An Investigation of Links Between Simple Sequences and Meiotic Recombination Hotspots

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy in

Molecular and Cellular Biology at the

University of Canterbury

by

Andrew Bagshaw

University of Canterbury

2008

Contents

List of tables

List of figures

Abstract

Previous evidence has shown that the simple sequences microsatellites and poly-purine/poly-pyrimidine tracts (PPTs) could be both a cause, and an effect, of meiotic recombination. The causal link between simple sequences and recombination has not been much explored, however, probably because other evidence has cast doubt on its generality, though this evidence has never been conclusive. Several questions have remained unanswered in the literature, and I have addressed aspects of three of them in my thesis.

First, what is the scale and magnitude of the association between simple sequences and recombination? I found that microsatellites and PPTs are strongly associated with meiotic double-strand break (DSB) hotspots in yeast, and that PPTs are generally more common in human recombination hotspots, particularly in close proximity to hotspot central regions, in which recombination events are markedly more frequent. I also showed that these associations can’t be explained by coincidental mutual associations between simple sequences, recombination and other factors previously shown to correlate with both.

A second question not conclusively answered in the literature is whether simple sequences, or their high levels of polymorphism, are an effect of recombination. I used three methods to address this question. Firstly, I investigated the distributions of two-copy tandem repeats and short PPTs in relation to yeast DSB hotspots in order to look for evidence of an involvement of recombination in simple sequence formation. I found no significant associations. Secondly, I compared the fraction of simple sequences containing polymorphic sites between human recombination hotspots and coldspots. The third method I used was generalized linear model analysis, with which I investigated the correlation between simple sequence variation and recombination rate, and the influence on the correlation of additional factors with potential relevance including GC-content and gene density. Both the direct comparison and correlation methods showed a very weak and inconsistent effect of recombination on simple sequence polymorphism in the human genome.

Whether simple sequences are an important cause of recombination events is a third question that has received relatively little previous attention, and I have explored one aspect of it. Simple sequences of the types I studied have previously been shown to form non-B-DNA structures, which can be recombinagenic in model systems. Using a previously described sodium bisulphite modification assay, I tested for the presence of these structures in sequences amplified from the central regions of hotspots and cloned into supercoiled plasmids. I found significantly higher sensitivity to sodium bisulphite in humans in than in chimpanzees in three out of six genomic regions in which there is a hotspot in humans but none in chimpanzees. In the DNA2 hotspot, this correlated with a clear difference in numbers of molecules showing long contiguous strings of converted cytosines, which are present in previously described intramolecular quadruplex and triplex structures. Two out of the five other hotspots tested show evidence for secondary structure comparable to a known intramolecular triplex, though with similar patterns in humans and chimpanzees. In conclusion, my results clearly motivate further investigation of a functional link between simple sequences and meiotic recombination, including the putative role of non-B-DNA structures.

Publications associated with this thesis

Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots. Bagshaw AT, Pitt JP, Gemmell NJ. BMC Genomics. 2006 Jul 18;7:179.

High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots. Bagshaw AT, Pitt JP, Gemmell NJ. BMC Genomics. 2008 Jan 28;9:49.

I wrote both of these publications and I also conceived, designed and carried out the experiments they describe. Joel Pitt wrote the computer programmes. Neil Gemmell contributed to the interpretation of the data and the writing of the papers.

Acknowledgments

I would like to thank: Professor Neil Gemmell for supervision, all of the members of the Molecular Ecology group in the Department of Biological Sciences at the University of Canterbury for generally fostering a pleasant working environment, Victor Wiebe and Svante Pääbo from the Max Planck Institute for Evolutionary Anthrolpology in Liepzig, Germany for donating chimpanzee DNA, Maggie Tisch and Jackie Healy for technical assistance with laboratory work, Vladimir Mencl for technical assistance with computing and Joel Pitt for computer programming. I obtained useful information from my collaboration with Dr. Mikael Brandström and Professor Hans Ellegren from the University of Uppsala, Sweden (see Chapter 5). Financial support for my project was provided by a University of Canterbury Doctoral Scolarship and a Royal Society of New Zealand Marsden grant (UOC206).

Chapter 1

General Introduction

1.1 Simple sequences

A large proportion of the DNA of higher eukaryotes does not encode any protein product [1], and much of this non-coding DNA consists of patterns recognizable by their repetitious, or simple nature. These simple sequences have been classified into distinct families with common features. Repeats with very long periodicity, in the order of more than 1000 base pairs (bp) are known as satellites, minisatellites are normally classed as having repeat units between 9 and 100 bp, and microsatellites consist of very short repeated sequence motifs of six bp or less [2]. This nomenclature is rooted in the discovery of these repetitive sequences, which occurred when ultracentrifugation was first applied to the separation of DNA by density, and some outlying, or satellite fractions were seen [2]. Subsequently, following the emergence of technology capable of determining the sequence of large segments of genomes, other simple sequences were found to be extremely common, including long tracts consisting of only one class of nucleotide (poly-purine/poly-pyrimdine) [3, 4], and self-propagating transposable elements [5].

Despite their high abundance in all genomes analyzed to date, the degree to which these simple sequences are functional, or parasitic, is still questionable, and how they evolve is not well understood. An opportunity to investigate these questions has arisen with the recent emergence of large amounts of DNA sequence and sequence annotation data in a variety of organisms. Study of the conservation and distribution relative to known functional elements of simple sequences can provide useful information about their function and evolution. In this thesis I describe an investigation into links between two types of simple sequences: microsatellites and poly-purine/poly-pyrimidine, and the fundamental genetic process meiotic recombination.

1.1.1 Microsatellites

Repeated copies of sequence motifs where the copies are adjacent and in the same orientation are known as direct tandem repeats. These are typically called short tandem repeats (STRs) when the motifs are 6 base pairs (bp) or less and are known as microsatellites when multiple copies of a motif are strung together in contiguous arrays [6]. Microsatellites are scattered throughout the genomes of all eukaryotes [7-10] and are much more common than expected by chance [11], appearing once every 2-30 kilo bases (kb) in the human genome, depending on selection criteria [12, 13].

The high abundance of microsatellites could be linked to their propensity to undergo frequent change of length mutations, which is thought to occur predominantly by replication errors due to strand misalignment in repetitive sequence (replication slippage) [10, 14]. Because of this, and the fact that they are rarely found in genes [15], microsatellites have traditionally been thought of as having no useful role in genome physiology [16]. Evidence is emerging, however, that they may have substantial functional importance [9, 17-19]. In the early 1990s a surprising degree of conservation of microsatellite loci across diverse species was reported [20-25], in one case over 470 million years of evolution [22], strongly suggesting selective constraint. Indeed, considerable evidence has implicated microsatellites in regulating gene expression [26-32], which has been the main focus of recent work on microsatellite functionality (reviewed in [19]). Microsatellites might also act as recombination signals, and the generality of this possibility has not been much explored despite clear evidence in its favour [33-36]. Another potentially important property of microsatellites is that they can affect the progress of DNA replication [37, 38], which might itself be important in the regulation of recombination [39, 40].

Microsatellites are also of interest because their high degree of array length polymorphism has made them convenient markers of genetic divergence for applications in genome mapping [41-43], gene hunting [44-46], forensics [47], deducing kinship [48], population genetics [49-51] and the study of the evolution of species [52-54]. These applications depend on assumptions about microsatellite evolution which, at present, are overly simplistic because of unexplained heterogeneity in mutation rates among loci, and an increased understanding of microsatellite evolution and mutational mechanisms is therefore being sought (reviewed in [10, 14]). A potentially useful line of investigation in this respect is the possibility that recombination can mutate microsatellites, since it is usually assumed that replication errors are primarily responsible for microsatellite variability [10], but recombination has been implicated in some microsatellite mutation events, including the extreme microsatellite instability seen in some human genetic diseases (reviewed in [55, 56]).

1.1.2 Poly-purine/poly-pyrimidine

Poly-purine/poly-pyrimidine tracts (PPTs) consist of purine nucleotides (Adenine or Guanine) on one strand of the DNA duplex and pyrimidine nucleotides (Thymine or Cytosine) on the other, complementary strand. They can be made up of repetitive patterns, for example poly-A and poly-AG, both of which are microsatellites as well as PPTs. The vast majority of PPTs do not consist of tandemly repeated sequence motifs [57], but I have classified them as simple sequences in this thesis on the basis that any given PPT can, by definition, only be made up of two possible nucleotide types, as opposed to four for normal DNA. Like microsatellites, PPTs are highly over-represented in eukaryotic genomes [3, 4], with their frequency in *S. cerevisiae* exceeding the level expected by chance by as much as 15-fold. [3].

PPTs are not used as genetic markers, so their evolution has received little attention, and whether or not they are commonly length polymorphic has not previously been determined. Mutational mechanisms that may be involved in maintaining high frequencies of PPTs in genomes are therefore unknown, but their abundance suggests the possibility of functional importance. Indeed, evidence from model organisms such as yeast indicates that, like microsatellites, PPTs could commonly function in regulating recombination [58], replication [59], and gene expression [60, 61], and they have also been implicated in genomic instability associated with human disease (reviewed in [62]). These effects have often been linked to the ability of PPTs with some GC-content readily to form stable intramolecular secondary structures under physiological conditions [58, 59, 63-66]. Interestingly, the stability of these structures can be sensitive to single nucleotide changes [67-69], and the exact sequence requirements for them to form *in vivo* are not well understood [65].

1.2. Meiotic recombination hotspots

1.2.1 The distribution of meiotic recombination events

The majority of higher organisms, including humans, have two equivalent copies of each chromosome, one coming from each parent. These homologous chromosomes cross over and exchange genetic information during meiotic cell division, the process by which new sperm cells and oocytes are created, resulting in heritable genetic recombination (reviewed in [70]). The discovery of this phenomenon dates back to the early 1900s, when crossing over was observed under a microscope, and it was noted that some traits are more often inherited together than others. Using the fruitfly *Drosophila melanogaster*, which has easily observable heritable morphological polymorphisms such as variable eye colour, as a model organism, it was discovered that the locations of genes responsible for these polymorphisms could be mapped to relative positions on chromosomes based on the fact that the closer they are on a chromosome, the less frequently a crossover resulting in separate inheritance will occur between them, and genetic traits are sill mapped by this method today (reviewed in [71]).

It was initially assumed that crossover locations are random, but in the 1980s evidence emerged from studies of recombination in the budding yeast *Saccharomyces cerevisiae* and the mouse *Mus musculus* that they have a non-random distribution, complicating the methodology of gene mapping. In the 1980s and '90s narrow hotspots of meiotic recombination were discovered and intensively studied at three loci in the *S. cerevisiae* genome, and several other recombination hotspots were also identified in yeast (reviewed in [39]). Studies in mice also reported the existence of areas in which crossovers occurred with elevated frequency [72-76]. Investigation of the generality of these observations was not immediately possible due to the fact that crossovers are very rare in any given chromosomal location, so it is labour-intensive to map crossover hotspots by traditional methods, particularly in mammals [77]. As a result, recombination maps of the human genome could still only be created with an average resolution of about one mega base at the beginning of this decade [78]. Finer resolution then became possible with the discovery through genome sequencing initiatives of the locations of increasing numbers of sequence polymorphisms, which could be used as genetic markers. The genome-wide recombination mapping studies in the early years of this decade accumulated further evidence that crossovers generally have a complex non-random distribution [78, 79].

The emergence of high-throughput genotyping technologies in the last five years has enabled an increasingly thorough characterization of the distribution of recombination events in complex organisms. The new technologies were initially used to observe recombinants directly by screening many thousands of sperm cells [80-84], and in yeast microarray technology has enabled identification of recombination hotspots by their tendency to bind with recombination-initiating proteins [85, 86]. These studies revealed that recombination is generally concentrated in hotspots of 1-2.5 kb, separated by as much as 50-100 kb of sequence that seldom recombines. Such hotspots have now been described in yeast, mice and humans (reviewed in [39, 87, 88]). They have also been found in other organisms, including chimpanzees [89, 90], plants (reviewed in [91]) and fruit flies [92], but they are less well characterized in these taxa.

Mapping the recombination landscape of an entire genome using sperm typing would be prohibitively labour-intensive with traditional techniques [77]. To achieve a genome-wide recombination map in humans, workers instead applied recently available high-density sequence polymorphism data to infer recombination events indirectly [89, 90, 93-97]. These methods take advantage of haplotypes: chromosomal regions in which particular polymorphisms are associated with one another in diverse individuals, to infer recombination hotspot locations at a fine scale in regions where these marker associations break down. Recombination rates averaged across many generations for the entire human genome can thus be deduced with high resolution, but some evidence suggests that these methods have only about 60% power to detect hotspots in the present generation [83, 84, 98, 99]. One problem with them is that the recombination landscape is polymorphic to some degree among individuals [82, 100-102], and evidence indicates that most hotspots are often [93, 97, 103], but not always [104], shared across populations. These complications do not, however, encumber studies directly observing recombination events between individual generations, and advances in the speed and economy of high-throughput genotyping techniques are beginning to make this approach to hotspot mapping more practical. A recent study taking advantage of these techniques revealed patterns of recombination broadly similar to those detected by sperm-typing and haploytype inference, with some hotspots used differently between the sexes [101].

1.2.2 Regulation of meiotic recombination hotspots

The molecular processes involved in meiotic hotspot recombination have been described in some detail, most deeply in the yeast *S. cerevisiae* (reviewed in [39, 105]). These include a meiosis-specific opening of the local chromatin structure, i.e. an unpacking of DNA, presumably allowing access to recombination machinery [106, 107], and a requirement for a chromosomal double-strand break (DSB) to initiate recombination, which is catalyzed by the protein Spo11 in yeast [86, 108, 109]. Surprisingly, however, the factors governing the locations of hotspots, and their widely varying activity levels, are poorly understood [101]. There are two obvious reasons for this. Firstly, sequence features defining hotspot locations have not been found, and secondly, evidence has shown that complex, multi-leveled interactions between sequence and non-sequence (epigenetic) factors are involved in the recombination process, and these have not yet been fully elucidated (reviewed in [39, 87, 88, 105]).

It has been known for some time that DSBs in the yeast genome occur within narrow 100-500 bp regions but are not sequence-specific [110-113]. Recent studies mapping the locations of hotspots in the human genome have also reported the apparent absence of a recombination-initiating consensus sequence [80, 81, 84], and similar results have been reported in mice [114, 115]. Although some sequence elements, including simple sequences and GC-content, have been found to correlate fairly strongly with mammalian recombination rate at mega base scales [79, 116], workers who first mapped the recombination landscape of the human genome at the kilo base level found no sequence elements correlating more than very weakly with recombination rate at this fine scale [93, 94]. These results contrast the situation in bacteria, in which a particular sequence motif is known to initiate recombination (reviewed in [117]).

The absence of a hotspot consensus sequence has combined with other lines of evidence to cast doubt on the idea that the locations of recombination events in eukaryotes are governed to any substantial degree by local sequences. The observation of sex-specific hotspot use indicates that epigenetic factors, such as differential expression of proteins involved in the recombination machinery [118], DNA methylation, and/or, modification of the DNA packaging proteins histones, are involved in the regulation of meiotic recombination [79, 101]. The importance of epigenetic factors is clear in any case, since a crossover at any given hotspot is rare, but at least one must be performed between every pair of homologous chromosomes at every meiosis to ensure accurate chromosomal segregation [119]. Epigenetic factors might only govern the frequency of hotspot use, rather than hotspot location, but a recent study showed that hotspot activity can change without any corresponding change in sequence, and the authors of this study proposed that this could be explained by changes in *trans* factors (not relating to the hotspot's own chromosome), distal *cis* sequences (from the same chromsome) or epigenetic changes such as DNA methylation [120].

Local sequences clearly do play some role in meiotic recombination hotspots, however, since changes in recombination activity levels have been linked to local single nucleotide changes in hotspot central regions [82, 100, 102, 115, 121], and hotspot regulation by factors operating both in *cis* and in *trans* has been demonstrated [115], though how these local sequences act to regulate hotspots is not well understood. Surprisingly little work in this area has been reported in the literature, notably excepting investigations of the well-studied *ade6* locus in the genome of the fission yeast *Schizosaccharomyces pombe*. A single nucleotide change at this site promotes transcription factor-binding, and also a recombination hotspot [122, 123]. Interestingly, the dependence on transcription factor binding shown for the *ade6* hotspot, and also for the *HIS4* hotspot in *Saccharomyces cerevisiae* [124], is not coupled to dependence on transcription [33, 125]. The mechanism for its involvement in recombination is unclear, but it could relate to modulation of chromatin structure [39].

If hotspots in higher organisms do, after all, require strict sequence motifs, this could explain why their locations are not consistent between humans and chimpanzees despite more than 98% sequence similarity between the two species [89, 90]. However, alleles which promote the initiation of recombination tend to be lost during the recombination process, so if specific motifs are required, there is a paradox as to how hotspots are maintained [100, 126, 127]. Theoretically, the paradox could be resolved if hotspots are regulated predominantly by *trans* acting factors, at least when they first appear in a genome [128]. This suggests the possibility that sequences near to, but outside, hotspots could have an important role in hotspot regulation, and, consistent with this, a study of the yeast *HIS2* hotspot region showed that as much as 11.5 kb of DNA from around the hotspot is necessary for its activity [129]. Flanking sequences are also required for transcription-factor-dependent recombination at the aforementioned fission yeast *ade6* hospot [130, 131], and at other loci [132, 133]. Their involvement could plausibly relate to some recombinagenic property of flanking sequence, to higher order chromosome structure, and/or to hotspot competition, in which increases in recombination activity at a location can cause lowered rates in neighbouring areas, and *vice versa* [110, 131, 132, 134-136].

Local and distal sequences could therefore act in concert to control hotspot locations, though the nature of this interaction is unknown. Another complication for any theory of hotspot control is that transcription factor binding is only involved in a subset of hotspots in yeast [39]. The potential of transcription factors in general to promote recombination has been shown to be context dependent [137], and doubt has been cast on the generality of its importance by the observation that recombination is reduced on average near genes in humans [94, 101]. Sequence features other than transcription factor binding sites have been found to be associated with recombination hotspots including GC-rich DNA [85], tandem repeats [33, 72, 82, 138], transposable elements [94], and some specific motifs less than 10 bp long [94]. Direct tests of whether these sequence features are functional in extant hotspots are lacking, with the notable exception of an often overlooked study showing that deletion of a 14 bp poly-A tract from the *S. cerevisiae ARG4* hot spot reduced its activity by 75% [33].

Observations that different kinds of sequence can stimulate recombination in model systems has led to the idea that DNA sequence could regulate recombination hotspots at a local level in several distinct ways, and a three-way division of hotspot control has been hypothesized [39]. The first category is alpha hotspots, which require transcription factor binding, as noted above. A second category, beta hotspots, was proposed in view of the observation that tandem repeats that exclude nucleosomes can stimulate hotspot activity, and interestingly also transcription, when inserted into a yeast chromosome [35, 139]. Finally, gamma hotspots were suggested to require replication pausing, and this was based primarily on reports that recombination in yeast is tightly linked, in space and time, to chromosomal DNA replication [119, 140]. The reasons for this are not fully understood, but it is possible that a paused replication fork could allow time for chromatin at recombination hotspots to be made receptive to the recombination machinery [39]. Certainly there is clear evidence that the repair of stalled replication forks involves recombination in all organisms from bacteria to humans [119, 141-143], though this link is also dependent on sequence context [40]. These three classes of hotspots may not be entirely distinct mechanistically, since transcription factor binding, nucleosome modulating sequences, and replication pausing might all act via a process which includes marking and preparation of histone DNA packaging proteins to potentiate recombination [39], and the importance of histone modification is supported by some recent evidence [144, 145].

1.3 Recombination can be mutagenic

In humans, recombination rate is correlated with genetic diversity at the fine scale of hotspots, and this may indicate a mutagenic effect [146]. Inter-species DNA sequence divergence also correlates with recombination rate, possibly also reflecting mutations linked to the recombination process [97, 146, 147]. There is some evidence that this hypothesized mutagenic effect is biased in favour of A/T → G/C single nucleotide substitutions, suggesting that recombination acts to increase GC-content by a process known as biased gene conversion [146, 148, 149]. This could account for observed correlations between GC-content and recombination rate, though these tend to be much stronger at broad than fine scales, suggesting that if recombination does drive increases in GC-content, the recombination landscape, which evolves quickly at a fine scales, must be more conserved at large scales [93]. Recombination might also be mutagenic due to errors in crossing over resulting in insertion or deletion mutations and in theory this is more likely to occur in tandem repeats due to their potential to misalign [150]. Recombination could also cause mutations in repetitive sequences without crossing over, perhaps as a result of sequence misalignment in recombination intermediate structures [151]. Recombination-associated mutations without evidence of crossing over have been shown to occur in minisatellites, [82, 152] and might also be an important factor in microsatellite mutability [150, 151]. This possibility has not yet been explored on a large scale, and most work on the link between recombination and simple sequence mutation has focussed on genetic instability implicated in human disease (Reviewed in [56, 62]).

1.4 Links between meiotic recombination and simple sequences

1.4.1 Simple sequences found in frequently recombining regions

Microsatellite abundance has been found to correlate with recombination rate measured across mega base scales in rats, mice and humans [116, 153] and the presence of a microsatellite has been noted in hotspots mapped at a finer scale of a few kilo bases [94, 114, 154, 155]. One recent study found an overall enrichment of some types of microsatellite in recombination hotspots throughout the human genome [94], but did not address the question of whether microsatellites in general are more common in recombination hotspots. This question is relevant to links between recombination and microsatellites because common mutational mechanisms are thought to act on all microsatellites, so patterns of distribution of the class of sequence as a whole are of some interest [10, 14]. In Chapters 2 and 3 of this thesis I detail investigations into the association between microsatellites and recombination hotspots in yeast and humans respectively. The work presented in Chapter 2 constituted the first report of a general enrichment of microsatellites in recombination hotspots in any species [138].

The question of whether PPTs are more common in recombination hotspots has also not been addressed elsewhere in the literature. A correlation between PPT abundance and broad scale recombination rate in humans, mice and rats has been noted [79, 116], and some short, poly-pu/py-rich motifs have been found to be enriched in human hotspots [94], but it was unknown, prior to publication of the work presented in Chapter 4 of this thesis, whether PPTs in general were associated with recombination hotspots [57].

Other relevant questions not addressed elsewhere are whether the broad scale correlation between recombination rate and simple sequence density is driven by large-scale or local effects, and whether it is attributable to co-variation with some third factor. In Chapters 3 and 4, I report the use of wavelet analysis and generalized linear models to address these questions, in relation to microsatellites and PPTs respectively, in the human genome. Also, in Chapters 2 and 4, I present analyses of the influence of transcription, promoter regions, GC-content and transposable elements on the associations between microsatellites, PPTs and recombination hotspots in yeast.

1.4.2 What biological processes underlie the association between frequently recombining regions and simple sequences?

An enrichment of simple sequences in recombination hotspots not attributable to a mutual correlation with any third factor would suggest a widespread direct, causal relationship between simple sequences and some aspect of the recombination process. This contention would be supported if the association were concentrated in hotspot central regions, in which recombination is most frequent [87, 115]. A causal link between recombination and simple sequences could involve a regulatory effect of the sequences on recombination hotspot locations or intensity levels, a recombination-mediated mutation bias, or a combination of both. Despite evidence for the existence of these processes, their prevalence is virtually unknown. This is because they have not been thoroughly investigated, probably as a result of studies that have cast doubt on the apparent likelihood of their occurring commonly. As summarized above (Section 1.2.2), evidence suggests that sequence patterns may not have a ubiquitous functional role in recombination hotspots, and this could be the reason why the generality of previous observations that simple sequences can affect recombination in yeast chromosomal and plasmid DNA [33-36, 58, 156-163] has apparently not been tested. Moreover, the question of whether recombination could drive microsatellite evolution through a mutagenic effect, which was raised as early as 1976 [150], remains incompletely explored today [10, 138]. This is presumably due to the existence of evidence that microsatellites usually mutate by replication errors rather than unequal recombination [10, 14], but there are reasons to think that such errors could be more frequent in recombination hotspots (see Section 5.1.1).

Recombination has been implicated in some cases of microsatellite mutability [34, 164, 165], and if this occurs commonly, an elevation of microsatellite polymorphism levels in recombination hotspots should be detectable. Whether or not this is the case has not previously been reported, and in Chapter 5 I present an investigation into the relationship between polymorphic microsatellites and recombination sites in the human genome. Similarly, in Chapter 6, I ask whether the association between recombination and PPTs could be driven by a mutation bias, and I detail an investigation into PPT polymorphism in relation to recombination hotspots, which has also not been reported elsewhere in the literature.

In Chapter 7 I present some preliminary results from an investigation of the possible functional role of PPTs in recombination hotspots. Evidence suggests that the tendency of PPTs to form non-B-DNA structures could mediate such a role [58, 63] (see Sections 4.4 and 7.1). I used sodium bisulphite to probe sequences from recombination hotspots for such secondary structures. I tested amplified DNA from humans and chimpanzees in regions in which there is a hotspot in humans but none in chimpanzees, with the idea that structural differences between the two species, occurring in spite of their very high degree of sequence similarity, would be a strong argument in favour of a functional role in recombination hotspots for non-B-DNA structures.

References

1. Flam F: **Hints of a language in junk DNA**. *Science* 1994, **266**:1320.
2. Tautz D: **Notes on the definition and nomenclature of tandemly repetitive DNA sequences**. *Exs* 1993, **67**:21-28.
3. Raghavan S, Burma PK, Brahmachari SK: **Positional preferences of polypurine/polypyrimidine tracts in Saccharomyces cerevisiae genome: implications for cis regulation of gene expression**. *J Mol Evol* 1997, **45**(5):485-498.
4. Behe MJ: **An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes**. *Nucleic Acids Res* 1995, **23**(4):689-695.
5. Feschotte C, Pritham EJ: **DNA transposons and the evolution of eukaryotic genomes**. *Annu Rev Genet* 2007, **41**:331-368.
6. Chambers GK, MacAvoy ES: **Microsatellites: consensus and controversy**. *Comp Biochem Physiol B* 2000, **126**(4):455-476.
7. Valdes AM, Slatkin M, Freimer NB: **Allele frequencies at microsatellite loci: the stepwise mutation model revisited**. *Genetics* 1993, **133**(3):737-749.
8. Charlesworth B, Sniegowski P, Stephan W: **The evolutionary dynamics of repetitive DNA in eukaryotes**. *Nature* 1994, **371**:215-220.
9. Li B, Xia Q, Lu C, Zhou Z, Xiang Z: **Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes**. *Genomics Proteomics Bioinformatics* 2004, **2**(1):24-31.
10. Buschiazzo E, Gemmell NJ: **The rise, fall and renaissance of microsatellites in eukaryotic genomes**. *Bioessays* 2006, **28**(10):1040-1050.
11. Pupko T, Graur D: **Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units**. *J Mol Evol* 1999, **48**(3):313-316.
12. Stallings RL, Ford AF, Nelson D, Torney DC, Hildebrand CE, Moyzis RK: **Evolution and distribution of (GT)n repetitive sequences in mammalian genomes**. *Genomics* 1991, **10**(3):807-815.
13. Lander E, Linton LM, Birren B, al. e: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.
14. Ellegren H: **Microsatellites: simple sequences with complex evolution**. *Nat Rev Genet* 2004, **5**(6):435-445.
15. Hancock JM: **The contribution of slippage-like processes to genome evolution**. *J Mol Evol* 1995, **41**(6):1038-1047.
16. Epplen JT: **On simple repeated GATCA sequences in animal genomes: a critical reappraisal**. *J Hered* 1988, **79**(6):409-417.
17. Li YC, Korol AB, Fahima T, Beiles A, Nevo E: **Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review**. *Mol Ecol* 2002, **11**(12):2453-2465.
18. Li YC, Korol AB, Fahima T, Nevo E: **Microsatellites within genes: structure, function, and evolution**. *Mol Biol Evol* 2004, **21**(6):991-1007.
19. Nakagama H, Higuchi K, Tanaka E, Tsuchiya N, Nakashima K, Katahira M, Fukuda H: **Molecular mechanisms for maintenance of G-rich short tandem repeats capable of adopting G4 DNA structures**. *Mutat Res* 2006, **598**(1-2):120-131.
20. Schlotterer C, Amos B, Tautz D: **Conservation of polymorphic simple sequence loci in cetacean species**. *Nature* 1991, **354**(6348):63-65.

21. FitzSimmons NN, Moritz C, Moore SS: **Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution**. *Mol Biol Evol* 1995, **12**(3):432-440.
22. Rico C, Rico I, Hewitt G: **470 million years of conservation of microsatellite loci among fish species**. *Proc R Soc Lond B Biol Sci* 1996, **263**(1370):549-557.
23. Ezenwa VO, Peters JM, Zhu Y, Arevalo E, Hastings MD, Seppa P, Pedersen JS, Zacchi F, Queller DC, Strassmann JE: **Ancient conservation of trinucleotide microsatellite loci in polistine wasps**. *Mol Phylogenet Evol* 1998, **10**(2):168-177.
24. Zhu Y, Queller DC, Strassmann JE: **A phylogenetic perspective on sequence evolution in microsatellite loci**. *J Mol Evol* 2000, **50**(4):324-338.
25. Eichler EE, Kunst CB, Lugenbeel KA, Ryder OA, Davison D, Warren ST, Nelson DL: **Evolution of the cryptic FMR1 CGG repeat**. *Nat Genet* 1995, **11**(3):301-308.
26. Struhl K: **Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast**. *Proc Natl Acad Sci U S A* 1985, **82**(24):8419-8423.
27. Uhlemann AC, Szlezak NA, Vonthein R, Tomiuk J, Emmer SA, Lell B, Kremsner PG, Kun JF: **DNA phasing by TA dinucleotide microsatellite length determines in vitro and in vivo expression of the gp91phox subunit of NADPH oxidase and mediates protection against severe malaria**. *J Infect Dis* 2004, **189**(12):2227-2234.
28. Curi RA, Oliveira HN, Silveira AC, Lopes CR: **Effects of polymorphic microsatellites in the regulatory region of IGF1 and GHR on growth and carcass traits in beef cattle**. *Anim Genet* 2005, **36**(1):58-62.
29. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M: **A polymorphic microsatellite that mediates induction of PIG3 by p53**. *Nat Genet* 2002, **30**(3):315-320.
30. Borrmann L, Seebeck B, Rogalla P, Bullerdiek J: **Human HMGA2 promoter is coregulated by a polymorphic dinucleotide (TC)-repeat**. *Oncogene* 2003, **22**(5):756-760.
31. Hammock EA, Young LJ: **Microsatellite instability generates diversity in brain and sociobehavioral traits**. *Science* 2005, **308**(5728):1630-1634.
32. Hammock EA, Young LJ: **Functional microsatellite polymorphism associated with divergent social structure in vole species**. *Mol Biol Evol* 2004, **21**(6):1057-1063.
33. Schultes NP, Szostak JW: **A poly(dA.dT) tract is a component of the recombination initiation site at the ARG4 locus in Saccharomyces cerevisiae**. *Mol Cell Biol* 1991, **11**(1):322-328.
34. Gendrel CG, Boulet A, Dutreix M: **(CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis**. *Genes Dev* 2000, **14**(10):1261-1268.
35. Kirkpatrick DT, Wang YH, Dominska M, Griffith JD, Petes TD: **Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes**. *Mol Cell Biol* 1999, **19**(11):7661-7671.
36. Treco D, Arnheim N: **The evolutionarily conserved repetitive sequence d(TG.AC)n promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis**. *Mol Cell Biol* 1986, **6**(11):3934-3947.
37. Hile SE, Eckert KA: **Positive correlation between DNA polymerase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences.** *J Mol Biol* 2004, **335**(3):745-759.
38. Hile SE, Eckert KA: **DNA polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellites sequences**. *Nucleic Acids Res* 2008, **36**(2):688-696.

39. Petes TD: **Meiotic recombination hot spots and cold spots**. *Nat Rev Genet* 2001, **2**(5):360-369.
40. Labib K, Hodgson B: **Replication fork barriers: pausing for a break or stalling for time?** *EMBO Rep* 2007, **8**(4):346-353.
41. Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E *et al*: **A comprehensive genetic map of the human genome based on 5,264 microsatellites**. *Nature* 1996, **380**(6570):152-154.
42. Ihara N, Takasuga A, Mizoshita K, Takeda H, Sugimoto M, Mizoguchi Y, Hirano T, Itoh T, Watanabe T, Reed KM *et al*: **A comprehensive genetic map of the cattle genome based on 3802 microsatellites**. *Genome Res* 2004, **14**(10A):1987-1998.
43. Dietrich WF, Miller J, Steen R, Merchant MA, Damron-Boles D, Husain Z, Dredge R, Daly MJ, Ingalls KA, O'Connor TJ: **A comprehensive genetic map of the mouse genome**. *Nature* 1996, **380**(6570):149-152.
44. Sibov ST, de Souza CL, Jr., Garcia AA, Silva AR, Garcia AF, Mangolin CA, Benchimol LL, de Souza AP: **Molecular mapping in tropical maize (Zea mays L.) using microsatellite markers. 2. Quantitative trait loci (QTL) for grain yield, plant height, ear height and grain moisture**. *Hereditas* 2003, **139**(2):107-115.
45. Goris A, Sawcer S, Vandenbroeck K, Carton H, Billiau A, Setakis E, Compston A, Dubois B: **New candidate loci for multiple sclerosis susceptibility revealed by a whole genome association screen in a Belgian population**. *J Neuroimmunol* 2003, **143**(1-2):65-69.
46. Dirlewanger E, Cosson P, Howad W, Capdeville G, Bosselut N, Claverie M, Voisin R, Poizat C, Lafargue B, Baron O *et al*: **Microsatellite genetic linkage maps of myrobalan plum and an almond-peach hybrid--location of root-knot nematode resistance genes**. *Theor Appl Genet* 2004, **109**(4):827-838.
47. Tamaki K, Jeffreys AJ: **Human tandem repeat sequences in forensic DNA typing**. *Leg Med (Tokyo)* 2005, **7**(4):244-250.
48. Webster MS, Reichart L: **Use of microsatellites for parentage and kinship analyses in animals**. *Methods Enzymol* 2005, **395**:222-238.
49. Schlotterer C, Pemberton J: **The use of microsatellites for genetic analysis of natural populations**. *Exs* 1994, **69**:203-214.
50. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW: **Genetic structure of human populations**. *Science* 2002, **298**(5602):2381-2385.
51. Hayano A, Yoshioka M, Tanaka M, Amano M: **Population differentiation in the Pacific white-sided dolphin Lagenorhynchus obliquidens inferred from mitochondrial DNA and microsatellite analyses**. *Zoolog Sci* 2004, **21**(9):989-999.
52. Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL: **High resolution of human evolutionary trees with polymorphic microsatellites**. *Nature* 1994, **368**(6470):455-457.
53. Meyer E, Wiegand P, Rand SP, Kuhlmann D, Brack M, Brinkmann B: **Microsatellite polymorphisms reveal phylogenetic relationships in primates**. *J Mol Evol* 1995, **41**(1):10-14.
54. Schlotterer C: **Genealogical inference of closely related species based on microsatellites**. *Genet Res* 2001, **78**(3):209-212.
55. Kovtun IV, McMurray CT: **Features of trinucleotide repeat instability in vivo**. *Cell Res* 2008, **18**(1):198-213.
56. Mirkin SM: **DNA structures, repeat expansions and human hereditary disorders**. *Curr Opin Struct Biol* 2006, **16**(3):351-358.

57. Bagshaw AT, Pitt JP, Gemmell NJ: **Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots**. *BMC Genomics* 2006, **7**:179.

58. Rooney SM, Moore PD: **Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells**. *Proc Natl Acad Sci U S A* 1995, **92**(6):2141-2144.

59. Dayn A, Samadashwily GM, Mirkin SM: **Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization**. *Proc Natl Acad Sci U S A* 1992, **89**(23):11406-11410.

60. Maiti AK, Brahmachari SK: **Poly purine.pyrimidine sequences upstream of the beta-galactosidase gene affect gene expression in Saccharomyces cerevisiae**. *BMC Mol Biol* 2001, **2**(1):11.

61. Lu Q, Teare JM, Granok H, Swede MJ, Xu J, Elgin SC: **The capacity to form H-DNA cannot substitute for GAGA factor binding to a (CT)n*(GA)n regulatory site**. *Nucleic Acids Res* 2003, **31**(10):2483-2494.

62. Bissler JD: **Triplex DNA and human disease**. *Frontiers in Bioscience* 2007, **12**:4536-4546.

63. Wells RD, Collier DA, Hanvey JC, Shimizu M, Wohlrab F: **The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences**. *Faseb J* 1988, **2**(14):2939-2949.

64. Radhakrishnan I, Patel DJ: **DNA triplexes: solution structures, hydration sites, energetics, interactions, and function**. *Biochemistry* 1994, **33**(38):11405-11416.

65. Raghavan SC, Chastain P, Lee JS, Hegde BG, Houston S, Langen R, Hsieh CL, Haworth IS, Lieber MR: **Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation**. *J Biol Chem* 2005, **280**(24):22749-22760.

66. Kohwi Y, Kohwi-Shigematsu T: **Altered gene expression correlates with DNA structure**. *Genes Dev* 1991, **5**(12B):2547-2554.

67. Bacolla A, Ulrich MJ, Larson JE, Ley TJ, Wells RD: **An intramolecular triplex in the human gamma-globin 5'-flanking region is altered by point mutations associated with hereditary persistence of fetal hemoglobin**. *J Biol Chem* 1995, **270**(41):24556-24563.

68. Boles TC, Hogan ME: **DNA structure equilibria in the human c-myc gene**. *Biochemistry* 1987, **26**(2):367-376.

69. Ulrich MJ, Gray WJ, Ley TJ: **An intramolecular triplex is disrputed by point mutations associated with hereditary persistence of fetal hemoglobin**. *Journal of Biological Chemistry* 1992, **267**:18649-18658.

70. Cromie GA, Smith GR: **Branching out: meiotic recombination and its regulation**. *Trends in Cell Biology* 2007, **17**(9):448-455.

71. Morton NE: **A history of association mapping**. *Methods in Molecular Biology* 2007, **376**:17-21.

72. Kobori JA, Strauss E, Minard K, Hood L: **Molecular analysis of the hotspot of recombination in the murine major histocompatibility complex**. *Science* 1986, **234**(4773):173-179.

73. Uematsu Y, Kiefer H, Schulze R, Fischer-Lindahl K, Steinmetz M: **Molecular characterization of a meiotic recombinational hotspot enhancing homologous equal crossing-over**. *Embo J* 1986, **5**(9):2123-2129.

74. Lafuse WP, Berg N, Savarirayan S, David CS: **Mapping of a second recombination hot spot within the I-E region of the mouse H-2 gene complex**. *J Exp Med* 1986, **163**(6):1518-1528.

75. Passmore HC, Kobori JA, Zimmerer EJ, Spinella DG, Hood L: **Molecular characterization of meiotic recombination within the major histocompatibility complex of the mouse: mapping of crossover sites within the I region**. *Biochem Genet* 1987, **25**(7-8):513-526.

76. Shiroishi T, Sagai T, Moriwaki K: **Sexual preference of meiotic recombination within the H-2 complex**. *Immunogenetics* 1987, **25**(4):258-262.

77. Arnheim N, Calabrese P, Nordborg M: **Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved**. *Am J Hum Genet* 2003, **73**(1):5-16.

78. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW *et al*: **Comparison of human genetic and sequence-based physical maps**

A high-resolution recombination map of the human genome. *Nature* 2001, **409**(6822):951-953.

79. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G *et al*: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**(3):241-247.

80. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex**. *Nat Genet* 2001, **29**(2):217-222.

81. May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys AJ: **Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX**. *Nat Genet* 2002, **31**(3):272-275.

82. Jeffreys AJ, Murray J, Neumann R: **High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot**. *Mol Cell* 1998, **2**(2):267-273.

83. Kauppi L, Stumpf MP, Jeffreys AJ: **Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region**. *Genomics* 2005, **86**(1):13-24.

84. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: **Human recombination hot spots hidden in regions of strong marker association**. *Nat Genet* 2005, **37**(6):601-606.

85. Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD: **Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae**. *Proc Natl Acad Sci U S A* 2000, **97**(21):11383-11390.

86. Cromie GA, Hyppa RW, Cam HP, Farah JA, Grewal SI, Smith GR: **A discrete class of intergenic DNA dictates meiotic DNA break hotspots in fission yeast**. *PLoS Genet* 2007, **3**(8):e141.

87. Kauppi L, Jeffreys AJ, Keeney S: **Where the crossovers are: recombination distributions in mammals**. *Nat Rev Genet* 2004, **5**(6):413-424.

88. Arnheim N, Calabrese P, Tiemann-Boege I: **Mammalian Meiotic Recombination Hot Spots**. *Annu Rev Genet* 2007, **41**:369-399.

89. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**(4):429-434.

90. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P *et al*: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**(5718):107-111.

91. Mezard C: **Meiotic recombination hotspots in plants**. *Biochem Soc Trans* 2006, **34**(Pt 4):531-534.

92. Dorer DR, Christensen AC: **A recombinational hotspot at the triplo-lethal locus of Drosophila melanogaster**. *Genetics* 1989, **122**(2):397-401.

93. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome**. *Science* 2004, **304**(5670):581-584.

94. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**(5746):321-324.

95. Crawford DC, Bhangale T, Li N, Hellenthal G, Rieder MJ, Nickerson DA, Stephens M: **Evidence for substantial fine-scale variation in recombination rates across the human genome**. *Nat Genet* 2004, **36**(7):700-706.

96. Zhang J, Li F, Li J, Zhang MQ, Zhang X: **Evidence and characteristics of putative human alpha recombination hotspots**. *Hum Mol Genet* 2004, **13**(22):2823-2828.

97. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM *et al*: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**(7164):851-861.

98. Yi S, Li WH: **Molecular Evolution of Recombination Hotspots and Highly Recombining Pseudoautosomal Regions in Hominoids**. *Mol Biol Evol* 2005.

99. Verhoeven KJ, Simonsen KL: **Genomic Haplotype Blocks May Not Accurately Reflect Spatial Variation in Historic Recombination Intensity**. *Mol Biol Evol* 2005, **22**(3):735-740.

100. Jeffreys AJ, Neumann R: **Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot**. *Nat Genet* 2002, **31**(3):267-271.

101. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M: **High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans**. *Science* 2008, **319**(5868):1395-1398.

102. Jeffreys AJ, Neumann R: **Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot**. *Hum Mol Genet* 2005, **14**(15):2277-2287.

103. Kauppi L, Sajantila A, Jeffreys AJ: **Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region**. *Hum Mol Genet* 2003, **12**(1):33-40.

104. Graffelman J, Balding DJ, Gonzalez-Neira A, Bertranpetit J: **Variation in estimated recombination rates across human populations**. *Hum Genet* 2007, **122**(3-4):301-310.

105. Nishant KT, Rao MR: **Molecular features of meiotic recombination hot spots**. *Bioessays* 2006, **28**(1):45-56.

106. Wu TC, Lichten M: **Meiosis-induced double-strand break sites determined by yeast chromatin structure**. *Science* 1994, **263**(5146):515-518.

107. Shenkar R, Shen MH, Arnheim N: **DNase I-hypersensitive sites and transcription factor-binding motifs within the mouse E beta meiotic recombination hot spot**. *Mol Cell Biol* 1991, **11**(4):1813-1819.

108. Sun H, Treco D, Schultes NP, Szostak JW: **Double-strand breaks at an initiation site for meiotic gene conversion**. *Nature* 1989, **338**(6210):87-90.

109. Steiner WW, Schreckhise RW, Smith GR: **Meiotic DNA breaks at the S. pombe recombination hot spot M26**. *Mol Cell* 2002, **9**(4):847-855.

110. Xu L, Kleckner N: **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* 1995, **14**(20):5115-5128.
111. Liu J, Wu TC, Lichten M: **The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate**. *Embo J* 1995, **14**(18):4599-4608.
112. de Massy B, Rocco V, Nicolas A: **The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in Saccharomyces cerevisiae**. *Embo J* 1995, **14**(18):4589-4598.
113. Xu F, Petes TD: **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* 1996, **143**(3):1115-1125.
114. Shiroishi T, Koide T, Yoshino M, Sagai T, Moriwaki K: **Hotspots of homologous recombination in mouse meiosis**. *Adv Biophys* 1995, **31**:119-132.
115. Baudat F, de Massy B: **Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot**. *PLoS Genet* 2007, **3**(6):e100.
116. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ: **Comparative recombination rates in the rat, mouse, and human genomes**. *Genome Res* 2004, **14**(4):528-538.
117. Smith GR: **Hotspots of homologous recombination**. *Experientia* 1994, **50**(3):234-241.
118. Kong A, al. E: **Sequence variants in the RNF212 Gene Associate with Genome-wide recombination rate**. *Science* 2008, **319**:1398-1401.
119. Baudat F, Keeney S: **Meiotic recombination: Making and breaking go hand in hand**. *Curr Biol* 2001, **11**(2):R45-48.
120. Neumann R, Jeffreys AJ: **Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation**. *Hum Mol Genet* 2006, **15**(9):1401-1411.
121. Schuchert P, Langsford M, Kaslin E, Kohli J: **A specific DNA sequence is required for high frequency of recombination in the ade6 gene of fission yeast**. *Embo J* 1991, **10**(8):2157-2163.
122. Kon N, Krawchuk MD, Warren BG, Smith GR, Wahls WP: **Transcription factor Mts1/Mts2 (Atf1/Pcr1, Gad7/Pcr1) activates the M26 meiotic recombination hotspot in Schizosaccharomyces pombe**. *Proc Natl Acad Sci U S A* 1997, **94**(25):13765-13770.
123. Wahls WP, Smith GR: **A heteromeric protein that binds to a meiotic homologous recombination hot spot: correlation of binding and hot spot activity**. *Genes Dev* 1994, **8**(14):1693-1702.
124. White MA, Dominska M, Petes TD: **Transcription factors are required for the meiotic recombination hotspot at the HIS4 locus in Saccharomyces cerevisiae**. *Proc Natl Acad Sci U S A* 1993, **90**(14):6621-6625.
125. White MA, Detloff P, Strand M, Petes TD: **A promoter deletion reduces the rate of mitotic, but not meiotic, recombination at the HIS4 locus in yeast**. *Curr Genet* 1992, **21**(2):109-116.
126. Nicolas A, Treco D, Schultes NP, Szostak JW: **An initiation site for meiotic gene conversion in the yeast Saccharomyces cerevisiae**. *Nature* 1989, **338**(6210):35-39.
127. Boulton A, Myers RS, Redfield RJ: **The hotspot conversion paradox and the evolution of meiotic recombination**. *Proc Natl Acad Sci U S A* 1997, **94**(15):8058-8063.

128. Peters A: **A combination of cis and trans control can solve the hotspot conversion paradox**. *Genetics* 2008, **[Eupub ahead of print]**(Feb 3).

129. Haring SJ, Halley GR, Jones AJ, Malone RE: **Properties of natural double-strand-break sites at a recombination hotspot in Saccharomyces cerevisiae**. *Genetics* 2003, **165**(1):101-114.

130. Steiner WW, Smith GR: **Optimizing the Nucleotide Sequence of a Meiotic Recombination Hotspot in Schizosaccharomyces pombe**. *Genetics* 2005.

131. Zahn-Zabal M, Lehmann E, Kohli J: **Hot spots of recombination in fission yeast: inactivation of the M26 hot spot by deletion of the ade6 promoter and the novel hotspot ura4-aim**. *Genetics* 1995, **140**(2):469-478.

132. Wu TC, Lichten M: **Factors that affect the location and frequency of meiosis-induced double-strand breaks in Saccharomyces cerevisiae**. *Genetics* 1995, **140**(1):55-66.

133. Borde V, Wu TC, Lichten M: **Use of a recombination reporter insert to define meiotic recombination domains on chromosome III of Saccharomyces cerevisiae**. *Mol Cell Biol* 1999, **19**(7):4832-4842.

134. Fan QQ, Xu F, White MA, Petes TD: **Competition between adjacent meiotic recombination hotspots in the yeast Saccharomyces cerevisiae**. *Genetics* 1997, **145**(3):661-670.

135. Ohta K, Wu TC, Lichten M, Shibata T: **Competitive inactivation of a double-strand DNA break site involves parallel suppression of meiosis-induced changes in chromatin configuration**. *Nucleic Acids Res* 1999, **27**(10):2175-2180.

136. Fukuda T, Kugou K, Sasanuma H, Shibata T, Ohta K: **Targeted induction of meiotic double-strand breaks reveals chromosomal domain-dependent regulation of Spo11 and interactions among potential sites of meiotic recombination**. *Nucleic Acids Res* 2008, **36**(3):984-997.

137. Mieczkowski PA, Dominska M, Buck MJ, Gerton JL, Lieb JD, Petes TD: **Global analysis of the relationship between the binding of the Bas1p transcription factor and meiosis-specific double-strand DNA breaks in Saccharomyces cerevisiae**. *Mol Cell Biol* 2006, **26**(3):1014-1027.

138. Bagshaw AT, Pitt JP, Gemmell NJ: **High frequency of microsatellites in S. cerevisiae meiotic recombination hotspots**. *BMC Genomics* 2008, **9**(1):49.

139. Wang YH, Griffith JD: **The [(G/C)3NN]n motif: a common DNA repeat that excludes nucleosomes**. *Proc Natl Acad Sci U S A* 1996, **93**(17):8863-8867.

140. Borde V, Goldman AS, Lichten M: **Direct coupling between meiotic DNA replication and recombination initiation**. *Science* 2000, **290**(5492):806-809.

141. Kuzminov A: **DNA replication meets genetic exchange: chromosomal damage and its repair by homologous recombination**. *Proc Natl Acad Sci U S A* 2001, **98**(15):8461-8468.

142. Michel B: **Replication fork arrest and DNA recombination**. *Trends Biochem Sci* 2000, **25**(4):173-178.

143. Michel B, Flores MJ, Viguera E, Grompone G, Seigneur M, Bidnenko V: **Rescue of arrested replication forks by homologous recombination**. *Proc Natl Acad Sci U S A* 2001, **98**(15):8181-8188.

144. Mieczkowski PA, Dominska M, Buck MJ, Lieb JD, Petes TD: **Loss of a histone deacetylase dramatically alters the genomic distribution of Spo11p-catalyzed DNA breaks in Saccharomyces cerevisiae**. *Proc Natl Acad Sci USA* 2007, **104**(10):3955-3960.

145. Merker JD, Dominska M, Greenwell PW, Rinella E, Bouck DC, Shibata Y, Strahl BD, Mieczkowski P, Petes TD: **The histone methylase Set2p and the histone deacetylase RpD3p repress meiotic recombination at the HIS4 meiotic recombination hotspot in Saccharomyces cerevisiae**. *DNA Repair* 2008, **7**(8):1298-1308.
146. Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G: **The influence of recombination on human genetic diversity**. *PLoS Genet* 2006, **2**(9):e148.
147. Hellmann I, Prufer K, Ji H, Zody MC, Paabo S, Ptak SE: **Why do human diversity levels vary at a megabase scale?** *Genome Res* 2005, **15**(9):1222-1231.
148. Galtier N, Piganeau G, Mouchiroud D, Duret L: **GC-content evolution in mammalian genomes: the biased gene conversion hypothesis**. *Genetics* 2001, **159**(2):907-911.
149. Galtier N: **Gene conversion drives GC content evolution in mammalian histones**. *Trends Genet* 2003, **19**(2):65-68.
150. Smith GP: **Evolution of repeated DNA sequences by unequal crossover**. *Science* 1976, **191**(4227):528-535.
151. Jakupciak JP, Wells RD: **Gene conversion (recombination) mediates expansions of CTG[middle dot]CAG repeats**. *J Biol Chem* 2000, **275**(51):40003-40013.
152. Jeffreys AJ, Neil DL, Neumann R: **Repeat instability at human minisatellites arising from meiotic recombination**. *Embo J* 1998, **17**(14):4147-4157.
153. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW *et al*: **Comparison of human genetic and sequence-based physical maps**. *Nature* 2001, **409**(6822):951-953.
154. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M: **High-resolution patterns of meiotic recombination across the human major histocompatibility complex**. *Am J Hum Genet* 2002, **71**(4):759-776.
155. Rana NA, Ebenezer ND, Webster AR, Linares AR, Whitehouse DB, Povey S, Hardcastle AJ: **Recombination hotspots and block structure of linkage disequilibrium in the human genome exemplified by detailed analysis of PGM1 on 1p31**. *Hum Mol Genet* 2004, **13**(24):3089-3102.
156. Murphy KE, Stringer JR: **RecA independent recombination of poly[d(GT)-d(CA)] in pBR322**. *Nucleic Acids Res* 1986, **14**(18):7325-7340.
157. Bullock P, Miller J, Botchan M: **Effects of poly[d(pGpT).d(pApC)] and poly[d(pCpG).d(pCpG)] repeats on homologous recombination in somatic cells**. *Mol Cell Biol* 1986, **6**(11):3948-3953.
158. Napierala M, Dere R, Vetcher A, Wells RD: **Structure-dependent recombination hot spot activity of GAA.TTC sequences from intron 1 of the Friedreich's ataxia gene**. *J Biol Chem* 2004, **279**(8):6444-6454.
159. Wahls WP, Wallace LJ, Moore PD: **The Z-DNA motif d(TG)30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture**. *Mol Cell Biol* 1990, **10**(2):785-793.
160. Napierala M, Parniewski P, Pluciennik A, Wells RD: **Long CTG.CAG repeat sequences markedly stimulate intramolecular recombination**. *J Biol Chem* 2002, **277**(37):34087-34100.
161. Jankowski C, Nag DK: **Most meiotic CAG repeat tract-length alterations in yeast are SPO11 dependent**. *Mol Genet Genomics* 2002, **267**(1):64-70.
162. Sutherland GR, Baker E, Richards RI: **Fragile sites still breaking**. *Trends Genet* 1998, **14**(12):501-506.

163. Freudenreich CH, Kantrow SM, Zakian VA: **Expansion and length-dependent fragility of CTG repeats in yeast**. *Science* 1998, **279**(5352):853-856.

164. Hashem VI, Rosche WA, Sinden RR: **Genetic recombination destabilizes (CTG)n.(CAG)n repeats in E. coli**. *Mutat Res* 2004, **554**(1-2):95-109.

165. Jakupciak JP, Wells RD: **Genetic instabilities of triplet repeat sequences by recombination**. *IUBMB Life* 2000, **50**(6):355-359.

Chapter 2

High frequency of microsatellites in *S. cerevisiae* meiotic recombination hotspots

Abstract

The yeast *S. cerevisiae* has been the model organism of choice for investigating the process of meiotic recombination in general. I examined in detail the relationship between the distribution of microsatellites and hotspots of meiotic double-strand breaks, the precursors of meiotic recombination, throughout the *S. cerevisiae* genome. I used a specially designed computer algorithm to investigate all tandem repeats with motif length (repeat period) between one and six base pairs, including repeats with only two copies, which have not previously been studied in relation to recombination. I found that long, A/T-rich mono-, di- and trinucleotide microsatellites are around twice as frequent in hot than non-hot intergenic regions. The associations are weak or absent for repeats with less than six copies, and also for microsatellites with 4-6 base pair motifs, but high-copy arrays with motif length greater than three are very rare throughout the genome. I present evidence that the association between high-copy, short-motif microsatellites and recombination hotspots is not driven by effects on microsatellite distribution of other factors previously linked to both recombination and microsatellites, including transcription, promoter regions, GC-content and transposable elements.

2.1 Introduction

An ideal model organism with which to examine the association between simple sequences and recombination is the yeast *S. cerevisiae*, since it is the eukaryote most amenable to genetic study due to its simplicity and short generation time, and its genome is extremely well annotated for various genomic features including recombination-initiating double-strand breaks (DSBs) [1]. Factors that could complicate an association between sequence features and recombination are likely to be less problematic in yeast since, for example, the locations of genes and their expression levels have been well characterized, making it possible to control for the possible influence of transcription and promoter regions. Also, transposable or other known repetitive elements are not likely to mediate a link between sequence patterns and recombination hotspots in yeast, since these elements are not enriched in yeast hotspots [1]. Surprisingly, in view of these considerations, there have been no published reports of sequence features generally associated with recombination hotspots in *S. cerevisiae*, other than simple sequences, as described in this thesis (see page xi, Publications associated with this thesis), and high GC-content [1].

This chapter describes in detail the association between microsatellites and hotspots of meiotic DSBs throughout the *S. cerevisiae* genome. Computer software capable of detecting short repeat arrays and examining microsatellites in detail was not available at the time this work was carried out. To enable this investigation, I therefore collaborated with a programming expert, Joel Pitt from Lincoln University, to design a computer algorithm capable of detecting all microsatellites, including two-copy repeats, as well as allowing mismatches to a specifiable degree, and reporting repeat location, length, purity, motif and GC content [2]. I examined all these aspects of microsatellites in relation to DSB hotspots. I included in the analysis repeats with two copies, which are almost invariably ignored by studies of microsatellites, probably because it is computationally intensive to detect them genome-wide, and also because microsatellites with less than six copies are rarely polymorphic so are not used as genetic markers [3-5]. The study of short repeat arrays is nevertheless of interest, because the origin of microsatellites has traditionally been thought to require random point mutations up to a minimum array length required for replication slippage to occur [6], but two-copy microsatellites are more common than expected by chance [7]. To explain this, it has been suggested that microsatellite formation could occur as a result of strand misaligment during DNA replication [8]. Theoretically, however, this mechanism requires formation of a stable loop in sequences with multiple repetitions [9, 10] (see Section 5.1), which is not required for the formation of repetitive sequences by unequal recombination [11]. If two-copy repeats are enriched in recombination hotspots, this would suggest that unequal recombination could be involved in microsatellite formation, and that microsatellites in general might ultimately be an effect of recombination.

2.2 Methods

2.2.1 Sequence and annotation databases used

I used DSB hotspot locations mapped by Gerton and co-workers throughout the *S. cerevisiae* genome using microarray analysis of meiotic DSB frequency [1]. This study identified 177 hotspots, which encompassed all previously known meiotic recombination hotspots in the species, and 40 coldspots. For the purposes of my analysis, I extended the hotspots and coldspots to include the intergenic regions (IGRs) adjacent to the open reading frames (ORFs) identified by Gerton and co-workers [1], since yeast hotspots are typically centred on IGRs, in which most DSBs occur [12], and IGRs in the *S. cerevisiae* genome average only 500 base pairs (bp). The hotspots as I defined them had a mean length of 3466 bp. The principal statistical comparisons I made were between hot and non-hot, rather than hot and cold regions. Two reasons motivated this. First, cold regions are too few to provide a sufficiently reliable picture of microsatellite density in view of the rarity of long microsatellites in the yeast genome, and second, it has been established that recombination frequencies are relatively very low in all experimentally tested regions outside hotspots [13, 14].

I took figures for transcriptional frequency from the study by Holstege and co-workers (1998) who mapped transcription activity in vegetative cells for all yeast ORFs [15]. For IGRs, I took the mean of the two adjacent ORFs. I downloaded yeast sequences and ORF locations from the Stanford website [16]. The GenBank accession numbers for the 16 yeast chromosomes are NC_001133 through NC_001148.

2.2.2 Detection of microsatellites

I detected microsatellites in the yeast genome using an algorithm written in C, which I designed in collaboration with Joel Pitt, who wrote the script [2]. The programme operates by initially generating databases of all non-overlapping repeats of two copies or greater for repeated motif sizes between two and six bp, and three copies or greater for mononucleotide arrays. I created separate databases for perfect repeats, arrays with a maximum of one mismatch allowed per ten bp of sequence matching expectation based on the consensus repeat motif, and arrays with a maximum of one mismatch per six bp. Microsatellites overlapping two regions were excluded from the analysis. This occurred for less than one percent of arrays overall.

2.2.3 Categorization of microsatellites

I categorized repeats by copy number into three main groups: two-copy (3-5 for mononucleotide runs), medium and long. The minimum copy number for long repeats was six, a figure that I used in view of a study showing that microsatellites with less than six copies are not highly polymorphic [3-5]. I also used an additional category of very long repeats in order to illustrate an observed trend towards longer microsatellites being more strongly associated with DSB hotspots. I set the minimum length for this category at 14 bp for mononucleotide runs, because a previous study showed functional importance for a 14 bp poly-A tract in the *S. cerevisiae Arg 4* recombination hotspot [17], and ten copies for other microsatellites, which I found to be close to the longest minimum for which significant associations were detectable due to the rarity of these repeats in the yeast genome as a whole (Table 2.1). In a survey of the motifs of microsatellites throughout the genome I found the abundance of short-motif, AT-rich repeats to be dramatically higher than other repeat types, so for the purpose of comparing hot with non-hot regions I divided microsatellites by motif length as well as by array length in order not to lose information about longer motifs. I also separated poly-A from poly-G, because poly-A is many-fold more frequent than poly-G in the yeast genome. For my principle analyses, I used 19 physically independent categories of motif and array length. This number did not include different mismatch categories, which were not fully distinct. Additionally, I investigated dinucleotide repeats divided into the following motif groups: AT/TA, AC/CA/TG/GT, AG/GA/TC/CT and CG/GC (Appendix A), and I examined sequence motifs of microsatellites with repeated motifs of 3-6 bp visually. I defined a complex microsatellite as a repeat array within five or ten bp of another microsatellite of the same or larger copy number group.

2.2.4 Statistical analysis

I did statistical comparisons between hot and non-hot regions (Mann-Whitney U Test, 2-tailed tests in call cases) using SPSS, and correlation analyses (Spearman's Rho) with SAS. This was necessary because SPSS (Version 11) does not have a facility for partial non-parametric correlation. I used non-parametric partial correlation analysis in preference to generalized linear models to test the influence of GC-content and transcriptional frequency on the association between DSB intensity and microsatellite frequency because Gerton and colleagues found that DSB intensity statistics were more consistent when ranked [1]. For direct statistical comparisons I initially tested the distribution of each sample for normality (Kolmogorov-Smirnov Test) and for all comparisons at least one sample was found to be non-normal. This was not correctable by standard transformations in the majority of cases, and associations were clearly identifiable with non-parametric tests, so I used these in all cases. Because repeats were divided into 19 physically independent categories for statistical testing, I used Bonferroni's correction for multiple tests to set the alpha level at 0.05/19 = 0.0026. Bonferroni's correction is particularly conservative in the case of this study, because statistical power declines with increasing numbers of categories due to the fact that there are proportionally fewer microsatellites in each category.

2.3 Results

2.3.1 Survey of microsatellites in the *S. cerevisiae* genome

I initially surveyed the distribution of microsatellites between coding and non-coding regions. In general, numbers of microsatellites of a substantial length are very much lower in coding open reading frames (ORFs) than in intergenic regions (IGRs), (Table 2.1), despite the fact that ORFs cover 73.5% of the genome. This is not surprising, since array length change mutations in microsatellites other than tri- or hexanucleotide repeats would cause frame-shifts in ORFs, destroying gene function. The trend is also present for short microsatellites, with the exception 3-5 bp mononucleotide runs, which have similar frequency in ORFs and IGRs, but this is likely to be due to coding sequence such as AAA (Lys), GTTTTA (Val Leu), GGG (Gly) or AGGGTT (Arg Val), because the vast majority of the short mononucleotide repeats genome-wide are only three bp long. In view of this pattern of distribution, and the known tendency of DSBs to concentrate almost exclusively in IGRs, I made statistical comparisons only between DSB-hot IGRs and DSB-non-hot IGRs.

Table 2.1: Microsatellite abundance in the *S. cerevisiae* genome
Total number of microsatellite repeats (6 copies or more) and percentage of regions with at least one repeat in the *S. cerevisiae* genome. The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus motif. A lower e value therefore results in the detection of more imperfect repeats.

Repeat type			IGRs				ORFs			
			Hot (n=473)		Non hot (n=5520)		Hot (n=297)		Non hot (n=5683)	
Motif length	Copy number	Mis-matches alloId	No. of repeats	% of IGRs with a rpt.	No. of repeats	% of IGRs with a rpt.	No. of repeats	% of ORFs with a rpt.	No. of repeats	% of ORFs with a rpt
1 (A)	6+	perfect	1277	83.1	12547	77.4	339	57.6	13556	74.7
		e=10	1236	82.2	12262	77.0	338	57.6	13495	74.8
		e=6	1470	85.4	15153	82.2	437	64.3	17657	80.8
	14+	perfect	79	15.6	409	6.99	4	1.35	30	0.475
		e=10	146	27.5	741	12.2	5	1.68	73	1.16
		e=6	173	31.9	917	14.7	7	2.02	132	2.16
1 (G)	6+	perfect	33	6.55	241	4.09	32	10.4	474	7.80
		e=10	32	6.34	240	4.08	32	10.4	474	7.80
		e=6	46	8.67	307	5.16	44	13.8	641	10.3
	14+	perfect	2	0.423	2	0.0362	0	0	0	0
		e=10	2	0.423	2	0.0362	0	0	0	0
		e=6	2	0.423	2	0.0362	0	0	0	0
2	6+	perfect	57	10.4	357	6.05	8	2.36	21	0.352
		e=10	100	18.7	668	11.1	15	4.38	137	2.32
		e=6	130	23.5	1016	16.3	24	7.07	246	4.12
	10+	perfect	19	3.81	117	2.08	3	1.01	6	0.106
		e=10	28	5.71	171	3.04	5	1.68	12	0.211
		e=6	33	6.77	213	3.77	5	1.68	16	0.282
3	6+	perfect	7	1.27	27	0.435	8	2.36	165	2.46
		e=10	11	2.11	66	1.12	20	5.39	316	4.43
		e=6	21	4.02	118	1.96	28	7.74	478	6.49
	10+	perfect	1	0.211	8	0.145	0	0	29	0.493
		e=10	3	0.634	17	0.308	0	0	64	1.09
		e=6	3	0.634	20	0.362	0	0	100	1.57
4	6+	perfect	0	0	5	0.0906	0	0	1	0.0176
		e=10	0	0	12	0.217	0	0	1	0.0176
		e=6	0	0	19	0.344	0	0	2	0.0352
	10+	perfect	0	0	1	0.0181	0	0	0	0
		e=10	0	0	1	0.0181	0	0	0	0
		e=6	0	0	1	0.0181	0	0	0	0
5	6+	perfect	0	0	2	0.0362	0	0	0	0
		e=10	1	0.211	4	0.0725	0	0	3	0.0528
		e=6	1	0.211	5	0.0906	0	0	4	0.0704
	10+	perfect	0	0	0	0	0	0	0	0
		e=10	0	0	0	0	0	0	0	0
		e=6	0	0	0	0	0	0	1	0.0176
6	6+	perfect	1	0.211	3	0.0543	0	0	3	0.0528
		e=10	1	0.211	21	0.326	2	0.673	15	0.246
		e=6	1	0.211	10	0.181	4	1.35	11	0.176
	10+	perfect	0	0	0	0	0	0	0	0
		e=10	0	0	9	0.145	1	0.337	1	0.0176
		e=6	0	0	4	0.0725	1	0.337	5	0.0704

2.3.2 Elevated microsatellite frequencies in meiotic DSB hotspots

Frequencies in meiotic recombination hot and non-hot IGRs of the *S. cerevisiae* genome of microsatellites of all classes, as defined in Section 2.2.3, including two-copy repeats, can be found in Appendix A, Table A.1. Several types of microsatellite have significantly different frequency in hot than non-hot intergenic regions (alpha, adjusting for 19 independent categories of microsatellite with Bonferroni's correction = 0.0026, Table 2.2). Repeat frequencies in the 40 coldspots are generally lower than in other non-hot regions, but these differences are not statistically significant (Appendix A, Table A1). The correlation between DSB intensity level assayed by Gerton and co-workers [1], and microsatellite frequency, is weak (Appendix A, Table A2), but several repeat types, especially long poly-A and dinucleotide microsatellites, are markedly more abundant in hotspots than non-hot regions (Figure 2.1, Table 2.2). This discrepancy suggested the possibility that weaker hotspots contain more microsatellites, so I compared repeat frequencies between the hottest half of hot IGRs and remaining hot IGRs. No significant differences are present between these two types of IGRs, with the exception of short poly-A runs, which are more frequent in the colder half of hot IGRs, by approximately 10% ($p<10^{-5}$, Mann-Whitney U Test).

Of the types of microsatellite I investigated, mononucleotide runs are by far the most common, and long mononucleotide arrays are highly over-represented in hotspots. Although poly-A ($n \geq 6$) is less than 28% enriched in hot IGRs, poly-A ($n \geq 14$) is 2-2.5 fold more common in hot IGRs, and poly-G ($n \geq 14$) is nearly five fold over-represented, though this statistic may be misleading as the total number of poly-G arrays is very low (Table 2.1). Short poly-G runs are somewhat enriched in hotspots, and short poly-A is under-represented, but these differences can partly be explained by elevated GC content in the studied hotspots, which has been shown previously [1], since correlations between DSB intensity and short mononucleotide repeat frequencies are up to 50% weaker when controlling for GC content using partial correlation analysis (Appendix A, Table A2). For long microsatellites other than poly-G, correlations with DSB intensity are generally increased when controlling for GC-content (Appendix A, Table A2).

Dinucleotide repeats of six copies or more, especially those with ten copies or more, are strongly associated with hot IGRs, with poly-AT the most abundant type of repeat involved (Figure 2.1, Table 2.2). Trinucleotide repeats of more than six copies are approximately twice as frequent in hot than non-hot IGRs ($p = 0.0027$ Mann-Whitney U Test). This association is not quite significant when using the conservative Bonferroni correction for multiple hypotheses (alpha = 0.0026, see Section 2.2.4), but trinucleotide microsatellites are much scarcer than mono- or dinucleotide repeats in the yeast genome (Table 2.1), so statistical power to detect effects on their distribution is lower.

More marginal associations are present for some other repeat types. Dinucleotide repeats with between three and five copies are significantly over-represented in hot compared with non-hot IGRs, ($p<10^{-4}$, Mann-Whitney U Test) but levels of enrichment, at 12-18%, are much lower than for longer microsatellites (Table 2.2). Frequency of two-copy repeats is not significantly different in hot compared with non-hot regions, despite great abundance of these repeats relative to longer microsatellites, and consequently high statistical power. Tetra- and pentanucleotide microsatellites show no significant associations at all, but these repeat types are relatively very rare throughout the yeast genome (Table 2.1).

Table 2.2: Types of microsatellite differing significantly in abundance between DSB hotspots and non-hot regions of the *S. cerevisiae* genome
Microsatellite types with a significant difference in frequency between hot and non-hot IGRs. Statistical comparisons were done with the Mann-Whitney U test (corrected alpha = 0.0026). The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus motif. A lower e value therefore results in the detection of more imperfect repeats. For all repeat size classes there was substantial overlap between mismatch types, so these were not considered separate hypotheses for statistical purposes.

Repeat type			**Hot IGRs**		**Non-hot IGRs**		**Freq. Ratio (hot/ non-hot)**	**P value**
Motif Length	Copy number	Mismatch type	Mean per kb freq.	SEM	Mean per kb freq.	SEM		
1 (A)	3 to 5	perfect	35.03	0.5459	39.86	0.1743	0.88	$<10^{-18}$
		e=10	34.32	0.5437	39.44	0.1743	0.87	$<10^{-20}$
		e=6	31.80	0.5173	36.69	0.1622	0.87	$<10^{-20}$
	6+	perfect	5.421	0.2013	4.615	0.0560	1.17	$<10^{-4}$
		e=10	5.239	0.1949	4.504	0.0549	1.16	$<10^{-4}$
		e=6	6.121	0.2097	5.530	0.0607	1.11	0.0017
	14+	perfect	0.4178	0.0595	0.1706	0.0114	2.45	$<10^{-11}$
		e=10	0.7332	0.0762	0.3112	0.0149	2.36	$<10^{-19}$
		e=6	0.8537	0.0802	0.3770	0.0164	2.26	$<10^{-21}$
1 (G)	3 to 5	perfect	9.180	0.3040	7.252	0.0755	1.27	$<10^{-11}$
		e=10	9.160	0.3039	7.238	0.0755	1.27	$<10^{-11}$
		e=6	8.886	0.3004	7.125	0.0745	1.25	$<10^{-9}$
	6+	e=6	0.1604	0.0303	0.0931	0.0068	1.72	0.0012
	14+	perfect	0.0035	0.0026	0.0007	0.0005	4.83	0.0018
		e=10	0.0035	0.0026	0.0007	0.0005	4.83	0.0018
		e=6	0.0035	0.0026	0.0007	0.0005	4.83	0.0018
2	3 to 5	perfect	4.665	0.2225	3.957	0.0608	1.18	$<10^{-4}$
		e=10	4.342	0.2179	3.683	0.0580	1.18	0.0003
		e=6	6.174	0.2302	5.517	0.0675	1.12	0.0002
	6+	perfect	0.3681	0.0738	0.1955	0.0139	1.88	0.0002
		e=10	0.5987	0.0879	0.3557	0.0187	1.68	$<10^{-5}$
		e=6	0.7967	0.1026	0.5287	0.0222	1.51	$<10^{-4}$
	10+	e=10	0.2215	0.0624	0.0931	0.0097	2.38	0.0016
		e=6	0.2522	0.0657	0.1086	0.0102	2.32	0.0013

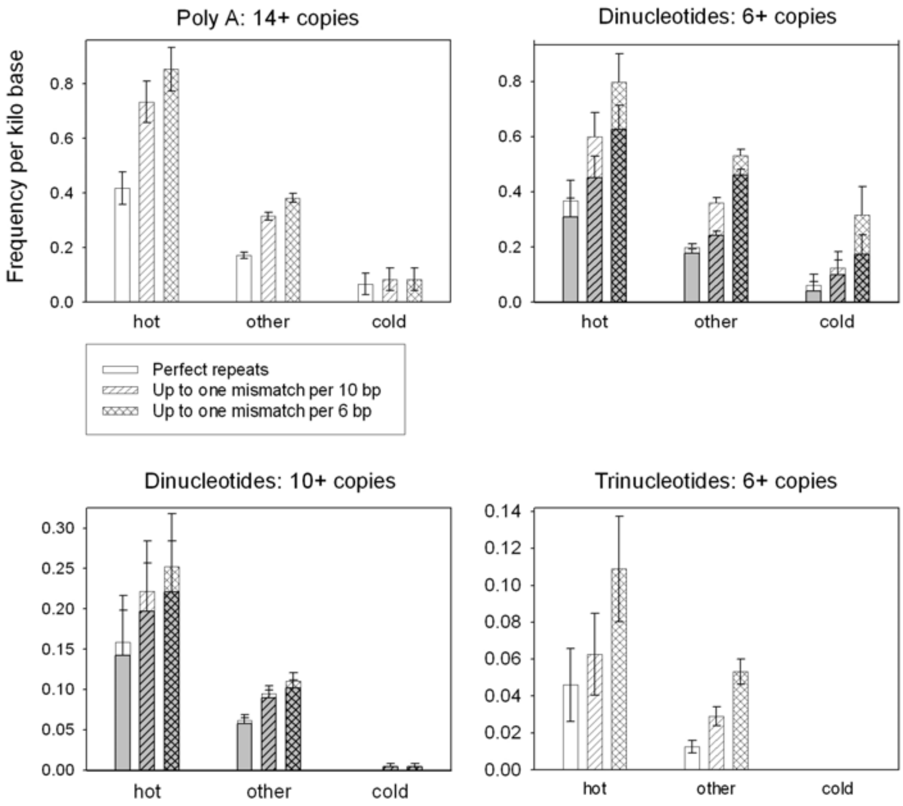


Figure 2.1: Level of enrichment of high-copy, short-motif microsatellites in *S. cerevisiae* recombination hotspots
Mean microsatellite frequencies in *S. cerevisiae* IGRs divided according to DSB intensity into 473 hot, 89 cold and 5431 other regions, which were all IGRs not categorized as either hot or cold. Poly-AT arrays comprised the majority of dinucleotide repeats and are highlighted in grey. Error bars are plus and minus one SEM.

2.3.3 Properties of hotspot-associated microsatellites

I compared microsatellite array length and purity (number of mismatches with respect to the consensus repeated motif) for repeats of at least six copies in hotspots and other regions of the yeast genome. I also compared the frequencies of insertion, substitution and deletion mismatches, with respect to the consensus repeated motifs, between hotspot-associated microsatellites and those in other regions. I found that poly-A and poly-G arrays are significantly longer in hot IGRs, but I saw no other significant differences in repeat length (Appendix A, Table A3). Microsatellites in hot and non-hot regions do not differ significantly in purity, but dinucleotide repeats in non-hot IGRs do show an elevated proportion of deletion mismatches (p=0.0006, Mann-Whitney U test).

To see whether any particular microsatellite motifs were associated with hotspots, I compared the most common motifs present in hot and non-hot regions. There are no obvious associations, but an extremely high over-representation of poly-purine/poly-pyrimidine motifs with only one G or C is clear among the most common motifs for low copy repeats in both hot and non-hot IGRs, and, interestingly also ORFs (Appendix A, Tables A4-A7). This is probably related to the enrichment of poly-purine/poly-pyrimidine tracts (PPTs) in the genome as a whole [18], and PPTs with internal tandem repeats comprise only a small proportion of total PPTs [19].

The GC-content of all repeats with at least six copies is strikingly low in IGRs throughout the genome, but there are no significant differences between hot and non-hot regions for microsatellite GC-content (Appendix A, Table A8). The low GC-content of microsatellites in yeast recombination hotspots is therefore due to the overall predominance of low-GC content microsatellites genome-wide.

2.3.4 Possible complicating factors

The influence of microsatellites on transcriptional frequency [20-26], and the mutagenic effect of transcription on microsatellites [27] suggest that factors relating to gene expression could affect microsatellite distribution. Theoretically, this could underlie the association between microsatellites and recombination hotspots in yeast, since transcriptional frequency (vegetative cells [15]) correlates with DSB intensity ($p<0.0001$, Spearman’s rank test). However, looking at the “hottest” IGRs and ORFs for transcriptional frequency in equivalent numbers to the numbers of recombination hot regions studied, I found that these regions overlap with recombination hotspots slightly less often than expected by chance, and the correlations between DSB intensity and frequency of microsatellites change very little when controlling for transcriptional frequency in partial correlation analysis (Appendix A, Table A2). DSBs have been shown to be more frequent in IGRs with two promoters (divergent transcription of flanking genes) than those with one (parallel transcription of flanking genes) or none (convergent transcription of flanking genes) [1]. I found that densities of some types of microsatellite do differ among IGRs with different numbers of promoters (Appendix A, Table A9). Significant differences are not present for longer microsatellites, however, with the exception of dinucleotide repeats, which are more common in IGRs with no promoters, though not significantly so when testing hot IGRs only. The level of enrichment of poly-A in hot over non hot IGRs does not differ by more than 5% between regions with zero, one and two promoters (two, one and zero 3’ untranslated regions respectively), so the association between poly-A and hotspots is not due to factors relating to the poly-A adenylation signal present in 3’ untranslated regions (UTRs).

