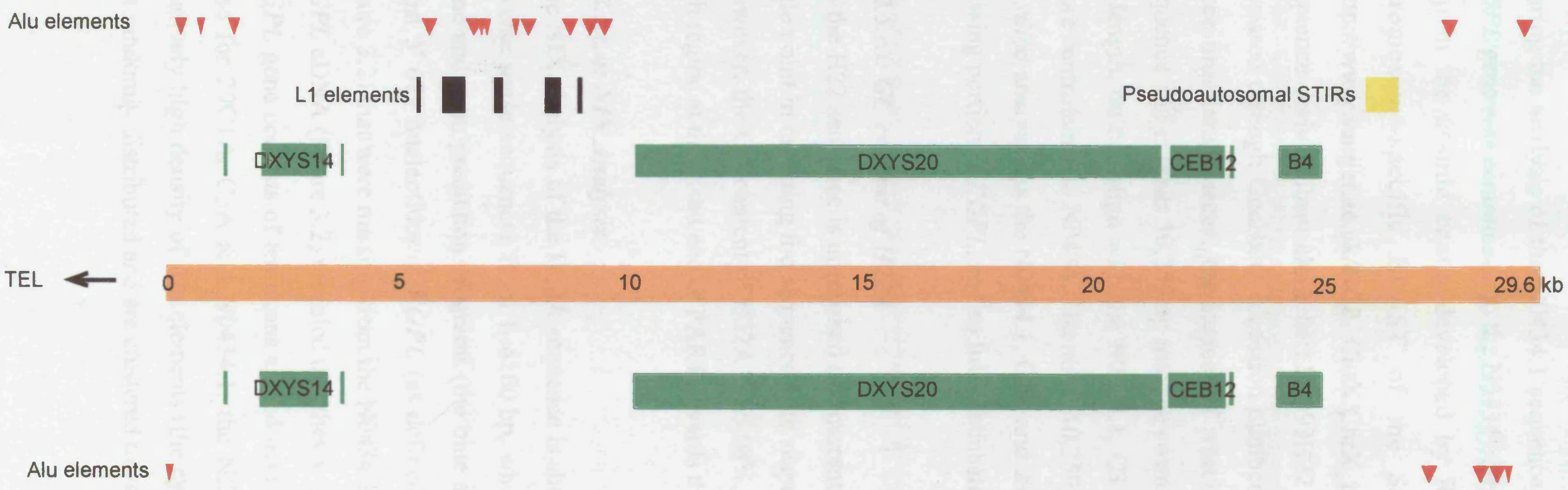


Figure 4.6: The genomic structure of 29C1A. Everything above the orange line is a feature found in the forward sense and everything below is a feature in the reverse sense. Alu elements are denoted by the red triangles and tandem repeats by the green blocks. The positions of L1 elements are marked with the black bars and the pseudoautosomal STIRs originally identified by Rouyer *et al.* (1990) are shown in yellow. Tandem repeat arrays that have already been characterised (see main text) are labelled and further information on these and the two small arrays that flank DXYS14 is provided in table 4.3.

4.2.5.c: Novel Sequence Proximal to N0434.1

During the analysis of the N0434.1 sequence it was noted that a significant portion of the *PGPL* gene was centromeric to the N0434 cosmid, and was thus expected to reside in the only gap in the cosmid contig developed by Rao *et al.* (1997) (figure 3.1). However, a chromosome-specific BLAST of the Sanger Centre X chromosome sequence (http://www.sanger.ac.uk/HGP/ ChrX/ChrX_blast_server.shtml), revealed some unfinished sequence derived from clone RP11-1293H22 (hereafter referred to as H22), which can be accessed through GenBank accession number BX546484. Currently, this clone consists of three unordered pieces, the largest of which overlaps the proximal end of the N0434.1 sequence and extends 10,259 bp into the cosmid contig gap. The other two pieces are 5–6 kb in length, do not align with the N0434.1, C3 or 29C1 sequences and so appear to lie even more centromeric to N0434. The novel 10,259 bp of H22 (H22A) were subjected to the same *in silico* analyses as the N0434.1, C3A and 29C1A sequences in the hope of identifying the missing portion of *PGPL*, and so fully delineating its genomic structure.

4.2.5.c.i: GC content of H22A

As the H22 sequence is unfinished and currently consists of three unordered pieces there was little point in orienting the sequence with respect to N0434.1 and the other PAR1 sequences. However, the GC content of H22A is 55.08%, which for the first time extends the very GC-rich region at the distal end of PAR1 towards the centromere.

4.2.5.c.ii: NIX Analysis

The NIX analysis of the H22A sequence is shown in figure 4.7. A closer examination of the reverse sense sequences from 1–4100 bp, where there is very good agreement between the gene and exon prediction programs (the blue and purple colours in figure 4.7), identified the most 5′ 679 nucleotides of *PGPL* (as defined by Gianfrancesco *et al.*, 1998 and shown in figure 3.2) that were missing from the N0434.1 sequence. Aligning these nucleotides with the *PGPL* cDNA (figure 3.2) revealed that they were split into four exons, meaning that the entire *PGPL* gene consists of ten exons spread over a genomic region of 14.3 kb. As has been the case for 29C1A, C3A and N0434.1, the NIX analysis has shown that H22A also has a relatively high density of Alu elements (the cyan colour). However, these Alu sequences are not randomly distributed and are clustered towards the proximal end of H22.

4.2.5.c.iii: Tandem Repeats Finder

This analysis identified another twelve novel tandem repeat arrays covering 33% of H22A (figure 4.8, table 4.4) and thereby provided compelling evidence that the minisatellite-rich region extends from the PAR1 telomere-adjacent interval and beyond *PGPL* towards the centromere.

4.2.5.c.iv: RepeatMasker

In order to identify the Alu subfamilies within H22A, the sequence was analysed with RepeatMasker, which recognised five AluJo elements and one copy each of the FLAM_A, AluJb, AluSg, AluSp and AluY subfamilies. As indicated by the NIX analysis (figure 4.7), these ten Alu elements were mainly clustered in the proximal 3.5 kb of H22A, which coincides with both a decrease in GC content to 43% and a lower density of tandem repeat arrays.

Table 4.4: The tandem repeats identified within the Sanger Centre H22A sequence.

Name [a, b]	*Length of repeat unit (bp)*	*Number of repeat units in 29C1*	*GC content (%)*
H22MS1 (*i*)	10	2.7	63
H22MS2 (*ii*)	65	16.1	48
H22(CACCC)1 (*iii*)	5	15	78
H22MS3 (*iv*)	38	6.4	63
H22MS4 (*v*)	90	3.2	68
H22(CGG)2 (*vi*)	3	9	96
H22MS5 (*vii*)	62	13.6	73
H22MS6 (*viii*)	36	2.6	39
H22(TATAA)3 (*ix*)	5	89	5
H22(TATAA)4 (*x*)	5	13	3
H22(TATAA)5 (*xi*)	5	28	6
H22(GGAA)6 (*xii*)	4	44	56

a: Nomenclature system for the novel tandem repeats is as follows: The minisatellites are named after the H22 region, MiniSatellite 1–6 in a distal to proximal order. The microsatellites are H22 (repeat unit) 1–6 in a distal to proximal order.

b: The italicised Roman numerals in this column refer to the positions of the tandem repeats within the H22A sequence as shown in figure 4.9.

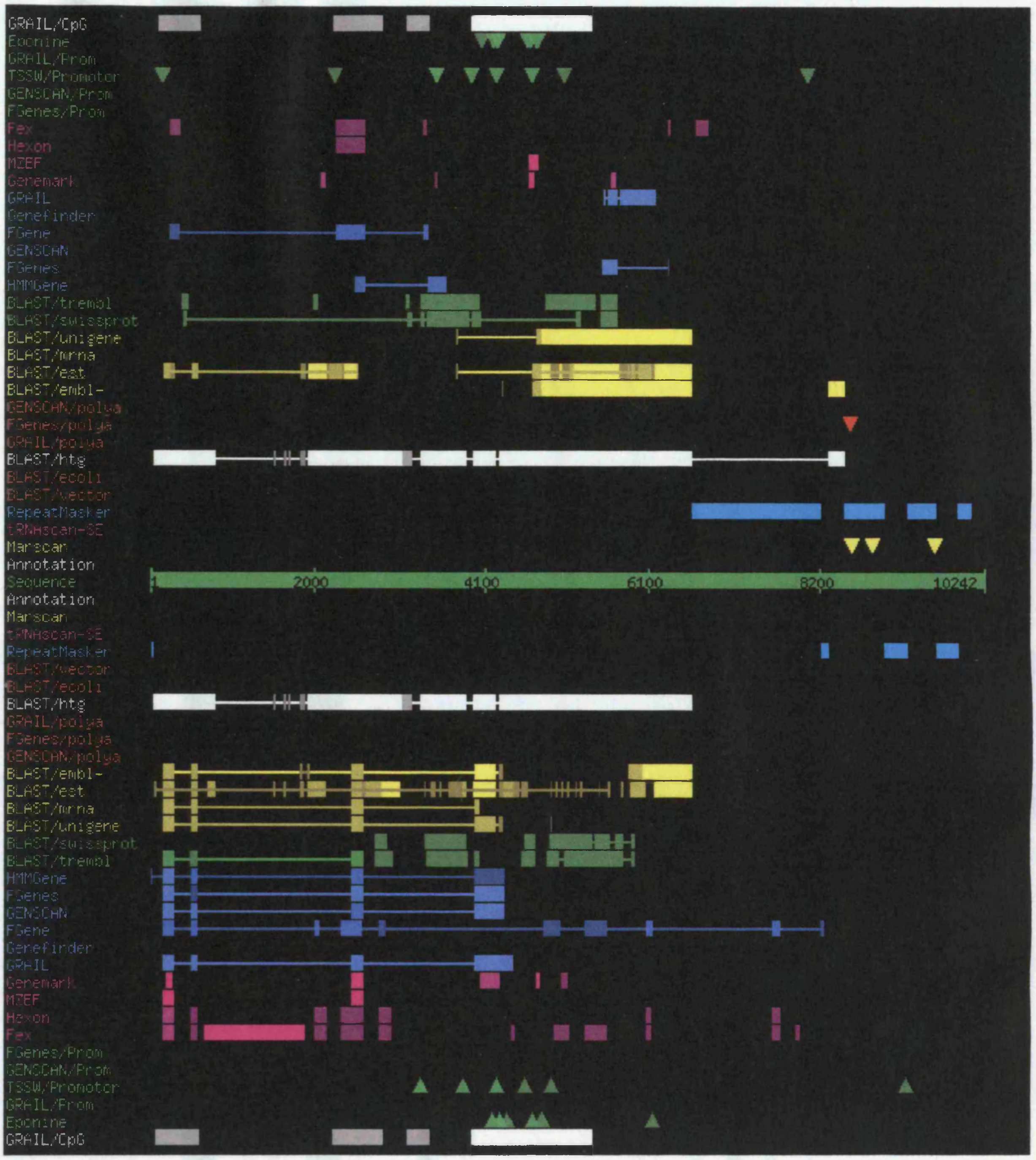


Figure 4.7: The NIX analysis of the H22A sequence, which lies proximal to N0434.1. The very good agreement between the gene and exon prediction programs (the blue and purple colours) indicated the positions of the *PGPL* nucleotides that were missing from N0434.1.

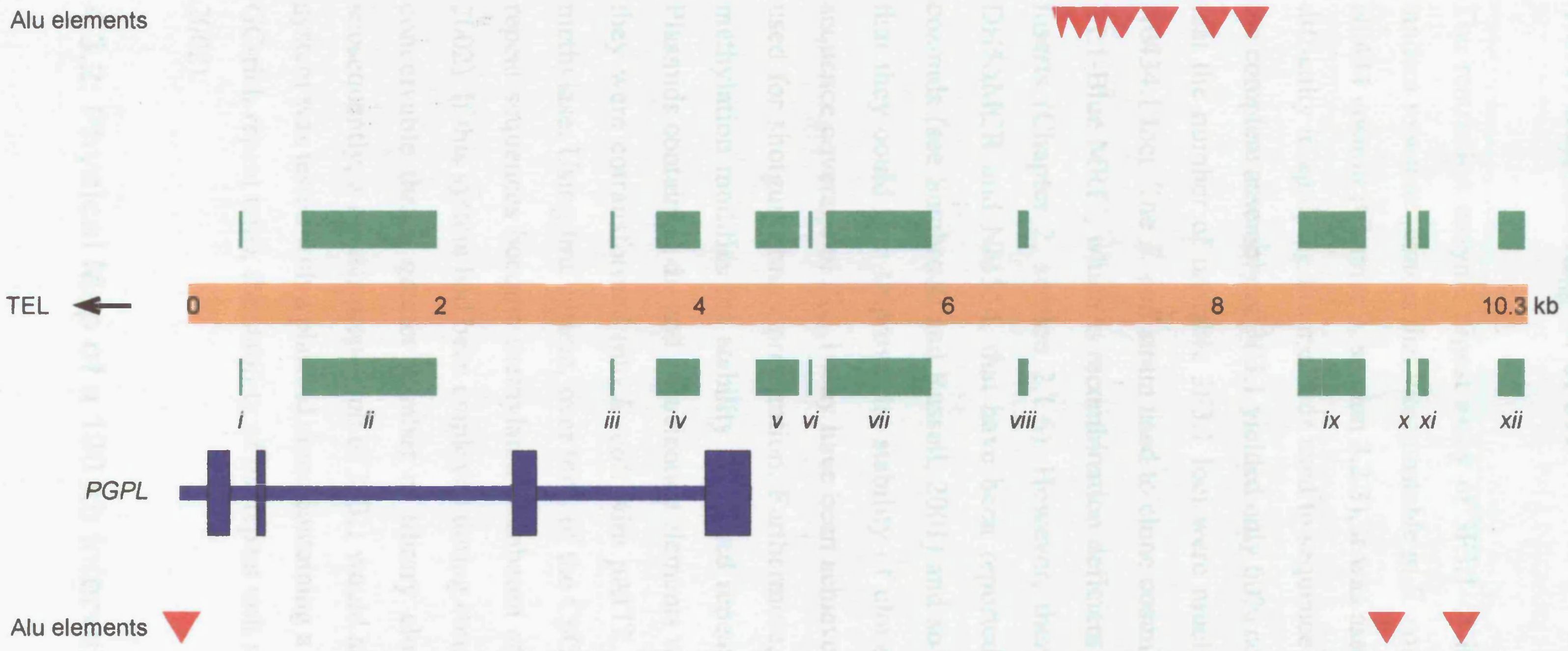


Figure 4.8: The genomic structure of H22A. Everything above the orange line is a feature found in the forward sense and everything below is a feature in the reverse sense. Alu elements are denoted by the red triangles and novel tandem repeats by the green blocks. The positions of the *PGPL* exons are denoted by the vertical blue bars. Roman numerals refer to descriptions of the tandem repeats within table 4.4.

4.3: Discussion

4.3.1: Sequencing of 3F3.1

The restriction enzyme digest assay of 3F3.1–.4 indicated that the 3F3 cosmid contained tandem repeat sequences that were unstable in *E. coli.* As this had also been suggested for the N0434 cosmid (Chapter 3, section 3.2.3), it was assumed that there would be relatively little difficulty in applying the methods used to sequence N0434.1 to sequencing 3F3.1. However, the complete assembly of 3F3.1 yielded only 60% coverage of the whole sequence, indicating that the number of unstable 3F3.1 loci were much greater, or more problematic, than the N0434.1 loci. The *E. coli* strain used to clone cosmid fragments for the shotgun libraries was XL1-Blue MRF′, which is recombination deficient and so ought to increase the stability of inserts (Chapter 2, section 2.1.6). However, there are other strains of *E. coli*, such as DH5αMCR and NM554, that have been reported to improve the stability of individual cosmids (see Sambrook and Russell, 2001) and so it would not be unreasonable to assume that they could also improve the stability of cloned plasmid inserts. Consequently, better sequence coverage of 3F3.1 may have been achieved if a different strain of *E. coli* had been used for shotgun library preparation. Furthermore, it has recently been reported that CpG methylation modifies the stability of cloned repeat sequences (Nichol and Pearson 2002). Plasmids containing di- and trinucleotide elements were seen to have increased stability when they were cotransformed into *E. coli* with pAIT2, a plasmid that expresses the *Sss*I CpG methylase. Using this system, over 90% of the CpG sites within the plasmids containing the repeat sequences become methylated (Renbaum *et al.*, 1990, cited in Nichol and Pearson, 2002). If this system had been employed during construction of the 3F3.1 shotgun library it is conceivable that a greater number of library clones would have been amplifiable and, subsequently, a greater proportion of 3F3.1 would have been sequenced. However, when this system was tested with a plasmid clone containing a minisatellite (228 copies of a 17 bp, 71% GC-rich repeat unit), the stability of the repeat unit was actually reduced (Nichol and Pearson 2002).

4.3.2: Physical Map of a 100 kb Interval at the Distal End of PAR1

The fact that, until recently, PAR1 had been largely overlooked by the Human Genome Sequencing Consortium was an undeniable source of frustration. The very recent production of PAR1 sequence data by the Sanger Centre was therefore most welcome, though the timing could have been a lot more convenient! As presented in figure 4.9, the current publicly available sequence in the PAR1 telomere–*PGPL* interval totals nearly 100 kb and has, for the first time, allowed the accurate physical mapping of a number of previously reported structures. The distance between the PAR1 telomere and the distal end of the known sequence (which is 2 kb distal to DXYS14) remains unclear, particularly as it is thought that DXYS14 and/or its distal flanking sequence is duplicated in some individuals at a point within the PAR1 telomere–DXYS14 interval (Inglehearn and Cooke, 1989; M. Hills, personal communication). Therefore, no attempt has been made to calculate the distances between the PAR1 telomere and any of the previously characterised or novel structures. However, this ought to be resolved relatively soon as the Sanger Centre is currently working on fosmid clone G248P87320F9, which reportedly extends towards the PAR1 telomere (M. Ross, personal communication).

4.3.2.a: Comparison of Regions Within the 100 kb Interval

The 100 kb of sequence at the distal end of PAR1 was derived from the 29C1, C3, N0434 and H22 clones. As the sequences within these clones overlapped it was possible to define four different subregions for the *in silico* analyses presented above and in Chapter 3 (figure 4.9). A number of the properties of these subregions are compared in table 4.5.

Table 4.5: A comparison of the 29C1A, C3A, N0434.1 and H22A sequences.

	29C1A	*C3A*	*N0434.1*	*H22A*	**Overall**
Length (kb)	29.5	23.7	33.4	10.3	**96.9**
GC content	53%	59%	53%	55%	**55%**
Alu content	15%	21%	33%	21%	**23%**
Tandem Repeat Content	53%	39%	22%	33%	**38%**

4.3.2.a.i: GC Content

The GC content of 29C1A, C3A, N0434.1 and H22A varies from 53% to 59%, the most GC-rich region being C3A. Over the whole interval, the average GC content is 55%, unambiguously placing the distal end of PAR1 in the H3 isochore family, the GC-richest 3–5% of human DNA (Bernardi, 1989).

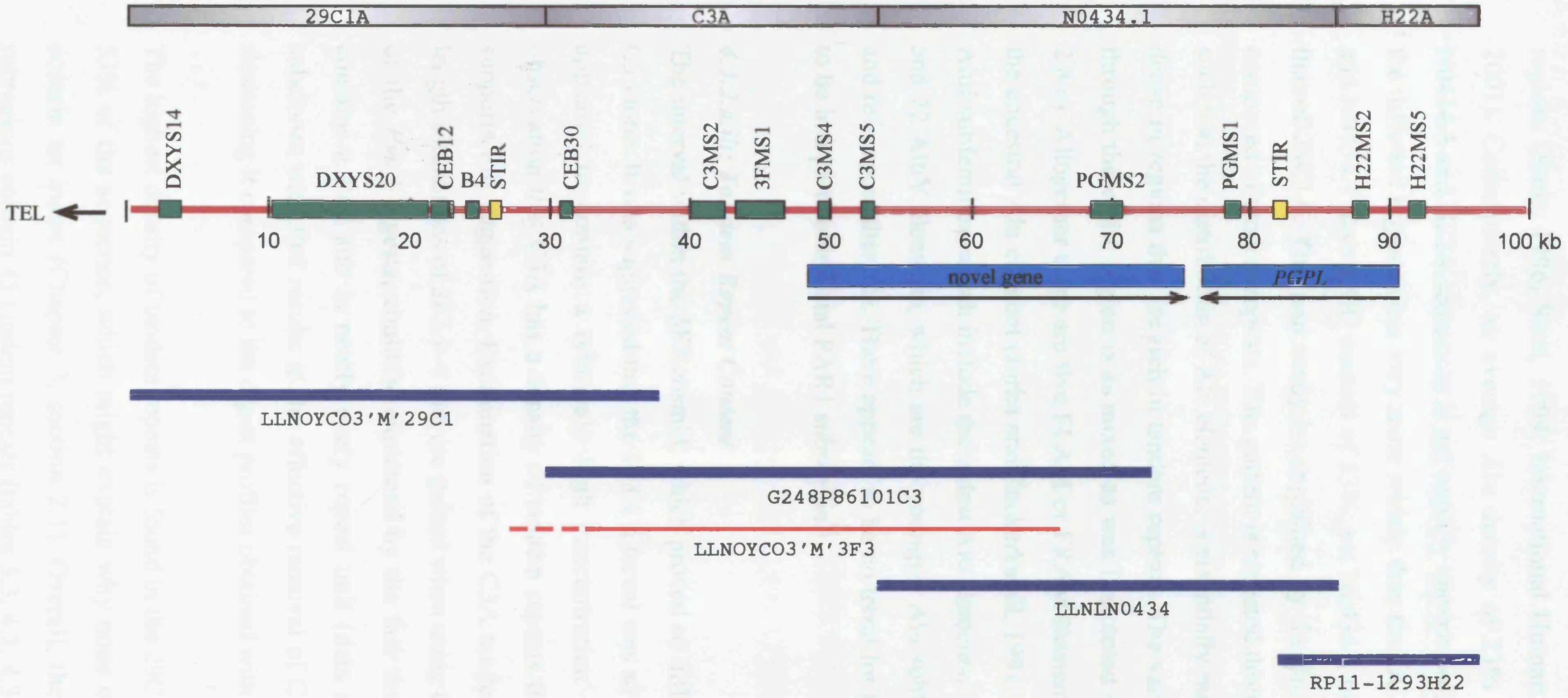


Figure 4.9: The physical map of a 100 kb interval at the distal end of PAR1. The interval extends from the *DXYS14* locus, which has been suggested to be only 3.6 kb from the PAR1 telomere (Baird and Royle, 1997), to the proximal end of the *PGPL* gene. The regions that were subjected to *in silico* analysis, as detailed in this chapter and Chapter 3, are marked by the grey boxes at the top of the diagram. The green boxes denote the positions of the major minisatellite loci within the interval and the positions of the pseudoautosomal STIRs are shown by the yellow boxes. The positions of *PGPL* and the novel gene are shown by the cyan boxes and the directions in which they are transcribed are denoted by the black arrows . Finally, the blue lines at the foot of the diagram show the contiguous clones that were used to sequence this interval. The red line marks the extent of cosmid 3F3. The telomeric end of 3F3 was not included in the 3F3.1 assembly, so the uncertainty surrounding its relative position is represented by the dashed line.

4.3.2.a.ii: Alu Element Density

Alu elements represent approximately 10% of the genome but preferentially locate to GC-rich regions (Smit, 1996, Smit, 1999; International Human Genome Sequencing Consortium, 2001). Consequently, an average Alu density of 23% across the GC-rich 29C1A, C3A, N0434.1 and H22A sequences is not entirely unexpected. However, the Alu density between the different regions does vary more widely than the GC content. For example, both 29C1A and N0434.1 have a GC content of 53%, yet N0434.1 has an Alu density that is over twice that of 29C1A. This can easily be explained by the fact that a large fraction of 29C1A is composed of tandem repeats. This pattern is repeated throughout the DXYS14–*PGPL* interval such that the distribution of Alu elements is essentially random, but their concentration is less dense in regions that are rich in tandem repeats. The variety of Alu subfamilies represented through the entire region is as mixed as was first noted within N0434.1 (Chapter 3, section 2.8.c). Altogether there are five FLAM or FRAM elements, believed to be the progenitors of the ancestral Alu element (Jurka and Zuckerkandl, 1991; Quentin, 1992), 14 members of the AluJ subfamilies, which include the oldest Alu elements, 36 members of the AluS subfamilies and 22 AluY elements, which are the youngest Alu subfamily (Kapitonov and Jurka, 1996 and references therein). There appears to be no trend for the oldest or youngest Alu elements to be in a particular distal PAR1 subregion.

4.3.2.a.iii: Tandem Repeat Content

The interval within the 3F3 cosmid, which proved so difficult to sequence, is covered by the C3 clone. It was suggested that the 3F3.1 interval was so problematic to sequence because it appeared to contain a relatively high concentration of unstable tandem repeats. The observation that C3A has a density of tandem repeats that is nearly twice that of N0434.1 supports this suggestion. Examination of the C3A tandem repeats revealed that the shorter length estimations of 3F3.1–4 that are gained when using the *Pst*I data, rather than the profiles of the *Pvu*II digests, could be explained by the fact that the C3MS2 tandem repeat array contains a *Pst*I site in nearly every repeat unit (data not shown). Digestion of the 3F3 subclones with *Pst*I results in the effective removal of C3MS2 from the digest profile, thus shortening it compared to the digest profiles obtained with other enzymes.

The highest density of tandem repeats is found in the 29C1A subregion, where they comprise 53% of the sequence, which might explain why none of the 29C1 subclones appeared to contain an insert (Chapter 3, section 2.1). Overall, the 29C1A, C3A, N0434.1 and H22 subregions contain 45 tandem repeats (tables 3.3, 4.2, 4.3 and 4.4) which make up over one-third of the sequences. There are a total of ten novel microsatellite sequences, two arrays of pseudoautosomal STIRs and 23 novel minisatellites, most of which are GC-rich. Thus, the distal PAR1 region is much more minisatellite-rich and extends more proximally than previously thought.

4.3.2.b: Identification of Previously Characterised Minisatellites

Previously characterised tandem repeats were identified through comparing their sequence (figure 4.1) with the structures that were detected by Tandem Repeats Finder (figures 4.4, 4.6, 4.8 and 4.9). For the first time this has allowed the physical distances between the previously reported distal PAR1 structures to be defined. The distance between DXYS14 and DXYS20 is 6.7 kb, in agreement with the previous estimates of between 6 and 40 kb (Rouyer *et al.*, 1986*b*; Brown, 1988) (Chapter 3, section 1.1.b.ii). Furthermore, the sizes of DXYS14 and DXYS20 in the 29C1 cosmid are 1.4 kb and 11.3 kb respectively, well within the size ranges observed in individuals (Inglehearn and Cooke, 1990; Page *et al.*, 1987). Assuming correct assembly of the 29C1 sequence at the Sanger Centre, the analysis also revealed that the original orientation of the tandem repeats identified by Vergnaud *et al.* (1993) was incorrect. The actual layout is shown correctly in figure 4.9 and demonstrates that the largest CEB12 repeat is only 263 bp proximal to DXYS20 and just 40 bp distal to CEB12 repeat 2. The monomorphic B4 minisatellite is slightly less than 1 kb proximal to CEB12 repeat 2. CEB30 was found in both the C3 and 29C1 sequences and is 5.8 kb proximal to B4.

4.3.2.b.i: Search for DXYS78

There is currently no sequence data available for the DXYS78 locus. However, it has been partially characterised and it is expected to be between 5 and 30 kb in length (Armour *et al.*, 1990) and suggested to contain sites for *Mbo*I and *Taq*I in a minority of its repeats and a *Hae*III recognition site in most repeats (J.A.L. Armour, personal communication). In addition, it was localised to an interval between 70 and 80 kb from the telomere (Henke *et al.*, 1993) and was thus expected to be identified within the C3A sequence. The best candidates within this region are C3MS2 and 3FMS1 (table 4.2), which are relatively large and contain a *Hae*III site in their repeat unit consensus sequences and an infrequently occurring *Mbo*I site. However, none of the C3MS2 and 3FMS1 repeat units within C3A contain a *Taq*I site, though some of the repeat units of 3FMS1 appear to have sequence that is only 1 bp removed from a *Taq*I recognition site. As it remained possible that the DXYS78 locus was actually more proximal than had been predicted, the novel tandem repeats of the H22A sequence were also searched for the aforementioned restriction enzyme recognition sites. Of these, the best candidate was H22MS5, which contained *Mbo*I, *Taq*I and *Hae*III sites. However, the *Taq*I site was contained within the consensus sequence and would thus have cut too frequently for the patterns seen by Armour *et al.* (1990) to be observed. In addition, *Hae*III sites were observed in only a few H22MS5 repeat units.

DXYS78 has not been identified, though two very good candidates have been suggested. Final identification of DXYS78 will not be possible until the definitive DXYS78, which has been partially characterised, is sequenced and compared with the candidate sequences in the DXYS14–*PGPL* interval. In this light, it is interesting to note that a plasmid subclone of DXYS78 is available for analysis (J.A.L. Armour, personal communication).

4.3.2.c: Relative Physical Positions of the Genes in the Distal 100 kb of PAR1

Analysis of the novel sequence in the PAR1 telomere–*PGPL* interval has allowed the entire genomic structures of both *PGPL* and the novel gene to be established (figure 4.9, also see Chapter 5). The physical position of the distal end of *PGPL* can now be defined as 73.3 kb proximal to DXYS14, whilst the distal end of the novel gene resides only 44.7 kb proximal to DXYS14. No other identifiable gene sequences are in the interval covered by the 29C1A, C3A, N0434.1 and H22A subregions. Hence, given the fact that DXYS14 may be as little as 3.6 kb from the PAR1 telomere (Baird and Royle, 1997), it is very probable that the novel gene is the most telomeric in PAR1.

4.4: Concluding Remarks

Extending and characterising the known sequence of the PAR1 telomere–*PGPL* interval has resulted in nearly 100 kb of annotated PAR1 subtelomeric sequence. The patterns noted after sequencing of the N0434.1 interval are repeated throughout the whole region, in that it is very GC-rich and contains a high density of both Alu sequences and tandem repeat arrays.

It is expected that coverage of distal PAR1 will be completed very shortly by the sequencing of a fosmid that is reported to extend towards the PAR1 telomere. This ought to reveal the extent of the suggested DXYS14 duplication and will allow the physical distances between the PAR1 telomere and the structures of the PAR1 subtelomere to be defined. In addition, extending the sequence towards the telomere will, for the first time, allow an accurate comparison of the distal region of PAR1 with the subtelomeric regions of both PAR2 and the autosomes. This will be interesting, as previous studies have suggested that there are conserved structural features between human subtelomeric regions. For example, it has been demonstrated that the subtelomeres of the 4p, 16p and 22q chromosomes consist of a mosaic of distal sequences that match many other chromosomes, and a more proximal region of longer uninterrupted matches to a few chromosome ends (Flint *et al.*, 1997*a*). Flint and colleagues suggest that this structural organisation is common to all chromosome ends. However, all experiments on the characterised PAR1 subtelomeric repeat arrays have indicated that they are PAR1-specific (e.g. Cooke *et al.*, 1985; Page *et al.*, 1987; Inglehearn and Cooke, 1990, Baird and Royle, 1997). Furthermore initial analyses of the annotated DXYS14–*PGPL* sequence have not identified any significant matches to the subtelomeres of other chromosomes.

As the assembly of the 3F3.1 cosmid was not completed, the overlapping region between N0434.1 and 3F3.1 was not investigated for polymorphisms that might have aided an analysis of LD in the region prior to an investigation of recombination. However, the availability of the PAR1 sequences recently released by the Sanger Centre do provide an extension of the primary tool for the discovery of SNPs in PAR1, thus providing a means to extend the high resolution analysis of recombination in PAR1 towards the telomere and thereby allowing a greater chance of identifying the putative distal boundary of recombination mentioned in Chapter 1, section 1.3.2.c.ii.

Chapter 5
Analysis of Genes in the N0434.1 Interval

5.1: Introduction

Following the publication of the first draft of the human genome, two independent analyses that relied heavily on gene-finding algorithms revealed it to contain about 30,000–40,000 genes (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001), much less than earlier estimates of 45,000–140,000 that had been based on EST clustering and detailed chromosomal analysis (Fields *et al.*, 1994; Dunham *et al.*, 1999; Liang *et al.*, 2000), and surprisingly similar to the number of genes in less complex organisms such as *Arabidopsis* and *C. elegans* (*C. elegans* Sequencing Consortium, 1998; *Arabidopsis* Genome Initiative, 2000). However, when the putative genes from the two human genome projects were compared, it was found that the numbers of novel genes predicted by both groups were largely nonoverlapping (Hogenesch *et al.*, 2001). Furthermore, a slightly more recent and in-depth study used all of the public database information on gene expression and estimated that the human genome contains 65,000–75,000 "transcriptional units," a term that reflects the fact that the vast majority of putative genes are still to be validated (Wright *et al.*, 2001). Still, as 4–22% of predicted genes are thought to be pseudogenes (Harrison *et al.*, 2002) the human gene count remains largely unsettled and a final total is likely to require a more integrated approach that combines *in silico* predictions and experimental validation. Hence, the work presented in this chapter provides a very small step in this direction by analysing the sequences of the novel gene predicted by the Nix analysis described in Chapter 3, section 3.2.8. In addition, the genomic structure of *PGPL* is described for the first time.

5.1.1: Known Genes in PAR1

Prior to the discovery of a novel gene in the N0434.1 interval (Chapter 3, section 2.8.a), thirteen complete genes plus the 5′ region of a fourteenth had been identified in PAR1 (Chapter 1, section 3.2.a; figure 1.8a) (table 5.1). Their genomic size is quite variable, from the single exon *ALTE* (*TRAMP*) gene (Esposito *et al.*, 1999) to the *CD99* (*MIC2*) gene that stretches over 52 kb of genomic DNA (Goodfellow *et al.*, 1986; Rappold, 1993) but their mRNA sizes are fairly similar, ranging only from 1.3 kb to 4.5 kb.

Table 5.1: The PAR1 genes. The green text denotes those observations that have been made as part of the work presented in this thesis. The references from which the rest of the data were obtained are as follows: *PGPL* (Gianfrancesco *et al.*, 1998); *PPP2R3B* (Schiebel *et al.*, 2000); *SHOX* (Rao *et al.*, 1997); *XE7* (Ellison *et al.*, 1992); *CSF2RA* (Rappold *et al.*, 1992); *CRLF2* (Tonozuka *et al.*, 2001); *IL3RA* (Milatovitch *et al.*, 1993); *SLC25A6* (*ANT3*) (Schiebel *et al.*, 1993; Slim *et al.*, 1993); *ASMTL* (Ried *et al.*, 1998); *DHRSXY* (Gianfrancesco *et al.*, 2001); *ASMT* (Yi *et al.*, 1993); *ALTE* (*TRAMP*) (Esposito *et al.*, 1999); *CD99* (*MIC2*) (Goodfellow *et al.*, 1996); *XG* (Ellis et al 1994*a*). The table is adapted from Rappold (1993).

Gene	*Physical location (kb)*[a]	*Orientation*	*Genomic size (kb)*	*Number of exons*	*Size of mRNA (kb)*	*Expression*
Novel gene	90	tel-cen	26	9	2.087	ubiquitous
PGPL	100	cen-tel	14.3	10	1.867	Predominantly skeletal muscle and heart, but detectable expression in all adult tissues investigated
PPP2R3B	200	cen-tel	36	13	1.864	Predominantly skeletal muscle and heart, but detectable expression in all adult tissues investigated
SHOX	500		25	6	*SHOXa* = 1.89 *SHOXb* = 1.349	Not ubiquitous. *SHOXb* expression is more restricted than that of SHOX*a*
XE7			11	6	3.23	Ubiquitous
CSF2RA	1270		>45	>5	1.8	Haemopoietic cells
IL3RA		tel-cen		>4	1.5	Haemopoietic cells
CRLF2					1.579	Not ubiquitous, but seen in heart, skeletal muscle, kidney and liver
SLC25A6	1300	cen-tel	5.9	4	1.3	Ubiquitous
ASMTL	1300–1350	cen-tel	50	13	2.085	Ubiquitous
DHRSXY	1700				2.554	Ubiquitous
ASMT	1700–1800		>6	10	1.25	Pineal gland, retina
ALTE	2200		Single exon	1	4.537	Predominantly skeletal muscle and heart, but detectable expression in all adult tissues investigated
CD99	2470	tel-cen	52	10	1.24	Ubiquitous
XG	2600	tel-cen				Haemopoietic cells and skin fibroblasts

a: The genes are presented from the most distal to the most proximal. Their physical location is defined as the distance of each gene from the PAR1 telomere.

5.1.1.a: Function of Known Genes in PAR1

Partly as a consequence of the initial omission of PAR1 from the major human genome sequencing projects, many of the PAR1 genes have not been fully characterised, or even discovered yet, and so the functions of their products are not fully understood. However, in most cases, putative functions have been ascribed and there is strong evidence for involvement of at least one of the PAR1 genes in a common mutant phenotype.

5.1.1.a.i: XG

XG, which encodes the Xg^a blood group antigen, spans the boundary between PAR1 and the sex-specific part of the X chromosome, such that the three exons at the 5′ end of the gene are situated in PAR1 and the remaining exons are in the X-specific region (Ellis *et al.*,1994*a*,*b*).

5.1.1.a.ii: CD99

The *CD99* gene product is a ubiquitous 32 kDa transmembrane glycoprotein (Gelin *et al.*, 1989). It has been suggested to be involved in a number of cellular events, including cell adhesion during haemopoietic cell differentiation (Hahn *et al.*, 1997), apoptosis of neuronal cells (Sohn *et al.*, 1998) and T-cell activation (Wingett *et al.*, 1999). In addition, it has been proposed that loss of *CD99* is associated with the pathogenesis of Hodgkin's disease (Kim *et al.*, 1998). The *XG* and *CD99* protein products are 48% similar to each other, indicating that they are evolutionarily related. Furthermore, *XG* and *CD99* are arranged head to tail (Ellis *et al.*, 1994*a*).

5.1.1.a.iii: ALTE

ALTE (originally referred to as *Tramp*) is a single-exon gene that appears to encode a protein of 694 amino acids that is similar to transposases of the Ac family (Esposito *et al.*, 1999). The Ac family of transposable elements has similar transposons, terminal inverted repeats (TIRs) of 11 bp, a duplicated target site of 8 bp and is known to include at least four members in organisms as diverse as maize, *Drosophila*, and fish (studies cited in Esposito *et al.*, 1999). Interestingly, the MER1 group of interspersed repeats (of which there are about 100,000 in the human genome) has TIRs and a duplicated site that is typical of the Ac elements so far described. Esposito and colleagues postulate that, if *ALTE* is the only member of the Ac family in the human genome, it may be the element that is responsible for the spread of the entire MER1 group, as first alluded to by Smit and Riggs (1996). The function of *ALTE* is yet to be described.

5.1.1.a.iv: ASMT

ASMT is the acetylserotonin methyltranferase gene and catalyses the final reaction in the synthesis of the hormone melatonin, which is secreted from the pineal gland in the brain (Yi *et al.*, 1993). Due to its tissue-specific expression in the brain and retina (table 5.1), *ASMT* is considered to be a candidate gene for psychiatric disorders, which is interesting in the light of several studies that have reported linkage of pseudoautosomal markers in schizophrenia patients (e.g. Collinge *et al.*, 1991; d'Amato *et al.*, 1992).

5.1.1.a.v: DHRSXY

DHRSXY encodes a putative protein of 330 amino acids, which shares homology with the short-chain dehydrogenase/reductase (SDR) family. Most SDR proteins are NAD- or NADP-dependent oxidoreductases but the *DHRSXY* function is as yet unknown (Gianfrancesco *et al.*, 2001).

5.1.1.a.vi: ASMTL

The 3′ part of *ASMTL* (exons 7 to 13) shows significant homology to *ASMT*, and its 5′ end is similar to the entire length of the *maf* gene of *Bacillus subtilis* and the *orfE* gene of *E. coli.* (The *orfE* gene is suggested to have a role in the cell growth and division of *E. coli* and *maf* appears necessary for the filamentation of *B. subtilis* cells (Wachi *et al.*, 1991; Butler *et al.*, 1993)). The two distinct domains of the resultant Asmtl protein are separated by a stretch of approximately 80 amino acids that has no significant homology to any known sequence. Thus, *ASMTL* appears to be the fusion product of two different full-length genes with different evolutionary origins (Ried *et al.*, 1998). The conservation of the putative *ASMT* catalytic domains within *ASMTL* argues for a methyltransferase activity of this enzyme too, but the significance of the maf/orfE domain is not clear.

5.1.1.a.vii: SLC25A6

Originally named *ANT3*, *SLC25A6* is a member of the ADP/ATP translocase family that is involved in cellular energy metabolism and which is the most abundant protein of the inner mitochondrial membrane (Klingenberg, 1981, cited in Slim *et al.*, 1993). *SLC25A6* encodes a protein that catalyses the ATP/ADP exchange between the mitochondrion and the cytosol and thereby plays an essential role in the energy metabolism of a eukaryotic cell (Cozens *et al.*, 1989).

5.1.1.a.viii: IL3RA

The ability of the IL3 cytokine to promote haemopoietic cell proliferation is dependent on it binding to its receptor (Lau and Zhang, 2000), which consists of an α and a β subunit. The interleukin receptor subunit α (*IL3RA*) gene is pseudoautosomal and is a characteristic member of the cytokine receptor family (Milatovich *et al.*, 1993). Both IL3 and *IL3RA* have been shown to have some kind of role in T cell and other blood malignancies (Renauld *et al.*, 1995) and may also contribute towards prostate cancer (Lau and Zhang, 2000).

5.1.1.a.ix: CSF2RA

Cells belonging to the monocyte/macrophage lineage are stimulated to grow and differentiate through the action of the haemopoietic growth factor, granulocyte-macrophage colony-stimulating factor (GM-CSF) (Gough *et al.*, 1990). The activity of GM-CSF is transduced through specific cell surface receptors that consist of an α and a β subunit. Interestingly, both the IL3 (above) and GM-CSF receptors share the same β subunit and the gene for the α subunit of the GM-CSF receptor (*CSF2RA*) is also in PAR1 (Gough *et al.*, 1990; Rappold *et al.*, 1992). In addition, it has been suggested that *CSF2RA* also has a role in prostate cancer, as it is expressed at an enhanced level in prostate carcinoma and is upregulated by androgen treatments in the prostatic cell line LNCaP (Rivas *et al.*, 1998; Lau and Zhang, 2000).

5.1.1.a.x: CRLF2

CRLF2 is a type I cytokine receptor, which has been mapped to the same pseudoautosomal region as *IL3RA* and *CSF2RA*, to which it also shows significant homology (Tonozuka *et al.*, 2001). The biological function of *CRLF2* is under investigation but it may exist simply as a part of the receptor for the cytokine TSLP.

5.1.1.a.xi: XE7

XE7 was identified from an inactive X cDNA library (Ellison *et al.*, 1992). Evidence suggests that two hydrophilic protein isoforms result from alternative splicing of *XE7* transcripts, though its function remains unknown.

5.1.1.a.xii: SHOX

SHOX is a highly conserved homeobox-containing gene, which has at least two alternative splice forms, *SHOXa* and *SHOXb*. The alternatively spliced products differ at their C-terminal ends and it is thought that this modifies the phosphorylation and binding properties of the protein. In addition, they do not have identical patterns of expression; *SHOXa* is widely expressed, but *SHOXb* expression is more confined and mainly seen in bone marrow fibroblasts. The data of Rao *et al.* (1997) clearly suggest that mutations in *SHOX* are a cause of both idiopathic growth retardation and the short stature phenotype of Turner syndrome patients. To this end it is quite possible that, in common with other homeodomain-containing proteins, *SHOX* functions as a transcription factor that regulates the activity of multiple target genes and so controls essential aspects of growth and development (Rao *et al.*, 1997).

5.1.1.a.xiii: PPP2R3B:

Protein kinases and phosphatases regulate many cellular functions, including signal transduction and cell division, and are divided into subfamilies according to their specificity, sensitivity to inhibitors, expression pattern and sequence (Cohen, 1997). *PPP2R3B* (*PPP2R3L*) encodes a protein phosphatase regulatory subunit and is a member of the PPP2 phosphatase subfamily (Schiebel *et al.*, 2000), which consists of Ser/Thr phosphatases that can be inhibited by okadaic acid. Schiebel *et al.* (2000) argue for an involvement of *PPP2R3B* in carcinogenesis, but its exact function remains unclear.

5.1.1.a.xiv: PGPL

Like *IL3RA* and *CSF2RA* (above), *PGPL* has also been suggested to play a role in or be influenced by oncogenesis in the prostate gland (Lau and Zhang, 2000). It is also worth noting that *PGPL* is the only human PAR gene that is highly conserved in the mouse genome, though the mouse homologue is autosomal (Gianfrancesco *et al.*, 2001). The other mouse homologues of human PAR genes (*dhrsxy*, *csf2ra* and *il3ra*) are poorly conserved and are also autosomal (Gianfrancesco *et al.*, 2001 and references therein). The *PGPL* gene has already been introduced (Chapter 3, section 1.1.d), but this discussion can now be developed (below) as sequencing of 100 kb at the distal end of PAR1 (Chapters 3 and 4) has, for the first time, enabled the genomic structure of *PGPL* to be determined (table 5.1).

5.1.2: A Novel PAR1 Gene

Of all the PAR1 genes so far identified, only *XE7* appears to be a 'single' gene. The others either have close relatives in the genome or are duplicated within PAR1. In fact, as noted by Ried *et al.* (1998), gene duplication is a recurrent theme in PAR1 and is seen three times; between *XGA*, *CD99* and a PAR1 pseudogene *MIC2R* (Smith and Goodfellow, 1994), between *IL3RA*, *CRLF2* and *CSF2RA* and between *ASMT* and *ASMTL*. Initial analyses of the putative novel PAR1 gene (Chapter 3, section 3.3.3.b) indicated that it too has close relatives in the genome, as it matched an as yet unidentified cDNA reported to be slightly similar to a

1-phosphatidylinositol phosphodiesterase precursor (GenBank accession number = AK002185) (figure 5.1).

5.1.2.a: Phosphatidylinositol-Specific Phospholipase C

A search of the databases showed that 1-phosphatidylinositol phosphodiesterase precursor is synonymous with phosphatidylinositol-specific (phosphoinositide-specific) phospholipase C (PI-PLC) (e.g. Swiss-Prot P14262). As the latter term is that which is most common in the literature, I will use that during the remainder of this thesis. Phosphodiesterases are actually an important group of enzymes that includes deoxyribonucleases, oxyribonucleases, restriction endonucleases and the exonuclease activity of polymerases (Griffith and Ryan, 1999).

Phosphatidylinositols (PIs) were first recognised as cellular signalling molecules during the 1950s (Hokin and Hokin, cited in Anderson *et al.*, 1999). The classical pathway transforms PI to phosphatidylinositol 4,5-biphosphate (PIP_2) by the successive actions of PI 4-kinases and PI-4-P 5-kinases. Activation of PI-PLCs takes place in many cells in response to a large number of extracellular signalling molecules such as hormones, growth factors and neurotransmitters (Berridge and Irvine, 1984; Nishizuka, 1986; Rhee and Bae, 1997) and, as such, PI-PLCs play a key role in initiating receptor-mediated signal transduction. PI-PLCs hydrolyse PIP_2 to generate two important second messenger molecules, inositol 1,4,5-triphosphate (IP_3), which mediates the release of intracellular Ca^{2+}, and diacylglycerol (DAG), which activates protein kinase enzymes (Berridge and Irvine, 1989; Berridge, 1993).

PI-PLCs have been isolated from a wide array of organisms, including bacteria, yeast, plants, insects and mammals (Singer *et al.*, 1997 and references therein; see Griffith and Ryan, 1999 for a review of bacterial PI-PLCs). Mammalian PI-PLCs are distinct from bacterial PI-PLCs due to the former's absolute requirement for Ca^{2+} and the latter's inability to hydrolyse the phosphorylated forms of PI, such as PIP_2. A number of mammalian PI-PLCs have been purified from a variety of tissues (Hoffman and Majerus, 1982; Ryu *et al.*, 1986; Rhee *et al.*, 1989) and they are currently divided into three subfamilies, PI-PLC-β, -γ and –δ (Rhee *et al.*, 1989) (figure 5.2). The β subfamily has been shown to interact with GTP-binding proteins (Taylor *et al.*, 1991), which is potentially very interesting given the juxtaposition of the novel gene and of *PGPL*, which is thought to encode a GTP-binding protein (Gianfrancesco *et al.*, 1998).

```
1        GTTTTTTAGGAAGAGTGTCCCGCAGAGACCCGGCGGGAGCTGCCAGGAGCTCTGGGATTC
61       CAGCGGCTGGAAGCCACCTGGGAAGCCTGGCCTCAGTGTGGAAGAGAAGGCAGCAGGATT
121      ATTACAGAACCTTGTGAAGCCAACGCGGGCAGCCGCCAGGAGCTGCAGACCGAGAGGATC
181      TCGTCCTTTCTTGCGGCCCAGGGAGACCAGGCCTTTCATTCTGGGCTCGAGACCAACAAT
241      TCGAATTCCGAACTCCCCCTGCGTGTGGGACTCAAGGTTGCCCAGGGCTCACCTCTGATG
                                                                   M
301      GGTGGGCAGGTGAGCGCTTCCAACAGCTTCTCGAGGCTGCACTGCAGAAATGCCAACGAG
          G  G  Q  V  S  A  S  N  S  F  S  R  L  H  C  R  N  A  N  E
361      GACTGGATGTCGGCACTGTGTCCCCGGCTCTGGGATGTGCCCCTCCACCACCTCTCCATC
          D  W  M  S  A  L  C  P  R  L  W  D  V  P  L  H  H  L  S  I
421      CCAGGGAGCCACGACACGATGACGTACTGCCTGAACAAGAAGTCCCCCATTTCTCACGAG
          P  G  S  H  D  T  M  T  Y  C  L  N  K  K  S  P  I  S  H  E
481      GAGTCCCGGCTGCTGCAGCTGCTGAACAAGGCCTTGCCCTGCATCACGCGCCCTGTCGTG
          E  S  R  L  L  Q  L  L  N  K  A  L  P  C  I  T  R  P  V  V
541      CTGAAATGGTCCGTCACCCAGGCACTGGACGTCACAGAGCAGCTGGATGCCGGGGTGCGG
          L  K  W  S  V  T  Q  A  L  D  V  T  E  Q  L  D  A  G  V  R
601      TACCTGGACCTGCGGATAGCCCACATGCTGGAGGGCTCGGAGAAGAACCTGCACTTTGTC
          Y  L  D  L  R  I  A  H  M  L  E  G  S  E  K  N  L  H  F  V
661      CATATGGTGTACACAACGGCGCTGGTGGAGGACACACTCACGGAAATCTCGGAGTGGCTG
          H  M  V  Y  T  T  A  L  V  E  D  T  L  T  E  I  S  E  W  L
721      GAGCGGCATCCACGCGAGGTGGTCATCCTGGCCTGCAGAAACTTCGAGGGGCTGAGCGAG
          E  R  H  P  R  E  V  V  I  L  A  C  R  N  F  E  G  L  S  E
781      GACCTGCACGAGTACCTGGTCGCCTGTATCAAGAACATCTTCGGGGACATGCTGTGTCCT
          D  L  H  E  Y  L  V  A  C  I  K  N  I  F  G  D  M  L  C  P
841      CGTGGGGAGGTGCCGACACTGCGGCAGCTGTGGTCCCGGGGCCAACAGGTCATCGTCTCC
          R  G  E  V  P  T  L  R  Q  L  W  S  R  G  Q  Q  V  I  V  S
901      TATGAAGACGAGAGCTCCTTGCGCCGGCACCACGAGCTGTGGCCAGGAGTCCCCTACTGG
          Y  E  D  E  S  S  L  R  R  H  H  E  L  W  P  G  V  P  Y  W
961      TGGGGAAACAGGGTGAAGACCGAGGCCCTCATCCGATACCTGGAGACCATGAAGAGCTGC
          W  G  N  R  V  K  T  E  A  L  I  R  Y  L  E  T  M  K  S  C
1021     GGCCGCCCAGGAGGGTTGTTCGTGGCCGGCATCAACCTCACGGAGAACCTGCAGTACGTT
          G  R  P  G  G  L  F  V  A  G  I  N  L  T  E  N  L  Q  Y  V
1081     CTGGCGCACCCGTCCGAGTCCCTGGAGAAGATGACGCTGCCCAACCTTCCGCGGCTGAGC
          L  A  H  P  S  E  S  L  E  K  M  T  L  P  N  L  P  R  L  S
1141     GCGTGGGTCCGAGAGCAGTGCCCGGGGCCGGGTTCACGGTGCACCAACATCATCGCGGGG
          A  W  V  R  E  Q  C  P  G  P  G  S  R  C  T  N  I  I  A  G
1201     GACTTCATCGGCGCAGACGGCTTCGTCAGTGACGTCATCGCGCTCAATCAGAAGCTGCTG
          D  F  I  G  A  D  G  F  V  S  D  V  I  A  L  N  Q  K  L  L
1261     TGGTGCTGACGGGACCCTTCTGAAGTTCGGGACGCGGCGGCTGCAGTTTCACCCCCGAAT
          W  C  ●
1321     TTCCAAATGGAGTTTCGCTCTCGTTGCCGAGGCTGGAGTGTAGTAGTGTGATCCTGGCTC
1381     ACTGCAACCTCCACCTCCCGGGTTCCAGCAAATTCTCCTGCCTCAGCCTCCCAAGTAGCT
1441     GGGATTGCAGGCGCCCGCCACCACGCCCGGATAATTTTTGTGTGTTTAGCAGAGACGGGG
1501     TTTCACCATGTTGGCCAGGCTGGTCTCGATCTCCTGACCTCAGGTGATCCACCCGCCTCG
1561     GCCTCCCAAAGTGCTGGGATGACAGGCGTGAGCCACCGCGCCCGGCCTATACCTCATTTT
1621     CTACATGTCGCTTGTTGGAGCTGCTGGTTCAAGTTCCCAGCCAGCCAATGGATGCCAGCA
1681     CCATTTTTACTCCCCTTTCCCAAGCAAATCGTGCATTTTTGTCTAACGAGAGACATCAGT
1741     TTCTCAGGATGATCCTCAAGAACGTTATGGAGTCCATGTTGCAATAGGTTCTCTTTGGGA
1801     CCTAATGACTCATTTTCCAAAAATCCGCTTCTACTTTTGGTACCCGGTTGCTACGGTGAA
1861     ATGAAGGTGCCCCGCATCCAGAAAGACGCACTCCTGGACCACAACCGGCGGCTACCTCAG
1921     CCCCACGGCTCTGCAGGATCAGGGCTCGGGCAGGCCCCGCGGAGATGAAGAATTTGCAGG
1981     GAGCCTCCCTGACTTCCGTCGGCTGTGAATCCTTGTCTGTCAGGGGCGTATCCACAAAAT
2041     CACCAAATTCATACAGATCGTTTAAATAAATGAACATCATTAAAGTC
```

Figure 5.1: The cDNA sequence of “hypothetical protein FLJ11323,” the uncharacterised database sequence that is reported to be slightly similar to a 1-phosphatidylinositol phosphodiesterase precursor and which exactly matches the novel gene exons in the N0434.1 interval. The putative translation of the open reading frame is denoted by the blue letters and the PI-PLC-X domain is underlined (see main text). The region highlighted in grey denotes the sequence of the probe that was used for expression analysis and the pink and red sequences respectively show the primer and *Dra*I site that were used to generate the probe (see main text).

Identification of the three-dimensional structure of PI-PLC-δ1 demonstrated that the two most highly conserved domains, the X and Y regions, converge to form two halves of an irregular TIM barrel (named after the more regular structure of this type first observed in triosephosphate isomerase) and thus form the catalytic core of PI-PLCs (Essen *et al.*, 1996). Other regions include the pleckstrin homology (PH) domain, which is found in over 100 other proteins and is thought to be responsible for membrane binding (Ferguson *et al.*, 1995). However, the isolated PH domain of PI-PLC-δ1 has a high affinity for PIP_2, indicating that these domains may have other functions (Lemmon *et al.*, 1995; Singer *et al.*, 1997). The function of the EF-hand region is yet to be identified (Rebecchi and Pentyala, 2000) but it has been determined that deletions in this region inactivate PI-PLC-δ1 (Nakashima *et al.*, 1995). The C2 motif contains Ca^{2+}-binding domains and may stabilise the structure of the catalytic unit (Essen *et al.*, 1996; Singer *et al.*, 1997; Rebecchi and Pentyala, 2000).

Considering this information, and given the fact that the sequencing projects described in Chapters 3 and 4 had allowed the identification of the novel gene's genomic structure, a small number of analyses were performed on the novel gene. It was hoped that these analyses would determine how similar the novel gene was to the PI-PLCs and that some clues as to its function would be obtained.

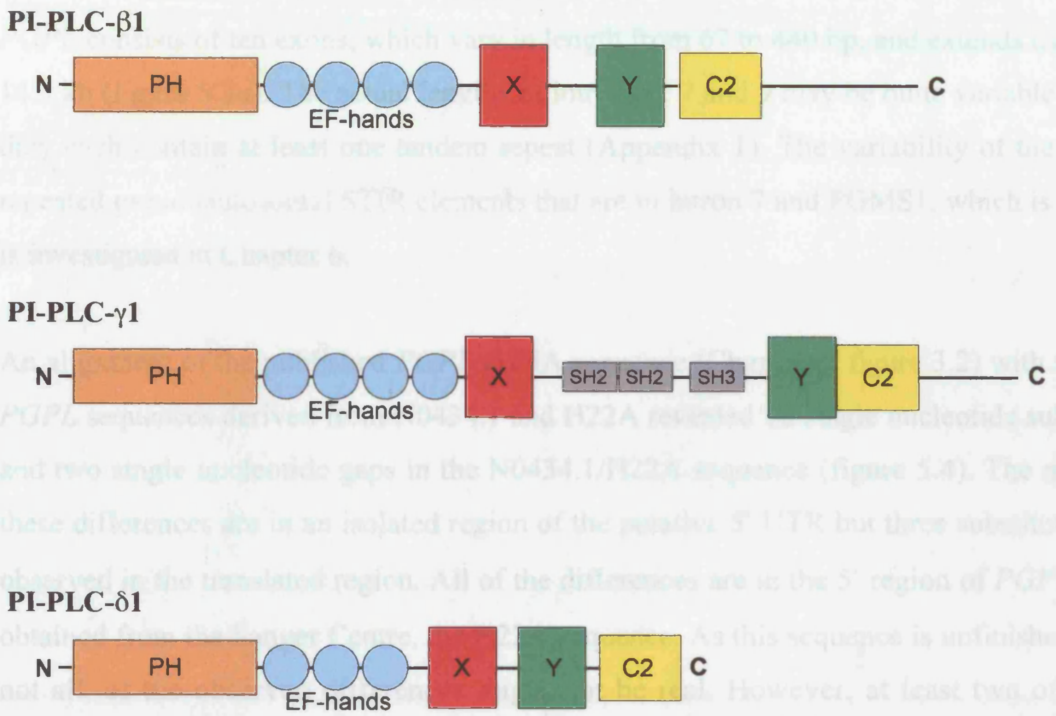


Figure 5.2: Linear representation of the three PI-PLC subfamilies and their different domains (adapted from Singer et al., 1997 and Rebecchi and Pentyala, 2000). The X and Y domains form the catalytic core of PI-PLCs and converge in the three-dimensional structure to form two halves of a TIM barrel motif (from triosephosphate isomerase). The PH domain is thought to have membrane-binding capabilities. PI-PLCs have up to four EF-hand motifs but their function is yet to be identified. C2 motifs have been identified in many proteins and are known to bind calcium ions. The PI-PLC-γ isozymes contain additional SH2 and SH3 domains that oversee interaction with phosphorylated growth factors and other signalling mediators.

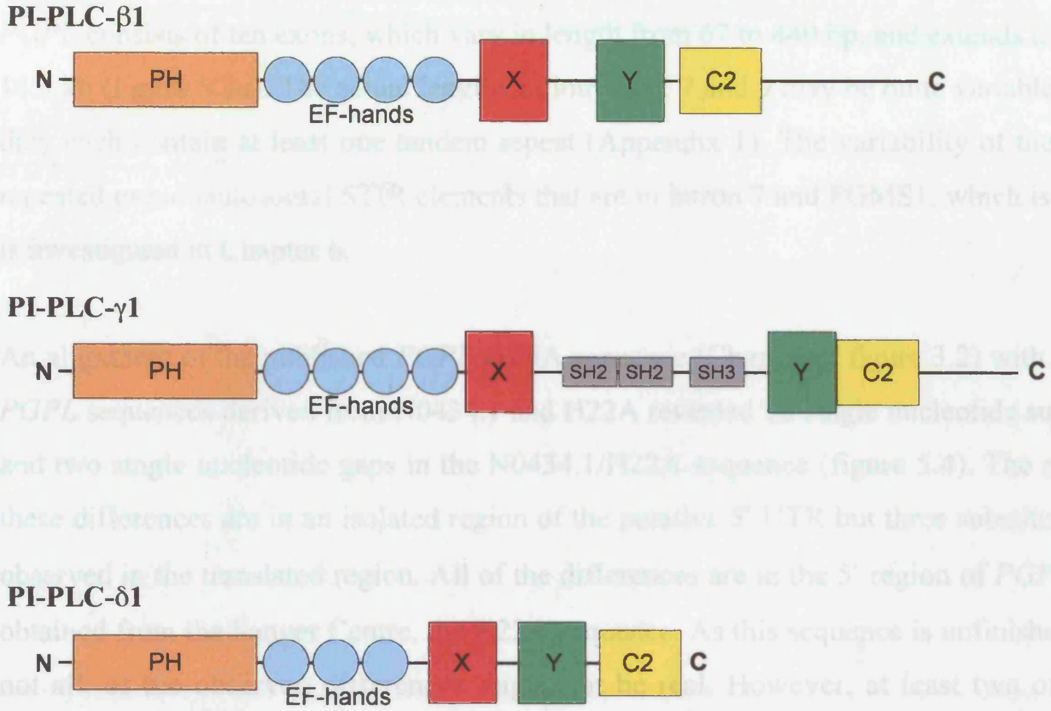


Figure 5.2: Linear representation of the three PI-PLC subfamilies and their different domains (adapted from Singer et al., 1997 and Rebecchi and Pentyala, 2000). The X and Y domains form the catalytic core of PI-PLCs and converge in the three-dimensional structure to form two halves of a TIM barrel motif (from triosephosphate isomerase). The PH domain is thought to have membrane-binding capabilities. PI-PLCs have up to four EF-hand motifs but their function is yet to be identified. C2 motifs have been identified in many proteins and are known to bind calcium ions. The PI-PLC-γ isozymes contain additional SH2 and SH3 domains that oversee interaction with phosphorylated growth factors and other signalling mediators.

5.2: Results

5.2.1: Genomic Organisation of the Genes in the N0434.1 Interval

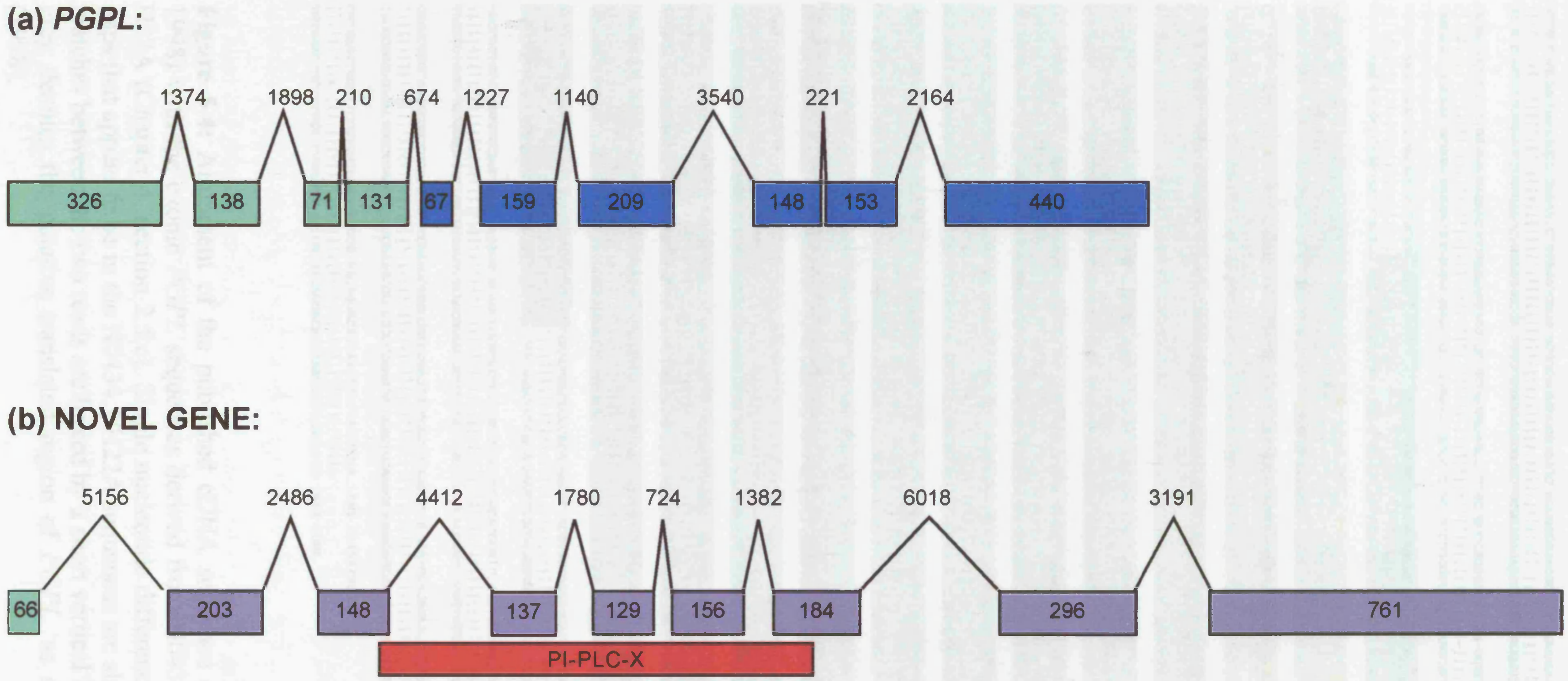
5.2.1.a: *PGPL*

PGPL consists of ten exons, which vary in length from 67 to 440 bp, and extends over at least 14.3 kb (figure 5.3*a*). The actual lengths of introns 6, 7 and 9 may be quite variable *in vivo* as they each contain at least one tandem repeat (Appendix 1). The variability of the tandemly repeated pseudoautosomal STIR elements that are in intron 7 and PGMS1, which is in exon 9, is investigated in Chapter 6.

An alignment of the published *PGPL* cDNA sequence (Chapter 3, figure 3.2) with the exonic *PGPL* sequences derived from N0434.1 and H22A revealed 12 single nucleotide substitutions and two single nucleotide gaps in the N0434.1/H22A sequence (figure 5.4). The majority of these differences are in an isolated region of the putative 5′ UTR but three substitutions were observed in the translated region. All of the differences are in the 5′ region of *PGPL* that was obtained from the Sanger Centre, the H22A sequence. As this sequence is unfinished, most, if not all, of the observed differences might not be real. However, at least two of the three putative substitutions in the translated region of *PGPL* are worthy of consideration; the most 5′ of these mutations is a G↔C transversion that would result in a glutamine to histidine replacement. The next mutation, at position 586 of the N0434.1/H22A sequence, is a C↔G transversion that would cause a conservative change from aspartic acid to glutamic acid. The putative T↔C transition at position 686 of the N0434.1/H22A sequence is a silent substitution at a leucine codon.

5.2.1.b: Novel Gene

Sequences derived from cosmid N0434 and produced by the Sanger Centre have shown the novel gene to be 26 kb in length and consist of nine exons, which vary in length from 66 to 761 bp (figure 5.3*b*). The genomic size of the novel gene is likely to vary between individual X and Y chromosomes, due to the presence of several tandem repeat arrays, as might be expected from a region that is rich in minisatellites (Chapter 3, section 3.2.8.b and Chapter 4, section 4.2.5.a.iii). Most notably intron 7, the largest novel gene intron, contains PGMS2, the variability of which is investigated in Chapter 6. The novel gene is predicted to encode a protein of weight 36.7 kDa.


(a) *PGPL*:

(b) NOVEL GENE:

Figure 5.3: Genomic organisation of (a) the *PGPL* and (b) novel genes. Both are presented in a 5′ to 3′ orientation. However, *in vivo*, *PGPL* reads from the centromere towards the short arm telomere whereas the novel gene, which is distal to *PGPL*, reads in the opposite direction and ceases within 1 kb of it. The exons of *PGPL* and the novel gene that are not contained within the N0434 cosmid are shown in green. The *PGPL* exons that are within N0434 are shown in cyan and those of the novel gene are shown in purple. The extent of the PI-PLC-X domain within the novel gene (see main text) is shown by the red box. Exon and intron lengths are given in bp.

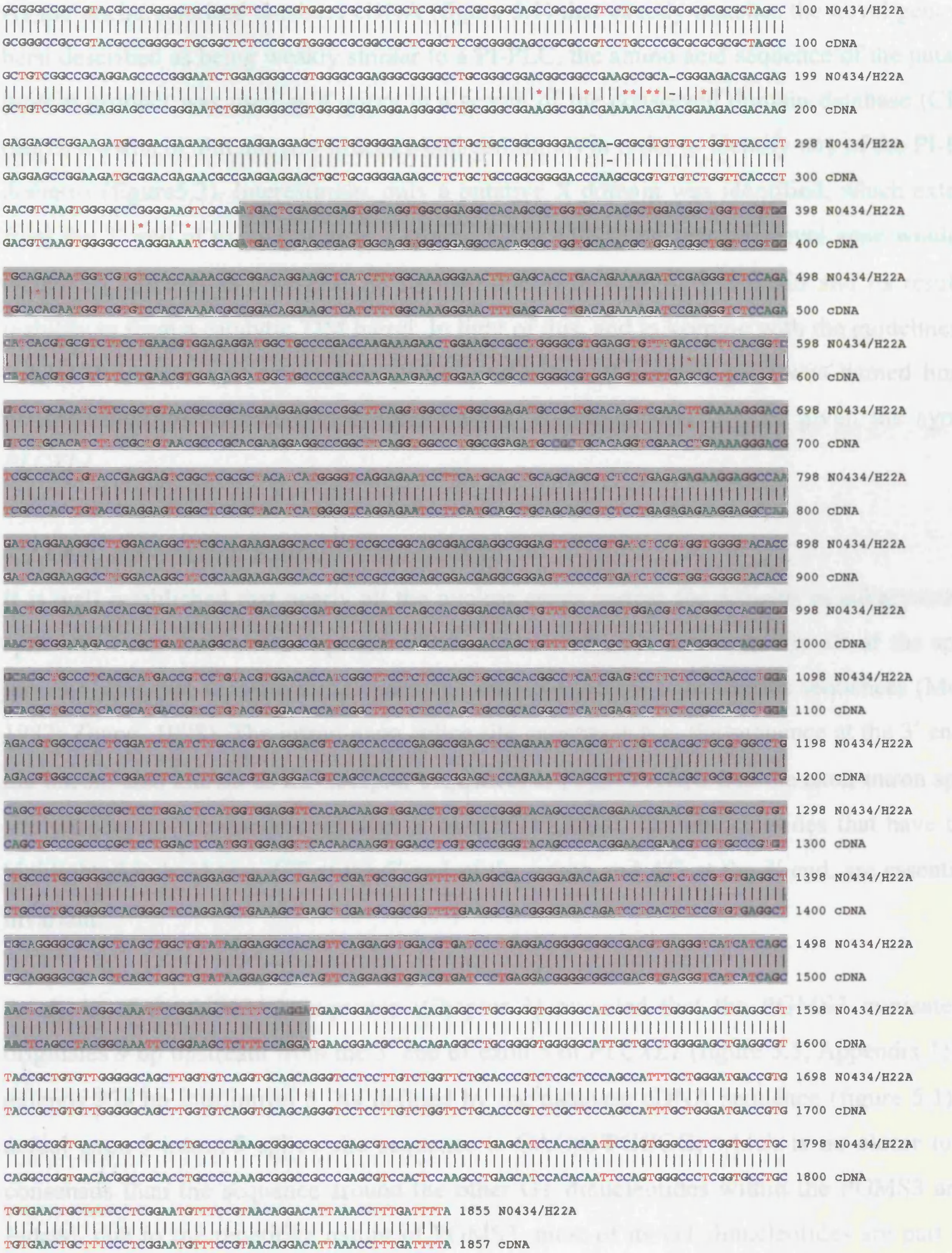


Figure 5.4: Alignment of the published cDNA sequence of *PGPL* (Gianfrancesco *et al.*, 1998) with the exonic *PGPL* sequences derived from N0434.1 (Chapter 3, section 2.8) and H22A (Chapter 4, section 2.5.c). Single nucleotide differences are marked by a red asterisk, gaps that appear to be in the N0434.1/H22A sequences are shown by short horizontal line and matches between the two reads are linked by a short vertical line. The sequence highlighted in grey denotes the putative translated region of *PGPL*, as defined by Gianfrancesco *et al.* (1998).

5.2.1.b.i: Identification of Conserved Domains

As the uncharacterised database cDNA (figure 5.1) that exactly matches the novel gene had been described as being weakly similar to a PI-PLC, the amino acid sequence of the putative protein product was used as a query in a search of the conserved domain database (CDD) (http://www.ncbi.nlm.nih.gov/structure/cdd/cdd.shtml) in order to identify any of the PI-PLC domains (figure5.2). Interestingly, only a putative X domain was identified, which extends from the 3′ end of exon 3 to exon 7 (figure 5.3*b*), indicating that the novel gene would be unable to function as a PI-PLC on its own due to its lack of Y domain and its resultant inability to form a catalytic TIM barrel. In light of this, and in keeping with the guidelines set out by the human genome nomenclature committee, the novel gene was named human phosphatidylinositol-specific phospholipase C X domain-like 1, and given the symbol *PLCXL1*.

5.2.1.b.ii: An Unusual Exon-Intron Boundary

It is well established that nearly all the nuclear genes coding for proteins in eukaryotes are split into exon and intron sequences. It is also well known that virtually all of the splice junction sites that separate these exons and introns conform to consensus sequences (Mount 1982; Zhang, 1998). The intron-exon splice site consensus (i.e. the sequence at the 3′ end of the intron, also known as the acceptor sequence) is $(Y)_n$NY**AG**|G and the exon-intron splice site consensus (the donor sequence) is MAG|**GT**RAGT. The dinucleotides that have been highlighted in bold text, GT at the 5′ end of the intron and AG at the 3′ end, are essentially invariant.

Analysis of the N0434.1 sequence (Chapter 3) revealed that the PGMS3 minisatellite originates 9 bp upstream from the 3′ end of exon 5 of *PLCXL1* (figure 5.5, Appendix 1) and extends 238 bp into intron 5. As defined by the database cDNA sequence (figure 5.1) the actual exon 5-intron 5 splice site sequence is GAG|**GT**GCGG, which is no closer to the consensus than the sequence around the other GT dinucleotides within the PGMS3 array. Indeed, due to the repetitive nature of PGMS3, most of its GT dinucleotides are part of a sequence that is exactly the same as the putative donor splice site, which raises the question as to how the transcriptional machinery manages to avoid generating frame-shifted messages with large insertions of one or more 62 bp repeats. It is possible that the inclusion of PGMS3 in the coding sequence of *PLCXL1* provides a mechanism for the generation of alternative gene products as there is an in frame stop codon (TGA) just 13 bp after the end of exon 5 (as defined by the database cDNA).

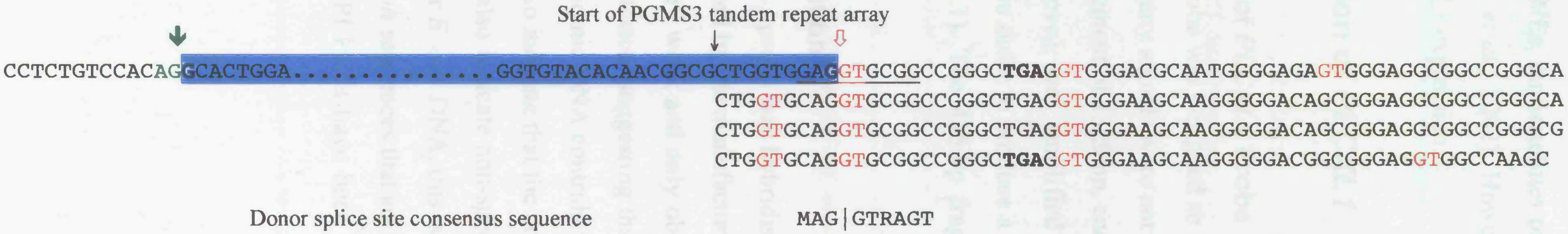

Start of PGMS3 tandem repeat array
CCTCTGTCCACAGGCACTGGA...............GGTGTACACAACGGCGCTGGTGGAGGTGCGGCCGGGCTGAGGTGGGACGCAATGGGGAGAGTGGGAGGCGGCCGGGCA
CTGGTGCAGGTGCGGCCGGGCTGAGGTGGGAAGCAAGGGGGACAGCGGGAGGCGGCCGGGCA
CTGGTGCAGGTGCGGCCGGGCTGAGGTGGGAAGCAAGGGGGACAGCGGGAGGCGGCCGGGCG
CTGGTGCAGGTGCGGCCGGGCTGAGGTGGGAAGCAAGGGGGACGGCGGGAGGTGGCCAAGC
Donor splice site consensus sequence
MAG | GTRAGT

Figure 5.5: Overlap of exon 5 of the novel gene and the PGMS3 repeat array. The four repeat units of PGMS3 have been aligned. The boundary between the 5′ end of exon 5 and the preceding intron is marked by the green, closed arrow and the conserved AG dinucleotide (in green). Most of the exonic sequence (highlighted in blue) has been excluded from this diagram but the proposed donor sequence at the exon-intron boundary has been underlined and the actual donor splice site, as defined by the database cDNA and its putative protein sequence (figure 5.1), is marked with a red, open arrow. Virtually all donor splice sites include a GT dinucleotide at the 5' end of the intron so all of the GT nucleotides within the repeat array have been marked in red. The in frame stop codons are denoted in bold.

5.2.1.b.iii: An Alu-Containing Exon

The ninth exon of *PLCXL1* comprises the 3′ UTR and includes an AluSq element that spans the final exon-intron border (Appendix 1). This is not in itself unusual because, in human mRNAs, SINEs, LINEs, minisatellites or microsatellites are found in approximately 36% of 3′ UTRs (Mignone *et al.*, 2002). However, it may be relevant to the post-transcriptional regulation of *PLCXL1* expression.

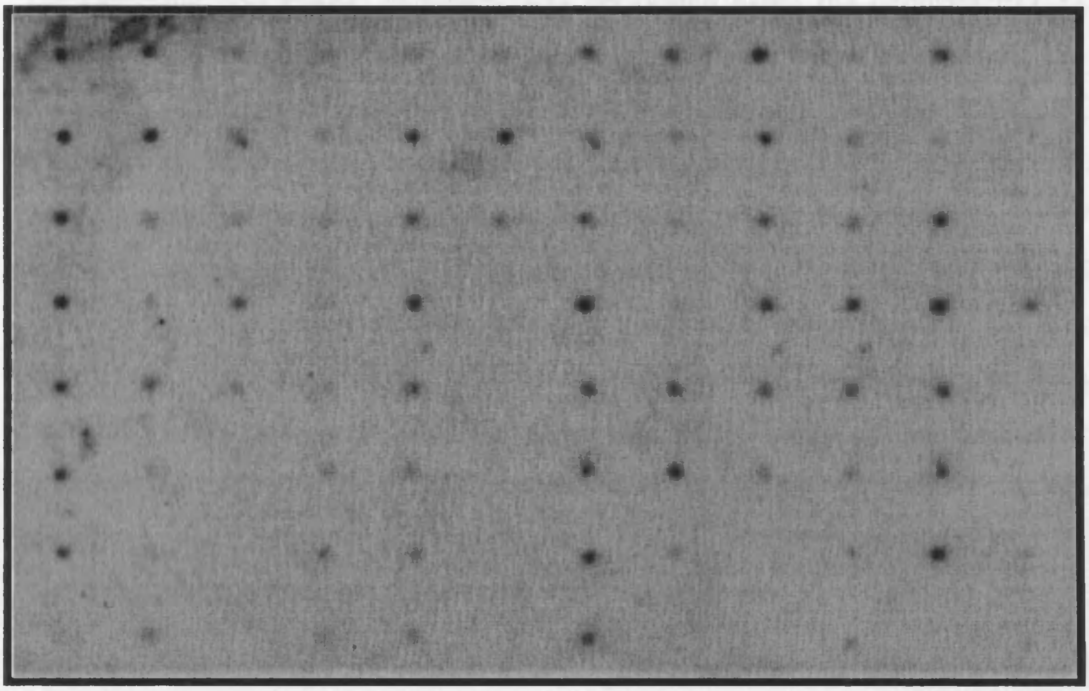
5.2.2: Expression of *PLCXL1*

5.2.2.a: Isolation of *PLCXL1* Probe

A *PLCXL1* gene probe was designed so as not to include any of the conserved X domain, in order to ensure that any signals were not due to cross-hybridisation with other PI-PLC genes. Consequently, a fragment of 1329 bp, encompassing the last 460 bp of exon 9 and none of its AluSq element (above), was amplified from 1 ng of N0434.1 DNA (table 2.3) and then digested with *Dra*I in order to produce a fragment of 435 bp that consisted entirely of exon 9 sequence (figure 5.1). The 435 bp fragment was purified by electroelution and ethanol precipitation.

5.2.2.b: Hybridisation

The isolated *PLCXL1* probe was hybridised to a Human Multiple Expression (MTE^{TM}) Array (Clontech), as directed by the manufacturer's instructions. The results are shown in figure 5.6. All signals were very weak and only obtained after 3–5 days exposure, but expression was found in all tissues tested, suggesting that this novel gene is a housekeeping gene. The faint signals from the genomic DNA controls further suggest that the gene belongs to a multi-gene family, but could also indicate that the probe is binding non-specifically. Hybridisation to the *E. coli* DNA could also indicate non-specific binding but, as there is no signal for yeast total RNA, yeast tRNA or *E. coli* DNA, this is probably unlikely. The signal could be explained by the presence of *E. coli* sequences that are homologous to *PLCXL1*, which is quite conceivable, given the fact that PI-PLCs have been identified in a wide variety of organisms (section 5.1.2.a, above).



	1	2	3	4	5	6	7	8	9	10	11	12
A	whole brain	cerebellum, left	substantia nigra	heart	oesophagus	colon, transverse	kidney	lung	liver	leukaemia, HL-60	foetal brain	yeast total RNA
B	cerebral cortex	cerebellum, right	nucleus accumbens	aorta	stomach	colon, descending	skeletal muscle	placenta	pancreas	HeLa S3	foetal heart	yeast tRNA
C	frontal lobe	corpus callosum	thalamus	atrium, left	duodenum	rectum	spleen	bladder	adrenal gland	leukaemia, K-562	foetal kidney	*E. coli* rRNA
D	parietal lobe	amygdala	pituitary gland	atrium, right	jejunum		thymus	uterus	thyroid gland	leukaemia, MOLT-4	foetal liver	*E. coli* DNA
E	occipital lobe	caudate nucleus	spinal cord	ventricle, left	ileum		peripheral blood leukocyte	prostate	salivary gland	Burkitt's lymphoma Raji	foetal spleen	poly r(A)
F	temporal lobe	hippo-campus		ventricle, right	ileocecum		lymph node	testis	mammary gland	Burkitt's lymphoma Daudi	foetal thymus	human C_0t-1 DNA
G	p.g.* of cerebral cortex	medulla oblongata		inter-ventricular septum	appendix		bone marrow	ovary		colorectal adeno-carcinoma SW480	foetal lung	human DNA 100 ng
H	pons	putamen		apex of the heart	colon, ascending		trachea			lung carcinoma A549		human DNA 500 ng

*paracentral gyrus

Figure 5.6: Hybridisation of a probe for the novel gene to a Human MTE^{TM} Array. The table shows the types of tissue on the array and their relative positions within it.

5.2.3: Conservation of *PLCXL1*

To investigate the possibility of *PLCXL1*-homologous sequences in other species, a search of the GenBank and EMBL databases was performed using the deduced *PLCXL1* protein sequence as a query. A number of matches were observed, with those showing the strongest homology found in the house mouse, *Mus musculus* (GenBank accession number BC039627), the brown rat, *Rattus norvegicus* (GenBank accession number XM_222258) and the zebrafish, *Danio rerio* (GenBank accession number BC054605), which respectively showed 72%, 69% and 39% identity to *PLCXL1* at the amino acid sequence level. A CDD BLAST search revealed that each of the putative orthologues contained a recognisable PI-PLC-X domain but none of the other conserved regions known to occur in PI-PLCs. Homology between *PLCXL1* and genes in *Xenopus laevis*, *Anopheles gambiae*, *D. melanogaster* and *C. elegans* was also observed. Intriguingly, the BLAST search also unveiled a second putative human *PLCXL1*-type gene, which is located at chromosome 3q13.13 (human hypothetical protein FLJ313579, GenBank accession number AK056141).

5.3: Discussion

5.3.1: *PGPL*

The characterisation of the genomic structure of *PGPL*, which is believed to encode a GTP-binding protein, into ten exons spread over at least 14.3 kb will allow a development of the analyses already performed on the *PGPL* cDNA sequence by Gianfrancesco *et al.* (1998). For example, Northern analysis detected three species of *PGPL* RNA in all human adult tissues tested. Now that the exon-intron structure of *PGPL* has been ascertained it should be possible to identify the structure of these RNA species and so determine the pattern of post-transcriptional modification. In addition, a more detailed analysis of the sequence flanking the 5′ end of *PGPL* would lead to the identification of its promoter sequences, which could then be used to determine if *PGPL* has the same transcription start site in different tissues. As this sequence (H22A) has not yet been finished, it would be prudent to wait before continuing along this path, especially as a number of differences between the cDNA sequence, as defined by Gianfranceso *et al.* (1998), and the Sanger Centre H22A sequence have already been seen. However, the conservation and the expression pattern of *PGPL* does suggest that *PGPL* does not have a specific alternatively spliced product or phenotypic effect in a particular tissue type (Gianfrancesco *et al.*, 1998).

5.3.2: *PLCXL1*

5.3.2.a: The Putative PI-PLC X Domain

There are at least five recognisable domains in mammalian PI-PLCs (figure 5.2) and, of these, only a putative X domain was identifiable in the amino acid sequence of the probable *PLCXL1* gene product. Whilst it is not unusual for some PI-PLCs to lack a PH domain (Rebecchi and Pentyala, 2000), without a Y domain it is impossible for *PLCXL1* to form the catalytic TIM barrel that is vital to the normal function of PI-PLCs. It is possible that *PLCXL1* represents a novel type of PI-PLC gene that, in a manner similar to *IL3RA*, *CSF2RA* and *CRLF2* (above), only encodes one subunit of a functional protein, but it is also clear that *PLCXL1* does not conform to the structure of any previously defined PI-PLC subfamily. Therefore, the putative X domain may have a functional significance of its own, which is perhaps reflected by the fact that similarities between mammalian and bacterial PI-PLCs are limited to just the N-terminal half and the X region (Griffith and Ryan, 1999).

5.3.2.b: Conservation and Expression of *PLCXL1*

The detection of genes with significant homology to *PLCXL1* in a variety of species suggests that *PLCXL1* has an important biological role. Moreover, the identification of a putative human autosomal gene that, like *PLCXL1* contains a PI-PLC X region as its only recognisable functional domain, indicates that *PLCXL1* is a member of a so far uncharacterised gene family, a supposition that is endorsed by the observation of putative *PLCXL1* homologues in the rat and mouse in addition to those listed in section 5.2.3 (data not shown). This also continues the pattern, seen throughout PAR1, of pseudoautosomal genes either being duplicated within the region or having close relatives elsewhere in the genome (this chapter, section 5.1.2).

PGPL was recently identified as being the only highly conserved PAR1 gene in the mouse genome (Gianfrancesco *et al.*, 2001) but the 72% amino acid identity (rising to 78% over the PI-PLC X domain) between human *PLCXL1* and the putative mouse *plcxl1* is slightly higher than the 68% amino acid identity that was reported between human *PGPL* and mouse *pgpl.* It thus appears as though there are at least two highly conserved human PAR1 genes in the mouse genome. As all of the previously identified mouse homologues of PAR1 genes have been shown to be autosomal (Gianfrancesco *et al.*, 2001), it will be interesting to determine the chromosomal location of the putative mouse *PLCXL1*, as this might have some bearing on interpreting the apparently separate and puzzling (Rappold, 1993; Graves *et al.*, 1998; Gianfrancesco *et al.*, 2001) evolutionary origins of the mammalian pseudoautosomal regions.

It has been shown that the β subfamily of PI-PLCs interacts with GTP-binding proteins (Taylor *et al.*, 1991) and that *PGPL* encodes a GTP-binding protein (Gianfrancesco *et al.*, 1998). *PLCXL1* is evidently not a member of the PI-PLC β subfamily but, given their similar expression patterns (table 5.1), their physical proximity and the identification of conserved mouse homologues, an investigation of the potential interaction between *PGPL* and *PLCXL1* is warranted. Genes in close proximity appear to be quite common in the human genome (Adachi and Lieber, 2002 and references therein) and are arranged in one of three fashions, listed here in order of their relative frequency (Adachi and Lieber, 2002); bidirectionally divergent (head-to-head), bidirectionally convergent (tail-to-tail) and tandemly (in the same direction). Only the divergent configuration will allow the potential utilisation of common promoter elements, so it is unlikely that *PGPL* and *PLCXL1*, which are arranged in a tail-to-tail fashion (Chapter 3, section 3.3.3.b; Chapter 4, figure 4.9), are transcriptionally co-regulated, though they could share enhancer elements.

5.3.2.c: Overlap of PGMS3 and *PLCXL1* Exon 5

The overlap of the PGMS3 minisatellite and *PLCXL1* exon 5 means that PGMS3 must be partially transcribed. Moreover, as the donor splice site of exon 5 forms part of the most upstream repeat unit, all subsequent repeat units provide potential splice donor sites that would effectively extend the length of exon 5, and consequently the size of the *PLCXL1* gene product. Whether these alternative splice sites are used or not is unknown as there are clearly a number of factors other than splice site consensus sequences that influence splicing; models proposed by Guo and Mount (1995) and McCullough and Berget (1997) have suggested that, in order to be recognised, a splice site must have a partner site an appropriate distance away, so that exon or intron definition is facilitated by the spacing. In addition, many sequences that match the consensus are incapable of acting as splice sites and there are many examples of active splice sites that match the consensus very poorly (Mount, 2000).

A tandem repeat overlapping an exon-intron border is by no means a unique phenomenon. Yang *et al.* (2000) observed a variable minisatellite in which the first 31 bp repeat unit originated 12 bp upstream of the 3′ end of exon 13 and extended into intron 13 of the human *CBS* gene. In this case alternate splicing was apparently prevented by a G→A substitution (so altering CAG|GT to CAG|AT) at the potential donor splice site of the second repeat unit. However, there are also examples of minisatellites that do provide multiple splice donor sites; the human interferon-inducible gene 6-16 contains a partially expressed minisatellite, whose first repeat unit spans the splice donor site of exon 2 (Turri *et al.*, 1995). The two downstream repeat units provide functional splice donor sites that extend exon 2 by 12 or 24 nt and insert four or eight amino acids respectively into the predicted gene product. The examples are not limited to human genes and have been observed in two amphioxus species, *Branchiostoma lanceolatum* and *Branchiostoma floridae* (Cañestro *et al.*, 2002). The alcohol dehydrogenase loci (*Adh3*) of these species contain minisatellites, termed *mirages*, that span 75% of the exon-intron boundaries and so duplicate the splice sites. Intriguingly, despite the high density of *mirages*, splicing appears not to be disturbed as no abnormal mRNA species were detected.

The in frame stop codon just 13 bp downstream of *PLCXL1* exon 5, and so upstream of the first potential alternate donor splice site, suggests that use of any of the alternate sites must be prevented in order to preclude the generation of nonsense messages. On the other hand, use of the first potential alternate donor splice site would include the in frame stop codon and so allow the translation of a truncated *PLCXL1* product. However, this is thought unlikely as it would lead to the truncation of the only recognisable domain within *PLCXL1* and the probable loss of function. Obviously this is worthy of further study, which could begin with an investigation of the expression of *PLCXL1* in human tissues by northern blot analysis; use of a probe derived from the 5′ end of the *PLCXL1* cDNA sequence, but not including any of the PI-PLC-X domain, might reveal more than one RNA species that might in turn provide clues as to any pattern of alternative splicing.

5.3.2.d: *PLCXL1* 3′ UTR

UTRs are very important in the post-transcriptional regulation of gene expression. They have been shown to influence the transport of mRNAs out of the nucleus (van der Velden and Thomas, 1999), modulate the subcellular localisation of mRNAs (Jansen, 2001) and it is known that mutations that alter the 3′ UTR can lead to serious pathology (Conne *et al.*, 2000). Consequently, the presence of an Alu element in the 3′ UTR of *PLCXL1* may have a functional significance. For example, analysis of the 3′ UTR of human MnSOD, which contains an Alu-like element, led to the suggestion that naturally occurring antisense RNA binds to Alu elements within mRNAs, and represses expression (Stuart *et al.*, 2000). The antisense RNA that binds Alu elements is likely to be 7SL RNA, which is evolutionarily related to the ancestral monomers that comprise Alu repeats (Chapter 3, section 3.3.2.a) and is also a component of the signal recognition particle (SRP) (Hsu *et al.*, 1995 and references therein). SRP also contains a complex of 9 kDa and 14 kDa polypeptide subunits (SRP9/14), which has been shown to bind with high affinity to synthetic mRNAs that contain interspersed Alu elements in their UTRs (Hsu *et al.*, 1995), an activity that has been referred to as Alu RNA-binding protein (Alu RBP). As Hsu and colleagues note, the effects of human Alu RBP on Alu-containing mRNAs deserves further examination *in vivo*. Additionally, Alu and L1 sequences in the 5′ UTR of a zinc finger gene, *ZNF177*, appear to exert a positive transcriptional enhancer effect, but repress translation of the gene (Landry *et al.*, 2001).

5.4: Concluding Remarks

The identification and localisation of *PLCXL1* increases the number of known genes in PAR1 to 15, including the 5′ region of *XG* and displaces *PGPL* from its previous position as the most telomeric PAR1 gene (Gianfrancesco *et al.*, 1998). Furthermore, the previous reports of interactions between GTP-binding proteins and PI-PLCs and the apparent identification of a mouse homologue of *PLCXL1*, which seems to be as highly conserved as the mouse homologue of *PGPL*, together with their close physical proximity suggests that the products of *PLCXL1* and *PGPL* may also associate in some way.

Although its function is as yet unknown, the availability of the complete cDNA sequence, and the detailed description of its genomic structure provided in this chapter, will allow further investigation of the biological role of *PLCXL1*, such as the detection of mutations related to human diseases. For example, *PLCXL1* would provide another target for those who have reported linkage of PAR1 markers in schizophrenia patients (e.g. Collinge *et al.*, 1991; d'Amato *et al.*, 1992), even though this issue remains somewhat contentious (e.g. Asherson *et al.*, 1992; Kalsi *et al.*, 1995; Ponnudurai, 1996). Furthermore, Loupart *et al.* (1995), have demonstrated loss of heterozygosity (LOH) for markers in the distal 500kb of PAR1 in breast cancer patients. Therefore, an analysis of the genes in this region is warranted in order to determine their possible relevance in human cancers and schizophrenia.

It remains to be seen as to whether or not *PLCXL1* represents a novel PI-PLC subfamily but, at the very least, it appears as though *PLCXL1* might provide a novel *in vivo* system for both the analysis of donor splice site choice and the exploration of UTR-mediated post-transcriptional regulation of gene expression.

Chapter 6
DNA Diversity in the N0434.1 Interval

6.1: Introduction

The N0434.1 sequence, obtained through the work presented in Chapter 3, provided an interval around the *PGPL* gene that could be targeted for the discovery of SNPs, an essential step in the high resolution analysis of recombination (Chapter 1, section 2.3). In addition, the variability of three N0434.1 tandem repeat arrays (table 3.3) was analysed to provide an indication of their instability as it has been suggested that unstable minisatellites can serve as occasional markers of local meiotic recombination hotspots (Jeffreys *et al.*, 1998*a*). The tandem repeat arrays selected for analysis were PGMS1, PGMS2, which are both minisatellites, and the pseudoautosomal STIR array. Both the SNP data and the results of the tandem repeat array analyses are presented in this chapter.

6.1.1: Measures of Genetic Variation

There are two commonly used measures of nucleotide variability in a population. The first, nucleotide diversity, π, is the average number of nucleotide differences between two sequences drawn at random from a population (Nei and Li, 1979; Tajima, 1983) and depends on both the number and frequency of polymorphic sites. Second, Watterson's θ (θ_w) is the proportion of polymorphic sites in a DNA segment, corrected for population sample size (Watterson, 1975) and, unlike π, does not depend on the frequency of SNPs. In addition, both π and θ_w can be used to indirectly estimate the mean population mutation rate according to the following relationship:

$$H = \frac{4N_e\mu}{1 + 4N_e\mu}$$

in which H is the heterozygosity (and is the factor estimated by θ_w and π), N_e is the effective population size and μ is the neutral mutation rate per base pair per generation. This relationship can also be applied to minisatellite variability data and so is pertinent to the analyses of PGMS1, PGMS2 and the STIR array (below). When $4N_e\mu$ is significantly less than one the relationship approximates to $H(\theta) = 4N_e\mu$. This is generally true in human populations, in which base mutation occurs at a frequency of 2×10^{-8} per nucleotide per gamete (Drake *et al.*, 1998) and N_e is usually estimated to be about 10,000 (Takahata, 1993; Zietkiewicz *et al.*, 1998; Zhao *et al.*, 2000).

Comparing the estimates of θ_w and π (and other measures of $\theta = 4N_e\mu$) can be used to assess departure from the standard neutral model of molecular evolution. This model assumes a hypothetical, randomly mating population of constant size in which genetic variation is neutral and follows the infinite sites model in which each new mutation occurs at a site not previously mutated (Kimura and Crow, 1964; Kimura, 1968; Kimura, 1985). One such assessment is Tajima's *D*, which compares values of θ_w and π from a single, non-recombining region of the genome (Tajima, 1989). A non-zero value of *D* indicates a departure from the neutral model and can occur if neutral polymorphisms are eliminated, and thus allowed to reach only low frequencies, as a result of the negative selection of deleterious mutations at linked sites (background selection). Alternatively, positive selection will remove older, high-frequency alleles and newer, low frequency alleles will hitchhike with the target of selection, which, like background selection, will result in a relative excess or deficiency of polymorphisms of various frequencies (Bamshad and Wooding, 2003).

6.1.2: Human Minisatellites

Minisatellites are generally GC-rich sequences of tandemly repeated DNA that consist of 6–100 bp repeat units stretching over 0.5–30 kb (for review see Bois and Jeffreys, 1999; Jeffreys *et al.*, 1999) and it has been shown that they include some of the most variable loci in the human genome (Buard and Vergnaud, 1994; Dubrova *et al.*, 1997; Tamaki *et al.*, 1999). Their variability stems, in part, from differences in the number of repeat units, and these length polymorphisms led to the discovery of minisatellites as highly informative markers for DNA fingerprinting (Jeffreys *et al.*, 1985). Furthermore, minisatellites also provided the first highly polymorphic, multiallelic markers for linkage studies (Nakamura *et al.*, 1987).

6.1.2.a: Instability at Minisatellite Loci

The instability of the most unstable minisatellites is strongly restricted to the germline (May *et al.*, 1996; Jeffreys and Neumann, 1997) and typically involves gene conversion-like repeat transfers between alleles (Jeffreys *et al.*, 1991; Armour *et al.*, 1993; Neil and Jeffreys, 1993). Hence, it is most likely that this recombinational interaction operates during meiotic recombination itself (Jeffreys *et al.*, 1998*a*,*b*; 1999). For most loci instability occurs preferentially in males, the most extreme example being the CEB1 locus, which has an extraordinarily high mutation rate (to new length alleles) of 15% per sperm, but only 0.2% per oocyte (Vergnaud *et al.*, 1991). Traditionally germline mutation at human minisatellites has been studied by pedigree analysis (e.g. Jeffreys *et al.*, 1988). However, this is an inefficient method due largely to the very low mutation rates at minisatellites, which are usually of the order of 1% per gamete (Bois and Jeffreys, 1999). Two more recently-developed approaches, small-pool PCR (SP-PCR) (Jeffreys *et al.*, 1994) and size-enrichment and SP-PCR (SESP-PCR) (Jeffreys and Neumann, 1997; Jeffreys *et al.*, 1997), allow the detection of new mutant molecules directly in genomic DNA and are therefore much more accurate and reliable tests of minisatellite mutation rates.

It has also been shown that minisatellite instability is not impeded by sequence variation between repeat units; e.g. two of the most unstable minisatellite loci detected to date, CEB1 and B6.7, can have 20 or more variant repeat types within a single allele array (Buard and Vergnaud, 1994; Tamaki *et al.*, 1999). This is in direct contrast to the situation seen at trinucleotide expansions and other simple tandem repeat (STR) loci, where sequence homogeneity between repeat units encourages repeat instability (Warren, 1996). There is significant variation between minisatellite alleles in sperm mutation rate, an effect that appears to operate in *cis*; for CEB1 and B6.7 instability is strongly affected by array size and increases steadily up to 40–50 repeats, above which it appears to reach a plateau (Buard *et al.*, 1998; Tamaki *et al.*, 1999). Minisatellites MS32 and MS205 show allele-specific variations that do not correlate with array length but appear to be associated with flanking haplotype (Jeffreys *et al.*, 1994; May *et al.*, 1996), such as the O1C variant at MS32, which has already been discussed (Chapter 1, section 2.3.a).

6.1.2.b: Minisatellites and Genes

The overlap between PGMS3 and exon 5 of *PLCXL1* (Chapter 5, sections 2.1.b.ii and 3.2.c) provided another example of a minisatellite within a gene coding sequence, but there are also examples of minisatellites which generate alleles that can influence the expression of neighbouring genes. For example, the insulin (*INS*) minisatellite, located 600 bp upstream of the transcriptional start site of the *INS* gene (Bell *et al.*, 1982), is likely to be the type 1 diabetes mellitus *IDDM2* locus (Stead *et al.*, 2000). Considering the high density of tandem repeat arrays in the N0434.1 interval, this may also be relevant to the expression of *PGPL* and *PLCXL1*.

6.1.2.c: AT-Rich Minisatellites

A small number of AT-rich minisatellites have been described in humans, including the Y chromosome-specific MSY1 (Jobling *et al.*, 1998) and the autosomal loci ApoB (Buresi *et al.*, 1996) and FRA16B (Yu *et al.*, 1997). They are quite different from GC-rich minisatellites, with internally palindromic repeat unit sequences, which have the potential to form hairpin structures that may contribute to repeat instability. In addition, their structure is typically modular, with similar variant repeats clustered into blocks, and not dispersed as in GC-rich minisatellites, indicating that mutation processes are likely to be different between the two types of loci.

6.1.3: SNPs

SNPs have only very recently gained favour as the marker of choice amongst molecular geneticists. In the 1990s, STRs (di-, tri- or tetranucleotide repeats) were employed in linkage analyses to identify a gene responsible for a monogenic disorder. STRs are ideal for this type of study as they show high levels of allelic variation in the number of repeat units, which can be typed by PCR amplification, and are distributed widely and evenly across the human genome (Gray *et al.*, 2000). However, interest in monogenic disorders has waned and linkage analysis is a poor system for the detection of the small effects of multiple gene variants contributing to complex, common diseases (Lander and Schork, 1994; Risch and Merikangas, 1996). Moreover, STR loci are probably too sparse for the association-based studies that are required to observe these small effects (Lander and Schork, 1994) and, in addition, their high level of polymorphism is thought to reflect high mutation rates (Chakraborty *et al.*, 1997), which would further hinder a population-based approach.

SNPs are the most common variable site and, as such, are expected to be present in the genome at a density high enough for a wide range of genetic studies. Indeed, there are at least three reasons for the recent upsurge of interest in SNPs. First, their high abundance, which is approximately one SNP every 1000–2000 nucleotides when comparing just two human DNA sequences (Li and Sadler, 1991; Nachman *et al.*, 1998; Wang *et al.*, 1998; Altshuler *et al.*, 2000; The International SNP Map Working Group, 2001), means that SNPs can be used as markers for mapping polygenic disease loci (Chakravarti, 2001). Pertinent to this is the realisation that the usefulness of SNPs for these association studies is determined by the extent of LD between them, so an understanding of the frequency and underlying patterns of association among SNPs unrelated to disease is essential for interpreting patterns of LD between markers and candidate disease genes (Nachman, 2001). Second, and in further contrast to STRs, SNPs have a relatively low rate of recurrent mutation, which means that two individuals sharing a variant allele can usually (but not exclusively) be marked with a common evolutionary heritage (Stoneking, 2001). SNPs are thereby stable indicators of human history and will greatly help in the identification of haplotypes for tracing migrations, relationships among ethnic groups and changes in population size of ancestral human populations. Third, the distribution of variation may shed light on the evolutionary processes at the molecular level by revealing the relative importance of selection, mutation, migration, recombination and genetic drift (Nachman, 2001). An added advantage, of course, is that SNPs are binary and so are well suited to high-throughput genotyping.

6.1.3.a: SNP Frequency Across the Human Genome

The average value of π in humans is approximately 7.5 x 10^{-4} (Li and Sadler, 1991; The International SNP Map Working Group, 2001; Reich *et al.*, 2002) and individual autosomes have very similar levels of heterozygosity, though chromosomes 15 and 21 do have values of π that are more than 10% diverged from the genome average. Whether this is biologically meaningful or not will require further investigation (The International SNP Map Working Group, 2001). The sex chromosomes have relatively low levels of diversity, which can be largely attributed to their lower effective population sizes (N_e) in $\theta = 4N_e\mu$ (The International SNP Map Working Group, 2001). When averaged across windows of 200 kb, amounts of heterozygosity show up to tenfold variation (The International SNP Map Working Group, 2001). However, on a finer scale, it has been observed that the amount of sequence variation is very similar between linked sites and, although this correlation declines with distance, it remains significant at distances of 100 kb (Reich *et al.*, 2002). This creates regions of 10–100 kb within the human genome that have fundamentally different levels of nucleotide heterozygosity (Zhao *et al.*, 2000; Reich *et al.*, 2002). These are thought to be determined largely by recombination hotspots that create blocks of LD, in which the correlated fixation or extinction of SNPs will lead to periodic intervals of high and low nucleotide diversity (Reich *et al.*, 2002; Kauppi *et al.*, 2003).

6.1.3.a.i: SNPs and Alu Elements

SNPs that have arisen through transitional changes are more common than transversions, with CpG dinucleotides showing the highest mutation rate, presumably due to the frequent deamination of 5′ methyl-deoxycytidine (Vogel, 1972; Duncan and Miller, 1980). Therefore, Alu elements, which have a high incidence of CpG dinucleotides (Batzer *et al.*, 1996), are predisposed to point mutations. In addition, it has been suggested that Alu elements undergo a large amount of gene conversion (Roy *et al.*, 2000), which could further contribute to SNP diversity as has been seen throughout the genome (Ardlie *et al.*, 2001; Frisse *et al.*, 2001). Thus, the high concentration of Alu elements in the N0434.1 interval (Chapter 3, section 3.2) is likely to have a significant affect on the nucleotide diversity of the region.

6.1.3.a.ii: SNPs in PAR1

Currently, very little is known about SNP density in PAR1. Very high levels of nucleotide diversity have been observed at both the telomere-adjacent region and at *PPP2R3B* (Baird *et al.*, 1995; Schiebel *et al.*, 2000). Schiebel and colleagues proposed that the elevated diversity could be partially explained by a high GC content that favours mutation at CpG dinucleotides, or by base-mutagenic recombination. However, the normal diversity seen in the *SHOX* region, which is recombinationally active and 53% GC-rich, suggests that differing levels of diversity cannot be explained by GC content or variable recombination rates, and they may instead reflect different population histories (May *et al.*, 2002). Therefore, investigating the nucleotide variability in the GC-rich N0434.1 region is not only an important step towards analysing LD and meiotic recombination in the interval, but may also help to clarify the reasons behind the variability of SNP frequency across the human genome, and particularly in PAR1.

6.2: Results

6.2.1: Variability of Tandem Repeat Arrays in the N0434.1 Interval

The three largest tandem repeat arrays in the N0434.1 interval, PGMS1, PGMS2 and the pseudoautosomal STIR elements, were analysed in order to determine their length variability. Each of these is located in the proximal half of the N0434.1 interval (figures 3.7, 4.9 and Appendix 1). Each array was PCR amplified from 10 ng of genomic DNA from forty unrelated UK men of North European descent, using primers that were designed from the flanking DNA of the array (table 2.5*a*). The PCR products were subjected to agarose gel electrophoresis, Southern blotted and hybridised to probes consisting purely of the relevant repeat array (table 2.5*b*).

6.2.1.a: PGMS1 and PGMS2

Both PGMS1 and PGMS2 consist of variant repeat units that are about 40% GC-rich and 50 bp in length (table 3.3, figure 6.1) and are thus typical minisatellite sequences.

6.2.1.a.i: Sizes of PGMS1 and PGMS2 Alleles

The hybridisation patterns of PGMS1 and PGMS2 are revealed in figure 6.2. Virtually all PGMS1 alleles are one of two sizes, which appear to differ by just one repeat unit. Alleles of the same size as the N0434.1 PGMS1 array, which is 1.13 kb (and 21 repeat units), have a relative frequency of 0.6, whereas the smaller allele has a relative frequency of 0.39. There is just one example of an allele outside of this size range, visible in donor 1, where there is an allele of approximately 0.81 kb, which is equivalent to about 15 repeat units. PGMS2 is more variable than PGMS1, though there are just two common alleles, which are either about 2.21 kb or 1.67 kb, which equates to approximately 45 or 34 repeat units respectively. There are nine alleles that are 1.9 kb in length and the size range of PGMS2 alleles is 1.4–7.5 kb, which is a considerably larger range than that seen at PGMS1, and suggests that PGMS2 can consist of about 150 repeat units *in vivo*.

6.2.1.a.ii: Estimation of Mutation Rates at PGMS1 and PGMS2

Under the standard neutral model of evolution (above), an estimation of the rate of mutation at minisatellite loci can be gained through the formula $H_O = 1/(1 + 4N_e\mu)$, where H_O is the homozygosity, N_e is the effective population size (10,000 in humans, see above) and μ is the

Figure 6.1: The internal heterogeneity of PGMS1 and PGMS2 repeat sequences in the N0434.1 cosmid. The sequences are presented in a PAR1 telomere to centromere orientation. Each array is shown with 50 bp of flanking DNA (in lower case), within which are the primers (in pink) used to amplify the minisatellites from a panel of unrelated UK men of North European descent. It was not possible to sequence across the whole of PGMS2 (Chapter 3, section 2.5.c), thus the string of Ns shown within the PGMS2 sequence here represent approximately 900 kb, which is about 18 repeat units.