Another factor that could complicate the association between hotspots and microsatellites is complex (tightly bunched or highly degenerate) repeats. My initial analysis left open this possibility, since my repeat-finding algorithm does not allow multiple consecutive mismatches within single microsatellites. I therefore looked at numbers of repeats within five and ten bp of other repeats, and compared levels of these complex microsatellites between hot and non-hot IGRs (Appendix A, Table A10). I found that numbers of microsatellites within complex repeats in IGRs are similar in hot and non-hot, or somewhat higher in non-hot, regions. Degenerate or complex repeats do not, therefore, affect the association between microsatellites and hot IGRs.

2.3.5 Microsatellite frequencies in hotspot flanking regions

Poly-purine/poly-pyrimidine tracts are enriched in hotspot flanking regions as far as two ORFs removed from hotspots (see Section 4.2.3). For microsatellites, however, there is no consistent evidence for a similar regional enrichment in these datasets (Appendix A, Table A11). This suggests that the association with recombination hotspots is less broad in scale for microsatellites than for PPTs. It is also possible, however, that the lower relative abundance of microsatellites could obscure a more general broad scale association than I was able to detect, since several repeat types have higher mean frequencies in hotspot flanking regions but the data are too sparse for statistical significance.

2.4 Discussion

The work presented in this chapter constituted the first published study on the relationship between microsatellites of all types, including low-copy repeats, and recombination hotspots, and the first report of a general enrichment of microsatellites in hotspots in *S. cerevisiae* [28] (see Page xi: Publications associated with this thesis). The level of enrichment is surprisingly high given that strong associations between microsatellites and recombination hotspots in other species have not been reported, and sequence features, including simple sequences, have not been found to be good predictors of fine scale recombination rates in humans [29, 30]. I showed that poly-A, poly-AT and other AT-rich repeats are highly over-represented in yeast hotspots, but A/T richness is a property of microsatellites throughout the genome and there is no significant difference between hot and non-hot IGRs for microsatellite GC-content. In humans, poly-AT and poly A are under-represented in highly recombining regions [30, 31] suggesting that this could partly explain the relative strength of the association between total microsatellites and hotspots in yeast (see Section 3.4). I also found, however, that poly-AC, poly-AG and poly-G are associated with yeast hot IGRs (Appendix A Table A1), so some effect on the distribution of microsatellites in general, and not just AT-rich ones, is operating in yeast to cause their over-representation in hotspots.

The association between microsatellites and recombination hotspots could be generated by a causal link between in the two, or it could be mediated by other factors coincident with both features. These could potentially include transposable elements, GC-content variations and the process of transcription, since this requires an opening of chromatin structure which might stimulate recombination. Known transposable elements are not an important factor in yeast, since they are not over-represented in hotspots from the dataset I used [1]. Effects of transcription or promoter regions also appear to be negligible, based on the observations that controlling for transcriptional frequency has no effect on the correlation between microsatellite abundance and DSB intensity, and that microsatellites are not more common in the most active promoter regions, or in IGRs with two promoters compared to those with one or none. Controlling for GC-content also does not reduce the magnitude of the correlation between microsatellite frequency and DSB intensity, except in the case of short poly-G runs. Some unknown feature of large-scale chromosome structure could perhaps favour both recombination and microsatellite formation, but a test of this possibility is not achievable with currently available data.

The results presented here therefore suggest the existence of a widespread causal link between microsatellites and the process of meiotic recombination in yeast. The two most obvious forms this link could take are a mutation bias caused by recombination acting to promote microsatellite formation and/or growth, and regulation of hotspot locations by simple sequences. I attempted to isolate evidence for a mutagenic effect of recombination on microsatellites by investigating low-copy repeats. These are not likely to be substantially affected by replication slippage, since this requires formation of a stable loop between newly replicated and template DNA strands [3-5], and there is no available evidence suggesting that short microsatellites could stimulate recombination. I did not find clear associations with hotspots for low-copy repeats, however, and while this does suggest that unequal recombination is not involved in microsatellite formation, previous evidence indicates that longer microsatellites have the potential to stimulate recombination [17, 32, 33], as well as to be mutated by it [34].

The question of whether either of these possible explanations has widespread importance has not been much explored. Studies at a chromosomal level in *S. cerevisiae* have shown that poly-A [17], poly-AC [33, 35] and pentanucleotide [32] arrays can affect meiotic recombination, and stimulation of recombination between plasmids by various types of repetitive sequence has been reported by numerous studies [36-43]. Only one extant meiotic recombination hotspot is currently known to be dependent on a microsatellite for full activity [17], and my data suggest that a substantial proportion of other hotspots might also be regulated by microsatellites, though other factors must simultaneously be involved in hotspot regulation (see Section 1.2.2). The existence of hotspots without local microsatellites does not rule out a functional role for the sequences in recombination, since it has been established that mechanisms of hotspot regulation are heterogeneous with respect to the role of local sequences [44-46] (see Section 1.2). Observations that meiotic DSBs are not sequence-specific [47-50] also do not rule out a role for microsatellites in regulating their locations; DSBs are known to avoid poly-A [48, 49], but poly-A can stimulate hotspot recombination [17]. Moreover, a role in regulating recombination events for sequences at distances as great as 11.5 kb has been demonstrated [51].

High frequencies of microsatellites in some regions outside hotspots show that the presence of a microsatellite is not sufficient to cause a hotspot. They do not, however, constitute conclusive evidence against the functional involvement of microsatellites in a substantial proportion of hotspots, since hotspots require multi-levelled processes acting in concert, including local sequences such as transcription factor binding sites, and local chromatin structure modifications, as well as contextual factors, which may include distal sequences and/or large scale chromosome structure (reviewed in [44-46]) (see Section 1.2). The ability of microsatellites to bind transcription factors [52], and to modulate chromatin structure [32, 53-58], therefore suggest two ways in which they could function to potentiate hotspot recombination, under the alpha and beta models respectively [44] (see Section 1.2.2). Another potential mechanism for microsatellite functionality in hotspots is replication pausing, since this may be causally involved in some recombination hotspots [44], and experimental evidence has linked it to microsatellites, in which it can also cause mutations [59-61]. Microsatellites could also, therefore, be functionally involved in hotspots under the gamma model [44] (see Section 1.2.2). Moreover, the obvious possibility that simple sequences in general could facilitate homologous recombination by helping to guide the recombinant DNA molecules into register is widely recognized.

It is also plausible that recombination is involved in some proportion of microsatellite mutations, and this could partly or fully drive the association between microsatellites and recombination hotspots even though recombination is apparently not commonly involved in microsatellite formation, based on my observation that two-copy repeats are not over-represented in hotspots. In model organisms, evidence has been found both for [33, 34, 62] and against [63, 64] a role for recombination, in the mutation of different types of microsatellite. The vast, presently unexplained, differences in mutation rates between loci (reviewed in [10, 65]) suggest the possible involvement of heterogeneous mutational mechanisms or regional mutation biases. If recombination is a factor in microsatellite variation, the further study of this relationship could potentially lead to more accurate prediction of the mutation rates of microsatellites, and consequently facilitate their use as genetic markers (see Section 1.1.1). In Chapter 5 I explore the possibility that microsatellite polymorphism is increased in frequently recombining areas of the human genome.

References

1. Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD: **Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae**. *Proc Natl Acad Sci U S A* 2000, **97**(21):11383-11390.
2. http://repeatfinder.sourceforge.net/.
3. Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B: **Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat**. *Am J Hum Genet* 1998, **62**(6):1408-1415.
4. Weber JL: **Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms**. *Genomics* 1990, **7**(4):524-530.
5. Zhu Y, Queller DC, Strassmann JE: **A phylogenetic perspective on sequence evolution in microsatellite loci**. *J Mol Evol* 2000, **50**(4):324-338.
6. Messier W, Li SH, Stewart CB: **The birth of microsatellites**. *Nature* 1996, **381**(6582):483.
7. Pupko T, Graur D: **Evolution of microsatellites in the yeast Saccharomyces cerevisiae: role of length and number of repeated units**. *J Mol Evol* 1999, **48**(3):313-316.

8. Zhu Y, Strassmann JE, Queller DC: **Insertions, substitutions, and the origin of microsatellites**. *Genet Res* 2000, **76**(3):227-236.
9. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution**. *Mol Biol Evol* 1987, **4**(3):203-221.
10. Ellegren H: **Microsatellites: simple sequences with complex evolution**. *Nat Rev Genet* 2004, **5**(6):435-445.
11. Smith GP: **Evolution of repeated DNA sequences by unequal crossover**. *Science* 1976, **191**(4227):528-535.
12. Baudat F, Nicolas A: **Clustering of meiotic double-strand breaks on yeast chromosome III**. *Proc Natl Acad Sci U S A* 1997, **94**(10):5213-5218.
13. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex**. *Nat Genet* 2001, **29**(2):217-222.
14. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: **Human recombination hot spots hidden in regions of strong marker association**. *Nat Genet* 2005, **37**(6):601-606.
15. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome**. *Cell* 1998, **95**(5):717-728.
16. ftp://genome-ftp.stanford.edu/pub/yeast/.
17. Schultes NP, Szostak JW: **A poly(dA.dT) tract is a component of the recombination initiation site at the ARG4 locus in Saccharomyces cerevisiae**. *Mol Cell Biol* 1991, **11**(1):322-328.
18. Raghavan S, Burma PK, Brahmachari SK: **Positional preferences of polypurine/polypyrimidine tracts in Saccharomyces cerevisiae genome: implications for cis regulation of gene expression**. *J Mol Evol* 1997, **45**(5):485-498.
19. Bagshaw AT, Pitt JP, Gemmell NJ: **Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots**. *BMC Genomics* 2006, **7**:179.
20. Struhl K: **Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast**. *Proc Natl Acad Sci U S A* 1985, **82**(24):8419-8423.
21. Uhlemann AC, Szlezak NA, Vonthein R, Tomiuk J, Emmer SA, Lell B, Kremsner PG, Kun JF: **DNA phasing by TA dinucleotide microsatellite length determines in vitro and in vivo expression of the gp91phox subunit of NADPH oxidase and mediates protection against severe malaria**. *J Infect Dis* 2004, **189**(12):2227-2234.
22. Curi RA, Oliveira HN, Silveira AC, Lopes CR: **Effects of polymorphic microsatellites in the regulatory region of IGF1 and GHR on growth and carcass traits in beef cattle**. *Anim Genet* 2005, **36**(1):58-62.
23. Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M: **A polymorphic microsatellite that mediates induction of PIG3 by p53**. *Nat Genet* 2002, **30**(3):315-320.
24. Borrmann L, Seebeck B, Rogalla P, Bullerdiek J: **Human HMGA2 promoter is coregulated by a polymorphic dinucleotide (TC)-repeat**. *Oncogene* 2003, **22**(5):756-760.
25. Hammock EA, Young LJ: **Microsatellite instability generates diversity in brain and sociobehavioral traits**. *Science* 2005, **308**(5728):1630-1634.
26. Hammock EA, Young LJ: **Functional microsatellite polymorphism associated with divergent social structure in vole species**. *Mol Biol Evol* 2004, **21**(6):1057-1063.

27. Wierdl M, Greene CN, Datta A, Jinks-Robertson S, Petes TD: **Destabilization of simple repetitive DNA sequences by transcription in yeast**. *Genetics* 1996, **143**(2):713-721.
28. Bagshaw AT, Pitt JP, Gemmell NJ: **High frequency of microsatellites in S. cerevisiae meiotic recombination hotspots**. *BMC Genomics* 2008, **9**(1):49.
29. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome**. *Science* 2004, **304**(5670):581-584.
30. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**(5746):321-324.
31. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G *et al*: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**(3):241-247.
32. Kirkpatrick DT, Wang YH, Dominska M, Griffith JD, Petes TD: **Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes**. *Mol Cell Biol* 1999, **19**(11):7661-7671.
33. Gendrel CG, Boulet A, Dutreix M: **(CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis**. *Genes Dev* 2000, **14**(10):1261-1268.
34. Jakupciak JP, Wells RD: **Genetic instabilities of triplet repeat sequences by recombination**. *IUBMB Life* 2000, **50**(6):355-359.
35. Treco D, Arnheim N: **The evolutionarily conserved repetitive sequence d(TG.AC)n promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis**. *Mol Cell Biol* 1986, **6**(11):3934-3947.
36. Murphy KE, Stringer JR: **RecA independent recombination of poly[d(GT)-d(CA)] in pBR322**. *Nucleic Acids Res* 1986, **14**(18):7325-7340.
37. Bullock P, Miller J, Botchan M: **Effects of poly[d(pGpT).d(pApC)] and poly[d(pCpG).d(pCpG)] repeats on homologous recombination in somatic cells**. *Mol Cell Biol* 1986, **6**(11):3948-3953.
38. Napierala M, Dere R, Vetcher A, Wells RD: **Structure-dependent recombination hot spot activity of GAA.TTC sequences from intron 1 of the Friedreich's ataxia gene**. *J Biol Chem* 2004, **279**(8):6444-6454.
39. Wahls WP, Wallace LJ, Moore PD: **The Z-DNA motif d(TG)30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture**. *Mol Cell Biol* 1990, **10**(2):785-793.
40. Napierala M, Parniewski P, Pluciennik A, Wells RD: **Long CTG.CAG repeat sequences markedly stimulate intramolecular recombination**. *J Biol Chem* 2002, **277**(37):34087-34100.
41. Jankowski C, Nag DK: **Most meiotic CAG repeat tract-length alterations in yeast are SPO11 dependent**. *Mol Genet Genomics* 2002, **267**(1):64-70.
42. Sutherland GR, Baker E, Richards RI: **Fragile sites still breaking**. *Trends Genet* 1998, **14**(12):501-506.
43. Freudenreich CH, Kantrow SM, Zakian VA: **Expansion and length-dependent fragility of CTG repeats in yeast**. *Science* 1998, **279**(5352):853-856.
44. Petes TD: **Meiotic recombination hot spots and cold spots**. *Nat Rev Genet* 2001, **2**(5):360-369.
45. Nishant KT, Rao MR: **Molecular features of meiotic recombination hot spots**. *Bioessays* 2006, **28**(1):45-56.

46. Kauppi L, Jeffreys AJ, Keeney S: **Where the crossovers are: recombination distributions in mammals**. *Nat Rev Genet* 2004, **5**(6):413-424.

47. Xu L, Kleckner N: **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* 1995, **14**(20):5115-5128.

48. Liu J, Wu TC, Lichten M: **The location and structure of double-strand DNA breaks induced during yeast meiosis: evidence for a covalently linked DNA-protein intermediate**. *Embo J* 1995, **14**(18):4599-4608.

49. de Massy B, Rocco V, Nicolas A: **The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in Saccharomyces cerevisiae**. *Embo J* 1995, **14**(18):4589-4598.

50. Xu F, Petes TD: **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* 1996, **143**(3):1115-1125.

51. Haring SJ, Halley GR, Jones AJ, Malone RE: **Properties of natural double-strand-break sites at a recombination hotspot in Saccharomyces cerevisiae**. *Genetics* 2003, **165**(1):101-114.

52. Lu Q, Teare JM, Granok H, Swede MJ, Xu J, Elgin SC: **The capacity to form H-DNA cannot substitute for GAGA factor binding to a (CT)n*(GA)n regulatory site**. *Nucleic Acids Res* 2003, **31**(10):2483-2494.

53. Prunell A: **Nucleosome reconstitution on plasmid-inserted poly(dA) . poly(dT)**. *Embo J* 1982, **1**(2):173-179.

54. Elgin SC: **The formation and function of DNase I hypersensitive sites in the process of gene activation**. *J Biol Chem* 1988, **263**(36):19259-19262.

55. Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin**. *Annu Rev Biochem* 1988, **57**:159-197.

56. Iyer V, Struhl K: **Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure**. *Embo J* 1995, **14**(11):2570-2579.

57. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA**. *J Mol Biol* 1986, **191**(4):659-675.

58. Otten AD, Tapscott SJ: **Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure**. *Proc Natl Acad Sci U S A* 1995, **92**(12):5465-5469.

59. Fouche N, Ozgur S, Roy D, Griffith JD: **Replication fork regression in repetitive DNAs**. *Nucleic Acids Res* 2006, **34**(20):6044-6050.

60. Hile SE, Eckert KA: **Positive correlation between DNA polymerase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences.** *J Mol Biol* 2004, **335**(3):745-759.

61. Hile SE, Eckert KA: **DNA polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellites sequences**. *Nucleic Acids Res* 2008, **36**(2):688-696.

62. Hashem VI, Rosche WA, Sinden RR: **Genetic recombination destabilizes (CTG)n.(CAG)n repeats in E. coli**. *Mutat Res* 2004, **554**(1-2):95-109.

63. Henderson ST, Petes TD: **Instability of simple sequence DNA in Saccharomyces cerevisiae**. *Mol Cell Biol* 1992, **12**(6):2749-2757.

64. Levinson G, Gutman GA: **High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12**. *Nucleic Acids Res* 1987, **15**(13):5323-5338.

65. Ellegren H: **Microsatellite mutations in the germline: implications for evolutionary inference**. *Trends Genet* 2000, **16**(12):551-558.

Chapter 3

Microsatellites in general are not over-represented in human meiotic recombination hotspots

Abstract

Aspects of the association between microsatellites and recombination in the human genome not examined by previous studies include its scale and magnitude, and the question of whether the association is mediated by other factors has not been addressed elsewhere. Surprisingly, in view of the high frequency of microsatellites in yeast double-strand break hotspots (see Chapter 2), I found no significant differences in microsatellite frequency between meiotic recombination hotspots and coldspots in humans, though there is a modest 10-20% enrichment of microsatellites with 2-5 bp repeat motifs in human hotspot central and flanking regions. The correlation between microsatellite abundance and recombination rate at a fine scale of 1 kb over 37 regions of the human genome, each of 32.8 mega bases, is very weak and inconsistent whether or not other factors expected to correlate with both variables are considered. Having found virtually no correlation between recombination and microsatellites at this fine scale, I used scale-specific wavelet correlation analysis to address the question of whether previously reported broad scale correlations between microsatellite abundance and recombination rate are due to factors operating on a larger scale. I found no substantial correlations at scales of one mega base or less, suggesting that the previously noted correlations could be due to factors operating on a scale larger than one mega base.

3.1 Introduction

It has been shown previously that microsatellite frequency correlates with recombination rate in rats, mice and humans at scales of five and ten mega bases (mb) [1]. Other studies have noted the presence of microsatellites in human [2, 3], and mouse [4] recombination hotspots. However, while one study showed an enrichment of poly-AG in hotspots and an enrichment of poly-AT in coldspots [5], it has not previously been reported whether microsatellites in general are associated with recombination at the level of hotspots genome-wide in any mammal. This question is of interest insofar as it is currently believed that all types of microsatellite mutate predominantly by replication slippage errors (reviewed in [6, 7]), though recombination has been implicated in disease-causing radical expansions of trinucleotide repeats (reviewed in [8, 9]). Previous studies have also not addressed the influences of scale, or other genomic factors, on the correlation between microsatellite frequency and recombination rate.

I initially compared microsatellite frequency between human hotspots and cold regions. I then used generalized linear models to examine the correlation between recombination rate and microsatellite abundance at a fine scale of one kilo base (kb), and to investigate the influence of mediating factors. The possible mediating factors I considered were GC-content, single nucleotide polymorphism (SNP) density and gene (exon) coverage, all of which were expected to correlate with both recombination and simple sequences in general, for the following reasons. Crossover locations from the study I utilized [5] were mapped using SNPs, so regions with very low SNP density could be less likely to contain mapped hotspots, and could also harbour lower or higher microsatellite frequencies. Microsatellite growth may be stunted by single base mutations interrupting repeat arrays, since this should reduce the opportunity for replication slippage [10], possibly resulting in lowered microsatellite frequencies in genetically diverse regions with a high frequency of SNPs. On the other hand, such regions might also correlate with a higher density of microsatellites, assuming they are less selectively constrained and therefore more likely to contain neutrally evolving simple sequences. GC-content could also vary with microsatellite abundance for the same reasons, because it correlates with genetic diversity, and also with recombination, in the human genome [11]. Moreover, it is possible that high GC-content recombination hotspots could replicate more slowly [12], potentially allowing more time for replication slippage mutations to occur in microsatellites. I also considered gene density, because recombination rate is elevated in regions with high gene density on a broad scale, but is lowered in close proximity to genes [5, 13], and it is well known that microsatellite abundance is very low in coding sequence, so the distribution of coding exons could also affect the correlation between microsatellites and recombination.

I investigated the scale of the association between recombination and microsatellites in the human genome in two ways. First, I examined microsatellite abundance in hotspot central and flanking regions. The central areas of hotspots are of particular interest in view of evidence that crossover frequencies increase markedly toward the mid points of hotspots in mammals [14, 15], and recombination-initiating double-stranded breaks (DSBs) are concentrated in 100-500 bp regions in yeast [16-18]. Moreover, sequence features of hotspot flanking regions have apparently not been examined by any previous study, and hotspot function can depend on flanking sequence at a distance of up to 11.5 kb [19]. Second, I examined scale-specific correlations between microsatellite frequency and recombination rate using wavelet analysis, a technique commonly employed for many applications in the analysis of frequency data [20]. Recently, it was adapted to the investigation of correlations between genomic features by Spencer *et al.*, [11], and I utilized aspects of their methodology here. An attractive feature of their method with respect to my question of correlation scale is that the wavelet transformation they used divides the variance in a dataset into specific scales in such a way that the effects of a particular scale can be deduced independently from those of other scales [20].

3.2 Methods

3.2.1 Sequence and annotation data used

I initially analysed microsatellite frequencies in 17 human meiotic recombination hotspots, which were the only ones to have been experimentally well characterized in humans, using sperm typing, (see Section 1.2.1), at the beginning of 2008. Two studies have reported multiple, well defined hotspots across broad contiguous regions. These are located in the MHC Class II region on human chromosome 6, in which seven hotspots were mapped over 292 kb [21-23], and in a 206 kb segment of human chromosome 1, in which eight hotspots were identified [24, 25]. In each region, areas between hotspots showed little or no evidence for recombination, and I define those areas here as cold regions. Two other human hotspots have been characterized experimentally by the mapping of substantial numbers of crossovers. These are located in the Beta-Globin gene cluster [26], adjacent to which a 90 kb non-recombining region has been identified [27] and in the pseudoautosomal region of the Y chromosome, in which a 9.9 kb section of the *SHOX* gene was assayed for recombination and found to contain a hotspot [28]. Overall, these 17 experimentally characterized hotspots average 1570 bp in length.

I also investigated microsatellite frequencies in relation to the fine-scale map of hotspots and recombination rates throughout the human genome published in 2005 by Myers and colleagues [5]. This recombination map was based on the use of statistical analysis of haplotype data to infer past recombination events, so I considered it separately because some evidence indicates that these methods are not always able to predict hotspots in the present generation [22, 25, 29, 30] (see Section 1.2.1), with a recent fine scale map of directly observed recombination rates suggesting that they are about 60 % accurate [13]. The advantage of the dataset from Myers *et al*., [5] is its size, with 9299 hotspots mapped to within 5 kb and an equivalent number of defined cold regions of equivalent size and SNP density, the genomic locations of all of which are available online [31].

The GenBank accession numbers for the human experimentally characterized hotspot sequences are: Beta-Globin hot spot: GI:37541814, Chromosome 1 hotspots: GI:37549514 and *SHOX* hotspot: U82668. For the MHC hot spots I used the 28 October 1999 version of the MHC class II region sequence, since that was the version to which the reported hotspot locations corresponded [21]. This version is available at the Sanger Centre website. I converted the sequence coordinates of all human genomic features studied into the latest genome build (NCBI Build 36/HG18) using the UCSC genome browser’s liftover facility [32]. This provided consistency in genome build between all the different variables included in the generalized linear model. Only one hotspot location of the 9299 hotspots originally mapped to within 5 kb [5] was lost in the conversion process. I obtained GC-content for the studied regions using software written in C [33]. I downloaded the sequence of the human genome (HG18), and exon locations, from the UCSC Genome Browser [34]. The SNP dataset I used [35] was the same as in the work presented in Chapter 5, and the rationale behind its employment is discussed there (Section 5.2.1).

3.2.2 Wavelet analysis

For wavelet and generalized linear model analyses, I first translated the locations of all genomic features of interest into non-overlapping 1 kb bins covering the entire genome. I did this using the galaxy bioinformatics toolset available from the UCSC Genome Browser, which was downloaded onto a stand-alone supercomputer. An average recombination rate for each bin was produced using a Java script (Lu LU, unpublished). I performed wavelet analysis in R (version 2.6.0) using scripts adapted from the analysis by Spencer *et al.*, [11]. The analysis requires that all variables have values across each of 2^n contiguous bins with no breaks in the data. Gaps in the human genome sequence build, separate chromosomes, and other gaps caused by failure of some recombination rate regions to convert between the latest genome build (HG18) and their original HG16 therefore necessitate separate analyses. A total of 37 regions for which there are data for each variable for 2^{15} contiguous bins (kb), are possible with the currently available data (Table 3.1). These included samples from 16 different chromosomes and a reasonable selection of short arm, long arm, centromere-proximal and telomere-proximal regions. I selected a region size of 2^{15} kb (32.8 mega bases) in preference to other possible sizes for three reasons. Firstly, it was employed in the study by Spencer *et al.*, on which my wavelet analysis method was based [11]. Secondly, only eight regions of 2^{16} kb would have been possible with the data available due to the need for contiguity in all variables. Thirdly, while smaller regions would have enabled more replication, and greater overall coverage of the genome, power to detect broad scale correlations obviously declines with decreasing size of studied regions.

Table 3.1: Regions of the human genome investigated with wavelet analysis
Showing locations, in HG18 coordinates (NCBI Build 36), of the 37 regions of 2^{15} kb for which contiguous data are available for all variables used in my analysis. The numbering scheme was based on descending order of the overall length of contiguous regions, which were then divided into non-overlapping sub-regions of 2^{15} kb each. Regions within 10 mega bases of a centromere or telomere are labelled as near to the feature.

Region no.	Chromosome	Start	End	Chromosome region
1	3	94988000	127756000	long arm, near centromere
2	3	127756000	160524000	mid long arm
3	3	160524000	193292000	long arm, near telomere
4	2	149499000	182267000	mid long arm
5	2	182267000	215035000	mid long arm
6	4	75672000	108440000	mid long arm
7	4	108440000	141208000	mid long arm
8	14	19166000	51934000	long arm, near centromere
9	14	51934000	84702000	mid long arm
10	1	29801000	62569000	mid short arm
11	1	62569000	95337000	mid short arm
12	2	21038000	53806000	mid short arm
13	2	53806000	86574000	short arm, near centromere
14	13	17921000	50689000	long arm, near centromere
15	13	50689000	83457000	mid long arm
16	3	47000	32815000	short arm, near telomere
17	3	32815000	65583000	mid short arm
18	6	95938000	128706000	mid long arm
19	12	75042000	107810000	mid long arm
20	8	86852000	119620000	mid long arm
21	1	147777000	180545000	mid long arm
22	15	26997000	59765000	long arm, near centromere
23	7	74604000	107372000	mid long arm
24	11	1170000	33938000	short arm, near telomere
25	5	74000	32842000	short arm, near telomere
26	10	81242000	114010000	mid long arm
27	5	49442000	82210000	long arm, near centromere
28	5	97613000	130381000	mid long arm
29	9	91719000	124487000	mid long arm
30	9	37000	32805000	short arm, near telomere
31	12	36143000	68911000	long arm, near centromere
32	11	95943000	128711000	long arm, near telomere
33	2	111009000	143777000	mid long arm
34	8	48310000	81078000	long arm, near centromere
35	7	478000	33246000	short arm, near telomere
36	18	16765000	49533000	long arm, near centromere
37	6	62237000	95005000	long arm, near centromere

Mathematically and computationally, I used the same wavelet analysis method described by Spencer and colleagues for investigating pair-wise correlations between wavelet decompositions of frequency variables (detail coefficients) [11]. The method employs the simplest discrete wavelet transformation known as the Haar wavelet function, which transforms a series of observations into a series of detail coefficients representing the difference between pairs of neighbouring observations, and also a smoothed version of the original signal [20]. I did not utilize the smoothed transformation here, because only the detail coefficients are relevant to the question of scale-specific correlation [11]. The advantage of the wavelet method I employed is that it divides the variance of correlates such that the influence of particular scales on the correlation is independent of the influence of other scales.

3.3 Results

3.3.1 Microsatellites are not associated with experimentally well-characterized recombination hotspots in humans

I initially compared microsatellite frequencies between the 17 human experimentally well-characterized human hotspots listed above (Section 3.2.1) and the defined cold regions from these studies, making the same division of microsatellite types, and using the same computer algorithm, described in Chapter 2 (Section 2.2). I found no significant enrichment of microsatellites of any type in these hotspots, or in their central regions, which I defined as 500 bp centred on the point mid-way between hotspot start and end coordinates. An enrichment of poly-AG in human hotspots has previously been reported [5], but I found no difference between hot and cold regions of any type of dinucleotide repeat (divided into motif groups as in Section 2.2.3). The division into 19 types of microsatellite could obscure overall patterns, for example if each type contributes a small, non-significant amount to an overall enrichment of microsatellites in hotspots, so I made a second, four-way categorization. The four categories were mononucleotide repeats of less than six copies, mononucleotide repeats of six copies or more, 2-6 bp motif repeats of less than six copies and 2-6 bp repeats of six copies or more. (see Section 2.1 for rationale behind the six copy limit). I found no significant differences between hotspots and cold regions, or between hotspot central regions and cold regions, for any of these microsatellite classes. Frequencies are in fact very similar between the two types of region, or even slightly higher in cold areas for some microsatellite types. (Table 3.2).

Table 3.2: Microsatellite frequencies in experimentally characterized human meiotic recombination hotspots and their adjacent cold regions.
Mean frequencies of four classes of microsatellite in 17 well-characterized human meiotic recombination hotspots and their intervening cold regions. Hotspot central regions were defined as 500 bp centred on the hotspot mid point. The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus motif. A lower e value therefore results in the detection of more imperfect repeats.

Microsatellite type			**Mean per kb frequency**				
Motif length	Copy number	Mismatch type	Hotspots	Hotspot central regions	Cold regions	Freq. ratio (hot/cold)	P value
1	under 6	perfect	48.8	48.8	48.6	1.00	0.843
		e=10	48.4	48.5	48.3	1.00	0.772
		e=6	46.6	46.6	46.4	1.00	0.692
1	6 or more	perfect	1.50	1.53	1.82	0.821	0.135
		e=10	1.43	1.53	1.80	0.793	0.072
		e=6	2.15	2.59	2.58	0.832	0.393
2 to 6	under 6	perfect	48.7	50.5	46.0	1.06	0.247
		e=10	48.1	50	45.4	1.06	0.235
		e=6	45.4	47.1	43.7	1.04	0.286
2 to 6	6 or more	perfect	0.198	0.235	0.201	0.985	0.043
		e=10	0.348	0.235	0.349	0.997	0.273
		e=6	1.10	0.706	0.626	1.76	0.902

3.3.2 A modest elevation of microsatellite frequency in human hotspot central and flanking regions

Although the human experimentally defined hotspots I investigated are located on four different chromosomes, the above results could be questioned due to the small sample size of 17 hotspots. I therefore repeated the investigation for a genome-wide dataset of 9299 hotspots mapped with at least 5 kb resolution by Myers and colleagues [36]. My repeat-finding algorithm is not capable of analyzing regions larger than about 1.5 mb due to its detailed mechanism, even when run on a supercomputer. I therefore used microsatellite locations reported by the well-known TRF algorithm [37], which are available online at the UCSC genome browser [34]. The microsatellites in this dataset are 25 bp or longer, with some mismatches allowed, and almost all of the mononucleotide repeats it contains are poly-A. I did not investigate low-copy repeats in relation to human hotspots genome-wide, but the lack of an association between short microsatellites and yeast DSB, and human experimentally characterized, hotspots suggested that this would not have been informative.

I compared abundance of the TRF-reported microsatellites between hotspots from the genome-wide dataset and a coldspot dataset of equivalent size, which had been defined by the authors of the study [5]. This comparison revealed no significant difference between the two types of regions for microsatellites with 2-5 bp motifs, or for mononucleotide repeats ($p>0.05$, Mann-Whitney U Test). There are also no significant differences when considering microsatellite coverage, i.e. the number of bases within each region covered by microsatellites, which should reflect array length as well as frequency. Repeating the comparison for hotspot central and flanking regions against their cold equivalents did, however, reveal a modest, 10-20% enrichment of microsatellites with 2-5 bp motifs both near hotspot mid points, and to each side of hotspots ($p<0.01$; Figure 3.1, Tables 3.3 and 3.4).

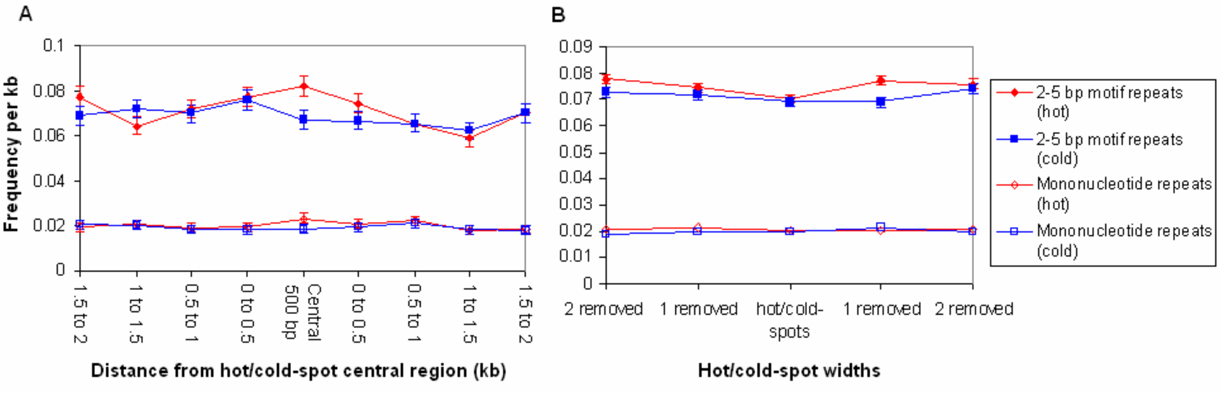


Figure 3.1: Distribution of microsatellites in relation to human hotspot central and flanking regions
Mean per kb frequency of microsatellites in relation to the central (A) and flanking (B) regions of human hotspots from the genome-wide dataset (mean hotspot width = 4070 bp). For the analysis of flanking regions, each hot/cold-spot was extended by one- (denoted “1 removed”) and two-fold (denoted “2 removed”) its own width on either side. In cases where this resulted in overlap between hot and cold areas, the cold ones were excluded from the analysis. Error bars are plus and minus one SEM or are not shown in cases where they are narrower than the symbol width.

Table 3.3 Enrichment of microsatellites in hotspot flanking regions
Ratios of mean microsatellite frequency in hotspots, hotspot flanking regions 0-1 and 1-2 hotspot widths removed from hotspots, and coldspots. Statistical comparisons were made with the Mann-Whitney U test (alpha = 0.05). Values for the means, and their standard errors, can be seen in Figure 3.1.

Microsat. motf length	Mean frequency ratio			P value (comparison with coldspots)		
	Hotspots/ coldspots	Hotspot flanks 0-1 widths removed/ coldspots	Hotspot flanks 1-2 widths removed/ coldspots	Hotspots	Hotspot flanks 0-1 widths removed	Hotspot flanks 1-2 widths removed
2-5 bp	1.01	1.10	1.11	n/s	0.0005	0.0082
1 bp	1.06	1.06	1.05	n/s	n/s	n/s

Table 3.4 Enrichment of microsatellites in hotspot central regions
Ratios of mean microsatellite frequency in hotspot central regions (denoted “hot CR”, defined as 500 bp centred on the hotspot mid point), hotspot non-central regions (denoted “hot non-CR”, defined as within hotspots but outside the central region) and coldspot central 500 bp regions (denoted “cold CR”). Statistical comparisons were made with the Mann-Whitney U test (alpha = 0.05). Values for the means, and their standard errors, can be seen in Figure 3.1.

Microsat. motf length	Mean frequency ratio		P value	
	Hot CR/ Hot non-CR	Hot CR/ cold CR	Hot CR vs Hot non-CR	Hot CR vs cold CR
2-5 bp	1.18	1.23	0.009	0.010
1 bp	1.18	1.27	n/s	N/s

3.3.3 The correlation between recombination rate and microsatellite frequency in the human genome

I found no significant correlation between microsatellite frequency and recombination rate among the 17 experimentally well-characterized human hotspots. In view of the small sample size, I extended the analysis to 37 separate 2^{15} kb (32.8 mb) regions of the human genome using wavelet analysis and generalized linear models. Initially, I used a generalized linear model to investigate the correlation between microsatellites and recombination rate at a scale of one kb. I then expanded the model to include other factors that could mediate the correlation (see Section 3.1). Microsatellites only occur once every 8-10 kb in the dataset I used, so their locations can be thought of as count data. The variance of this dataset is similar to its mean, so a generalized linear model with a poisson error distribution is indicated [38]. It is well known that the distribution of microsatellites is clustered in some genomic regions, so

I used a model with relaxed restrictions on data dispersion [39]. Results from this analysis are shown in Table 3.5. Recombination only significantly predicts microsatellites with two to five base pair motifs in two out of the 37 regions. Mononucleotide repeats are not significantly predicted by recombination rate, but they are positively predicted by GC-content in most of the studied regions.

Table 3.5 Generalized linear model predicting microsatellite abundance in the human genome
Results from a GLM analysis (quasipoisson family in R, link = log) predicting microsatellite frequency at a scale of one kb. The analysis was repeated for 37 regions of 2^{15}kb (32.8 mb) spanning 16 chromosomes in the human genome, and mean statistics over all regions are shown, with standard errors. Two types of model were employed, one with recombination rate on its own predicting microsatellite abundance (denoted “single”) and another with the additional predictors GC-content, exon coverage and SNP density (denoted “multiple”). The rightmost column shows numbers of regions with a significant positive (pos) or negative (neg) effect of the predictor (by Student’s T test; Bonferroni-corrected alpha=0.00135 for the single regression model and 0.000338 for the multiple model). Overall significance was calculated by Stouffer’s method [40] in cases where the direction of correlation was consistent across all regions, and “inc” indicates that some regions showed negative effects and others showed positive effects.

Motif Length	**Predictor**	**Regr. model**	**Estimated Coeff.**		**T**		**Pr(T>|t|)**	**# sig pos(neg)**
			Mean	Mean SEM	Mean	SEM		
1	Recombination	single	0.0005	0.013	0.169	0.206	inc	0(0)
		multiple	-0.0062	0.014	-0.319	0.189	inc	0(2)
	GC-content	multiple	4.14	0.540	7.60	0.498	$<10^{-300}$	25(0)
	Exons	multiple	-0.0016	0.0006	-2.93	0.176	$<10^{-55}$	0(8)
	SNPs	multiple	-0.045	0.0329	-1.34	0.213	inc	0(2)
2-5	Recombination	single	0.0090	0.0072	1.35	0.206	inc	2(0)
		multiple	0.0080	0.0072	1.23	0.225	inc	4(0)
	GC-content	multiple	0.0052	0.3563	-0.0181	0.388	inc	4(9)
	Exons	multiple	-0.0011	0.0003	-3.59	0.191	$<10^{-93}$	0(18)
	SNPs	multiple	0.068	0.013	5.31	0.664	inc	26(2)

To investigate the correlation between microsatellite abundance and recombination rate at scales greater than one kb, I used wavelet analysis. Figures 3.2 and 3.3 show plots of the scale-specific correlation coefficient (Kendall’s rank test) at scales between two kb and one mb, increasing in exponentials of two. I performed these correlations using a wavelet detail coefficient method, which divides variance in a sample into scales so that correlations at particular scales are independent of those at other scales [11, 20]. The analysis revealed that no scale is responsible for more than a negligibly weak correlation between microsatellites and recombination rate when all regions are considered, though there is a slight upward trend in the correlation coefficients at the one mb scale for microsatellits with 2-5 bp motifs (Figures 3.2 and 3.3). To test whether the absence of significant correlations is due to a lack of statistical power resulting from the sparse distribution of microsatellites, I repeated the pair-wise non-parametric correlation analysis using other variables shown by multiple regression to predict microsatellite abundance. I found that microsatellites with 2-5 bp repeat motifs are significantly correlated with SNP density at a scale of 2 kb in 21 out of 37 regions, and 35 out of 37 regions show significant positive correlations at the 2 kb level between mononucleotide repeats and GC-content.