PGMS1

```
      cggcatccagctggagcttgctttcttattggtagggagacctgtacccc
TTGACTGGCA-GCACAGATTAGGCACCTGTTGTGCGCACAGTCAGAAATGTA--TT
TTGACTGTCAAGTGCAGATTAGGCACCTGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCAAGCGCAGATTAGGCACCTGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCA-GCACAGATTAGGCACCTGTTGTATGCACAGTCAGAAATGTACATT
TTGACTGTCA-GTGCAGATTAGGCACCTGTTGTATGCACAG----AAATGTACATT
TTGGCTGTCAAGCACAGATTAGGCACCTGTTGTATGGTCAG----AAATGTACATT
TTCACTGTCA-GCATA-ATTAGGCACCTGTTGTATCCACAGTCAGAAATGTACATT
--GAGTGTCA-GCACACATTAGGCACCTGTTGTATGCACAGTCAGAAATGTACATT
TTGTCAGCACAG-----ATTAGGCAACTGTTGTATGCA--GTCAGAAATGTA--TT
TTTACTGTCAAGCACAGATTAGGCACCTGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCA-GCACAGATTAGGCACCTGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCA-GTGCAGATTAGGCACCTGTTGTATGCACAG----AAATGTACATT
TTGGCTGTCAAGCACAGATTAGGCACCTGTTGTATGGTCAG----AAATGTACATT
TTCACTGTCA-GCATA-ATTAGGCACCTGTTGTATCCACAGTCAGAAATGTATATT
TTGAGTGTCA-GCACAGATTAGGCACCTGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCA-GCACAGATTAGGCACCTGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCA-GCACAGATTAGGCACCTGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCAAGCACAGATTAGGCAACTGTTGTATGCA--GTCAGAAATGTA--TT
TTTACTGTCAAGCACAGATTAGGCACCCGTTGTATGCA--GTCAGAAATGTACATT
TTGACTGTCA-GCACAGATTAGGCACCTGTTGTATGCACAGTCACAAATGTAGATT
TTGACTCTCAAGCGCAGATTAGGCACCTCTTGTATGCACAGTCACAAATGTACATT
Tatgcaaacccattcatctcgtctgtacatcctaaagctctcggggatctc
```

Consensus repeat, 53bp

```
TTGACTGTCAGCACAGATTAGGCACCTTGTATGCACAGTCAGAAATGTACATT
```

PGMS2

```
      agcctcataagagatggcacagacttagtgcccaagacaacagcattctt
--CCT-GTCTATC-ACATGGGGGATAAGAATG-TGGACATGTTT-GGGGCCATATT
ATTCT-GTCTCCC-ACATGGG--ATTAGGACG-TGGACATCTTT-GGGGCCA--TT
ATCCT-GTCTACC-TCATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
AATCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGCCA--TT
ATTCT-GTCTCCC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTCCC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTACGACA-TGGACATCTTT-GGGGACA---T
ATTCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGCCA--TT
ATTCT-GTCTCACCTCATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCTTGTCTATC-ACATGGGA-ATTAGGACGGTGGACATCTTTTGGGGCCA--TT
ATTCT-GTCTACC-TCATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTCCC-ACATGGGN-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-ATCTCCC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGCCA--TT
ATTCT-GTCTACC-TCATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTACC-TCATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGA-ATTAGGACG-TGGACATCTTT-GGGGCCA--TT
ATTCT-GNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

NNTCT-GTCTCCC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGATG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTCCC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTCCC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGCCA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGATG-TGGACATCTTT-GGGGACA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGACG-TGGACATCTTT-GGGGCCA--TT
ATTCT-GTCTATC-ACATGGGG-ATTAGGATG-TGGACATCTTT-GGGGACA--TT
ATTtctcccacacggtgttacgacgtgagcatctttggggttgtctactgccc
```

Consensus repeat, 49bp

```
ATTCTGTCTATCACATGGGGATTAGGACGTGGACATCTTTGGGGACATT
```

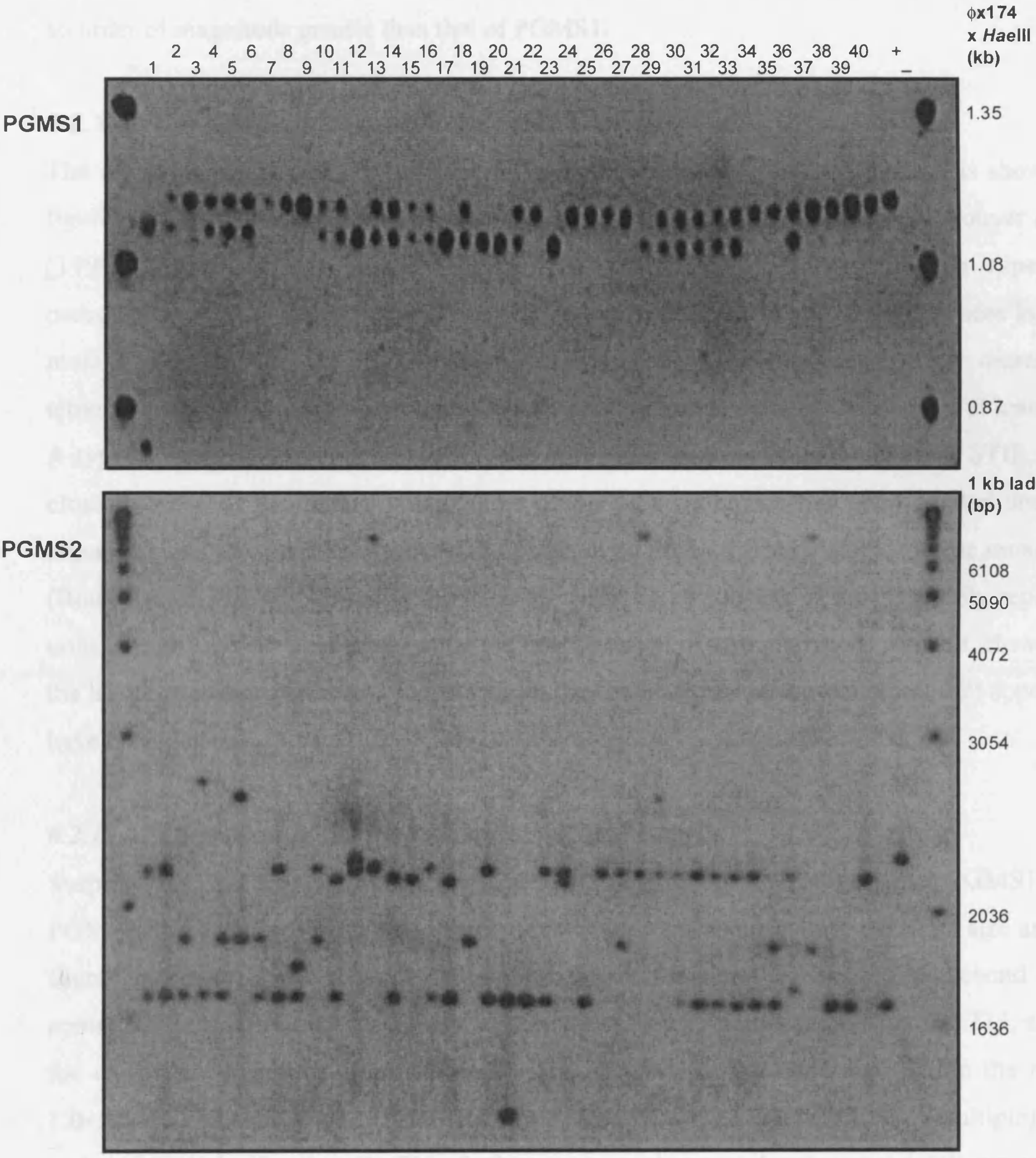


Figure 6.2: Variability of the PGMS1 and PGMS2 minisatellites in 40 unrelated UK semen donors. Positive controls (+) were amplified from 1.3 pg of isolated N0434.1 DNA, and a reaction that had no input of template DNA was used as a negative control (-).

mutation rate. The observed homozygosity at PGMS1 is 0.525 and thus, the approximate rate of mutation at PGMS1 is 2 x 10^{-5} events per gamete. In contrast, the homozygosity at PGMS2 is 0.2, which provides an estimate of the PGMS2 mutation rate at 1 x 10^{-4} events per gamete, an order of magnitude greater than that of PGMS1.

6.2.1.b: The N0434.1 Pseudoautosomal STIR Array

The sequence of the pseudoautosomal STIR array within the N0434.1 interval is shown in figure 6.3, together with the STIR element consensus sequence as defined by Rouyer *et al.* (1990). As previously stated (Chapter 3, section 1.1.c) there are two types of pseudoautosomal STIR, A-type and B-type, which can be distinguished by differences in their most 5′ 80 bp. Previously reported pseudoautosomal STIR arrays are either dimers or tetramers and all include at least one B-type monomer that is usually preceded by at least one A-type monomer (Rouyer *et al.*, 1990). The only known exception is found at the STIR array close to the *DXYS15* locus, where most of the first monomer has been deleted and the remaining elements within the array all appear to be derived from a single B-type monomer (Rouyer *et al.*, 1990). Similarly, the N0434.1 STIR array consists of four tandemly repeated units, which are very alike, suggesting the amplification of a single type B element. However, the leader sequence of the first repeat unit and the most 3′ part of the last repeat unit appear to have been deleted

6.2.1.b.i: Allele Sizes of the Pseudoautosomal STIR Array

Surprisingly, the array is moderately variable (figure 6.4). However, like PGMS1 and PGMS2, there are two main allele sizes. Most alleles are about 1.4 kb, the same size as that found in N0434.1, and which is seen at a relative frequency of 0.625. The second most common allele size is approximately 5.1 kb, which has a relative frequency of 0.275, and is the equivalent of about 15 repeat units. All other alleles are contained within the range 1.0–5.5 kb, and their sizes suggest that allele size differences result from the (multiple) gain or loss of whole repeat units. This further suggests that pseudoautosomal STIR arrays can consist of 16 tandemly repeated units, which would be four or eight times the size of the dimers and tetramers that have been reported so far (Rouyer *et al.*, 1990).

6.2.1.b.ii: Estimation of N0434.1 Pseudoautosomal STIR Mutation Rate

The observed homozygosity at the N0434.1 pseudoautosomal STIR array is 0.475 (figure 6.4). Consequently, the formula $H_O = 1/(1 + 4N_e\mu)$ provides an estimated mutation rate of 3 x 10^{-5} events per molecule, which is very similar to that of PGMS1.

Figure 6.3: Tandemly repeated STIR elements in the N0434.1 interval. The four different monomers (N0434.1/1–4) have been divided and read (5′ to 3′) centromere to PAR1 telomere to ease comparison with the consensus sequence, shown above in lower case, and their sequences are continuous from the first line to the last. The consensus sequence derived by Rouyer *et al.*, (1990) is divided over the first 80 bp to show the two different leader sequences. All type B leader sequences have been highlighted. The array is shown with 50 bp of flanking DNA (in grey). As previous experiments had shown that a generic pseudoautosomal STIR probe was also able to detect autosomal STIRs (Rouyer *et al.*, 1990), two primers were designed in order to generate a probe that would only detect the N0434.1 array. Their positions are underlined.

```
aacaacagnaayttatyytctcccagtcctgnngaccaggagtctgagatcaaggrgtytcaggrccryrctccctcc-------                     type A

                                                                                     rgaggctctagggga             consensus

-rggttctgggrrttagg-ayrtgracayrrctttyygrg---rrccactgttcaatcca-ttacaattgtatccagttccttct                      type B

                                        cacaggaaccagctctaccacagttccagaaagcgggagctgaccacagaCTCTAGGGGA             N0434.1/1

CAGGTTCCAGATGATCAATACATGGACAGGTCTTTT-GTGGGGGGCCACAGTTCAGTTCACTT-CAGTTGGATCCAGTTCCTTCTGGAGGCTCTAGGGGA             N0434.1/2

CAGGTTCCAGATGATCAATACATGGACAGGTCTTTT-GTGGGGGGCCACAGTTCAGTTCACTT-CAGTTGGATCCAGTTCCTTCTGGAGGCTCTAGGGGA             N0434.1/3

CAGGTTCCAGATGATCAATACATGGACAGGTCTTTT-GTGGGGGGCCACAGTTCAGTTCACTT-CAGTTGGATCCAGTTCCTTCTGGAGGCTCTAGGTGA             N0434.1/4


gggtccttcctgcctctcccagctcctgggggctccaggcrtccctgggcttgtggccgcatcactccagtctctgcctccgtctccacgtggccttctc             consensus

GGGTCCTTCCTGCCTCTCCCAGCTCCTGGGGGCTCCAGGCGTCCCTGGGCTTGAGGCCGCATCACTCCAGTCTCTGCCTCTGTCT----GTGGCCTTCTC             N0434.1/1

GGGTCCTTCCTGCCTCTCCCAGCTCCTGGGGGCTCCAGGCGTCCCTGGGCTTGTGGCCGCATCACTCCAGTCTCTGCCTCTGTCT----GTGGCCTCCTC             N0434.1/2

GGGTCCTTCCTGCCTCTCCCAGCTCCTGGGGGCTCCAGGCGTCCCTGGGCTTGTGGCCGCATCACTCCAGTCTCTGCCTCTGTCT----GTGGCCTTCTT             N0434.1/3

GGGTCCTTCCTGCCTCTCCCAGCTCCTGGGGGCTCCAGGCGTCCCTGGGCTTGTGGCCGCACCACTCCAGTCTCTGCCTCCATCTCCACGTGGCCTTC--             N0434.1/4


ctctgtgtctgtgtctcctcttctgtctcttanaaggacacctgtcattggatttagrgyccaccct a-at----ccargaygatytcatctcaagatcc             consensus

CTCTGTGTCTGTGTCTCCTCTTCTGTCTCTTAGAAGGACAGTAGTCATTAGATTTAGGGTCCACCCTA-AT----CCAGGATGATCTCATTTC-AGATCT             N0434.1/1

CTCTGTGTCTGTGTCTCCTCTTCTGTCTCTTAGAAGGACACCTGTCATTAGATTTAGGGTCCACCCTA-AT----CCAGGATGATCTCATCTCCAGATCT             N0434.1/2

CTCTGTGTCTGTGTCTCCTCTTCTGTCTCTTAGAAGGACAGTAGTCATTAGATTTAGGGTCCACCCTA-AT----CCAGGATGATCTCATTTC-AGATCT             N0434.1/3

-TCTGTGTCTGTGTCTCCT---------------AGGACACCTGTCATTGCATTTAGGGGCCACC-TAGATaaatccagggtaacctcatcttaactaca             N0434.1/4


ttracttaa--ttayatytgcaaagaccctatttccaaayrr----grtcycattcn          consensus

TCCACTTAA--TCACATCTGCAGAGACCCTGTTTCCAAATAA----TGTCCCATTCA          N0434.1/1

TCCACTTAA--TCACATCTGCAGAGACCCTGTTTCCAAATAA----TGTCCCATTCA          N0434.1/2

TCCACTTAA--TCACATCTGCAGAGACCCTGTTTCCAAATAA----TGTCCCATTCA          N0434.1/3

Tctgcaaaaactccatttcca
```

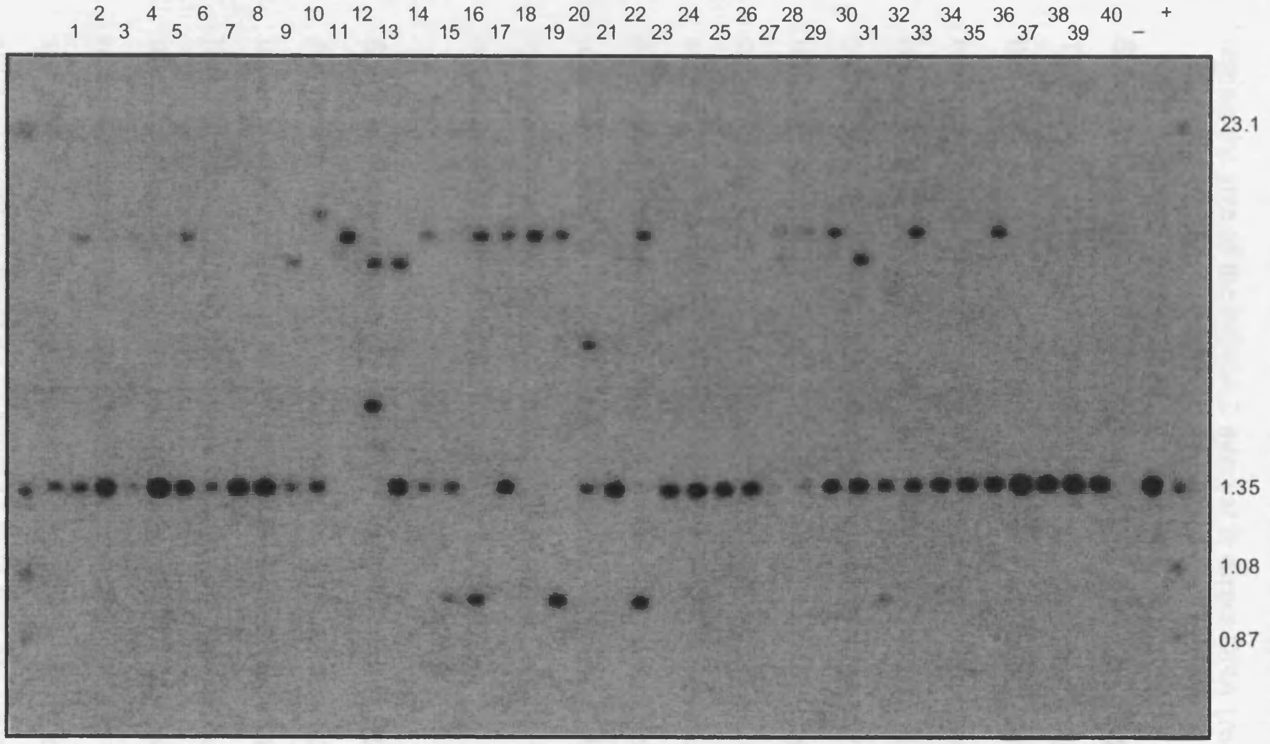


Figure 6.4: The variability of the pseudoautosomal STIR array within the N0434.1 interval in a panel of 40 unrelated UK men of North European descent.

6.2.2: Identification and Analysis of SNPs in the N0434.1 Interval

Four methods are commonly used for SNP detection; the identification of single strand conformation polymorphisms (SSCPs) (Orita *et al.*, 1989), heteroduplex analysis (Lichten and Fox, 1983), variant detector arrays (VDAs) (Wang *et al.*, 1998) and direct DNA sequencing. Of these, the simplest and most efficient method for the identification of novel SNPs in a region the size of the N0434.1 interval is direct DNA (re-) sequencing.

6.2.2.a: Design and Amplification of N0434.1 Regions for Resequencing

The N0434.1 sequence (Appendix 1) was used as a template to design a series of primers for the PCR amplification of twelve regions of the N0434.1 interval. These regions were then resequenced to identify SNPs. As it had been established that at least a subset of the tandem repeat arrays within the N0434.1 interval are moderately variable (above), care was taken to ensure that the twelve amplicons did not encompass any tandem repeat array lest these would impede later alignments of sequences and thereby hinder SNP identification. This consideration had to be balanced against including as much sequence of the N0434.1 sequence as possible, which inevitably led to the overlap of a number of amplicons (figure 6.5). Each resequencing amplicon was PCR amplified from the genomic DNA of eight unrelated UK men of North European descent (tables 2.5 and 2.6) and purified by electroelution and ethanol precipitation whilst taking care to prevent cross-contamination between amplicons from different donor DNAs.

6.2.2.b: Resequencing and SNP Identification Strategy

All twelve resequencing amplicons were first sequenced with the primers that had been used in their amplification. The primary N0434.1 sequence then allowed design of further primers, listed in table 2.5, that were used to complete the resequencing of each of the twelve regions in each of the eight donors. However, as N0434.1 has a very high density of Alu elements, the resequencing strategy was impeded in two ways. First, it was very difficult to design locus-specific primers due to the sequence similarity of different Alu elements and, second, Alu elements contain poly(A) or poly(T) tails that are usually impossible to sequence through, although this problem can occasionally be surmounted if the sequencing reaction begins very close to the poly(A/T) tail. Therefore, a number of resequencing amplicons were subamplified to create smaller and less complex intervals that could be sequenced more easily (figure 6.6*a*). Regrettably, some regions remained impossible to resequence completely. In total the sequence reads across all twelve regions in the eight donors amounted to over 200,000 bp and covered approximately 25 kb (75%) of N0434.1. All sequence reads were assembled and

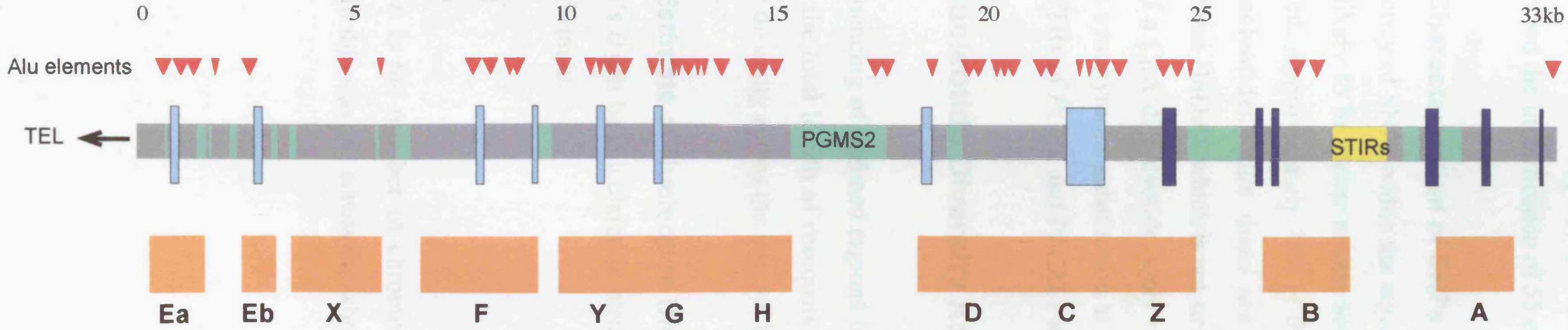


Figure 6.5: The amplicons used to resequence the N0434.1 interval in eight UK semen donors of North European origin. The grey bar represents the N0434.1 sequence, the light blue rectangles denote *PLCXL1* exons, the dark blue rectangles mark the positions of *PGPL* exons and the green bars show the positions of the tandem repeat arrays. Alu elements are marked wth the red triangles above the N0434.1 sequence and the positions of the resequencing amplicons are represented by the orange bars.

aligned with the N0434.1 primary sequence using ABI AutoAssembler software (figure 6.6*b*), which allowed the identification of 55 evenly distributed SNPs (table 6.1, figure 6.7).

6.2.2.c: Characterisation of SNPs

The frequency of SNPs within the region is approximately 1 SNP per 600 bp, though the true figure is likely to be closer to one SNP per 450 bp as only 75% of N0434.1 was actually resequenced. Approximately 75% of the SNPs (42 of 55) are transitions, 21 of which are at CpG dinucleotides, plus there are twelve transversions and one insertion/deletion polymorphism (InDel), which is not strictly a SNP as it is a polymorphism for the presence or absence of a CpA dinucleotide. Only one SNP was identified in coding sequence, 2793C/T, which is a synonymous substitution at a glycine codon; a further three SNPs were identified in the 3′ UTRs of *PGPL* and *PLCXL1* and 23 were found in Alu sequences (table 6.1).

6.2.3: Nucleotide Diversity Across the N0434.1 Interval

The resequencing of defiined regions in the N0434.1 interval resulted in the discovery of 55 SNPs. As the total length of resequenced DNA is also known, it is possible to estimate the nucleotide diversity across the interval by calculating θ_w and π.

6.2.3.a: Estimate of Nucleotide Diversity Using θ_w

Watterson's θ can be calculated in terms of *K*, the observed number of variant sites (SNPs), using the formula:

$$\theta = \frac{1}{L}\frac{K}{\sum_{i=1}^{n-1}\frac{1}{i}}$$

in which *n* is the number of chromosomes analysed and *L* is the total sequence length screened. In this way, θ_w across the whole N0434.1 interval is estimated to be 6.3 x 10^{-4}.

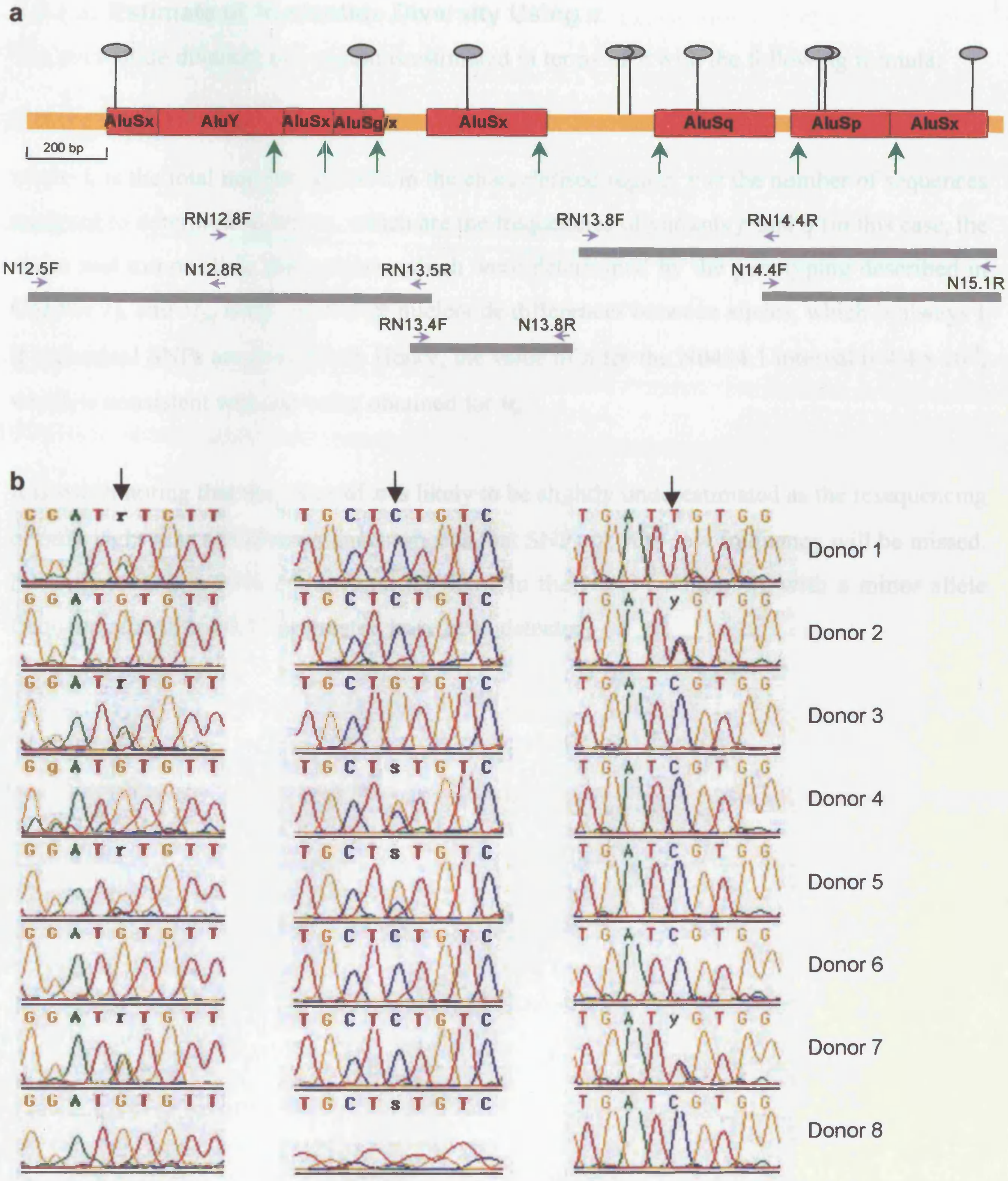


Figure 6.6: An example of the resequencing strategy for SNP discovery used in the N0434.1 interval. (a) Resequencing amplicon H is denoted by the orange line. Alu elements are shown by the red boxes and their poly(A/T) tails are marked by green arrows. Sub-amplicons that had to be PCR amplified due to the high density of Alu elements are shown by the grey lines and the positions of primers used for amplifications and resequencing are marked by the purple arrows. (b) Three examples of SNP identification after resequencing the N0434.1 interval in eight unrelated men and aligning the sequence reads using AutoAssembler software. SNP positions are marked with black arrows .

6.2.3.a: Estimate of Nucleotide Diversity Using *π*

The nucleotide diversity of a region is estimated in terms of *π* with the following formula:

$$\pi = \frac{1}{L}\frac{n}{n-1}\sum_{p,q} x_p x_q \Pi_{pq}$$

where *L* is the total number of SNPs in the characterised region, *n* is the number of sequences analysed to determine x_p and x_q, which are the frequencies of variants *p* and *q* (in this case, the major and minor allele frequencies, which were determined by the genotyping described in Chapter 7), and Π_{pq} is the number of nucleotide differences between alleles, which is always 1 if individual SNPs are considered. Hence, the value of *π* for the N0434.1 interval is 4.4×10^{-4}, which is consistent with the value obtained for θ_w.

It is worth noting that the value of *π* is likely to be slightly underestimated as the resequencing of only eight men (16 chromosomes) means that SNPs of very low frequency will be missed. Nevertheless, it is 95% certain that all SNPs in the N0434.1 interval, with a minor allele frequency (mAF) of 0.17 or greater, have been detected.

Table 6.1: SNPs in the N0434.1 interval. All were identified in a panel of 8 UK semen donors of North European origin and are denoted as transitions (Ti), transversions (Tv) or insertion/deletion polymorphisms (ID). Transitions at CpG dinucleotides are underlined. SNP nomenclature reflects their nucleotide position from the most distal end of the N0434.1 sequence and the two possible alleles at that position.

SNPs telomeric to PGMS2			*SNPs centromeric to PGMS2*		
423C/T	(Ti)	in Alu	18530A/C	(Tv)	in Alu
1107A/G	(Ti)	in Alu	18610C/T	(Ti)	in Alu
2357C/T	(Ti)		18634C/T	(Ti)	in Alu
2458C/T	(Ti)		18664C/G	(Tv)	in Alu
2793C/T	(Ti)	in coding sequence	18970A/G	(Ti)	
3103C/T	(Ti)		19963C/G	(Tv)	in Alu
3370+/-	(ID)		20723C/G	(Tv)	
4383G/T	(Tv)		21309C/G	(Tv)	in Alu
4589A/G	(Ti)		22030A/G	(Ti)	in *PLCXL1* 3´ UTR
4722C/T	(Ti)	in Alu	22145A/G	(Ti)	
4823A/G	(Ti)	in Alu	22186C/T	(Ti)	
5230A/G	(Ti)		23133A/G	(Ti)	in Alu
7225C/T	(Ti)		23483A/T	(Tv)	in *PGPL* 3´ UTR
7298A/G	(Ti)		23757A/G	(Ti)	in *PGPL* 3´ UTR
8020C/T	(Ti)		24151C/T	(Ti)	in Alu
8248C/T	(Ti)	in Alu	26645C/T	(Ti)	
8429C/G	(Tv)		26901A/G	(Ti)	
9771C/G	(Tv)		27689C/G	(Tv)	in Alu
11501A/G	(Ti)	in Alu	27919A/G	(Ti)	
12237A/G	(Ti)	in Alu	30381C/T	(Ti)	
12657A/G	(Ti)	in Alu	31679C/T	(Ti)	
13351C/T	(Ti)	in Alu	31697A/T	(Tv)	
13642A/G	(Ti)	in Alu	31703C/T	(Ti)	
13994A/G	(Ti)		31717C/G	(Tv)	
14051A/G	(Ti)		32258A/G	(Ti)	
14209C/T	(Ti)	in Alu	32451C/T	(Ti)	
14537A/G	(Ti)	in Alu			
14567A/G	(Ti)	in Alu			
14893A/G	(Ti)	in Alu			

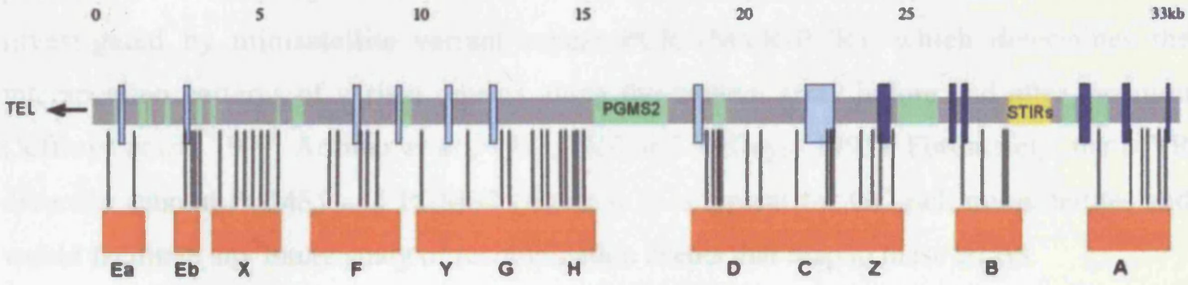


Figure 6.7: The distribution of SNPs in the N0434.1 interval. Adapted from figure 6.5. SNPs are shown by the vertical black lines between the N0434.1 sequence and the resequencing amplicons.

6.3: Discussion

The work presented in this chapter forms the first overview of diversity in the *PGPL* region. Not only did this include the identification of SNPs, which are essential for analyses of LD and recombination (Chapters 7 and 8), but also involved studies of tandem repeat DNA, which potentially associates with recombinationally active DNA.

6.3.1: PGMS1 and PGMS2 Variability

Strictly speaking, the minisatellite studies presented in this chapter were merely investigations into the different allele sizes that were present in a relatively small panel of unrelated men, and not a reliable system for determining mutation rates. Indirect population genetic estimates of the PGMS1 and PGMS2 mutation rates were obtained, demonstrating that both are likely to be only moderately unstable, as the mutation rates are well below the values of 1% or higher observed at classic unstable minisatellite loci. Furthermore, PGMS1 and PGMS2 both have a limited number of allele sizes compared to more unstable loci. Therefore, it appears as though the N0434.1 interval does not contain a highly variable minisatellite that might have evolved from a local meiotic recombination hotspot, or more generally from recombinationally active DNA, as has been shown at MS32 (Jeffreys *et al.*, 1998*a*).

6.3.1.a: Future Development of Minisatellite Studies

The moderate variability suggested that further analyses of PGMS1 and PGMS2 would not be worthwhile at this stage. However, it was noted that if the region around the minisatellites was determined to be recombinationally active, and exchange points appeared to be within PGMS1 or PGMS2, it would be wise to return to the relevant array to determine both the positions of the exchange events and the structures of the crossover products. This could be investigated by minisatellite variant repeat PCR (MVR-PCR), which determines the interspersion patterns of variant repeats along the tandem array before and after mutation (Jeffreys *et al.*, 1991; Armour *et al.*, 1993; Neil and Jeffreys, 1993). Fortunately, the MVR diversity seen at PGMS1 and PGMS2 (figure 6.1) is typical for GC-rich minisatellites and would facilitate any future study of recombination events that map to these arrays.

There are twenty other novel minisatellites within N0434.1, C3A, 29C1A and H22A, for which nothing is known other than their sequence (tables 3.3, 4.2, 4.3, 4.4). Therefore, an initial investigation into their variability, such as that presented in this chapter for PGMS1 and

PGMS2, may provide further insights into the variability of the distal PAR1 region. For example, a continuing pattern of allele size variability, as has been seen at PGMS1, PGMS2 (this chapter), CEB12, CEB30 (Vergnaud *et al.*, 1993), and more dramatically at DXYS14 and DXYS20 (Page *et al.*, 1987; Inglehearn and Cooke, 1990), at most of the other novel tandem repeat loci in the region will have implications for the length of PAR1 as a whole. If PAR1 is significantly variable in its length, there must then be a system to constrain the variability and allow pre-meiotic pairing, for example, to proceed as normal. Clearly these points are only speculative and a study of minisatellite variability in PAR1 would only offer a part of the data needed to fully analyse the structure and mechanisms of this extremely important, yet relatively neglected, area of the human genome.

6.3.2: N0434.1 Pseudoautosomal STIR Array

The work presented in this chapter included the first investigation of length variability at a STIR array. The results were intriguing, but perhaps not surprising given the fact that the STIR array is essentially a GC-rich minisatellite with an unusually large repeat unit (figure 6.3). Therefore, its variability is suggested to be caused by the same kind of recombinational process that destabilises more typical minisatellites. In addition, it is noteworthy that even the limited internal heterogeneity of the STIR array repeats (figure 6.3) is sufficient to allow the molecular structure of the STIR array instability to be investigated by MVR-PCR. The study has also raised further questions as to a putative biological role for STIRs, particularly as these sequences have been discovered in most orders of mammals (W. Schempp and B. Weber, cited in Petit *et al.*, 1990; see Chapter 3, section 3.1.1.c).

The protermini of chromosomes are the regions where initiation of homologue pairing preferentially takes place prior to meiotic recombination (e.g. Laurie and Hultén, 1985, see Chapter 1, section 1.2.2), and thus support high levels of meiotic recombination itself (NIH/CEPH Collaborative Mapping Group, 1992; Weissenbach *et al.*, 1992; Mohrenweiser *et al.*, 1998; Kong *et al.*, 2002). The colocalisation of GC-rich minisatellites (Royle *et* al., 1988) and recombination nodules and chiasmata (Hulten, 1974; Maudlin and Evans, 1980; Solari, 1980) to the end regions of chromosomes, led to the suggestion that minisatellites are the sequences within proterminal regions that support homologue recognition (Ashley, 1994; Sybenga, 1999). However, in PAR1, this putative function could now be ascribed to pseudoautosomal STIRs, which appear to represent a more complex and conserved class of element than the monomeric autosomal STIRs (Chapter 3, section 3.1.1.c) (Rouyer *et al.*, 1990) and might, therefore, have a role that is biologically more important. Due to the restricted homology between the X and Y chromosomes, PAR1 is the site of an obligatory crossover during male meiosis (Burgoyne, 1982). Therefore, the increased sequence conservation between pseudoautosomal STIRs might be a reflection of their essential function within PAR1 to support homologous pairing of the X and Y chromosomes at male meiosis and thus allow recombination to proceed. This putative STIR function within the autosomes may not be so important as autosomal homologues can pair along their entire length. Alternatively, pseudoautosomal STIRs might serve to prevent exchanges between non-homologous chromosome ends and thereby maintain the integrity of the essential PAR1 region from one generation to the next. In view of this, it is interesting to note the suggestion that chromosome ends can be exchanged at a low rate between non-homologous autosomes (Wilkie *et al.*, 1991), which is likely to be mediated by conserved sequence features between subtelomeric regions such as those identified by Flint *et al.* (1997). However, this process has not been observed at PAR1 and, indeed, PAR1 does not even cross-hybridise to the subterminal repeats of other human chromosome ends (Royle, 1995).

Clearly, any homologue-recognition function that is attributed to pseudoautosomal STIRs must be reconciled with the variability that has been demonstrated for the N0434.1 STIR array. Perhaps as the variability and (inferred) instability are only very modest, any such function may not be impeded. Alternatively, the variability of STIRs may stem from their suggested role within meiotic recombination and, as initiators or supporters of homologous pairing, they may serve as occasional markers of recombination hotspots, as has been suggested previously for GC-rich minisatellites (Jeffreys *et al.*, 1998*a*). These suggestions could be tested directly by charting the relationship between the location of STIRs and the positions of recombination initiation events, which could be mapped through a gene conversion assay (e.g. see Chapter 9, section 9.4 and figure 9.1).

6.3.3: Nucleotide Variability in the N0434.1 Interval

6.3.3.a: SNP Distribution

Approximately 25 kb of the N0434.1 interval was resequenced, identifying 55 SNPs and demonstrating that the frequency of SNPs across the region is approximately one per 450 bp. Over 40% (23/55) of the SNPs were identified within Alu elements but it would be misleading to think of this as significant, as Alu elements constitute approximately 40% of the nucleotides within the resequencing amplicons. Ordinarily, the high incidence of CpG dinucleotides within Alu elements (Batzer *et al.*, 1996) predisposes them to point mutations and has been suggested as a mechanism for the prevention of Alu-mediated nonhomologous recombination events (Batzer and Deininger 2002). However, in the N0434.1 interval, only six of the 23 SNPs that are within Alu elements are at CpG dinucleotides (table 6.1), which suggests that the overall level of nucleotide diversity within the region is not actually influenced by the Alu elements.

6.3.3.b: Estimations of Nucleotide Variability

The estimations of diversity across the N0434.1 interval ($\theta_w = 6.3 \times 10^{-4}$; $\pi = 4.4 \times 10^{-4}$) are very similar to those obtained from large genome-wide surveys, in which $\pi = 5.4 \times 10^{-4}$ (Cargill *et al.*, 1999) to 8.5×10^{-4} (Halushka *et al.*, 1999). Very little is known about the levels of nucleotide diversity in PAR1, but the values of θ_w and π within the N0434.1 interval and those determined for the *SHOX* region ($\theta_w = 6.3 \times 10^{-4}$; $\pi = 8.7 \times 10^{-4}$) (May *et al.*, 2002) contrast with the far higher levels of diversity reported at *PPP2R3B* (figure 1.8*a*), where π was shown to be an order of magnitude bigger, at 5.4×10^{-3} (Schiebel *et al.*, 2000), and at the Xp/Yp telomere-adjacent region, which has a SNP frequency of one per 65 bp in Caucasians and one per 50 bp in Africans (Baird *et al.*, 1995). Schiebel and colleagues (2000) suggested that levels of high diversity resulted from either high GC-content that favours mutation at CpG dinucleotides or from base-mutagenic recombination, a proposition that was later contradicted by the normal level of diversity around *SHOX*, which is situated in a recombinationally active and GC-rich region of PAR1 (May *et al.*, 2002). The GC-rich N0434.1 interval is also expected to be recombinationally active (Lien *et al.*, 2001) and has now been shown to have a normal level of nucleotide diversity, which argues further against GC-content and base-mutagenic recombination being the causes of differing levels of diversity in the genome. Although resolving this issue is far from complete, different evolutionary histories of genomic regions, rather than population genetic processes, are likely to have a major role in shaping nucleotide diversity profiles across the genome, as suggested by May *et al.* (2002) and demonstrated at the MHC (Gaudieri *et al.*, 2000; Kauppi *et al.*, 2003). For example, in a recombinationally suppressed region of DNA (LD block), allele fixation or extinction at different SNPs will not occur independently; the loss of a haplotype from a population will result in loss of all variants restricted to that haplotype. Thus, recombinationally inactive LD blocks might occassionally be in periods of low or high diversity that are caused by chance or through processes such as selective sweeps (Kauppi *et al.*, 2003). In contrast, recombinationally active DNA should show relatively stable levels of nucleotide diversity as SNPs would be able to escape correlated extinctions by being constantly reshuffled onto different haplotypes.

6.4: Concluding Remarks

The analysis of diversity in the N0434.1 region has shown that PGMS1, PGMS2 and the pseudoautosomal STIR array are variable and has provided an estimate of their mutation rates. The list of loci within PAR1 that have length variability continues to grow, but whether or not this causes significant length variability in PAR1 as a whole remains to be established. The first study of variability at a STIR array has suggested that pseudoautosomal STIRs could be destabilised by the same kind of recombinational processes that mutate more typical minisatellites to new length alleles and established that this could be investigated through MVR-PCR. In addition, the study has raised further questions as to a putative biological role for pseudoautosomal STIRs, which is suggested to be some kind of homologue-recognition or pairing role prior to meiotic recombination. However, it is clear that further analysis of other STIR arrays has to take place before drawing any firm conclusions.

It has been demonstrated that the GC-rich N0434.1 region has a genome-average level of nucleotide diversity, which is further evidence that PAR1 is not globally enriched in SNPs. With specific regard to this thesis, the novel SNPs have provided the tool for determining the pattern of LD across the N0434.1 interval, the next step in the analysis of meiotic recombination in the region.

Chapter 7

Linkage Disequilibrium in the N0434.1 Interval

7.1: Introduction

There are very few clues as to the overall linkage disequilibrium (LD) pattern and distribution of recombination events within PAR1, though low resolution data from families and from typing single sperm (Lien *et al.*, 2000) do suggest that male crossovers are fairly randomly distributed across PAR1, with only modest regional variation in recombination efficiency. In addition, the intense LD immediately adjacent to the PAR1 telomere (Baird *et al.*, 1995) suggests that there is a boundary of recombination proximal to the telomere, but this, and the work that led to the discovery of the *SHOX* meiotic recombination hotspot (May *et al.*, 2002), remain the only high resolution surveys of LD within PAR1. Consequently, the SNPs within the N0434.1 interval (Chapter 6, table 6.1) were subjected to LD analysis to generate the first high resolution patterns of LD in the region; any evidence of a sudden decline of LD would suggest that hotspotting, as seen at *SHOX* (May *et al.*, 2002) and in the MHC II region (Jeffreys *et al.*, 2000; 2001), is also a feature of recombination around the *PGPL* and *PLCXL1* genes. Also, SNPs in the N0434.1 interval could be analysed for association with those in the telomere-adjacent region to potentially narrow down the interval containing the putative boundary of recombination.

7.1.1: Measuring Linkage Disequilibrium

LD is the non-random association of alleles at closely linked loci (Chapter 1, section 1.2.4) and, although it is a relatively simple concept, a satisfactory single measure of LD has yet to be determined. Most measures of LD capture the strength of association between a pair of bi-allelic sites and the most widely used pairwise measures of LD are $|D'|$ (Lewontin, 1988) and Δ^2 (also denoted r^2) (reviewed in Devlin and Risch, 1995), which are derived as follows:

For two bi-allelic markers (such as a pair of SNPs), one with alleles A, a and the other with alleles B, b, the LD parameter is defined as $D = p_{AB} - p_A p_B$, where p_{AB} is the population frequency of the haplotype AB, and p_A and p_B refer to the allele frequencies of the two loci. To avoid dependence on allele frequencies D can be scaled as $D' = D/D_{max}$, where D_{max} is the lesser of $p_A p_B$ or $p_a p_b$ if D is negative, or $p_a p_B$ or $p_A p_b$ if D is positive. As D (and thus D') is positive or negative depending only on the arbitrary labelling of alleles, the absolute value, $|D'|$, which varies from 0 to 1, is normally used. The case of $|D'| = 1$ is known as complete LD and will only occur if two loci have not been separated by historical recombination (or recurrent mutation), meaning that a maximum of three out of the four possible two-locus haplotypes are observed. Free association of loci is indicated when $|D'| = 0$, and values of $|D'| < 1$ indicate that the complete ancestral LD has been disrupted. However, the relative magnitude of $|D'|$ values between 0 and 1 has no clear interpretation (Ardlie *et al.*, 2002). Furthermore, $|D'|$ is artificially inflated in small samples, particularly when rare alleles are examined, so intermediate values are of little use for measuring the strength of LD or comparing levels of LD between studies.

The measure Δ^2 represents another scaling of D and is the statistical correlation between alleles at two sites. It is obtained by dividing D^2 by the product of the four allele frequencies at the two loci. So, for the markers above, $\Delta^2 = D^2/(p_A p_a p_B p_b)$. Perfect LD (the case of complete association) is indicated by $\Delta^2 = 1$, which cannot occur unless the allele frequencies at both loci are equal and only two of the possible four haplotypes are observed. The Δ^2 measure has several properties that make it more useful than $|D'|$. For example, under the standard neutral model of molecular evolution (Chapter 6, section 6.1.1), population genetics theory provides a simple relationship between Δ^2, the effective population size (N_e), the recombination rate per unit distance (r) and inter-marker distance (d) as follows: $\Delta^2 = 1/(1 + 4N_e rd)$ (Ohta and Kimura, 1971). Furthermore, in contrast to $|D'|$, intermediate values of Δ^2 are easily interpretable as they are inversely proportional to the sample size required to detect statistically significant LD between two loci. Hence, a Δ^2 value is related to the amount of information provided by one locus about the other (Kruglyak, 1999; Pritchard and Przeworski, 2001; Weiss and Clark, 2002). Δ^2 also shows much less bias upwards in small samples but, due to its dependency on allele frequency, is typically lower than $|D'|$ for any chromosomal distance (Weiss and Clark, 2002). However, measures of $\Delta^2 < 1$ can occur even without recombination but values of $|D'| < 1$ signal historical recombination events and are thus more useful for the analyses presented in this chapter and Chapter 8.

7.2: Results

7.2.1: Genotyping

Diploid genotypes of the SNPs in the N0434.1 interval were generated by allele-specific oligonucleotide (ASO) hybridisation in a panel of 50 unrelated UK semen donors of north European descent. Genotyping semen donors allows the identification of suitable individuals for sperm crossover analysis (Chapter 8).

The N0434.1 regions that were used to identify SNPs (Chapter 6, section 6.2.2.b) were PCR amplified from the genomic DNA of each man in the semen donor panel. As a high proportion of the N0434.1 SNPs are within Alu sequences (Chapter 6, table 6.1), the target sequences for some ASOs were duplicated, meaning that it was sometimes necessary to generate smaller, less complex amplicons to ensure unambiguous genotyping. Usually, these sub-amplicons were the same as those that had been generated to facilitate SNP identification, and are detailed in Chapter 2, table 2.6. ASOs were designed for both alleles of each SNP (table 2.7) and, for each SNP, a dot blot of the relevant PCR products was hybridised sequentially with one ASO followed by the other (figure 7.1). This approach provides full, internal control for differences in DNA loading per dot, and makes genotyping easier. Furthermore, the method is simple, robust and allowed the unambiguous genotyping of 98.6% of N0434.1 SNPs. The very few incidences in which the genotype could not be determined were a result of weak signals on the dotblots (probably caused by poor yield of the PCR product). This high accuracy is very important because even a modest error rate will create the appearance of "rare variant" haplotypes that do not exist in nature (Gabriel *et al.*, 2002). The full diploid genotypes of all 55 N0434.1 SNPs in all 50 semen donors is shown in figure 7.2. None of the SNPs showed significant departure from Hardy-Weinberg equilibrium (data not shown).

7.2.1.a: Distribution of Minor Allele Frequencies

The minor allele frequencies (mAFs) at each N0434.1 SNP site were determined from the genotype data (table 7.1) and their distribution was compared to that of 35,989 SNPs discovered on chromosome 21 (Patil *et al.*, 2001) (figure 7.3). Within N0434.1 nearly 75% of the SNPs have a mAF that is greater than 0.3 and SNP 4823A/G is the only example of a singleton, i.e. a SNP with a minor allele that is observed only once. Under the neutral model of molecular evolution, and given the estimation of nucleotide diversity within N0434.1

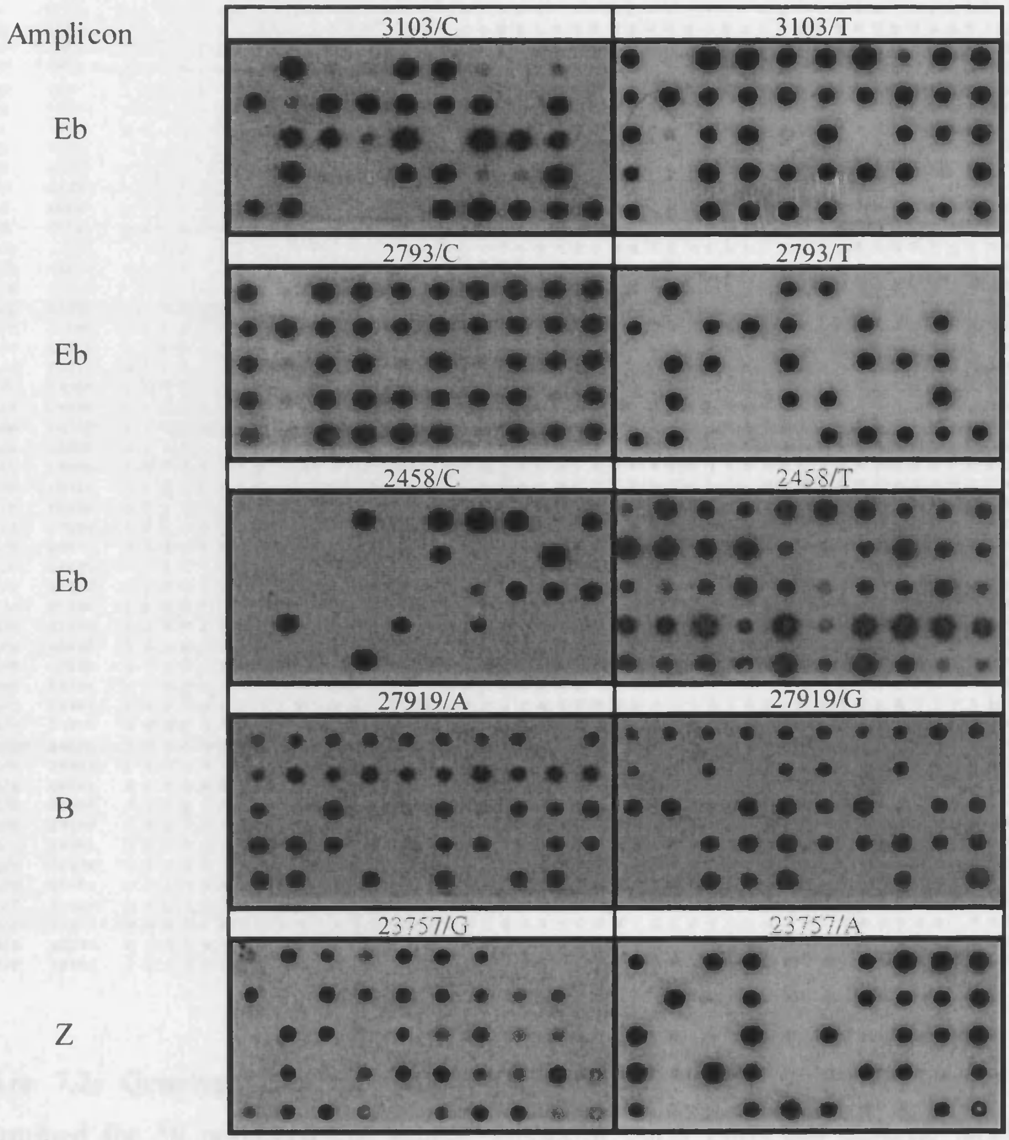


Figure 7.1: Examples of SNP genotyping in the N0434.1 interval by ASO hybridisation. Resequencing amplicons used in Chapter 6 to identify SNPs were PCR amplified from 50 unrelated UK semen donors. The PCRs were dot blotted in order 1–50, reading left to right and top to bottom. As each resequencing amplicon contained multiple SNP sites, four dotblots of each amplicon were prepared. Single dotblots were then probed sequentially with the ASO for one allele of a relevant SNP followed by the other. Hence, in the top panel above, a dotblot of the Eb amplicon has been probed first with the ASO specific to 3103/C and then with the ASO for 3103/T; donor 1 is a TT homozygote, donor 2 is a CC homozygote and donor 5 is a heterozygote.

name	pos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
423T/C	423	C	T	C	C	H	H	C	C	C	C	H	C	H	C	H	H	H	C	H	C	C	T	H	C	T	C	T	H	H	C	C	T	C	C	H	H	C	C	T	C	H	T	C	C	C	C	T	H	H	H
1107G/A	1107	H	G	H	H	G	G	H	A	A	A	G	H	G	A	G	G	H	H	H	A	A	G	G	A	G	H	G	H	H	H	A	G	H	H	G	G	H	H	G	H	H	G	G	H	A	H	G	G	H	H
2357C/T	2357	H	C	H	H	C	C	H	T	T	T	C	H	C	H	C	C	H	H	H	T	T	C	C	T	C	H	C	H	H	H	H	C	H	H	C	H	H	H	C	H	H	C	C	H	T	C	C	C	H	H
2458C/T	2458	T	C	T	T	H	H	T	T	T	T	H	T	H	H	H	H	H	T	H	T	T	C	H	T	C	T	C	H	H	T	T	C	T	T	H	H	T	T	C	T	H	C	T	T	T	H	C	H	H	H
2793C/T	2793	C	T	C	C	H	H	C	C	C	C	H	C	H	H	H	C	H	C	H	C	C	T	H	C	T	C	T	H	H	C	C	T	C	C	H	H	C	C	T	C	H	T	C	C	C	H	T	H	H	H
3103C/T	3103	H	C	H	H	C	C	H	T	T	T	C	T	C	H	C	C	H	H	H	T	T	C	C	T	C	H	C	H	H	T	T	C	T	H	C	C	H	H	C	H	H	C	C	H	T	H	C	C	H	H
3370+/-	3370	+	-	+	H	H	H	+	+	+	+	H	+	H	H	H	?	H	+	H	+	+	-	H	+	-	+	-	H	H	+	+	-	+	+	H	H	+	+	-	+	H	-	+	+	+	H	-	H	H	H
4383G/T	4383	G	T	G	H	H	H	G	G	G	G	H	G	H	H	H	?	H	G	H	G	G	T	H	G	T	G	T	H	H	G	G	H	G	G	H	H	G	G	T	G	H	T	G	G	G	T	T	H	H	H
4589A/G	4589	H	G	H	H	G	G	H	A	A	A	G	H	G	H	G	?	H	H	H	A	A	G	G	A	G	A	G	H	H	H	A	G	H	H	G	G	H	H	G	H	H	G	G	H	A	G	G	G	H	H
4722C/T	4722	T	C	T	T	H	H	T	T	T	T	H	T	H	H	H	?	H	T	H	T	T	C	H	T	C	T	C	H	H	T	T	C	T	T	H	H	T	T	C	T	H	C	T	T	T	H	C	H	H	H
4823A/G	4823	A	A	A	H	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A
5230G/A	5230	G	G	H	G	H	H	H	G	G	G	G	G	G	G	H	?	G	H	?	?	G	G	G	G	G	H	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G
7225C/T	7225	H	C	H	H	C	C	H	T	T	H	C	T	H	H	C	C	H	H	H	T	H	C	C	T	C	H	C	H	H	H	T	H	H	H	H	H	T	H	C	T	H	C	H	H	H	H	C	H	H	H
7298A/G	7298	H	A	G	H	H	H	G	G	G	H	A	G	H	H	H	H	H	G	H	G	G	A	A	G	A	G	A	G	H	H	G	H	H	G	H	H	G	H	A	G	H	H	G	H	G	H	A	H	H	H
8020C/T	8020	H	C	H	H	C	C	H	T	T	H	C	T	H	H	C	C	H	C	H	T	T	H	C	T	C	H	H	H	H	H	T	H	H	H	H	H	T	H	C	T	H	C	H	H	T	H	C	H	H	H
8248C/T	8248	H	C	H	H	C	C	H	T	T	H	C	T	C	T	C	H	H	C	H	H	H	H	C	H	C	H	H	?	H	H	H	H	H	H	H	H	H	H	C	T	H	T	H	H	T	H	C	H	H	H
8429C/G	8429	C	C	H	H	H	H	H	C	C	C	H	C	C	C	H	H	C	H	C	C	C	C	H	C	C	H	C	C	C	C	C	C	C	H	C	C	C	H	C	C	C	C	H	H	C	C	C	C	C	C
9771C/G	9771	C	G	C	C	H	H	C	C	C	H	H	H	H	H	H	C	H	H	G	H	C	H	H	H	G	C	G	H	H	H	C	H	H	H	H	H	C	H	G	C	H	G	C	C	H	H	H	H	G	H
11501A/G	11501	H	H	H	H	H	G	H	A	H	G	G	H	H	H	G	?	H	G	H	H	A	H	G	H	G	H	G	G	H	H	?	H	H	H	H	H	A	H	G	A	H	G	H	H	H	H	G	H	H	H
12237A/G	12237	A	H	G	H	H	G	H	A	H	G	H	A	H	H	G	?	H	H	A	A	A	H	H	A	H	H	H	G	H	H	?	H	H	G	H	A	A	G	A	G	H	G	H	A	A	H	G	H	A	H
12657G/A	12657	G	G	H	G	H	H	H	G	H	H	G	H	G	G	H	?	G	A	H	H	G	G	G	G	G	H	H	H	H	G	H	A	G	H	G	H	G	G	G	G	G	G	H	G	H	G	H	G	G	G
13351T/C	13351	H	H	H	H	H	C	H	T	H	C	H	C	C	H	C	?	H	C	H	H	T	C	C	T	C	?	?	C	H	C	H	H	C	H	H	H	T	T	C	T	H	C	H	H	T	C	C	H	H	H
13642A/G	13642	H	A	H	H	H	H	H	A	H	H	H	A	A	A	H	G	A	H	A	A	A	A	H	A	H	H	A	H	H	A	H	H	A	H	A	A	A	A	A	A	A	A	H	H	A	A	A	A	A	A
13994A/G	13994	H	H	A	A	A	H	A	A	A	H	H	A	H	H	H	?	H	A	A	A	A	G	G	A	G	A	H	H	A	A	A	H	A	A	H	A	A	A	G	A	H	G	A	H	A	H	G	H	A	G
14051A/G	14051	A	A	H	H	H	H	H	A	H	H	A	A	A	A	H	?	A	H	A	A	A	A	A	A	A	H	A	H	H	A	H	A	A	H	A	A	A	A	A	A	A	A	H	A	A	A	A	A	A	A
14209C/T	14209	C	C	H	C	H	H	H	C	H	H	C	C	C	C	H	?	C	H	C	C	C	C	C	C	C	H	C	H	H	C	H	C	C	H	C	C	C	C	C	C	C	C	H	C	C	C	C	C	C	C
14537G/A	14537	H	H	G	H	G	H	G	H	G	H	H	G	H	H	H	?	H	G	G	G	G	H	A	G	A	G	H	H	G	G	G	H	G	G	H	G	G	G	A	G	H	A	G	H	G	H	A	H	G	H
14567G/A	14567	H	H	A	H	A	H	A	H	A	H	H	G	G	H	H	?	H	H	H	H	A	H	G	A	G	H	G	H	A	H	A	H	H	A	H	H	A	A	G	A	H	G	A	H	A	G	G	H	H	H
14893G/A	14893	G	G	H	G	H	H	H	G	H	H	G	G	G	G	H	G	G	H	G	G	G	G	G	G	G	H	G	H	H	G	H	G	G	H	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G
18530A/C	18530	H	A	H	A	H	C	H	A	H	C	H	C	C	H	C	H	H	C	H	H	A	A	C	A	H	C	H	C	H	H	H	H	H	H	H	H	H	H	H	H	C	C	H	H	H	C	C	H	H	H
18610C/T	18610	H	T	H	H	H	C	H	T	H	H	H	C	H	H	C	H	H	H	H	C	T	T	C	T	H	C	T	C	H	H	H	H	H	H	T	T	T	T	T	T	C	C	H	H	T	C	C	T	H	T
18634C/T	18634	H	T	H	T	H	C	H	T	H	H	H	C	H	H	C	H	H	H	H	C	T	T	C	T	H	C	T	C	H	H	H	H	H	H	T	T	T	T	T	T	C	C	H	H	T	C	C	T	H	T
18664C/G	18664	H	C	C	C	C	H	C	C	H	H	C	G	H	H	G	H	H	C	H	H	C	C	G	C	H	H	C	G	H	H	H	H	H	C	C	C	C	C	C	C	G	G	C	H	C	G	G	C	H	C
18970A/G	18970	H	?	?	G	A	A	H	G	G	?	H	A	A	H	A	H	H	A	A	?	A	G	A	G	H	A	H	A	A	?	H	H	H	H	G	A	G	G	G	?	A	A	H	H	G	A	A	G	A	G
19963C/G	19963	H	H	C	C	H	H	C	C	C	H	H	G	H	H	H	H	H	C	G	H	H	C	G	C	H	H	C	H	H	H	C	H	H	C	C	H	C	C	C	C	G	G	C	H	C	G	G	C	G	C
20723G/C	20723	H	H	G	G	H	H	G	G	G	H	H	C	C	H	H	H	H	G	C	H	H	G	C	G	H	H	G	H	H	H	G	H	H	G	G	C	G	G	G	G	C	C	G	H	G	C	C	G	C	G
21309G/C	21309	H	H	G	G	H	H	G	G	G	H	H	C	H	H	H	H	H	G	C	H	H	G	C	G	H	H	G	H	H	H	G	H	H	G	G	H	G	G	G	G	C	C	G	H	G	C	C	G	C	G
22030G/A	22030	H	H	G	G	H	H	G	G	G	H	H	A	H	H	H	H	H	G	A	H	H	G	A	G	G	H	G	H	H	H	G	G	H	G	G	H	G	G	G	G	A	A	G	H	G	A	A	G	A	G
22145A/G	22145	A	H	A	A	H	H	A	A	A	H	A	G	H	H	H	H	H	A	G	H	A	A	G	A	H	H	A	A	H	H	A	A	H	A	A	H	A	A	A	A	G	G	A	H	A	G	G	A	G	A
22168C/T	22168	H	H	T	T	H	H	T	T	T	H	H	C	H	H	H	H	H	T	C	H	H	T	C	T	H	H	T	H	H	H	T	H	H	T	T	H	T	T	T	T	C	C	T	H	T	C	C	T	C	T
23133G/A	23133	H	H	G	G	H	H	G	G	G	H	H	A	H	H	H	H	A	G	A	H	H	G	A	G	H	H	G	H	H	H	G	H	H	G	G	H	G	G	G	G	A	A	G	H	G	A	A	G	A	G
23483A/T	23483	H	H	T	T	H	H	T	T	T	T	H	A	H	H	H	H	A	T	A	H	H	T	A	T	H	T	T	H	H	H	T	H	H	T	T	H	T	T	T	T	H	A	T	H	T	A	A	T	A	T
23757G/A	23757	G	G	G	H	G	H	H	H	G	H	G	G	G	G	G	?	G	G	H	G	G	G	G	G	G	G	G	H	H	H	G	H	G	G	H	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G
24151C/T	24151	H	H	C	C	H	H	C	C	C	H	H	T	T	H	H	?	T	H	T	T	H	C	T	C	H	C	C	T	H	H	H	H	H	C	C	H	C	C	C	C	T	T	C	H	C	H	T	C	T	C
26645C/T	26645	H	H	H	H	H	H	H	H	T	H	H	C	H	C	H	H	C	H	C	C	H	T	H	T	T	H	T	C	H	H	C	C	H	T	T	H	H	T	H	H	C	H	?	H	T	C	C	H	C	T
26901G/A	26901	G	G	H	G	G	G	H	H	G	G	G	G	G	A	G	G	?	G	?	G	G	G	?	G	G	G	G	?	G	G	A	A	G	G	G	G	H	G	H	H	?	?	?	G	G	G	G	H	G	G
27689C/G	27689	C	H	C	H	H	H	C	C	C	C	C	C	C	C	H	C	H	H	H	C	H	C	H	C	C	C	C	H	H	C	C	C	C	C	C	H	C	C	C	C	C	H	C	C	C	C	?	C	H	C
27919G/A	27919	H	H	G	H	H	H	G	G	G	H	H	A	H	H	H	H	A	H	A	A	H	G	A	G	G	G	G	A	H	H	H	H	H	G	G	H	G	G	G	G	A	A	G	H	G	A	A	G	A	G
30381T/C	30381	T	T	H	T	H	H	H	T	H	T	T	T	T	T	H	H	T	H	T	T	T	T	T	T	T	H	T	T	H	T	T	T	T	H	T	T	T	T	T	T	T	T	H	H	T	T	T	T	T	T
31679C/T	31679	H	C	C	C	C	C	C	C	C	C	T	C	C	H	C	H	H	C	C	H	C	C	C	H	C	C	C	C	C	C	C	H	C	H	C	C	C	H	C	C	H	C	H	H	T	H	C	C	C	C
31697A/T	31697	H	H	A	H	A	A	A	A	A	H	H	T	H	H	H	H	T	H	T	H	H	A	H	A	A	A	H	T	H	A	H	H	A	H	A	T	A	A	A	A	T	H	H	H	A	T	H	A	T	A
31703C/T	31703	H	H	C	H	C	C	C	C	C	H	H	T	H	H	H	H	T	H	T	H	H	C	H	C	C	C	C	T	H	C	H	H	C	H	C	H	C	C	C	C	T	H	H	H	C	T	H	C	T	C
31717G/C	31717	G	G	G	G	H	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	G	G	G	G	G	G	G	H	G	G	G	G	G	G	H	G	G	G	G	G
32258G/A	32258	H	H	A	H	A	A	A	A	A	H	H	G	H	H	H	H	G	H	G	G	H	A	G	A	A	A	A	G	H	A	H	H	A	G	H	H	A	A	?	A	G	H	H	H	A	G	H	A	G	A
32451T/C	32451	H	C	H	H	T	T	T	T	T	H	C	C	H	H	H	H	C	H	H	C	H	T	C	H	T	H	H	C	H	H	H	H	H	H	H	C	T	H	T	T	C	H	H	H	C	C	H	T	H	T

Figure 7.2: Genotype data for the 55 SNPs in the N0434.1 interval. Genotypes were determined for 50 unrelated UK semen donors of north European descent, labelled 1–50 (above). The SNPs are listed from the most distal to the most proximal and their nomenclature reflects their nucleotide position counted from the distal end of the N0434.1 sequence (Appendix 1). Letters denote the alternative alleles at each SNP, with homozygotes shown as a single base and heterozygotes marked with a red H. Presence or absence polymorphisms are given the symbols + or -. Uncertain genotypes are denoted by a question mark.

Table 7.1: Minor allele frequencies (mAFs) of SNPs in the N0434.1 interval.

SNPs telomeric to PGMS2		*SNPs centromeric to PGMS2*	
*Name**	*mAF*	*Name**	*mAF*
423C/T	0.33	18530C/A	0.43
1107G/A	0.4	18610T/C	0.46
2357C/T	0.38	18634T/C	0.45
2458T/C	0.35	18664C/G	0.37
2793C/T	0.33	18970A/G	0.44
3103C/T	0.42	19963C/G	0.36
3370+/-	0.36	20723G/C	0.4
4383G/T	0.36	21309G/C	0.4
4589G/A	0.4	22030G/A	0.36
4722T/C	0.35	22145A/G	0.33
4823A/G	0.01	22168T/C	0.38
5230G/A	0.12	23133G/A	0.39
7225C/T	0.45	23483T/A	0.34
7298G/A	0.41	23757G/A	0.13
8020C/T	0.48	24151C/T	0.42
8248C/T	0.46	26645C/T	0.48
8429C/G	0.15	26901G/A	0.2
9771C/G	0.43	27689C/G	0.16
11501G/A	0.42	27919G/A	0.42
12237A/G	0.47	30381T/C	0.13
12657G/A	0.25	31679C/T	0.17
13351C/T	0.41	31697A/T	0.36
13642A/G	0.23	31703C/T	0.36
13994A/G	0.32	31717G/C	0.05
14051A/G	0.16	32258A/G	0.4
14209C/T	0.15	32451C/T	0.45
14537G/A	0.32		
14567A/G	0.45	**minor allele written second*	
14893G/A	0.14		

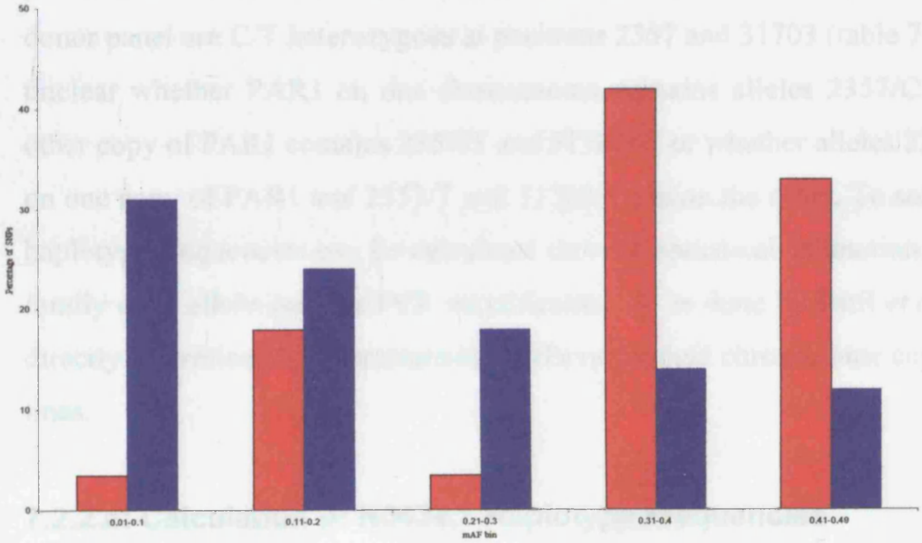


Figure 7.3: The distribution of mAFs of all 55 SNPs in the N0434.1 interval (red) compared to the distribution of 35,989 SNPs discovered on chromosome 21 (blue) (Patil *et al.*, 2001).

($\theta_w = 6.3 \times 10^{-4}$, $\pi = 4.4 \times 10^{-4}$) (Chapter 6, section 6.2.3), the observed number of singletons is significantly lower than expected (Fu and Li, 1993).This reflects the fact that only eight men (16 chromosomes) from one population were used for SNP identification (Chapter 6, section 6.2.2.b), which inevitably reduces the power to identify rare as compared to common SNPs. The more complete SNP ascertainment on chromosome 21 performed by Patil *et al.* (2001) demonstrated that singletons accounted for 32% of the SNPs they identified (figure 7.3). However, compared to the Patil *et al.* data, it does seem as though the N0434.1 interval is actually a region where common markers are over represented; the probability of detecting all SNPs with a mAF ≥ 0.1 in 16 chromosomes is 81% [$P = (1 - (0.9)^{16}) \times 100)$]. Furthermore, the relatively low number of SNPs that fall into the 0.21–0.3 mAF bin indicates that they are curiously under represented compared to the Patil *et al.* data, as it is more than 97% probable that all SNPs with a mAF ≥ 0.2 can be detected in 16 chromosomes.

In terms of using LD analysis to screen N0434.1 for evidence of historical recombination, the fact that most of the N0434.1 SNPs are high frequency is actually advantageous; SNPs with low mAF tend to represent relatively recent mutations on young haplotypes that have not had time to recombine and so can give misleadingly high LD values.

7.2.2: Linkage Disequilibrium in the N0434.1 Interval

The construction of haplotypes from diploid data is complicated by the fact that the linkage phase between any two SNPs is unknown. For example, a number of the men in the semen donor panel are C/T heterozygous at positions 2357 and 31703 (table 7.1). In these cases, it is unclear whether PAR1 on one chromosome contains alleles 2357/C and 31703/C and the other copy of PAR1 contains 2357/T and 31703/T, or whether alleles 2357/C and 31703/T are on one copy of PAR1 and 2357/T and 31703/C are on the other. To solve the phase problem, haplotype frequencies can be calculated through statistical estimation, direct inference from family data, allele-specific PCR amplification or, as done by Patil *et al.* (2001), they can be directly determined by characterising SNPs on haploid chromosome copies held in hybrid cell lines.

7.2.2.a: Calculation of N0434.1 Haplotype Frequencies

Analysis software, written by A.J. Jeffreys in TrueBASIC 4.1, estimated N0434.1 haplotype frequencies from the unphased diploid genotype data (figure 7.2) via a maximum-likelihood approach. Considering the example provided above (section 7.1.1) for two bi-allelic markers, one with alleles *A*, *a* and the other with alleles *B*, *b*, the program works as follows: from the observed allele frequencies at each locus pair it varies one haplotype frequency (p_{AB}) from 0 to the lesser value of p_A or p_B. From this, the corresponding values for each of the other three possible haplotypes are defined. The probablities (P) of getting the observed genotype data at the two loci are then computed and the value of p_{AB} that gives the maximum probability (P_{max}) of getting the observed genotype data is determined to produce the maximum likelihood values of all four possible haplotypes p_{AB}, p_{Ab}, p_{aB} and p_{ab}. These haplotype frequencies are then used to calculate $|D'|$ and Δ. The probability (P) is then re-calculated for haplotype frequencies at linkage equilibrium (i.e. $p_{AB} = p_A p_B$, etc) to give a value of P at equilibrium (P_{eq}). Calculating P_{eq} / P_{max} thus provides LR, the relative likelihood of obtaining the data under the most likely value of $|D'|$ versus the probability of obtaining the data when $|D'| = 0$. LR is independent of $|D'|$ and Δ measurements and thereby forms a statistical test of significance of association.

7.2.2.b: LD Analysis of N0434.1 with the PAR1 Telomere-Adjacent Region

LD analysis was initially restricted to the distal part of the N0434.1 region, using the SNPs telomeric to PGMS2 that have a mAF ≥ 0.18 (table 7.1), and three high frequency SNPs from the telomere-adjacent region (data provided by M. Hills). The LD plot (figure 7.4) shows that the distal 7.5 kb of the N0434.1 interval forms a recognisable LD block that is in free association with the telomere-adjacent region. This suggests historical recombinational activity in the 50–60 kb long GC-rich intervening region that contains a high density of both Alu sequences and tandem repeats arrays (Chapter 4). To the right of this LD block (i.e. further into the N0434.1 interval), the extent of LD declines with physical distance forming a pattern that is quite different to that observed in both *SHOX* and the MHC II region. Across MHC II 40–90 kb blocks of strong LD are abruptly separated by small intervals of free association, corresponding to local meiotic recombination hotspots. (Jeffreys *et al.*, 2001) (Chapter 1, figure 1.7). In contrast, the pattern throughout the *SHOX* region is one of very rapid decay of LD with distance, with many instances of closely linked (< 1 kb) markers in free association, which typically prevents the use of LD breakdown as a method to localise putative hotspots (May *et al.*, 2002) (Chapter 1, figure 1.8*c*). Within the N0434.1 region, there are two exceptions to the relatively close association observed between most markers. This could have been caused by inaccurate determination of genotypes, but this situation is highly unlikely as none of the SNPs deviate significantly from Hardy-Weinberg equilibrium (section 7.2.1, above).

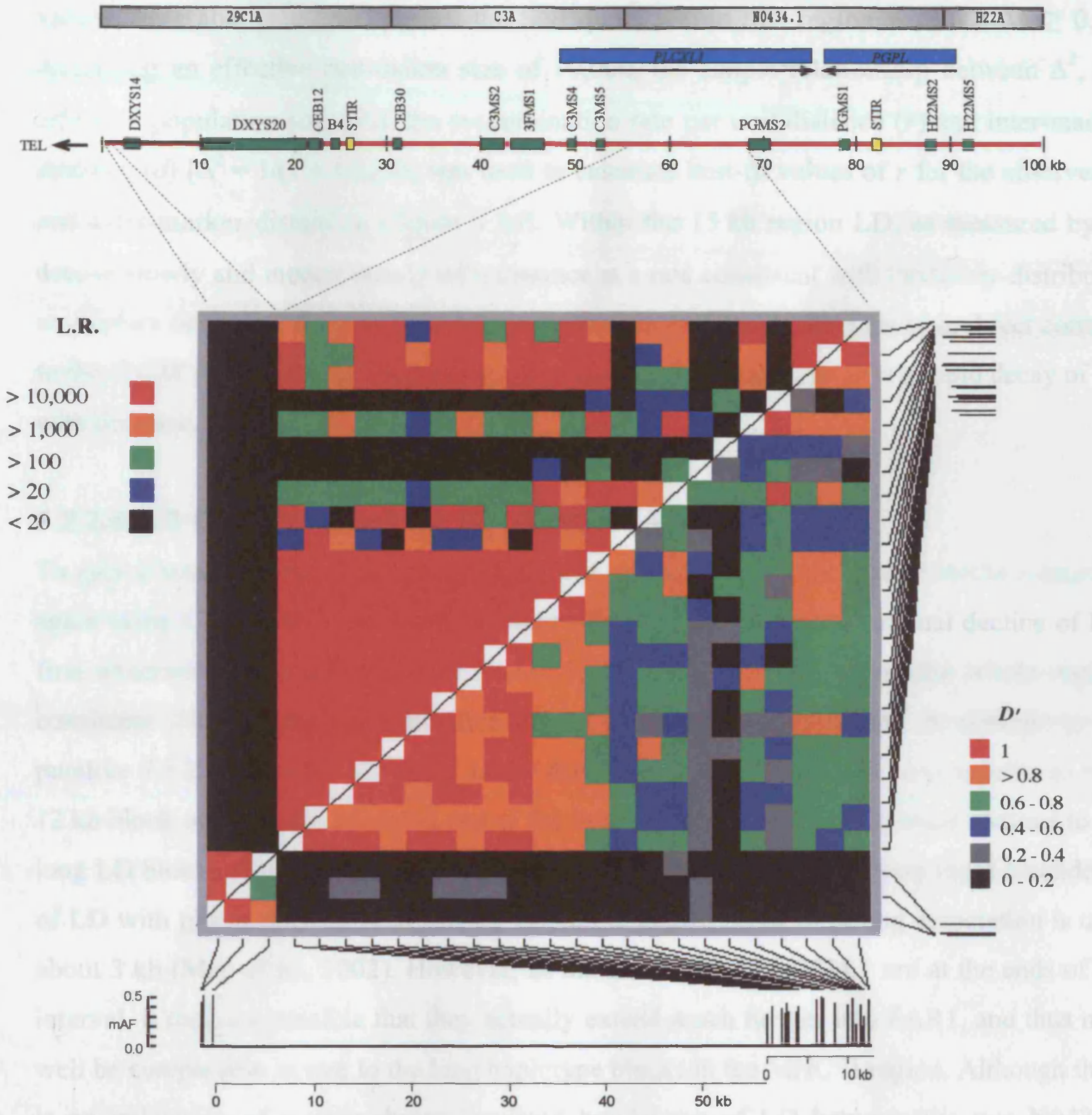


Figure 7.4: The LD pattern across the 15 kb region that is immediately distal to PGMS2 and the demonstration that this region is in free association with markers near the telomere. The top part of the diagram is derived from figure 4.9 and has been included to define the telomere-adjacent and N0434.1 regions that have been included in this analysis. $|D'|$ measures of complete LD between all pairs of markers with mAFs of at least 0.18 are shown in the bottom right triangle of the plot. $|D'| = 1$ for marker pairs showingonly two or three haplotypes (shown in red), $|D'|$ values less than 1 indicate SNP pairs with all four haplotypes(and thus suggest historical recombinational activity), with values of zero (in black) indicating pairs in free association. The top left triangle shows the likelihood ratio (LR) of association versus linkage equilibrium and thus the significance of each comparison. Points are plotted as rectangles centred on each pair of SNPs, shown below and to the right of the plot.

7.2.2.c: Gradual Decline of LD with Physical Distance in Distal N0434.1

To investigate the decay of LD within the N0434.1 region distal to PGMS2 more closely, Δ values were also calculated for all the SNP pairs within this region with a mAF ≥ 0.18. Assuming an effective population size of 10,000, the simple relationship between Δ^2, the effective population size (N_e), the recombination rate per unit distance (r) and inter-marker distance (d) [$\Delta^2 = 1/(1 + 4N_erd)$] was used to calculate best-fit values of r for the observed Δ and inter-marker distances, (figure 7.5*a*). Within the 15 kb region LD, as measured by Δ, decays slowly and monotonously with distance at a rate consistent with randomly-distributed crossovers occurring close to the genome average rate of 1 cM/Mb. This is in direct contrast to the *SHOX* region (figure 7.5*b*) where, even outside the hotspot, there is a rapid decay of LD with distance.

7.2.2.d: LD Pattern Across the Whole N0434.1 Interval

To gain a broader view of LD, the analysis was extended across the whole N0434.1 interval, again using SNPs with a mAF ≥ 0.18 (figure 7.6*a*, *b*). The slow and gradual decline of LD, first observed in the distal part of N0434.1, appears to extend across the whole region, consistent with a mean recombination rate of just 2 cM/Mb. However, in addition to the putative 7.5 kb LD block identified in the distal part of N0434.1, there also appears to be a 12 kb block of LD at the proximal end of the interval. Again, this in in marked contrast to the long LD blocks of the MHC II region (Jeffreys *et al.*, 2001), and to the very rapid breakdown of LD with physical distance at *SHOX*, where the largest block of strong association is only about 3 kb (May *et al.*, 2002). However, as the LD blocks in N0434.1 are at the ends of the interval, it remains possible that they actually extend much further into PAR1, and thus may well be comparable in size to the long haplotype blocks in the MHC II region. Although there is no indication of a very abrupt localised breakdown of LD between the two N0434.1 putative LD blocks, LD does decay relatively quickly with physical distance across the moderately variable minisatellite PGMS2. Using a subset of the N0434.1 SNPs, the decline of Δ with physical distance suggested that the rate of recombination across the PGMS2 region increases to about 13 cM/Mb (data not shown). It is possible that this indicates a localised region that has undergone historical recombination and might therefore contain a recombination hotspot.

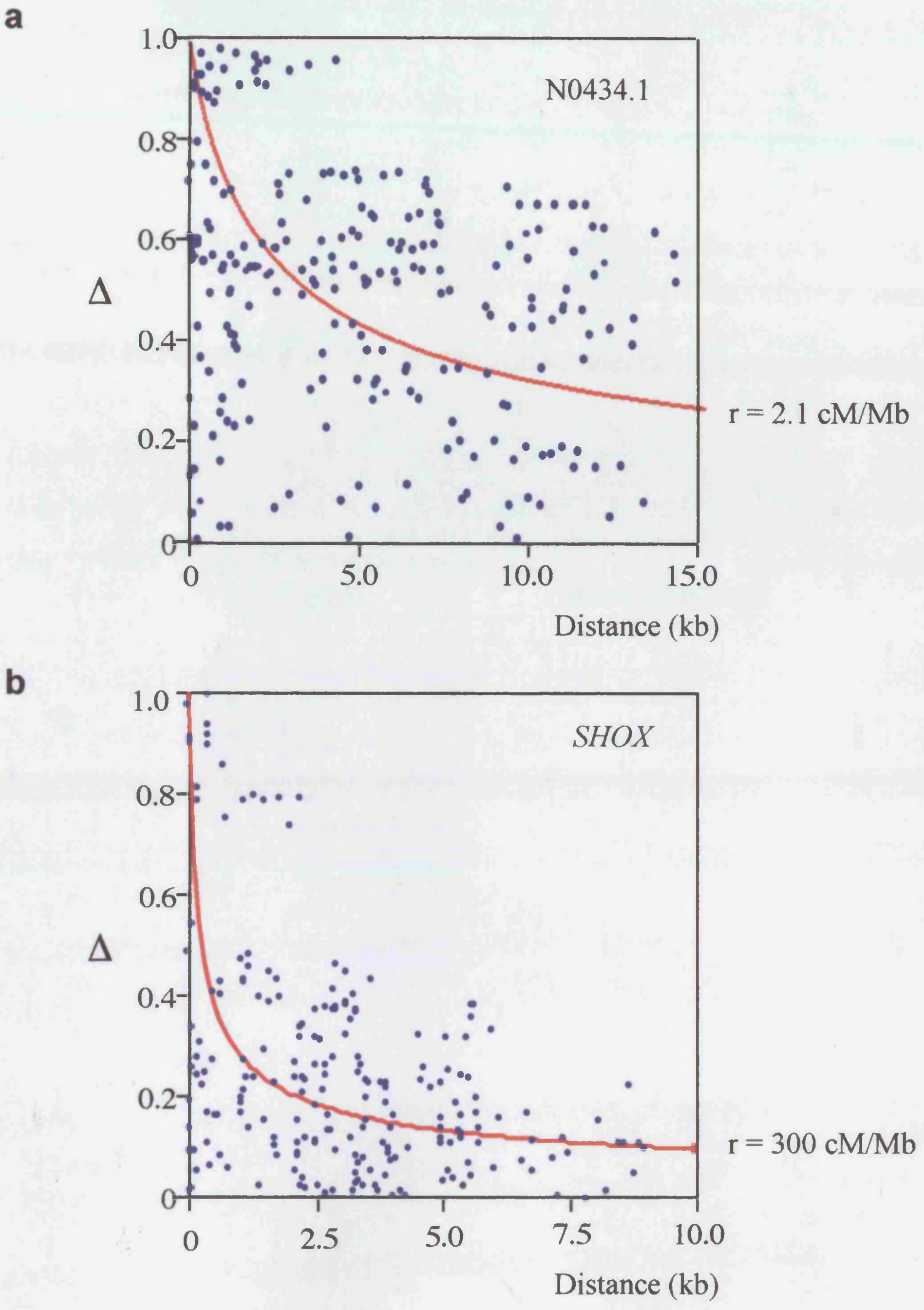
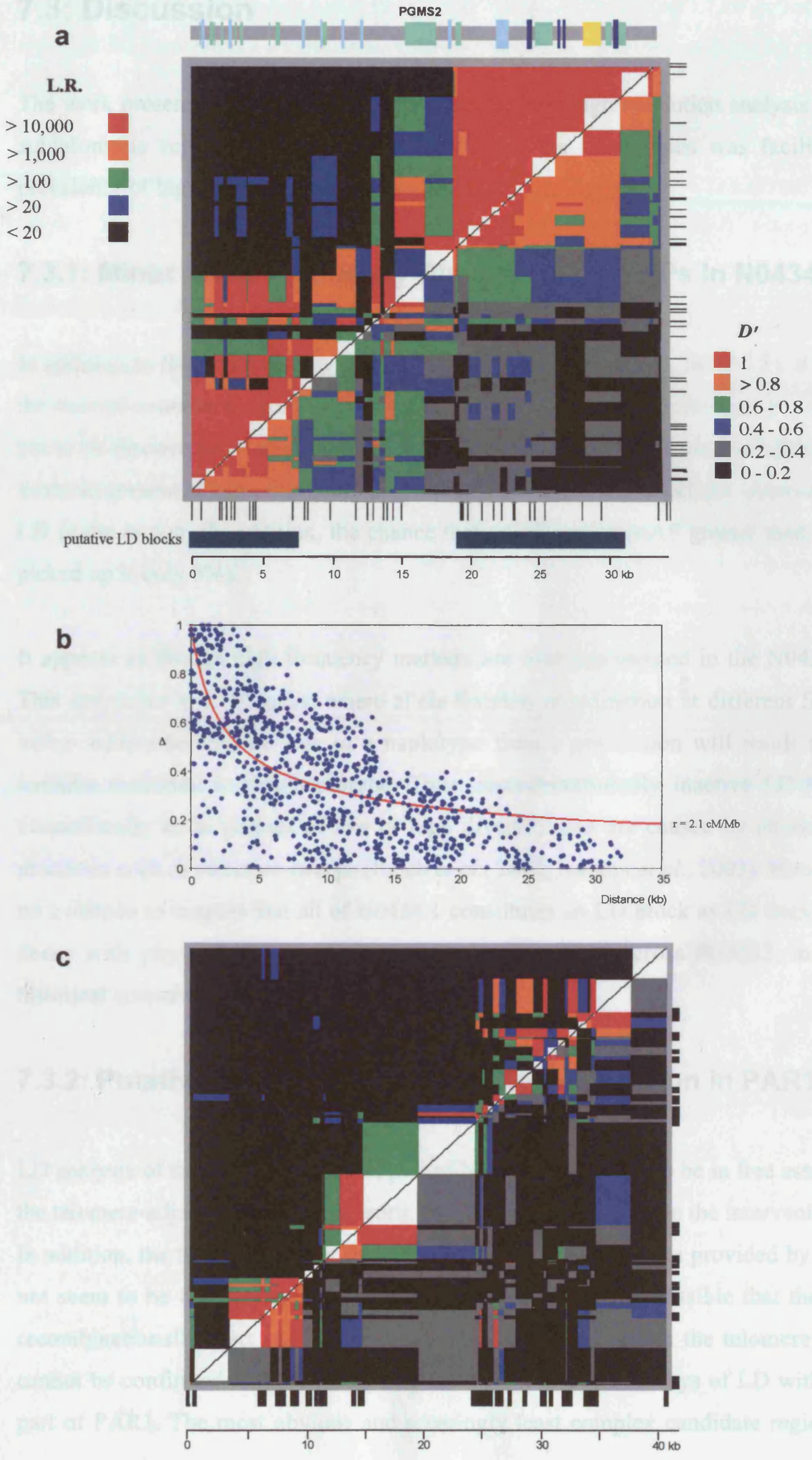


Figure 7.5: Decline of Δ (a measure of absolute LD) with physical distance across (a) the distal 15 kb of the N0434.1 interval and (b) a 10 kb interval in the SHOX region. The red lines in each graph are the best-fit values of the sex-averaged recombination rate per unit distance, calculated using the relationship between Δ, effective population size (assumed to be 10,000), recombination rate per unit distance and inter-marker distance (see main text).

Figure 7.6: The patterns of LD across the N0434.1 interval and *SHOX*. As in figure 7.4, $|D'|$ measures of LD between all pairs of markers with mAFs of at least 0.18 are shown in the bottom right triangle of each plot (*a* and *c*) and LR values are top left. Values of $|D'| = 1$ (in red) indicate pairs of SNPs in complete LD and values of zero (in black) indicates SNP pairs in free association. SNP positions are shown below and to the right of the plots. (a) LD across N0434.1. The top part of the diagram shows the relative positions of the strucures within N0434.1 (Chapter 3). *PLCXL1* exons are pale blue, *PGPL* exons are dark blue, tandem repeat sequences are shown in green and the STIR array in yellow. The positions of putative LD blocks are marked below the plot. (b) The decline of Δ with physical distance across the entire N0434.1 interval. The best-fit value of recombination rate across the interval, assuming an effective population size of 10,000, is approximately 2 cM/Mb, i.e. the same rate that was estimated for just the distal part of N0434.1 (figure 7.5*a*). (c) The LD pattern across *SHOX*, the only other region of PAR1 to be surveyed in this way at high resolution (Chapter 1, section 1.3.2.c.i), is very different to that of N0434.1 (see main text).


a
PGMS2
L.R.
> 10,000
> 1,000
> 100
> 20
< 20
D′
1
> 0.8
0.6 - 0.8
0.4 - 0.6
0.2 - 0.4
0 - 0.2
putative LD blocks
0
5
10
15
20
25
30 kb
b
1
0.8
0.6
0.4
0.2
0
Δ
r = 2.1 cM/Mb
0
5
10
15
20
25
30
35
Distance (kb)
c
0
10
20
30
40 kb

7.3: Discussion

The work presented in this chapter constitutes the first high resolution analysis of LD in the subtelomeric region of PAR1. Determination of the LD pattern was facilitated by the prevalence of high frequency SNPs.

7.3.1: Minor Allele Frequency Distribution of SNPs in N0434.1

In addition to the SNPs already identified in N0434.1 (Chapter 6, table 6.1), it is likely that the interval contains a significant number of rare SNPs, i.e. those with a mAF ≤ 0.1, that have yet to be discovered. However, as all SNPs with a mAF ≤ 0.17 were excluded from the LD analyses presented here, these rare polymorphisms would not affect the observed patterns of LD in the region. (In addition, the chance that a SNP with a mAF greater than 0.17 was not picked up is only 5%).

It appears as though high frequency markers are over represented in the N0434.1 interval. This can occur in LD blocks, where allele fixation or extinction at different SNPs will not occur independently; the loss of a haplotype from a population will result in loss of all variants restricted to that haplotype. Thus, recombinationally inactive LD blocks might occassionally be in periods of low or high diversity that are caused by chance or through processes such as selective sweeps (Reich *et al.*, 2002; Kauppi *et al.*, 2003). However, there is no evidence to suggest that all of N0434.1 constitutes an LD block as LD does significantly decay with physical distance in the interval, particularly across PGMS2, indicating that historical crossovers have occurred in this region.

7.3.2: Putative Boundary of Meiotic Recombination in PAR1

LD analysis of the SNPs in the distal part of N0434.1 showed it to be in free association with the telomere-adjacent region, suggesting recombinational activity in the intervening 50–60 kb. In addition, the three markers at the telomere adjacent region (data provided by M. Hills) do not seem to be in complete LD (figure 7.4). Thus, it is quite possible that the supposedly recombinationally inert region of PAR1 extends no further than the telomere, though this cannot be confirmed without continuing the high resolution surveys of LD within the distal part of PAR1. The most obvious and seemingly least complex candidate regions for these studies are the 6.7 kb region between DXYS14 and DXYS20, the 15 kb region between B4 and C3MS2 (see figure 7.4) and a re-evaluation of LD in the telomere-adjacent region itself.

7.3.3: Gradual Decline of LD within the N0434.1 Interval

Within the distal part of the N0434.1 interval, LD (as measured by Δ) decays slowly and reasonably monotonously with physical distance. Using the relationship $\Delta^2 = 1/(1 + 4N_e rd)$, the rate of meiotic recombination within this interval is estimated at 2.1 cM/Mb, which is close to the genome average rate. Extending the analysis of LD across all of N0434.1 suggested that there is no significant deviation from this rate, other than across the PGMS2 minisatellite (below), and indicated that there is no abrupt breakdown of LD within the region, arguing against the existence of localised recombination hotspots observed previously at *SHOX* and in the MHC II region. However, there is evidence suggesting the presence of two LD blocks, of length 7.5 kb and 12 kb. Block structure has not been observed at *SHOX* (May *et al.*, 20020), but has been seen in the MHC II region, where blocks extend for up to 90 kb. It is possible that the whole N0434.1 region is part of an inter-hotspot region, which is potentially very interesting because the close-to genome average rate of recombination in the region, suggested by the gradual decay of LD, would be directly testable by high resolution analysis of crossover in sperm DNA. Therefore, for the first time, it might be feasible to determine how crossovers are distributed outside of a hotspot. This type of investigation is technically impossible in the LD blocks of the MHC II region (figure 1.7) as the extremely low rates of crossover would require the use of impractical amounts of sperm DNA to pick up the very few recombination events that are taking place and would cause serious PCR problems.

Across PGMS2 there is a greater breakdown of LD that is superimposed on the pattern of putative LD blocks and gradual LD decline seen throughout N0434.1. Therefore, if the N0434.1 interval does contain a hotspot of meiotic recombination, the LD pattern indicates that it is likely to be in, or close to, the moderately variable PGMS2 minisatellite. This is noteworthy as hotspot-driven minisatellite instability was observed at MS32, adjacent to the first human meiotic recombination hotspot to be defined at the molecular level (Jeffreys *et al.*, 1998*a*) (Chapter 1, section 1.2.5.a).

7.4: Concluding Remarks

The genotyping and LD analysis of SNPs within the N0434.1 interval has indicated that the region is in free association with the telomere-adjacent region. This suggests that the putative boundary of PAR1 meiotic recombination is located in the intervening minisatellite-rich region, which is 50–60 kb long. Characterisation of this interval (Chapter 4) has revealed two regions that could be targeted for the further SNP identification and LD analyses that will be required if the supposed boundary is to be identified.

Prior to analysing LD within N0434.1, one might have expected to see a pattern very similar to that observed at *SHOX* (figure 7.6*b*), where the very rapid decay of LD with physical distance is consistent with the evidence suggesting that PAR1 is a (male-specific) recombination hot domain (Rouyer *et al.*, 1986*a*; Page *et al.*, 1987; Rappold, 1993; Lien *et al.*, 2000). However, within N0434.1, association declines gradually with physical distance, at a rate that appears to be consistent with randomly distributed crossovers occurring at a rate not dissimilar to the genome average rate of 1 cM/Mb. This represents a kind of "halfway house" between the prodigious rate of LD decay at *SHOX* and the much more gradual decline of LD within the MHC II haplotype blocks.

If meiotic recombination hotspots occur in the N0434.1 interval, the greater breakdown of LD across PGMS2 provides the most likely location. Crossover analysis of single sperm DNA molecules will determine whether hotspotting as seen at MHC II and *SHOX* is also a feature of recombination around the *PGPL* and *PLCXL1* genes.

Chapter 8

Sperm Crossover Analysis in the N0434.1 Interval

8.1: Introduction

The analysis of SNP genotypes in the N0434.1 interval did provide some evidence of LD block structure, which does not occur in the *SHOX* region, but there was no indication of an abrupt breakdown of LD, as has been observed previously within the MHC II region. However, LD did appear to decay relatively quickly with physical distance across the PGMS2 minisatellite, which might suggest the existence of a localised recombination hotspot (figure 7.6). As LD patterns can only be used to indirectly infer the distribution of crossing-over events (Wall and Pritchard, 2003), a picture of what is truly happening can only be gained through the recovery of crossover molecules directly from sperm DNA. Thus far, LD and sperm crossover analyses have been successfully applied in the identification of recombination hotspots in the MHC II and *SHOX* regions (Jeffreys *et al.*, 2001; May *et al.*, 2002; also see Chapter 1, sections 1.2.5.b and 1.3.2.c.i). The earlier observation of hotspot-driven minisatellite instability at MS32 (Jeffreys *et al.*, 1998*a*) suggested that the most interesting N0434.1 region (in terms of crossover) would be around the interval of greater breakdown of LD across the PGMS2 minisatellite. This chapter describes the assays that were used to detect crossover in this region.

8.1.1: Crossover Detection Strategy

An overview of the sperm crossover analysis strategy has already been provided in Chapter 1 (figure 1.5). The first and most obvious step is to use LD patterns to target a region for sperm crossover analysis. A man is then identified with multiple heterozygous SNPs across the target region and two rounds of allele-specific repulsion-phase PCR are used to selectively amplify recombinant molecules from multiple aliquots of sperm DNA. However, this technique is far from trivial, and there are a number of factors that must be accounted for, and steps that must be taken, before the actual experiment can take place.

8.2: Results

8.2.1: Attempts to Detect Crossovers in a 10 kb PGMS2 Region

8.2.1.a: Defining Primary PCR Interval and Selection of Suitable Donor

Defining the interval for the first (primary) round of repulsion-phase allele-specific PCR, and identifying the individual in which crossover analysis is to take place, must consider both the optimal size for primary PCR (about 7 kb) and inclusion of the maximum number of informative (heterozygous) SNPs. As it is essential that the sites used for both primary and secondary PCR are heterozygous, examination of the genotypes of men in the semen donor panel (Chapter 7, figure 7.2) very quickly rendered most individuals unsuitable for this particular crossover assay. Consideration of primary PCR amplicon size was also affected by the allele size variability of PGMS2 (Chapter 6, figure 6.2). As a result, a 9.3–9.8 kb interval within donor 5, containing 16 informative SNPs, was considered the most suitable for sperm crossover analysis (figure 8.1*a*).

8.2.1.b: Allele-Specific Primer Optimisations

In order to design a highly efficient and very specific system for the detection of crossover molecules, allele-specific primers (AS-primers) were designed in the forward sense for 12237A/G, 12657A/G, 13351C/T and in the reverse sense for 22168C/T, 22145A/G, 22030A/G, 21309C/G and 20723C/G. The optimal annealing temperature of each AS-primer was determined (table 2.8; figure 8.1*a*) using individuals that had been shown by genotyping to be homozygous for one allele or the other (figure 7.2), and that had PGMS2 allele sizes similar those of donor 5. The identification of a suitable crossover molecule detection system was complicated by the number of selector sites that are in Alu sequences (figure 8.1*a*), which impeded the efficiency and specificity of the AS-primers, presumably because they were also binding to other Alu sequences in the region, and elsewhere in the genome. Occasionally, it was possible to reduce this problem by the addition of a synthetic CCCC 5′ extension to the primer or by modifying the primer with a synthetic 20 nt 5′ tail (see table 2.8) and using both the AS-primer and a primer identical to the 5′ tail to drive amplification (Jeffreys *et al.*, 1991). However, it remained impossible to amplify either efficiently or specifically from a number of SNP sites, despite trying up to six designs of certain AS-primers. Nevertheless, it appeared as though a workable system would involve a primary PCR selecting for 12237/A or G and 22145/A, and a secondary PCR that would detect crossovers taking place between 12657/A or G and 21309/G (figure 8.1*b*).

Figure 8.1: The optimisation of AS-primer annealing temperatures and the determination of donor 5 linkage phase within the crossover analysis interval. (a) The crossover analysis interval within donor 5. Informative SNPs are marked by green circles and homozygous SNPs are marked with grey circles, *PLCXL1* exons are denoted by the blue rectangles and positions of Alu elements are shown by red rectangles. AS-primers were designed for all SNPs that are marked with a red arrow and the positions of the universal primers used in the annealing temperature optimisation experiments are shown by the black arrows. (b) Southern blots showing the results of the AS-primer annealing temperature optimisations. The interval defined by an AS-primer and the relevant universal primer (XN13.5F or XN20.6R) was PCR amplified at annealing temperatures 56, 59, 62 and 65°C (shown left-to-right for each batch of four bands in the figure) from the genomic DNA of men known to be homozygous for one SNP allele or the other. As a positive control, and under the same conditions, the universal primers were also used in a PCR amplification. Thus, the top left blot shows that 12237/FA (and XN20.6R) can amplify across PGMS2 at temperatures of 56 and 59°C with reasonable efficiency, but at the lowest temperature there is a loss of specificity as the interval can also be amplified from a 12237 G homozygote. On the other hand, the results indicate that the efficiency and specificity of 12237/FG is excellent at all annealing temperatures between 56 and 65°C, though there is a slight loss of specificity at 56°C, shown by the positive signal after amplification at this temperature in a 12237 A homozygote. In the positive controls, the efficiency of amplification from both the 12237 A and 12237 G homozygote decreases with increasing temperature. (c) The AS-PCR and dotblotting strategy used to establish linkage phase between 12237, 12657, 21309 and 22145 in donor 5. (*i*) Allele-specific and universal primers were used to PCR amplify the intervals as shown (1, 2 and 3) from donor 5. Each interval was then dotblotted and ASO-hybridised for the internal SNP. Universal primers (black arrows) were used to PCR amplify individuals homozygous for one or the other internal SNP allele, and these smaller intervals were also dotblotted and ASO-hybridised for use as positive controls. (*ii*) Linkage phase of donor 5 between the selector sites for crossover. Homozygous SNPs, and those between 21309 and 22145, are not shown. This determined that the primary PCR would involve 12237/FA and 22145/RA/Ad and that the secondary PCR would use 12657/FG and c21309/RG.

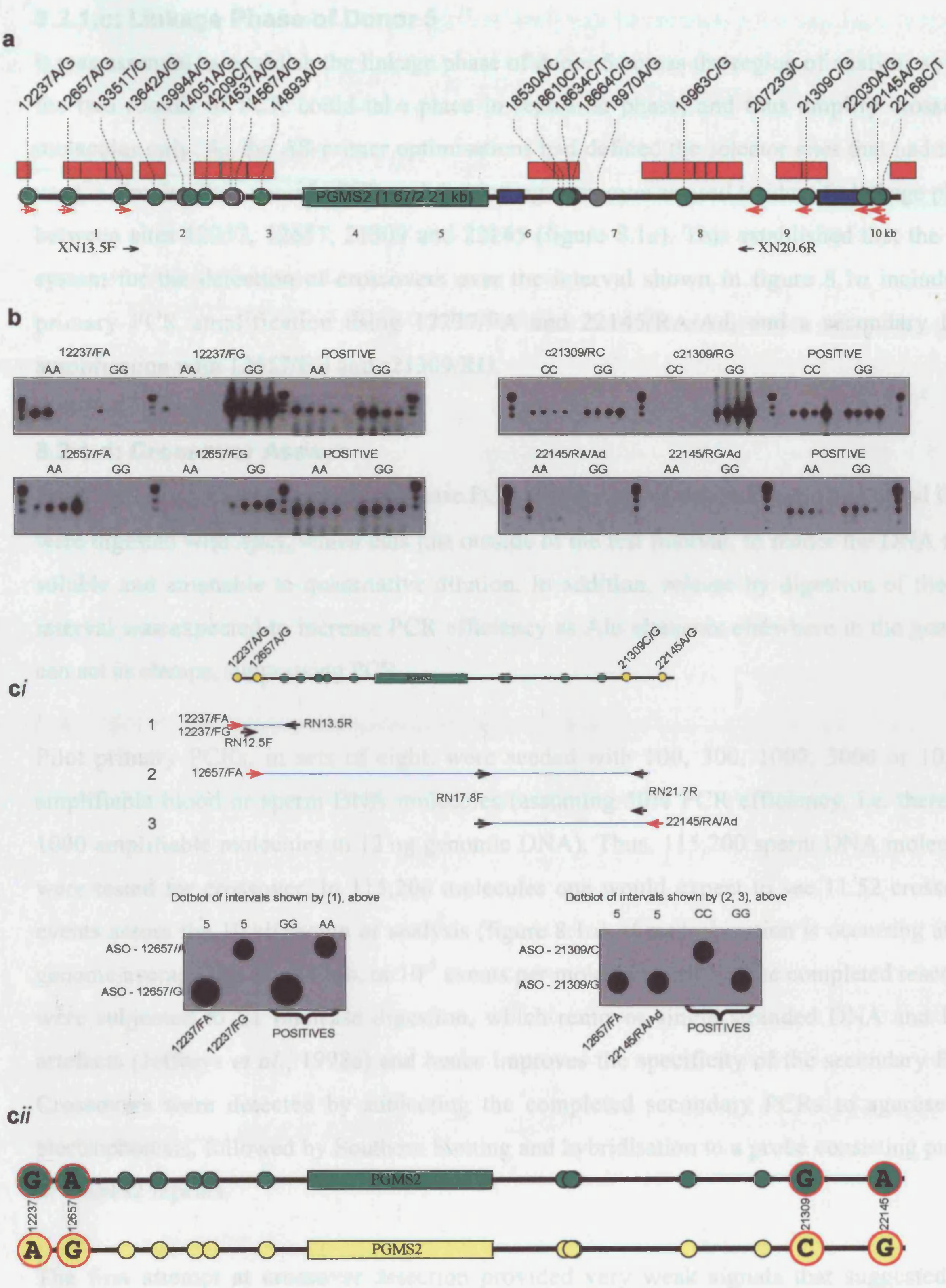


Figure 8.1: Optimisation of AS-primer annealing temperature and determination of donor 5 linkage phase within the crossover analysis interval.

8.2.1.c: Linkage Phase of Donor 5

It was essential to establish the linkage phase of donor 5 across the region of analysis so that the two rounds of PCR could take place in repulsion phase, and thus amplify crossover molecules only. As the AS-primer optimisations had defined the selector sites that had to be used, a simple allele-specific PCR and dotblotting strategy was used to identify linkage phase between sites 12237, 12657, 21309 and 22145 (figure 8.1*c*). This established that the best system for the detection of crossovers over the interval shown in figure 8.1*a* included a primary PCR amplification using 12237/FA and 22145/RA/Ad, and a secondary PCR amplification with 12657/FG and c21309/RG.

8.2.1.d: Crossover Assays

Prior to the two rounds of repulsion-phase PCR, 20 μg each of donor 5 sperm and blood DNA were digested with *Apa*I, which cuts just outside of the test interval, to render the DNA fully soluble and amenable to quantitative dilution. In addition, release by digestion of the test interval was expected to increase PCR efficiency as Alu elements elsewhere in the genome can act as clamps, suppressing PCR.

Pilot primary PCRs, in sets of eight, were seeded with 100, 300, 1000, 3000 or 10,000 amplifiable blood or sperm DNA molecules (assuming 50% PCR efficiency, i.e. there are 1000 amplifiable molecules in 12 ng genomic DNA). Thus, 115,200 sperm DNA molecules were tested for crossover. In 115,200 molecules one would expect to see 11.52 crossover events across the 10 kb region of analysis (figure 8.1*a*), if recombination is occurring at the genome average rate (1 cM/Mb, or 10^{-5} events per molecule per kb). The completed reactions were subjected to S1 nuclease digestion, which removes single stranded DNA and PCR artefacts (Jeffreys *et al.*, 1998a) and hence improves the specificity of the secondary PCR. Crossovers were detected by subjecting the completed secondary PCRs to agarose gel electrophoresis, followed by Southern blotting and hybridisation to a probe consisting purely of PGMS2 repeats.