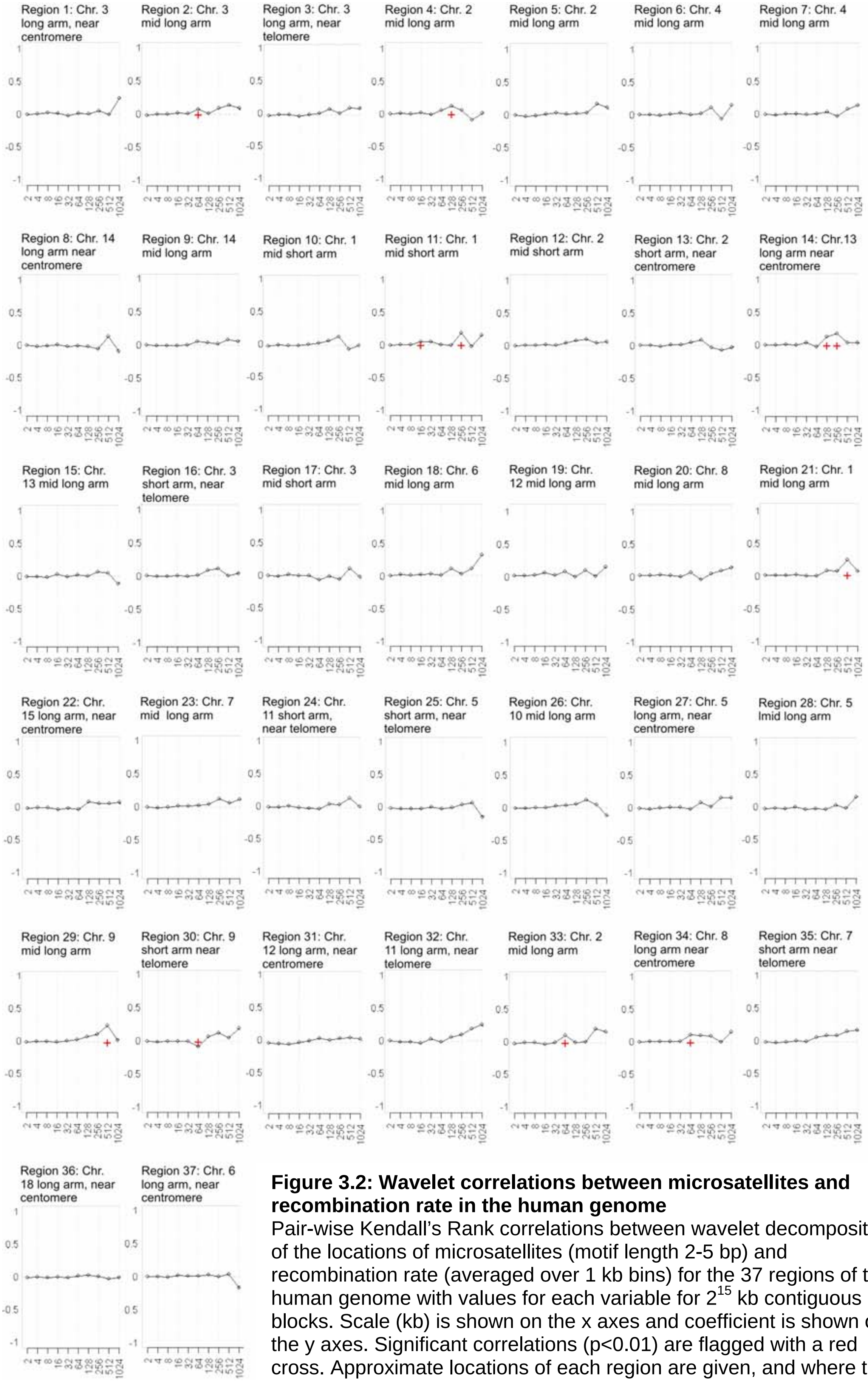
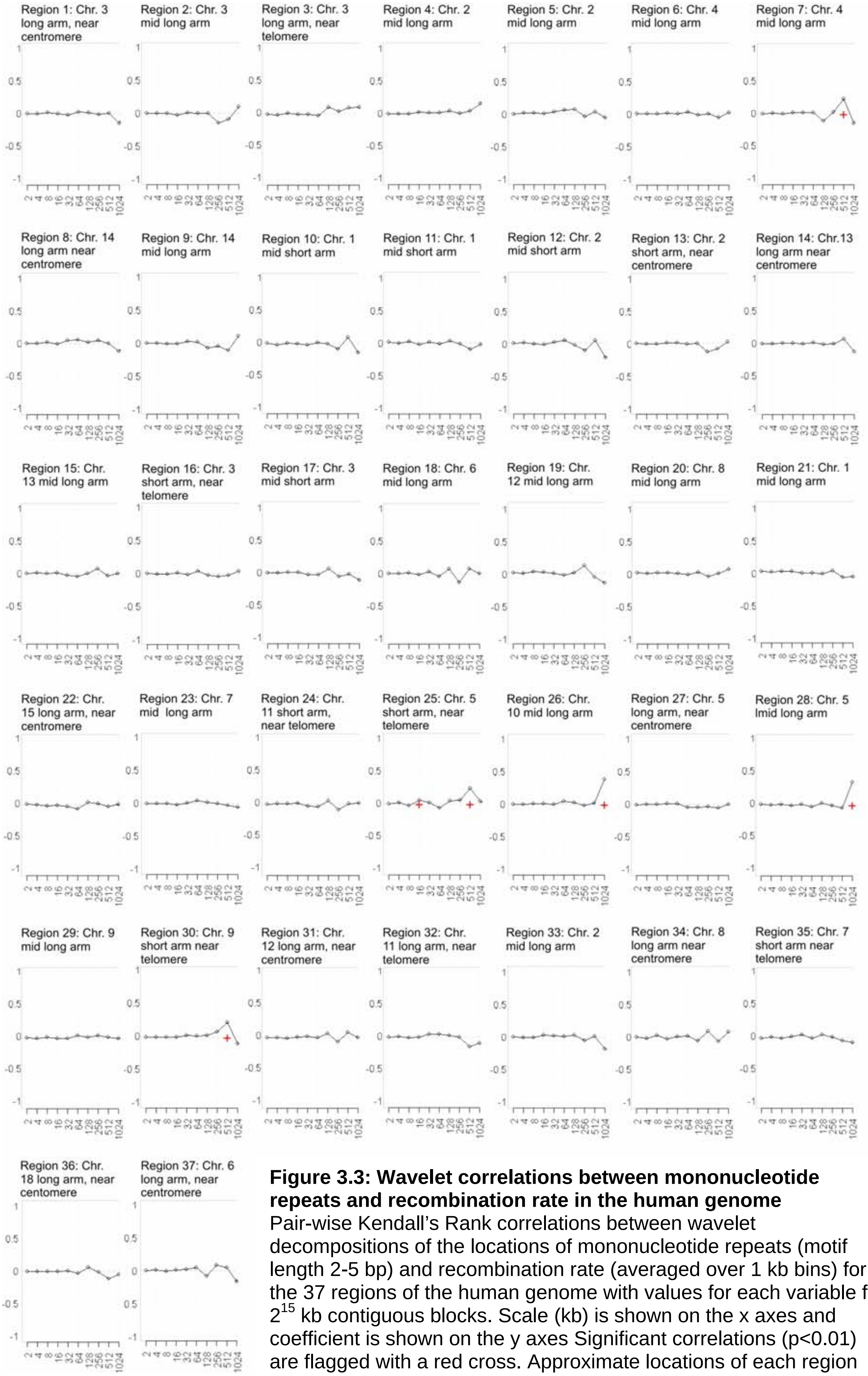


Figure 3.2: Wavelet correlations between microsatellites and recombination rate in the human genome
Pair-wise Kendall's Rank correlations between wavelet decompositions of the locations of microsatellites (motif length 2-5 bp) and recombination rate (averaged over 1 kb bins) for the 37 regions of the human genome with values for each variable for 2^{15} kb contiguous blocks. Scale (kb) is shown on the x axes and coefficient is shown on the y axes. Significant correlations (p<0.01) are flagged with a red cross. Approximate locations of each region are given, and where they are within 10 mb of a centromere or telomere they are labeled as near to that feature.



Region 1: Chr. 3 long arm, near centromere

Region 2: Chr. 3 mid long arm

Region 3: Chr. 3 long arm, near telomere

Region 4: Chr. 2 mid long arm

Region 5: Chr. 2 mid long arm

Region 6: Chr. 4 mid long arm

Region 7: Chr. 4 mid long arm

Region 8: Chr. 14 long arm near centromere

Region 9: Chr. 14 mid long arm

Region 10: Chr. 1 mid short arm

Region 11: Chr. 1 mid short arm

Region 12: Chr. 2 mid short arm

Region 13: Chr. 2 short arm, near centromere

Region 14: Chr.13 long arm near centromere

Region 15: Chr. 13 mid long arm

Region 16: Chr. 3 short arm, near telomere

Region 17: Chr. 3 mid short arm

Region 18: Chr. 6 mid long arm

Region 19: Chr. 12 mid long arm

Region 20: Chr. 8 mid long arm

Region 21: Chr. 1 mid long arm

Region 22: Chr. 15 long arm, near centromere

Region 23: Chr. 7 mid long arm

Region 24: Chr. 11 short arm, near telomere

Region 25: Chr. 5 short arm, near telomere

Region 26: Chr. 10 mid long arm

Region 27: Chr. 5 long arm, near centromere

Region 28: Chr. 5 lmid long arm

Region 29: Chr. 9 mid long arm

Region 30: Chr. 9 short arm near telomere

Region 31: Chr. 12 long arm, near centromere

Region 32: Chr. 11 long arm, near telomere

Region 33: Chr. 2 mid long arm

Region 34: Chr. 8 long arm near centromere

Region 35: Chr. 7 short arm near telomere

Region 36: Chr. 18 long arm, near centomere

Region 37: Chr. 6 long arm, near centromere

Figure 3.3: Wavelet correlations between mononucleotide repeats and recombination rate in the human genome
Pair-wise Kendall's Rank correlations between wavelet decompositions of the locations of mononucleotide repeats (motif length 2-5 bp) and recombination rate (averaged over 1 kb bins) for the 37 regions of the human genome with values for each variable for 2^{15} kb contiguous blocks. Scale (kb) is shown on the x axes and coefficient is shown on the y axes Significant correlations (p<0.01) are flagged with a red cross. Approximate locations of each region are given, and where they are within ten mb of a centromere or telomere they are labeled as near to that feature

3.4 Discussion

In view of the strikingly strong association between microsatellites and recombination hotspots in yeast, it was surprising to find that no general association between microsatellites and recombination is present in humans. This result does not rule out the possibility that recombination drives microsatellite evolution by mutation in both species, since humans and chimpanzees do not have a large proportion of hotspot locations in common [42, 43], so hotspots in the human lineage may not spend enough time one place to give rise to a substantial association with microsatellites. A mutation bias could have brought about the previously reported 5-10 mb scale correlation between microsatellite abundance and recombination rate [44] if recombination rate is constrained at a large scale. A similar hypothesis was proposed with respect to the broad scale association between recombination and GC-content, which is not substantially reproduced at the fine scale of hotspots [45]. The rationale for this was that some evidence suggests that recombination could increase local GC-content *via* biased gene conversion (see Section 1.3), so the broad scale association with GC-content could reflect a more constrained recombination landscape at larger scales [45]. Other evidence has suggested that very large scale recombination rates may be constrained [11], but a test of this hypothesis has not yet been reported.

If recombination drives microsatellite abundance and is constrained at scales of less than one mb, however, my scale-specific wavelet analysis should have revealed significant correlations, but I found none. This apparent discrepancy with the previous report of a correlation between microsatellite density and human recombination rate [1] is presumably due to methodological differences between the two studies. The most obvious of these is that I used wavelet analysis, which might be a less powerful technique. However, its least powerful variant, the non-parametric wavelet detail coefficient analysis, is powerful enough to detect correlations between microsatellite density and factors other than recombination consistently across most regions. Although power is presumably lower for the sporadically distributed variable recombination, the near zero correlation coefficients seen for most regions and scales when correlating microsatellites and recombination rate with this form of wavelet analysis, and the almost complete lack of significant results, are consistent with the finding that microsatellite density is not elevated in recombination hotspots. Taken together, these results indicate that the association between microsatellites and recombination at scales of one mb or less is very weak. This suggests that the previous report of a correlation between microsatellite density and recombination rate [1], which used window sizes of five mb and ten mb, detected very broad scale correlations not substantially reproduced at finer scales, pointing to a link between microsatellites and the previously noted very broad scale variation in recombination rate termed recombination “jungles” and “deserts” [5, 45-47]. This hypothesis requires further testing, because gaps in the data I used precluded the analysis of a substantial number of regions larger than 32 mb using the wavelet method. These gaps often resulted from failure of recombination rate coordinates to convert between genome builds, so the latest recombination map [13], which has coordinates based on the latest genome build, should enable testing of larger regions.

Although I did not detect a correlation between microsatellite frequency and recombination rate at scales of one mb or less in humans, nor a significant association between microsatellites and human recombination hotspots, I did find a modest elevation of microsatellite abundance in hotspot central and flanking regions. This enrichment is in the order of 10-20%, which is clearly not of sufficient magnitude to be considered suggestive of a widespread causal link between microsatellites and recombination in the human genome, but it does suggest that such a link could exist for a small proportion of hotspots. The hotspots predicted from haplotype inference methods are not well-characterized enough for their true mid points to be known, so it could be argued that the central tendency I observed for microsatellites might be due to stochastic variation. However, the hotspot dataset is large enough (n = 9298) that averages taken over all the hotspots should reflect a mutual cancelling out of variations of the true hotspot mid points either side of the mid points I defined, i.e. halfway between hotspot start and end coordinates. Crossover frequencies increase toward the mid points of hotspots in mammals [14, 15], so unequal recombination events causing microsatellites to mutate, and sequences responsible for regulating recombination at a local level, are likely to be concentrated in hotspot central regions. If the association is due to a mutation bias, microsatellite polymorphism should therefore be elevated in hotspot central regions, and I report a test of this possibility in Chapter 5. An alternative hypothesis is that a proportion of microsatellites are functionally involved in recombination at some hotspots, and this possibility is supported by some previous evidence (see Section 2.4). My finding of elevated microsatellite frequency in hotspot flanking regions is also consistent with a functional role for microsatellites, at least in some areas, in regulating the presently unexplained phenomenon of the control of hotspots by their sequence context [19].

In terms of a putative function for microsatellites in hotspots, an interesting difference between humans and yeast is the strong association between poly-A and recombination in yeast (see Section 2.3), contrasting with the marginally negative association I have seen in the human genome, which has also been reported previously for both poly-A [47] and poly-AT [5]. Why this is the case is not clear. It could perhaps be related to base composition and its links with recombination, or to differences between the two species in selective constraint of poly-A in hotspots, perhaps resulting from a functional role of poly-A in gene expression (see Section 8.1). Because GC-rich poly-purine/poly-pyrimidine tracts (PPTs) have some similarities with poly-A, notably an ability to modulate chromatin structure [48-50], it is possible that a functional role for poly-A in yeast recombination, which has been demonstrated [50], could, in the human lineage, have been displaced by GC-rich PPTs. I have investigated in detail the relationship between recombination and PPTs in humans and yeast, and the results are presented in Chapter 4.

References

1. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ: **Comparative recombination rates in the rat, mouse, and human genomes**. *Genome Res* 2004, **14**(4):528-538.
2. Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M: **High-resolution patterns of meiotic recombination across the human major histocompatibility complex**. *Am J Hum Genet* 2002, **71**(4):759-776.
3. Rana NA, Ebenezer ND, Webster AR, Linares AR, Whitehouse DB, Povey S, Hardcastle AJ: **Recombination hotspots and block structure of linkage disequilibrium in the human genome exemplified by detailed analysis of PGM1 on 1p31**. *Hum Mol Genet* 2004, **13**(24):3089-3102.
4. Shiroishi T, Koide T, Yoshino M, Sagai T, Moriwaki K: **Hotspots of homologous recombination in mouse meiosis**. *Adv Biophys* 1995, **31**:119-132.
5. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**(5746):321-324.
6. Buschiazzo E, Gemmell NJ: **The rise, fall and renaissance of microsatellites in eukaryotic genomes**. *Bioessays* 2006, **28**(10):1040-1050.
7. Ellegren H: **Microsatellite mutations in the germline: implications for evolutionary inference**. *Trends Genet* 2000, **16**(12):551-558.
8. Jakupciak JP, Wells RD: **Genetic instabilities of triplet repeat sequences by recombination**. *IUBMB Life* 2000, **50**(6):355-359.
9. Pearson CE, Nichol Edamura K, Cleary JD: **Repeat instability: mechanisms of dynamic mutations**. *Nature Reviews Genetics* 2005, **6**(10):729-742.

10. Kruglyak S, Durrett R, Schug MD, Aquadro CF: **Distribution and abundance of microsatellites in the yeast genome can Be explained by a balance between slippage events and point mutations**. *Mol Biol Evol* 2000, **17**(8):1210-1219.
11. Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G: **The influence of recombination on human genetic diversity**. *PLoS Genet* 2006, **2**(9):e148.
12. Petes TD: **Meiotic recombination hot spots and cold spots**. *Nat Rev Genet* 2001, **2**(5):360-369.
13. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M: **High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans**. *Science* 2008, **319**(5868):1395-1398.
14. Baudat F, de Massy B: **Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot**. *PLoS Genet* 2007, **3**(6):e100.
15. Kauppi L, Jeffreys AJ, Keeney S: **Where the crossovers are: recombination distributions in mammals**. *Nat Rev Genet* 2004, **5**(6):413-424.
16. Xu L, Kleckner N: **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* 1995, **14**(20):5115-5128.
17. de Massy B, Rocco V, Nicolas A: **The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in Saccharomyces cerevisiae**. *Embo J* 1995, **14**(18):4589-4598.
18. Xu F, Petes TD: **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* 1996, **143**(3):1115-1125.
19. Haring SJ, Halley GR, Jones AJ, Malone RE: **Properties of natural double-strand-break sites at a recombination hotspot in Saccharomyces cerevisiae**. *Genetics* 2003, **165**(1):101-114.
20. Keitt TH, Urban, D.L.: **Scale-specific inference using wavelets**. *Ecology* 2005, **86**(9):2497-2504.
21. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex**. *Nat Genet* 2001, **29**(2):217-222.
22. Kauppi L, Stumpf MP, Jeffreys AJ: **Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region**. *Genomics* 2005, **86**(1):13-24.
23. Jeffreys AJ, Ritchie A, Neumann R: **High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot**. *Hum Mol Genet* 2000, **9**(5):725-733.
24. Jeffreys AJ, Murray J, Neumann R: **High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot**. *Mol Cell* 1998, **2**(2):267-273.
25. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: **Human recombination hot spots hidden in regions of strong marker association**. *Nat Genet* 2005, **37**(6):601-606.
26. Holloway K, Lawson VE, Jeffreys AJ: **Allelic recombination and de novo deletions in sperm in the human beta-globin gene region**. *Hum Mol Genet* 2006, **15**(7):1099-1111.

27. Schneider JA, Peto TE, Boone RA, Boyce AJ, Clegg JB: **Direct measurement of the male recombination fraction in the human beta-globin hot spot**. *Hum Mol Genet* 2002, **11**(3):207-215.

28. May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys AJ: **Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX**. *Nat Genet* 2002, **31**(3):272-275.

29. Yi S, Li WH: **Molecular Evolution of Recombination Hotspots and Highly Recombining Pseudoautosomal Regions in Hominoids**. *Mol Biol Evol* 2005.

30. Verhoeven KJ, Simonsen KL: **Genomic Haplotype Blocks May Not Accurately Reflect Spatial Variation in Historic Recombination Intensity**. *Mol Biol Evol* 2005, **22**(3):735-740.

31. www.stats.ox.ac.uk/mathgen/Recombination.html.

32. http://genome.ucsc.edu/cgi-bin/hgLiftover.

33. http://repeatfinder.sourceforge.net/.

34. http://genome.ucsc.edu/.

35. http://naetet.net/millssnps.txt.gz.

36. Bagshaw AT, Pitt JP, Gemmell NJ: **Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots**. *BMC Genomics* 2006, **7**:179.

37. Benson G: **Tandem repeats finder a program to analyze DNA sequences**. *Nucleic Acids Res* 1999, **27**(2):573-580.

38. Quinn GP, Keough MJ: **Experimental Design and Data Analysis for Biologists**: Cambridge University Press; 2002.

39. Crawley MJ: **Statistics: An Introduction to Using R**: John Wiley and Sons; 2005.

40. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RMJ: **The American Soldier, Vol 1: Adjustment during army life.** Princeton: Princeton University Press; 1949.

41. Carducci B, Hedges JR, Beal JC, Levy RC, Martin M: **Emergency phenytoin loading by constant intravenous infusion**. *Ann Emerg Med* 1984, **13**(11):1027-1031.

42. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**(4):429-434.

43. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P *et al*: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**(5718):107-111.

44. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ, Hellmann I *et al*: **Comparative recombination rates in the rat, mouse, and human genomes***Genome Res* 2004, **14**(4):528-538.

45. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome**. *Science* 2004, **304**(5670):581-584.

46. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW *et al*: **Comparison of human genetic and sequence-based physical maps**. *Nature* 2001, **409**(6822):951-953.

47. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G *et al*: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**(3):241-247.

48. Elgin SC: **The formation and function of DNase I hypersensitive sites in the process of gene activation**. *J Biol Chem* 1988, **263**(36):19259-19262.

49. Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin**. *Annu Rev Biochem* 1988, **57**:159-197.

50. Schultes NP, Szostak JW: **A poly(dA.dT) tract is a component of the recombination initiation site at the ARG4 locus in Saccharomyces cerevisiae**. *Mol Cell Biol* 1991, **11**(1):322-328.

Chapter 4

Poly-purine/poly-pyrimidine tracts are associated with meiotic recombination hotspots in humans and yeast

Abstract

This chapter details an investigation into the scale and magnitude of the association between poly-purine/poly-pyrimidine tracts (PPTs) and recombination in humans and the yeast *S. cerevisiae*. I found that PPTs are highly over-represented in hotspots of the meiotic recombination initiating lesions double-strand breaks (DSBs), in the yeast genome. They are also significantly enriched in human meiotic crossover hotspots, though the level of enrichment is somewhat lower in humans than in yeast. A notable feature of the association between PPTs and hotspots common to both species is that it becomes more marked with increasing tract length, and this trend is stronger for high GC-content PPTs. These observations suggest a possible link with non-B-DNA structures, and this possibility is discussed.

Using generalized linear models, I found that the fine scale correlation between recombination intensity and PPT frequency is still significant when other factors that could mediate the correlation, including gene density, single nucleotide polymorphism density, and GC-content, are considered. The correlation is quite weak at a fine scale, and wavelet analysis showed that it is stronger at scales broader than hotspots, indicating that there are regional factors influencing the association between PPTs and recombination. However, I also found that PPTs are highly enriched in 500 base pair regions spanning hotspot mid points. Recombination activity is known to be most frequent in these areas, relative to the remaining parts of hotspots, so this observation indicates that regional factors are not solely responsible for the association between PPTs and recombination hotspots and suggests the existence of a localized causal link between PPTs and recombination. I also found that all three single

nucleotide polymorphisms previously shown to be associated with human hotspot activity changes occur within sequence contexts of 14 bp or longer that are 85% or more poly-purine/poly-pyrimidine and at least 70% G/C. This again suggested the possibility of non-B-DNA structures, and sensitivity to single nucleotide changes has previously been shown for these structures.

4.1 Introduction

Poly-purine/poly-pyrimidine tracts (PPTs) are an interesting class of sequence because of their unusual structural propensities, which may be linked to widespread functional importance, the exploration of which is currently in its infancy (see Sections 1.1.2 and 7.1). Studies have shown that PPT density correlates with recombination rate in humans at scales of several hundred thousand to several million base pairs (bp) [1, 2], suggesting that one function of PPTs could be in regulating recombination. Consistent with this possibility, a study showed that a PPT with secondary structure stimulated recombination between two closed circular DNA molecules (plasmids) [3]. This need not reflect a role for PPTs in inter-chromosomal recombination, the regulation of which is complex (see Section 1.2.2). If PPTs were closely associated with meiotic recombination sites in general, however, the hypothesis of a widespread causal link between PPTs and recombination would be supported. Functionality in one recombination hotspot has been demonstrated in yeast for poly-A [4], as noted in Sections 1.2.2 and 2.4, but PPTs with some GC-content are distinct from poly-A in their ability to form intramolecular secondary structures [5]. Poly-A is stiff and cannot form intramolecular structures [5, 6], though it can form inter-molecular three-stranded aggregates [7].

This chapter is based on published work [8] (see page xi, Publications associated with this thesis). It also includes an extension of this work, investigating the scale and magnitude of the association between PPTs and recombination, and the possible influence on the association of other factors expected to correlate with both. These questions have not been addressed elsewhere in the literature. To examine the scale of the association between recombination and PPT density, I used direct comparison of hotspots and their central and flanking regions with cold areas, combined with generalized linear models to investigate the influence of other factors, and also wavelet analysis to assess large scale effects. The analysis techniques used in the work presented in this chapter, and the rationale behind their application, were as described in Chapter 3 (see Sections 3.1 and 3.2.2).

4.2 Methods

Meiotic recombination hotspots investigated were as in Chapters 2 and 3, namely the *S. cerevisiae* hotspots reported by Gerton and colleagues [9], the 17 well characterized human hotspots known to date [10-15] , and the fine scale recombination map from the genome-wide study by Myers *et al*., (2005) [16]. Additionally, I investigated the relationship between PPT locations and recombination at the finest possible scale using 76 double-strand break (DSB) sites mapped with high resolution on *S. cerevisiae* chromosome 3 [17]. As expected, PPTs, like microsatellites, are much less frequent in genes than in intergenic regions of the yeast genome (data not shown), so I limited the investigation of yeast hotspots to a comparison between hot intergenic regions (IGRs), in which DSBs are known to be concentrated [17], and non-hot IGRs.

To detect PPTs in DNA sequence, I collaborated with a computer programmer to design a pattern-matching algorithm in the C language, and he wrote the programme [18]. Details of sequence and sequence annotation data retrieval were identical to Section 2.2.1 for yeast sequences, and section 3.2.1 for human sequences. The wavelet analysis methods were as described in Section 3.2.2. I performed other statistical comparisons as described in Chapters 2 and 3. I excluded PPTs overlapping two regions from all analyses, with the exception of those that used the variable PPT coverage rather than tract frequency.

4.3 Results

4.3.1 High frequency of PPTs in yeast DSB hotspots

I initially used a 12 bp minimum length for PPT searches. This was based on the fact that a 12 bp PPT has been shown to form a stable intramolecular quadruplex [19], and in my search of the literature I did not find reports of shorter sequences forming intramolecular structures. I found that the frequency of PPTs of at least 12 bp is 65 % elevated yeast hot intergenic regions (IGRs) compared with non-hot IGRs ($p<10^{-23}$, Mann-Whitney U test). The 40 DSB coldspots identified by Gerton *et al.*, [9] have generally lower frequencies than other non-hot regions (Figure 4.1), but the differences are not statistically significant.

Changing the minimum length limit for PPT searches alters the level of association between PPT frequency and hotspots, with a clear trend towards an increased enrichment in hotspots for longer tracts (Table 4.1, Figure 4.1). The enrichment is significant when applying minimum length limits as low as seven bp, but not lower. Raising the minimum size limit above 12 bp markedly increases the hot/non-hot mean PPT frequency ratio (Table 4.1, Figure 4.1), which reaches six for PPTs of at least 35 bp, the highest minimum for which there is a statistically significant difference between hot and non-hot IGRs. This does not indicate that only very long PPTs are associated with hotspots, however, since tracts of between 12 and 19 bp are significantly more common in hotspots, by 56% ($p<10^{-15}$, Mann-Whitney U Test).

PPTs with some mismatches are much more common than pure PPTs, and mismatched tracts can form secondary structures [5, 18, 20, 21], so I repeated the searches allowing some mismatches. This reduces the level of enrichment of PPTs in hotspots for any given length limit but increases the maximum length limit for which a statistically significant difference between hot and non-hot IGRs is detectable (Table 4.1, Figures 4.1 and 4.2). Another relevant factor to consider is PPT GC-content, because high GC-content PPTs show a greater readiness to form the secondary structures intramolecular triplexes [5, 22-24], and intramolecular quadruplexes require rows of guanine residues [25, 26]. To address the question of PPT GC-content in relation to hotspots, I divided PPTs into those with less than the mean GC-content for all PPTs (31.4%) and those with more. I repeated all analyses for high GC content tracts, and I found that their level of enrichment is not substantially different from that of all tracts considered together, except for very long tracts, for which it is somewhat higher (Table 4.1, Figure 4.2). PPTs in yeast hotspots do not differ significantly in GC content from those in non-hot IGRs, however ($p>0.05$, Mann-Whitney U test), and, as noted previously (Section 2.3.1), these hotspots contain highly elevated frequencies of poly-A.

Table 4.1: PPT frequencies in yeast intergenic regions
Mean per kb frequencies of PPTs in 473 hot and 5520 non-hot IGRs in the *S. cerevisiae* genome. Standard errors are also shown (SEM) and p values are for the Mann-Whitney U Test. Data for minimum length limits of 12, 20, and, for each type of PPT, one higher limiter, are shown. This third limiter is the highest, for each GC-content/mismatch type of PPT for which a significant difference is detectable between hot and non-hot regions and for which all preceding (lower) limiters also give a significant difference. The e value indicates the number of bases, in any part of a PPT, within which no more than one mismatch was allowed. A lower e value therefore means more mismatches were allowed.

Type of PPT			**Hot IGRs**		**Non-hot IGRs**		**Freq. ratio (hot/ non-hot)**	**P value**
Lower length limit	Mismatch type	GC-content type	Mean per kb freq.	SEM	Mean per kb freq.	SEM		
12	All tracts	perfect	3.93	0.17	2.38	0.0403	1.65	$<10^{-23}$
		e=10	9.01	0.235	7.02	0.0659	1.28	$<10^{-17}$
		e=5	10.8	0.245	9.18	0.0712	1.17	$<10^{-11}$
	High GC	perfect	1.32	0.103	0.848	0.0246	1.55	$<10^{-7}$
		e=10	4.08	0.169	3.08	0.0445	1.32	$<10^{-10}$
		e=5	5.25	0.187	4.25	0.0511	1.23	$<10^{-7}$
20	All tracts	perfect	0.584	0.0593	0.238	0.0125	2.45	$<10^{-19}$
		e=10	1.88	0.1129	1.07	0.0263	1.76	$<10^{-16}$
		e=5	2.6	0.136	1.64	0.0325	1.59	$<10^{-14}$
	High GC	perfect	0.107	0.0228	0.0538	0.0054	2	$<10^{-4}$
		e=10	0.682	0.0704	0.366	0.0149	1.86	$<10^{-9}$
		e=5	1.05	0.0892	0.657	0.0207	1.61	$<10^{-8}$
35	All tracts	perfect	0.0754	0.0318	0.0119	0.0024	6.34	0.0005
35	High GC	perfect	0.0079	0.0042	0.0007	0.0004	10.82	$<10^{-5}$
50	All tracts	e=10	0.032	0.0166	0.0044	0.0015	7.31	0.0001
40	High GC	e=10	0.0148	0.0064	0.0034	0.0011	4.32	0.0001
55	All tracts	e=5	0.0262	0.0126	0.0048	0.0013	5.5	$<10^{-5}$
60	High GC	e=5	0.0069	0.0041	0.0007	0.0004	10.1	0.00013

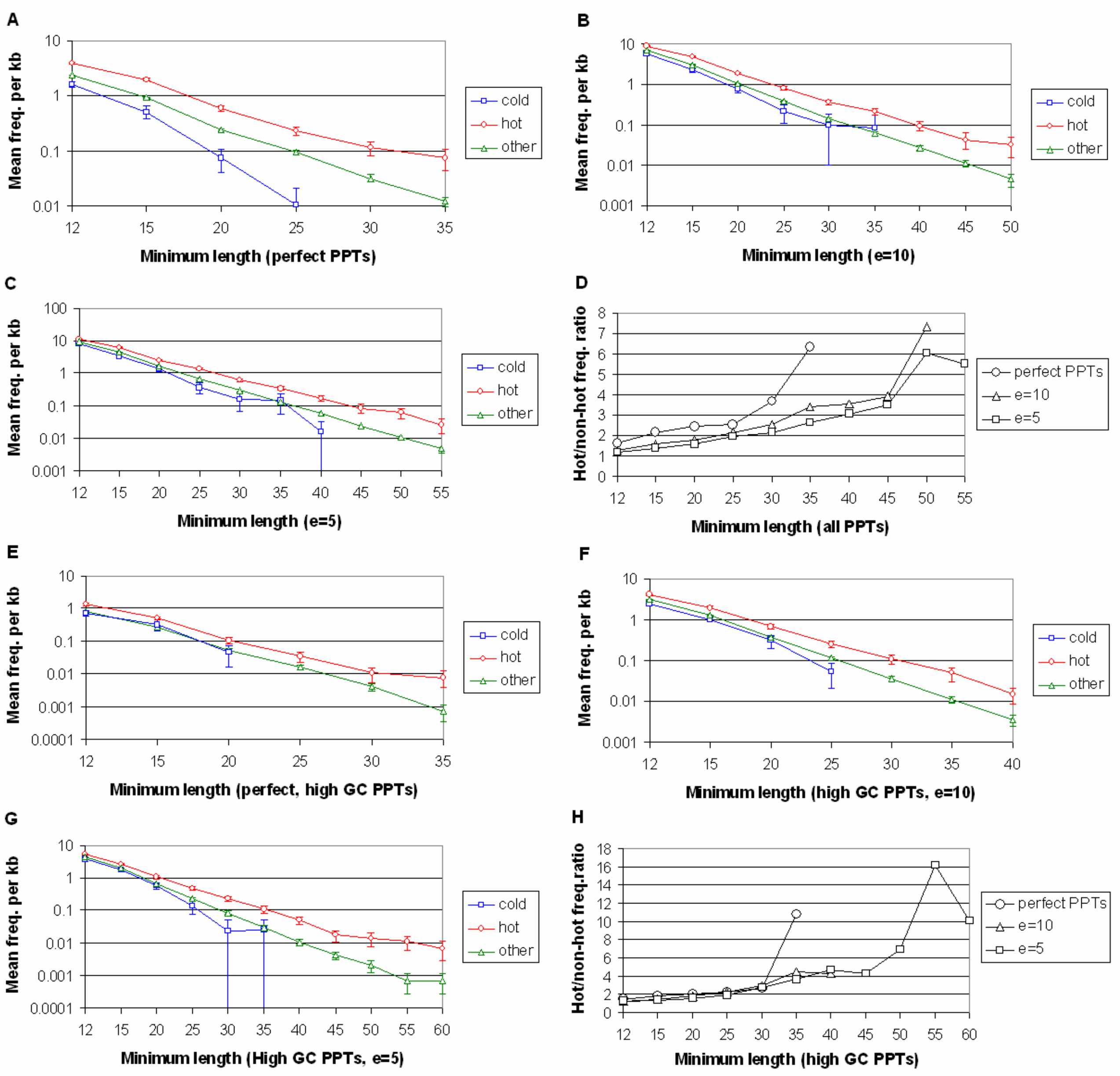


Figure 4.1 - Enrichment of PPTs in yeast hot intergenic regions.
Mean frequencies of PPTs in 473 hot, 89 cold and 5431 other regions of the *S. cerevisiae* genome, which were all IGRs not categorized either as hot or cold. Plots A, B, C and D are for PPTs with no GC-content restriction and plots E, F, G and H are for high GC-content PPTs (those with more than the genome mean PPT GC-content of 31.4%). Separate plots are shown for perfect PPTs (A and E), PPTs with one mismatch allowed per 10 bp (B and F) and PPTs with one mismatch allowed per 5 pb (C and G). In each of these plots, mean frequency is plotted on a logarithmic scale. The ratio of hot and non-hot mean PPT frequencies is also shown, for each mismatch class of PPT, in D and H. The highest limiter shown for each PPT type in all plots is the highest for which a significant difference is detectable between hot and non-hot regions for which all preceding (lower) limiters also give a significant difference (p<0.01, Mann-Whitney U Test). Error bars are plus and minus one SEM, or are not shown in cases where they are smaller than the chart symbol size, or for parts D and H, which simply plot the ratio of mean frequencies between hot and non-hot IGRs. The e value indicates the number of bases, in any part of a PPT within which no more than one mismatch was allowed. A lower e value therefore means more mismatches were allowed.

The high frequency of PPTs in yeast hotspots does not result from a simple presence/absence relationship, since most non-hot IGRs contain at least one PPT of at least 12 bp, and the ratio of hot IGRs containing at least one PPT to the number of non-hot IGRs containing at least one PPT is considerably less, for each particular lower length limit, than the hot/non-hot mean PPT frequency ratio (Table 4.2). Mean region length does not differ significantly between hot and non-hot IGRs (p>0.05, Mann-Whitney U Test), so the observed PPT frequency enrichment is due to multiple PPTs occurring in hot IGRs

Table 4.2: Percentage of intergenic regions with at least one PPT
Showing the percentage of IGRs with at least one PPT among 473 hot and 5520 non-hot IGRs. High GC-content PPTs were those with greater GC-content than the genome mean for PPTs (31.4%). The e value indicates the number of bases within which no more than one mismatch to the PPT motif was allowed. A lower e value therefore indicates that more mismatches were allowed.