The first attempt at crossover detection provided very weak signals that suggested the presence of only four molecules in which crossover might have taken place (data not shown). The experiment described above was repeated twice but it was not possible to detect any more crossover events. This suggested that either the AS-PCR system was not working, or that the rate of crossover in the region was very low. The AS-PCR system was assayed by doping 1000 and 3000 molecule inputs of donor 5 blood DNA with a series of positive control DNA inputs (i.e. 20, 10, 5, 2, 1 0.5 and 0.2 molecules of blood DNA from a donor in recombinant phase relative to donor 5 across the region of analysis). In parallel, 1000 and 3000 molecule inputs of donor 5 blood and sperm DNA that had not been doped were also subjected to AS-PCR amplification. This assay not only provided signals in the positive controls, demonstrating that the AS-PCR system worked, but also identified a number of putative crossover molecules in sperm DNA (figure 8.2*b*). Unfortunately, there was also significant bleed-through of a signal from blood DNA. The quality of the assay was clearly erratic, which strongly suggested that it needed fine tuning before it could successfully be applied to crossover detection. A number of the conditions were altered in turn, including: (i) a reduction in the number of input molecules to decrease the bleed-through of signal from blood DNA; (ii) alterations to the number of primary and secondary PCR cycles in order to take advantage of the expectation that secondary PCR would be relatively more efficient (figure 8.1*b*); (iii) tertiary PCR amplifications with universal primers in an attempt to boost the overall efficiency of amplification (figure 8.2*a*); (iv) alterations to $MgCl_2$ concentration; (v) the use of different *Taq* polymerases; (vi) the design and optimisation of AS-primers selecting for SNPs at 13642 and 20723. Occasionally it appeared as though an efficient and specific system of crossover detection had been identified (e.g. figure 8.2*c*) but it was impossible to develop a system that remained robust. It seemed as though the highly repetitive nature of the region, coupled with the length of the primary PCR interval, ensured that a PGMS2 crossover assay would remain frustratingly inconsistent at best, and impossible at worst.

As a system to successfully assay crossovers within the region defined by SNPs at 12657 and 21309 (figure 8.1*a*) was not identified (despite the many attempts alluded to above), none of the conditions employed in these assays have been fully described in this thesis. In addition, the strategies used to determine the optimal annealing temperatures of AS-primers used in subsequent crossover analyses (below) remained the same as those outlined above, and are not described again.

8.2.2: Crossover Analysis within a Truncated PGMS2 Region

Although a workable crossover assay appeared to be unattainable, there was clear evidence to suggest that crossovers were occurring within the region of analysis shown in figure 8.1*a*. However, the reliability of crossover detection was so inconsistent that it was impossible to

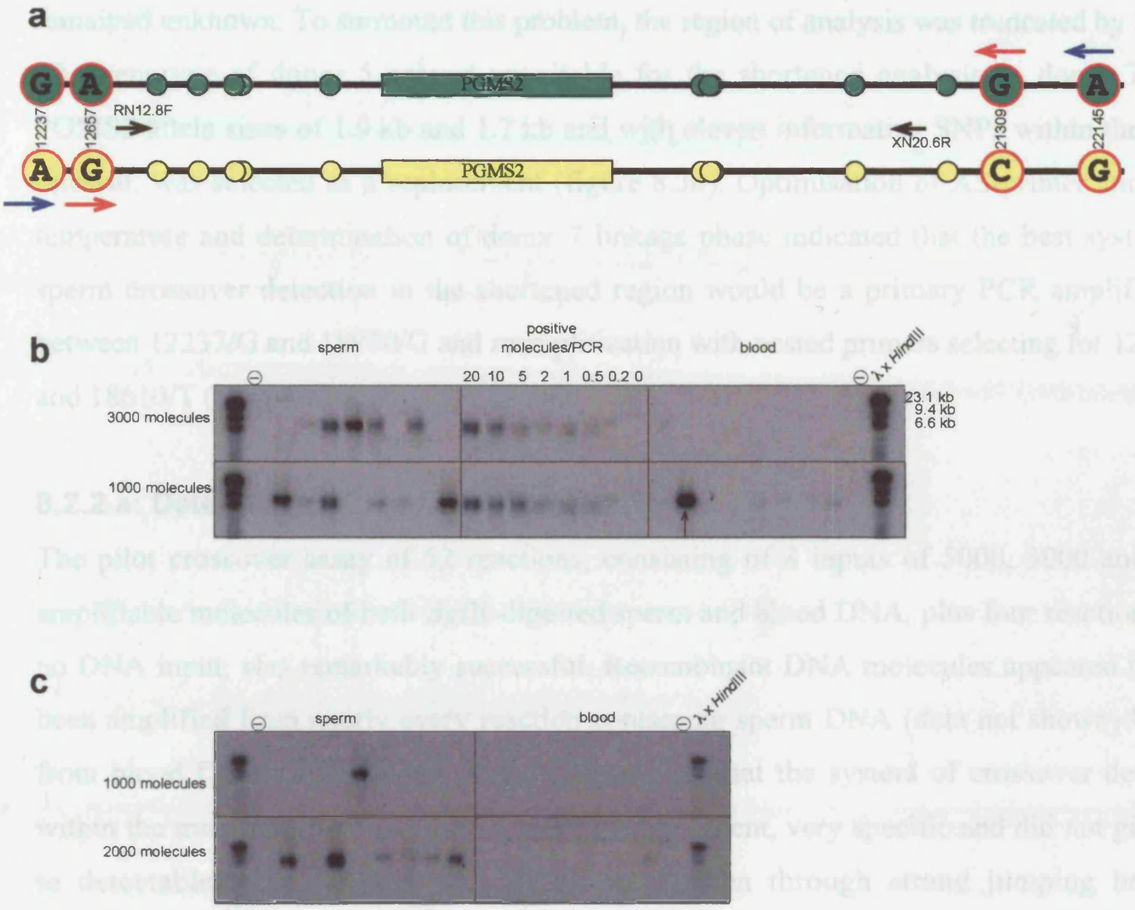


Figure 8.2: Crossover assays across the PGMS2 minisatellite in donor 5. (a) Linkage phase of donor 5. Positions of the primary PCR primers selecting for 12237/A and 22145/A are shown by the blue arrows and the secondary PCR primers are marked by the red arrows. Universal primers that were occasionally used to boost the overall efficiency of crossover assay attempts are marked by the black arrows. (b) Assays to determine the efficiency of the crossover detection system. DNA-free negative controls are marked with minus signs. The intense signals in sperm DNA for both 1000 and 3000 molecule inputs are presumably derived from genuine recombinant molecules, while the signals of decreasing intensity in the positive controls are due to decreasing inputs of positive control DNA (blood genomic DNA from a donor in recombinant phase relative to donor 5). The signal in one of the 3000 molecule blood DNA inputs (marked by the vertical arrow) could be a result of inconsistent specificity of the AS-primers, or be derived from PCR artefacts that arise by strand jumping between haplotypes. (c) Very occasionally, fine tuning of the crossover assay resulted in intense signals in sperm DNA, and little or no evidence of signals in blood DNA, indicating that an efficient and specific system of crossover detection had been achieved.

identify enough crossover molecules as to make a mapping strategy worthwhile. Therefore, the rate of recombination and the pattern of crossover distribution over the region of PGMS2 remained unknown. To surmount this problem, the region of analysis was truncated by 3.2 kb. The genotype of donor 5 proved unsuitable for the shortened analysis so donor 7, with PGMS2 allele sizes of 1.9 kb and 1.7 kb and with eleven informative SNPs within the assay interval, was selected as a replacement (figure 8.3*a*). Optimisation of AS-primer annealing temperature and determination of donor 7 linkage phase indicated that the best system for sperm crossover detection in the shortened region would be a primary PCR amplification between 12237/G and 18970/G and reamplification with nested primers selecting for 12657/A and 18610/T (figure 8.3*b*) (Chapter 2, table 2.12).

8.2.2.a: Detection of Crossover Molecules

The pilot crossover assay of 52 reactions, consisting of 8 inputs of 5000, 3000 and 1000 amplifiable molecules of both *Bgl*II-digested sperm and blood DNA, plus four reactions with no DNA input, was remarkably successful. Recombinant DNA molecules appeared to have been amplified from nearly every reaction containing sperm DNA (data not shown) but not from blood DNA (figure 8.3*c*). This demonstrated that the system of crossover detection within the truncated PGMS2 interval was highly efficient, very specific and did not give rise to detectable PCR artefacts that might have arisen through strand jumping between haplotypes. Moreover, it strongly suggested that the rate of crossover in the region was much higher than 13 cM/Mb, as had been suggested by the LD data (Chapter 7, section 7.2.2.c). Repeating the pilot assays, with reduced numbers of molecules, showed that the number of positive signals did not drop significantly until sperm DNA input was reduced to about 250 amplifiable molecules (3 ng genomic DNA) (figure 8.3*c*).

8.2.2.b: Mapping Crossover Break Points

Maximum likelihood analysis (software written by A.J. Jeffreys in TrueBASIC 4.1) applied to the results of the pilot assays, in which a total of 87,200 amplifiable sperm DNA molecules were analysed for crossover, suggested that the rate of exchange between 12657/A and 18610/T was approximately 5.5 x 10^{-3} crossovers per molecule (data not shown). Consequently, some PCRs, particularly those containing more than 200 molecules, are likely to contain more than one crossover event, which would impede mapping of crossover breakpoints. Therefore, 120 aliquots of digested sperm DNA, each expected to contain 0.4, 0.5 or 0.66 crossover molecules (respectively 0.88 ng, 1.1 ng or 1.44 ng DNA inputs) were subjected to the two rounds of repulsion-phase AS-PCR shown in figure 8.3*b*. Positive

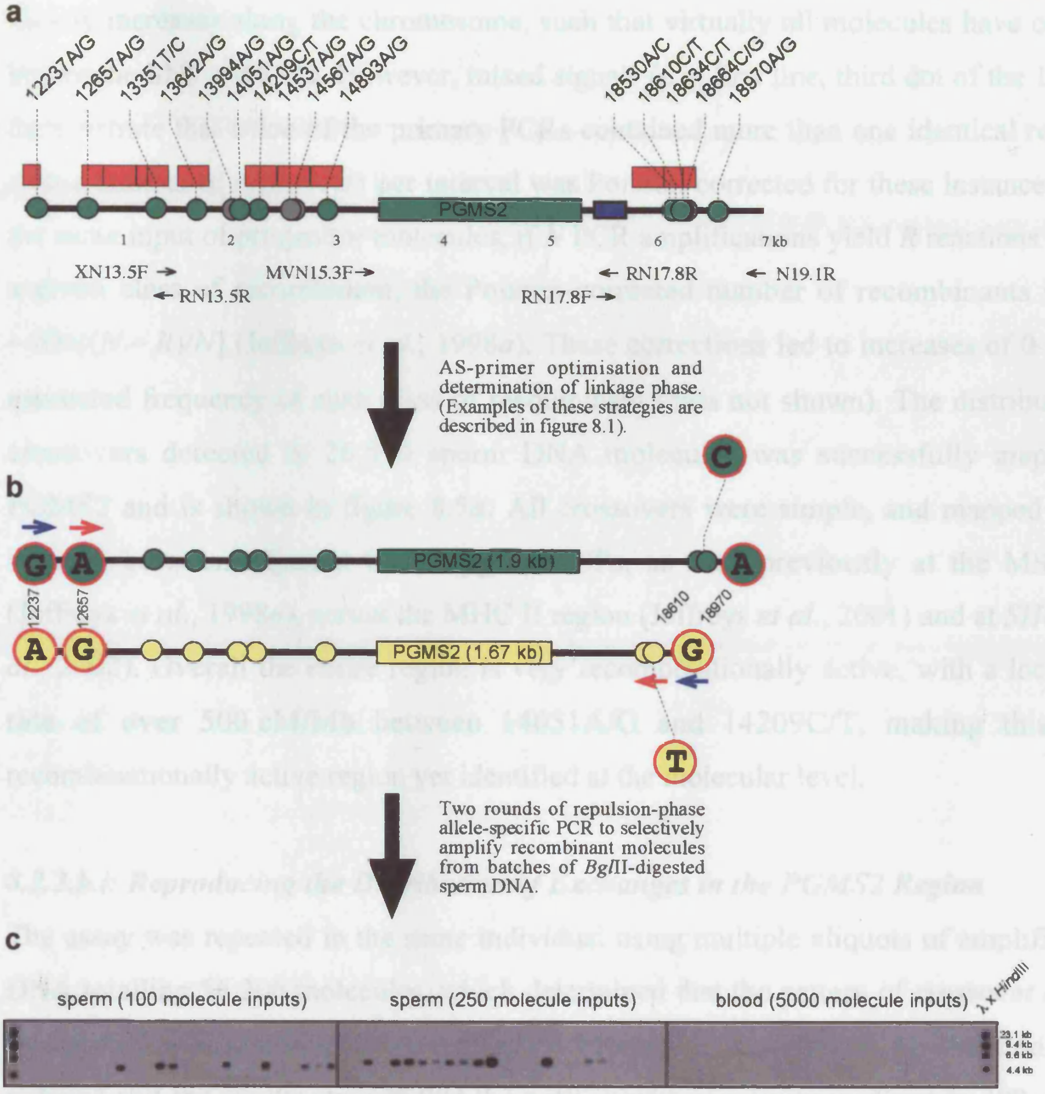

a
12237A/G
12657A/G
13351T/C
13842A/G
13994A/G
14051A/G
14209C/T
14531A/G
14567A/G
14893A/G
18530A/C
18610C/T
18634C/T
18684C/G
18970A/G
PGMS2
1
2
3
4
5
6
7 kb
XN13.5F
RN13.5R
MVN15.3F
RN17.8F
RN17.8R
N19.1R
AS-primer optimisation and determination of linkage phase. (Examples of these strategies are described in figure 8.1).
b
C
G
A
PGMS2 (1.9 kb)
A
12237
12657
18610
18970
A
G
PGMS2 (1.67 kb)
G
T
Two rounds of repulsion-phase allele-specific PCR to selectively amplify recombinant molecules from batches of BglII-digested sperm DNA.
c
sperm (100 molecule inputs)
sperm (250 molecule inputs)
blood (5000 molecule inputs)
λ x HindIII
23.1 kb
9.4 kb
6.6 kb
4.4 kb

Figure 8.3: High resolution sperm crossover analysis across the PGMS2 region. (a) The truncated crossover analysis interval within donor 7. Informative SNPs are marked by green circles and homozygous SNPs are marked with grey circles, *PLCXL1* exon 8 is denoted by the blue rectangle and Alu elements are shown by red rectangles. The positions of universal primers used in AS-primer optimisations (Chapter 2, table 2.9) and the determination of donor 7 linkage phase (Chapter 2, table 2.11) between 12237A/G, 12657A/G, 18610C/T, 18634C/T, 18970A/G and PGMS2 allele sizes are marked with black arrows. Note that these strategies were similar to those used in ascertainment of donor 5 linkage phase and AS-primer optimisations (figure 8.1). (b) Linkage phase of donor 7. The positions of first round AS-primers are marked with blue arrows and those used in the second round are denoted by the red arrows. (c) Examples of crossover detection in *Bgl*II-digested sperm and blood DNA (*Bgl*II cuts much closer to this truncated test interval than *Apa*I, which was used to digest donor 5 DNA prior to earlier attempts to detect crossovers). Secondary PCR products were detected by agarose gel electrophoresis and Southern blot hybridisation using a probe consisting entirely of PGMS2 repeats. The signals in sperm DNA are derived from genuine recombinant molecules. No signals were observed in blood DNA, even when PCRs were seeded with 5000 amplifiable molecules.

secondary PCRs were reamplified using universal PCR primers, (figure 8.4*a*) and the status of internal SNP sites was determined by ASO typing (figure 8.4*b*). The number of molecules that had crossed from the 12657/A (green) haplotype to the 18610/T (yellow) haplotype clearly increases along the chromosome, such that virtually all molecules have crossed over before the SNP at 18530. However, mixed signals (e.g. first line, third dot of the 13351 blots) demonstrate that some of the primary PCRs contained more than one identical recombinant, so the number of crossovers per interval was Poisson-corrected for these instances. Thus, for the same input of progenitor molecules, if N PCR amplifications yield R reactions positive for a given class of recombinant, the Poisson-corrected number of recombinants is given by $-N\ln[(N-R)/N]$ (Jeffreys *et al.*, 1998*a*). These corrections led to increases of 0–62% in the estimated frequency of each class of recombinant (data not shown). The distribution of 109 crossovers detected in 26,520 sperm DNA molecules was successfully mapped across PGMS2 and is shown in figure 8.5*a*. All crossovers were simple, and mapped to a single interval between adjacent heterozygous SNPs, as seen previously at the MS32 hotspot (Jeffreys *et al.*, 1998*a*), across the MHC II region (Jeffreys *et al.*, 2001) and at *SHOX* (May *et al.*, 2002). Overall the entire region is very recombinationally active, with a localised peak rate of over 500 cM/Mb between 14051A/G and 14209C/T, making this the most recombinationally active region yet identified at the molecular level.

8.2.2.b.i: Reproducing the Distribution of Exchanges in the PGMS2 Region

The assay was repeated in the same individual using multiple aliquots of amplifiable sperm DNA totalling 56,200 molecules, which determined that the pattern of crossover distribution was remarkably reproducible (figure 8.5*b*). Moreover, the intervals of 95% confidence are reduced and the results suggest that the peak crossover activity is closer to 300 cM/Mb, but that this activity extends over a slightly wider interval (13642A/G to 14209C/T) than previously suggested by the first assay (figure 8.5*a*).

8.2.2.b.ii: Identification of PGMS2 Allele Sizes After Crossover

Virtually all recombinant molecules cross from the green to yellow haplotype before reaching the SNP at 18530 (figure 8.4*b*). However, it was not known whether these crossovers occurred on the distal or the proximal side of PGMS2, or if they were taking place within the minisatellite itself. Therefore, as a preliminary investigation, the PGMS2 array was amplified from a subset of the positive secondary PCRs used to map the crossover events (above) and visualised by agarose gel electrophoresis with ethidium bromide staining (figure 8.6). Remarkably, nearly all of the PGMS2 alleles in recombinant molecules were either 1.9 kb or 1.7 kb, the respective lengths of the green and yellow haplotypes (figure 8.3*b*).

Figure 8.4: Mapping of crossover breakpoints in the PGMS2 region. (a) Recombinant molecules amplified by the two rounds of AS-PCR (figure 8.3*b*) are bounded by alleles 12657/A (of the "green" parental haplotype) and 18610/T ("yellow" parental haplotype). Green and yellow haplotypes were revealed by AS-PCR and ASO-typing (Chapter 2, table 2.13). Alleles at internal SNP sites (lilac circles) were identified by tertiary PCRs between primer pairs XOT13.0F & N15.1R; RN13.8F & N15.1R and N18.4F & XOT18.5R, dotblotting and ASO typing (Chapter 2, table 2.14), which thus mapped the point of crossover from the green to yellow haplotype. As most internal SNPs are in Alu sequences, tertiary PCRs were divided into three amplicons to avoid hybridisation of certain ASOs to the wrong Alu elements. (b) Examples of dotblots used to map crossover breakpoints. The last six spots on the bottom line of each blot were loaded with four tertiary PCRs from blood DNA and two reactions with no DNA input. Clearly, there has been no bleed through of either parental haplotype. Dotblots have been arranged top-to-bottom in order of the SNPs nearest to 12657 to those nearest 18610, and those dotblots detecting alleles of the green haplotype have been placed on the left. In a 12657 to 18610 direction, the dotblots show that the number of yellow haplotype alleles clearly increases such that the crossover breakpoint of virtually all the recombinant molecules detected in this assay occurs before the SNP at 18530.

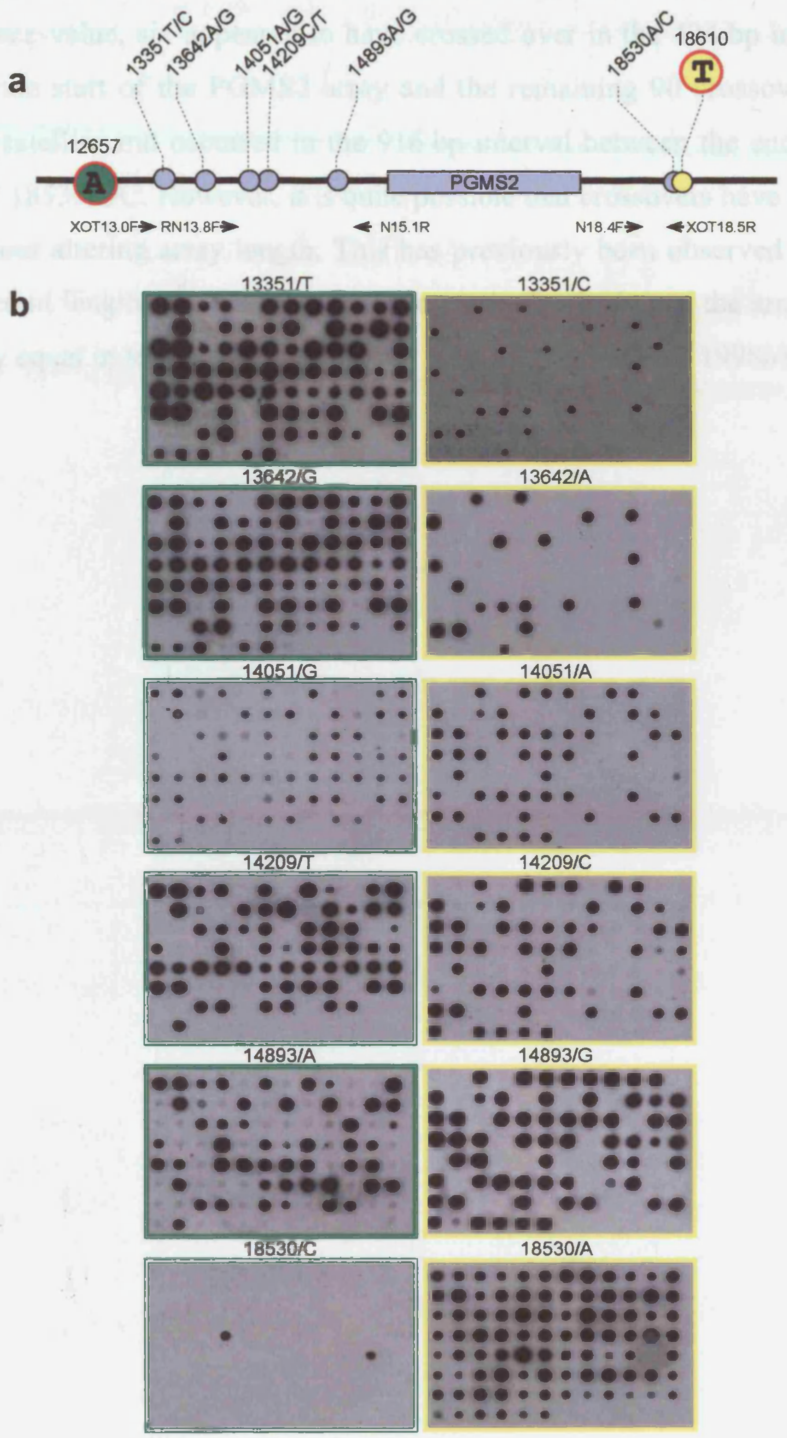


Figure 8.4: Mapping of crossover breakpoints in the PGMS2 region.

Of the 96 crossovers that were originally mapped to the interval containing PGMS2 (figure 8.5*b*), only two appeared to have an allele that differed in size from the parental haplotypes. At face-value, six appeared to have crossed over in the 424 bp interval between 14893A/G and the start of the PGMS2 array and the remaining 90 crossovers have also avoided the minisatellite and occurred in the 916 bp interval between the end of the PGMS2 array and SNP 18530A/C. However, it is quite possible that crossovers have taken place within PGMS2 without altering array length. This has previously been observed at MS32, where alleles of different lengths can pair in-register over the beginning of the array, yielding a recombinant array equal in length to the progenitor allele (Jeffreys *et al.*, 1998*b*).

Figure 8.5: Observed distribution of sperm recombinants across the PGMS2 test interval. Heterozygous markers are indicated above each plot. The observed numbers of crossovers mapping to each interval, shown above the histograms, were used to estimate the recombinational efficiency of each interval in cM/Mb. The error bars show the 95% confidence intervals. (a) Distribution of 109 crossovers amplified from an initial input of 26,520 donor 7 sperm DNA molecules (40 x 120 molecules; 40 x 91 molecules; 40 x 73 molecules; and the pilot assays of 8 x 1000 molecules; 8 x 500 molecules; 8 x 300 molecules and 8 x 100 molecules). (b) Distribution of 263 crossovers detected in 56,200 amplifiable sperm DNA molecules (144 x 250 molecules; 144 x 125 molecules; 40 x 55 molecules).

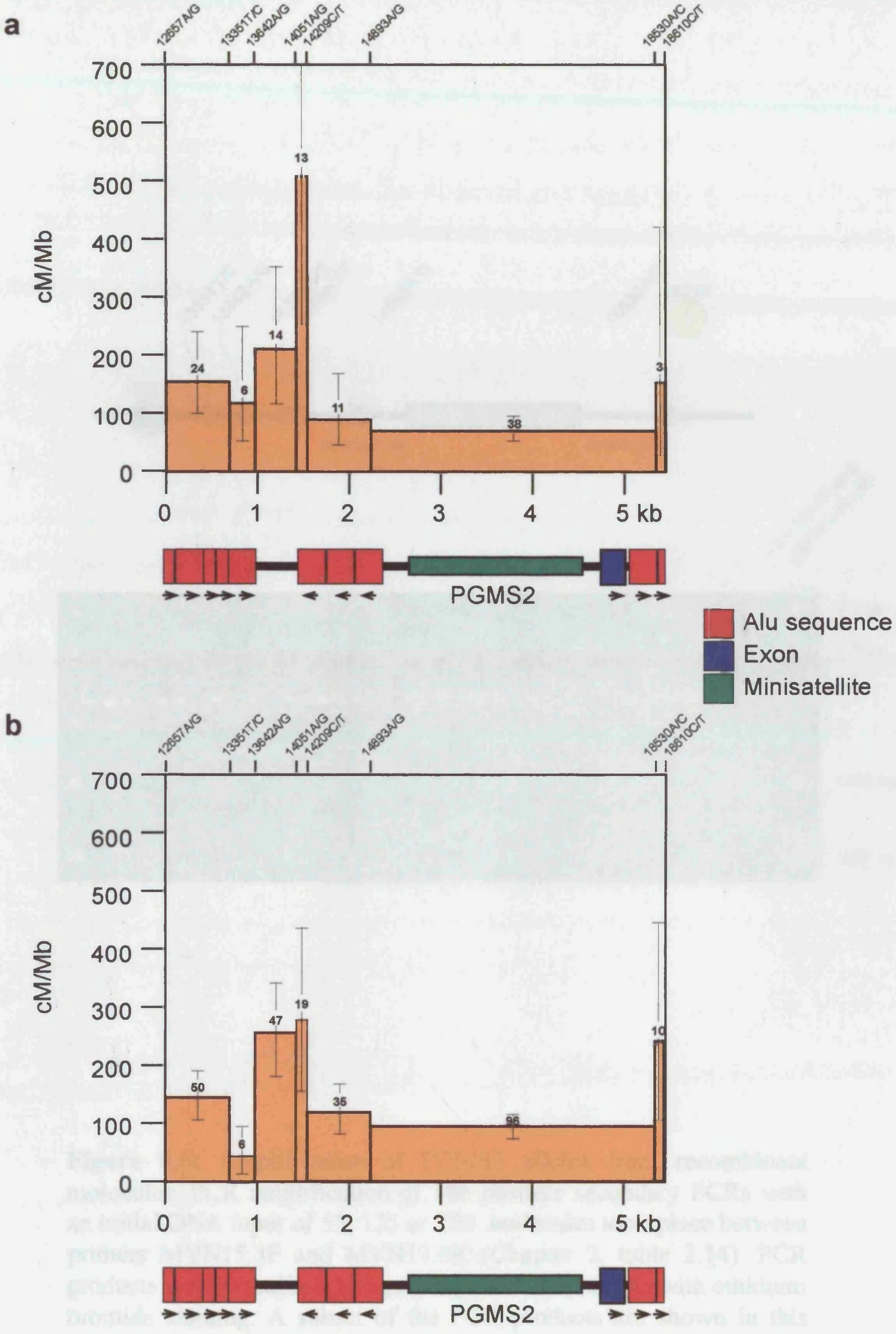


Figure 8.5: Observed distribution of sperm recombinants across the PGMS2 test interval

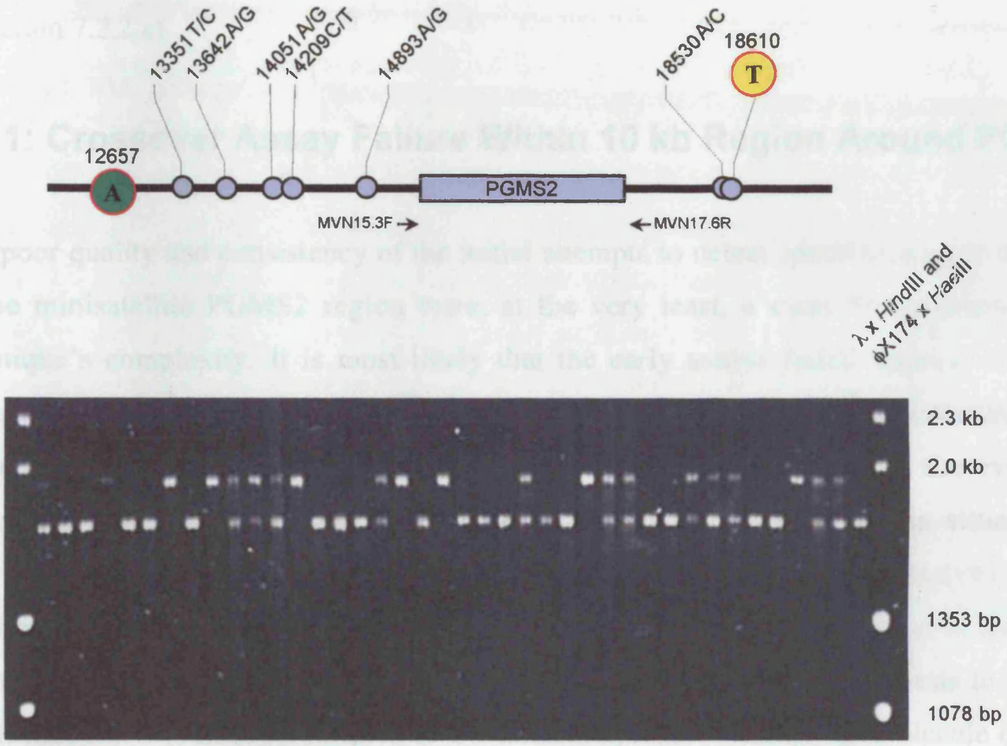


Figure 8.6: Amplification of PGMS2 alleles from recombinant molecules. PCR amplification of 186 positive secondary PCRs with an initial DNA input of 55, 125 or 250 molecules took place between primers MVN15.3F and MVN17.6R (Chapter 2, table 2.14). PCR products were visualised by agarose gel electrophoresis with ethidium bromide staining. A subset of the PCR products are shown in this diagram.

8.3: Discussion

This chapter has presented the successful use of sperm crossover analysis to map the distribution of sperm crossover events in the PGMS2 region. However, the profile that emerged is unlike anything previously observed and, moreover, is widely different from the population-based value of recombination rate, which was estimated from LD decay (Chapter 7, section 7.2.2.c).

8.3.1: Crossover Assay Failure Within 10 kb Region Around PGMS2

The poor quality and consistency of the initial attempts to detect sperm crossover molecules in the minisatellite PGMS2 region were, at the very least, a clear demonstration of the technique's complexity. It is most likely that the early assays failed because of the low efficiency of the AS-primers used for the first round of PCR (figure 8.1*b*). Unfortunately, the linkage phase of the donor most suitable for this analysis, coupled with the even worse efficiency and lack of specificity of other AS-primers, meant that this situation was unavoidable. It is worth noting that this was the first time sperm crossover analysis had been attempted in such a large and Alu-dense region and, of course, continuing to design and optimise other AS-primers for the selector sites, or persisting with adjustments to the assay itself, may have eventually resulted in a workable system of crossover molecule detection. However, the number of attempts that were made suggests that this technique is not applicable within relatively long regions that are rich in repeat sequences.

8.3.2: Successful Detection of Crossovers

8.3.2.a: Distribution of Crossovers

The excellent efficiency and specificity of the assay used in the truncated PGMS2 interval allowed the recovery of recombinant molecules from several batches of sperm DNA. Even at relatively high inputs, no recombinant molecules were recovered from parallel analyses of blood DNA, indicating that those derived from sperm were genuine meiotic crossover molecules, and not PCR artefacts (figure 8.3*c*). In addition, the fact that no crossovers were picked up in nearly 100,000 molecules of blood DNA demonstrates that the rate of mitotic recombination across the PGMS2 region must be very low (and certainly less than 0.1 cM/Mb). Initial quantification and mapping of 109 crossover breakpoints was impaired by the relatively wide 95% confidence intervals around the calculated rates of recombination (figure 8.5*a*). However, the distribution pattern proved highly reproducible, as was shown by mapping a further 263 crossovers, amplified from 56,200 sperm molecules (figure 8.5*b*). Indeed, both Fisher's exact test ($P = 0.25$) and the Kolmogorov-Smirnov test ($P = 0.912$) showed that there was no significant difference between the two observed crossover distributions.

8.3.2.a.i: Rate of Recombination in the PGMS2 Region

The recombination frequency over the 6.2 kb region corresponds to a genetic length of 0.6 cM, equal to about 100 cM/Mb. This is over 110 times the mean autosomal crossover rate at male meiosis, over five times higher than the average rate in PAR1 and very different from the sex-averaged population estimate of about 13 cM/Mb derived from LD decay (Chapter 7, section 7.2.2.c). A Chi-squared test showed that exchanges were not randomly distributed across the region (data not shown). It appears as though there is a 90–95 cM/Mb uniform crossover rate but an excess of exchanges between 13642A/G and 14209C/T, where activity is 260–280 cM/Mb. There is therefore evidence for a very localised hotspot of meiotic recombination superimposed on a very high background rate. However, only 15% of the PGMS2 crossovers cluster into the 566 bp region of highest activity, compared to 81% and 95% of crossovers that respectively localise to the 1–2 kb hotspots in the *TAP2* and *SHOX* regions. Thus, it is possible that crossovers within the PGMS2 region are more-or-less randomly distributed and that the reproducible narrow region of enhanced activity is a site of preferred crossover resolution.

8.3.2.a.ii: Crossover Activity in the PGMS2 Minisatellite

Within the MS32 minisatellite, equal crossovers occur at a frequency that is about 40% that of the unequal crossover rate (Jeffreys *et al.*, 1998*b*). Therefore, if crossovers are taking place within the PGMS2 array, one would expect a significant proportion of them to associate with changes in array length. However, virtually all PGMS2 alleles in recombinant molecules are equal in length to the progenitor alleles, which suggests that crossovers are actually avoiding the PGMS2 minisatellite. This would be unprecedented as the three other minisatellite loci analysed so far show significant levels of exchange in both the flanking DNA and within the repeat array, leading to unequal and equal crossovers between alleles (Jeffreys *et al.*, 1998*a*,*b*; Buard *et al.*, 2000; M. Panayi, unpublished data). Furthermore, the points of exchange at one of these minisatellites define the MS32 hotspot, which extends into the polar end of the repeat array (Jeffreys *et al.*, 1998*a*; Chapter 1, section 1.2.5.a) and there is provisional evidence for a very similar hotspot upstream of MS31 (M. Panayi, unpublished data). Furthermore, if crossovers do not occur within PGMS2, the relative rate of recombination in the flanking regions would be very high indeed (figure 8.5). This is noteworthy because both the PGMS2 flanking regions and the 566 bp region of peak activity are the only extended intervals of single copy DNA within the region of analysis (figure 8.5). However, the possibility that a significant number of equal exchanges occur within PGMS2 cannot be discounted. An MVR-PCR strategy (Chapter 6, figure 6.3.1.a), which would identify the internal structure of PGMS2 alleles in crossover molecules as compared to progenitor alleles, would determine whether exchanges were going into the repeat array without altering array length, or if they were avoiding the minisatellite.

8.3.2.a.iii: PGMS2 Region is Highly Recombinationally Active

The region around PGMS2 is the most recombinationally active region yet identified at the molecular level but there are few clues within the sequence as to why this is so. For example, there are no GT microsatellites in the region of PGMS2, which were found to have a positive correlation with regions of high recombination along chromosome 22 (Majewski and Ott, 2000) and nor was there any evidence for a palindromic sequence akin to that implicated in the elevation of recombination rate in a 12 kb region of the human adenosine deaminase gene (Cruciani *et al.*, 2003). However, the PGMS2 region is 52% GC-rich and, as noted previously (Chapter 1, section 1.2.5.c), there is a significant positive correlation between recombination rate and GC content in both *S. cerevisiae* (Gerton *et al.*, 2000) and humans (Yu *et al.*, 2001). Furthermore, it was noted recently that regions with a high CpG fraction (but low GC and poly(A/T) content) tend to have the highest recombination rates (Kong *et al.*, 2002). Within the PGMS2 crossover analysis interval CpG dinucleotides account for 15% of the sequence, which is notable as they are normally greatly under-represented in human DNA, occurring at a frequency of only 0.8% (International Human Genome Sequencing Consortium, 2001).

There is a particularly high concentration of Alu elements within the PGMS2 region, accounting for 43% of the sequence, as compared to 33% of the N0434.1 interval as a whole (Chapter 3, section 3.3.2) and just 10% of the human genome (Smit, 1996; International Human Genome Sequencing Consortium, 2001). Whether this is significant or not is not known, but Alu elements are known to facilitate unequal homologous recombination events (Batzer and Deininger, 1999) and two of the meiotic recombination hotspots within the MHC II region, including the *DNA3* hotspot, which is the most active hotspot so far identified in MHC II, are centred on Alu elements. However, as already noted (above) the localised interval of peak activity within the PGMS2 region is actually centred on an interval of single copy DNA.

It is intriguing to wonder what conclusions might have been drawn had the results of the first crossover assay been reproducible. Only four putative recombinant molecules were amplified from 115,200 sperm molecules (section 8.2.1.d), suggesting a rate of crossover in the region that is both below genome average and well beneath the rate that was eventually determined. If the four putative crossovers had been subjected to a third round of amplification, it is possible that ASO-typing might have revealed the presence of more than one type of recombinant in each case, but it remains likely that the rate of crossover in the PGMS2 region would have been severely underestimated. Alternatively, the second assay may have revealed a localised, but extended, gene conversion hotspot contained within the interval defined by the selector sites of the first assay. Consequently, the first assay would have detected only a few crossovers and the number of events detected by the truncated (second) assay would have been inflated because one of the selector sites was within the interval of gene conversion (see Chapter 9, figure 9.1). However, the data indicate that the gene conversion tracts in the PGMS2 region would have to be over 4 kb, which is difficult to reconcile with observations at *SHOX* and in the MHC II region, where tracts are only 300 bp long (A.J. Jeffreys and C.A. May, manuscript in preparation).

8.3.2.a.iv: PCR Efficiency and Rate of Crossover

For any PCR amplification from which a rate of crossover or mutation has been invoked there are two variables that should be taken into account – the accuracy of quantification and the actual PCR efficiency. Therefore, in order to confirm the rate of crossover as calculated across the PGMS2 region, it would be necessary to determine the exact number of amplifiable molecules that were screened. For a SP-PCR experiment this would be resolved by the determination of single molecule PCR efficiency. A number of PCR reactions (60–90), each containing a known amount (equivalent to about one molecule) of the DNAs analysed by SP-PCR would be amplified as in the SP-PCR reaction, and the PCR products would be detected by agarose gel electrophoresis and Southern blot hybridisation. Poisson analysis, applied to the number of reactions in which a PCR product is observed, would then provide a figure for the average number of amplifiable molecules per known amount of DNA. From this, the single molecule PCR efficiency per 6 pg DNA (assumed to be equivalent to one diploid genome) can be calculated (e.g. Jeffreys *et al.*, 1994).

Correcting for single molecule PCR efficiency by Poisson analysis of single molecule dilutions is rather more difficult when faced with a crossover assay because the primers are designed to amplify only recombinant molecules. Hence, the vast majority of molecules are unamplifiable by definition. Therefore, if a Poisson correction is to take place, the variables of the actual reaction have to be changed, i.e. at least one primer must be changed in order to have a system that, in an ideal world, would amplify all of the DNA molecules present. However, this primer would be in a slightly different location, would have a different nucleotide composition and almost certainly have an efficiency different to that of the AS-primer. The alternative is to assume that the DNA quantification (section 8.2.1.d, above) is correct and then to use an estimate of PCR efficiency. A large number of Poisson analyses over intervals of 5–10 kb have shown that there is approximately one amplifiable molecule per 12 pg of DNA, giving a single molecule PCR efficiency of 50%, and for intervals of less than 5 kb the PCR efficiency rises to approximately 80% (A.J. Jeffreys, personal communication). Therefore, given this precedent, it was thought reasonable to assume that the crossover assay, which took place over an interval of 6.2 kb, would also have a PCR efficiency of 50%.

However, the most interesting result gained through the work presented in this chapter is not the rate of crossover but how the crossovers are distributed. Importantly, conclusions with regard to the distribution of crossovers are not affected by the efficiency of the PCR reaction.

8.4: Concluding Remarks

Clearly the highly reproducible distribution of crossover events within the PGMS2 region is not one that fits the classic hotspot profile, but nor does it fit a pattern of random distribution. The evidence in favour of a meiotic recombination hotspot may have been strengthened had it been possible to develop a workable crossover assay over the longer initial interval of choice. If a hotspot of meiotic recombination has been revealed, it is much narrower than all previously identified hotspots, which are 1–2 kb, and is flanked by DNA that is also very recombinationally active. As studies of hotspots outside of PAR1 have indicated that they are shared between males and females (Jeffreys *et al.*, 2001), it could be that PAR1 recombination hotspots are also shared but that male-specific crossover occurs at very high rates in interhotspot regions. This might explain why low-resolution linkage analyses have consistently indicated that the female contribution to crossover proficiency in PAR1 is negligible (e.g. Rouyer *et al.*, 1986*a*; Page *et al.*, 1987), but is not totally compatible with the observations in the *SHOX* region, where the rate of recombination outside of the hotspot drops to approximately 6% of the peak activity (C.A. May, personal communication).

As the data are also consistent with a relatively uniformly hot region in which crossovers are either avoiding PGMS2 or occurring within the array without altering its length, further analysis of the minisatellite, through an MVR-PCR strategy, is required before it can be confirmed whether or not crossover is taking place within the repeat array. Fortunately, the repeat unit variability of PGMS2 is amenable to MVR-PCR (Chapter 6, figure 6.1), so this would not present too great a challenge.

Finally, it would be useful to repeat the crossover analysis in at least one other man to determine whether or not the distribution pattern is reproducible between individuals. In addition, it would have been useful to perform a reciprocal crossover analysis (i.e. from the yellow to green haplotype in figure 8.3*b*) to see if crossing over in the PGMS2 region was fully reciprocal, as has been seen at all but one of the MHC II hotspots (Jeffreys and Neumann, 2002) and at *SHOX* (May *et al.*, 2002). This should be feasible.

Chapter 9
Discussion

9.1: Introduction

Sequencing of the N0434.1 cosmid (Chapter 3) provided the primary tool for the variety of studies that have been presented in this thesis, the main objective of which has been to carry out a high resolution analysis of DNA diversity and recombination in the *PGPL* gene region. SNPs identified in the N0434.1 interval (Chapter 6) were genotyped in 50 unrelated UK semen donors of north European descent and used in a LD analysis to screen for evidence of historical recombination (Chapter 7). The LD pattern then led to the identification of a target region, which encompassed the novel and moderately variable PGMS2 minisatellite, for the recovery of sperm crossover molecules and mapping of exchange points (Chapter 8). In addition, the long-awaited production of PAR1 sequence data by the Human Genome Sequencing Consortium allowed analysis of the PAR1 subtelomeric sequence to be extended over an interval of nearly 100 kb (Chapter 4). This enabled the full establishment of the entire genomic structures of both the *PGPL* gene and *PLCXL1*, a novel gene suggested to be related to the phosphatidylinositol-specific phospholipase C (PI-PLC) gene family (Chapter 5).

9.2: *PGPL* and *PLCXL1*

Although PAR1 had been neglected by the Human Genome Sequencing Consortium until very recently (June 2003), a relatively small amount of genomic sequence around the *SHOX* gene, located approximately 500 kb from the PAR1 telomere (Rao *et al.*, 1997), had allowed markers within a 43 kb interval to be subjected to a high resolution analysis of recombination. This revealed an extremely rapid decay of LD with physical distance, as expected for a region very active in (male) recombination, but also showed that crossovers clustered into a highly localised hotspot about 2 kb wide, flanked by recombinationally much less active DNA (May *et al.*, 2002). To see if this pattern held true for other regions in PAR1, the *PGPL* gene was selected for further study. *PGPL* had been mapped to the interval contained within the N0434 cosmid and the cDNA sequence of the gene had been published (Gianfrancesco *et al.*, 1998). Sequencing of N0434, which was thought to be approximately 80–110 kb from the PAR1 telomere (Rao *et al.*, 1997), revealed that most of the 5′ end of *PGPL* lies proximal to the

N0434 interval and also led to the identification of *PLCXL1*, a novel gene that appears to be the most telomeric in PAR1.

PGPL, which appears to encode a GTP-binding protein (Gianfrancesco *et al.*, 1998), consists of ten exons between 67 bp and 440 bp in length that are spread over about 14.3 kb. Similarly, there are nine exons, 66–761 bp long, in the *PLCXL1* gene, which extends over approximately 26 kb. However, the genomic size of the each gene is likely to vary between individual X and Y chromosomes due to the presence of several intronic tandem repeat arrays. Specifically, both PGMS1 and the pseudoautosomal STIR array, which vary in size from 0.8–1.2 kb and 1.0–5.5 kb respectively (Chapter 6, figures 6.2*a*; 6.4), are located within *PGPL*. Intron 7 of *PLCXL1* contains PGMS2, which has been shown to vary in size from 1.4–7.5 kb (Chapter 6, figure 6.2*b*). These repeats must therefore be transcribed.

The genomic structure of *PLCXL1* provided two other interesting features. First, the PGMS3 minisatellite overlapped the 3′ end of *PLCXL1* exon 5, suggesting that the minisatellite was not only transcribed but also, because each minisatellite repeat unit provided at least one additional potential splice donor site, that it could extend the length of exon 5 by up to 240 bp. Second, the 3′ UTR of *PLCXL1* contains an Alu element, which potentially has a functional significance as UTRs are very important in the post-transcriptional regulation of gene expression (van der Velden and Thomas, 1999; Conne *et al.*, 2000; Jansen, 2001). Both of these features have been described in other genes (see Chapter 5, sections 5.3.2.c and 5.3.2.d), and are therefore not unusual, but *PLCXL1* may provide a novel system for their further analysis.

It was initially thought that *PLCXL1* was a member of the phosphatidylinositol-specific phospholipase C (PI-PLC) gene family. PI-PLCs play a key role in initiating receptor-mediated cellular signal transduction, have been isolated from a wide variety of organisms and contain at least five recognisable domains, named X and Y (which combine to form the catalytic core of PI-PLCs), PH, EF-hands and C2 (Chapter 5, section 5.1.2.a). However, only a putative X domain was identified within *PLCXL1*, which indicated that *PLCXL1* could be part of a so far uncharacterised gene family that is rather distantly related to PI-PLCs. In addition, the observation of a series of putative *PLCXL1* homologues in a variety of organisms suggested that *PLCXL1* is likely to have an important biological function.

Finally, the close physical proximity of *PGPL* and *PLCXL1*, coupled with the previous reports of interactions between GTP-binding proteins and PI-PLCs (Taylor *et al.*, 1991), suggests that an investigation of a potential interaction between the *PGPL* and *PLCXL1* products is warranted.

9.3: Tandem Repeats

Sequencing of the N0434 cosmid and analysis of sequences produced at the Sanger Centre confirmed previous reports that had suggested the distal 100 kb of PAR1 was largely composed of unrelated, variable, GC-rich tandem repeats (e.g. Cooke *et al.*, 1985; Page *et al.*, 1987; Vergnaud *et al.*, 1993), a feature that has also been noted at the distal regions of many human autosomes (Royle *et al.*, 1988; Flint *et al.*, 1997*a,b*). Indeed, tandem repeats account for 38% of the DNA within the PAR1 subtelomeric region and, for the first time, it was possible to accurately define the physical distances between the previously reported distal PAR1 structures such as *PGPL*, DXYS14, DXYS20 and the CEB12 and CEB30 repeats (Chapter 4, figure 4.9).

Three of the novel tandem repeats within the N0434.1 interval, PGMS1, PGMS2 and the pseudoautosomal STIR array, were selected for investigation. Each had only a limited number of allele sizes and was moderately unstable, as indirect population genetic estimates of their mutation rates were well below the values of 1% or higher observed at classic unstable minisatellite loci. Thus, it appeared as though the N0434.1 interval did not contain a highly unstable minisatellite that might have served as a marker for a local recombination hotspot, as had been shown at MS32 (Jeffreys *et al.*, 1998*a*). However, this was not an indication that the N0434.1 interval did not contain a region of localised intense recombination activity, as all of other hotspots identified so far are not associated with tandemly repeated DNA (Jeffreys *et al.*, 2000, 2001; May *et al.*, 2002).

The length variability of a STIR array had not been investigated prior to the study presented in Chapter 6. Essentially STIR arrays are GC-rich minisatellites with unusually large repeat units, and it is possible that the variability is caused by the same kind of recombinational processes that destabilise more typical minisatellites (Jeffreys *et al.*, 1991; Armour *et al.*, 1993; Neil and Jeffreys, 1993). The fact that STIRs have been observed in most orders of mammals (W. Schempp and B. Weber, cited in Petit *et al.*, 1990) suggests that they have some kind of biological function. Rouyer and colleagues (1990) thought that STIR arrays were likely to have a role in the recognition and initiation of pairing between homologous chromosomes prior to meiotic recombination. As STIR arrays within pseudoautosomal regions appear to be a more conserved and complex class of element than the autosomal STIRs (Chapter 3, section 3.1.1.c), I have suggested that a STIR-mediated homologue-recognition role may be specific to PAR1 and that pseudoautosomal STIRs, as well as highly variable GC-rich minisatellites, may serve as occasional markers of recombination hotspots (Chapter 6, section 6.3.2).

There are a number of variable minisatellites (e.g. Simmler *et al.*, 1985, 1987; Rouyer *et al.*, 1986*a*,*b*; Armour *et al.*, 1990; Klink *et al.*, 1993) and at least four other pseudoautosomal STIR arrays throughout PAR1 (Rouyer *et al.*, 1990), which must have implications for the length variability of PAR1 as a whole. The much-needed PAR1 sequence data that has only very recently become available will allow the systems that must exist to limit the variability of PAR1 length to be investigated more completely.

9.4: High Resolution Analysis of Recombination

The two-tier strategy for the high resolution analysis of recombination (Chapter 1, figure1.5) in the *PGPL* region began with the identification of diploid genotypes of 55 N0434.1 SNPs in a panel of 50 unrelated UK semen donors of north European descent (Chapter 7, figure 7.2). The subsequent LD analysis did provide some evidence for LD blocks, of at least 7.5 kb and 12 kb, in the region but, overall, LD decayed gradually with physical distance at a rate that seemed consistent with randomly distributed crossovers occurring at close to the genome average autosomal rate of recombination.

LD analysis had suggested that the rate of recombination across the PGMS2 region increased to about 13 cM/Mb (Chapter 7, section 7.2.2.c), but sperm crossover analysis established that the true rate was closer to 100 cM/Mb, five times higher than the PAR1 average rate of recombination and over 110 times faster than the mean autosomal rate at male meiosis. These observations cannot simultaneously be true and thus create an interesting paradox: The relationship between the LD measure Δ, and recombination frequency r, ($\Delta^2 = 1/(1 + 4N_e r)$, Chapter 7, section 7.1.1), means that in order to match the recombination rate derived from LD measures with the much higher rate established by sperm crossover analysis, the effective population size, N_e, must be small. If true, the PGMS2 region must therefore have had a very recent common ancestor. If this were true, one would expect to see a low level of nucleotide diversity, but the PGMS2 region has a genome-average level of nucleotide diversity (Chapter 6, section 6.2.3), suggesting that the effective population size cannot be small. Linkage analyses have consistently indicated that the female contribution to recombination in PAR1 is negligible (Rouyer *et al.*, 1986*a*) but if the PGMS2 region is recombinationally active in females, then the overall rate of recombination in the region would be even higher. Therefore, one would need to invoke a greater decrease in population size in order to match the recombination rates derived from the LD and sperm crossover analyses.

The remarkably reproducible distribution of crossover events within the PGMS2 region does not match any of the patterns that have been observed previously either at *SHOX*, or in the MHC II region (Jeffreys *et al.*, 2000; 2001; May *et al.*, 2002). The crossover profile suggested that there might be a very narrow localised hotspot of meiotic recombination, of width 500 bp or less, where activity rises to at least 300 cM/Mb, superimposed on a uniform rate of 100 cM/Mb. However, all previously identified hotspots have a standard width of 1–2 kb, in which recombination initiation sites, gap expansion (resection), repair and resolution by conversion or crossover (see DSB repair model, Chapter 1, figure 1.3) all appear to be contained (Jeffreys and Neumann, 2002; A.J. Jeffreys and C.A. May, manuscript in preparation). Furthermore, all previously identified hotspots are surrounded by recombinationally much less active DNA. The data that I have presented on crossover distribution in the PGMS2 region (Chapter 8, figure 8.5) only defines the sites of crossover resolution. Therefore it is possible that the narrow region of apparently enhanced recombination in the PGMS2 region is actually a preferred site for termination of events that have initiated elsewhere.

Analysis of previously identified recombination hotspots has shown that all gene conversion tracts are short (average length of 300 bp or less) and contained within the hotspots (Jeffreys and Neumann, 2002; A.J. Jeffreys and C.A. May, manuscript in preparation) Therefore, if the PGMS2 region had exhibited a clear LD block structure and if the distribution of crossovers in the region had matched the "classic" hotspot profile, then precedence would have suggested that all of the molecules detected by the crossover assay (Chapter 8, section 8.2.2) were genuine crossovers and not conversions. However, the unusual LD pattern and crossover distribution seen in the PGMS2 region requires the serious consideration of the possibility that lengthy gene conversion tracts are being scored as crossovers. Therefore, an essential next step in the analysis of recombination in the PGMS2 region will be to carry out a gene conversion assay.

A gene conversion assay is essentially half a crossover assay, in which AS-primers directed to two sites from one haplotype at one end of a target region and universal primers at the other end of the target region are used to PCR amplify small pools of sperm DNA (figure 9.1). Intervening SNP sites are then typed by ASO hybridisation. There are three possible outcomes to a conversion assay in the PGMS2 region (figure 9.1*a*,*b*,*c*). First, conversion tracts might be very short, as has been observed before (Jeffreys and Neumann, 2002; A.J. Jeffreys and C.A. May, manuscript in preparation), and thereby not only indicate that the crossovers were genuine but also provide information on sites of initiation. This would help to resolve whether the narrow interval of enhanced recombination in the PGMS2 region was genuinely a narrow hotspot or if it was a preferred site for resolution of crossover events. Second, conversion tracts in the PGMS2 region might be relatively long, suggesting that some or all of the molecules that had been scored as crossovers were actually gene conversion events that had swept across 5′or 3′ selector sites. Third, if the assay fails to detect any conversions this would indicate either that all crossovers in the region are genuine and are not accompanied by gene conversion, or that there are no crossovers in the region, but that conversion tracts are very long. If it is assumed that the interval of gene conversion is long, such that it includes the reverse selector sites of the truncated crossover assay (Chapter 8, section 8.2.2), but is contained within the interval defined by the selector sites of the first crossover assay (Chapter 8, section 8.2.1), this could provide an alternative explanation for the poor results of the first assay. Thus, the very few events detected by the first assay could have been genuine crossovers, but most, if not all, of the events scored as crossovers by the second assay may actually have been gene conversions (figure 9.1).

The flanking DNA at the polar end of the MS32 minisatellite contains a meiotic recombination hotspot that extends into the repeat array itself (Chapter 1, figure 1.6) (Jeffreys *et al.*, 1998*a*) and there is additional evidence for a similar hotspot at the MS31 minisatellite (M. Panayi, unpublished data). Within the MS32 minisatellite, equal crossovers that do not alter array length occur at a frequency that is about 40% that of the unequal crossover rate (Jeffreys *et al.*, 1998*b*). Therefore, if crossovers are taking place within the PGMS2 array, one would expect a significant proportion of them to associate with changes in array length. However, as changes in array length are not seen (Chapter 8, section 8.2.2.b.ii), it appears as though crossovers are either avoiding the minisatellite or that PGMS2 alleles only undergo equal crossover. If crossovers are avoiding PGMS2, this would mean that the rate of recombination in the regions of flanking DNA would be much higher than 100 cM/Mb (Chapter 8, figure 8.5). Thus, the only significant intervals of single copy DNA in the PGMS2 region would also be the intervals of intense, localised recombination. However, like the region of enhanced activity that has been identified, the flanking DNA intervals are very narrow, and therefore do not fit the model in which initiation, repair and termination all occur

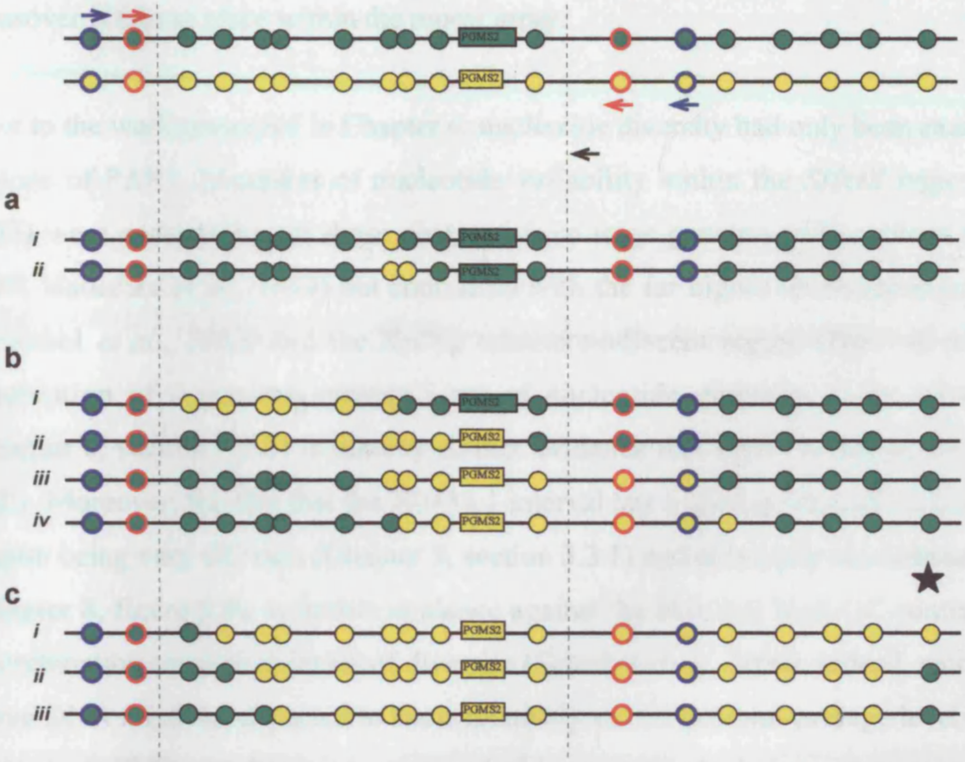


Figure 9.1: Possible results of a conversion assay in the PGMS2 region. The diagram has been simplified such that the distance between markers is not to scale and there has been no representation of conversion (or crossover) occurring within the repeat array. A representation of the crossover assay (Chapter 8, figure 8.3*b*), which used two rounds of AS-PCR (red and blue arrows), is shown in the top part of the diagram. Thus, exchange point mapping occurred in the interval bounded by the vertical dashed lines. The rest of the diagram shows the possible results of a conversion assay following PCR between the two selector sites on the left hand side and the universal primer marked by the black arrow. Selector sites are highlighted by red or blue outlines. (a) Very short conversion tracts, as observed previously (Jeffreys and Neumann, 2002; Jeffreys and May, manuscript in preparation), would provide information on recombination initiation sites and would indicate that a conversion model to explain the (assumed) crossovers is unlikely. (b) Observation of long, multiple-site conversions (*i* and *ii*) would suggest that molecules previously scored as crossovers may in fact be conversions that have swept across selector sites (e.g. *iii* and *iv*). (c) If the assay detects no conversions, either all molecules scored in the sperm crossover assay are genuine, and never accompanied by conversion (*i*), or crossovers do not actually occur in the PGMS2 region and, instead, conversion tracts are very long indeed (*ii* and *iii*). Interestingly, if conversion tracts are very long, this might explain why the first crossover assay (Chapter 8) detected so few crossover molecules and the truncated assay appeared to detect so many. Taking the SNP position marked by the star as the reverse selector site in the first crossover assay, only *i* would have been amplified and scored as a crossover, but *i*, *ii* and *iii* would all have been scored as crossovers by the truncated assay.

in a "standard" 1–2 kb hotspot. A conversion assay (above and figure 9.1) would help to clarify this issue, particularly with respect to sites of initiation, but further analysis of the minisatellite, through an MVR-PCR strategy, will be required to confirm whether or not crossover is taking place within the repeat array.

Prior to the work presented in Chapter 6, nucleotide diversity had only been examined in three regions of PAR1. Measures of nucleotide variability within the *SHOX* region (May *et al.*, 2002) were consistent with those obtained from large genome-wide surveys (Cargill *et al.*, 1999; Halushka *et al.*, 1999) but contrasted with the far higher levels reported at *PPP2R3B* (Schiebel *et al.*, 2000) and the Xp/Yp telomere-adjacent region (Baird *et al.*, 1995). The observation of a genome-average level of nucleotide diversity in the N0434.1 interval (Chapter 6, section 6.2.3) is thereby further evidence that PAR1 is not globally enriched in SNPs. Moreover, the fact that the N0434.1 interval has a normal level of nucleotide diversity, despite being very GC-rich (Chapter 3, section 3.3.1) and extremely recombinationally active (Chapter 8, figure 8.5), is further evidence against the idea that high GC content and rates of recombination cause high levels of diversity (Schiebel *et al.*, 2000). Indeed, recombinationally active DNA might be expected to show relatively stable, genome-average levels of nucleotide diversity as SNPs are able to escape correlated extinctions by being constantly reshuffled onto different haplotypes (Kauppi *et al.*, 2003). In contrast, recombinationally inactive LD blocks will undergo correlated fixation or extinction of SNPs that lead to periodic intervals of high and low nucleotide diversity. This is supported by observations in the *PPP2R3B* region where, despite the very high level of nucleotide diversity (Schiebel *et al.*, 2000), there is preliminary evidence for limited haplotype diversity; i.e. markers appear to be in strong LD (A. Webb, unpublished data).

9.4.1: Evidence Against Ectopic Recombination

Alu elements represent approximately 10% of the human genome and preferentially locate to GC-rich regions (Smit, 1996, 1999; International Human Genome Sequencing Consortium, 2001). The distal 100 kb of PAR1 is part of the most GC-rich fraction of the human genome, so it is not surprising that Alu elements are over-represented in this region. The average Alu density across the distal 100 kb of PAR1 is 23%, but rises to 33% in the N0434.1 interval, where the analysis of sperm crossover took place. It is clear that many crossovers are resolved within Alu elements in the PGMS2 region (Chapter 8, figure 8.5). Therefore Alu elements are not a barrier to resection and branch migration during recombination. Consequently, Alu elements will constitute a significant proportion of single-stranded DNA during recombination and may thereby be able to promote ectopic recombination events either in the Alu-dense N0434.1 interval, leading to localised duplications and deletions, or elsewhere in the genome. However, there are two lines of evidence against this. First, the sperm crossover analysis provided no evidence whatsoever for crossover accompanied by rearrangement, which would have been reflected by differently-sized PCR amplicons during ASO-mapping of crossover breakpoints. Second, at the population level, 100 chromosomes were analysed to establish the diploid genotypes of 55 N0434.1 SNPs (Chapter 7, section 7.2.1) and the pattern of LD across the region. Again, all PCR amplicons generated were of the expected size, providing no evidence for a shift in haplotype structure. This latter observation can be used to determine that the upper limit of the frequency of selectively neutral Alu-promoted ectopic recombination events in the PGMS2 region is 1.5 x 10^{-6}*, which is 4000 times less than the frequency of homologous recombination in the region (Chapter 8, section 8.3.2.a.i). Hence, there appears to be a mechanism that prevents ectopic exchanges between Alu elements. In humans, homologous chromosomes are not paired prior to meiosis (Chapter 1, section 1.1.1.d), but the apparent absence of ectopic exchanges between Alu elements does suggest that pairing occurs in the very early stages of meiosis, before crossover takes place. Alternatively, mismatch repair could act to reject ectopic exchanges between diverged Alus during the recombination process itself.

*(95% confidence intervals applied to the observation of zero ectopic crossover events in 100 chromosomes means that the maximum number of such events is three. Therefore, according to Hardy-Weinberg equilibrium, $2pq$ = (2 x 0.03 x 0.97) = 0.058. This value of heterozygosity was then be applied to the relationship $H_o = (1/(1+4N_e\mu)$ (Chapter 6, section 6.2.1.a.ii) to obtain an estimate of μ, assuming N_e to be 10,000)

9.5: Final Comments

Prior to the analysis of recombination in the *PGPL* region, it appeared as though hotspots dominated human recombination by structuring human DNA diversity into haplotype blocks that are stable in different human populations (Jeffreys *et al.*, 2001, May *et al.*, 2002; Reich *et al.*, 2002; Kauppi *et al.*, 2003). However, the observation of a very recombinationally active region of the genome in which crossover events do not cluster into the "classic" hotspot profile, suggests that the same rules of recombination cannot be applied to the genome as a whole. In addition, the lack of correlation between LD and sperm crossover patterns in the *PGPL* region of PAR1 is not unique, as shown by a recent investigation at the MS32 hotspot. Despite a peak activity of nearly 120 cM/Mb, this hotspot was found to be located within a large block of substantial LD (A.J. Jeffreys, unpublished). Together, these observations are further evidence that quantitative measures of LD cannot be used as a reliable substitute for quantitative measures of recombination rate. A specific consequence of this observation across the PGMS2 region must be that the extent of useful LD (Chapter 1, section 1.22) in the region of *PGPL* and *PLCXL1* is likely to be relatively small, and an association study in this region would require a relatively dense set of markers.

Within this thesis I have described a series of analyses that culminated in determining the pattern of crossover distribution in a region close to the *PGPL* gene (table 9.1). I have discovered a novel gene and numerous novel minisatellites and been able to develop the first accurate physical map of the PAR1 subtelomere. In addition I have been able to establish that over 30 kb of the distal PAR1 region is in free association with the telomere, has a normal level of diversity and contains the most recombinationally active region of DNA that has so far been identified at the molecular level. However, the distribution of crossovers within this region is unlike anything that has seen before. Thus, rather than helping to establish unified hotspot-based rules that govern meiotic recombination in the human genome, it appears that situation is even more complex than previously thought.

Table 9.1: A summary of the main results of the work presented in this thesis

Chapter 3 **Sequencing the *PGPL* Region**	• N0434.1 is GC, Alu and tandem repeat rich. • Identification of *PLCXL1*, a novel gene. • 5′ end of *PGPL* is not within N0434.1 interval.
Chapter 4 **Extending and Characterising the Known Sequence in the PAR1 Telomere-*PGPL* Interval**	• Not possible to sequence all of cosmid 3F3. • Sanger Centre produced distal PAR1 sequence: • All of distal PAR1 is GC, Alu and tandem repeat rich. • Identification and physical mapping of previously reported distal PAR1 minisatellites. • *PLCXL1* is probably the most telomeric PAR1 gene.
Chapter 5 **Analysis of Genes in the N0434.1 Interval**	• Established complete exon-intron structure of *PGPL* and *PLCXL1*. • *PLCXL1* appears related to PI-PLC gene family. • *PLCXL1* is highly conserved in mouse.
Chapter 6 **DNA Diversity in the N0434.1 Interval**	• PGMS1 and PGMS2 minisatellites are only moderately variable. • STIR array shows length variability. • Nucleotide variability in region is normal.
Chapter 7 **Linkage Disequilibrium in the N0434.1 Interval**	• Region contains an over-representation of high frequency SNP markers. • Some evidence of LD blocks. • Overall LD pattern consistent with recombination occurring at a rate close to the autosomal genome average.
Chapter 8 **Sperm Crossover Analysis in the N0434.1 Interval**	• Targeted PGMS2 region for analysis. • Most recombinationally active region yet identified. • Distribution of crossover events does not fit the classic hotspot profile.

Appendix 1

Sequence of Cosmid N0434.1

N0434.1 was sequenced as described in Chapter 3. Presented here is the entire sequence, which begins 53263nt from the start of cosmid 29C1 (see Chapter 4) and has been annotated as follows:

Alu   L1   Exon   STIR   Tandem repeat   SNP

```
                                                                          PLCXL1 exon 2-->
1       TGAGTTCATGTAGCTGGTGTTGGCTTAGGGTCTGGAGGAAGGCTTTTGGGAAGATGTAAATAAGAACAAAATCTGCAGC  80

81      CACCTGGGAAGCCTGGCCTCAGTGTGGAAGAGAAGGCAGCAGGATTATTACAGAACCTTGTGAAGCCAACGCGGGCAGCC  160
161     GCCAGGAGCTGCAGACCGAGAGGATCTCGTCCTTTCTTGCGGCCCAGGGAGACCAGGCCTTTCATTCTGGGCTCGAGACC  240
241     AACAATTCGAATTCCGAACTCCCCCTGCGTGTGGGACTCAAGGTGGGTTTGCAGTTTGCAGGCAGCTGAAGTTTGTCTCT  320

321     TCTCCAGGAGGCCGGGGCTTCTTCCCTTCCTCTCTGTCCCATTTCTTTTTTCTTGAGACAGAGTCTCACTTTGTCACCCA  400
                            423C/T
401     GGCTGGAGTGCAATGGTGTGATTGTGGCTCACTACAGCCCCCGCCTCCCGGGTTCAAGCCTCAGCCTCCTGAGTAGCTGG  480
481     GATTACAGGCGTGCGCCACCACGGCCGGCTACTTTTTGTACTTTTAGTAGAAATGGGGTTTCACCATGTTGGCCAGGCTG  560
561     GTCTCGAACCCCTGACCTCAGGTGATCCACCCACCTCAGCCTCCCAAAGTGATGGGATGACAGGCGTGAGCCACCGTGCC  640
641     CGGCCCCTCCAGGTCTCATTTCTAAGAGGAGGCCTCAGGTCCACCAGGAAACATTCCTCAGATGTGAAACTGTCAACAGG  720
        <--AluSx
                                                            NMS1
721     CTGATTTCTGGGCTCAAGATCAACAATTCTAATTGATTTGATTAAATCAATTAGATCTAATGATTTTAATCTAATCAGTT  800
                    AluSq-->
801     TTAATCTAAATGATTAAAAATCTTACATACATTGCCGGGCGTGGTGGCGGACGCCTGTAATCCCAGCTACTCTGGAGGCT  880
881     GAGGCAGGAGAATTGCTTGAACCTGGGAGGCGGAGGTTGCAGTGAGCCGAGATTGCATCATTGCACTCCAGCCTGGGTAA  960
                                                                  AluSq-->
961     CAAGAGTGAAACCCTGTCTTTAAAAAAAACAAAACAAAACAAAAAAAAAACCGTACATACAGCTGGGCACCGTGGCTCAC  1040
                                                                     1107G/A
1041    GCCAGTAATCCCAGCACTTTGGGAGGCCGAGGCAGGCAGATCACCTGAGGTCAGGAGTTCAAGACTAGCCTGACCAAGAT  1120
1121    AGTGAAACCCCGTCTCTACCAAAAATACAAAAATTAAGCAGGTGTGGTGGCGGGCGCCTGTAATCCCAGCTACTCTGGAG  1200
1201    GCTGAGGCAGGAGAATTGCTTGAACCTGGGAGGCGGAGGTTGCAGTGAGCCGCGATCGCGCCATTGCAGTCCAGCCTGGG  1280
1281    CAACGAGAGGGAAACTGTGTCAAAAAAAAAAAAAAAAAGACCAACCAAAAAAGTTATATACACTTCAGAGGCAGAGAAAGA  1360
                                                                 N(TA)1
1361    ATTTACAAGTTGTCTAAAATGTCCTTATGGAAAGGGTCACTTCCCTTATTTTCAACAGTATATTATATATATATACTTAT  1440
1441    ATATGTATATATATAGTGATGTATATATGTATATATGTTATGTATGTGTTATATATGTCTATATTATATATGTATATATGTT  1520
1521    ATACATGTATGTTATATATATATTATATATATATTATATATGTATATATTATATGTATAATATATATATTATATGTATATAT  1600
                                                                 N(TG)2
1601    TATATATGTTATATATATATGTTATGTATATAATATGTATATATGTATATATTCTGTTATGTGTGTATGTGTGTGTGTGT  1680
                                                                           AluY/Sp/q-->
1681    GTGTGTGTAGGCCACACGGACACACGTGGAGTGGTTTTAAGGAGCGGAGAGTTTAATAGGAAAGAAGGGAGGTCCGGGCA  1760
1761    GTGGCTCACGCCTGTAATCCCAGCACCGCGGAGGTTGCGGTGAGCCGAGATCGCGCCATTTCACTGCAGCCTGGGCAACA  1840
1841    AGAGCGAAACTGCGTCTCAAAAAAAAAAAAAACCAAGGCGAGAAGGCAGAAAGAAGTGGCTCCCCTGGACTGAGACAGAGG  1920
1921    GACGGGGGCTCCAAACCCCAGGCAGGAGACCAGCCCGTGTTATACGGTGCCTGGAGGAGGCGTGACTCATTTGCATAGCG  2000
2001    CTGAGGGGATTGGTCTGACCAGGCCTGTCATTCACGTAGCCCGCGAAAAACCTGGCCCGCCCACCCCAGTTCCGTAATAT  2080
                                                NMS2
2081    GCAAATGTAGGGCGCCATGATGTTCCACACCCCTGAGGGTAGTGGGGGCGGCCGTGGTGTCAGGCCCGGGTGGGGGCGGC  2160
2161    CTTGGTGTCAGGCCCGGGTGGGGCGGCCGTTGGTGCCAAGCCCGGGTGGGGTCGGCCGTGTTGCCAGGCCCGGGTGGG  2240
2241    GGCGGCCTTGGTGTCAGGCCCGGGTGGGGCGGCCTTGGTGTCAGGCACGGGTGGGGGCGAGGGCAAGAGGGCAACGGTG  2320
                                     2357C/T
2321    GGAATCGCCATGTGGGCTGGACCAGCTAAAGGCTCGCATTTGCATATTAAAGGTTGCCAAGCCGGGTCTAAGAGCCAGGG  2400
                                                                2458C/T
2401    CTTTCACGCTAGACAAGAAACATTTTTTGGAGCTGCAAAAAATGCTTCCAAGGACCCCTTTTCCTCTCTCTCTCTCTCTT  2480
2481    TTTTTTTTTTCTTTTTTGGATATGGAGTTTCACTCTTGTTGCCCAGCCTGGAGTGCAGTGGCGGGATCTCCGCTCACTGT  2560
2561    AACCTCCACCTCCCGGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATGACAGGTGCCACAGGTGCCCA  2640
2641    CCACCACGCCCAGCTAATTTTTGTATTTTTAGTAGAGATGGGGTTTTGTCAAGTTGGCCAGGCTGGTCTCGAACTCCTGA  2720
                                                     PLCXL1 exon 3-->          2793C/T
2721    CCTGAAGTCACCGTAAAAAACCTCTTTTCCTCTTCTCCTTCCTCAGGTTGCCCAGGGCTCACCTCTGATGGGTGGGCAGG  2800
        <--AluSq
2801    TGAGCGCTTCCAACAGCTTCTCGAGGCTGCACTGCAGAAATGCCAACGAGGACTGGATGTCGGCACTGTGTCCCCGGCTC  2880
2881    TGGGATGTGCCCCTCCACCACCTCTCCATCCCAGGTGAGGTTGGGGTGGGGCAGGGGCCGTTGCCTCTATCCCAGGTGAC  2960
2961    GGCAGGGTGGGGCGGGAGCTGTGGCGTCTGCATTCCCCAGTGGGGACGCTGGCTGTAGGAGCAATCCTGGGTGTAGGGAA  3040
                                                                3103C/T
3041    ATCCTAGAAAGCACACTGATGTGGACACACACATACACACAGTGCACACGTGTGTGCATACACATATGTACATAGTCATG  3120
         N(CA)3
3121    GACATGCACAGACATGCATACACACAGGCATGCACACATGTGCACACATACACAGGCATACATGCACACACGCACACACG  3200
3201    CACACATGCACATCTGCACACGCACATCTGCACACACACATGCACATCTGCACACACACATGCACATCTGCACACATACG  3280
3281    CATGTGCATACTGTGTTGGTCTGTTCTCACGCTGCTAACAAAGACATACCCGAGACTGGGTAATTTATAAAAGAGGTTTA  3360
```

```
              3370+/-(CA)
3361     ATGGACTCA--GTTCCACCTGGCTGGGGAGGCCTCACAATCACGGCAGAAGGTGAATGCGGAGCAAAGTCACATGTTACAG
3440
3441     GTGGCAGCCAACAGAGCGTGTGCTGGAGAACTGCCCTTTGTACAACCATCAGATCTCGTGAGACTTATTCACGATCACGA 3520
3521     GAACAGCATGGGAAAGACCCGCCCCGTGATTCGGTGACCTCCCACCAGGTCCCTCCCATGACAGATGGGAATTATGGGAG 3600
                                                                             N(CA)4
3601     CTACAATTCAAGATGAGATTTGGGTGGGGACACAGCCGAGCCATATCACATACATAGGTGCACATAGGCACACATGCAGG 3680
3681     CACACACAGACACATACATACACACACACATGCATGCGCACACACATACATGCATGCACACATGCACATCCCCACACATA 3760
3761     TGCACATGGATGCATATACATGCACACACACAGGTGTACACACATGCAGGCACATGTATACATGGATGTAGGCACATACT 3840
3841     AATGCACACATGTATACATGCACATAGGCTCATACACACATGGATGCACATACACATGCATGCACATGTGCGGGTGTGTA 3920
3921     CACATGCACATGCACACGTGTACACATGTACACACATGCACACATGTATATATATATTTGCACCTATGCACACATATATA 4000
4001     CACCTGCAGATATACACAGGTGTAGACATGTACATGCACAGACACACGTGCACACATAAACATTTGCACCTATGCACGCA 4080
4081     GAGACACGTCTATGCAATGCATGTATACACGTGTACACACGTACATGCATGCACACAGACATACACATGTACACATTTGC 4160
4161     ACCTATTCACATAGACATACATGAATGCACATATGCACACGTGTACACAGGCACACATGCACATGCACACGCGTGTACAC 4240
4241     ACATGCACACACGTAGACACACGTCCATGCACACAGGAACACAGGTACATGTTTAAACACAGGCACACACATGCACACTC 4320
                                                                    4383T/G
4321     ACATATAGGCATACGGACACACATACACATGCACACAGATACGTATTGTCAACCTAAGGAAATAAACTTAGACAAAATTA 4400
4401     ATATAAGTAGGGGGTTTCCTTGAGCCACATTTGAGGACTGCAGCCCAGGAAACCCTCCAACTCGCCTTAGGAAGTGGCTG 4480
4481     TGGAGAACAACGTCGAGACGGGAGCTTCTAAAGAAAGTGCACATCAGGAGAGGGGGTGATGATGAAAGCTGTTTTTCAGG 4560

                                 4589A/G
4561     AATTCATGTGGGTTACTGAAGTGGCATCGAGCACTGATCGCGTTCTTCCTTGCAATGTAGGGAATGAGTTATGGTGCCCC 4640
4641     AGTGTGACATCCTGAGGTTAATTTTTTTTTTTTGAGACGGAGTTTTGCTCTTATCGCCCAGGCTCGAGTGCAGTGGCTCC 4720
         4722C/T
4721     ACCTCAGCTCATCGCAACCTCTGCCTCCTGGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGG 4800
                            4823A/G
4801     CACGCTCCACTATGCCTGGCTGATTTTTGTATTTTTAGCAGAGATGGGGTTTCACCATGTTGGTCAAGCTGGTCTCGAAC 4880
4881     TCCCGACCTCAGGCAATCCACCTGCCTCGGCCTCCCAAAGTGCTGGCATTGCAGACATGAGCTATTGCACCCGACATCTT 4960
                                                                            <--AluSp
4961     AGGTTAATTTGTAGAGATGCGGGGCATCTGTCAGTCTAGAGCTGACCTAGTAAGTAGCTTCCAGGGGTGATACTGTAGCA 5040
5041     GGACCAGCCGCAGACAAAACTCAGACACCGGATTAAAGAAAGAAGAGGTTTCTTTGGCCGGGAGCTTCGGCAGACTCACA 5120
5121     TCTTAAGAGCCGAGCTCCTTAAAAAAGAAATTCTTGGCCTTTTTAAAGGCTTACAACTCTGAGGGGTCCACGTGAAGGGG 5200
                                    5230A/G
5201     TCGTGATAAATCAAGCAAGTGTGGGGAACGTGACTGGGGGCTACATGTGTCAGCTAACAGAACAGAAAGTTTTGTAACGC 5280
5281     TTTTTCATACAACGTCTGGCATTTACAGATAACAGAAGTGGTTTAGGTTATGGATTGATATTATTTTAACTCCCAGGGCT 5360
5361     GGGTGGTGGTGCCAAGGTTGTCTGGCTATTTATCTTACTTCTGTTTCTTTCCAACCTTTTGCTTTCTCCTGTCTTATAAA 5440
5441     CTAGGCAAGGTGGGGGTAGGGGGCAGCAGGAGAAGTACTGGTCTCCTTCCTTAACACCTTAGCTCAAGATCGGGGGAGCA 5520
                                                NMS3
5521     GGTGACTGTGGCTTCGTTCCAATGCCTCTCGGGGTCTGGGGATTTAAAGGGGGTCCACGTTCCTCAGAGGATGTAAAGGG 5600
                                          FLAM_C-->
5601     GCTTCACGTTCCTCAGGGGATTTAAAGGGGGAAGGCCGGGCGTGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAAGC 5680
5681     CAAGGCAGGAGGATCACTTGAGTTTAGGAGTTTGAGAGCAGCCTGGGCAACATAGTGAGGCCCCATCTCTGCAAATATGT 5760
5761     ATATACACACACACATTTATATGTAAGTGTACATATTTATATATAAACATACATATTTATATTTATATATAATATATAAA 5840
5841     TAAGTAAATGTAAAATATAAAATATATATCATATATTTATATAAATATATGATACATAAGTACATTTATATATTATATTT 5920
5921     ATAAACATAATATGTGGATATATATTTTATATATTTATATAAATACATATTTATGCAAAAATATGTATTTATGTAAATAT 6000
            NMS4
6001     ATACTAATATATGTTTATATATAGATATATATACACTATATATGTTTATATATAGAATATATACTATATATGTTTATATA 6080
6081     TAGATATAGATACTATATATGTTTATATATAGATATAGATACTATATATGTTTATATATAGATATAGATACTATATATGT 6160
6161     TTATATATAGATATAGATACTATATATGTTTATATATAGATATAGATACTATATATGTTTATATATAGATATAGATACTA 6240
6241     TATATGTTTATGGATATAGATACTATATATCTTTGTTTATAGATATAGATACTATATATGTTTATATATAGATATATGTC 6320
6321     ATATATATCTTAAATATATATATCTTAAATATAGATATATTTATATCTATATTTAAGAAATAACAACTACGTGTCATAGG 6400
6401     CAGGAAATGCAAACGTCAGCCATCAGTCAAACTCTACAAAGGTCTGTGTTTCTTGCGGTTACCATAACAAGTGAGCACAC 6480
6481     GTCAGGAGGCGGCAAAGAACAGAAATTTACTCTCTCCCAGTTCTAGATACAGAAATCTGAAGTCAAGATATGGGCAGGAC 6560
6561     CACACGTCCTCCTGAGGCTCTAGGGGAGGGTCCTTCCTGCCTCTCCCAGCTCCTGGGGGCTCCAGGCATCCCTGGGCTTG 6640
6641     TGGCCGCATCACTCCAGTCTCTGCCTCCATCTCCACGTGGCCTCCTCCTCTGGGTCTGTGTCTCCTAAGGACACCTGTCA 6720
6721     TTGCATTTAGGACCCAGCACAAAGCCAAGATGACTTCATCTTGAGCATTTTGAGAGCATGTGTAAAGACCCTATTTCCAA 6800
6801     ACAGGATCCCATTCACAGGTTCTGGGCAGGGGGTTAGGACTTGAACAGCTCTTTTGGGGATCACCACTCAAGTCATTGAA 6880
6881     ATTGTATTCAGTTCCTTCTAGAGGCTCTAGAGTGGGGGATCCTTCCTGCCTCTCCCAGCTCCTGGGGGCTCCAGGCATCC 6960
6961     CTGGGCTTGTGGCCGCATCACTCCAGTCTCTGCCTCCGTCTCCACGAGGCCTCCTCCTCTGTGTCTGTCTCCTCCTCTTG 7040
7041     TGTCTTAGACGGACACCTGTCACTGGATTTGGGGGCCCGCCCTACTCCAGGATGATCTCATTTCAACCTAATGATATCTG 7120
7121     CAGAGACCCTGTTTCAACACGGGTCGTGTTCCCAGGTTCCAGTGGACGTGAGTTTTGGGGGTCAGCGTGCACCCCTCCCG 7200
                              7225C/T
7201     CCATACCTGTGCCTGTGCTGTGGGCGGGCAGCAGCCTGTGCTGTAGGCGGGCTGTGGAGCGACTCACAGCAGGTGGCGGG 7280
                       7298A/G                       PLCXL1 exon 4-->
7281     GACGGACTCGTGGTGACATGTCCGCGTGTGTGCTTTGCTCTCAGGGAGCCACGACACGATGACGTACTGCCTGAACAAGA 7360
7361     AGTCCCCCATTTCGCACGAGGAGTCCCGGCTGCTGCAGCTGCTGAACAAGGCCTTGCCCTGCATCACGCGCCCTGTCGTG 7440
7441     CTGAAATGGTCCGTCACCCAGGTACGGTCTGTGCCCCGTGCTGCTGACCTGGCCTGTCAGCTATGTGGGGCCCACGGCCC 7520
7521     CGTGGTGCTGAAATGGCCCCTCACCCAGGTAGGGTTTGAGTGGTGCCCTGTGGGCTCAGGAGGCAGGTTCCTATCCCAGC 7600
7601     TGCGGAGACAGGATTAAAACAAAACCTGCTGCCCACCCAGAACCCCTCCCCAGAGGTGGAAGAGAGGGGAACAGTGTTTT 7680
7681     TATTTATTTATTTTTTGAGACAGGGTCTCACTCTTCTTACCCGGGCTGGAGTGCAGTGGCGTGATCTCGGCTCACTGCAA 7760
7761     CCTCCGCTGCCTGGGTTCAAGCGATTCTCCTGCCTCATCCTCCTAAGTACCTGGGTCTACAGGCACCTGTCACCACACCT 7840
7841     GGCTAATTTTTTGTATTTTTAGTAGAGATGGGGTTTTACCGTGTTAGTCAGGATGGTCTTGATCTCCTGACCTCATGATC 7920
```

```
7921    TGCCCGCCCCAGCCTCCGAAAGTGCTGGGATTACAGGCGTGAGCCACCGCGCCCGGCCCAGCATTTTATTAAATGGCGTA 8000
                                                           <--AluY
                         8020C/T
8001    AAGCTAATGCCGTGAGCATCGCAGGCCTCCACTGAGAGCTGCAGAGGGAGAGGGTCTCGCCCTGTTGCATCCCAGCAGAG 8080
8081    ACAACACCTTTCTTTTTCTTTCTTTTTCTTTTTTTTTTTTTTGAGACGGAGTCTCACTGCCACCCAGGCTGGAGTGCAGT 8160
8161    TGGTGCCATCTCGGCTCACTGCAAGCTCCGCCTCCCGGGTTCACGCCATTCTCCTGTCTCAGCCTCCTGAGTAGCTGGGA 8240
              8248C/T
8241    CTACAGGCTCCCGCGACCACGCCCGGCTAATTTTTGGTATTTTGAGTAGAGTCAGGGTTTCACTGTGTTAGCCAGGATGG 8320
8321    TCTCGAACTCCTGACCTCGTGATCCACCCGCCTCAGCCTCCCAAAGTGCTGGGATTACAGGCGTCAGCCACCGCGCCCGG 8400
                                 8429C/G
8401    CCGAGACAACCCCTTTCATGCCTGCACTCAAGACAGACAAGGACTCATTTCTCGTGTCTTTAGGACCTGAGGCGGGTCTG 8480
   <--AluY
8481    CAGTTTGCAGTCAGGCATCTGCTAGTCCCTGGGAAACAGGGAGACAGGCCCCTTTCTCCCTAATAATCACATTCATTCAT 8560
8561    TCCTTTAGAGATGGAGTCTCGCTCTCTCACCCAGACTGGAGTGCGGTGGTGCGATCTCAGCTTCCTGCAGCCTGGGCCTC 8640
8641    CCAGGATCAAAGGATCCTCCTTCCTCAGCCTCCCGGGCAGCTGGGACTACAGGTGTACACCACCACACCCAGGCTGATTT 8720
                        AluY-->
8721    TTAAAATTTTTAGTAGAGGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGTGGGTGGATCA 8800

               <--AluSg/x
8801    CGAGGTCAGGAGATCGAGACCGTCCTGGCTAACATGTTGAAACCACATCTCTACTAAAAATACAAAACAAAATTAGCTGG 8880
8881    GCGCGGTGGCGGGCGCCTGCAGTCCCAGCTACTGGGGAGGCTGAGGCAGGAGAACGGCATGAACCCGGGAGGCGGAGCTT 8960
8961    GCAGTGAGCCCAGATTGCACCACTGCACTCCAGCCTGGGTGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAGGC 9040
9041    CAGGTCCCCGGAGATGAGGTTTTCCAAATGCAGGTCCCCGTGGCTGCAGCGCTCCCAGCACATACAAACAGCTGTTGTTA 9120
9121    CGACATCCACGTCCCAGTGGGCAGCAGGACATGCGGGTGGGCCAACCCCACAGCCCATTCCTGAGCCCCCCAGACGCGGC 9200
                                               PLCXL1 exon 5-->
9201    GCTCAGCCAGGCAGACATGACAGCAGCCTCTGTCCACAGGCACTGGACGTCACAGAGCAGCTGGATGCCGGGGTGCGGTA 9280
                                                                                   PGMS3
9281    CCTGGACCTGCGGATAGCCCACATGCTGGAGGGCTCGGAGAAGAACCTGCACTTTGTCCATATGGTGTACACAACGGCGC 9360
9361    TGGTGGAGGTGCGGCCGGGCTGAGGTGGGACGCAATGGGGAGAGTGGGAGGCGGCCGGGCACTGGTGCAGGTGCGGCCGG 9440
9441    GCTGAGGTGGGAAGCAAGGGGGACAGCGGGAGGCGGCCGGGCACTGGTGCAGGTGCGGCCGGGCTGAGGTGGGAAGCAAG 9520
9521    GGGGACAGCGGGAGGCGGCCGGGCGCTGGTGCAGGTGCGGCCGGGCTGAGGTGGGAAGCAAGGGGGACGGCGGGAGGTGG 9600
9601    CCAAGCTCCCGTGGGGATGAAACAGCCACATGCAGATGTGGACAGGAAACGCCCGGTCTTTAATGGAAGGGTGACGTCAC 9680
9681    CTATACACCAGACAGGAGACACTGACCCCGCCAATCCGTTACGGAGATTTCTTTTTTTTTTGCTTTTTTTTTTTGGAGAT 9760
                9771C/G
9761    GGAGTCTTGCTGTGTCGCCCAGGCTGGAGTGCAGTGGTGCAATCTCAGCTCACTGCAACCTCTGCCTCCTGGGTTCAGGC 9840
9841    AATTCTCCTGCCTCAGCCTTCCAAGTAGCTGGGATTACAGACGCCGGACACCACGCCCAGCTAATTTTTGTATTTTTAGT 9920
9921    AGATACGGGGTTTCACCATGTTGGCTAGGCTGGTCTCAAACTCCTAACCTCAGGTGATCCACCCGCCTCGGCCTCCCAAA 10000
10001   TTGCTGGGATGACAGGCGTGAGCCGCCACACCCCGCCTTCCCTGTGGTGTTGGAGTCGTGCAGGACTCAGCCCAGCACCC 10080
                                     <--AluSx
                  PLCXL1 exon 6-->
10081   CCCTCCCCAGGACACACTCACGGAAATCTCGGAGTGGCTGGAGCGGCATCCACGCGAGGTGGTCATCCTGGCCTGCAGAA 10160
10161   ACTTCGAGGGGCTGAGCGAGGACCTGCACGAGTACCTGGTCGCCTGTATCAAGAACATCTTCGGGGACATGCTGTGTCCT 10240
10241   CGTGGGGTGAGGAGGGGAAGGATATCCGCACGTCTCTCCCGGGGCAGGGGCATCGTCAGCCTCAGACTCCATTTGCTGTG 10320
10321   CCCTGGGTGTGGGTGCTGCCATGATTCCTGTAGCAGGTAAAGCCGTCGAACGGGGGCTGCCTGCTCTCCCGCAGTGTGGG 10400

                                                                                     AluSc-->
10401   GGGCCCTGGCTGACCCGGTGGGGTGGCTCCTGGTGAGATCTGCTTCCCTTGACGGGTTTAGAAATGTGTGCAGCGTCAGC 10480
10481   CGGGCGCGGTGGCTCACGCCTGTCATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACGAGGTCAAGAGATCGAGAC 10560
10561   TATCCTGGCCAACATGGTGAAACCTGGTGTCTACTAAAAATACAACAATTAGGCTGGGTGTGGTGGCTCACGCTTGTCAT 10640
                                            AluSc-->
10641   CCCAGGACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAAGAGATTGAGACTATCCTGGCCAACATGGTGAAACCCC 10720

                                                  AluY-->
10721   GTCTCTACTAAAAATACCAAAATTATCTGGGTGTGGCGGTGGGTGTGCCGGACGCAGTGGCTCACACCTGTCATCCCAGG 10800
10801   ACTTTGGGAGGCCGAGGCGGGCGGATCATGAGGTCAGGACATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCT 10880
10881   ACTAACACAAAATACAAACAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGA 10960
10961   GAATGGTGTGAACCCGGGGGGTGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCATTCCAGCCTGGGAGACAGAGCAGG 11040
11041   ATCCGTCTCAAAAAATAAAAAATAAAAATAAATAAATATATAAAAATAGAGACAGGGTCACCCTGTGCAGCCCAGGCTCG 11120
11121   AATTCCTGGCCTCGAGTGATCCTCCCGCCGTGTCCTCCCAAAGTGCTGGGATTACAGGCATGAGCCCCTGCACCTGGCCA 11200
                                                                             <--FLAM_C

11201   TACTTAATTTATTTATTTATTTGAGAGGGAGTCTCGGTCTGTCTCCCAGGCTGGAGTGCAATAGCGAAACCTCGGCTCAC 11280
11281   CGCAACCTCTGCCTCCCAGGTTCAAACGATTCTCCTACCTCAGCCTTCTGACTAGCCGGGATGACAGGCGTGCACCACCG 11360
11361   CACCCGGCCAATTTTTGTATTTTTAGTAGAGATGGGGTTTCACCATGTTGCCCGGGCTGGTCTCAAACTCCTGACCTCAG 11440
                                                                 11501A/G
11441   GTGATCCACCCACCTCGGCCTCCCAAAGTGCTGGGATGACAGGCGTCAGCCACTGCGCCCAGCCAGAATTTCATTTTTGG 11520
                                                                <--AluSx
11521   TTTATACTCGTGTTGGGCCGGTGTCCCGACTTCCCAGCCCTCCGCTCTCCCAGGTGCCCCCGGCTCTCCTCTCTCCCCTG 11600
                                  PLCXL1 exon 7-->
11601   CACCCCTTAACTCTGGTCCTTTGCAGGAGGTGCCGACACTGCGGCAGCTGTGGTCCCGGGGCCAACAGGTCATCGTCTCC 11680
11681   TATGAAGACGAGAGCTCCTTGCGCCGGCACCACGAGCTGTGGCCAGGAGTCCCCTACTGGTGGGGAAACAGGGTGAAGAC 11760
11761   CGAGGCCCTCATCCGATACCTGGAGACCATGAAGAGCTGCGGCCGCCCAGGTACCAGGTCGCCCCTCGTGGGGGTAGATT 11840
```

```
11841  CCACACAGCCTCCCGTGACGCCCTGCGGCAGGCCGGGTCCACATGGACTTGCCTTCACTTTTACGTGAAAACAATTTAAA 11920
                  AluSx-->
11921  AGGAATCCATGAGCTGGGCGTGGTAATCCCTGTACCTGTAATCCCAGCACTTCGGGAGGCCGAGGCGGGTGGATCACCTG 12000
12001  AGGTCGGGAGTTTGAGACCAGCCTGGCCAACATAGTGAGACCCCAACTTTACTAGAAATAAAACTTAGCGGCCGGGCGCA 12080
                                                                         AluSq/x-->
12081  GTGGTTCATGCCTGTCATCCTAGCACTTTGGGAGGCTGAGGCGGGCAGATCACCTGAGGTCGGGAGTTCGAGACCGGCCT 12160
                                                5´ truncated L1MB5 -->        12237A/G
12161  GACCAATGTGATGAAACCCTGTCTCTACTTAAGAAGCACCCAGGGAGAAAGCAGTCCATGTACATTGGCGGGGGAAGGA 12240
12241  TCAACAAGGTGTGGTCCATCCACGCGGTGGAATATTACACAGCCATGAAAAAGAACGAGGCTCTGACAAGGATGCAGCGG 12320
12321  GCACAAACCTTAAGACATCACGCTCAGTGAGAGAAGCCAGACACAAAAGGACACGTAGTGTGTGAATCCATTTACAGGAA 12400
12401  ATGCCCAGAACATGCCAATCCAGAGACAGAAAGAGGATTTGTGGTTGCCGGGGGCTGCGGAGGGGAAATGAGGACTGACT 12480
12481  GCTTCATGGAAACAGGGTCTCTCCTTTTAGGACGATGAGAATGTTCTGGAACTAGGGAGAGGTCGGAGTTGCACAACGTG 12560
12561  AATCCACTTTATTTTTTTTAGTTAATTCTTTGATTAGAGAAGGGGTCTCACTGTGTTGGCCAGGCTGGTCTCAAACTCCT 12640
             AluSx--> 12657A/G
12641  GGTGTCAGCCAGGTGCGGTGGCTCACACCTGTGATCACAGCACTTTGGGAGGCCAAGGTGGGTGGATCACCTGAGGTCAG 12720

                                                                               AluY-->
12721  GAGTTTGGAGACCAGCCTGGCCAACATGGCGAAACCTCTTCTCTAGCAAAAATACAAAAATTAGCGGGCACAGGCCGGGC 12800
12801  GCGGTGGCTCACGCCTGTAATCCCAGCACTTCGGGAGGCCGAGGCGGGCGGATCATGAGGTCAGGAGATCGAGACCATCC 12880
12881  CGGCTAACATGGAGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAATCCCAG 12960
12961  CTACTCCGGAGGCTGAGGTAGGAGAATGGTGTGAACTCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCACGCCACTGAAC 13040

                                                        AluSx-->
13041  TCCAGCCTGGGCGACAGAGTGGAACTCTGTCTCAAAAAAAAAAAAAAAAAATAGGCACCTGTAATCCCATCTACTTGGGAGG 13120
13121  CTGAGGCAGGAGAAATGCTTGAATCCAGGAGGCGGAGGTTGCTGTGATCCATGGTCGCGCCACTTGCACTCCAGCCTGGG 13200

                                                      AluSg/x-->
13201  CGACAGGAGCAAGACTCCATCTCAAAACAAAACAAAAAAGAGAAAAATTAGCCAGGCATGGTGATGCGCGCCTATAATCC 13280
                                                                         13351C/T
13281  CAGCTACTCAGGAGGCTGAGGCAGGGGAATCACTTGAACCAGGAAGGCGGAGGTTGCAGTGGGCCGAGATTGAGCCAGTG 13360
13361  TACTCCAGCCTGGGCAACAGAGCGAGACTCTGTCTCAAAACAAAACAAGACAAAATCAAACTTCTGGGGTCAAGCGATCT 13440
                                                          AluSx-->
13441  GCTTGAACCACCATTCCCAGCCGAGTTGTGCAATTTAAAATCGTGAATTTCGGTGCAGTGGCTCACGCCTGTAATCCCAC 13520
13521  CACTTTGGGAGGCCGAGGCAGGTGGATTAATTGAGATCAGGAGTTTGCGACCAGTCTGGTGAAACCCCATCTCTAGTAAA 13600
                                            13642A/G
13601  AATACAAAAAGATTTAGCTGGGCGTGGTTGTGTGCACCTGTAATCCCAGCTCCTCAGGAGGGTTAGGCAGGAGAATTGCT 13680
13681  TGAACCCGGGGGCGGAGGCTGCCGTGAACCGAAATCGGGCCACAGCACTCCAGCCTGGGCAACAGAGCGAGACTCCATCT 13760
13761  TGGGAAAAAAAAAAAAAGGTAACTTTTTTGTACGTGAATTTCACCACAATTTTTGTAAAAAAGCCACCGACGGGGTGTGA 13840
13841  GGTGCAGACAGCGTCGGCAGCCACGGCCCCGTGTCCTCCGGAGGCAGGAGCCTGAGGGAGGGAGGGAGGAAGGTCCTCCC 13920
                                                                            13994A/G
13921  CGCGGGGACGGTGGCAGGTGGGGCCGTCTCGGTCCTGCAGCCTGCGGCTGGGTTCTCTCAAGGTGATCTTCACAGCCGCG 14000
                                                       14051A/G
14001  AGGGCTGCTTCCCACAGGGGCAGCATGAAACCCAGTCACGGGCCGGGCAAAGGAGCAGGTTTTTTCAACTTGGTCTCTTG 14080
14081  CCCTCACACAGGAACGAGGTTAGAGACTTCAGAAATGTCTTTTTTTTTTTTTTTGAGATGGAGTCTCACTCTTGTTGCCCC 14160
                                                    14209C/T
14161  GGTCGGAGTGCAGTGGCGCGATCTCGGCTCACTGCAACCTCCGTCTCCCGGGTTCAAGCCATTCTCCTGCCTCAGCCTCC 14240
14241  TGAGTAGCTGGGACTACAGGCGCCCGCCACCACACCCGGCTAATTTTTTTTTGTATTTTTAGTAGAGACAGGGTTTCACC 14320
14321  ATGTTGGTCGGGCTGGTCTCGAACTCCTGACCTCATGATCTGCCCGTCTCAGTCTCCCAAAGTGCTGGGATGACAGGCGT 14400
14401  GAGCCACGGTGCCAGGCGGAAAATTTTTTTGTTTGTGTGTTTGTTTTTTGAGACAGAGTTTTGCTCTTGTTGCCCAGGCT 14480
               <--AluSq

                                                            14537G/A
14481  AGAGTGCAATGGCGCAATCTTGGCTCACTGCAGCCTTCGCCTCCTGGGTTCGAGCAGTTCTCCCGCCTCAGCCTCCCACC 14560
          14567G/A
14561  ACGCCCGGCTAATTTTTTTTATTTTTAGTAGAGATGGGGTTTCTCCATGTTGATCAGGCTGGTCTCGAACTCCCGACCTC 14640
14641  AGGTGATCTGCCCGCCTTGGCCTCCCAAACTGGTGGGATTACAGGCATGAGCCACTGCACCCGGCCTTTTTTTGTTTGTT 14720
                                                                 <--AluSp

14721  TTGTTTTGTTTTTGAGACAGAGTCTTGCTGTGTCACCCAGGCTGGAGTGCAATGGCGTGATCTCGGCTCACTGCAACACT 14800
14801  CCACCTCCTGGATTCAAGTGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACGGGCACGTACCACCATGCTGAGC 14880
                14893G/A
14881  TAATTTTTGGATGTGTTTTTACTAGAGAAAGGGTTTCAGCGTGTTGGTCAGGCTGGTCTTGAACTCCTGACCTCAGATGA 14960
14961  TCCACCGGCCTCAACCTCTCAACGTGCTGGGATGACAGGCATGAGCCACCGTGCCCAGCCTAGAAATACGTGTTAAAAAC 15040
                                                           <--AluSx
15041  CACTGAGAGAACTTGGCCATAAACGTCTGTTGACTTTGCCGTCTCTGGCCGTGGAATCAGTCCAGCTGGTCAACACAGGG 15120
15121  GCACTGGAACCTACCTCTGGGGGAGCTTCTTCAAATACGGCTTGCAGGAGATTCCGGCTCAGCAGATGTAGGTGGAGCCC 15200
15201  CCAAATTTGCATTTCTAACTATTGCAGGGACCCCACTTTGAGAATCACTGGTCAACGTCAACCCTCTCTATAAGCCTCAT 15280
                                                 PGMS2
15281  AAGAGATGGCACAGACTTAGTGCCCAAGACAACAGCATTCTTCCTGTCTATCACATGGGGGATAAGAATGTGGACATGTT 15360
15361  TGGGGCCATATTATTCTGTCTCCCACATGGGATTAGGACGTGGACATCTTTGGGGCCATTATCCTGTCTACCTCATGGGG 15440
15441  ATTAGGACGTGGACATCTTTGGGGACATTAATCTGTCTATCACATGGGGATTAGGACGTGGACATCTTTGGGGCCATTAT 15520
15521  TCTGTCTCCCACATGGGGATTAGGACGTGGACATCTTTGGGGACATTATTCTGTCTCCCACATGGGGATTAGGACGTGGA 15600
15601  CATCTTTGGGGACATTATTCTGTCTATCACATGGGGATTACGACATGGACATCTTTGGGGACATTATTCTGTCTATCACA 15680
```

```
15681  TGGGGATTAGGACGTGGACATCTTTGGGGCCATTATTCTGTCTCACCTCATGGGGATTAGGACGTGGACATCTTTGGGGA 15760
15761  CATTATTCTTGTCTATCACATGGGAATTAGGACGGTGGACATCTTTTGGGGCCATTATTCTGTCTACCTCATGGGGATTA 15840
15841  GGACGTGGACATCTTTGGGGACATTATTCTGTCTATCACATGGGGATTAGGACGTGGACATCTTTGGGGACATTATTCTG 15920
15921  TCTCCCACATGGGNATTAGGACGTGGACATCTTTGGGGACATTATTCTGTCTATCACATGGGGATTAGGACGTGGACATC 16000
16001  TTTGGGGACATTATTCTATCTCCCACATGGGGATTAGGACGTGGACATCTTTGGGGACATTATTCTGTCTATCACATGGG 16080
16081  GATTAGGACGTGGACATCTTTGGGGCCATTATTCTGTCTACCTCATGGGGATTAGGACGTGGACATCTTTGGGGACATTA 16160
16161  TTCTGTCTATCACATGGGGATTAGGACGTGGACATCTTTGGGGACATTATTCTGTCTACCTCATGGGGATTAGGACGTGG 16240
16241  ACATCTTTGGGGACATTATTCTGTCTATCACATGGGAATTAGGACGTGGACATCTTTGGGGCCATTATTCTGNNNNNNNN 16320
16321  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16400
16401  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16480
16481  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16560
16561  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16640
16641  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16720
16721  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16800
16801  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16880
16881  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 16960
16961  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 17040
17041  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 17120
17121  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN 17200
17201  NNNNNNNNNNNNNNNNNNNNTCTGTCTCCCACATGGGGATTAGGACGTGGACATCTTTGGGGACATTATTCTGTCTATCA 17280
17281  CATGGGGATTAGGATGTGGACATCTTTGGGGACATTATTCTGTCTCCCACATGGGGATTAGGACGTGGACATCTTTGGGG 17360
17361  ACATTATTCTGTCTCCCACATGGGGATTAGGACGTGGACATCTTTGGGGACATTATTCTGTCTATCACATGGGGATTAGG 17440
17441  ACGTGGACATCTTTGGGGCCATTATTCTGTCTATCACATGGGGATTAGGATGTGGACATCTTTGGGGACATTATTCTGTC 17520
17521  TATCACATGGGGATTAGGACGTGGACATCTTTGGGGCCATTATTCTGTCTATCACATGGGGATTAGGATGTGGACATCTT 17600
17601  TGGGGACATTATTCTGTCTCCCACACGGTGTTACGACGTGAGCATCTTTGGGGTTGTCTACTGCCCACCACGCTTTATAA 17680
17681  GCAAAGCTCACCCAATTTCCTTGTTGGACATGGTGCTTTCAACTCTTAATTCCTGAGATGCGAACCTCTAATAATGTGAC 17760
                                                                         PLCXL1 exon 8->
17761  TAGGAGGGAGAAACAGGCGGGTGAGGCCCGTGACCGTGTAACCTCTCCCCACCCTCACCGTTGCAGGAGGGTTGTTCGTG 17840
17841  GCCGGCATCAACCTCACGGAGAACCTGCAGTACGTTCTGGCGCACCCGTCCGAGTCCCTGGAGAAGATGACGCTGCCCAA 17920
17921  CCTTCCGCGGCTGAGCGCGTGGGTCCGAGAGCAGTGCCCGGGGCCGGGTTCACGGTGCACCAACATCATCGCGGGGACT 18000
18001  TCATCGTCGCAGACGGCTTCGTCAGTGACGTCATCGCGCTCAATCAGAAGCTGCTGTGGTGCTGACGGACCCTTCTGAA 18080
18081  GTTCGGGACGCGGCGGCTGCAGTTTCACCCCCGAATTTCCAAGTATTGTGACTTTGTTTGGGCCAAATGTTGGTGATCAT 18160
                                                       AluSg-->
18161  AGGACCGATGATAATACGTTTTCATTTTCTTTAAAATAGAGATGGGGTGGCTGGGCGTGGTGACTTCGCCTGTCTTCCCA 18240
18241  GCACTTTGGGAGGCCGAGGTGGGTGGATCATGAGGTCAGGAGCTTGAGAGCAGCCTGACCAACATGGTGAAATCCCGTCT 18320
18321  CTACTAAAAATACAAAACTTAGCTGGGTGTGTGGCAGGCGCCTGTAGTCCCAGTTACTCGGGAGGCTCAGGCAGGAGAAC 18400
18401  TGCTTGAAGCCGGGAGGTGGAGGTTGCATTGAGCTGACATCGTGCCACTGCACTCCAGTCTGAATGATAGACCGAGACTC 18480

                                    AluSg-->                 18530C/A
18481  CATCTCAAAAGAAAAAAAAACAGCCGAGTGTGCAGTGACTCACGCCTGTCATCCCAGCACTTTGGGAGGCGAAGGCGGGT 18560
                                                             18610C/T              18634C/T
18561  GGAACATGAGGTCAGGAGTTCGAGACCAGCCTGACCAACATGGTGAAACTCGTCTCTATTAAAAACACAAAAATTACCTG 18640
                          18664C/G
18641  GGTGCGTGATGGGCACCTGTAGTCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGTGGAGGTT 18720
18721  GCATTGAGCTGAGATCGTGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCATCTCAAAAAAAAAAAAAAAAAAA 18800
18801  GATGGGGTCTCTCTATGTTGGCCAGGTTGGTCTTGAACTCCTGGCCTCAAGTGATCCTCCCACCTCAGCCTCCCAGAGTG 18880
18881  CTGGGATGACAGTCAAGAACCACCATGGCAGCCCATAATATGTTTTCTTATTTCTGTATTCTCCTTGCTGTGGCGTCTGG 18960
                    <--AluJo/FLAM
              18970A/G
18961  AGCCCTTACAGACCCAGGGAGACACTATCCTCCCACAGCTCACTAATTACTAAACACAGTGACAGCCCACTTAGGAGCCG 19040
                                                            NMS5
19041  GCCTCCCCTGTCAGCCAACCCCTCAGCTAGTTCTCACACCAAGCCTATAGATATATATGTGTGTGTATATATGTGTGTTT 19120
19121  ATGTGTCTGTGTGTATATATGTGTATATATCGGTGTGTATATATGTATGTATATGTTTATACATGTGTATGTGTGTATGT 19200
19201  GTGTGTATATGTATGTGTGTATATATGTATGTGTGCATATGTGTATACGTGTATGCATACATGTATATGTGTATGCATGT 19280
19281  ATATGTGTATGTGTACATGTATATGTGTGTATACATGTATGTGTGTATGCGTGTATACGTGTATGTATACATGTATATGT 19360
19361  GTGTATGCGTGTATATACACACGTATACATATATACGTGCGTGTGTATGTGTATATATATATATGTGTATATATATATTT 19440
19441  TTTGAGGAGTCTCACTCTGTCACCCAGGCTGGAGTGCAACGGCGAGATCTCGGCTCACCGCAACCTCCGCCTCCCTGGTT 19520
19521  CAAGCAATTCTCCTGCCTCCGCCTCCCGAGTAGCTGGGATGACAGGCATGTGCCACCACACCCGACTAATTTCATATATT 19600
19601  TAGTAGAGACGGGGTTTCTCCGCGTTGGTCAGGCCGTTCTCAAACTTCTGACCTCAGGTGATCTGCCCGCCTCGACCTCC 19680
19681  CAAAGTGCTGGGATGACAGGCATGAGCCACTGTGCCCAGCCGCCGATATTTCTTTAAATTATGTTTAGGGACAGGGTCTT 19760
                                       <--AluSx

19761  TTTCTGTCACCCAGGCTGGAGGGCAGTGGTACAGTCATAGCTCACTGCAGCCTCAACCTCCTGGGCTCAAGCGATCCTCT 19840
19841  CAGCTCAGCCTCCCGTGTAGCTGGGGCTCCAGGGACACACCCCCACTGCTGGCTAATTTTTGTATTTTTGGTAGAGTCAG 19920
                                              19963C/G
19921  GGTTTCACCACATGGCACAGTCTGGTCTCAAATTCCTGGGCTCCAGTGATCCTCCCACCTTAGCTTCCAGAGTGGCCAGG 20000
20001  ATCACAGGCAGGCACCACCATGCCCAGGTAATTTTATTTTTTTGTAGAGACGTGGTCTGGCTATGTTGTCCAGGGTGGTC 20080
                          <--AluJb
20081  TCAAACTCCTGGGCTCAACTGATCCTCCCACCTCAACCTCTGCCATAGCCAGGACCTCAGGTGTCAGCCACCACACCCAC 20160
                                                                            <--FLAM_A
20161  AGCTAATTTTTTTTTGTAGAGATGGGGTCTGGCTCTGTTGCTCAGGCTGGTGTGTAGTAGGGGCACAGTCATAGCTCACTG 20240
```

```
20241   TAACCTTCAACGCCTGGGCTGAAGCAATTCTCCCGCCTCAGCTTCCCAAGTAACTGGGAGTATAGGTGTACACCACCATG 20320
20321   CCCAGCTAATTATTTAGTAGTAGTAGGAGGAGTATTATGTTTGAGATGGAGTCTCGCTCTGTCGCCCAGGCTGGAGTGCA 20400
             <--FRAM

20401   GTGGCGCAGTCTCAGCTCACTGCAACCTCTGCCTCCCAGGTTCAAGCAATTCTTGTGCCTCAGCCTCCTGAGTAGCTGGG 20480
20481   ATTACAGGCGCCCGCCACCGCGCCTGGCTAACTTTTGTATTTTTAGCAGAGATGGGGTTTCACCATGTTGCCCAGGCTGG 20560
20561   TCTTGAACTCCCGACCTCAGGTGATCCATCCGCATTGGCCTCCCAAAGTGCTGGGATCACAGGCGTGAGCCACCGCACCT 20640
20641   GGCCTCAAGCCAGTATTTCCCTGGCCCTAAATCATTCCTGGCTGGGTCCCAGCCCAGTAGAGCCAGCCCCCCAGCCCGGA 20720
  <--AluSx
    20723C/G
20721   GTGCTACTGAAGTGTTCAAAAGTTGTCAATCCTCAGCCGTTCCCTAGTCCTGTTGCCCGGTTCTGCCGAAGCCCCTTGAA 20800
20801   GGCTGTGGCCTGGGCTGTTCCCTCATTCATTCCTGCCTCCCGAGCCAAACCCAGGTATCCGCTTGTGCCCTGCGTGGTGT 20880
20881   GGCAGCGTCTTCTCTCGGGAAGGGTCAGGAGTAATTTCTTCTTTCAGTGGCCTCTCTGTGCTAGTCCCGGTCACCTCCGT 20960
20961   GAATTAAAGTCCTACAGGTACAAGGGAGACCCCCCCCCACGGAAGGCGCCCCCAGTCCGTGTGGGAGACTCGCACACCG 21040
21041   GTTTTCTGCACAGGTTTCTTTCTGCCTCTGAAGCGTGAACGGTCCTAGTTTCAGACGCAGATCCTGCAAATACTTTTTTT 21120
21121   TGTTTTTTTGAGATGGATTTTCCCTCTTGTTGTTCAGGCTGGAGTGCAATGGCACGATCTCAGCTCACTGCAACCTCTGC 21200
21201   CTCCCGGGTTCAAGTGATTCTCCTGCCTCAGCCTCCCGAGGAGCTGAGATTACAGGCGCGTGCCACCGTGCCTGGTTAAT 21280
      TTTGTATTTGTATTTTTATTTATGTTTTGAGACGGAGTTTCGCTCTCGTTGCCGAGGCTGGAGTGTAGTAGTGTGATCCT?
      GGCTCACTGCAACCTCCACCTCCCGGGTTCCAGCAAATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGATTGCAGGCGCC
21281 CGCCACCACGCCCGGATAATTTTTGTGTGTTTAGCAGAGACGGGGTTTCACCATGTTGGCCAGGCTGGTCTCGATCTCCT 21360
21361   GGCTCACTGCAACCTCCACCTCCCGGGTTCCAGCAAATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGATTGCAGGCGCC 21440
21441   CGCCACCACGCCCGGATAATTTTTGTGTGTTTAGCAGAGACGGGGTTTCACCATGTTGGCCAGGCTGGTCTCGATCTCCT 21520
21521   GACCTCAGGTGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATGACAGGCGTGAGCCACCGCGCCCGGCCTATACCTC 21600
                                                                          <--AluSq
21601   ATTTTCTACATGTCGCTTGTTGGAGCTGCTGGTTCAAGTTCCCAGCCAGCCAATGGATGCCAGCACCATTTTTACTCCCC 21680
21681   TTTCCCAAGCAAATCGTGCATTTTTGTCTAACGAGAGACATCAGTTTCTCAGGATGATCCTCAAGAACGTTATGGAGTCC 21760
21761   ATGTTGCAATAGGTTCTCTTTGGGACCTAATGACTCATTTTCCAAAAATCCGCTTCTACTTTTGGTACCCGGTTGCTACG 21840
21841   GTGAAATGAAGGTGCCCCGCATCCAGAAAGACGCACTCCTGGACCACAACCGGCGGCTACCTCAGCCCCACGGCTCTGCA 21920
21921   GGATCAGGGCTCGGGCAGGCCCCGCGGAGATGAAGAATTTGCAGGGAGCCTCCCTGACTTCCGTCGGCTGTGAATCCTTG 22000
                                   22030A/G
22001   TCTGTCAGGGGCGTATCCACAAAATCACCGAATTCATACAGATCGTTTAAATAAATGAACATCATTAAAGTCAAATATGA 22080
                                                                          22145A/G
22081   GTATGAATTTTATTACCACCAATGCAGCCAAGACACCTCTGGCAGCTTTCAGGATAGCACGCCAGAAGCATCTTTAGAAA 22160
            22168C/T  AluSg-->
22161   ATGTTAATTCAGGAGGCCGGGTGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGTGGGCGGATCACAA 22240
22241   GGTCAAGAGTTCGAGACCAGCCTGACCGACATGGTGAAAATACAAAAAATTACTAAATATACAAAAATAATATATAAATT 22320

                                      AluY-->
22321   ATAAATATATAAGAATACTAAAAATATAAAAAATTAGCCAGGCATGGTGGTGGGGGCCTGTAGTCCCAGCTACTCAAGAG 22400
22401   GCTGAGGCAGGAGAATGGTGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGACTGCACCACTGCACTCCAGCCTGGA 22480

                                                            AluSx-->
22481   TGACAGAGTGACACTCATTCCGTCTAAAAAAAAAAAAAAAAGTTAATTCAGGCCGGGCACAGTGGCTCCGCCTGTAATCC 22560
22561   CAGCACTTTGGGAGGCCGAGTTGGGTGGATCACCTGAGGTCAGGAGTTTGAGACCAGCCTGACCGACATGCTGAAAACCC 22640
22641   ATCTCTACTAAAAATGCAAAAAATTAGCCGGGCGTGGTGGCGGGCGCCTGTAATCCCAGCTACTTGGGAGGTTGAGGCAG 22720
22721   GAGAACTGTCTGAACTCAAGAGGCAGAGGTTGCAGTGAACTGAGATCGCACCACTGTACTCCAGCCTGGGTGACAGAGCG 22800
22801   AGATTCTGTCTCAAAACATACAAGGCATTTTGTTTTCCCGTTGATGGAGACGGCTAATGTGCGTGTAACGGCTGCACAGC 22880
22881   CTGGCCACACGCAGGTGAAATTCTCTCTCTGCATCTCTTAGTGGATGGTCTGTGACACATCACCGTCTGGTTTGTTTGTT 22960
22961   TTGAGACGGAGTCTCGCTCTGTCTTCCAGGCTGGAGTGCAGTGGCGCGATCTTGGCTCACTGCAACCTCCGCCTCCCGGG 23040
23041   TTCATGCCATCCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCGCCCGCCACCACCCCCGGCTAATTTTTTGTA 23120
              23133A/G
23121   TTTTTAGCAGAGGTGGGGTTTCACCATGTTAGCCAGGATGGTCTGGATCTCCTGACCTCGTGATCCACCCACCTCAGCCT 23200
23201   CTCAAAGTGCTGGGATTACAGGCGTGAGCCACCGTGCCCGTCCTCACGGTCTGGTTTGAAGCTGCTTCTTTAGTAAAACT 23280
                                           <--AluY
23281   ATTTGCTTTCCCTTCTACTTTTGTGGAAGGGTTCTCTGTGCTGCCGGGAAACCTGATTTTTCGTCATTTCCCCGACACCA 23360
23361   CCATGGGAAACGAGACCATCTGTGAACACAGACAGCCGGGCGGAGGGGCCGTCGGTGCCCACCAGGGCCACGGCTCACGG 23440
                                                 23483A/T
23441   CAGGTGCAGGAGGAACTGGAAATGCTGCTCACGGAAGTAAAATCAAAGGTTTAATGTCCTGTTACGGAAACATTCCGAGG 23520
23521   GAAAGCAGTTCACAGCAGGCACCGAGGGCCCACTGGAATTGTGTGGATGCTCAGGCTTGGAGTGGACGCTCGGGCGGCCC 23600
23601   GCTTTGGGGCAGGTGCGGCCGTGTCACCGGCCTGCACGGTCATCCCAGCAAATGGCTGGGAGCGAGACGGGTGCAGAACC 23680
                                                                                23757A/G
23681   AGACAAGGAGGACCCTGCTGCACCTGACACCAAGCTGCCCCCAACACAGCGGTAACGCCTCAGCTCCCCAGGCAGCGATG 23760
23761   CCCCCACCCCGCAGGCCTCTGTGGGCGTCCGTTCATCCTGGAAAGAGCTTCCGGAATTTGCCGTAGGCTGAGTTGCTGAT 23840
23841   GATGACCCTCACGTCGGCCGCCCCGTCCTCAGGGATCACGTCCACCTCCTGAACTGTGGCCTCCTTATACAGCCAGCTGG 23920
                                                                     <--PGPL exon 10
23921   GCACAGATGCGCGTTGTATGGAGACAAGCAGAACCCGTAAGTATTTGCTTAGTTTCATGATAAATAATTACGCTAAAAAG 24000
24001   AGCTTAGCTCAAACCATTCATCAGACCGTCCTGTTTCCTTTTGTTTTTTTTTTTTTTTGAGACGGAGTCTCACTCTGTCG 24080
                                                                             24151C/T
24081   CGTAGGCTGGAGTGCAGTGGCGCGATCTCAGCTCACTGCAAGCTCCACCTCCCGGGTTCAAGTGATTCTCCTGCCTCAGC 24160
24161   CTCCCGAGTAGCTGGGACTACAGGTGCATGCCACCACACCTGGCTAATTTTTTGTGTTTTTAGTAGAGACGGGGTTTCAC 24240
24241   CGTGTTAGCCAGGATGGTCTCGATCTCCTGACTTCGTGATCCACCCGCCTCGGCCTCCCAAAGTGCTGGGATGACAGGCA 24320
```

```
24321   TGAGCCACTGCGCCCGGCTTTTATTTTTATTTTTTTTTTGAGACAGAGTCTCGCTCTGTCGCCAGGCTGGGGTGCAG 24400
                   <--AluY

24401   TGGCACGATCTTGGCTCACTGCAACCTCGGCCTCCTGGGTTCCAGCAATTCTCCGGCCTTAGCCTCCCGAGTAGCTGGGA 24480
24481   CTACAGGTGCCCGCCACTGCGCCCGGCTAATTTTTTGTATTTTTATTAGAGACGGGGTTTCACCGTGTTAGCCAGGCTGG 24560
24561   TCTCGATCTCCTGACCTTGTGGTCCGCCCACCTCGGCCTCCCAACGTGTTGGGATTACAGGTGTGAGCCACCCCACCTGG 24640
                                                                                 <--AluY

24641   GGTGGTAACTTTTTATTCTTTGTAGAGATGGGGTCTCACCATGTTGCCCAGCCTGGCCTCAAACTCCTCTCAGCTCAAGC 24720
24721   AATCCTCCTGCCTCGGCCTCCCAAAGTCTTGGGGTTACAGGCCTGTGCCACGGCATCCAGCTGGAGCTTGCTTTCTTATT 24800
                                                              <--AluJo
                            PGMS1
24801   GGTAGGGAGACCTGTACCCCTTGACTGGCAGCACAGATTAGGCACCTGTTGTGCGCACAGTCAGAAATGTATTTTGACTG 24880
24881   TCAAGTGCAGATTAGGCACCTGTTGTATGCAGTCAGAAATGTACATTTTGACTGTCAAGCGCAGATTAGGCACCTGTTGT 24960
24961   ATGCAGTCAGAAATGTACATTTTGACTGTCAGCACAGATTAGGCACCTGTTGTATGCACAGTCAGAAATGTACATTTTGA 25040
25041   CTGTCAGTGCAGATTAGGCACCTGTTGTATGCACAGAAATGTACATTTTGGCTGTCAAGCACAGATTAGGCACCTGTTGT 25120
25121   ATGGTCAGAAATGTACATTTTCACTGTCAGCATAATTAGGCACCTGTTGTATCCACAGTCAGAAATGTACATTGAGTGTC 25200
25201   AGCACACATTAGGCACCTGTTGTATGCACAGTCAGAAATGTACATTTTGTCAGCACAGATTAGGCAACTGTTGTATGCAG 25280
25281   TCAGAAATGTATTTTTACTGTCAAGCACAGATTAGGCACCTGTTGTATGCAGTCAGAAATGTACATTTTGACTGTCAGCA 25360
25361   CAGATTAGGCACCTGTTGTATGCAGTCAGAAATGTACATTTTGACTGTCAGTGCAGATTAGGCACCTGTTGTATGCACAG 25440
25441   AAATGTACATTTTGGCTGTCAAGCACAGATTAGGCACCTGTTGTATGGTCAGAAATGTACATTTTCACTGTCAGCATAAT 25520
25521   TAGGCACCTGTTGTATCCACAGTCAGAAATGTATATTTTGAGTGTCAGCACAGATTAGGCACCTGTTGTATGCAGTCAGA 25600
25601   AATGTACATTTTGACTGTCAGCACAGATTAGGCACCTGTTGTATGCAGTCAGAAATGTACATTTTGACTGTCAGCACAGA 25680
25681   TTAGGCACCTGTTGTATGCAGTCAGAAATGTACATTTTGACTGTCAAGCACAGATTAGGCAACTGTTGTATGCAGTCAGA 25760
25761   AATGTATTTTTACTGTCAAGCACAGATTAGGCACCCGTTGTATGCAGTCAGAAATGTACATTTTGACTGTCAGCACAGAT 25840
25841   TAGGCACCTGTTGTATGCACAGTCACAAATGTAGATTTTGACTCTCAAGCGCAGATTAGGCACCTCTTGTATGCACAGTC 25920
25921   ACAAATGTACATTTGATGCAAACCCATTCATCTCGTCTGTACATCCTAAAGCTCTCGGGGATCTCACAGCTCCTTGTGCA 26000
26001   CCCACGAAGAGCCCGTTTCAGAGCCAGAGACAGGCATCCAAAGCACCATCCCGTCTCCTGCCCCTGCAGGCCGCTCACCT 26080
26081   GAGCTGCGCCCCTGCGAGCCTCACACGGAGAGTGAGGATCTGTCTCCCCGTCGCCTTCAAAACCGCCGCATCGAGCTCAG 26160
26161   CTTTCAGCTCCTGGAGCCCGTGGCCCCGCAGGGCAGACACGGGCACGACGTTCGGTTCCGTGGGGCTGTACCTGCAAGGG 26240
                                                                 <--PGPL exon 9
26241   TGGGGATGTCACAGGCCCCGCTCAGCGTCGGGGCGGCCGGACGAAATCAGGGTCCCCAGGAGTCCACTGCCCACGGGGCA 26320
26321   CAGTCTGGGGCCACTCCCTGTGTCCTGACTGCCACCGCTGCGGTTCACACGAGGAGACGGGGCATCTCCCCACCCGGCTC 26400

26401   CAGCGCGTGCAGGGGAAGGAGACGCTTGCGGACCCCAGGGCCGGACTCACCCGGGCACGAGGTCCACCTTGTTGTGAACC 26480
26481   TCCACCATGGAGTCCAGGAGCGGGGCGGGCAGCTGCAGGCCACGCAGCGTGGACAGAACGCTGCATTTCTGGAGCTCCGC 26560
26561   CTCGGGGTGGCTGACGTCCCTCACGTGCAAGATGAGATCCTGTGGGCCGGGCCGTGGGGTCAGAGCTGCGGAGCCTCTGG 26640
                                 <--PGPL exon 8
         26645C/T
26641   TCCCTGACCCCAAGCTTGCAGACAGGCCCAGGAGAGGGGCTCACACGAGCTCCCAACGACAGGCTGGGCATGGGAGGTAC 26720
26721   GCCTGTGTGCAGGCCCCTCGGACACCCCAGGACGGGGGCTCCTAGACCAACAGTGGACGCGAGCCCACCCGGCTGCACTT 26800
26801   ACCCAACCTTCCAAGCCACAGGCAGCAGCTCCGCACCCCCAGACCCACACGCAAGGGGGTGCCATATATGAGCACCCAAC 26880
                         26901A/G
26881   CACCACCCACCCACCCAGGGGGTCTTCATTCAAAGCTTTAGGGCGGCTTCATCCTCCTGTGAAATGTCTTTTAACAGCGG 26960
26961   AATTATTTCCTCTTTAAAGGATGCTTTTTTTCTCACGTTCAAAAAAAATCTTACAGGCAGGCTCCTTACACAAAATTTGA 27040
27041   AAAACACAGGGAAAAAAAAATAAAGCCGTGTGTAATTCTCCAACTTATACTACCGGGGTATCCACGTCTACTTTTTGTTT 27120
                                        AluJb-->
27121   GTTTGGATGCACTTAATACAAATAATATTTTTCCTGTAACTGAGGCACTTTGGGAGGTGGCTTGAGCCCAGAAGTTTGAG 27200
27201   ACCAGCCTGGGCAAGAGAGTGAGGCCGTTTCTACAGAAAGTACAAAAATTAGCCATGGCCTGGTTGTGCGTGTCTGTGGT 27280
27281   CTCAGCTACTCAGGAGGCTGAGGTGGGAGGATCACTTGAGCCCAGGAGGTCGAGGCTGCAGTGAGCCGAGATCATACCAC 27360
27361   TGCGGTTCAGTCTGGGTGACAGAGCGAGACCCTGTCTCTAAAGAAGAAAAGTAAAAACAAAAAAAATAATTTCATCCCAG 27440
27441   GAAAGTTTTACTTTTTTTTTTTTTTTTTTTTTGAGACGAGTCTCGCTCTGTCACACAGGCTGGAGTGCAGTGGCGCGATC 27520
27521   TCAGCTCACTGCAAGCTCCGCCTCCCGGGTTCAGCCATTCTCCTGCCTCAGCCTCTCTTGAGTAGCTGGGACTACAGGCA 27600
27601   CCCGCCACCATGCCTGGCTAATTTTTTATATTTTTAGTAGAGACGGGATTTCGCCGTGGTCTCGATCTCCTGACCTGAAG 27680
             27689C/G
27681   TGATCCACCCGCCTCAACCTCCCAAAGTGCTGGGATTGCAGGCGTGAGCCACCACACCCGGTCCATAATTTATTGTCGGG 27760
                                                                <--AluY
27761   AGGAGTCGAAAGCGGAGTCCAGGCTCCGGGCGGGGTTCAGTCCCATCTCCTCAAGGAGGTGGCAGCCGCGTCCGTTCTTT 27840
                                                                                   27919A/G
27841   GGGACATTTGCTGCTTCTCCCTCAGGGCAAAAAACAAAGCCGTAGCCTGAATGTGACAATCTCACACCTTGTTTTCCTGC 27920
27921   CTTGTTGCTTGACAATATTTCCCCGTGCTCTTCATGCACTTGGAAAGTCTACGGTACGGATGGAGTGTGCAGCTCACTCA 28000
28001   GCACCCTACGGCCGGGGGCAGTTCCGGAGCCAAACAGCACCCCGCCCCCAAATCCACATCCACCAGCAGCCTCAGAATGG 28080
28081   GACCCTACCCAGAATTAAGATCTCTGCGGATGCAGCTGGTGAAGATGAGGTCAGGGTGGAGCAGAGTGGGCCTTAAATCC 28160
28161   AACGACCGACCGGTATGTTTACGACAGAAAGAAAGAGATGTGGGGCAGACACAGAAGAGAAGGGGACCTTGCTTGGAAAT 28240
28241   GGAGTTTTTGCAGATGTAGTTAAGATGAGGTTACCCTGGATTTATCTAGGTGGCCCCTAAATGCAATGACAGGTGTCCTA 28320
28321   GGAGACACAGACACAGAGAAGGCCACGTGGAGATGGAGGCAGAGACTGGAGTGGTGCGGCCACAAGCCCAGGGACGCCTG 28400
28401   GAGCCCCCAGGAGCTGGGAGAGGCAGGAAGGACCCTCACCTAGAGCCTCCAGAAGGAACTGGATCCAACTGAAGTGAACT 28480
28481   GAACTGTGGCCCCCCACAAAAGACCTGTCCATGTATTGATCATCTGGAACCTGTGAATGGGACATTATTTGGAAACAGGG 28560
28561   TCTCTGCAGATGTGATTAAGTGGAAGATCTGAAATGAGATCATCCTGGATTAGGGTGGACCCTAAATCTAATGACTACTG 28640
28641   TCCTTCTAAGAGACAGAAGAGGAGACACAGACACAGAGAAGAAGGCCACAGACAGAGGCAGAGACTGGAGTGATGCGGCC 28720
28721   ACAAGCCCAGGGACGCCTGGAGCCCCCAGGAGCTGGGAGAGGCAGGAAGGACCCTCCCCTAGAGCCTCCAGAAGGAACTG 28800
28801   GATCCAACTGAAGTGAACTGAACTGTGGCCCCCCACAAAAGACCTGTCCATGTATTGATCATCTGGAACCTGTGAATGGG 28880
```

```
28881  ACATTATTTGGAAACAGGGTCTCTGCAGATGTGATTAAGTGGAAGATCTGGAGATGAGATCATCCTGGATTAGGGTGGAC 28960
28961  CCTAAATCTAATGACAGGTGTCCTTCTAAGAGACAGAAGAGGAGACACAGACACAGAGGAGGAGGCCACAGACAGAGGCA 29040
29041  GAGACTGGAGTGATGCGGCCACAAGCCCAGGGACGCCTGGAGCCCCCAGGAGCTGGGAGAGGCAGGAAGGACCCTCCCCT 29120
29121  AGAGCCTCCAGAAGGAACTGGATCCAACTGAAGTGAACTGAACTGTGGCCCCCCACAAAAGACCTGTCCATGTATTGATC 29200
29201  ATCTGGAACCTGTGAATGGGACATTATTTGGAAACAGGGTCTCTGCAGATGTGATTAAGTGGAAGATCTGAAATGAGATC 29280
29281  ATCCTGGATTAGGGTGGACCCTAAATCTAATGACTACTGTCCTTCTAAGAGACAGAAGAGGAGACACAGACACAGAGGAG 29360
29361  AAGGCCACAGACAGAGGCAGAGACTGGAGTGATGCGGCCTCAAGCCCAGGGACGCCTGGAGCCCCCAGGAGCTGGGAGAG 29440
29441  GCAGGAAGGACCCTCCCCTAGAGTCTGTGGTCAGCTCCCGCTTTCTGGAACTGTGGTAGAGCTGGTTCCTGTGGTTTCTG 29520
                         <--STIRs
29521  GCCCCCATTTGTGGCCCCTTTTGCAGTGACAGCCTCAGGACCCCACAGGTGTTTCCAGGGTCTCGGTGTTCATCGATCCT 29600
29601  CAGGCATCACCGTGCTGTGTCCGCCACTGCCCGCACGTCTGTGACACGAACGGTCACTCTGTCTGTCAGGGGTCTGTGAC 29680
29681  GTGGAGCAGGTGCACCCGGGAACACGCTTATGTGTTCAGGCACATAAGCCGGGAGAGGACTGACGGCCGACTTCCTATTT 29760
29761  CTACACGGGCTCACGGCGGCAAGAACATGGCCCTGAGTGTCGGTGAGGCTGGGGCAACCGCTCAGCCTCTCGCCCGGGGA 29840
                                                      NMS6
29841  CGTGGGAGGCTTGGGGACACTAAGTCTGAGCCCCAGGAGGCAGTTGTTGTTCCGGCCCCGAGTTGGGTGGGTGTCTGAG 29920
29921  GGCCCGGCCCCTGGCTTTGTGTGTGTCTGAGTGCCTGGTCCCCGTGCCCAGTGGGTGTCCGAGGGCCTGGCCCCTGGGCT 30000
30001  GAGTGGGTGTCCGAGGGCCCGGCCCCTGGGCTGAGTGGGTGTCCGAGGGCCCGACCCCTGGCTGTGTGTGTCTGAGTGCC 30080
30081  CAGCCCCGTGCCGAATGGGTGTCCGAGCACCCGATCCCCGGCCGTCCCCACGCTCACCGAGTGGCCACGTCTTCCAGGG 30160
30161  TGGCGGAGAAGGACTCGATGAGGCCGTGCGGCAGCTGGGAGAGGAAGCCGATGGTGTCCACGTACAGGACGGTCATGCGT 30240
30241  GAGGGCAGCGTGCCCGCGTGGGCCGTGACGTCCAGCGTGGCAAACAGCTGGTCCCGTGGCTGGATGGCGGCATCGCCCGT 30320
                                                               30381C/T
30321  CAGTGCCTTGATCAGCGTGGTCTTTCCTAGGAGGGCGTGGAGGTCAGGGCGCTGCAGAGATCCCTGCGTCCCAACTCCTA 30400
                    <--PGPL exon 7
30401  GCGCCGCAGCCAGGCCCTCCAAATGGAGACCACGGCACCCTCTGGGGGCCGCACCCCGGGCTACACGGTCCCTGAGCTC 30480
30481  CTCGGGCACCCCGGGCCAGACCCGACGCGTGGACAAAGCCACCGTCGCTGTTCTCGGGCCGATGGAGCGGGGCCGCCGT 30560
                                                    NMS7
30561  GCGGACACGGGGGAGATGGTGGTGTGGACGGGTGTGCGTGTGAAGGCGTGGATGGGAGTGAGGTCGTCTGTGTAGATTTG 30640
30641  GCAATGGCGTAGATGGTGGGATGGTGTAGGTGGGGTGGTGGTGCTGGTGTGGATGGGAGGATGGTGTAGACACAGGGGAG 30720
30721  GTGATGGTGTAGATGGGTGGATATAGATACGGCAGTGGTGCCGGTGTTGACAGAAGGATGGTGTAGACAGAGGAGAGGCG 30800
30801  ATGGTGTAGACAGATGGGTGGATTGTATAGACGGGTGGTGGTATAGATGGGGCAGTGGTGCCAGTGTAGACAGGAGGACG 30880
30881  GTGCAGACGAAGGGGACATTGTGGTGTACACGGGTGGATTATGTAGACGGGGTGGTGGTATAGCTGGGGCAGTGGTGCCA 30960
30961  GTGTAGACAGGAGGACGGTGCAGACAAAGGGAACATTGTGGTGTACACGGGTGGATTATGTAGACGGGGTGGTGGTATAG 31040
31041  ATGGGGTAGTGGTGCTGGTGTAGACAGGAGGATGGTATAGACAGAGGAGAGGTGATGGTGTAGATGGGTGGATTGTGTAG 31120
31121  ACAGGGTGGTGGTATAGATAGGGCAGTGGGGATGGTCTAGACAGGAGGATGGTGTAGATGGAGGGGAGATTGTGGTATAG 31200
31201  ACAGGGTGGTAAAGCTCTGGATGGGAGGATGGGGTAGATGACATCCTCACCTTGGCTGCTGTACCCCACACAGAGACCAC 31280
31281  AGGAAACCCGTCTCCTGGGTGAGCTCTCACAGCAGCTGGTGTACCCCACACGGCGACCATGGGAACCCCCTCTCCTGGGC 31360
31361  ACGTGCTCACCGCAGCTGTCGTACGGCACCACTGAGACGACAGGGACCCCCTGCCCTCCCCCGGGCGAGTCCTCACCGGT 31440
31441  GACACGGAGACCGCGGAAGGCCCCTCCCCTGGGCGCGTGCTCACCGCAGTTGGTGTACCCCACCACGGAGATCACGGGGA 31520
31521  ACTCCCGCCTCGTCCGCTGCCGGCGGAGCAGGTGCCTCTTCTTGCGAAGCCTGTCCAAGGCCTTCCTGATCTTGGCCTCC 31600
                                                                                31679/T
31601  TTCTCTCTCAGGAGACGCTGCTGCAGCTGCATGAAGGATTCTCCTAAAAGACACCCGAGATGGTGAGGCGGTGAGCCACG 31680
                                   <--PGPL exon 6
              31697A/T  31703C/T    31717C/G
31681  CCGGGAAAGGCACAAGTGCGGGCGGTGCCGCGGAGGGTCTGCGGGGGCCCGGGGCCTGCTCCCGCTCCAGCATCTGGGGG 31760
31761  CCCGGCTGCTTTCCCCAGCAGCGGCCCCGACACGTCCTGTTCCATGAGAAAGGGCTCTGCTCGGCGGAACGGGATCTCGC 31840
31841  CTCCGGGTGTCCACCCTGTGAGGTCAACGAGGGCTTCCCGTGACTCGGGGAGAAAGTAAAGGCCAGATCACTACGGAGGC 31920
31921  CTCCGAGCGGGACACGGCCCAGGACGTCATCGTGAGCGGGCTCAGTGGGGCCCCTTTCACATCCGAAGTCCCAGAACCTG 32000
32001  GGAATGGGACCTGATTTGGAAATATGGTCTTTGTAGGGTCTCAAAATATCACCCTAGGTTAGGGTGGGCCCTAAGGCAAC 32080
32081  GACCTGTGTCCTTCTAAGAAGCAGAGACCGGAGTGACGAGGCCACAAGCCTGGGACGCCTGGAGCCCCCAGGAGCTGGGA 32160
32161  GAGGCAGGAAGGACCCTGCCCTAGAGCCTCCAGAGGGAAGTGGATACAAGTGTAGCGGATTGAACTGTGGTCCTCCTAAA 32240
                  32258A/G
32241  AGATCTGTCCACATCCTAATTCCCAGAACTGGTCAATGGGACCTTATTTGAAAATAAGGTGTTTGTCATAACTGAAGTAT 32320
32321  CTAGAGATGCGACGATTCTGGATTAGGGCGGGCCATAAATGCAATGACATGTGTCCTTGTAAGAGACAGAAGAGGAGACA 32400
                                                    32451C/T
32401  CAGACACAGAGGAGGAGGCCACGTGGAGACGGAGGCAGAGACTGGAGTGACGCGGCCACAAACCCAGGGATGCCTGGAGC 32480
32481  CCCCAGGAGCTGGGAGAGGCAGAAGGAGCCCCTGGAGGGAGCGCAGCCCTGTCCTCACCTTGATCTCAGAGTTTTGGTGT 32560
32561  CCAGGAGGGGAGAGCATAGATCTCTGCTGTTTAAATCCCCAGTTTGTGGCTACTGATTTAGGCAAAACCTCCAAAAACCA 32640
32641  CCAGCATCTCCTATCTCCTCAAATGCTACCAGACGGAGCCCCCAGCCAGCCCCCACCCTGTTGGAATCTTCCAGGGCTGA 32720
32721  GGGTCTCCTATGTCCTCAAATGCTACCAGGAGACGGAGACCCCAGCTGGACCCCACCCTGTTGGAATCTTCCAGGGCTGA 32800
32801  GAAGGGTGGCTGCCGCTGACAACCGCATTCCGAGGACCCTCTGGACGCCGCGCCCGGCCCGAGTTACCTGACCCCATGA 32880
32881  TGTAGCGCGAGCCGACTCCTCGGTACAGGTGGGCGACGTCCCTTTTCAAGTTCGACCTGGTGTGGGAACGGAGTGGCTC 32960
                                             <--PGPL exon 5
32961  GGTCTCTGCGGACGCTGTCTCCCTCCCGGCAGAGGTCGAGGTGAAGGTGAAGGGGGCACTGAGCTGGGCCTCTGGGCCGA 33040
33041  AGGGAGGAGCCGCTGTTGCGGCCCCGCTCCGCGTGTAGCTCACACACTGTGAGTGACGCGTGTCCCGGGCCGAGGGGACC 33120
33121  TGCCTAGTGGGGAGCCGGGGGCAGGCAGTGGGGGCAACGGGCCCCGGAGAGAGAGCCCACACGGGGCTTTTCTTCTTTTT 33200
33201  TTTTTTTTTTAGACGGAGTCTCGCTCTGTCACCCAGGCTGGAGTGCAGTGGCGTGATCTCGGCTCACTGCAAGCTCCACC 33280
                                                      Overlap with H22-->
33281  TCTGGGGTTCACGCCATTCTCCTGCCTCAGCCTCCTGAGTAGCTGGGATTTCAGGCGCCCGCCACCACGCCCGGCTAATT 33360
33361  TTTTTCTTGTATTTTTAGTACAGACGGGGTTTCACCGTGTTGGCCAGGATGGTCTCGATCTCGTGACCTCGT........ 33440
                                                                   <--AluY
```