Type of PPT			**% of IGRs with at least one PPT**	
Lower length limit	GC-content Type	Mismatch type	Hot	Non-hot
12	All tracts	perfect	76.32	58.08
		e=10	92.6	87.04
		e=5	94.66	91.64
	High GC	perfect	40.8	29.78
		e=10	75.9	66.7
		e=5	81.82	75.27
20	All tracts	perfect	24.1	10.31
		e=10	52.85	35.18
		e=5	62.79	47.53
	High GC	perfect	6.13	2.59
		e=10	26.22	15.47
		e=5	37.21	25.02
35	All tracts	perfect	2.11	0.65
35	High GC	perfect	0.85	0.07
50	All tracts	e=10	1.27	0.24
40	High GC	e=10	1.27	0.24
55	All tracts	e=5	1.69	0.31
60	High GC	e=5	0.63	0.05

4.3.2 Association of PPTs with individual non-hotspot DSB sites in the yeast genome

Baudat and Nicolas (1997) mapped meiotic DSBs throughout chromosome 3 of the genome of the yeast *S. cerevisiae* and identified 70 IGRs subject to at least one DSB [17]. Overall, these DSB-containing IGRs averaged 567 bp in length. I found that a 15 bp lower PPT length limit gives the strongest association between PPT frequency and DSB sites based on p value, and PPTs of at least 15 bp are significantly enriched in these areas (p=0.000791; mean per kb frequencies 1.83 in DSB-containing IGRs and 0.925 in IGRs without a DSB). Most of the 70 DSB-containing IGRs have very low levels of DSBs (see Figure 2 in ref [17]), and 48 of them occur outside hotspots reported in the genome-wide survey by Gerton and co-workers [9]. It is therefore likely that many of them reflect non-hotspot background recombination events, since these have been found to occur with low frequency between hotspots [10-12]. For PPTs of at least 15 bp, the mean frequency per kb is 1.70 in DSB-containing IGRs outside hotspots reported by Gerton *et al*. [9]. This is significantly greater than the mean per kb frequency of 0.925 for IGRs without a DSB (p=0.00262, Mann-Whitney U Test).

4.3.3 The association of PPTs with yeast hotspots is also extended to hotspot flanking regions

I compared frequencies of PPTs of at least 12 bp between flanking IGRs one, two, three and four ORFs removed from hotspots with remaining non-hot IGRs. The level of enrichment of PPTs in hotspot flanking IGRs is substantially reduced compared with hotspots themselves, but is still significant in comparison with remaining non-hot regions. IGRs one ORF removed from hotspots have 28% higher PPT frequency than remaining non-hot IGRs (p=0.003, Mann-Whitney U test) and IGRs two ORFs removed have a 21% enrichment of PPTs (p=0.02, Mann-Whitney U test). The mean distance encompassed by the hot spot-containing regions in which PPTs are enriched is just over 11.5 kb.

4.3.4 No effect of transcription or promoter regions on the association between PPTs and DSB hotspots

PPTs have been implicated in the regulation of gene expression [27, 28], and I found that transcriptional frequency in vegetative cells [29] correlates with DSB intensity (p<0.0001, Spearman’s rank test). However, looking at the “hottest” IGRs and ORFs for transcriptional frequency in equivalent numbers to the numbers of recombination hot regions studied, I found that the number of these that overlap with recombination hotspots is lower than random expectation (see Section 2.3.2), and the correlations between DSB intensity and frequency of PPTs hardly change at all when controlling for transcriptional frequency in partial correlation analysis (Table 4.3). Furthermore, frequencies of PPTs, classed as in Tables 4.1 and 4.2, are not significantly different between the 473 hottest IGRs for transcriptional frequency and remaining IGRs.

DSBs have been shown to be more frequent in IGRs with two promoters (divergent transcription of flanking genes) than those with one (parallel transcription of flanking genes) or none (convergent transcription of flanking genes) [9]. Therefore, if PPTs are associated with promoters for reasons not connected with recombination, this could coincidentally increase their association with DSB hotspots in yeast, since IGRs in the *S. cerevisiae* genome average only about 500 bp. I did find significant differences in PPT frequency between IGRs with different numbers of promoters, but as is the case for microsatellites (see Section 2.3.3), the differences are relatively small and inconsistent. For example, PPTs of at least 12 bp are most common in IGRs with no promoters, but only by 2-3% ($p<10^{-6}$, Kruskal-Wallis ANOVA). PPTs of at least 20 bp are most common in IGRs with one promoter, an excess of around 14% ($p<10^{-7}$, Kruskal-Wallis ANOVA). High GC-content PPTs are most common in IGRs with two promoters, but the difference is again relatively slight, e.g. 11 % for high GC-content tracts of at least 12 bp ($p<10^{-19}$, Kruskal-Wallis ANOVA). Moreover, there is no substantial or consistent difference in the association between PPTs and DSB hotspots between IGRs with two promoters and IGRs without a promoter (Table 4.3). A larger number of PPT classes are significantly associated with hotspots when considering only IGRs with one promoter, but this is probably because this type of region covers slightly more of the genome than the other two types of regions combined, allowing greater statistical power, because it does not contain higher hot/non-hot PPT frequency ratios. An exception to this is some classes of very long PPTs, but these elements are very rare in the genome as a whole (see Table 4.2).

Table 4.3: The effect of promoter regions on the association between PPTs and yeast DSB hotspots
Showing results from an analysis of PPT frequency in hot vs non-hot IGRs performed as for the results presented in Table 4.1, with the exception that IGRs were divided according to the number of promoters they contain, i.e. hot IGRs with no promoters were compared with non-hot IGRs with no promoters, etc. The e value indicates the number of bases, in any part of a PPT, within which no more than one mismatch was allowed. A lower e value therefore means more mismatches were allowed. Also shown is the total amount of the genome, in kb, covered by IGRs of each respective type.

PPT type			Number of Promoters (total genomic coverage of IGR type)					
			Zero (57 kb)		One (160 kb)		Two (100 kb)	
Lower length limit	GC-content type	Mismatch type	Mean freq. ratio (hot/ non hot)	P value	Mean freq. ratio (hot/ non hot)	P value	Mean freq. ratio (hot/ non hot)	P value
12	All tracts	perfect	1.76	$<10^{-5}$	1.63	$<10^{-11}$	1.59	$<10^{-6}$
		e=10	1.30	0.0008	1.26	$<10^{-9}$	1.27	$<10^{-5}$
		e=5	1.18	n/s	1.17	$<10^{-6}$	1.17	$<10^{-4}$
	High GC	perfect	1.68	n/s	1.58	0.0002	1.35	n/s
		e=10	1.21	n/s	1.29	$<10^{-5}$	1.36	0.0001
		e=5	1.15	n/s	1.22	$<10^{-4}$	1.25	$<10^{-4}$
20	All tracts	perfect	2.10	0.0003	2.36	$<10^{-9}$	2.87	$<10^{-6}$
		e=10	1.64	$<10^{-4}$	1.70	$<10^{-6}$	1.90	$<10^{-7}$
		e=5	1.48	0.0037	1.58	$<10^{-6}$	1.67	0.0044
	High GC	perfect	2.06	n/s	2.00	0.0014	1.76	n/s
		e=10	2.26	0.0004	1.71	0.0003	1.79	0.0015
		e=5	1.41	n/s	1.68	$<10^{-5}$	1.53	$<10^{-6}$
35	All tracts	perfect	0.77	n/s	11.4	0.0007	4.59	n/s
35	High GC	perfect	n/a	n/s	8.18	n/s	14.1	0.0005
50	All tracts	e=10	0	n/s	10.3	$<10^{-5}$	1.57	n/s
40	High GC	e=10	6.30	n/s	3.74	n/s	4.23	n/s
55	All tracts	e=5	0	n/s	11.3	$<10^{-6}$	1.88	n/s
60	High GC	e=5	n/a	n/s	8.10	0.0025	11.8	n/s

4.3.5 The correlation between DSB intensity and PPT frequency in yeast

Gerton *et al.*, mapped DSB concentration for the whole genome, not just for hotspots, so I analysed the genome-wide correlation between PPT frequency and DSB intensity (Table 4.4). Correlations are significant and positive but quite weak, with coefficients no greater than 0.16 (Spearman's rho). When controlling for GC-content the correlations are not affected in the case of all PPTs considered together, but are somewhat reduced though still significant when only high GC-content PPTs are considered. The weakness of the correlations suggested that the enrichment of PPTs in hotspots might be stronger for less active hotspots. I tested this possibility by dividing hot IGRs into two equal-sized classes, warm and very hot, based on their DSB intensity. None of the PPT classes considered in Tables 4.1 and 4.2 differ significantly in frequency between warm and very hot IGRs.

Table 4.4: The correlation between PPT frequency and DSB intensity in yeast
Showing statistics for the genome-wide correlation between DSB intensity and PPT frequency for the types of PPT considered in Tables 4.1 and 4.2. I controlled for GC-content and transcriptional frequency using non-parametric partial correlation analysis. The e value indicates the number of bases, in any part of a PPT, within which no more than one mismatch was allowed. A lower e value therefore means more mismatches were allowed.

Type of PPT			**Correlation statistics (Spearman's rho)**					
Lower length limit	GC-content type	Mismatch type	DSB intensity vs PPT frequency		Controlling for regional GC-content		Controlling for transcriptional frequency	
			Coeff.	P value	Coeff.	P value	Coeff.	P value
12	All tracts	perfect	0.155	<.0001	0.162	<.0001	0.151	<.0001
		e=10	0.123	<.0001	0.137	<.0001	0.114	<.0001
		e=5	0.117	<.0001	0.127	<.0001	0.107	<.0001
	High GC	perfect	0.0911	<.0001	0.0662	<.0001	0.0916	<.0001
		e=10	0.114	<.0001	0.0778	<.0001	0.110	<.0001
		e=5	0.118	<.0001	0.0748	<.0001	0.111	<.0001
20	All tracts	perfect	0.115	<.0001	0.111	<.0001	0.114	<.0001
		e=10	0.134	<.0001	0.135	<.0001	0.130	<.0001
		e=5	0.137	<.0001	0.137	<.0001	0.132	<.0001
	High GC	perfect	0.0539	<.0001	0.0399	0.0021	0.0569	<.0001
		e=10	0.0657	<.0001	0.0467	0.0003	0.0665	<.0001
		e=5	0.107	<.0001	0.0776	<.0001	0.106	<.0001
35	All tracts	perfect	0.029	0.0254	0.0276	0.0329	0.0308	0.0198
35	High GC	perfect	0.0506	<.0001	0.0445	0.0006	0.0474	0.0003
50	All tracts	e=10	0.0473	0.0003	0.0447	0.0006	0.046	0.0005
40	High GC	e=10	0.057	<.0001	0.0498	0.0001	0.0541	<.0001
55	All tracts	e=5	0.0531	<.0001	0.0491	0.0002	0.0539	<.0001
60	High GC	e=5	0.039	0.0026	0.0345	0.0077	0.0388	0.0034

4.3.6 PPT frequency is elevated in experimentally characterized human hotspots

The association between PPT frequency and meiotic recombination hotspots is somewhat weaker in humans than in yeast, but it is qualitatively similar between the two species in some respects. The 17 experimentally well-characterized human meiotic recombination hotspots (see Section 3.2.1) do not show significant enrichment for PPTs of at least 12 bp, or for lower size minima, but, as in yeast, the hot/cold frequency ratio increases as the minimum tract length is raised, and significant differences are present for length minima 13-15 (Table 4.5). When PPTs of over 20 bp are excluded from this analysis, the difference is no longer significant (p=0.052, Mann-Whitney U Test), but PPTs of 13-19 bp are significantly enriched in these hotspots, by 78%, when only high GC-content tracts are considered (p=0.008, Mann-Whitney U test). Only one of the 17 experimentally well-characterized human hotspots does not have a perfect PPT of more than 12 bp. This hotspot is located in an intergenic region of the Class II MHC complex with the closest gene being *DPAI*. The hotspot does contain a 12 bp high GC-content PPT with one mismatch within its central region, defined as the 500 bp centred on its mid point.

Contrasting the situation in yeast, human hotspot-associated PPTs have elevated GC-content (45.4%) compared with those in cold regions (37.5%; p=0.001, Mann-Whitney U Test). Enrichment in human hotspots is higher for PPTs with greater than the overall mean PPT GC-content for the studied regions (38%) than for low GC-content PPTs (Table 4.5). This difference is apparently not consistently linked to variation in overall GC-content, since the tested hot and cold regions have, on average, almost exactly the same GC content (ratio=1.00).

An interesting question is whether PPTs are enriched in the central regions of hotspots, since crossover rates increase sharply for markers close to hotspot mid points [30, 31], and recombination-initiating double-stranded breaks (DSBs) are concentrated in regions of 100-500 bp in yeast [32-34]. Enrichment of PPTs in hotspot central regions could not be investigated in yeast, since some reports have shown an increase in recombination at the 5' ends of genes, but this is not a general phenomenon (reviewed in [35]), and it was not investigated by the genome-wide yeast DSB mapping study [9]. In humans, I defined a hotspot central region as 500 bp centred on the hotspot mid point. I found that PPT frequencies are somewhat higher in these regions than in the remaining parts of hotspots, and while the differences are not significant for length minima 12-15, which other data show is due to the small sample size of 17 hotspots combined with the small region size (see section 4.3.7), the difference for high GC tracts of at least 20 bp is significant (Table 4.5).

Table 4.5: PPT frequencies in 17 experimentally well-characterized human hotspots compared with their adjacent cold regions.
Mean per kb frequencies of PPTs in 17 experimentally well-characterized hotspots, the 500 bp regions centred on their mid points, and their adjacent cold regions. Standard errors are also shown (SEM) and p values are for the Mann-Whitney U Test (alpha=0.05, since the classes of PPTs investigated are not fully independent). Data for minimum length limits of 12-15 and 20 are shown.

Type of PPT		Hotspots		Hotspot central regions		Non-hot regions		Mean freq. ratio	P value	
GC-content type	Lower length limit	Mean per kb freq.	SEM	Mean per kb freq.	SEM	Mean per kb freq.	SEM	Hot/ non-hot	Hot v non-hot	Hot central regions vs non-hot
All tracts	12	2.546	0.455	3.059	0.86	2.233	0.346	1.14	n/s	n/s
	13	2.024	0.441	2.588	0.762	1.333	0.194	1.518	0.044	n/s
	14	1.496	0.273	1.882	0.606	0.986	0.16	1.517	0.035	n/s
	15	1.21	0.233	1.294	0.513	0.734	0.098	1.649	0.036	n/s
	20	0.308	0.157	0.353	0.353	0.271	0.04	1.138	n/s	$<10^{-4}$
High GC	12	1.831	0.333	2.118	0.757	1.494	0.381	1.226	n/s	n/s
	13	1.481	0.324	1.882	0.717	0.795	0.182	1.863	0.008	n/s
	14	1.039	0.208	1.294	0.541	0.533	0.142	1.949	0.017	n/s
	15	0.867	0.215	0.941	0.518	0.361	0.072	2.4	n/s	n/s
	20	0.227	0.154	0.353	0.353	0.092	0.023	2.47	n/s	0.002

Two of the experimentally characterized hotspots from human chromosome 1 show little or no evidence for historical recombination events, indicating that they have recently appeared in the genome [12]. These hotspots have higher PPT densities than average for human hotspots for some tract length minima, but because they are both within 2 kb of other hotspots, comparisons are probably not meaningful in view of the weak distal associations between hotspots and PPT density which are apparent in yeast (see Section 4.3.3), a pattern that is also evident in relation to human hotpots from a genome-wide dataset derived from haplotype inference methods (see Section 4.3.8). One of the hotspots in the MHC Class II region predicted from statistical analysis of haplotype data was not found to be present in sperm, indicating that it could recently have become extinct [11]. This region was reported as spanning exon II of the HLA-DPB1 gene, so I investigated PPT density in a 2 kb region centred around that exon. I found its PPT frequency to be about average for the human experimentally characterized hotspots, with five PPTs longer than 12 bp. Four of these have at least 38% GC-content, which is above the average number present in the other human hotspots.

4.3.7 Elevated PPT frequency in a genome-wide set of human hotspots derived from haplotype inference methods

The genome-wide dataset of hotspot and coldspot locations mapped to within 5 kb or less reported by Myers *et al.,* [16] could provide a more complete picture of the association between hotspots and PPTs in humans. Investigating the frequency of PPTs of the types considered in Tables 4.1 and 4.2, I found highly significant associations of PPTs with hotspots, which are quite modest in magnitude compared with what I found for yeast and human experimentally characterized hotspots (Table 4.6). Perfect PPTs of at least 12 bp are only 6% more common in hotspots ($p<10^{-22}$, Mann-Whitney U Test) and this increases to 15% for tracts with greater than the overall mean PPT GC-content ($p<10^{-22}$, Mann-Whitney U Test). The pattern of association with respect to PPT length is, however, similar to that seen for yeast and human experimentally well-characterized hotspots in that the level of enrichment increases with increasing tract length, though it declines somewhat for tract lengths of over 50 bp (Figure 4.2). Looking at all possible length minima between 12 and 100 bp, with three different mismatch allowance parameters (perfect, a maximum of one per 5 bp and a maximum of one per 10 bp), for all PPTs, and for high GC-content PPTs considered separately, I found the highest level of enrichment for high GC-content PPTs of at least 50 bp with one mismatch allowed per 10 bp (Figure 4.2). However, only 7.3% of hotspots contain a tract of these specifications. Similarly to the other hotspot datasets I investigated, the genome-wide human hotspots from Myers *et al.,* [16] are not only associated with very long tracts. PPTs of between 12 and 19 bp are enriched in these hotspots, by 5.3 % when no GC-content restriction is applied ($p<10^{-14}$, Mann-Whitney U test), and by 13.2% for high GC-content tracts ($p<10^{-39}$, Mann-Whitney U test).

Table 4.6: PPT frequencies in human hot- and coldspots from a genome-wide dataset Showing mean per kb frequencies of PPTs in 9298 hotspots and 9292 coldspots mapped to within 5 kb (mean length 4070 bp) using haplotype inference methods. Standard errors are also shown (SEM) and p values are for the Mann-Whitney U Test. Data for minimum length limits of 12, 20, and, for each GC-content/mismatch type of PPT, one higher limiter, are shown. This third limiter is the highest for each PPT mismatch/GC-content type for which a significant difference is detectable between hot and non-hot regions for which all preceding (lower) limiters also give a significant difference ($p<0.01$). The e value indicates the number of bases, in any part of a PPT, within which no more than one mismatch was allowed. A lower e value therefore means more mismatches were allowed.

Type of PPT			**Hot**		**Cold**		**Freq. ratio (hot/ non-hot)**	**P value**
Lower length limit	GC-content type	Mismatch type	Mean per kb freq.	SEM	Mean per kb freq.	SEM		
12	All tracts	perfect	1.942	0.0083	1.829	0.0077	1.06	$<10^{-22}$
		e=10	7.066	0.0149	6.853	0.0149	1.03	$<10^{-24}$
		e=5	9.648	0.0166	9.388	0.0166	1.03	$<10^{-28}$
	High GC	perfect	1.082	0.007	0.942	0.0062	1.15	$<10^{-48}$
		e=10	4.271	0.0159	3.809	0.0154	1.12	$<10^{-100}$
		e=5	5.783	0.0196	5.19	0.0194	1.11	$<10^{-104}$
20	All tracts	perfect	0.26	0.0032	0.23	0.0028	1.13	$<10^{-7}$
		e=10	0.856	0.0053	0.787	0.0049	1.09	$<10^{-18}$
		e=5	1.429	0.0067	1.332	0.0064	1.07	$<10^{-24}$
	High GC	perfect	0.11	0.0024	0.084	0.0018	1.31	$<10^{-13}$
		e=10	0.499	0.0043	0.415	0.0038	1.2	$<10^{-47}$
		e=5	0.839	0.0057	0.706	0.0052	1.19	$<10^{-67}$
30	all tracts	perfect	0.069	0.0018	0.061	0.0015	1.13	0.0062
40	high GC	perfect	0.017	0.0009	0.013	0.0007	1.32	0.0033
60	all tracts	e=10	0.02	0.0009	0.017	0.0007	1.21	0.0086
65	high GC	e=10	0.012	0.0007	0.009	0.0005	1.32	0.0017
65	all tracts	e=5	0.021	0.0009	0.017	0.0007	1.23	0.0052
80	high GC	e=5	0.011	0.0006	0.008	0.0005	1.34	0.0048

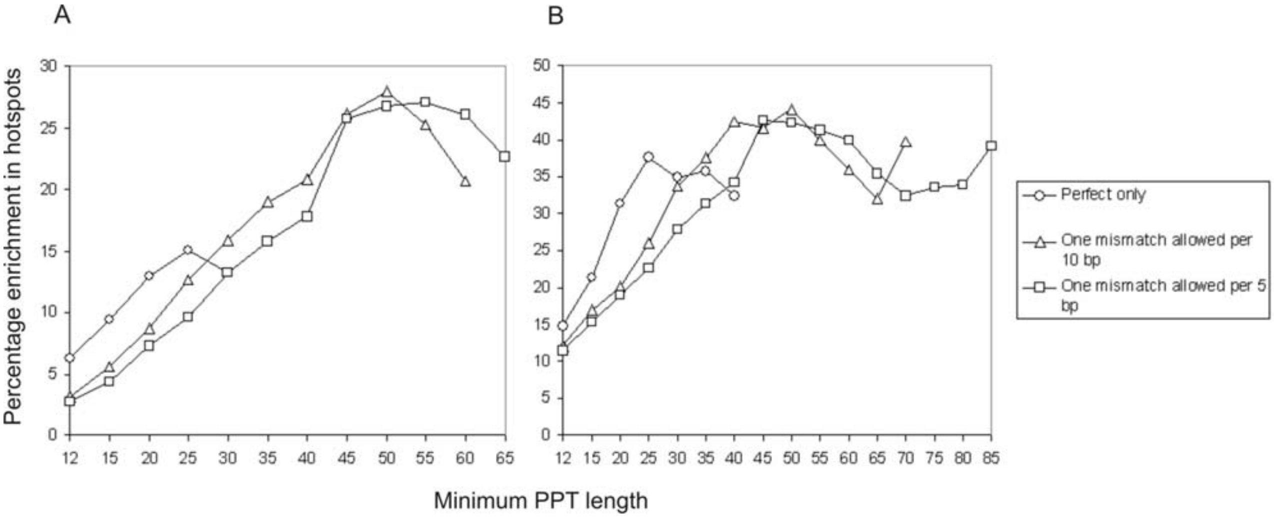


Figure 4.2: PPT enrichment in hotspots from the genome-wide dataset vs minimum tract length
Per kb frequencies of PPTs in 9298 hotspots and 9292 coldspots were averaged for the two types of regions and the percentage excess of the mean for hotspots over coldspots is plotted here for all PPTs (A) and for tracts with above the mean PPT GC-content (B). Data points are only shown for PPT types with a significant difference between hot- and coldspots ($p<0.01$, Mann-Whitney U test) for which all preceding (lower) limiters also give a significant difference.

4.3.8 The scale of the association between PPTs and human hotspots from the genome-wide dataset

I investigated the scale of the association between PPT density and hotspots from the genome-wide dataset [16] by comparing hotspot central and flanking regions with the corresponding areas of coldspots (Figures 4.3 and 4.4, Tables 4.7 and 4.8). For this analysis I used PPT coverage rather than many different size limiters in view of the above observations that PPT enrichment in hotspots increases with increasing tract length. Coverage of PPTs is significantly elevated in the 500 bp regions centred on hotspot mid points relative to the remaining parts of hotspots (Table 4.7). The elevation is 2-3 fold for high GC-content PPTs of at least 50 bp with one mismatch allowed per 10 bp ($p<10^{-8}$, Mann-Whitney U Test), the PPT search parameters that yielded the greatest level of enrichment in hotspots generally (Figure 4.2). Coverage of PPTs of at least 12 bp is also significantly elevated in hotspot central regions but the levels of enrichment are considerably lower than for high GC-content tracts of more than 50 bp considered separately (Table 4.7). In order to see if these associations could be related to elevated GC-content in central regions of hotspots, I compared mean overall GC-content in hotspots as a whole to that in hotspot central regions. The means are the same to within one percent and do not differ significantly, based on to the Mann-Whitney U Test, and also the T-Test (p>0.05).

Table 4.7 Elevated PPT coverage in the central regions of human hotspots from the genome-wide dataset
Ratios of mean PPT coverage, expressed as the proportion of bases covered, in hotspot central regions (denoted "hot CR", defined as 500 bp centred on the hotspot mid point), hotspot non-central regions (denoted "hot non-CR", defined as within hotspots but outside the central region) and coldspot central 500 bp regions (denoted "cold CR"). Statistical comparisons were made with the Mann-Whitney U test. Values for the means, and their standard errors, can be seen in Figures 4.3 and 4.4.

PPT type	Mean frequency ratio		P value	
	Hot CR vs Hot non-CR	Hot CR vs cold CR	Hot CR vs Hot non-CR	Hot CR vs cold CR
12 bp+ all	1.12	1.17	$<10^{-8}$	$<10^{-10}$
12 bp+ high GC	1.20	1.35	$<10^{-10}$	$<10^{-19}$
50 bp+ high GC	1.90	2.82	$<10^{-7}$	$<10^{-8}$

Coverage of PPTs 12 bp and longer in hotspot flanking regions 0-1 and 1-2 hotspot widths removed from hotspots is significantly elevated relative to coldspots, but the enrichment is modest, ranging between 2% and 6% (Table 4.8). High GC-content PPTs of at least 50 bp with one mismatch allowed per 10 bp do not cover significantly more, on average, of hotspot flanking regions than coldspots (Table 4.8).

Table 4.8 Elevated PPT coverage in the flanking regions of human hotspots from the genome-wide dataset
Ratios of mean PPT coverage in hotspots, hotspot flanking regions 0-1 and 1-2 hotspot widths removed from hotspots, and coldspots. Where hotspot flanking regions overlapped with coldspots, the coldspots were excluded from the analysis. Statistical comparisons were made with the Mann-Whitney U test (alpha = 0.01). Values for the means, and their standard errors, can be seen in Figures 4.3 and 4.4

PPT type	Mean frequency ratio			P value (comparison with coldspots)		
	Hotspots/ coldspots	Hotspot flanks 0-1 widths removed/ coldspots	Hotspot flanks 1-2 widths removed/ coldspots	Hotspots	Hotspot flanks 0-1 widths removed	Hotspot flanks 1-2 widths removed
12 bp+ all	1.07	1.03	1.02	$<10^{-18}$	0.0002	0.0055
12 bp+ high GC	1.17	1.06	1.03	$<10^{-39}$	$<10^{-6}$	n/s
50 bp+ high GC	1.51	1.13	1.15	$<10^{-7}$	n/s	n/s

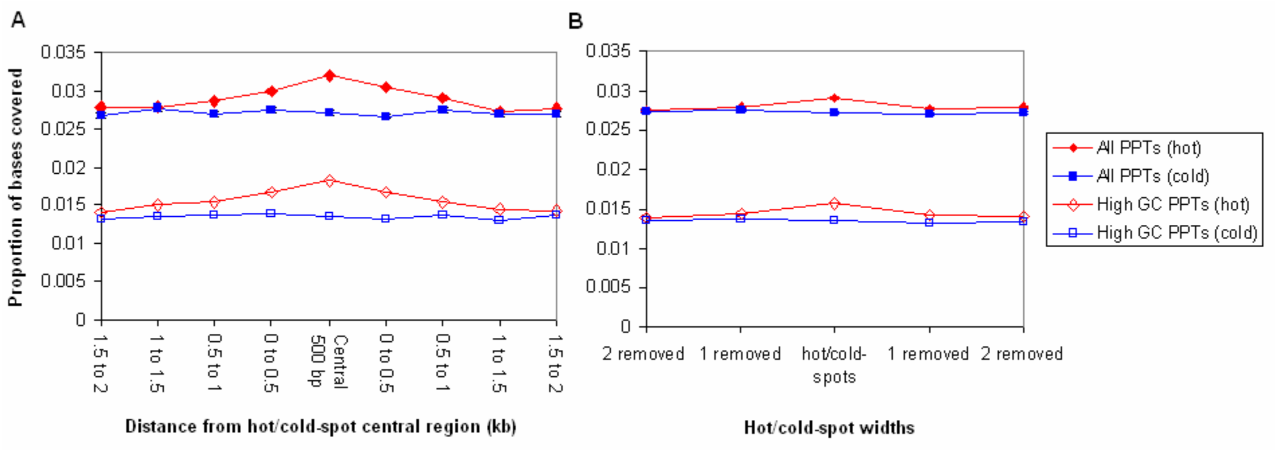


Figure 4.3 The scale of the association between hotspots and PPT coverage for tracts of at least 12 bp.
Mean proportion of bases covered by PPTs of at least 12 bp in relation to the central (A) and flanking (B) regions of human hotspots from the genome-wide dataset (mean hotspot width = 4070 bp, n=9298). For the analysis of flanking regions, each hot/cold-spot was extended for one-, and two-fold its own width on either side. Error bars (one SEM) were narrower than the symbol widths in every case, so are not shown.

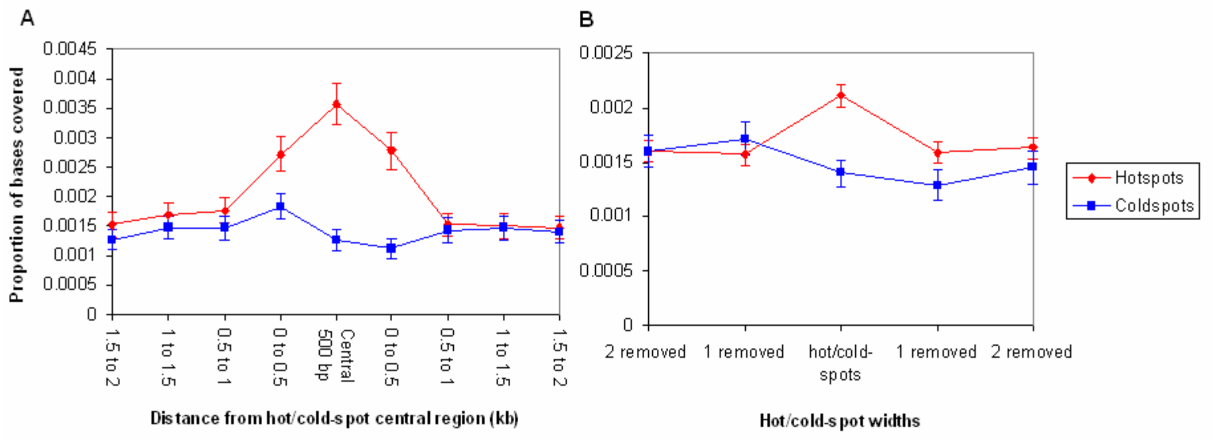


Figure 4.4 The scale of the association between hotspots and PPT coverage for high GC-content tracts of at least 50 bp.
Mean proportion of bases covered by high GC-content PPTs of at least 50 bp (one mismatch allowed per 10 bp) in relation to the central (A) and flanking (B) regions of human hotspots from the genome-wide dataset (mean hotspot width = 4070 bp, n=9298). For the analysis of hotspot flanking regions, each hot/cold-spot was extended for one-, and two-fold its own width on either side. In cases where this resulted in overlap between hot and cold regions, the cold ones were excluded from the analysis. Error bars are plus and minus one SEM.

4.3.9 Investigation of factors that could mediate the association between human recombination hotspots and PPTs

Having found that PPTs are over-represented in human hotspots I turned to the question of whether this association could have arisen coincidentally. Theoretically, this could occur as a result of factors associated with both hotspots and PPTs, and I investigated this possibility with linear models. The association could also be influenced by factors operating on a scale larger than hotspots, such as broad scale features of chromosome structure, and I investigated this possibility using wavelet analysis. The rationale behind these methods was as described in Sections 3.1 and 3.2.

Using linear models, I examined the extent to which the correlation between PPT coverage and recombination rate at a scale of 1 kb can be explained by SNP density, exon coverage and GC-content. I restricted this analysis to the 37 2^{15} kb (32.8 mb) regions of the human genome for which continuous data are available for all variables under consideration (see Section 3.2.2). This provided consistency with my wavelet analysis, which could only be done on these regions, excluded areas of the genome that were poorly annotated for one or more of the variables in question, and enabled evaluation of regional effects. PPTs of at least 12 bp are much more common than microsatellites (see Section 3.3) and their distribution can be approximately normalized by log transformation, so I used a linear model predicting the log-transformed variable. I did separate tests for prediction of PPTs by recombination rate, and by recombination rate with the additional predictors exon coverage, GC-content and SNP density (Table 4.9). I found that recombination rate is a consistent, but weak, predictor of PPT coverage, which is more strongly predicted by GC-content, judging by coefficient values. The regression coefficients for recombination are reduced by around 45 %, but are still highly significant, when correcting for the other factors.

Table 4.9: Predicting PPT coverage at a scale of one kilo base
Results from linear model analyses predicting coverage of PPTs of at least 12 bp, and high GC-content PPTs of at least 12 bp, in 37 2^{15} kb regions of the human genome. Prior to analysis, all variables were averaged for contiguous one kb windows covering the tested areas. Separate models were run for prediction of PPTs by recombination rate alone (denoted “Single”) and with the additional predictors exon coverage, density of SNPs and GC-content (denoted “Multiple”). The rightmost column shows numbers of regions with a significant positive (pos) or negative (neg) effect of the respective predictor (by Student’s T test; Bonferroni-adjusted alpha=0.00135 for the single regression model and 0.000338 for the multiple model). Overall significance was calculated by Stouffer’s method [36] in cases where the direction of correlation was consistent across all regions, and “inc” indicates that some regions showed negative effects and others positive effects.

PPT type	**Model type**	**Predictor**	**Estimated Coeff.**		**T**		**Pr(T>|t|)**	**# sig pos(neg)**
			Mean	mean SEM	Mean	SEM		
12 bp+	Single	Recombination	0.0060	0.0012	5.14	0.282	$<10^{-196}$	35(0)
	Multiple		0.0034	0.0012	2.92	0.208	$<10^{-56}$	13(0)
	Multiple	Exons	-0.0001	0.0000	-3.48	0.292	$<10^{-85}$	0(16)
	Multiple	GC-content	1.27	0.0537	23.4	0.944	$<10^{-300}$	37(0)
	Multiple	SNPs	0.0005	0.0027	0.17	0.294	inc	1(0)
12 bp+, high GC	Single	Recombination	0.0107	0.0013	9.60	0.501	$<10^{-210}$	36(0)
	Multiple		0.0058	0.0013	5.34	0.317	$<10^{-167}$	33(0)
	Multiple	Exons	-0.00003	0.00003	-0.95	0.446	inc	2(4)
	Multiple	GC-content	2.21	0.0561	44.2	1.283	$<10^{-300}$	37(0)
	Multiple	SNPs	0.0127	0.0029	4.72	0.293	$<10^{-167}$	28(0)

A relatively strong correlation between PPTs and very broad scale recombination rates has been noted previously in mammals [1, 2]. Using wavelet analysis, I addressed the question of whether this correlation can be attributed more to broad or fine scale interactions. Scale-specific wavelet detail coefficient analysis is ideal for this purpose, since it allows correlations at particular scales to be evaluated independently from the influence of other scales (see Section 3.2.2). The results show broad, medium and fine scale correlations, generally stronger at medium to broad scales (Figures 4.5 and 4.6). Substantial variation among regions in the strength of the correlation between PPT coverage and recombination is also clear, and the variation is not obviously related to proximity to centromeres or telomeres, nor to a difference between chromosomal long and short arms.

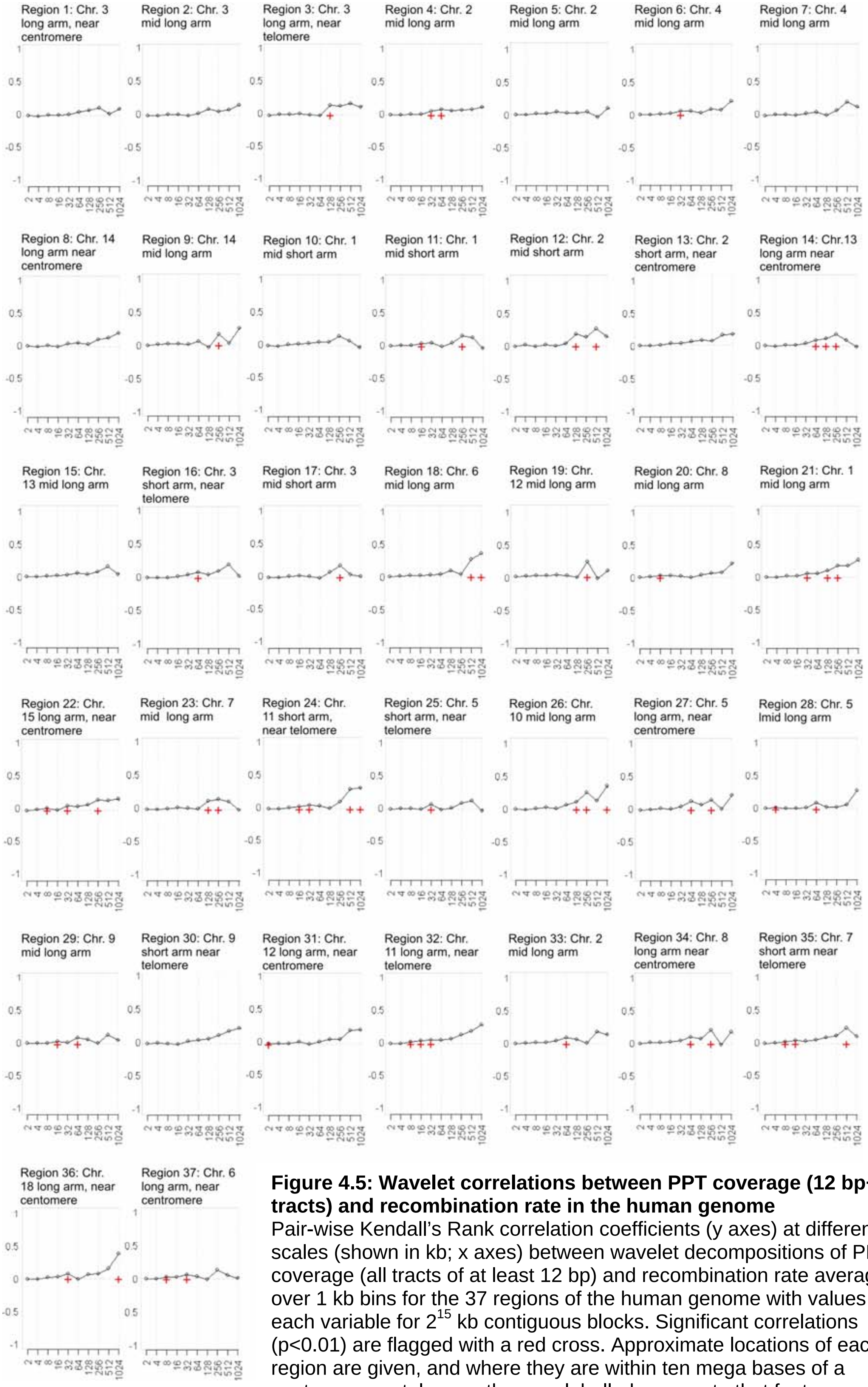
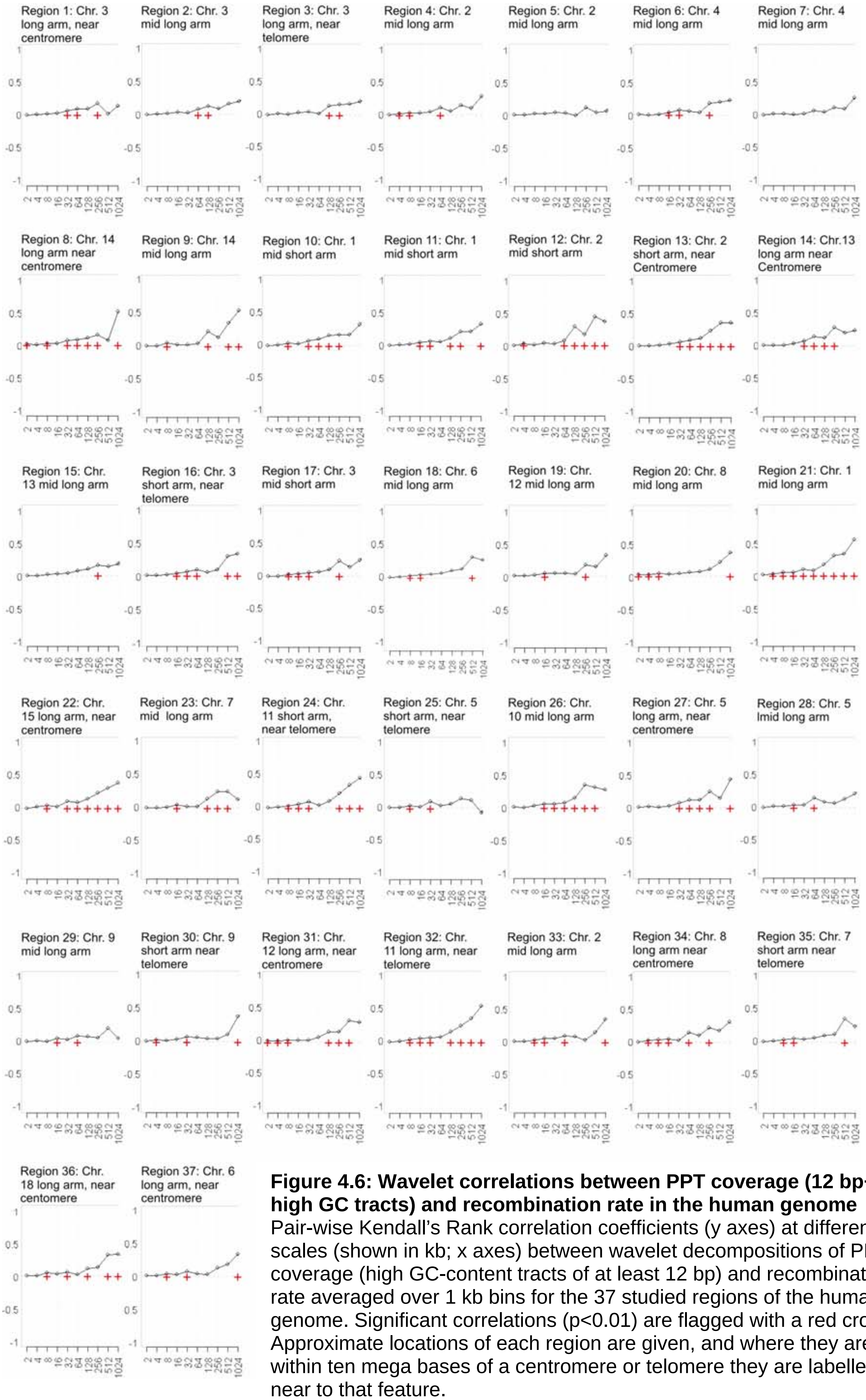


Figure 4.5: Wavelet correlations between PPT coverage (12 bp+ tracts) and recombination rate in the human genome
Pair-wise Kendall's Rank correlation coefficients (y axes) at different scales (shown in kb; x axes) between wavelet decompositions of PPT coverage (all tracts of at least 12 bp) and recombination rate averaged over 1 kb bins for the 37 regions of the human genome with values for each variable for 2^{15} kb contiguous blocks. Significant correlations ($p<0.01$) are flagged with a red cross. Approximate locations of each region are given, and where they are within ten mega bases of a centromere or telomere they are labelled as near to that feature.