References

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., Anderson, G. G., Zhang, Y., Lench, N. J., Carey, A., Cardon, L. R., Moffatt, M. F. and Cookson, W. O. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**: 191-197.

Adachi, N. and Lieber, M. R. (2002). Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**: 807-809.

Agarwal, S. and Roeder, G. S. (2000). Zip3 provides a link between recombination enzymes and synaptonemal complex proteins. *Cell* **102**: 245-255.

Alani, E., Thresher, R., Griffith, J. D. and Kolodner, R. D. (1992). Characterization of DNA-binding and strand-exchange stimulation properties of y-RPA, a yeast single-strand-DNA-binding protein. *J. Mol. Biol.* **227**: 54-71.

Allers, T. and Lichten, M. (2001). Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* **106**: 47-57.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

Anderson, L. K., Hooker, K. D. and Stack, S. M. (2001). The distribution of early recombination nodules on zygotene bivalents from plants. *Genetics* **159**: 1259-1269.

Anderson, L. K., Offenberg, H. H., Verkuijlen, W. M. and Heyting, C. (1997). RecA-like proteins are components of early meiotic nodules in lily. *Proc. Natl. Acad. Sci. USA* **94**: 6868-6873.

Anderson, R. A., Boronenkov, I. V., Doughman, S. D., Kunz, J. and Loijens, J. C. (1999). Phosphatidylinositol phosphate kinases, a multifaceted family of signaling enzymes. *J. Biol. Chem.* **274**: 9907-9910.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408**: 796-815.

Ardlie, K., Liu-Cordero, S. N., Eberle, M. A., Daly, M., Barrett, J., Winchester, E., Lander, E. S. and Kruglyak, L. (2001). Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am. J. Hum. Genet.* **69**: 582-589.

Ardlie, K. G., Kruglyak, L. and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**: 299-309.

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., Povey, S., Jeremiah, S. and Jeffreys, A. J. (1990). Systematic cloning of human minisatellites from ordered array charomid libraries. *Genomics* **8**: 501-512.

Asherson, P., Parfitt, E., Sargeant, M., Tidmarsh, S., Buckland, P., Taylor, C., Clements, A., Gill, M., McGuffin, P. and Owen, M. (1992). No evidence for a pseudoautosomal locus for schizophrenia. Linkage analysis of multiply affected families. *Br. J. Psychiatry* **161**: 63-68.

Ashley, T. (1994). Mammalian meiotic recombination: a reexamination. *Hum. Genet.* **94**: 587-593.

Badge, R. M., Yardley, J., Jeffreys, A. J. and Armour, J. A. (2000). Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum. Mol. Genet.* **9**: 1239-1244.

Bailis, J. M. and Roeder, G. S. (1998). Synaptonemal complex morphogenesis and sister-chromatid cohesion require Mek1-dependent phosphorylation of a meiotic chromosomal protein. *Genes Dev..* **12**: 3551-3563.

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.

Baker, S. M., Plug, A. W., Prolla, T. A., Bronner, C. E., Harris, A. C., Yao, X., Christie, D. M., Monell, C., Arnheim, N., Bradley, A., Ashley, T. and Liskay, R. M. (1996). Involvement of mouse Mlh1 in DNA mismatch repair and meiotic crossing over. *Nat. Genet.* **13**: 336-342.

Bamshad, M. and Wooding, S. P. (2003). Signatures of natural selection in the human genome. *Nat. Rev. Genet.* **4**: 99-111.

Barlow, A. L. and Hultén, M. A. (1998). Crossing over analysis at pachytene in man. *Eur. J. Hum. Genet.* **6**: 350-358.

Bass, H. W., Marshall, W. F., Sedat, J. W., Agard, D. A. and Cande, W. Z. (1997). Telomeres cluster *de novo* before the initiation of synapsis: a three-dimensional spatial analysis of telomere positions before and during meiotic prophase. *J. Cell. Biol.* **137**: 5-18.

Batzer, M. A. and Deininger, P. L. (2002). Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**: 370-379.

Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., Schmid, C. W., Zietkiewicz, E. and Zuckerkandl, E. (1996). Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42**: 3-6.

Batzer, M. A., Rubin, C. M., Hellmann-Blumberg, U., Alegria-Hartman, M., Leeflang, E. P., Stern, J. D., Bazan, H. A., Shaikh, T. H., Deininger, P. L. and Schmid, C. W. (1995). Dispersion and insertion polymorphism in two small subfamilies of recently amplified human Alu repeats. *J. Mol. Biol.* **247**: 418-427.

Baudat, F., Manova, K., Yuen, J. P., Jasin, M. and Keeney, S. (2000). Chromosome synapsis defects and sexually dimorphic meiotic progression in mice lacking Spo11. *Mol. Cell* **6**: 989-998.

Baudat, F. and Nicolas, A. (1997). Clustering of meiotic double-strand breaks on yeast chromosome III. *Proc. Natl. Acad. Sci. USA* **94**: 5213-5218.

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.

Benson, F. E., Baumann, P. and West, S. C. (1998). Synergistic actions of Rad51 and Rad52 in recombination and DNA repair. *Nature* **391**: 401-404.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res..* **27**: 573-580.

Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P. C., Nicolas, A. and Forterre, P. (1997). An atypical topoisomerase II from *Archaea* with implications for meiotic recombination. *Nature* **386**: 414-417.

Bernardi, G. (1989). The isochore organization of the human genome. *Ann. Rev. Genet.* **23**: 637-661.

Berridge, M. J. (1993). Inositol trisphosphate and calcium signalling. *Nature* **361**: 315-325.

Berridge, M. J. and Irvine, R. F. (1984). Inositol trisphosphate, a novel second messenger in cellular signal transduction. *Nature* **312**: 315-321.

Berridge, M. J. and Irvine, R. F. (1989). Inositol phosphates and cell signalling. *Nature* **341**: 197-205.

Bickel, S. E., Wyman, D. W. and Orr-Weaver, T. L. (1997). Mutational analysis of the *Drosophila* sister-chromatid cohesion protein ORD and its role in the maintenance of centromeric cohesion. *Genetics* **146**: 1319-1331.

Bishop, D. K., Park, D., Xu, L. and Kleckner, N. (1992). DMC1: a meiosis-specific yeast homolog of *E. coli* recA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell* **69**: 439-456.

Blumental-Perry, A., Zenfirth, D., Klein, S., Onn, I. and Simchen, G. (2000). DNA motif associated with meiotic double-strand break regions in *Saccharomyces cerevisiae. Embo Reports* **1**: 232-238.

Bois, P. and Jeffreys, A. J. (1999). Minisatellite instability and germline mutation. *Cell. Mol. Life Sci.* **55**: 1636-1648.

Borde, V., Goldman, A. S. and Lichten, M. (2000). Direct coupling between meiotic DNA replication and recombination initiation. *Science* **290**: 806-809.

Borde, V., Wu, T. C. and Lichten, M. (1999). Use of a recombination reporter insert to define meiotic recombination domains on chromosome III of *Saccharomyces cerevisiae. Mol. Cell Biol.* **19**: 4832-4842.

Britten, R. J., Baron, W. F., Stout, D. B. and Davidson, E. H. (1988). Sources and evolution of human Alu repeated sequences. *Proc. Natl. Acad. Sci. USA* **85**: 4770-4774.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. and Weber, J. L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am. J. Hum. Genet.* **63**: 861-869.

Brown, W. R. (1988). A physical map of the human pseudoautosomal region. *Embo J.* **7**: 2377-2385.

Brown, W. R. (1989). Molecular cloning of human telomeres in yeast. *Nature* **338**: 774-776.

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., Collick, A., Brown, J. and Jeffreys, A. J. (2000). Somatic versus germline mutation processes at minisatellite CEB1 (D2S90) in humans and transgenic mice. *Genomics* **65**: 95-103.

Buard, J. and Vergnaud, G. (1994). Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *Embo J* **13**: 3203-3210.

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.

Burgoyne, P. S. (1982). Genetic homology and crossing over in the X and Y chromosomes of mammals. *Hum. Genet.* **61**: 85-90.

Butler, Y. X., Abhayawardhane, Y. and Stewart, G. C. (1993). Amplification of the *Bacillus subtilis* maf gene results in arrested septum formation. *J. Bacteriol.* **175**: 3139-3145.

Cañestro, C., Gonzalez-Duarte, R. and Albalat, R. (2002). Minisatellite instability at the Adh locus reveals somatic polymorphism in *Amphioxus. Nucleic Acids Res..* **30**: 2871-2876.

Cao, L., Alani, E. and Kleckner, N. (1990). A pathway for generation and processing of double-strand breaks during meiotic recombination in *S. cerevisiae. Cell* **61**: 1089-1101.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q. and Lander, E. S. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**: 231-238.

Carpenter, A. T. (1975). Electron microscopy of meiosis in *Drosophila melanogaster* females: II. The recombination nodule - a recombination-associated structure at pachytene? *Proc. Natl. Acad. Sci. USA* **72**: 3186-3189.

Carpenter, A. T. (1987). Gene conversion, recombination nodules, and the initiation of meiotic synapsis. *Bioessays* **6**: 232-236.

Carrington, M., Stephens, J. C., Klitz, W., Begovich, A. B., Erlich, H. A. and Mann, D. (1994). Major histocompatibility complex class II haplotypes and linkage disequilibrium values observed in the CEPH families. *Hum. Immunol.* **41**: 234-240.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012-2018.

Cervantes, M. D., Farah, J. A. and Smith, G. R. (2000). Meiotic DNA breaks associated with recombination in *S. pombe*. *Mol. Cell* **5**: 883-888.

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J. and Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc. Natl. Acad. Sci. USA* **94**: 1041-1046.

Chakravarti, A. (2001). To a future of genetic medicine. *Nature* **409**: 822-823.

Chakravarti, A., Buetow, K. H., Antonarakis, S. E., Waber, P. G., Boehm, C. D. and Kazazian, H. H. (1984). Nonuniform recombination within the human beta-globin gene cluster. *Am. J. Hum. Genet.* **36**: 1239-1258.

Chu, S. and Herskowitz, I. (1998). Gametogenesis in yeast is regulated by a transcriptional cascade dependent on Ndt80. *Mol. Cell* **1**: 685-696.

Chua, P. R. and Roeder, G. S. (1997). Tam1, a telomere-associated meiotic protein, functions in chromosome synapsis and crossover interference. *Genes Dev.*. **11**: 1786-1800.

Chua, P. R. and Roeder, G. S. (1998). Zip2, a meiosis-specific protein required for the initiation of chromosome synapsis. *Cell* **93**: 349-359.

Ciccodicola, A., D'Esposito, M., Esposito, T., Gianfrancesco, F., Migliaccio, C., Miano, M. G., Matarazzo, M. R., Vacca, M., Franze, A., Cuccurese, M., Cocchia, M., Curci, A., Terracciano, A., Torino, A., Cocchia, S., Mercadante, G., Pannone, E., Archidiacono, N., Rocchi, M., Schlessinger, D. and D'Urso, M. (2000). Differentially regulated and evolved genes in the fully sequenced Xq/Yq pseudoautosomal region. *Hum. Mol. Genet.* **9**: 395-401.

Cohen, P. T. (1997). Novel protein serine/threonine phosphatases: variety is the spice of life. *Trends Biochem. Sci.* **22**: 245-251.

Collinge, J., Delisi, L. E., Boccio, A., Johnstone, E. C., Lane, A., Larkin, C., Leach, M., Lofthouse, R., Owen, F., and Poulter, M. (1991). Evidence for a pseudoautosomal locus for schizophrenia using the method of affected sibling pairs. *Br. J. Psychiatry* **158**: 624-629.

Collins, A. (2000). Linkage disequilibrium in maps of SNPs and other markers. *GeneScreen* **1**: 59-61.

Conne, B., Stutz, A. and Vassalli, J. D. (2000). The 3' untranslated region of messenger RNA: A molecular 'hotspot' for pathology? *Nat. Med.* **6**: 637-641.

Cooke, H. J., Brown, W. R. and Rappold, G. A. (1985). Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature* **317**: 687-692.

Cooke, H. J. and Smith, B. A. (1986). Variability at the telomeres of the human X/Y pseudoautosomal region. *Cold Spring Harb. Symp. Quant. Biol.* **51** : 213-219.

Cozens, A. L., Runswick, M. J. and Walker, J. E. (1989). DNA sequences of two expressed nuclear genes for human mitochondrial ADP/ATP translocase. *J. Mol. Biol.* **206**: 261-280.

Cruciani, F., Bernardini, L., Santolamazza, P., Modiano, D., Torroni, A. and Scozzari, R. (2003). Linkage disequilibrium analysis of the human adenosine deaminase (ada) gene provides evidence for a lack of correlation between hot spots of equal and unequal homologous recombination. *Genomics* **82**: 20-33.

Cullen, M., Erlich, H., Klitz, W. and Carrington, M. (1995). Molecular mapping of a recombination hotspot located in the second intron of the human TAP2 locus. *Am. J. Hum. Genet.* **56**: 1350-1358.

Cullen, M., Noble, J., Erlich, H., Thorpe, K., Beck, S., Klitz, W., Trowsdale, J. and Carrington, M. (1997). Characterization of recombination in the HLA class II region. *Am. J. Hum. Genet.* **60**: 397-407.

Cullen, M., Perfetto, S. P., Klitz, W., Nelson, G. and Carrington, M. (2002). High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am. J. Hum. Genet.* **71**: 759-776.

d'Amato, T., Campion, D., Gorwood, P., Jay, M., Sabate, O., Petit, C., Abbar, M., Malafosse, A., Leboyer, M., and Hillaire, D. (1992). Evidence for a pseudoautosomal locus for schizophrenia. II: Replication of a non-random segregation of alleles at the DXYS14 locus. *Br. J. Psychiatry* **161**: 59-62.

D'Esposito, M., Ciccodicola, A., Gianfrancesco, F., Esposito, T., Flagiello, L., Mazzarella, R., Schlessinger, D. and D'Urso, M. (1996). A synaptobrevin-like gene in the Xq28 pseudoautosomal region undergoes X inactivation. *Nat. Genet.* **13**: 227-229.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**: 229-232.

Dawson, E., Abecasis, G. R., Bumpstead, S., Chen, Y., Hunt, S., Beare, D. M., Pabial, J., Dibling, T., Tinsley, E., Kirby, S., Carter, D., Papaspyridonos, M., Livingstone, S., Ganske, R., Lohmussaar, E., Zernant, J., Tonisson, N., Remm, M., Magi, R., Puurand, T., Vilo, J., Kurg, A., Rice, K., Deloukas, P., Mott, R., Metspalu, A., Bentley, D. R., Cardon, L. R. and Dunham, I. (2002). A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544-548.

de Massy, B. and Nicolas, A. (1993). The control in *cis* of the position and the amount of the ARG4 meiotic double-strand break of *Saccharomyces cerevisiae. Embo J.* **12**: 1459-1466.

de Massy, B., Rocco, V. and Nicolas, A. (1995). The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in *Saccharomyces cerevisiae. Embo J.* **14**: 4589-4598.

Dernburg, A. F., McDonald, K., Moulder, G., Barstead, R., Dresser, M. and Villeneuve, A. M. (1998). Meiotic recombination in *C. elegans* initiates by a conserved mechanism and is dispensable for homologous chromosome synapsis. *Cell* **94**: 387-398.

Detloff, P., White, M. A. and Petes, T. D. (1992). Analysis of a gene conversion gradient at the HIS4 locus in *Saccharomyces cerevisiae. Genetics* **132**: 113-123.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**: 311-322.

Devlin, C., Tice-Baldwin, K., Shore, D. and Arndt, K. T. (1991). RAP1 is required for BAS1/BAS2- and GCN4-dependent transcription of the yeast HIS4 gene. *Mol. Cell Biol.* **11**: 3642-3651.

Diaz, R. L., Alcid, A. D., Berger, J. M. and Keeney, S. (2002). Identification of residues in yeast Spo11p critical for meiotic DNA double-strand break formation. *Mol. Cell Biol.* **22**: 1106-1115.

Dolganov, G. M., Maser, R. S., Novikov, A., Tosto, L., Chong, S., Bressan, D. A. and Petrini, J. H. (1996). Human Rad50 is physically associated with human Mre11: identification of a conserved multiprotein complex implicated in recombinational DNA repair. *Mol. Cell Biol.* **16**: 4832-4841.

Drake, J. W., Charlesworth, B., Charlesworth, D. and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics* **148**: 1667-1686.

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.

Duncan, B. K. and Miller, J. H. (1980). Mutagenic deamination of cytosine residues in DNA. *Nature* **287**: 560-561.

Dunham, I., Shimizu, N., Roe, B. A., Chissoe, S., *et al.*[†] (1999). The DNA sequence of human chromosome 22. *Nature* **402**: 489-495. [†] *see paper for full list of 217 authors*

Egel, R. (1995). The synaptonemal complex and the distribution of meiotic recombination events. *Trends Genet.* **11**: 206-208.

Egel-Mitani, M., Olson, L. W. and Egel, R. (1982). Meiosis in *Aspergillus nidulans*: another example for lacking synaptonemal complexes in the absence of crossover interference. *Hereditas* **97**: 179-187.

Eijpe, M., Heyting, C., Gross, B. and Jessberger, R. (2000). Association of mammalian SMC1 and SMC3 proteins with meiotic chromosomes and synaptonemal complexes. *J. Cell Sci.* **113**: 673-682.

Ellis, N. A., Goodfellow, P. J., Pym, B., Smith, M., Palmer, M., Frischauf, A. M. and Goodfellow, P. N. (1989). The pseudoautosomal boundary in man is defined by an Alu repeat sequence inserted on the Y chromosome. *Nature* **337**: 81-84.

Ellis, N. A., Tippett, P., Petty, A., Reid, M., Weller, P. A., Ye, T. Z., German, J., Goodfellow, P. N., Thomas, S. and Banting, G. (1994a). PBDX is the XG blood group gene. *Nat. Genet.* **8**: 285-290.

Ellis, N. A., Ye, T. Z., Patton, S., German, J., Goodfellow, P. N. and Weller, P. (1994b). Cloning of PBDX, an MIC2-related gene that spans the pseudoautosomal boundary on chromosome Xp. *Nat. Genet.* **6**(4): 394-400.

Ellison, J. W., Ramos, C., Yen, P. H. and Shapiro, L. J. (1992). Structure and expression of the human pseudoautosomal gene XE7. *Hum. Mol. Genet.* **1**: 691-696.

Engebrecht, J., Hirsch, J. and Roeder, G. S. (1990). Meiotic gene conversion and crossing over: their relationship to each other and to chromosome synapsis and segregation. *Cell* **62**: 927-937.

Engebrecht, J. and Roeder, G. S. (1990). MER1, a yeast gene required for chromosome pairing and genetic recombination, is induced in meiosis. *Mol. Cell Biol.* **10**: 2379-2389.

Esposito, T., Gianfrancesco, F., Ciccodicola, A., Montanini, L., Mumm, S., D'Urso, M. and Forabosco, A. (1999). A novel pseudoautosomal human gene encodes a putative protein similar to Ac-like transposases. *Hum. Mol. Genet.* **8**: 61-67.

Essen, L. O., Perisic, O., Cheung, R., Katan, M. and Williams, R. L. (1996). Crystal structure of a mammalian phosphoinositide-specific phospholipase C delta. *Nature* **380**: 595-602.

Fan, Q. Q. and Petes, T. D. (1996). Relationship between nuclease-hypersensitive sites and meiotic recombination hot spot activity at the HIS4 locus of Saccharomyces cerevisiae. *Mol. Cell Biol.* **16**: 2037-2043.

Fan, Q. Q., Xu, F., White, M. A. and Petes, T. D. (1997). Competition between adjacent meiotic recombination hotspots in the yeast *Saccharomyces cerevisiae*. *Genetics* **145**: 661-670.

Ferguson, K. M., Lemmon, M. A., Sigler, P. B. and Schlessinger, J. (1995). Scratching the surface with the PH domain. *Nat. Struct. Biol.* **2**: 715-718.

Feunteun, J. (1998). Breast cancer and genetic instability: the molecules behind the scenes. *Mol. Med. Today* **4**: 263-267.

Fields, C., Adams, M. D., White, O. and Venter, J. C. (1994). How many genes in the human genome? *Nat. Genet.* **7**: 345-346.