Region 1: Chr. 3 long arm, near centromere

Region 2: Chr. 3 mid long arm

Region 3: Chr. 3 long arm, near telomere

Region 4: Chr. 2 mid long arm

Region 5: Chr. 2 mid long arm

Region 6: Chr. 4 mid long arm

Region 7: Chr. 4 mid long arm

Region 8: Chr. 14 long arm near centromere

Region 9: Chr. 14 mid long arm

Region 10: Chr. 1 mid short arm

Region 11: Chr. 1 mid short arm

Region 12: Chr. 2 mid short arm

Region 13: Chr. 2 short arm, near Centromere

Region 14: Chr.13 long arm near Centromere

Region 15: Chr. 13 mid long arm

Region 16: Chr. 3 short arm, near telomere

Region 17: Chr. 3 mid short arm

Region 18: Chr. 6 mid long arm

Region 19: Chr. 12 mid long arm

Region 20: Chr. 8 mid long arm

Region 21: Chr. 1 mid long arm

Region 22: Chr. 15 long arm, near centromere

Region 23: Chr. 7 mid long arm

Region 24: Chr. 11 short arm, near telomere

Region 25: Chr. 5 short arm, near telomere

Region 26: Chr. 10 mid long arm

Region 27: Chr. 5 long arm, near centromere

Region 28: Chr. 5 Imid long arm

Region 29: Chr. 9 mid long arm

Region 30: Chr. 9 short arm near telomere

Region 31: Chr. 12 long arm, near centromere

Region 32: Chr. 11 long arm, near telomere

Region 33: Chr. 2 mid long arm

Region 34: Chr. 8 long arm near centromere

Region 35: Chr. 7 short arm near telomere

Region 36: Chr. 18 long arm, near centomere

Region 37: Chr. 6 long arm, near centromere

Figure 4.6: Wavelet correlations between PPT coverage (12 bp+ high GC tracts) and recombination rate in the human genome
Pair-wise Kendall's Rank correlation coefficients (y axes) at different scales (shown in kb; x axes) between wavelet decompositions of PPT coverage (high GC-content tracts of at least 12 bp) and recombination rate averaged over 1 kb bins for the 37 studied regions of the human genome. Significant correlations (p<0.01) are flagged with a red cross. Approximate locations of each region are given, and where they are within ten mega bases of a centromere or telomere they are labelled as near to that feature.

4.3.10 Sliding window analysis of the distributions of PPTs and GC-content relative to experimentally characterized human hotspots

I further investigated the scale of the association between PPTs and human hotspots using sliding window analysis. Sliding window plots of the density variations of high GC-content PPTs of at least 12 bp across the two regions in which multiple human hotspots have been well characterized experimentally (see Section 3.2.1) are shown in Figure 4.9. Peaks in PPT density often occur close to hotspots. The association is weaker when PPTs of all GC contents are considered, and when the lower PPT size limit is raised the association becomes stronger for some hotspot regions but weaker for others (Appendix B, Figures B1 and B2). Comparing the variation in PPT density over these regions (Figure 4.9) with the variation in GC-content (Figure 4.10) it is apparent that hotspots in the MHC class II studied region are associated with broad scale elevation of both PPT abundance and GC-content, but hotspots in the studied region of chromosome 1 are associated with PPTs most obviously at a fine scale, and not with GC-content at any scale.

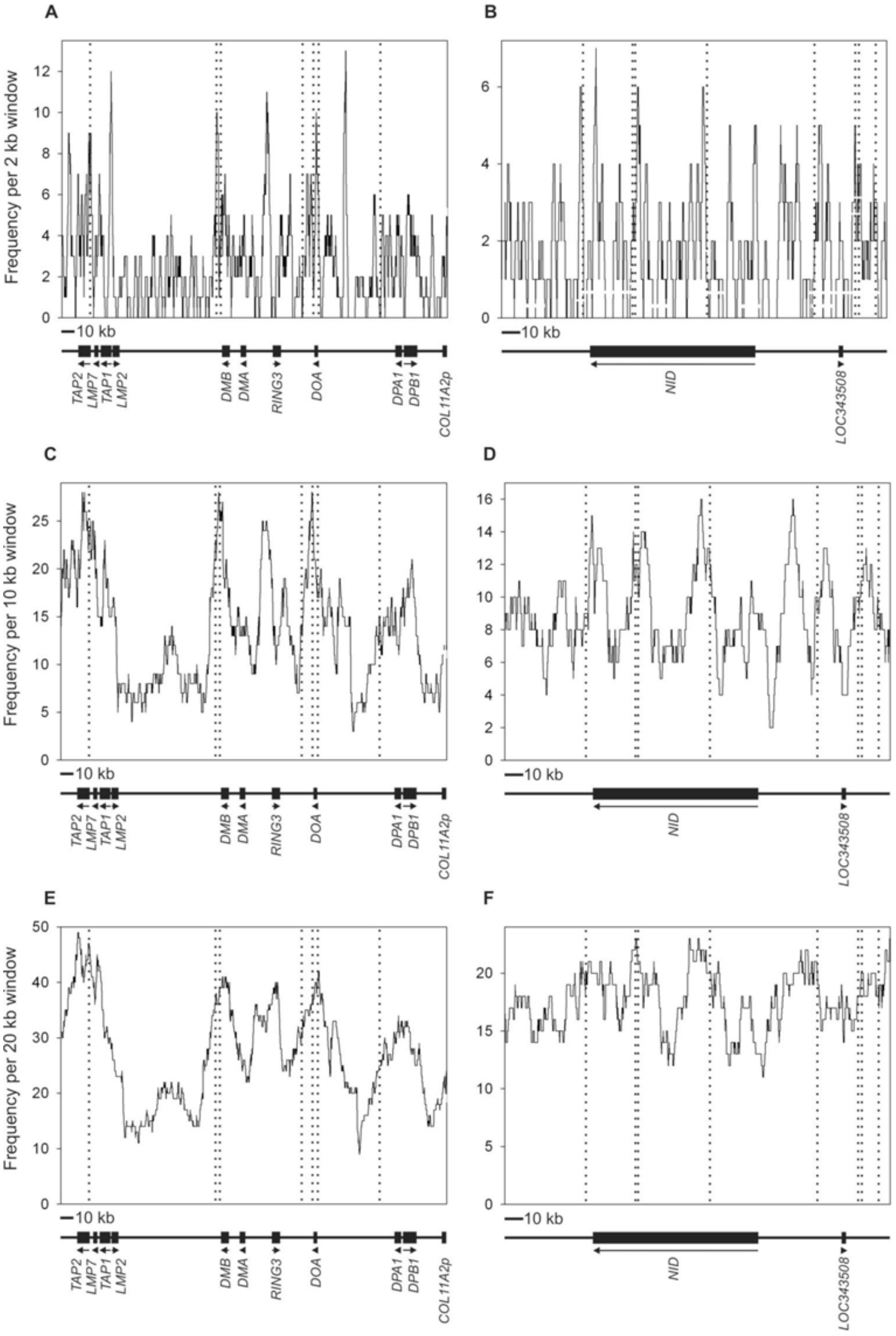


Figure 4.9: Densities of high GC-content PPTs relative to human hotspot locations. Sliding window plots of the densities of PPTs of at least 12 bp, with GC-contents above the mean for PPTs in these regions, relative to hotspot locations in the two contiguous areas of the human genome over which multiple hotspots have been well characterized experimentally: A, C and E: a 292 kb region of the human MHC Class II region in which 7 hotspots have been mapped [10, 11] and B, D and F: a 206 kb region of human chromosome 1 in which 8 hot have been mapped [12]. Sliding window plots with different window sizes are shown: 2 kb (A and B), 10 kb (C and D) and 20 kb (E and F). Vertical dotted lines represent hotspot mid point locations. Sliding windows moved in steps of 100 bp. Locations of genes in are shown below the plots with arrows indicating direction of transcription.

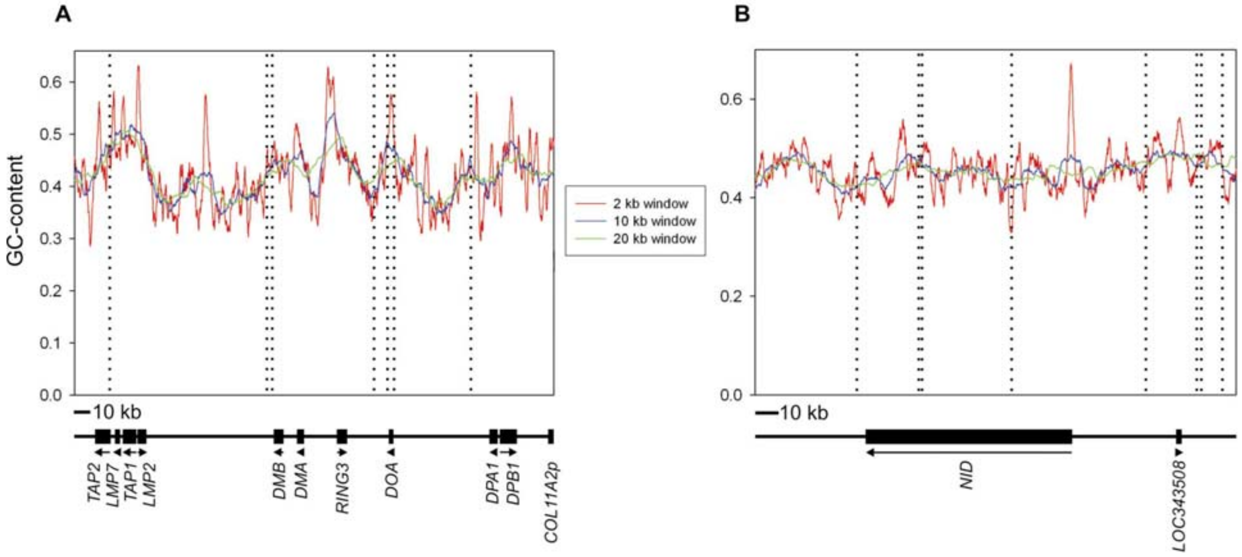


Figure 4.10: GC-content variation in two human hotspot-containing regions.
GC-content plotted in sliding windows of 2 (red), 10 (blue) and 20 (green) kb relative to recombination hotspots in the MHC Class II region (A) and a 206 kb region of chromosome 1 (B). Vertical dotted lines represent hotspot mid point locations. Sliding windows moved in steps of 100 bp. Locations of genes in are shown below the plots with arrows indicating direction of transcription.

4.3.11 Sequence changes associated with recombination occur in poly-purine-rich contexts

Until 2008, there were only three known cases in humans of single nucleotide changes associated with altered recombination levels in hotspots [37-39]. All three polymorphisms are associated with several-fold reductions in recombination frequency and are located close to the estimated hotspot mid points. I noticed that each of these polymorphisms occurs within 3 bp of the end of a sequence 14 bp or longer consisting of 85% or more poly-pu/py and at least 70% G/C (Table 4.10).

Table 4.10: Poly-purine-rich sequence contexts of polymorphisms associated with hotspot activity in humans.
Sequence contexts of the three polymorphisms associated with reduced recombination frequencies in human hotspots. The recombination-suppressing alleles are shown in lower case.

Hot spot	References	Sequence context	Distance from hotspot mid point (bp)
MS32	[39]	(G/c)GTGGGAAGGGTGG	151
*NID*1	[12, 38]	CC(C/t)CCCACCCCACCCC	64
*DNA*2	[10, 37]	AGGGGGCAGCAACAGGG(A/g)GG	166

4.4 Discussion

Work presented in this chapter constituted the first published report of an association between PPTs and recombination hotspots in any species [8] (see page xi, Publications associated with this thesis). The most striking aspect of the association, common to both humans and the yeast *S. cerevisiae*, is that its magnitude increases with increasing tract length, and in both species this tendency is more marked for high GC-content tracts. These observations suggest the possibility of a link with non-B-DNA structures, since while the sequence requirements for these structures to form are not fully understood, they are preferentially formed by PPTs under physiological conditions [3, 18, 20, 23, 24, 41, 42], though mismatches can be tolerated [18, 20, 24], and the tendency for them to form increases with both increasing tract length [43, 44], and increasing tract GC-content [5]. Other factors could also drive the association between PPTs and recombination hotspots, however, and I have attempted to test as many of these as practicable.

In yeast, I was able to show that the association is not significantly influenced by frequently transcribed, GC-rich, or promoter regions. It is also not mediated by known transposable elements, since these are not over-represented in the yeast hotspots I studied [9]. While PPTs are highly enriched in yeast DSB hotspots, however, the coefficients of the correlation between DSB intensity and recombination rate genome-wide are generally quite weak. This clearly results in part from the fact that most PPTs occur outside hotspots, though at lower frequency. The other factor I identified as contributing to this apparent discrepancy was that PPTs are not associated preferentially with the most active hotspots, judging by the fact that there is no significant difference in PPT abundance between the most DSB-enriched half of hotspots and remaining hotspots.

Using linear models, I found that recombination rate is also a weak predictor of density of PPTs of at least 12 bp at a scale of 1 kb in humans. This was not surprising considering the modest enrichment of PPTs of this length type in hotspots from the hapolotype inference recombination map of the human genome I utilized and, although correcting for potential mediating factors reduces the coefficients contributed by recombination as a predictor by about 45%, recombination remains a significant predictor of PPTs, showing that these factors cannot explain the link between PPTs and recombination at a fine scale. However, the strongest predictor of PPTs at a scale of 1 kb is GC-content, and elucidation of the association between PPTs and recombination in the human genome is also complicated by the results of my scale-specific wavelet correlation analysis, which showed that influences operating on a larger scale than hotspots are important. This was also indicated by the observed substantial variation in the magnitude and significance of the correlation among the 37 studied regions. The reason for this is unclear, since an effect of chromosome arm is not apparent, and there are strongly and weakly correlating telomere-proximal and centromere-proximal regions. The regional effects might, therefore, be linked to the previously reported very large scale variation in recombination rate known as recombination “deserts” and “jungles” [1, 16, 45, 46]. I was not able to investigate scales larger than one mega base with wavelet analysis due to breaks in the data, but this would be possible with more recent recombination maps of the human genome (see Sections 3.3.2 and 3.4).

The most obvious potential large-scale influence on the correlation between recombination and PPTs is GC-content variation. High GC-content might slow the progress of DNA replication, potentially stimulating recombination [35], and possibly also giving rise to a mutation bias driving increased PPT density, since stalled replication is mutagenic to microsatellite PPTs [47]. However, several arguments strongly suggest that a mutation bias driven by GC-content, or other factors operating on a large scale, cannot adequately explain the patterns I found. Firstly, correlation analysis considered in isolation may be misleading for sporadically distributed variables such as recombination. Although the correlation between PPT density and GC-content is apparently quite strong, the potential to detect an effect of recombination in regions that seldom recombine, which encompass the vast majority of the genome, is low. In comparison with recombination rate, GC-content does not vary sporadically, and a slight increase in GC-content associated with a large proportion of PPTs and hotspots, which would not be surprising in the case of high GC-content PPTs, might give rise to the strong predictive power of GC-content relative to recombination in view of the fact that only a relatively small proportion of PPTs is associated with hotspots. Another argument against the possibility that high GC-content mediates a general mutagenic effect driving the association between hotspots and PPTs is that sliding window analysis showed that the experimentally characterized hotspots on human chromosome 1 are associated with PPTs at a fine scale of 2 kb, but not with high GC-content at any scale. This indicates that the mutual association between GC-content, PPTs and recombination hotspots is region-specific. Given the inconsistency of the association, it seems unlikely that GC-content generally operates to drive a mutation bias, giving rise to high PPT density in hotspots. I have tested this possibility further in Chapter 6 by investigating whether high GC-content is generally associated with increased levels of PPT polymorphism.

In relation to the influence of large-scale factors on the correlation between PPT density and recombination shown by wavelet analysis, two points should be considered. Firstly, the broad scale correlations I observed using wavelet analysis could be due in part to a preferential association of PPTs with hotspots that occur in clusters, rather than a distal association with recombination events, and this is supported by the fact that PPTs are only slightly enriched in hotspot flanking regions. Secondly, the experimentally characterized hotspots on human chromosome 1 are associated with PPTs at a fine scale despite that fact that PPT density in the region as a whole is lower than in the MHC class II hotspot-containing region studied (Figure 4.9). This suggests the possibility that the apparent relative increase in strength of the correlation at broad scales compared with fine scales could be inflated by high PPT density in some non-hot areas of broad PPT-rich regions.

The strongest argument against the hypothesis that regional indirect factors determine the association between PPTs and recombination hotspots is that PPT density increases with proximity to hotspot central regions. Furthermore, no such central tendency is seen for GC-content. The association between hotspots and PPT frequency is somewhat weaker in humans than in yeast, but a central tendency of PPT enrichment within hotspots is quite marked in the human genome. The genome-wide dataset of hotspots I studied was derived from haplotype data [16], so the hotspots are not well enough characterized for their true mid points to be known, but the dataset is large enough (n=9298) that overall averages should reflect a mutual cancelling out of variations of the true mid points either side of the mid points I defined, i.e. halfway between hotspot start and end coordinates. Moreover, a central tendency of PPT enrichment is also evident in the 17 experimentally well-characterized hotspots I studied, including a significantly elevated frequency, in their 500 bp central regions, of high GC-content PPTs of at least 20 bp. Whether a similar central tendency occurs in yeast could not be tested with the available data, since some yeast hotspots are concentrated at the 5' ends of genes rather than in the central regions of IGRs (reviewed in [35]), but a close association of PPTs with recombination-initiating DSBs is suggested by my finding that PPTs are significantly enriched in yeast IGRs that harbour very low frequencies of DSBs and have not been defined as recombination hotspots. The increased average PPT coverage in the central 500 bp of human hotspots is also suggestive of a close relationship between recombination and PPTs, because crossover frequencies increase toward the mid points of hotspots in mammals [30, 31].

Recombination-mediated mutations causing PPTs to form and/or grow and a stimulatory effect of PPTs on recombination are two possible ways in which a causal link between PPTs and recombination could occur. I could find no evidence that short PPTs are associated with hotspots, suggesting that PPT formation is not often caused by recombination. As is the case for microsatellites, however, it is possible that recombination could directly drive PPT length increase *via* a mutation bias, and this could give rise to the broad scale correlation pattern I observed if recombination evolves quickly at a fine scale but is constrained at broader scales (see Section 3.4). If a mutation bias is at work, local increases in PPT polymorphism levels should be detectable near recombination sites, and I have explored this possibility in Chapter 6. It seems unlikely that a mutation bias is the sole explanation of the associations I have found, however, because the recombination landscape is short-lived in evolutionary time [48, 49], so to drive the association between PPTs and hotspot central regions, hotspots must recur at the same chromosomal locations, and no other type of sequence has previously been found to be associated with these narrow 500 bp regions, with the exception of a modest enrichment of microsatellites (see Section 3.3).

The data presented in this chapter therefore suggest that PPTs may have a widespread functional role in recombination hotspots. The fact that most PPTs occur outside hotspots, and the consequently weak correlation between recombination rate and PPT frequency, are consistent with a functional role of PPTs in at least some hotspots if only some tracts are functional, and/or if high PPT frequency is only one among several factors working together in hotspot control. Other factors clearly are involved in regulating recombination (see Section 1.2.2). The influence of non-sequence (epigenetic) factors is shown by sex-specific hotspot use [40], and sequences outside hotspots are also important [50-55]. How these factors operate in conjunction with local sequence to regulate recombination is essentially unknown, and their involvement does not rule out a functional role for PPTs, which could occur by several plausible mechanisms, and has been directly demonstrated, in one case, for poly-A [4].

High PPT density in itself could in be involved in view of the fact that PPTs can stick together *via* Hoogsteen (non-Watson-Crick) base pairing interactions, [7, 25] and it has been suggested that these interactions might help homologous chromosomes to align prior to meiotic recombination [25]. Another possible scenario is that PPTs could potentiate hotspots by binding proteins that interact with recombination machinery. PPTs can bind transcription factors [28, 56], so they might have a role in alpha hotspots (see Section 1.2.2). The discovery of several poly-pu/py-rich motifs of 5-9 bp in association with hotspots [16] might be linked to such a role. Binding with other types of proteins could also mediate a function of PPTs in the recombination process, for example intramolecular quadruplex secondary structures, which are formed by poly-purine-rich, GC-rich sequences [24] can bind the nuclear matrix-associated type III intermediate filament proteins [57], suggesting a role for these elements in higher-order chromosome structure.

An involvement of the non-B-DNA structures intramolecular quadruplexes, and/or intramolecular triplexes could occur in several other ways. The structures include some single-stranded DNA, which could itself be recombinagenic [3]. Furthermore, they have been implicated in creating nucleosome-free regions of chromatin [58, 59], suggesting that they might be involved in regulating hotspots under the beta model (see Section 1.2.2). This was suggested as the reason behind the functional involvement of poly-A at the yeast *ARG4* hotspot in view of observations that poly-A can exclude nucleosomes, though without forming a secondary structure [4]. Finally, the potential of PPTs to cause replication pausing [20, 60, 61], (reviewed in [62]), suggests that they could also be involved in creating gamma hotspots (see Section 1.2.2). Intramolecular triplex formation might actually occur predominantly during DNA replication, since it has been proposed that strand displacement during replication could bring three strands into close proximity causing triplex formation and resultant replication pausing [60].

The exact conditions needed for PPTs to form secondary structures on chromosomes are not yet known, but immunocytological evidence has shown that intramolecular triplexes do occur on human chromosomes *in vivo* [63]. Interestingly, intramolecular structure formation by PPTs can be sensitive to single nucleotide changes [64-66]. One reason for structural variation is likely to be the supply of torsional energy on the chromosome. The requirement for this is the reason why longer PPTs are more likely to form intramolecular triplexes *in vivo*, because the torsional energy required reduces with increasing PPT length [44]. This energy is limited, which is the explanation of the observation that two non-B-DNA structure-forming PPTs within 1500 bp of each other cannot both form structures on plasmid DNA simultaneously [64]. Potentially, this could be one explanation for the phenomenon of local competition between hotspots [32, 51, 54, 67, 68], and it illustrates a way in which high PPT density in some regions outside hotspots is consistent with a function for the sequences in hotspot recombination.

Although very long PPTs have greater potential to form intramolecular triplexes, GC-rich PPTs as short as 12 bp have been shown to form intramolecular quadruplexes [19]. It is therefore suggestive that all three human polymorphisms shown to affect recombination occur within 3 bp of the end of a sequence 14 bp or longer consisting of 85% or more poly-pu/py and at least 70% G/C. The wider generality of this observation will soon be tested, with the emergence of increasing amounts of data on polymorphic hotspots genome-wide [40].

The results presented in this chapter therefore indicate that the possibility of a widespread causal link between PPTs and recombination should be considered plausible. In particular, a functional role for PPTs in hotspots should be further explored. It remains possible that recombination and PPTs could be linked causally by a mutation bias, and in Chapter 6 I detail an investigation of the possibility that recombination mediates a mutation bias in PPTs. In Chapter 7 I present preliminary results from a test of the non-B-DNA structure-forming potential of sequence amplified from human hotspot central regions.

References

1. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G *et al*: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**(3):241-247.
2. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ: **Comparative recombination rates in the rat, mouse, and human genomes**. *Genome Res* 2004, **14**(4):528-538.
3. Rooney SM, Moore PD: **Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells**. *Proc Natl Acad Sci U S A* 1995, **92**(6):2141-2144.

4. Schultes NP, Szostak JW: **A poly(dA.dT) tract is a component of the recombination initiation site at the ARG4 locus in Saccharomyces cerevisiae**. *Mol Cell Biol* 1991, **11**(1):322-328.
5. Hanvey JC, Klysik J, Wells RD: **Influence of DNA sequence on the formation of non-B right-handed helices in oligopurine.oligopyrimidine inserts in plasmids**. *J Biol Chem* 1988, **263**(15):7386-7396.
6. Drew HR, Travers AA: **DNA bending and its relation to nucleosome positioning**. *J Mol Biol* 1985, **186**(4):773-790.
7. Kiyama R, Nishikawa N, Oishi M: **Enrichment of human DNAs that flank poly(dA).poly(dT) tracts by triplex DNA formation**. *J Mol Biol* 1994, **237**(2):193-200.
8. Bagshaw AT, Pitt JP, Gemmell NJ: **Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots**. *BMC Genomics* 2006, **7**:179.
9. Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD: **Inaugural article: global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae**. *Proc Natl Acad Sci U S A* 2000, **97**(21):11383-11390.
10. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex**. *Nat Genet* 2001, **29**(2):217-222.
11. Kauppi L, Stumpf MP, Jeffreys AJ: **Localized breakdown in linkage disequilibrium does not always predict sperm crossover hot spots in the human MHC class II region**. *Genomics* 2005, **86**(1):13-24.
12. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: **Human recombination hot spots hidden in regions of strong marker association**. *Nat Genet* 2005, **37**(6):601-606.
13. Schneider JA, Peto TE, Boone RA, Boyce AJ, Clegg JB: **Direct measurement of the male recombination fraction in the human beta-globin hot spot**. *Hum Mol Genet* 2002, **11**(3):207-215.
14. Wall JD, Frisse LA, Hudson RR, Di Rienzo A: **Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates**. *Am J Hum Genet* 2003, **73**(6):1330-1340.
15. May CA, Shone AC, Kalaydjieva L, Sajantila A, Jeffreys AJ: **Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX**. *Nat Genet* 2002, **31**(3):272-275.
16. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**(5746):321-324.
17. Baudat F, Nicolas A: **Clustering of meiotic double-strand breaks on yeast chromosome III**. *Proc Natl Acad Sci U S A* 1997, **94**(10):5213-5218.
18. Raghavan SC, Chastain P, Lee JS, Hegde BG, Houston S, Langen R, Hsieh CL, Haworth IS, Lieber MR: **Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation**. *J Biol Chem* 2005, **280**(24):22749-22760.
19. Matsugami A, Ouhashi K, Kanagawa M, Liu H, Kanagawa S, Uesugi S, Katahira M: **New quadruplex structure of GGA triplet repeat DNA--an intramolecular quadruplex composed of a G:G:G:G tetrad and G(:A):G(:A):G(:A):G heptad, and its dimerization**. *Nucleic Acids Res Suppl* 2001(1):271-272.

20. Dayn A, Samadashwily GM, Mirkin SM: **Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization**. *Proc Natl Acad Sci U S A* 1992, **89**(23):11406-11410.
21. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S: **Quadruplex DNA: sequence, topology and structure**. *Nucleic Acids Res* 2006, **34**(19):5402-5415.
22. Hanvey JC, Shimizu M, Wells RD: **Intramolecular DNA triplexes in supercoiled plasmids. II. Effect of base composition and noncentral interruptions on formation and stability**. *J Biol Chem* 1989, **264**(10):5950-5956.
23. Simonsson T: **G-quadruplex DNA structures--variations on a theme**. *Biol Chem* 2001, **382**(4):621-628.
24. Dai J, Dexheimer TS, Chen D, Carver M, Ambrus A, Jones RA, Yang D: **An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution**. *J Am Chem Soc* 2006, **128**(4):1096-1098.
25. Sen D, Gilbert W: **Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis**. *Nature* 1988, **334**(6180):364-366.
26. Qin Y, Hurley LH: **Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions**. *Biochimie* 2008.
27. Maiti AK, Brahmachari SK: **Poly purine.pyrimidine sequences upstream of the beta-galactosidase gene affect gene expression in Saccharomyces cerevisiae**. *BMC Mol Biol* 2001, **2**(1):11.
28. Lu Q, Teare JM, Granok H, Swede MJ, Xu J, Elgin SC: **The capacity to form H-DNA cannot substitute for GAGA factor binding to a (CT)n*(GA)n regulatory site**. *Nucleic Acids Res* 2003, **31**(10):2483-2494.
29. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome**. *Cell* 1998, **95**(5):717-728.
30. Baudat F, de Massy B: **Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot**. *PLoS Genet* 2007, **3**(6):e100.
31. Kauppi L, Jeffreys AJ, Keeney S: **Where the crossovers are: recombination distributions in mammals**. *Nat Rev Genet* 2004, **5**(6):413-424.
32. Xu L, Kleckner N: **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* 1995, **14**(20):5115-5128.
33. de Massy B, Rocco V, Nicolas A: **The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in Saccharomyces cerevisiae**. *Embo J* 1995, **14**(18):4589-4598.
34. Xu F, Petes TD: **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* 1996, **143**(3):1115-1125.
35. Petes TD: **Meiotic recombination hot spots and cold spots**. *Nat Rev Genet* 2001, **2**(5):360-369.
36. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RMJ: **The American Soldier, Vol 1: Adjustment during army life.** Princeton: Princeton University Press; 1949.
37. Jeffreys AJ, Neumann R: **Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot**. *Nat Genet* 2002, **31**(3):267-271.

38. Jeffreys AJ, Neumann R: **Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot**. *Hum Mol Genet* 2005, **14**(15):2277-2287.

39. Jeffreys AJ, Murray J, Neumann R: **High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot**. *Mol Cell* 1998, **2**(2):267-273.

40. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M: **High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans**. *Science* 2008, **319**(5868):1395-1398.

41. Kohwi Y, Kohwi-Shigematsu T: **Altered gene expression correlates with DNA structure**. *Genes Dev* 1991, **5**(12B):2547-2554.

42. Radhakrishnan I, Patel DJ: **DNA triplexes: solution structures, hydration sites, energetics, interactions, and function**. *Biochemistry* 1994, **33**(38):11405-11416.

43. Collier DA, Wells RD: **Effect of length, supercoiling, and pH on intramolecular triplex formation. Multiple conformers at pur.pyr mirror repeats**. *J Biol Chem* 1990, **265**(18):10652-10658.

44. Michel D, Chatelain G, Herault Y, Brun G: **The long repetitive polypurine/polypyrimidine sequence (TTCCC)48 forms DNA triplex with PU-PU-PY base triplets in vivo**. *Nucleic Acids Res* 1992, **20**(3):439-443.

45. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW *et al*: **Comparison of human genetic and sequence-based physical maps**. *Nature* 2001, **409**(6822):951-953.

46. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome**. *Science* 2004, **304**(5670):581-584.

47. Hile SE, Eckert KA: **Positive correlation between DNA polymerase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences.** *J Mol Biol* 2004, **335**(3):745-759.

48. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**(4):429-434.

49. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P *et al*: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**(5718):107-111.

50. Haring SJ, Halley GR, Jones AJ, Malone RE: **Properties of natural double-strand-break sites at a recombination hotspot in Saccharomyces cerevisiae**. *Genetics* 2003, **165**(1):101-114.

51. Wu TC, Lichten M: **Factors that affect the location and frequency of meiosis-induced double-strand breaks in Saccharomyces cerevisiae**. *Genetics* 1995, **140**(1):55-66.

52. Borde V, Wu TC, Lichten M: **Use of a recombination reporter insert to define meiotic recombination domains on chromosome III of Saccharomyces cerevisiae**. *Mol Cell Biol* 1999, **19**(7):4832-4842.

53. Steiner WW, Smith GR: **Optimizing the Nucleotide Sequence of a Meiotic Recombination Hotspot in Schizosaccharomyces pombe**. *Genetics* 2005.

54. Zahn-Zabal M, Lehmann E, Kohli J: **Hot spots of recombination in fission yeast: inactivation of the M26 hot spot by deletion of the ade6 promoter and the novel hotspot ura4-aim**. *Genetics* 1995, **140**(2):469-478.

55. Mieczkowski PA, Dominska M, Buck MJ, Gerton JL, Lieb JD, Petes TD: **Global analysis of the relationship between the binding of the Bas1p transcription factor and meiosis-specific double-strand DNA breaks in Saccharomyces cerevisiae**. *Mol Cell Biol* 2006, **26**(3):1014-1027.
56. Sandaltzopoulos R, Mitchelmore C, Bonte E, Wall G, Becker PB: **Dual regulation of the Drosophila hsp26 promoter in vitro**. *Nucleic Acids Res* 1995, **23**(13):2479-2487.
57. Li G, Tolstonog GV, Traub P: **Interaction in vitro of type III intermediate filament proteins with triplex DNA**. *DNA Cell Biol* 2002, **21**(3):163-188.
58. Elgin SC: **The formation and function of DNase I hypersensitive sites in the process of gene activation**. *J Biol Chem* 1988, **263**(36):19259-19262.
59. Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin**. *Annu Rev Biochem* 1988, **57**:159-197.
60. Samadashwily GM, Dayn A, Mirkin SM: **Suicidal nucleotide sequences for DNA polymerization**. *Embo J* 1993, **12**(13):4975-4983.
61. Krasilnikova MM, Mirkin SM: **Replication stalling at Friedreich's ataxia (GAA)n repeats in vivo**. *Mol Cell Biol* 2004, **24**(6):2286-2295.
62. Rao BS: **Regulation of DNA replication by homopurine/homopyrimidine sequences**. *Mol Cell Biochem* 1996, **156**(2):163-168.
63. Ohno M, Fukagawa T, Lee JS, Ikemura T: **Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies**. *Chromosoma* 2002, **111**(3):201-213.
64. Bacolla A, Ulrich MJ, Larson JE, Ley TJ, Wells RD: **An intramolecular triplex in the human gamma-globin 5'-flanking region is altered by point mutations associated with hereditary persistence of fetal hemoglobin**. *J Biol Chem* 1995, **270**(41):24556-24563.
65. Boles TC, Hogan ME: **DNA structure equilibria in the human c-myc gene**. *Biochemistry* 1987, **26**(2):367-376.
66. Ulrich MJ, Gray WJ, Ley TJ: **An intramolecular triplex is disrputed by point mutations associated with hereditary persistence of fetal hemoglobin**. *Journal of Biological Chemistry* 1992, **267**:18649-18658.
67. Fan QQ, Xu F, White MA, Petes TD: **Competition between adjacent meiotic recombination hotspots in the yeast Saccharomyces cerevisiae**. *Genetics* 1997, **145**(3):661-670.
68. Ohta K, Wu TC, Lichten M, Shibata T: **Competitive inactivation of a double-strand DNA break site involves parallel suppression of meiosis-induced changes in chromatin configuration**. *Nucleic Acids Res* 1999, **27**(10):2175-2180.

Chapter 5

Does a mutation bias drive the association between microsatellites and recombination?

Abstract

If recombination is mutagenic to microsatellites, there should be an increase in microsatellite polymorphism levels in frequently recombining regions. Recombination can also maintain neutral polymorphism in general by interrupting the effects of natural selection, but this effect should be seen at broader scales than recombination hotspots as well as within them. In collaboration with workers from Uppsala University, Sweden, I have tested the association between recombination and microsatellite polymorphism in the human genome. I used a published set of over 400,000 insertion/deletion (indel) polymorphisms that were derived from shotgun sequencing initiatives. Defining as polymorphic all microsatellties harbouring at least one indel, I initially examined the relationship between recombination hotspots and polymorphic microsatellites. I found that the fraction of microsatellites harbouring polymorphisms is not elevated in hotspots from a genome-wide dataset compared with coldspots of equivalent number and size. It is slightly increased in hotspot flanking regions, but not in hotspot central regions. A generalized linear model predicting polymorphic microsatellites at a scale of one kilo base while correcting for microsatellite distribution, indels, single nucleotide polymorphisms, GC-content and gene density showed that recombination predicts microsatellite polymorphism very weakly and inconsistently, with some regions showing a slight negative association. Taken together, these results suggest that it is unlikely that recombination, or any property of recombination hotspots, is commonly mutagenic to microsatellites in the human genome.

5.1 Introduction

5.1.1 Background

The idea that meiotic recombination could drive the evolution of repetitive sequences through unequal crossing over between chromosomes dates back to at least 1976, and was initially based on the fact that homologous chromosomes are theoretically more likely to misalign at direct sequence repeats [1] (Figure 5.1 part A). Unequal crossover has been observed to occur between satellite repeats, which have extremely long periodicity, but studies of microsatellite mutations that have reported checking for exchange of flanking markers, which would be expected in unequal crossing over, have almost invariably found no evidence of this (reviewed in [2]). A more recent idea is that gene conversions, unequal recombination events that do not involve reciprocal exchange of information between chromosomes, and hence do not cause exchange of flanking markers, could cause change of array length mutations in microsatellites, as they do in minisatellites [3-5]. Aberrant meiotic recombination events without exchange of flanking markers have been implicated in cases of extreme instability at some microsatellite loci implicated in human genetic disease, (reviewed in [2, 6]), and these may occur as a result of strand slippage in recombination intermediates [7] (Figure 5.1 part C). Evidence has, however, counted against recombination-linked processes being considered a significant factor in microsatellite evolution. Microsatellite instability was not found to be reduced in recombination deficient strains of *E. coli* [8] or *S. cerevisiae* [9] and similar microsatellite mutation rates have been reported for the non-recombining human Y chromosome and the autosomes [10-12], though interpretation of the latter result is problematic because the Y chromosome undergoes intramolecular recombination [13]. These findings gave rise to the theory that the predominant mechanism of microsatellite mutation is strand slippage during DNA replication, which involves misaligment of repetitive sequences between newly replicated and template DNA strands [14-16] (Figure 5.1 part B). The theory is currently quite well entrenched, judging by the fact that evidence of increased microsatellite divergence between humans and chimpanzees on the Y chromosome compared with the autosomes was interpreted as a putative mutagenic effect of heterozygosity, rather than of recombination [17].