Flint, J., Bates, G. P., Clark, K., Dorman, A., Willingham, D., Roe, B. A., Micklem, G., Higgs, D. R. and Louis, E. J. (1997a). Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum. Mol. Genet.* **6**: 1305-1313.

Flint, J., Thomas, K., Micklem, G., Raynham, H., Clark, K., Doggett, N. A., King, A. and Higgs, D. R. (1997b). The relationship between chromosome structure and function at a human telomeric region. *Nat. Genet.* **15**: 252-257.

Fox, M. E., Virgin, J. B., Metzger, J. and Smith, G. R. (1997). Position- and orientation-independent activity of the *Schizosaccharomyces pombe* meiotic recombination hot spot M26. *Proc. Natl. Acad. Sci. USA* **94**: 7446-7451.

Freije, D., Helms, C., Watson, M. S. and Donis-Keller, H. (1992). Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science* **258**: 1784-1787.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J. and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* **69**: 831-843.

Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133**: 693-709.

Fujitani, Y., Mori, S. and Kobayashi, I. (2002). A reaction-diffusion model for interference in meiotic crossing over. *Genetics* **161**: 365-372.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science* **296**: 2225-2229.

Gaudieri, S., Dawkins, R. L., Habara, K., Kulski, J. K. and Gojobori, T. (2000). SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity. *Genome Res.* **10**: 1579-1586.

Gelin, C., Aubrit, F., Phalipon, A., Raynal, B., Cole, S., Kaczorek, M. and Bernard, A. (1989). The E2 antigen, a 32 kd glycoprotein involved in T-cell adhesion processes, is the MIC2 gene product. *Embo J.* **8**: 3253-3259.

Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O. and Petes, T. D. (2000). Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **97**: 11383-11390.

Gianfrancesco, F., Esposito, T., Montanini, L., Ciccodicola, A., Mumm, S., Mazzarella, R., Rao, E., Giglio, S., Rappold, G. and Forabosco, A. (1998). A novel pseudoautosomal gene encoding a putative GTP-binding protein resides in the vicinity of the Xp/Yp telomere. *Hum. Mol. Genet.* **7**: 407-414.

Gianfrancesco, F., Sanges, R., Esposito, T., Tempesta, S., Rao, E., Rappold, G., Archidiacono, N., Graves, J. A., Forabosco, A. and D'Urso, M. (2001). Differential divergence of three human pseudoautosomal genes and their mouse homologs: implications for sex chromosome evolution. *Genome Res.* **11**: 2095-2100.

Gilbertson, L. A. and Stahl, F. W. (1996). A test of the double-strand break repair model for meiotic recombination in *Saccharomyces cerevisiae*. *Genetics* **144**: 27-41.

Goldman, A. S. and Hultén, M. A. (1993). Meiotic analysis by FISH of a human male 46,XY,t(15;20)(q11.2;q11.2) translocation heterozygote: quadrivalent configuration, orientation and first meiotic segregation. *Chromosoma* **102**: 102-111.

Goldman, A. S. and Lichten, M. (2000). Restriction of ectopic recombination by interhomolog interactions during *Saccharomyces cerevisiae* meiosis. *Proc. Natl. Acad. Sci. USA* **97**: 9537-9542.

Goodfellow, P. J., Darling, S. M., Thomas, N. S. and Goodfellow, P. N. (1986). A pseudoautosomal gene in man. *Science* **234**: 740-743.

Gough, N. M., Gearing, D. P., Nicola, N. A., Baker, E., Pritchard, M., Callen, D. F. and Sutherland, G. R. (1990). Localization of the human GM-CSF receptor gene to the X-Y pseudoautosomal region. *Nature* **345**: 734-736.

Graves, J. A., Wakefield, M. J. and Toder, R. (1998). The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Hum. Mol. Genet.* **7**: 1991-1996.

Gray, I. C., Campbell, D. A. and Spurr, N. K. (2000). Single nucleotide polymorphisms as tools in human genetics. *Hum. Mol. Genet.* **9**: 2403-2408.

Grelon, M., Vezon, D., Gendrot, G. and Pelletier, G. (2001). *AtSPO11-1* is necessary for efficient meiotic recombination in plants. *Embo J* **20**: 589-600.

Griffith, O. H. and Ryan, M. (1999). Bacterial phosphatidylinositol-specific phospholipase C: structure, function, and interaction with lipids. *Biochim. Biophys. Acta* **1441**: 237-254.

Grushcow, J. M., Holzen, T. M., Park, K. J., Weinert, T., Lichten, M. and Bishop, D. K. (1999). *Saccharomyces cerevisiae* checkpoint genes MEC1, RAD17 and RAD24 are required for normal meiotic recombination partner choice. *Genetics* **153**: 607-620.

Guacci, V., Koshland, D. and Strunnikov, A. (1997). A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in *S. cerevisiae*. *Cell* **91**: 47-57.

Guillon, H. and de Massy, B. (2002). An initiation site for meiotic crossing-over and gene conversion in the mouse. *Nat. Genet.* **32** 296-299.

Guo, M. and Mount, S. M. (1995). Localization of sequences required for size-specific splicing of a small *Drosophila* intron *in vitro*. *J. Mol. Biol.* **253**: 426-437.

Gyapay, G., Morissette, J., Vignal, A., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M. and Weissenbach, J. (1994). The 1993-94 Genethon human genetic linkage map. *Nat. Genet.* **7**: 246-339.

Haber, J. E. (2000). Partners and pathwaysrepairing a double-strand break. *Trends Genet.* **16**: 259-264.

Hahn, J. H., Kim, M. K., Choi, E. Y., Kim, S. H., Sohn, H. W., Ham, D. I., Chung, D. H., Kim, T. J., Lee, W. J., Park, C. K., Ree, H. J. and Park, S. H. (1997). CD99 (MIC2) regulates the LFA-1/ICAM-1-mediated adhesion of lymphocytes, and its gene encodes both positive and negative regulators of cellular adhesion. *J. Immunol.* **159**: 2250-2258.

Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., Cooper, R., Lipshutz, R. and Chakravarti, A. (1999). Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**: 239-247.

Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone, P., Echols, N., Johnson, T. and Gerstein, M. (2002). Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.* **12**: 272-280.

Hassold, T. J., Sherman, S. L., Pettay, D., Page, D. C. and Jacobs, P. A. (1991). XY chromosome nondisjunction in man is associated with diminished recombination in the pseudoautosomal region. *Am. J. Hum. Genet.* **49**: 253-260.

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics* **117**: 331-341.

Henke, A. and Rappold, G. (1993). PA2.1 detects a TaqI polymorphism in the pseudoautosomal region. *Hum. Mol. Genet.* **2**: 339.

Herickhoff, L., Stack, S. and Sherman, J. (1993). The relationship between synapsis, recombination nodules and chiasmata in tomato translocation heterozygotes. *Heredity* **71**: 373-385.

Higgs, D. R., Vickers, M. A., Wilkie, A. O., Pretorius, I. M., Jarman, A. P. and Weatherall, D. J. (1989). A review of the molecular genetics of the human alpha-globin gene cluster. *Blood* **73**: 1081-1104.

Hofmann, S. L. and Majerus, P. W. (1982). Identification and properties of two distinct phosphatidylinositol-specific phospholipase C enzymes from sheep seminal vesicular glands. *J. Biol. Chem.* **257**: 6461-6469.

Hogenesch, J. B., Ching, K. A., Batalov, S., Su, A. I., Walker, J. R., Zhou, Y., Kay, S. A., Schultz, P. G. and Cooke, M. P. (2001). A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* **106**: 413-415.

Hollingsworth, N. M. and Byers, B. (1989). HOP1: a yeast meiotic pairing gene. *Genetics* **121**: 445-462.

Hollingsworth, N. M., Goetsch, L. and Byers, B. (1990). The HOP1 gene encodes a meiosis-specific component of yeast chromosomes. *Cell* **61**: 73-84.

Hollingsworth, N. M., Ponte, L. and Halsey, C. (1995). MSH5, a novel MutS homolog, facilitates meiotic reciprocal recombination between homologs in *Saccharomyces cerevisiae* but not mismatch repair. *Genes Dev.*. **9**: 1728-1739.

Holmquist, G. P. (1992). Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.* **51**(1): 17-37.

Hsu, K., Chang, D. Y. and Maraia, R. J. (1995). Human signal recognition particle (SRP) Alu-associated protein also binds Alu interspersed repeat sequence RNAs. Characterization of human SRP9. *J. Biol. Chem.* **270**: 10179-10186.

Hubert, R., MacDonald, M., Gusella, J. and Arnheim, N. (1994). High resolution localization of recombination hot spots using sperm typing. *Nat. Genet.* **7**: 420-424.

Hultén, M. (1974). Chiasma distribution at diakinesis in the normal human male. *Hereditas* **76**: 55-78.

Hultén, M., Lawrie, N. M. and Laurie, D. A. (1990). Chiasma-based genetic maps of chromosome 21. *Am. J. Med. Genet. Suppl.* **7**: 148-154.

Hunter, N. and Borts, R. H. (1997). Mlh1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis. *Genes Dev.* **11**: 1573-1582.

Hunter, N. and Kleckner, N. (2001). The single-end invasion: an asymmetric intermediate at the double-strand break to double-holliday junction transition of meiotic recombination. *Cell* **106**: 59-70.

Inglehearn, C. F. and Cooke, H. J. (1990). A VNTR immediately adjacent to the human pseudoautosomal telomere. *Nucleic Acids Res.* **18**: 471-476.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

Jansen, R. P. (2001). mRNA localization: message on the move. *Nat. Rev. Mol. Cell Biol.* **2**: 247-256.

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. (1997a). Spontaneous and induced minisatellite instability. *Electrophoresis* **18**: 1501-1511.

Jeffreys, A. J., Kauppi, L. and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* **29**: 217-222.

Jeffreys, A. J., MacLeod, A., Tamaki, K., Neil, D. L. and Monckton, D. G. (1991). 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. (1997b). Somatic mutation processes at a human minisatellite. *Hum. Mol. Genet.* **6**(1): 129-32; 134-6.

Jeffreys, A. J. and Neumann, R. (2002). Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat. Genet.* **31**: 267-271.

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., Ritchie, A. and Neumann, R. (2000). High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum. Mol. Genet.* **9**: 725-733.

Jeffreys, A. J., Royle, N. J., Wilson, V. and Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**: 278-281.

Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, 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. and Thein, S. L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature* **316**: 76-79.

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.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G. and Todd, J. A. (2001). Haplotype tagging for the identification of common disease genes. *Nat. Genet.* **29**: 233-237.

Johzuka, K. and Ogawa, H. (1995). Interaction of Mre11 and Rad50: two proteins required for DNA repair and meiosis-specific double-strand break formation in *Saccharomyces cerevisiae*. *Genetics* **139**(4): 1521-1532.

Jones, G. H. (1984). The control of chiasma distribution. *Symp. Soc. Exp. Biol.* **38**: 293-320.

Jorde, L. B. (2000). Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**: 1435-1444.

Jurka, J. and Zuckerkandl, E. (1991). Free left arms as precursor molecules in the evolution of Alu sequences. *J. Mol. Evol.* **33**: 49-56.

Kaback, D. B., Guacci, V., Barber, D. and Mahon, J. W. (1992). Chromosome size-dependent control of meiotic recombination. *Science* **256**: 228-232.

Kaback, D. B., Steensma, H. Y. and de Jonge, P. (1989). Enhanced meiotic recombination on the smallest chromosome of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **86**: 3694-3698.

Kadyk, L. C. and Hartwell, L. H. (1992). Sister chromatids are preferred over homologs as substrates for recombinational repair in *Saccharomyces cerevisiae*. *Genetics* **132**: 387-402.

Kalsi, G., Curtis, D., Brynjolfsson, J., Butler, R., Sharma, T., Murphy, P., Read, T., Petursson, H. and Gurling, H. M. (1995). Investigation by linkage analysis of the XY pseudoautosomal region in the genetic susceptibility to schizophrenia. *Br. J. Psychiatry* **167**: 390-393.

Kapitonov, V. and Jurka, J. (1996). The age of Alu subfamilies. *J. Mol. Evol.* **42**: 59-65.

Kauppi, L., Sajantila, A. and Jeffreys, A. J. (2003). Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.* **12**: 33-40.

Keeney, S. (2001). Mechanism and control of meiotic recombination initiation. *Curr. Top. Dev. Biol.* **52**: 1-53.

Keeney, S., Giroux, C. N. and Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* **88**: 375-384.

Kim, S. H., Choi, E. Y., Shin, Y. K., Kim, T. J., Chung, D. H., Chang, S. I., Kim, N. K. and Park, S. H. (1998). Generation of cells with Hodgkin's and Reed-Sternberg phenotype through downregulation of CD99 (Mic2). *Blood* **92**: 4287-4295.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**: 624-626.

Kimura, M. (1985). *Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, U.K.

Kimura, M. and Crow, J. F. (1964). The Number of Alleles That Can Be Maintained in a Finite Population. *Genetics* **49**: 725-738.

Kirkpatrick, D. T., Fan, Q. and Petes, T. D. (1999a). Maximal stimulation of meiotic recombination by a yeast transcription factor requires the transcription activation domain and a DNA-binding domain. *Genetics* **152**: 101-115.

Kirkpatrick, D. T., Wang, Y. H., Dominska, M., Griffith, J. D. and Petes, T. D. (1999b). Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes. *Mol. Cell Biol.* **19**: 7661-7771.

Kjeldgaard, M., Nyborg, J. and Clark, B. F. (1996). The GTP binding motif: variations on a theme. *Faseb J.* **10**: 1347-1368.

Klein, F., Mahr, P., Galova, M., Buonomo, S. B., Michaelis, C., Nairz, K. and Nasmyth, K. (1999). A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. *Cell* **98**: 91-103.

Klink, A., Wapenaar, M., van Ommen, G. J. and Rappold, G. (1993). AK1 detects a VNTR locus in the pseudoautosomal region. *Hum. Mol. Genet.* **2**: 339.

Kohli, J. and Bahler, J. (1994). Homologous recombination in fission yeast: absence of crossover interference and synaptonemal complex. *Experientia* **50**: 295-306.

Koller, P. C. and Darlington, C. D. (1934). The genetical and mechanical properties of the sex chromosomes. 1. *Rattus norvegicus*. *J. Genet.* **29**: 159.

Kon, N., Krawchuk, M. D., Warren, B. G., Smith, G. R. and Wahls, W. P. (1997). Transcription factor Mts1/Mts2 (Atf1/Pcr1, Gad7/Pcr1) activates the M26 meiotic recombination hotspot in *Schizosaccharomyces pombe*. *Proc. Natl. Acad. Sci. USA* **94**: 13765-13770.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R. and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nat. Genet.* **31**: 241-247.

Kowalczykowski, S. C. (2000). Initiation of genetic recombination and recombination-dependent replication. *Trends Biochem. Sci.* **25**: 156-165.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22**: 139-144.

Laan, M. and Paabo, S. (1997). Demographic history and linkage disequilibrium in human populations. *Nat. Genet.* **17**: 435-438.

Lamb, N. E., Freeman, S. B., Savage-Austin, A., Pettay, D., Taft, L., Hersey, J., Gu, Y., Shen, J., Saker, D., May, K. M., Avramopoulos, D., Petersen, M. B., Hallberg, A., Mikkelsen, M., Hassold, T. J. and Sherman, S. L. (1996). Susceptible chiasmate configurations of chromosome 21 predispose to non-disjunction in both maternal meiosis I and meiosis II. *Nat. Genet.* **14**: 400-405.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265**: 2037-2048.

Landry, J. R., Medstrand, P. and Mager, D. L. (2001). Repetitive elements in the 5' untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics* **76**: 110-116.

Latos-Bielenska, A. and Vogel, W. (1990). Frequency and distribution of chiasmata in Syrian hamster spermatocytes studied by the BrdU antibody technique. *Chromosoma* **99**: 267-272.

Lau, Y. F. and Zhang, J. (2000). Expression analysis of thirty one Y chromosome genes in human prostate cancer. *Mol. Carcinog.* **27**: 308-321.

Laurie, D. A., Hultén, M. and Jones, G. H. (1981). Chiasma frequency and distribution in a sample of human males: chromosomes 1, 2, and 9. *Cytogenet. Cell Genet.* **31**: 153-166.

Laurie, D. A. and Hultén, M. A. (1985). Further studies on chiasma distribution and interference in the human male. *Ann. Hum. Genet.* **49**: 203-214.

Lawrie, N. M., Tease, C. and Hultén, M. A. (1995). Chiasma frequency, distribution and interference maps of mouse autosomes. *Chromosoma* **104**: 308-314.

Le Roux, M. G., Pascal, O., Lostanlen, A., Berard, I., Vergnaud, O. and Moisan, J. P. (1994). VNTR at the DXYS14 locus. *Hum. Mol. Genet.* **3**: 389.

Lee, J. Y. and Orr-Weaver, T. L. (2001). The molecular basis of sister-chromatid cohesion. *Annu. Rev. Cell Dev. Biol.* **17**: 753-777.

Lemmon, M. A., Ferguson, K. M., O'Brien, R., Sigler, P. B. and Schlessinger, J. (1995). Specific and high-affinity binding of inositol phosphates to an isolated pleckstrin homology domain. *Proc. Natl. Acad. Sci. USA* **92**: 10472-10476.

Lewontin, R. C. (1988). On measures of gametic disequilibrium. *Genetics* **120**: 849-852.

Li, H. H., Gyllensten, U. B., Cui, X. F., Saiki, R. K., Erlich, H. A. and Arnheim, N. (1988). Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* **335**: 414-417.

Li, L. and Hamer, D. H. (1995). Recombination and allelic association in the Xq/Yq homology region. *Hum. Mol. Genet.* **4**: 2013-2016.

Li, W. H. and Sadler, L. A. (1991). Low nucleotide diversity in man. *Genetics* **129**: 513-523.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L. and Quackenbush, J. (2000). Gene index analysis of the human genome estimates approximately 120,000 genes. *Nat. Genet.* **25**: 239-240.

Lichten, M. and Goldman, A. S. (1995). Meiotic recombination hotspots. *Annu. Rev. Genet.* **29**: 423-444.

Lichten, M. J. and Fox, M. S. (1983). Detection of non-homology-containing heteroduplex molecules. *Nucleic Acids Res.* **11**: 3959-3971.

Lien, S., Szyda, J., Schechinger, B., Rappold, G. and Arnheim, N. (2000). Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* **66**: 557-566.

Lin, Y. and Smith, G. R. (1994). Transient, meiosis-induced expression of the rec6 and rec12 genes of *Schizosaccharomyces pombe*. *Genetics* **136**: 769-779.

Liu, J., Wu, T. C. and Lichten, M. (1995). The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate. *Embo J.* **14**: 4599-4608.

Loidl, J., Scherthan, H., Den Dunnen, J. T. and Klein, F. (1995). Morphology of a human-derived YAC in yeast meiosis. *Chromosoma* **104**: 183-188.

Loupart, M. L., Adams, S., Armour, J. A., Walker, R., Brammar, W. and Varley, J. (1995). Loss of heterozygosity on the X chromosome in human breast cancer. *Genes Chromosomes Cancer* **13**: 229-238.

Lydall, D., Nikolsky, Y., Bishop, D. K. and Weinert, T. (1996). A meiotic recombination checkpoint controlled by mitotic checkpoint genes. *Nature* **383**: 840-843.

Maguire, M. P., Paredes, A. M. and Riess, R. W. (1991). The desynaptic mutant of maize as a combined defect of synaptonemal complex and chiasma maintenance. *Genome* **34**: 879-887.

Majewski, J. and Ott, J. (2000). GT repeats are associated with recombination on human chromosome 22. *Genome Res.* **10**: 1108-1114.

Malone, R. E., Kim, S., Bullard, S. A., Lundquist, S., Hutchings-Crow, L., Cramton, S., Lutfiyya, L. and Lee, J. (1994). Analysis of a recombination hotspot for gene conversion occurring at the HIS2 gene of *Saccharomyces cerevisiae*. *Genetics* **137**: 5-18.

Mao-Draayer, Y., Galbraith, A. M., Pittman, D. L., Cool, M. and Malone, R. E. (1996). Analysis of meiotic recombination pathways in the yeast *Saccharomyces cerevisiae*. *Genetics* **144**: 71-86.

Maudlin, I. and Evans, E. P. (1980). Chiasma distribution in mouse oocytes during diakinesis. *Chromosoma* **80**: 49-56.

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.

May, C. A., Shone, A. C., Kalaydjieva, L., Sajantila, A. and Jeffreys, A. J. (2002). Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nat. Genet.* **31**: 272-275.

McCullough, A. J. and Berget, S. M. (1997). G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell. Biol.* **17**: 4562-4571.

McKee, A. H. and Kleckner, N. (1997). A general method for identifying recessive diploid-specific mutations in *Saccharomyces cerevisiae*, its application to the isolation of mutants blocked at intermediate stages of meiotic prophase and characterization of a new gene SAE2. *Genetics* **146**: 797-816.

McKim, K. S. and Hayashi-Hagihara, A. (1998). mei-W68 in Drosophila melanogaster encodes a Spo11 homolog: evidence that the mechanism for initiating meiotic recombination is conserved. *Genes Dev.* **12**: 2932-2942.

Michaelis, C., Ciosk, R. and Nasmyth, K. (1997). Cohesins: chromosomal proteins that prevent premature separation of sister chromatids. *Cell* **91**: 35-45.

Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biol* **3**: 1-10.

Milatovich, A., Kitamura, T., Miyajima, A. and Francke, U. (1993). Gene for the alpha-subunit of the human interleukin-3 receptor (IL3RA) localized to the X-Y pseudoautosomal region. *Am. J. Hum. Genet.* **53**: 1146-1153.

Miyazaki, W. Y. and Orr-Weaver, T. L. (1992). Sister-chromatid misbehavior in *Drosophila* ord mutants. *Genetics* **132**: 1047-1061.

Mizuno, K., Hasemi, T., Ubukata, T., Yamada, T., Lehmann, E., Kohli, J., Watanabe, Y., Iino, Y., Yamamoto, M., Fox, M. E., Smith, G. R., Murofushi, H., Shibata, T. and Ohta, K. (2001). Counteracting regulation of chromatin remodeling at a fission yeast cAMP response element-related recombination hotspot by stress-activated protein kinase, cAMP-dependent kinase and meiosis regulators. *Genetics* **159**: 1467-1478.

Mohrenweiser, H. W., Tsujimoto, S., Gordon, L. and Olsen, A. S. (1998). Regions of sex-specific hypo- and hyper-recombination identified through integration of 180 genetic markers into the metric physical map of human chromosome 19. *Genomics* **47**: 153-162.

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.

Morimoto, T., Loh, P. C., Hirai, T., Asai, K., Kobayashi, K., Moriya, S. and Ogasawara, N. (2002). Six GTP-binding proteins of the Era/Obg family are essential for cell growth in Bacillus subtilis. *Microbiology* **148**: 3539-3552.

Mortimer, R. K. and Fogel, S. (1969). *Genetical interference and gene conversion.* Mechanisms in Recombination pp. 263-275. *Ed. Grell, R. F.* Plenum Press, New York, N.Y.

Morton, N. E., Lindsten, J., Iselius, L. and Yee, S. (1982). Data and theory for a revised chiasma map of man. *Hum. Genet.* **62**(3): 266-270.

Mount, S. M. (1982). A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**: 459-472.

Mount, S. M. (2000). Genomic sequence, splicing, and gene annotation. *Am. J. Hum. Genet.* **67**: 788-792.

Nachman, M. W. (2001). Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet.* **17**: 481-485.

Nachman, M. W., Bauer, V. L., Crowell, S. L. and Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. *Genetics* **150**: 1133-1141.

Nairz, K. and Klein, F. (1997). *mre11S* - a yeast mutation that blocks double-strand-break processing and permits nonhomologous synapsis in meiosis. *Genes Dev.* **11**: 2272-2290.

Nakagawa, T. and Kolodner, R. D. (2002). *Saccharomyces cerevisiae* Mer3 is a DNA helicase involved in meiotic crossing over. *Mol. Cell Biol.* **22**: 3281-3291.

Nakagawa, T. and Ogawa, H. (1997). Involvement of the MRE2 gene of yeast in formation of meiosis-specific double-strand breaks and crossover recombination through RNA splicing. *Genes Cells* **2**: 65-79.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. and et al. (1987). Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**: 1616-1622.

Nakashima, S., Banno, Y., Watanabe, T., Nakamura, Y., Mizutani, T., Sakai, H., Zhao, Y., Sugimoto, Y. and Nozawa, Y. (1995). Deletion and site-directed mutagenesis of EF-hand domain of phospholipase C-delta 1: effects on its activity. *Biochem. Biophys. Res. Commun.* **211**: 365-369.

Nasmyth, K., Peters, J. M. and Uhlmann, F. (2000). Splitting the chromosome: cutting the ties that bind sister chromatids. *Science* **288**: 1379-1385.

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA* **76**: 5269-5273.

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.

New, J. H., Sugiyama, T., Zaitseva, E. and Kowalczykowski, S. C. (1998). Rad52 protein stimulates DNA strand exchange by Rad51 and replication protein A. *Nature* **391**: 407-410.

Nichol, K. and Pearson, C. E. (2002). CpG methylation modifies the genetic stability of cloned repeat sequences. *Genome Res.* **12**: 1246-1256.

Nicolas, A., Treco, D., Schultes, N. P. and Szostak, J. W. (1989). An initiation site for meiotic gene conversion in the yeast *Saccharomyces cerevisiae*. *Nature* **338**: 35-39.

NIH/CEPH Collaborative Mapping Group (1992). A comprehensive genetic linkage map of the human genome. *Science* **258**: 67-86.

Nilsson, N.-O. and Pelger, S. (1991). The relationship between natural variation in chiasma frequencies and recombination frequencies in barley. *Hereditas* **115**: 121-126.

Nishizuka, Y. (1986). Studies and perspectives of protein kinase C. *Science* **233**: 305-312.

Ohta, K., Nicolas, A., Furuse, M., Nabetani, A., Ogawa, H. and Shibata, T. (1998). Mutations in the MRE11, RAD50, XRS2, and MRE2 genes alter chromatin configuration at meiotic DNA double-stranded break sites in premeiotic and meiotic cells. *Proc. Natl. Acad. Sci. USA* **95**: 646-651.

Ohta, K., Shibata, T. and Nicolas, A. (1994). Changes in chromatin structure at recombination initiation sites during yeast meiosis. *Embo J.* **13**: 5754-5763.

Ohta, K., Wu, T. C., Lichten, M. and Shibata, T. (1999). Competitive inactivation of a double-strand DNA break site involves parallel suppression of meiosis-induced changes in chromatin configuration. *Nucleic Acids Res.* **27**: 2175-2180.

Ohta, T. and Kimura, M. (1971). Behavior of neutral mutants influenced by asociated overdominant loci in finite populations. *Genetics* **69**: 247-260.

Orita, M., Iwahana, H., Kanazawa, H., Hayashi, K. and Sekiya, T. (1989). Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc. Natl. Acad. Sci. USA* **86**: 2766-2770.

Paabo, S. (2003). The mosaic that is our genome. *Nature* **421**: 409-412.

Page, D. C., Bieker, K., Brown, L. G., Hinton, S., Leppert, M., Lalouel, J. M., Lathrop, M., Nystrom-Lahti, M., de la Chapelle, A. and White, R. (1987). Linkage, physical mapping, and DNA sequence analysis of pseudoautosomal loci on the human X and Y chromosomes. *Genomics* **1**: 243-256.

Paques, F. and Haber, J. E. (1999). Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae. Microbiol. Mol. Biol. Rev.* **63**: 349-404.

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.

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stern, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P. and Cox, D. R. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**: 1719-1723.

Pecina, A., Smith, K. N., Mezard, C., Murakami, H., Ohta, K. and Nicolas, A. (2002). Targeted stimulation of meiotic recombination. *Cell* **111**: 173-184.

Pelttari, J., Hoja, M. R., Yuan, L., Liu, J. G., Brundell, E., Moens, P., Santucci-Darmanin, S., Jessberger, R., Barbero, J. L., Heyting, C. and Hoog, C. (2001). A meiotic chromosomal core consisting of cohesin complex proteins recruits DNA recombination proteins and promotes synapsis in the absence of an axial element in mammalian meiotic cells. *Mol. Cell Biol.* **21**: 5667-5677.

Pentao, L., Wise, C. A., Chinault, A. C., Patel, P. I. and Lupski, J. R. (1992). Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nat. Genet.* **2**: 292-300.

Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* **2**: 360-369.

Petit, C., Levilliers, J., Rouyer, F., Simmler, M. C., Herouin, E. and Weissenbach, J. (1990). Isolation of sequences from Xp22.3 and deletion mapping using sex chromosome rearrangements from human X-Y interchange sex reversals. *Genomics* **6**: 651-658.

Petit, C., Levilliers, J. and Weissenbach, J. (1988). Physical mapping of the human pseudo-autosomal region; comparison with genetic linkage map. *Embo J.* **7**: 2369-2376.

Petrini, J. H., Walsh, M. E., DiMare, C., Chen, X. N., Korenberg, J. R. and Weaver, D. T. (1995). Isolation and characterization of the human MRE11 homologue. *Genomics* **29**: 80-86.

Plug, A. W., Peters, A. H., Keegan, K. S., Hoekstra, M. F., de Boer, P. and Ashley, T. (1998). Changes in protein composition of meiotic nodules during mammalian meiosis. *J. Cell Sci.* **111**: 413-423.

Ponnudurai, R. (1996). Failure to support a pseudoautosomal locus for schizophrenia. *Psychiatry Res.* **62**: 281-284.

Porter, S. E., White, M. A. and Petes, T. D. (1993). Genetic evidence that the meiotic recombination hotspot at the HIS4 locus of *Saccharomyces cerevisiae* does not represent a site for a symmetrically processed double-strand break. *Genetics* **134**: 5-19.

Prinz, S., Amon, A. and Klein, F. (1997). Isolation of COM1, a new gene required to complete meiotic double-strand break-induced recombination in *Saccharomyces cerevisiae*. *Genetics* **146**: 781-795.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**: 1-14.

Quentin, Y. (1992). Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. *Nucleic Acids Res.* **20**: 487-493.

Rao, E., Weiss, B., Fukami, M., Rump, A., Niesler, B., Mertz, A., Muroya, K., Binder, G., Kirsch, S., Winkelmann, M., Nordsiek, G., Heinrich, U., Breuning, M. H., Ranke, M. B., Rosenthal, A., Ogata, T. and Rappold, G. A. (1997). Pseudoautosomal deletions encompassing a novel homeobox gene cause growth failure in idiopathic short stature and Turner syndrome. *Nat. Genet.* **16**: 54-63.

Rappold, G., Willson, T. A., Henke, A. and Gough, N. M. (1992). Arrangement and localization of the human GM-CSF receptor alpha chain gene CSF2RA within the X-Y pseudoautosomal region. *Genomics* **14**: 455-461.

Rappold, G. A. (1993). The pseudoautosomal regions of the human sex chromosomes. *Hum. Genet.* **92**: 315-324.

Rappold, G. A., Klink, A., Weiss, B. and Fischer, C. (1994). Double crossover in the human Xp/Yp pseudoautosomal region and its bearing on interference. *Hum. Mol. Genet.* **3**: 1337-1340.

Rattray, A. J. and Symington, L. S. (1993). Stimulation of meiotic recombination in yeast by an ARS element. *Genetics* **134**: 175-188.

Raymond, W. E. and Kleckner, N. (1993). RAD50 protein of *S.cerevisiae* exhibits ATP-dependent DNA binding. *Nucleic Acids Res.* **21**: 3851-3856.

Rebecchi, M. J. and Pentyala, S. N. (2000). Structure, function, and control of phosphoinositide-specific phospholipase C. *Physiol. Rev.* **80**: 1291-1335.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R. and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature* **411**: 199-204.

Reich, D. E., Schaffner, S. F., Daly, M. J., McVean, G., Mullikin, J. C., Higgins, J. M., Richter, D. J., Lander, E. S. and Altshuler, D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**: 135-142.

Renauld, J. C., Kermouni, A., Vink, A., Louahed, J. and Van Snick, J. (1995). Interleukin-9 and its receptor: involvement in mast cell differentiation and T cell oncogenesis. *J Leukoc. Biol.* **57**: 353-360.

Rhee, S. G. and Bae, Y. S. (1997). Regulation of phosphoinositide-specific phospholipase C isozymes. *J. Biol. Chem.* **272**: 15045-15048.

Rhee, S. G., Suh, P. G., Ryu, S. H. and Lee, S. Y. (1989). Studies of inositol phospholipid-specific phospholipase C. *Science* **244**: 546-550.

Ried, K., Rao, E., Schiebel, K. and Rappold, G. A. (1998). Gene duplications as a recurrent theme in the evolution of the human pseudoautosomal region 1: isolation of the gene ASMTL. *Hum. Mol. Genet.* **7**: 1771-1778.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science* **273**: 1516-1517.

Rivas, C. I., Vera, J. C., Delgado-Lopez, F., Heaney, M. L., Guaiquil, V. H., Zhang, R. H., Scher, H. I., Concha, II, Nualart, F., Cordon-Cardo, C. and Golde, D. W. (1998). Expression of granulocyte-macrophage colony-stimulating factor receptors in human prostate cancer. *Blood* **91**: 1037-1043.

Rockmill, B. and Roeder, G. S. (1988). RED1: a yeast gene required for the segregation of chromosomes during the reductional division of meiosis. *Proc. Natl. Acad. Sci. USA* **85**: 6057-6061.

Rockmill, B., Sym, M., Scherthan, H. and Roeder, G. S. (1995). Roles for two RecA homologs in promoting meiotic chromosome synapsis. *Genes Dev.* **9**: 2684-2695.

Roeder, G. S. (1997). Meiotic chromosomes: it takes two to tango. *Genes Dev.* **11**: 2600-2621.

Romanienko, P. J. and Camerini-Otero, R. D. (1999). Cloning, characterization, and localization of mouse and human SPO11. *Genomics* **61**: 156-169.

Ross, L. O., Maxfield, R. and Dawson, D. (1996). Exchanges are not equally able to enhance meiotic chromosome segregation in yeast. *Proc. Natl. Acad. Sci. USA* **93**: 4979-4083.

Ross-Macdonald, P. and Roeder, G. S. (1994). Mutation of a meiosis-specific MutS homolog decreases crossing over but not mismatch correction. *Cell* **79**: 1069-1080.

Rouyer, F., de la Chapelle, A., Andersson, M. and Weissenbach, J. (1990). An interspersed repeated sequence specific for human subtelomeric regions. *Embo J.* **9**: 505-514.

Rouyer, F., Simmler, M. C., Johnsson, C., Vergnaud, G., Cooke, H. J. and Weissenbach, J. (1986a). A gradient of sex linkage in the pseudoautosomal region of the human sex chromosomes. *Nature* **319**: 291-295.

Rouyer, F., Simmler, M. C., Vergnaud, G., Johnsson, C., Levilliers, J., Petit, C. and Weissenbach, J. (1986b). The pseudoautosomal region of the human sex chromosomes. *Cold Spring Harb. Symp. Quant. Biol.* **51**: 221-228.

Roy, A. M., Carroll, M. L., Nguyen, S. V., Salem, A. H., Oldridge, M., Wilkie, A. O., Batzer, M. A. and Deininger, P. L. (2000). Potential gene conversion and source genes for recently integrated Alu elements. *Genome Res.* **10**: 1485-1495.

Royle, N. J. (1995). The proterminal regions and telomeres of human chromosomes. *Adv. Genet.* **32**: 273-315.

Royle, N. J., Clarkson, R. E., Wong, Z. and Jeffreys, A. J. (1988). Clustering of hypervariable minisatellites in the proterminal regions of human autosomes. *Genomics* **3**: 352-360.

Royle, N. J., Hill, M. C. and Jeffreys, A. J. (1992). Isolation of telomere junction fragments by anchored polymerase chain reaction. *Proc. R. Soc. Lond. B. Biol. Sci.* **247**: 57-67.

Ryu, S. H., Cho, K. S., Lee, K. Y., Suh, P. G. and Rhee, S. G. (1986). Two forms of phosphatidylinositol-specific phospholipase C from bovine brain. *Biochem. Biophys. Res. Commun.* **141**: 137-144.

Saadallah, N. and Hultén, M. (1983). Chiasma distribution, genetic lengths, and recombination fractions: a comparison between chromosomes 15 and 16. *J. Med. Genet.* **20**: 290-299.

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.

Sambrook, J., Fritsch, E.F. and Maniatis T. (1989). *Molecular Cloning, A Laboratory Manual*, 2nd edition. Cold Spring Harbour Laboratory Press, New York.

Sambrook, J. and Russel, D. W. (2001). *Molecular Cloning, A Laboratory Manual* 3rd edition. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, N.Y.

Scherthan, H., Bahler, J. and Kohli, J. (1994). Dynamics of chromosome organization and pairing during meiotic prophase in fission yeast. *J. Cell Biol.* **127**: 273-285.

Scherthan, H., Weich, S., Schwegler, H., Heyting, C., Harle, M. and Cremer, T. (1996). Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. *J. Cell Biol.* **134**: 1109-1125.

Schiebel, K., Meder, J., Rump, A., Rosenthal, A., Winkelmann, M., Fischer, C., Bonk, T., Humeny, A. and Rappold, G. (2000). Elevated DNA sequence diversity in the genomic region of the phosphatase PPP2R3L gene in the human pseudoautosomal region. *Cytogenet. Cell Genet.* **91**: 224-230.

Schiebel, K., Weiss, B., Wohrle, D. and Rappold, G. (1993). A human pseudoautosomal gene, ADP/ATP translocase, escapes X-inactivation whereas a homologue on Xq is subject to X-inactivation. *Nat. Genet.* **3**: 82-87.

Schuchert, P., Langsford, M., Kaslin, E. and Kohli, J. (1991). A specific DNA sequence is required for high frequency of recombination in the ade6 gene of fission yeast. *Embo J.* **10**: 2157-2163.

Schwacha, A. and Kleckner, N. (1994). Identification of joint molecules that form frequently between homologs but rarely between sister chromatids during yeast meiosis. *Cell* **76**: 51-63.

Schwacha, A. and Kleckner, N. (1995). Identification of double Holliday junctions as intermediates in meiotic recombination. *Cell* **83**: 783-791.

Schwacha, A. and Kleckner, N. (1997). Interhomolog bias during meiotic recombination: meiotic functions promote a highly differentiated interhomolog-only pathway. *Cell* **90**: 1123-1135.

Sears, D. D., Hegemann, J. H. and Hieter, P. (1992). Meiotic recombination and segregation of human-derived artificial chromosomes in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* **89**: 5296-5300.

Sherman, J. D., Herickhoff, L. A. and Stack, S. M. (1992). Silver staining two types of meiotic nodules. *Genome* **35**: 907-915.

Shi, Q., Spriggs, E., Field, L. L., Ko, E., Barclay, L. and Martin, R. H. (2001). Single sperm typing demonstrates that reduced recombination is associated with the production of aneuploid 24,XY human sperm. *Am. J. Med. Genet.* **99**: 34-38.

Shinohara, A., Ogawa, H., Matsuda, Y., Ushio, N., Ikeo, K. and Ogawa, T. (1993). Cloning of human, mouse and fission yeast recombination genes homologous to RAD51 and recA. *Nat. Genet.* **4**: 239-243.

Shinohara, A., Ogawa, H. and Ogawa, T. (1992). Rad51 protein involved in repair and recombination in *S. cerevisiae* is a RecA-like protein. *Cell* **69**: 457-470.

Shinohara, A. and Ogawa, T. (1998a). Stimulation by Rad52 of yeast Rad51-mediated recombination. *Nature* **391**: 404-407.

Shinohara, A. and Ogawa, T. (1999). Rad51/RecA protein families and the associated proteins in eukaryotes. *Mutat. Res.* **435**: 13-21.

Shinohara, A., Shinohara, M., Ohta, T., Matsuda, S. and Ogawa, T. (1998b). Rad52 forms ring structures and co-operates with RPA in single-strand DNA annealing. *Genes Cells* **3**: 145-156.

Simmler, M. C., Johnsson, C., Petit, C., Rouyer, F., Vergnaud, G. and Weissenbach, J. (1987). Two highly polymorphic minisatellites from the pseudoautosomal region of the human sex chromosomes. *Embo J.* **6**: 963-969.

Simmler, M. C., Rouyer, F., Vergnaud, G., Nystrom-Lahti, M., Ngo, K. Y., de la Chapelle, A. and Weissenbach, J. (1985). Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes. *Nature* **317**: 692-697.

Singer, W. D., Brown, H. A. and Sternweis, P. C. (1997). Regulation of eukaryotic phosphatidylinositol-specific phospholipase C and phospholipase D. *Annu. Rev. Biochem.* **66**: 475-509.

Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H. and Deininger, P. (1987). Clustering and subfamily relationships of the Alu family in the human genome. *Mol. Biol. Evol.* **4**: 19-29.

Slim, R., Levilliers, J., Ludecke, H. J., Claussen, U., Nguyen, V. C., Gough, N. M., Horsthemke, B. and Petit, C. (1993). A human pseudoautosomal gene encodes the ANT3 ADP/ATP translocase and escapes X-inactivation. *Genomics* **16**: 26-33.

Smit, A. F. (1996). The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.* **6**: 743-748.

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**: 657-663.

Smit, A. F. and Riggs, A. D. (1996). Tiggers and DNA transposon fossils in the human genome. *Proc. Natl. Acad. Sci. USA* **93**: 1443-1448.

Smith, A. V. and Roeder, G. S. (1997). The yeast Red1 protein localizes to the cores of meiotic chromosomes. *J. Cell Biol.* **136**: 957-967.

Smith, K. N., Penkner, A., Ohta, K., Klein, F. and Nicolas, A. (2001). B-type cyclins CLB5 and CLB6 control the initiation of recombination and synaptonemal complex formation in yeast meiosis. *Curr. Biol.* **11**: 88-97.

Smith, M. J. and Goodfellow, P. N. (1994). MIC2R: a transcribed MIC2-related sequence associated with a CpG island in the human pseudoautosomal region. *Hum. Mol. Genet.* **3**(9): 1575-1582.

Sohn, H. W., Choi, E. Y., Kim, S. H., Lee, I. S., Chung, D. H., Sung, U. A., Hwang, D. H., Cho, S. S., Jun, B. H., Jang, J. J., Chi, J. G. and Park, S. H. (1998). Engagement of CD99 induces apoptosis through a calcineurin-independent pathway in Ewing's sarcoma cells. *Am. J. Pathol.* **153**: 1937-1945.

Solari, A. J. (1980). Synaptosomal complexes and associated structures in microspread human spermatocytes. *Chromosoma* **81**: 315-337.

Stead, J. D., Buard, J., Todd, J. A. and Jeffreys, A. J. (2000). Influence of allele lineage on the role of the insulin minisatellite in susceptibility to type 1 diabetes. *Hum. Mol. Genet.* **9**: 2929-2935.

Stoneking, M. (2001). Single nucleotide polymorphisms. From the evolutionary past. *Nature* **409**: 821-822.

Storlazzi, A., Xu, L., Cao, L. and Kleckner, N. (1995). Crossover and noncrossover recombination during meiosis: timing and pathway relationships. *Proc. Natl. Acad. Sci. USA* **92**: 8512-8516.

Storlazzi, A., Xu, L., Schwacha, A. and Kleckner, N. (1996). Synaptonemal complex (SC) component Zip1 plays a role in meiotic recombination independent of SC polymerization along the chromosomes. *Proc. Natl. Acad. Sci. USA* **93**: 9043-9048.

Stuart, J. J., Egry, L. A., Wong, G. H. and Kaspar, R. L. (2000). The 3' UTR of human MnSOD mRNA hybridizes to a small cytoplasmic RNA and inhibits gene expression. *Biochem. Biophys. Res. Commun.* **274**: 641-648.

Sugiyama, T., New, J. H. and Kowalczykowski, S. C. (1998). DNA annealing by RAD52 protein is stimulated by specific interaction with the complex of replication protein A and single-stranded DNA. *Proc. Natl. Acad. Sci. USA* **95**: 6049-6054.

Sun, H., Treco, D., Schultes, N. P. and Szostak, J. W. (1989). Double-strand breaks at an initiation site for meiotic gene conversion. *Nature* **338**: 87-90.

Sun, H., Treco, D. and Szostak, J. W. (1991). Extensive 3'-overhanging, single-stranded DNA associated with the meiosis-specific double-strand breaks at the ARG4 recombination initiation site. *Cell* **64**: 1155-1161.

Sung, P. (1994). Catalysis of ATP-dependent homologous DNA pairing and strand exchange by yeast RAD51 protein. *Science* **265**: 1241-1243.

Sung, P. (1997). Yeast Rad55 and Rad57 proteins form a heterodimer that functions with replication protein A to promote DNA strand exchange by Rad51 recombinase. *Genes Dev.* **11**: 1111-1121.

Sung, P. and Robberson, D. L. (1995). DNA strand exchange mediated by a RAD51-ssDNA nucleoprotein filament with polarity opposite to that of RecA. *Cell* **82**: 453-461.

Sybenga, J. (1999). What makes homologous chromosomes find each other in meiosis? A review and an hypothesis. *Chromosoma* **108**: 209-219.

Sym, M., Engebrecht, J. A. and Roeder, G. S. (1993). ZIP1 is a synaptonemal complex protein required for meiotic chromosome synapsis. *Cell* **72**: 365-378.

Sym, M. and Roeder, G. S. (1995). Zip1-induced changes in synaptonemal complex structure and polycomplex assembly. *J. Cell Biol.* **128**: 455-466.

Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. and Stahl, F. W. (1983). The double-strand-break repair model for recombination. *Cell* **33**: 25-35.

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. and Kwok, P. Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.* **25**: 324-328.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**(2): 437-60.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**(3): 585-95.

Takahata, N. (1993). Allelic genealogy and human evolution. *Mol. Biol. Evol.* **10**: 2-22.

Tamaki, K., May, C. A., Dubrova, Y. E. and Jeffreys, A. J. (1999). Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7. *Hum. Mol. Genet.* **8**: 879-888.

Taylor, S. J., Chae, H. Z., Rhee, S. G. and Exton, J. H. (1991). Activation of the beta 1 isozyme of phospholipase C by alpha subunits of the Gq class of G proteins. *Nature* **350**: 516-518.

Tease, C., Hartshorne, G. M. and Hultén, M. A. (2002). Patterns of meiotic recombination in human fetal oocytes. *Am. J. Hum. Genet.* **70**: 1469-1479.

The International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Thompson, D. A. and Stahl, F. W. (1999). Genetic control of recombination partner preference in yeast meiosis. Isolation and characterization of mutants elevated for meiotic unequal sister-chromatid recombination. *Genetics* **153**: 621-641.

Tonozuka, Y., Fujio, K., Sugiyama, T., Nosaka, T., Hirai, M. and Kitamura, T. (2001). Molecular cloning of a human novel type I cytokine receptor related to delta1/TSLPR. *Cytogenet. Cell Genet.* **93**: 23-25.

Tsubouchi, H. and Ogawa, H. (1998). A novel mre11 mutation impairs processing of double-strand breaks of DNA during both mitosis and meiosis. *Mol. Cell Biol.* **18**: 260-268.

Turri, M. G., Cuin, K. A. and Porter, A. C. (1995). Characterisation of a novel minisatellite that provides multiple splice donor sites in an interferon-induced transcript. *Nucleic Acids Res.* **23**: 1854-1861.

Umezu, K. and Kolodner, R. D. (1994). Protein interactions in genetic recombination in Escherichia coli. Interactions involving RecO and RecR overcome the inhibition of RecA by single-stranded DNA-binding protein. *J. Biol. Chem.* **269**: 30005-30013.

van der Velden, A. W. and Thomas, A. A. (1999). The role of the 5' untranslated region of an mRNA in translation regulation during development. *Int. J. Biochem. Cell Biol.* **31**: 87-106.

van Endert, P. M., Lopez, M. T., Patel, S. D., Monaco, J. J. and McDevitt, H. O. (1992). Genomic polymorphism, recombination, and linkage disequilibrium in human major histocompatibility complex-encoded antigen-processing genes. *Proc. Natl. Acad. Sci. USA* **89**: 11594-11597.

van Heemst, D. and Heyting, C. (2000). Sister chromatid cohesion and recombination in meiosis. *Chromosoma* **109**: 10-26.

Vazquez Nin, G. H., Flores, E., Echeverria, O. M., Merkert, H., Wettstein, R. and Benavente, R. (1993). Immunocytochemical localization of DNA in synaptonemal complexes of rat and mouse spermatocytes, and of chick oocytes. *Chromosoma* **102**: 457-463.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., *et al.*,[†] (2001) The sequence of the human genome. *Science* **291**: 1304-1351.
[†] *see paper for full list of 273 authors*

Vergnaud, G., Gauguier, D., Schott, J. J., Lepetit, D., Lauthier, V., Mariat, D. and Buard, J. (1993). *Detection, cloning, and distribution of minisatellites in some mammalian genomes.* DNA Fingerprinting: State of the Science pp. 47-57. *Eds. Pena, S. D. J., Chakraborty, R., Epplen, J. T., Jeffreys A. J.* Birkhauser Verlag Basel, Switzerland.

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.

Vermeesch, J. R., Petit, P., Kermouni, A., Renauld, J. C., Van Den Berghe, H. and Marynen, P. (1997). The IL-9 receptor gene, located in the Xq/Yq pseudoautosomal region, has an autosomal origin, escapes X inactivation and is expressed from the Y. *Hum. Mol. Genet.* **6**: 1-8.

Vogel, F. (1972). Non-randomness of base replacement in point mutation. *J. Mol. Evol.* **1**: 334-367.

Wachi, M., Doi, M., Ueda, T., Ueki, M., Tsuritani, K., Nagai, K. and Matsuhashi, M. (1991). Sequence of the downstream flanking region of the shape-determining genes mreBCD of *Escherichia coli. Gene* **106**: 135-136.

Wahls, W. P. (1998). Meiotic recombination hotspots: shaping the genome and insights into hypervariable minisatellite DNA change. *Curr. Top. Dev. Biol.* **37**: 37-75.

Wall, J. D. and Pritchard, J. K. (2003). Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* **73**: 502-515.

Wallace, B. M. and Hultén, M. A. (1985). Meiotic chromosome pairing in the normal human female. *Ann. Hum. Genet.* **49**: 215-226.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., and Lander, E. S. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077-1082.

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R. and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* **71**: 1227-1234.

Wang, Y. H. and Griffith, J. D. (1996). The [(G/C)3NN]n motif: a common DNA repeat that excludes nucleosomes. *Proc. Natl. Acad. Sci. USA* **93**: 8863-8867.

Warren, S. T. (1996). The expanding world of trinucleotide repeats. *Science* **271**: 1374-1375.

Watanabe, Y. and Nurse, P. (1999). Cohesin Rec8 is required for reductional chromosome segregation at meiosis. *Nature* **400**: 461-464.

Watanabe, Y., Yokobayashi, S., Yamamoto, M. and Nurse, P. (2001). Pre-meiotic S phase is linked to reductional chromosome segregation and recombination. *Nature* **409**: 359-363.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**: 256-276.

Weiner, B. M. and Kleckner, N. (1994). Chromosome pairing via multiple interstitial interactions before and during meiosis in yeast. *Cell* **77**: 977-991.

Weiss, K. M. and Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**: 19-24.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. and Lathrop, M. (1992). A second-generation linkage map of the human genome. *Nature* **359**: 794-801.

White, M. A., Detloff, P., Strand, M. and Petes, T. D. (1992). A promoter deletion reduces the rate of mitotic, but not meiotic, recombination at the HIS4 locus in yeast. *Curr. Genet.* **21**: 109-116.

White, M. A., Dominska, M. and Petes, T. D. (1993). Transcription factors are required for the meiotic recombination hotspot at the HIS4 locus in *Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA* **90**: 6621-6625.

White, M. A., Wierdl, M., Detloff, P. and Petes, T. D. (1991). DNA-binding protein RAP1 stimulates meiotic recombination at the HIS4 locus in yeast. *Proc. Natl. Acad. Sci. USA* **88**: 9755-9759.

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, C., Nguyen, H. T. and Schmid, C. W. (1987). Existence of at least three distinct Alu subfamilies. *J. Mol. Evol.* **26**: 180-186.

Wingett, D., Forcier, K. and Nielson, C. P. (1999). A role for CD99 in T cell activation. *Cell Immunol.* **193**: 17-23.

Wright, F. A., Lemon, W. J., Zhao, W. D., Sears, R., Zhuo, D., Wang, J. P., Yang, H. Y., Baer, T., Stredney, D., Spitzner, J., Stutz, A., Krahe, R. and Yuan, B. (2001). A draft annotation and overview of the human genome. *Genome Biol.* **2**: 1-18.

Wu, T. C. and Lichten, M. (1994). Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science* **263**: 515-518.

Wu, T. C. and Lichten, M. (1995). Factors that affect the location and frequency of meiosis-induced double-strand breaks in *Saccharomyces cerevisiae. Genetics* **140**: 55-66.

Xu, F. and Petes, T. D. (1996). Fine-structure mapping of meiosis-specific double-strand DNA breaks at a recombination hotspot associated with an insertion of telomeric sequences upstream of the HIS4 locus in yeast. *Genetics* **143**: 1115-1125.

Xu, L. and Kleckner, N. (1995). Sequence non-specific double-strand breaks and interhomolog interactions prior to double-strand break formation at a meiotic recombination hot spot in yeast. *Embo J.* **14**: 5115-5128.

Xu, L., Weiner, B. M. and Kleckner, N. (1997). Meiotic cells monitor the status of the interhomolog recombination complex. *Genes Dev.* **11**: 106-118.

Yang, F., Hanson, N. Q., Schwichtenberg, K. and Tsai, M. Y. (2000). Variable number tandem repeat in exon/intron border of the cystathionine beta-synthase gene: a single nucleotide substitution in the second repeat prevents multiple alternate splicing. *Am. J. Med. Genet.* **95**: 385-390.

Yauk, C. L., Bois, P. R. and Jeffreys, A. J. (2003). High-resolution sperm typing of meiotic recombination in the mouse MHC E_β gene. *Embo J.* **22**: 1389-1397.

Yi, H., Donohue, S. J., Klein, D. C. and McBride, O. W. (1993). Localization of the hydroxyindole-O-methyltransferase gene to the pseudoautosomal region: implications for mapping of psychiatric disorders. *Hum. Mol. Genet.* **2**: 127-131.

Yip, S. P., Lovegrove, J. U., Rana, N. A., Hopkinson, D. A. and Whitehouse, D. B. (1999). Mapping recombination hotspots in human phosphoglucomutase (PGM1). *Hum. Mol. Genet.* **8**: 1699-1706.

Young, J. A., Schreckhise, R. W., Steiner, W. W. and Smith, G. R. (2002). Meiotic recombination remote from prominent DNA break sites in *S. pombe. Mol. Cell* **9**: 253-263.

Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A. J., Deloukas, P., Olsen, A., Doggett, N. A., Ghebranious, N., Broman, K. W. and Weber, J. L. (2001). Comparison of human genetic and sequence-based physical maps. *Nature* **409**: 951-953.

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.

Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7**: 919-932.

Zhao, Z., Jin, L., Fu, Y. X., Ramsay, M., Jenkins, T., Leskinen, E., Pamilo, P., Trexler, M., Patthy, L., Jorde, L. B., Ramos-Onsins, S., Yu, N. and Li, W. H. (2000). Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. *Proc. Natl. Acad. Sci. USA* **97**: 11354-11358.

Zickler, D. and Kleckner, N. (1999). Meiotic chromosomes: integrating structure and function. *Annu. Rev. Genet.* **33**: 603-754.

Zickler, D., Moreau, P. J., Huynh, A. D. and Slezec, A. M. (1992). Correlation between pairing initiation sites, recombination nodules and meiotic recombination in Sordaria macrospora. *Genetics* **132**: 135-148.

Zietkiewicz, E., Yotova, V., Jarnik, M., Korab-Laskowska, M., Kidd, K. K., Modiano, D., Scozzari, R., Stoneking, M., Tishkoff, S., Batzer, M. and Labuda, D. (1998). Genetic structure of the ancestral population of modern humans. *J. Mol. Evol.* **47**: 146-155.