One interesting study does, however, suggest that a possible role of recombination hotspots in driving microsatellite evolution has not been given sufficient attention. A poly-AC tract inserted in the *ARG4* hotspot both influenced recombination and showed a high rate of mutation [18]. There are no reported tests of the generality of these findings, presumably because of the evidence discussed above suggesting replication slippage to be the principle, or sole, mechanism of microsatellite mutation. However, a possible mutagenic effect of hotspot regions need not require recombination to take place. For example, if there are sequences in gamma hotspots that cause replication pausing, as has been suggested [19] (see Section 1.2.2), they could also mediate microsatellite mutability, since pausing at a replication fork would create more time for newly replicated and template strands to misalign resulting in slippage mutations, and replication fork stalling has been shown to cause microsatellite mutations [20, 21]. Moreover, as mentioned above, strand slippage mutations of microsatellites could occur in recombination intermediate structures [7] (Figure 5.1 part C), and the generality of this phenomenon has not been investigated.

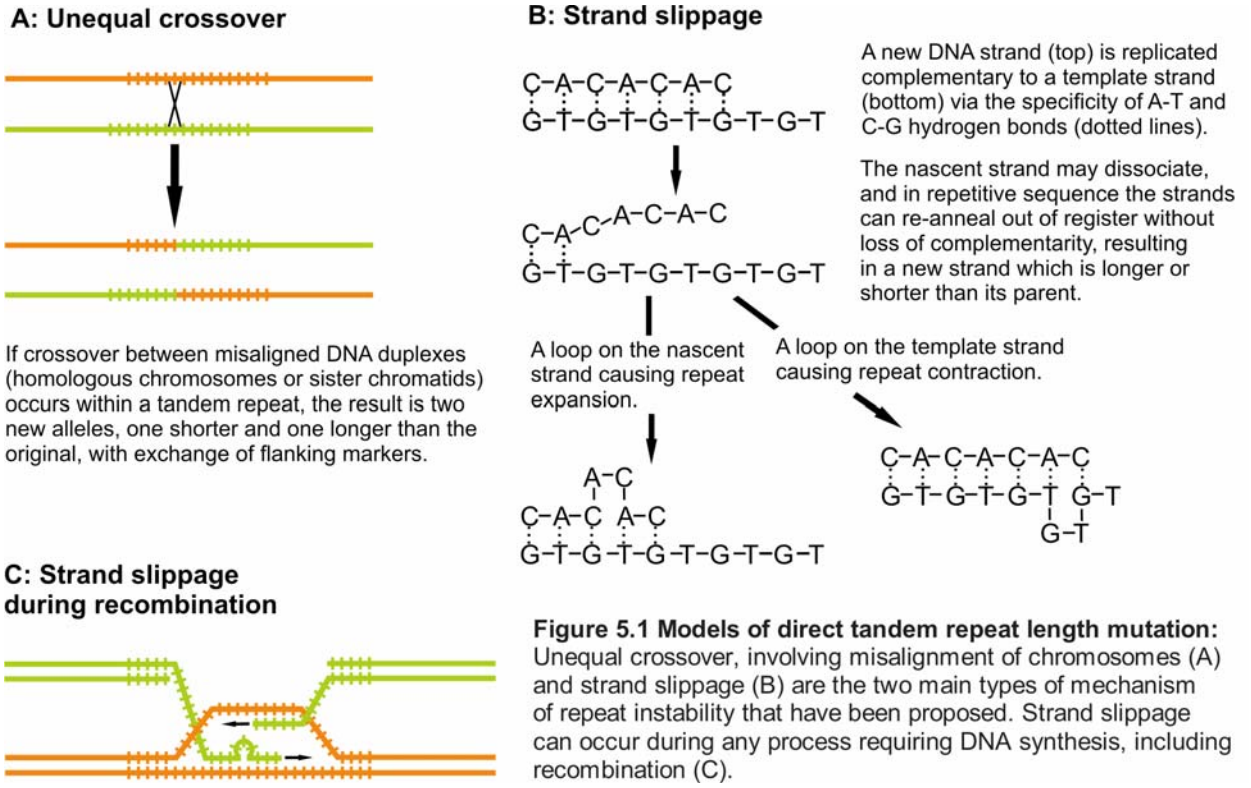

A: Unequal crossover
If crossover between misaligned DNA duplexes (homologous chromosomes or sister chromatids) occurs within a tandem repeat, the result is two new alleles, one shorter and one longer than the original, with exchange of flanking markers.
B: Strand slippage
C-A-C-A-C-A-C
G-T-G-T-G-T-G-T-G-T
A new DNA strand (top) is replicated complementary to a template strand (bottom) via the specificity of A-T and C-G hydrogen bonds (dotted lines).
C-A-C-A-C-A-C
G-T-G-T-G-T-G-T-G-T
The nascent strand may dissociate, and in repetitive sequence the strands can re-anneal out of register without loss of complementarity, resulting in a new strand which is longer or shorter than its parent.
A loop on the nascent strand causing repeat expansion.
A loop on the template strand causing repeat contraction.
A-C
C-A-C A-C
G-T-G-T-G-T-G-T-G-T
C-A-C-A-C-A-C
G-T-G-T-G-T G-T
G-T
C: Strand slippage during recombination

Figure 5.1 Models of direct tandem repeat length mutation: Unequal crossover, involving misalignment of chromosomes (A) and strand slippage (B) are the two main types of mechanism of repeat instability that have been proposed. Strand slippage can occur during any process requiring DNA synthesis, including recombination (C).

If recombination hotspots do drive microsatellite evolution by mutation, they should harbour elevated rates of microsatellite polymorphism. Inferring a mutagenic effect from such an association is not straightforward, however, because recombination is expected to maintain neutral polymorphism by interrupting the effects of selection on linked mutations. This should occur both for advantageous mutations (hitch-hiking), and also for deleterious mutations (background selection). In the absence of recombination, all neutral sequence variants on the same chromosome as a positively selected mutation would tend to increase at the expense of a population's neutral variants on that chromosome. Selection would also act to drive toward extinction all polymorphisms on the same chromosome as a disadvantageous mutation. Recombination operates to limit the scale of these effects to portions of chromosomes by transferring mutations acted on by selection onto different genetic backgrounds. The size of these portions is dependent on the location and frequency of recombination crossover breakpoints on a chromosome, so that chromosome segments that frequently recombine are expected to harbour higher levels of polymorphism than non-recombining regions.

In accordance with this principle, correlations between general genetic diversity and broad scale recombination rate have been seen in humans [22, 23]. Such a correlation has not been found for human microsatellite polymorphism [24, 25], but a correlation between recombination rate and microsatellite polymorphism was found in *Drosophila melanogaster* [26]. The fact that microsatellite mutation rates are several orders of magnituude lower in *Drosophila* [27] might explain this discrepancy, because high mutation rates are expected to reduce the apparent effect of long-range hitchhiking in infrequently recombining regions [28, 29]. This suggests that an activity of recombination acting to interrupt the effects of hitchhiking and selective sweeps on microsatellite polymorphism levels may not actually be detectable in humans. However, it is almost certain that these previous studies have never adequately tested the effect of recombination hotspots, because a random sample of loci, such as the studies selected [24, 25], is unlikely to pick out a substantial proportion of microsatellites in recombination hotspots, given the relative rarity of hotspots in general and the fact that human hotspots do not have excessively high microsatellite frequencies (see Section 3.3).

5.1.2 Collaborative work

I initially solicited the assitance of two colleagues, Dr Mikael Brandström and Professor Hans Ellegren from the University of Uppsala, Sweden, to investigate the relationship between microsatellite polymorphism and recombination hotspots in the human genome (Brandström, Bagshaw, Gemmell and Ellegren, unpublished). The first dataset analysed, by Dr Brandström, was the ALFRED database [30], which contains microsatellite allele frequency information from population surveys. Polymorphism data were extracted from 282 microsatellite loci spread across the human genome. Dr Brandström found that four measures of microsatellite polymorphism: allele span, number of alleles and heterozygosity, do not differ significantly from random expectation at hotspot-associated loci, though the ALFRED microsatellites are slightly but significantly more common in hotspots than expected by chance. This result is apparently inconsistent with a substantial mutagenic effect of recombination hotspots on microsatellites, but its interpretation is not straightforward because markers contained within allele frequency databases are likely initially to have been selected on the basis of known high heterozygosity in order to increase their potential to provide information about genetic divergence [31]. This could give rise to an ascertainment bias, which might mask an effect of recombination on degree of microsatellite polymorphism. To overcome this problem, Dr Brandström investigated a set of about 400,000 insertion/deletion (indel) polymorphisms within tandem repeats identified from shotgun sequencing initiatives [32]. This dataset should be free from ascertainment bias, but it does not allow estimation of the degree of microsatellite polymorphism, because loci can only be scored as polymorphic or not polymorphic. Dr. Brandström's microsatellite polymorphism statistic, which was derived, for each studied region, by dividing the number of indels overlapping with microsatellites by the total number of microsatellites, was found to be 14% higher, on average, in recombination hotspots, which is a statistically significant over-representation (Brandström, Bagshaw, Gemmell and Ellegren, unpublished).

5.1.3 Questions addressed in this chapter

The results from the collaborative work described above suggested an effect of recombination hotspots on microsatellite polymorphism, but they left two main questions unanswered, and I address these in this chapter. First, dividing the number of indels overlapping with microsatellites by the total number of microsatellites does not always give the number of length-polymorphic microsatellites, because there are a substantial number of loci with multiple indels mapped within them. This could bias the results because length changes in microsatellites are usually insertions or deletions of units of the consensus repeated motif [15] and in an uninterrupted (perfect) repeat array, only one of these can possibly be mapped per locus. Multiple indels in a microsatellite can therefore only be detected if there are interruptions in the repeat array, and the theoretical likelihood of detecting them increases with the degree of imperfection in the microsatellite. Imperfection in microsatellites varies both in terms of number and type of interruptions, so this bias would be difficult to control for. Dr Brandstrom's result might reflect a relatively small number of hotspot associated miicrosatellites harbouring multiple indels, perhaps due to extreme length and repeat imperfection. While this remains to be tested, an alternative approach is to ask whether there is an elevation in recombination hotspots of numbers of polymorphic microsatellites relative to total microsatellites. This is clearly expected if recombination or some property of its hotspots does mutate microsatellites with substantial frequency. I therefore scored microsatellites with at least one indel mapping within them as polymorphic, and compared the average magnitude in hotspots and coldspots of the statistic derived by dividing the number of polymorphic microsatellites thus defined by the total number of microsatellites in each tested region. The second question I address here is what is the scale of the correlation between microsatellite polymorphism and recombination? If recombination is mutagenic to microsatellites, the strongest correlation should presumably be seen at the fine scale of hotspots. Such a correlation might still reflect an effect of recombination acting to maintain polymorphism by interrupting the effects of selection, but this should also be evident at larger scales [33]. The third main question I address here is whether there are other factors, associated with both recombination and polymorphism, underlying an apparent link between microsatellite polymorphism and recombination hotspots.

5.1.4 Methodological rationale

My second and third questions were recently addressed, not for microsatellites but for general genetic diversity, using wavelet analysis to assess correlations over multiple genomic scales, in conjunction with multiple regression to control for various known effectors of polymorphism [33] (see Sections 3.1 and 3.3.2). In this chapter I have used wavelet analysis to measure scale-specific correlations between microsatellite polymorphism and recombination rate, with the idea that a broad scale correlation not present at fine scales would indicate effects of selection rather than a mutation bias. I also used a generalized linear model to investigate the influence on the correlation of other factors expected to correlate with both polymorphic microsatellites and recombination, including GC-content, single nucleotide polymorphism (SNP) density, gene (exon) coverage, density of indel polymorphisms occurring outside microsatellites, and density of monomorphic microsatellites. The rationale for the use of these variables was as follows. GC-content has previously been shown to correlate with both genetic diversity and recombination rate [33], so could mediate a link between polymorphic microsatellites and recombination rate. SNP density is likely to reflect influences on polymorphism in general, such as regional selective constraint, and also methodological artefacts of polymorphism detection. It might correlate with recombination rate as well, for at least three reasons. Firstly, frequently recombining regions should harbour more polymorphic loci due to recombination interrupting the effects of selection. Secondly, recombination might cause an elevated rate of single nucleotide mutations (see Section 1.3). Thirdly, the recombination map I used was derived from haplotype inference based on SNPs, potentially causing some degree of coincidental association between mapped recombination rate and SNP density [34]. I controlled for exon density because it is likely to affect polymorphism due to selective constraint, and indel density, which I defined as all indels from the dataset I utilized [32] not found within microsatellites, in order to assess the influence of factors specifically influencing indels. I corrected for the frequency, in each studied region, of monomorphic microsatellites to provide a control for influences on microsatellite distribution in general, independently of the other control variables.

In a separate analysis, I compared microsatellite polymorphic fraction, i.e. the proportion of total microsatellites containing at least one indel, between hotspots and coldspots, including hotspot central and flanking regions. Direct comparison of hotspots and coldspots does not allow control of the influence of other variables, but the coldspot dataset I used had originally been selected to have similar SNP density to the studied hotspots [34], and the central regions of the hotspots from this dataset do not have elevated GC-content (see Section 4.3.8). Direct comparison is unlikely to be affected by low statistical power to detect correlations between sparsely distributed variables and it is relevant in several respects to the question of causality. Firstly, an association between microsatellite polymorphism and recombination with a magnitude equal or greater in hotspot flanking regions than in hotspots would obviously be much more likely to reflect an effect of selection than a mutation bias. On the other hand, an enrichment of microsatellite polymorphism in hotspot central regions, such as I observed for total microsatellites (see Section 3.2), would suggest the existence of a mutagenic effect, since crossovers and gene conversion tracts are most concentrated near hotspot mid points [35, 36]. Finally, an elevation of the level of microsatellite polymorphism in hotspots not reflected in a fine scale correlation between recombination rate and polymorphic microsatellites would presumably reflect an effect of hotspots not directly related to recombination, such as replication pausing, assuming this doesn’t always lead to recombination.

5.2 Methods

I used the same genome-wide human fine-scale recombination map and microsatellite locations detailed in Chapter 3, namely the haplotype inference map and hotspot locations determined to within 5 kb, and coldspots of equivalent size and number, as reported by Myers *et al*., [34], and microsatellites predicted by the TRF algorithm with default search parameters [37] (see Section 3.2.1). I extracted sequence and annotation data and prepared them for generalized linear model and wavelet analyses by binning into 1 kb windows as described in Chapter 3 (Sections 3.2.1 and 3.2.2). Polymorphism datasets I used in this chapter were as follows. Indels were as reported by Mills and colleagues [32], who mapped them by comparing sequences from 36 diverse humans, generated by shotgun re-sequencing initiatives [38, 39]. The indel map they produced consists of just over 400,000 polymorphisms, the majority of which are short, and deletions of more than ten base pairs comprise less than 5 % of the total. [32]. Here I analysed indels and insertions separately in order to investigate length polymorphism in general in the first instance, and also the secondary hypothesis of a recombination-mediated bias in favour of microsatellite insertions. Insertions had been mapped relative to the chimpanzee genome, but do not comprise approximately half of the indels from the dataset because the majority of the indels could not be mapped to unique positions on the chimpanzee genome [32]. Where indels mapped within microsatellites predicted by TRF using is default parameters [37], they had been labelled as such in the Mills *et al*., dataset. I filtered these polymorphic microsatellites to include only arrays with repeated motif sizes of five base pairs or less. For my control variable SNPs I used a dataset of just over 3.3 million polymorphic loci that had been extracted in parallel with the indel dataset, enabling separation of the two types of polymorphism for analysis, and the SNP dataset is available online [40].

5.3 Results

I initially compared microsatellite polymorphic fraction between hotspots and coldspots. Polymorphic fraction was defined as the number of microsatellites in a region containing at least one indel polymorphism (polymorphic microsatellites), divided by the total number of microsatellites in the region. Averaging this value over all hotspots (n=9298) and all coldspots (n=9283), I found no significant differences, either for repeats with 2-5 bp motifs, or for mononucleotide repeats (p>0.07, Mann-Whitney U Test). Testing loci with insertion polymorphisms relative to the chimpanzee reference sequence separately, I found that the hot/cold ratio of means is slightly lower for these elements than for all indel-polymorphic microsatellites considered together, and the difference between hot and coldspots is not significant (p>0.3, Mann-Whitney U test).

I also found no significant enrichment of polymorphic fraction in hotspot 500 bp central regions compared with hotspot non-central regions, i.e. all areas of hotspots outside the central 500 bp. Plotting the distributions of polymorphic microsatellites and total microsatellites in relation to hotspot central regions shows clearly that no central tendency is present for these elements (Figure 5.2). The previously noted modest enrichment of microsatellites in human hotspot flanking regions (see Section 3.2.3) is, however, reflected in a small increase in polymorphic fraction in these regions, which is significant for repeats with 2-5 bp motifs, but not for mononucleotide repeats (Table 5.1, Figure 5.3).

Table 5.1 Elevated microsatellite polymorphism in hotspot flanking regions
Ratios of polymorphic fraction in hotspots, and hotspot flanking regions 0-1 and 1-2 hotspot widths removed from hotspots (mean hotspot width = 4070 bp), compared with coldspots. In cases where hotspot flanking regions overlapped with coldspots, the coldspots were excluded from the analysis. Statistical comparisons were made with the Mann-Whitney U test (alpha = 0.01). Values for the means, and their standard errors, can be seen in Figure 5.2.

	Mean frequency ratio			P value (comparison with coldspots)		
Microsat. motf length	Hotspots/ coldspots	Hotspot flanks 0-1 widths removed/ coldspots	Hotspot flanks 1-2 widths removed/ coldspots	Hotspots	Hotspot flanks 0-1 widths removed	Hotspot flanks 1-2 widths removed
2-5 bp	1.09	1.14	1.10	n/s	0.002	n/s
1 bp	1.24	1.21	1.12	n/s	n/s	n/s

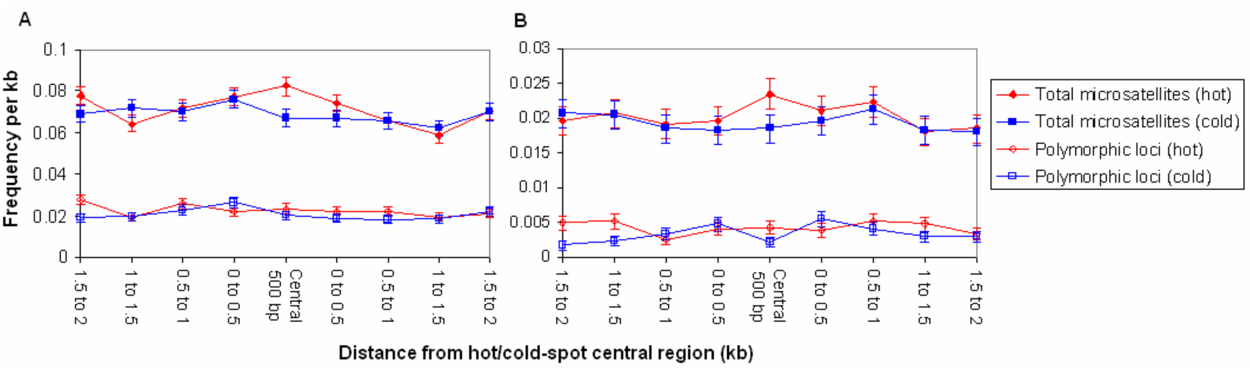


Figure 5.2: Distribution of polymorphic and total microsatellites in relation to human recombination hotspot central regions
Mean per kb frequencies of monomorphic (solid symbols) and polymorphic (empty symbols) microsatellites in relation to hotspot central regions for repeats with 2-5 bp motifs (A) and mononucleotide repeats (B). Error bars are plus and minus one SEM or are not shown in cases where they are narrower than the symbol widths.

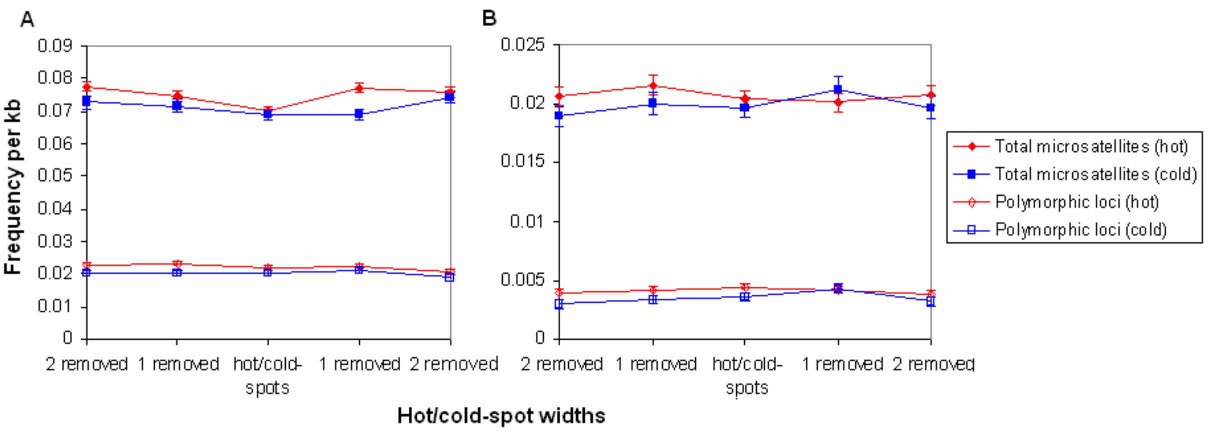


Figure 5.3: Distribution of polymorphic and total microsatellites in relation to human recombination hotspot flanking regions
Mean per kb frequencies of monomorphic (solid symbols) and polymorphic (empty symbols) microsatellites in relation to hotspots and hotspot flanking regions for repeats with 2-5 bp motifs (A) and mononucleotide repeats (B). For the analysis of flanking regions, each hot/cold-spot was extended by one- (denoted “1 removed”) and two-fold (denoted “2 removed”) its own width on either side (mean hotspot width = 4070 bp). In cases where this resulted in overlap between hot and cold areas, the cold ones were excluded from the analysis. Error bars are plus and minus one SEM or are not shown in cases where they are narrower than the symbol widths.

A potential complicating factor in my direct comparison analysis is that the coldspot locations I used were originally selected to have SNP densities close to those of the hotspots analysed [34], and I used a different SNP dataset to that employed in the hotspot analysis by

Myers *et al.* [32]. If polymorphic loci in general from the dataset I utilized were more common in coldspots, a bias would be indicated that could reduce the apparent magnitude of an elevation of microsatellite polymorphism in hotspots, since polymorphic microsatellites correlate with SNP density (see Table 5.2). In fact, this is not the case, because the indels and SNPs from the dataset I used are more common in hotspots than coldspots (Figure 5.4). This should, if anything, increase the apparent association of microsatellite polymorphism with hotspots, assuming that polymorphic microsatellites co-vary with general genetic diversity.

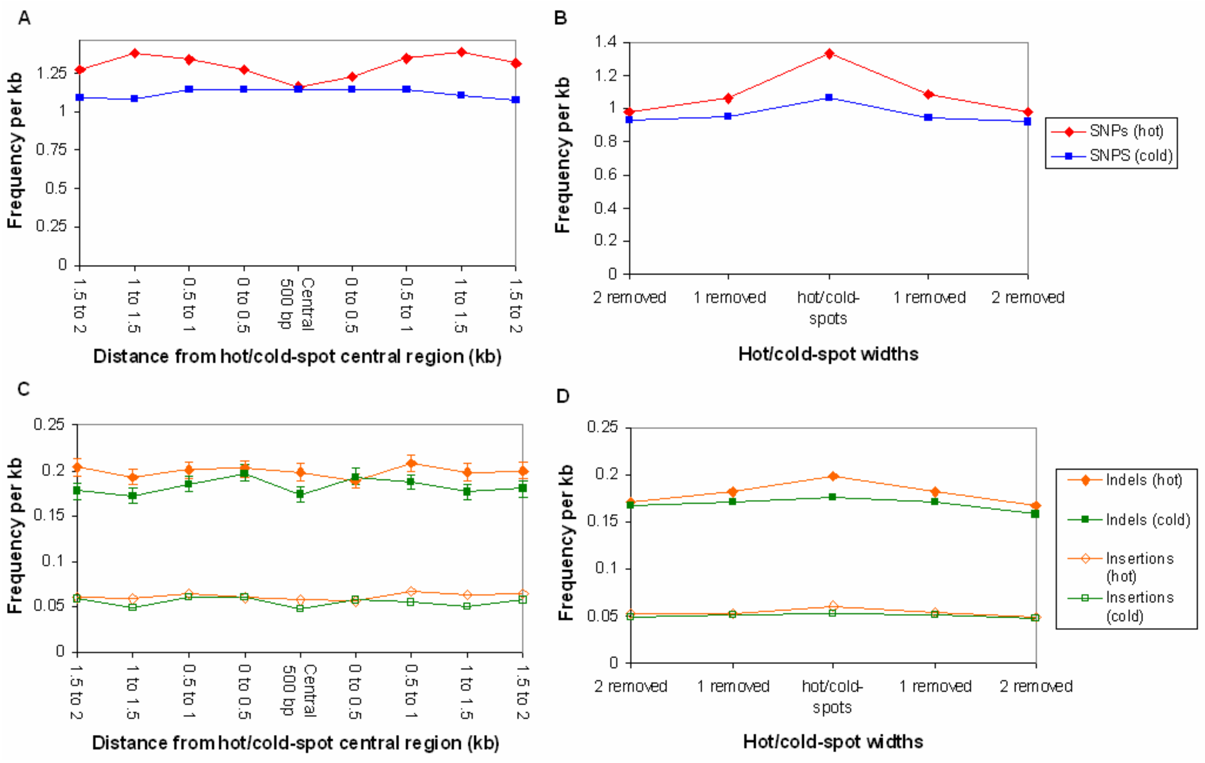


Figure 5.4: Distribution of total polymorphisms from the dataset used [32] in relation to human hotspot central and flanking regions
Mean per kb frequencies of total polymorphisms in relation to hotspot central (A and C) and flanking (B and D) regions for SNPs (A and B) and insertion/deletion (indel) polymorphisms (C and D). For the analysis of flanking regions, each hot/cold-spot was extended by one- (denoted “1 removed”) and two-fold (denoted “2 removed”) its own width on either side (mean hotspot width = 4070 bp). In cases where this resulted in overlap between hot and cold areas, the cold ones were excluded from the analysis. Error bars are plus and minus one SEM or are not shown in cases where they are narrower than the symbol widths.

Based on a direct hotspot/coldspot comparison analysis, therefore, microsatellite polymorphism is not associated with recombination, suggesting that recombination does not drive microsatellite evolution by mutation. In order to test this possibility more rigorously, I investigated the correlation between polymorphic microsatellites and recombination rate in the human genome at a fine scale of one kilo base, correcting for other features expected to influence such a correlation (see above, Section 5.1). Polymorphic microsatellites are rare at this scale (Table 5.2), so I used a generalized linear model with a poisson error distribution, and no restriction on dispersion (quasipoisson family in R, link = log). I repeated the analysis for 37 separate 2^{15} kb, (32.8 mega bases) regions of the human genome (see Section 3.2.2). These regions had been selected in view of the requirement of wavelet analysis for contiguous data across all studied windows for all variables. My use of them for the generalized linear models therefore provided consistency with my wavelet analysis, avoided areas of the genome poorly annotated for the variables in question, and enabled evaluation of regional effects. The results clearly show that SNP density is the strongest and most consistent predictor of polymorphic microsatellites and that recombination is a very weak and inconsistent predictor (see Table 5.3). Only one out of 37 regions shows significant prediction of polymorphic microsatellites by recombination when using a Bonferroni-adjusted alpha level of 0.000225, and there is no consistency in the direction of correlation across regions. The contribution made to the model by recombination is actually negative, though non-significant, for 6 of the 37 regions when considering microsatellites with 2-5 bp repeated motifs containing indel polymorphisms, and 8 of 37 regions when only considering 2-5 bp motif microsatellites containing insertion polymorophisms. These numbers are even higher for mononucleotide repeats: 20 of 37 for indel polymorphic tracts and 24 of 37 for mononucleotide repeats harbouring insertion polymorphisms.

Table 5.2 Abundance and distribution of polymorphic microsatellites analysed in the generalized linear models

Statistics relating to the distribution and abundance of polymorphic microsatellites from the dataset I used for wavelet and generalized linear model analyses are shown here. Deletion or insertion biases should not be inferred from these data because insertions do not comprise half of total indels due to the fact that not all indels could be mapped to unique positions on the chimpanzee genome (see Section 5.2).

Microsat. Motif length	Polymorphism type	Mean per kb frequency	Variance	Sum
1 bp	Indels	0.0035	0.0035	4227
	Insertions	0.0015	0.0015	1827
2 to 5 bp	Indels	0.0189	0.0197	22899
	Insertions	0.0068	0.0070	8193

Table 5.3: Predicting polymorphic microsatellites at a scale of one kilo base
Results from generalized linear model analyses (poisson error distribution with no restriction on dispersion) predicting polymorphic microsatellites at a scale of 1 kb in 37 2^{15} kb regions of the human genome. The control variable indels consisted of all indels located outside microsatellites and the control variable denoted “microsats” was calculated for each 1 kb region by subtracting the number of polymorphic microsatellites from the total number of microsatellites. The rightmost column shows numbers of regions with a significant positive (pos) or negative (neg) effect of the respective predictor (by Student’s T test; Bonferroni-adjusted alpha=0.000225). Overall significance was calculated by Stouffer’s method [41] in cases where the direction of correlation was consistent across all regions, and “inc” indicates that some regions showed negative effects and others positive effects.

Polymor-phism type	Repeat motif length	Predictor	Estimated Coeff.		T		Pr(T>|t|)	# sig pos(neg)
			Mean	Mean SEM	Mean	SEM		
Indel	1 bp	Recombination	-0.0103	0.0346	0.0851	0.209	inc	1(0)
		Exons	-0.0013	0.0016	-0.385	0.162	inc	0(0)
		GC-content	3.25	1.42	2.33	0.238	inc	7(0)
		SNPs	0.175	0.0340	5.13	0.287	$<10^{-200}$	29(0)
		Microsats	-0.0230	0.703	0.093	0.131	inc	0(0)
		Indels	0.0934	0.198	0.771	0.185	inc	0(0)
Insertion	1 bp	Recombination	-0.0293	0.0621	-0.109	0.151	inc	0(0)
		Exons	-0.02	0.0772	-0.175	0.121	inc	0(0)
		GC-content	3.28	2.19	1.54	0.158	inc	3(0)
		SNPs	0.174	0.0523	3.44	0.239	$<10^{-84}$	19(0)
		Microsats	0.559	0.800	0.872	0.203	inc	2(0)
		Indels	0.146	0.275	0.755	0.186	inc	1(0)
Indel	2 to 5 bp	Recombination	0.0113	0.0125	1.01	0.186	inc	1(0)
		Exons	-0.0010	0.0006	-1.54	0.158	inc	0(0)
		GC-content	-1.14	0.676	-1.73	0.285	inc	0(3)
		Microsats	0.432	0.122	3.73	0.234	inc	18(0)
		SNPs	0.178	0.0138	12.9	0.670	$<10^{-300}$	35(0)
		Indels	0.313	0.0561	5.74	0.332	$<10^{-245}$	31(0)
Insertion	2 to 5 bp	Recombination	0.0129	0.0207	0.816	0.206	inc	0(0)
		Exons	-0.0008	0.0009	-0.599	0.176	inc	0(0)
		GC-content	-1.37	1.14	-1.22	0.182	inc	0(0)
		Microsats	0.508	0.194	2.86	0.245	inc	1(0)
		SNPs	0.185	0.0223	8.50	0.462	$<10^{-300}$	14(0)
		Indels	0.309	0.0956	3.56	0.278	$<10^{-89}$	1(0)

Having found no substantial fine scale correlation between microsatellite polymorphism and recombination, I tested the correlation at broader scales using wavelet analysis (see Section 3.2.2). This analysis showed no clear correlation between microsatellite polymorphic fraction (as defined above) and recombination rate at any scale (Figures 5.5 and 5.6). A significant positive association at scales of 256 kb to one mega base is evident in a small number of regions, but no inference can be made from this result because 370 scale/factor correlations were tested in total so three or four significant associations would be expected by chance given an alpha level of 0.01.

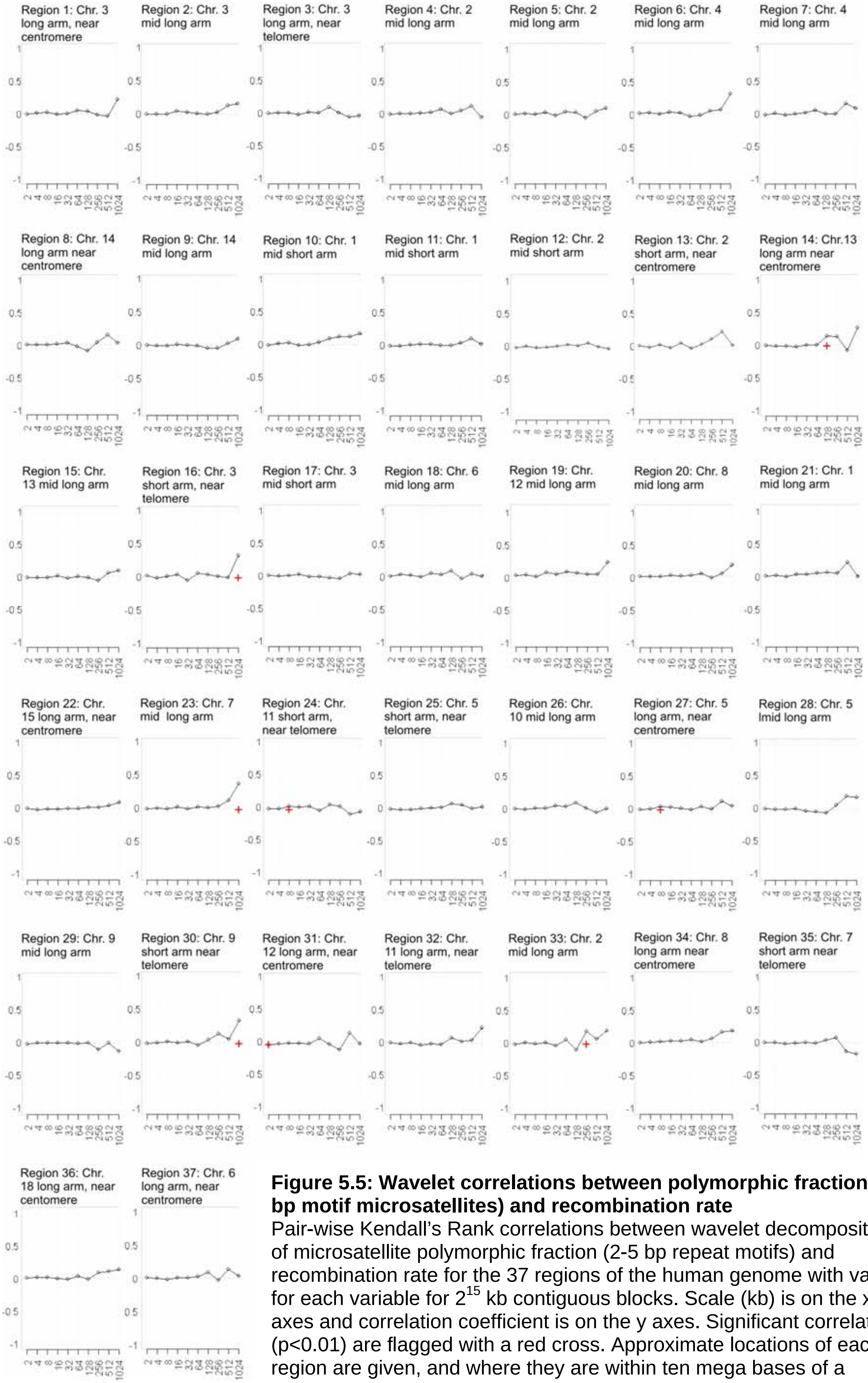


Figure 5.5: Wavelet correlations between polymorphic fraction (2-5 bp motif microsatellites) and recombination rate

Pair-wise Kendall's Rank correlations between wavelet decompositions of microsatellite polymorphic fraction (2-5 bp repeat motifs) and recombination rate for the 37 regions of the human genome with values for each variable for 2^{15} kb contiguous blocks. Scale (kb) is on the x axes and correlation coefficient is on the y axes. Significant correlations (p<0.01) are flagged with a red cross. Approximate locations of each region are given, and where they are within ten mega bases of a centromere or telomere they are labeled as near to that feature.

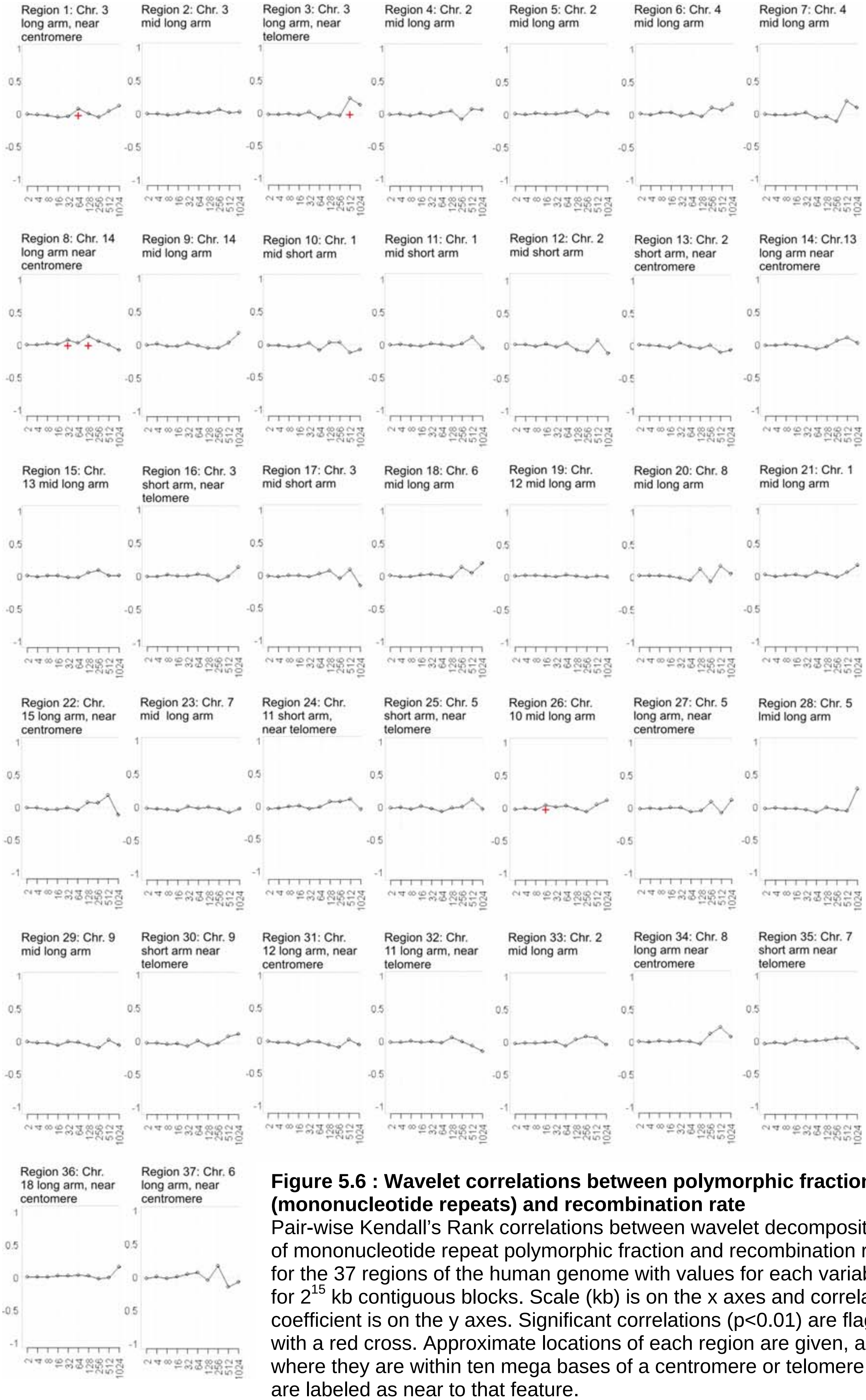


Figure 5.6 : Wavelet correlations between polymorphic fraction (mononucleotide repeats) and recombination rate

Pair-wise Kendall's Rank correlations between wavelet decompositions of mononucleotide repeat polymorphic fraction and recombination rate for the 37 regions of the human genome with values for each variable for 2^{15} kb contiguous blocks. Scale (kb) is on the x axes and correlation coefficient is on the y axes. Significant correlations (p<0.01) are flagged with a red cross. Approximate locations of each region are given, and where they are within ten mega bases of a centromere or telomere they are labeled as near to that feature.

5.4 Discussion

In my collaborative work (see Section 5.1.2), indels occurring in microsatellites were found to be significantly enriched in meiotic recombination hotspots compared with randomly selected regions of the genome. This could indicate a recombination-mediated mutation bias, but the data presented in this chapter, combined with several theoretical arguments, suggest that it can be explained by recombination acting to interrupt the effects of selection on neutral linked polymorphism rather than by a mutagenic effect. I found no significant difference between hotspots and coldspots in the fraction of microsatellites containing at least one indel, and no significant prediction of polymorphic microsatellites by recombination rate when correcting for potential mediating factors, except in one out of 37 32.8 mega base regions of the human genome. This result is clearly not be expected if recombination is commonly mutagenic to microsatellites, since a mutation bias should presumably manifest itself to some extent in most or all of the 37 regions, which are each more than 32 mega bases in size, providing considerable statistical power at the 1 kb level. Because a substantial number of regions show non-significant negative prediction of microsatellite polymorphism by recombination, any effect of recombination must be region-specific, and this is more suggestive of selection than mutation, since it is reasonable to assume that not all regions of the human genome have been subject to selection to an equal extent. Recombination acting on polymorphism through selection should be seen at scales larger than hotspots, but wavelet analysis showed no correlations between microsatellite polymorphic fraction and recombination rate at any scale. This analysis might have quite low power to detect correlations for sparsely distributed variables such as microsatellites (see Section 3.4), but hotspot flanking regions do show a modest but significant increase in microsatellite polymorphic fraction, suggesting a slight distal effect of recombination on microsatellite polymorphism. This is supported by the fact that the enrichment is not extended to hotspots themselves, and there is no elevation of microsatellite polymorphic fraction in hotspot central regions, which might be expected if recombination mutates microsatellites, since crossovers and gene conversions are markedly more frequent toward hotspot mid points [35, 36]. Moreover, a central tendency in the distribution of total microsatellites is present in human recombination hotspots (see Section 3.3.2).

Considered together, the results from my generalized linear model and direct comparison analyses therefore indicate that neither recombination, nor any property of recombination hotspots, commonly has an effect on microsatellite mutability in the human genome. Several arguments might suggest that this should not be considered conclusive evidence against the existence of a recombination-mediated mutation bias on microsatellites, but none of these is particularly strong. Firstly, my use of total microsatellites as detected by the TRF algorithm [37], could provide an incomplete picture. Possibly, some types of microsatellites are more mutable by recombination, or some other feature of its hotspots, than others. This seems unlikely to obscure completely a mutagenic effect of recombination given the current models of microsatellite mutation [15]. Replication slippage or recombination errors resulting from strand or chromosome misalignment should theoretically be increased in tandem repeats regardless of their motif. While some motifs or motif sizes are no doubt more mutable than others, there is no reason to think that any particular class of microsatellites should have reduced mutability in recombination hotspots compared with cold regions.

A second potential argument against the conclusion of no mutation bias is that the association between microsatellites and recombination in humans is weak, so could perhaps be driven by a mutation bias so weak as to be undetectable by the methods I have employed here. This argument must be considered in view of the fact that, while total microsatellites are more common in hotspot central and flanking regions (see Section 3.3), the association is present in hotspot flanking but not central regions for polymorphic microsatellites. The observation that there is no difference between hotspots and coldspots in microsatellite polymorphic fraction is unlikely to be due to insufficient statistical power, since the analysis included over 9000 hotspots and coldspots and 1903 polymorphic microsatellites. The generalized linear model analysis should also have high statistical power because it included over 1.2 million 1 kb regions and 27,126 polymorphic microsatellites. One possible weakness of this analysis is that I did not control for sequence read depth, which could give rise to some regional variation in overall polymorphism levels due to the fact that high GC-content regions of genomes are more difficult to clone and sequence [33]. However, I accounted for this possibility by correcting for SNP density, effectively controlling for artefacts introduced during the original process of polymorphism detection, since SNPs are ubiquitously many-fold more common than polymorphic microsatellites. I controlled for any effect specific to indels by using indels occurring outside microsatellites as an additional control variable. Furthermore, I also controlled for GC-content, which showed very weak and inconsistent effects in all cases.

A fourth argument that a mutation bias could drive the association between microsatellites and recombination despite the results presented in this chapter is that the polymorphic nature of the recombination landscape in primates [42-46] might obscure an association between microsatellites and recombination hotspots due to the fact that hotspots do not exist at any particular genomic location for a sufficient time to drive high microsatellite frequencies by mutation. This argument is not compelling because human recombination hotspots are long-lived enough to be associated with general genetic diversity [33], and the hotspots from the dataset I used must have been in similar genomic positions for hundreds of generations in order to produce their observed effects on haplotype patterns [44, 47]. If recombination does drive microsatellite evolution by mutation, this amount of time should therefore have been enough to cause significantly increased levels of polymorphism in hotspot-associated microsatellites among the 36 diverse humans tested in the polymorphism study I utilized, since microsatellites are highly mutable sequences in general, with a mutation rate of around one per thousand generations for a typical human locus [11, 48, 49] (see Section 1.1.2).

Given the evidence that recombination causes GC-biased single nucleotide changes (see Section 1.3), a possible explanation for the fact that microsatellites are strongly associated with recombination hotspots in yeast but not in humans is that AT-rich microsatellites have been interrupted by A/T to G/C mutations and thus are either no longer detectable as microsatellites in the case of poly-AT, or have become PPTs with some GC-content in the case of poly-A (see Sections 3.4 and 8.1). The discrepancy between the two species in the relationship between microsatellites and recombination hotspots may not, therefore, be attributable to the labile nature of the primate recombination landscape. The results presented in this chapter therefore suggest that research into a potential functional role for microsatellites in recombination hotspots is more likely to be fruitful than a further search for a recombination-mediated mutation bias, though direct observation of very large numbers of microsatellite mutations will be necessary to conclusively disprove the mutation bias theory.

References

1. Smith GP: **Evolution of repeated DNA sequences by unequal crossover**. *Science* 1976, **191**(4227):528-535.
2. Pearson CE, Nichol Edamura K, Cleary JD: **Repeat instability: mechanisms of dynamic mutations**. *Nature Reviews Genetics* 2005, **6**(10):729-742.
3. Jeffreys AJ: **Spontaneous and induced minisatellite instability in the human genome**. *Clin Sci (Lond)* 1997, **93**(5):383-390.
4. Jeffreys AJ, Tamaki K, MacLeod A, Monckton DG, Neil DL, Armour JA: **Complex gene conversion events in germline mutation at human minisatellites**. *Nat Genet* 1994, **6**(2):136-145.
5. Richard GF, Paques F: **Mini- and microsatellite expansions: the recombination connection**. *EMBO Rep* 2000, **1**(2):122-126.
6. Kovtun IV, McMurray CT: **Features of trinucleotide repeat instability in vivo**. *Cell Res* 2008, **18**(1):198-213.
7. Jakupciak JP, Wells RD: **Gene conversion (recombination) mediates expansions of CTG[middle dot]CAG repeats**. *J Biol Chem* 2000, **275**(51):40003-40013.
8. Levinson G, Gutman GA: **High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in Escherichia coli K-12**. *Nucleic Acids Res* 1987, **15**(13):5323-5338.
9. Henderson ST, Petes TD: **Instability of simple sequence DNA in Saccharomyces cerevisiae**. *Mol Cell Biol* 1992, **12**(6):2749-2757.
10. Gusmao L, Sanchez-Diz P, Calafell F, Martin P, Alonso CA, Alvarez-Fernandez F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML *et al*: **Mutation rates at Y chromosome specific microsatellites**. *Hum Mutat* 2005, **26**(6):520-528.
11. Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, Kruger C, Krawczak M, Nagy M, Dobosz T *et al*: **Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs**. *Am J Hum Genet* 2000, **66**(5):1580-1588.
12. Nebel A, Filon D, Hohoff C, Faerman M, Brinkmann B, Oppenheim A: **Haplogroup-specific deviation from the stepwise mutation model at the microsatellite loci DYS388 and DYS392**. *Eur J Hum Genet* 2001, **9**(1):22-26.
13. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC: **Abundant gene conversion between arms of palindromes in human and ape Y chromosomes**. *Nature* 2003, **423**(6942):873-876.
14. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution**. *Mol Biol Evol* 1987, **4**(3):203-221.
15. Buschiazzo E, Gemmell NJ: **The rise, fall and renaissance of microsatellites in eukaryotic genomes**. *Bioessays* 2006, **28**(10):1040-1050.
16. Ellegren H: **Microsatellites: simple sequences with complex evolution**. *Nat Rev Genet* 2004, **5**(6):435-445.
17. Kayser M, Vowles EJ, Kappei D, Amos W: **Microsatellite length differences between humans and chimpanzees at autosomal Loci are not found at equivalent haploid Y chromosomal Loci**. *Genetics* 2006, **173**(4):2179-2186.
18. Gendrel CG, Boulet A, Dutreix M: **(CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis**. *Genes Dev* 2000, **14**(10):1261-1268.
19. Petes TD: **Meiotic recombination hot spots and cold spots**. *Nat Rev Genet* 2001, **2**(5):360-369.

20. Fouche N, Ozgur S, Roy D, Griffith JD: **Replication fork regression in repetitive DNAs**. *Nucleic Acids Res* 2006, **34**(20):6044-6050.
21. Hile SE, Eckert KA: **Positive correlation between DNA polymerase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences.** *J Mol Biol* 2004, **335**(3):745-759.
22. Payseur BA, Nachman MW: **Gene density and human nucleotide polymorphism**. *Mol Biol Evol* 2002, **19**:336-349.
23. Nachman MW: **Single nucleotide polymorphisms and recombination rate in humans**. *Trends Genet* 2001, **17**:481-485.
24. Huang QY, Xu FH, Shen H, Deng HY, Liu YJ, Liu YZ, Li JL, Recker RR, Deng HW: **Mutation patterns at dinucleotide microsatellite loci in humans**. *Am J Hum Genet* 2002, **70**(3):625-634.
25. Payseur BA, Nachman MW: **Microsatellite variation and recombination rate in the human genome**. *Genetics* 2000, **156**(3):1285-1298.
26. Schug MD, Hutter CM, Noor MA, Aquadro CF: **Mutation and evolution of microsatellites in Drosophila melanogaster**. *Genetica* 1998, **102-103**(1-6):359-367.
27. Schlotterer C, Ritter R, Harr B, Brem G: **High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates**. *Mol Biol Evol* 1998, **15**(10):1269-1274.
28. Schlotterer C, Wiehe T: **Microsatellites, a neutral marker to infer selective sweeps**. In: *Microsatellites Evolution and Applications.* Edited by Goldstein D, Schlotterer C. Oxford: Oxford University Press; 1999: 238-248.
29. Slatkin M: **Hitchhiking and associative overdominance at a microsatellite locus**. *Mol Biol Evol* 1995, **12**(3):473-480.
30. Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP *et al*: **ALFRED: the ALelle FREquency Database. Update**. *Nucleic Acids Res* 2003, **31**(1):270-271.
31. Selkoe KA, Toonen RJ: **Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers**. *Ecol Lett* 2006, **9**(5):615-629.
32. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE: **An initial map of insertion and deletion (INDEL) variation in the human genome**. *Genome Res* 2006, **16**(9):1182-1190.
33. Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G: **The influence of recombination on human genetic diversity**. *PLoS Genet* 2006, **2**(9):e148.
34. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**(5746):321-324.
35. Baudat F, de Massy B: **Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot**. *PLoS Genet* 2007, **3**(6):e100.
36. Kauppi L, Jeffreys AJ, Keeney S: **Where the crossovers are: recombination distributions in mammals**. *Nat Rev Genet* 2004, **5**(6):413-424.
37. Benson G: **Tandem repeats finder a program to analyze DNA sequences**. *Nucleic Acids Res* 1999, **27**(2):573-580.
38. Consortium TIH: **The International HapMap Project**. *Nature* 2003, **426**:789-796.
39. Group TISMW: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms**. *Nature* 2001, **409**:928-933.
40. http://naetet.net/millssnps.txt.gz.

41. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RMJ: **The American Soldier, Vol 1: Adjustment during army life.** Princeton: Princeton University Press; 1949.

42. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M: **High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans**. *Science* 2008, **319**(5868):1395-1398.

43. Neumann R, Jeffreys AJ: **Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation**. *Hum Mol Genet* 2006, **15**(9):1401-1411.

44. Jeffreys AJ, Neumann R, Panayi M, Myers S, Donnelly P: **Human recombination hot spots hidden in regions of strong marker association**. *Nat Genet* 2005, **37**(6):601-606.

45. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**(4):429-434.

46. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P *et al*: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**(5718):107-111.

47. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P: **The fine-scale structure of recombination rate variation in the human genome**. *Science* 2004, **304**(5670):581-584.

48. Xu X, Peng M, Fang Z: **The direction of microsatellite mutations is dependent upon allele length**. *Nat Genet* 2000, **24**(4):396-399.

49. Ellegren H: **Heterogeneous mutation processes in human microsatellite DNA sequences**. *Nat Genet* 2000, **24**(4):400-402.

Chapter 6

Is the association between poly-purine/poly-pyrimidine tracts and recombination driven by a mutation bias?

Abstract

If recombination is mutagenic to PPTs, it should drive up the frequency of PPT polymorphism in recombination hotspots. I investigated the association between recombination and PPT polymorphism using methods similar to those described in Chapter 5. I mapped three different kinds of polymorphism: indels, insertions and SNPs, to each of three kinds of PPT: tracts of at least 12 bp, high GC-content tracts of at least 12 bp (defined as more GC-rich than the average PPT) and high GC-content tracts of at least 50 bp. Initially, I compared the frequency of these PPT polymorphisms between hotspots and coldspots throughout the human genome, controlling for PPT coverage and total polymorphism density. I found that three of the nine types of PPT polymorphism are significantly enriched in hotspots, namely indels occurring in PPTs of at least 12 bp, SNPs occurring in PPTs of at least 12 bp, and SNPs occurring in high GC-content PPTs of at least 12 bp. These differences can largely be explained by other factors associated with both recombination and PPT sequence polymorphism, however, because generalized linear models incorporating gene density and GC-content in addition to total polymorphism density and PPT coverage as control variables show very weak and inconsistent prediction of PPT polymorphism by recombination rate, which is not statistically significant in the case of PPT indels. In a substantial number of regions of the human genome, recombination predicts PPT polymorphism in a negative direction, indicating that arguments similar to those given in Section 5.4 are applicable, suggesting that the association between PPT polymorphism and

recombination is largely caused by recombination acting to interrupt the effects of selection rather than a general mutation bias.

6.1 Introduction

Having found that PPTs are strongly associated with meiotic recombination hotspots in both humans and yeast (see Section 4.3), it was of interest to determine whether this association could be driven by a mutation bias, because evidence against such a bias would suggest that PPTs are not primarily an effect of recombination. Investigating recombination-mediated mutation biases in the human genome is not straightforward, however, for the reasons already outlined (see Section 5.1.1). In particular, any correlation found could result from recombination acting to preserve neutral polymorphism, by interrupting the purgative effects of selective sweeps and background selection, as well as from mutagenic activity. However, the absence of an association between recombination and PPT polymorphism would suggest that no substantial mutation bias exists.

In this chapter I present a test of the hypothesis that PPT polymorphism is affected by meiotic recombination in the human genome. The methods I employed, and the rationale for their use, were the same as for my investigation of the effect of recombination on microsatellite polymorphism (See Sections 5.1.3 and 5.2), with the following exceptions. I investigated three types of polymorphism in three types of PPT. These were insertions, insertion/deletions (indels) and single nucleotide polymorphisms (SNPs) occurring in PPTs of at least 12 bp, high GC-content PPTs of at least 12 bp (defined as more GC-rich than the average PPT), and high GC-content PPTs of at least 50 bp with one mismatch allowed per 10 bp. This third class was the PPT type most enriched in hotspots from the genome-wide dataset I have utilized in this thesis [1] (see Section 4.3.7) so it was of particular interest to determine the effect of recombination hotspots on its length variability.

The genomic features I studied in the work presented in this chapter were polymorphic sites occurring within PPTs. This contrasted my approach to the question of the relationship between microsatellite polymorphism and recombination (Chapter 5), which was to investigate microsatellites harbouring at least one polymorphism. The reason for the different approach was that frequency of polymorphic sites occurring in a particular type of sequence is clearly more informative of the degree of its variability than the proportion of tracts containing at least one polymorphism. Frequency of polymorphic sites could not be used for microsatellites due to a potential bias arising from their very low complexity (see Section 5.1.2). This bias also applies to microsatellite PPTs, but, as I have reported elsewhere, the proportion of total PPTs consisting of short tandem repeats is very low [2].

As in Chapter 5, I used direct comparison of hotspots with coldspots, additional comparisons for hotspot central and flanking regions, and generalized linear models to investigate the correlation between recombination rate and PPT polymorphism frequency at a scale of 1 kb while accounting for other possible effectors of polymorphism. I also used wavelet analysis to test the possibility of a scale-specific broad scale correlation between PPT polymorphism and recombination, since this would likely be caused by recombination acting to interrupt the effects of selection on genetic diversity in general (see Section 5.1.1).

6.2 Methods

I used the same genome-wide human fine-scale recombination map as for previous investigations in this thesis, namely the hotspot locations determined to within 5 kb, the coldspots of equivalent size and number, and the fine scale recombination rates reported by Myers *et al*., [1] (see Sections 3.1 and 3.2). Other sequence and sequence annotation data, and data analysis methods, were as described in Section 5.2, with the following exceptions. I mapped the locations of indels, insertions and SNPs within PPTs using the galaxy bioinformatics software available from the UCSC genome browser [3], which was downloaded onto a stand-alone supercomputer. I detected PPT locations throughout the human genome using the same computer algorithm described in Section 4.2 (Joel Pitt 2003, unpublished).

6.3 Results

I first asked whether PPTs are more polymorphic, relative to DNA in their immediate vicinity, in meiotic recombination hotspots or coldspots in the human genome. I excluded from this analysis all regions containing no PPTs, since these could bias the results due to the fact that presence of a PPT is required for a non-zero PPT polymorphism statistic. To obtain a statistic reflecting PPT polymorphism in each region, I divided the fraction of each type of polymorphism (SNPs, indels and insertions) occurring within PPTs by the total number of bases covered by PPTs in the region. In total, I analysed nine PPT/polymorphism combinations (see Section 6.1). Combinations with PPT polymorphism statistics differing significantly between hotspots and coldspots are shown in Table 6.1. I found no significant differences for insertion polymorphisms, and PPT indels are not significantly enriched in hotspots when low GC-content PPTs are excluded, but SNPs are over-represented in hotspot-associated PPTs of at least 12 bp, including high-GC content PPTs considered separately. I saw no additional significant differences when not accounting for total regional polymorphism i.e. using the number of PPT polymorphisms divided by the PPT base coverage as the test statistic for each region.

Table 6.1: Enrichment of PPT polymorphisms in human recombination hotspots
Showing PPT/polymorphism combinations over-represented in recombination hotspots throughout the human genome. The mean statistic over all hotspots (n=9298) divided by the mean statistic for coldspots (n=9283) is shown, and statistical comparisons were made using the Mann-Whitney U test (alpha = 0.0055 with Bonferroni's correction for nine PPT/polymorphism combinations tested).

PPT type	Polymorphism type	Hot/cold ratio of mean polymorphism stat.	P value
12 bp+	indels	1.08	0.00011
12 bp+	SNPs	1.07	$<10^{-6}$
12 bp+, high GC	SNPs	1.12	$<10^{-6}$

When limiting the investigation to high GC-content PPTs of at least 50 bp (one mismatch allowed per 10 bp), which is the class of PPT most enriched in hotspots from the genome-wide dataset (see Section 4.3.7), PPT polymorphisms are more common in coldspots, though the differences are not statistically significant, despite the fact that 485 PPT polymorphisms were analysed in total (Table 6.2). Only 61 of these were insertions, which at first sight suggests the possibility of a deletion bias for this type of element, such as has been seen for very long microsatellites [4]. In fact there is no indication of such a bias, however, because only about half of the indels from the dataset I used could be mapped to unique positions on the chimpanzee genome [5], and only about half of these were insertions, so insertions comprise only just over one quarter of indels from the dataset employed here.

I compared polymorphism statistics for each of the hotspot-associated PPT polymorphism types between 500 bp regions centred on hotspot mid points relative to hotspot non-central regions, and I found no significant differences (p>0.2, Mann-Whitney U test). Investigating the flanking regions of hotspots, as described in Chapters 3, 4 and 5, I found that the hotspot-associated PPT polymorphism types are not significantly enriched in these regions relative to coldspots.

Table 6.2: Hotspot/coldspot comparison of polymorphism levels in high GC-content PPTs of at least 50 bp
Showing the ratio, for each type of polymorphism, of mean polymorphism statistics averaged over 9298 hotspots and 9283 coldspots, considering only high-GC-content PPTs of at least 50 bp. P value was determined by the Mann-Whitney U test. Total numbers of PPT-polymorphisms included in this comparison are also shown

Type of polymorphism	Hot/cold ratio of mean polymorphism stat.	P value	Total number of polymorphic loci
Indels	0.87	0.60	266
Insertions	0.74	0.58	61
SNPs	0.98	0.32	219

To determine whether the enrichment of PPT polymorphism levels in recombination hotspots is reflected in a genome-wide correlation with recombination rate, I used generalized linear models and wavelet analysis. The methodological rationale for these techniques was as for previous analyses in this thesis (see Sections 3.1 and 3.2.2). Initially, I determined that PPT polymorphisms are rare at the scale of 1 kb used in the generalized linear models, and that their distributions at this scale have variances greater than their means (Table 6.3), so I used a poisson error distribution with no restriction on dispersion. I repeated the analysis for the 37 separate regions of 2^{15} kb, (32.8 mega bases) of the human genome analysed previously (see Section 3.2.2). The results from the generalized linear models show that recombination rate is a weak and inconsistent predictor of PPT polymorphism in all cases (Table 6.4). In no region is it a significant positive predictor of PPT indels, though seven regions show significance for PPT SNPs. In all cases, the direction of the regression coefficient for recombination is inconsistent, with substantial numbers of regions showing slight negative effects (Table 6.5). GC-content is also a weak and inconsistent predictor of all PPT polymorphism types (Table 6.4), and also contributes negatively to the models in a substantial number of regions (Table 6.5). This contrasts with the fairly strong and consistent correlation between GC-content and total PPT coverage (see Section 4.3.9).

Table 6.3 Abundance and distribution of polymorphism types predicted in the generalized linear models
I binned all variables into contiguous 1 kb windows covering 37 2^{15} kb regions prior to GLM and wavelet analyses (total number of 1 kb bins=1,212,416). Statistics relating to the distribution and abundance of PPT polymorphism types thus binned are shown here.

PPT type	Polymorphism Type	Mean per kb freq.	Variance	Sum
12 bp +	Indels	0.0428	0.0706	51835
12 bp +	SNPs	0.0236	0.0378	28655
12 bp +, high GC	SNPs	0.0105	0.0158	12750

Table 6.4: Predicting PPT polymorphism at a scale of one kilo base
Results from generalized linear model analyses (quasipoisson family in R, link=log) predicting PPT polymorphism types found to be enriched in recombination hotspots. 37 2^{15} kb regions of the human genome were studied in this analysis. Prior to analysis, all variables were averaged for contiguous one kb windows covering the test regions. The control variable indels consisted of all indels located outside PPTs and the control variable SNPs was all SNPs occurring outside PPTs, except for the analysis of PPT indels, for which total SNPs were used. The rightmost column shows numbers of regions with a significant positive (pos) or negative (neg) effect of the respecitve predictor (by Student's T test; adjusted alpha=0.000225). Overall significance was calculated by Stouffer's method [6] in cases where the direction of correlation was consistent across all 37 regions, and "inc" indicates that some regions showed negative effects and others positive effects.

Poly-morphism type	PPT type	Predictor	Estimated Coeff.		T		Pr(T>|t|)	# sig pos(neg)
			Mean	Mean SEM	Mean	SEM		
Indels	12 bp+	Recombination	0.0043	0.0097	0.621	0.264	inc	0(0)
	all tracts	PPTs	0.0101	0.0003	35.1	0.854	$<10^{-300}$	37(0)
		Exons	-0.0002	0.0003	-0.470	0.203	inc	0(1)
		GC-content	0.730	0.487	1.50	0.408	inc	6(1)
		Indels	0.218	0.0347	6.82	0.545	$<10^{-279}$	32(0)
		SNPs	0.187	0.0096	19.4	0.760	$<10^{-300}$	37(0)
SNPs	12 bp+	Recombination	0.0161	0.0108	1.79	0.262	inc	7(0)
	all tracts	PPTs	0.0108	0.0003	32.8	1.23	$<10^{-300}$	37(0)
		Exons	-0.0002	0.0004	-0.362	0.222	inc	1(0)
		GC-content	0.377	0.608	0.602	0.272	inc	3(1)
		SNPs	0.198	0.0108	18.9	1.22	$<10^{-300}$	37(0)
SNPs	12 bp+	Recombination	0.0198	0.0149	1.75	0.283	inc	7(0)
	high GC	PPTs	0.0120	0.0005	24.7	1.02	$<10^{-300}$	37(0)
		Exons	-0.0003	0.0006	-0.240	0.208	inc	0(1)
		GC-content	1.77	0.855	2.17	0.204	inc	7(0)
		SNPs	0.175	0.0191	13.2	1.13	inc	35(1)

Table 6.5: Numbers of regions with negative prediction of PPT polymorphisms by recombination or GC-content in the generalized linear models.
Among 37 regions tested, this table lists the number showing a negative direction of prediction of PPT polymorphism by recombination or GC-content in generalized linear models for each of the three PPT/polymorphism combinations significantly enriched in human recombination hotspots. Numbers of regions for which the negative prediction is significant (p<0.01 by Student's T test) are bracketed.

Polymorphism type	PPT type	Number of regions with (non-significant) negative predictor	
		Recombination	GC-content
Indels	12 bp+	14	10 (1)
SNPs	12 bp+	6	13 (1)
SNPs	12 bp+, high GC	5	3

Using wavelet analysis as described (see Section 3.2.2) I investigated the correlation between PPT polymorphism and recombination rate at scales between 2 kb and one mega base. I found few significant results, but those present tended to be concentrated at scales larger than 2 kb (Table 6.6).

Table 6.6: Summary of significant pair-wise wavelet correlations between recombination and PPT polymorphism
Pair-wise Kendall's Rank correlations were carried out on wavelet decompositions of PPT/polymorphism combinations significantly enriched in human hotspots for 37 2^{15} kb regions of the human genome as described (see Section 3.2.2). For each 1 kb region, the variable tested was the number of polymorphisms of each respective type overlapping with PPTs, divided by PPT coverage in the region. A total of 1110 factor/scale combinations were tested. The left hand number in each pair is the total number of significant positive correlations (alpha = 0.01) for that factor/scale combination. The right hand number is the total number of significant negative correlations. A total of 40 significant results were found with the alpha level of 0.01, but eleven would be expected by chance at this level.

PPT type	Polymorphism Type	Scale of correlation (kb)									
		2	4	8	16	32	64	128	256	512	1024
12 bp+	Indels	0/1	0/0	0/0	0/1	2/0	0/0	0/0	1/0	4/0	2/0
12 bp+	SNPs	0/0	0/0	1/0	1/0	1/0	1/0	3/0	3/0	3/0	4/0
12 bp +,high GC	SNPs	0/0	1/0	2/0	1/0	2/1	1/0	1/0	1/0	1/0	2/0

6.4 Discussion

The association between recombination and PPT polymorphism I have shown here is weak and inconsistent, so similar arguments to those given in Chapter 5 suggest that it is predominantly caused by recombination interrupting the effects of background selection and hitch-hiking, rather than generating new mutations (see Section 5.4). These arguments are further supported by the fact that the wavelet detail coefficient analysis revealed some sparse positive correlations between recombination rate and PPT polymorphism at scales between 4 and 1024 kb, but none at the fine scale of 2 kb (Table 6.4).

Further suggesting that recombination does not drive a mutation bias to cause the association between recombination hotspots and PPTs, I found no significant association between recombination hotspots and PPT insertions, and no significant prediction of PPT indels by recombination rate when correcting for factors expected to correlate with both. However, 7 of 37 regions do show significant positive prediction of SNPs occurring within PPTs at a fine scale of 1 kb. My results might therefore reflect weakly preferential mutation of a subset of PPTs by recombination. How and why this could occur is unclear. In view of their relatively low complexity, strand misalignment during DNA replication and chromosomal misaligment during recombination causing unequal crossover are possible mechanisms for indel mutability of PPTs (see Section 5.1). It is not clear, however, why recombination should act in these ways on PPTs, given that it apparently does not do so in microsatellites (see Section 5.3), because microsatellites are much less complex, on average, than PPTs and are therefore theoretically more likely to be subject to mutations arising from sequence misalignment. Another factor that could drive the mutability of PPTs is their propensity to form non B-DNA structures (See Section 1.1.3). These structures can cause mutations due to replication pausing and resultant slippage [7]. Replication pausing could also stimulate recombination, as been demonstrated in model systems [8-11], and it has been suggested as a potential factor in hotspot regulation, i.e. the gamma hotspot model [12] (see Section 1.2.2). Replication pausing can also promote base misincorporations [13], so it might explain the increased frequency of SNPs in hotspot-associated PPTs.

In Chapter 4, I showed that PPT density correlates quite strongly with GC-content, suggesting that high GC-content could drive a mutation bias, giving rise to the association between recombination and PPTs. If this is the case there should be a consistent correlation between PPT polymorphism and GC-content, but I found that the predictive ability of GC-content is very inconsistent among PPT polymorphism types, and regional effects are evident including a substantial number of negative associations. The association of high GC-content with PPT polymorphism is certainly weaker and less consistent than its association with PPTs in general (see Table 4.9). This corroborates other evidence that high GC-content does not generally mediate the link between recombination and PPTs (see Section 4.4).

Because I did find some significant associations between PPT polymorphism and recombination, however, the results presented in this chapter cannot rule out the possibility that a mutation bias drives the association between recombination and PPTs. Nevertheless, this possibility seems remote considering the evidence I have presented here, and in Chapter 4. In view of the weakness of the association between recombination and PPT polymorphism I observed, any mutation bias must have produced the enrichment of PPTs in hotspots over a very long evolutionary time scale. Because recombination hotspots are short-lived in the primate lineage [14, 15], this would entail multiple recurrence, in evolutionary time, of hotspots at very similar genomic positions. These regions must in fact be almost identical given that PPTs are most highly enriched in the central regions of hotspots (see Section 4.3.7). If this is the case, it seems highly likely that some sequence element in these regions has the property of potentiating hotspot recombination, and PPTs are a plausible candidate for such a sequence (see Section 4.4).

References

1. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**(5746):321-324.
2. Bagshaw AT, Pitt JP, Gemmell NJ: **Association of poly-purine/poly-pyrimidine sequences with meiotic recombination hot spots**. *BMC Genomics* 2006, **7**:179.
3. http://genome.ucsc.edu/.
4. Ellegren H: **Heterogeneous mutation processes in human microsatellite DNA sequences**. *Nat Genet* 2000, **24**(4):400-402.
5. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE: **An initial map of insertion and deletion (INDEL) variation in the human genome**. *Genome Res* 2006, **16**(9):1182-1190.
6. Stouffer SA, Suchman EA, Devinney LC, Star SA, Williams RMJ: **The American Soldier, Vol 1: Adjustment during army life.** Princeton: Princeton University Press; 1949.

7. Hile SE, Eckert KA: **Positive correlation between DNA polymerase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences.** *J Mol Biol* 2004, **335**(3):745-759.
8. Kuzminov A: **DNA replication meets genetic exchange: chromosomal damage and its repair by homologous recombination**. *Proc Natl Acad Sci U S A* 2001, **98**(15):8461-8468.
9. Baudat F, Keeney S: **Meiotic recombination: Making and breaking go hand in hand**. *Curr Biol* 2001, **11**(2):R45-48.
10. Michel B: **Replication fork arrest and DNA recombination**. *Trends Biochem Sci* 2000, **25**(4):173-178.
11. Michel B, Flores MJ, Viguera E, Grompone G, Seigneur M, Bidnenko V: **Rescue of arrested replication forks by homologous recombination**. *Proc Natl Acad Sci U S A* 2001, **98**(15):8181-8188.
12. Petes TD: **Meiotic recombination hot spots and cold spots**. *Nat Rev Genet* 2001, **2**(5):360-369.
13. Fry M, Loeb LA: **A DNA polymerase alpha pause site is a hot spot for nucleotide misinsertion**. *Proc Natl Acad Sci U S A* 1992, **89**(2):763-767.
14. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**(4):429-434.
15. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P *et al*: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**(5718):107-111.

Chapter 7

Probing the secondary structure of meiotic recombination hotspot sequences with sodium bisulphite

Abstract

In recent years evidence has accumulated that the functions of DNA can be determined by its structure as well as by its sequence. Work in this area has mostly been focussed on the possible role of DNA secondary structure in gene expression, and no reports have yet pointed to the possibility of a general link between DNA structural variation and meiotic recombination hotspots. In Chapter 4 I showed that simple sequences likely to have structure-forming potential are highly over-represented close to recombination sites, suggesting that the functional involvement of secondary structure in meiotic recombination should be investigated further. In this chapter, I present some preliminary results from this investigation. I have probed the non-B-DNA structure-forming potential, in supercoiled plasmids, of sequences amplified from human recombination hotspot central regions. I have also tested the orthologous regions of the chimpanzee genome previously shown not to contain hotspots. I used a sodium bisulphite modification assay, which causes deamination of cytosine residues in single stranded DNA, such as is formed in the non-B-DNA structures intramolecular triplexes, quadruplexes and cruciforms. Out of six hotspots tested, I found that sensitivity to sodium bisulphite is significantly higher in humans than in chimpanzees in three, though only in one, the DNA2 hotspot, is this correlated with a difference in numbers of molecules showing long contiguous strings of converted cytosines, which would be expected in intramolecular quadruplex and triplex structures.

7.1 Introduction

In general, the functionality of DNA structure has been explored very little in comparison with the functionality of DNA sequence, and the biological significance of structural variations is not yet well understood. One reason for this is that they were discovered quite recently. Prior to 1979 it was assumed that DNA *in vivo* uniformly consisted of a right-handed double helix, known as B-DNA, based on A:T and G:C base pairs, as described by Watson and Crick [1]. This paradigm began to change during the decade 1979-1989 with the discovery of five major variations on DNA structure, each of which is preferentially formed by sequences of relatively low complexity in the presence of negative supercoiling in the DNA duplex, i.e. torsional or twisting energy (reviewed in [2]). The first structural variant to be described, using crystallography, was Z-DNA, a left-handed double helix formed by microsatellites with alternating purines and pyrimidines [3]. Immunocytological assays showed that this form of DNA often occurs near active genes in eukaryotes, suggesting functional significance [4]. Soon after the discovery of Z-DNA, studies using the enzyme nuclease S1, which attacks single-stranded DNA, revealed the existence of other structures. The first of these was the cruciform, which consists of intrastrand Watson-Crick base pairs in inverted (self-complementary) repeat sequences [5]. Cruciforms are not as stable under physiological conditions as Z-DNA, except in very AT-rich sequences [6] and their possible functional significance has been explored relatively little [2]. The next non-B-DNA structure to be discovered was the intramolecular triple-helix, or triplex, which is formed by fold-back interactions between three strands, giving rise to a substantial amount of single-stranded DNA and resultant S1 nuclease sensitivity [7, 8]. The fold-back interactions involved in intramolecular triplex formation are mediated by Hoogsteen base pairing, a hydrogen bonding interaction first described over 50 years ago to explain the appearance of three-stranded aggregates in DNA *in vitro* [9]. An alternative structure involving fold-back Hoogsteen interactions is the intramolecular quadruplex, first discovered in 1988, which requires rows of consecutive guanine residues [10]. Evidence has pointed to a role for intramolecular triplexes and quadruplexes in DNA function, including the regulation of gene expression (reviewed in [11-13]), though conclusive proof of their functional importance has remained elusive. It has, however, been demonstrated that intermolecular triplexes, formed between duplex DNA sequences and olignucleotides designed to bind to them *via* cognate Hoogsteen base pairing interactions, can modulate both gene expression and recombination, allowing targeted manipulation of these processes (reviewed in [14]). Interestingly, recombination between two DNA duplexes can be stimulated by distal binding of a triplex-forming oligonucleotide, in one case as far away as 4000 bp [15]. Following the discovery of intramolecular triplexes and quadruplexes, a fifth major type of non-B-DNA structure, the DNA unwinding element, which is a stably unwound configuration formed by long AT-rich sequences, was described [16]. This type of structure may also be functionally important since it can be formed by sequence found near replication origins [17].

The most interesting of these structural variants with regard to meiotic recombination hotspots are intramolecular triplexes and quadruplexes, because these structures are preferentially formed by GC-rich, poly-purine/poly-pyrimidine-rich sequences, and, as I have shown, sequences of this type are closely associated with meiotic recombination sites in both humans and yeast (see Section 4.3). These results do not necessarily indicate that recombination hotspots have structure-forming potential, however, because sequence properties required for intramolecular triplexes and quadruplexes to form are less well understood than for the other three types of non-B-DNA structure described above. The canonical form of intramolecular triplex, known as H-DNA, requires a poly-purine motif with internal mirror symmetry [18], but non-mirror-symmetric PPTs can also adopt an intramolecular triplex, and a substantial proportion of mismatches to the poly-purine/poly-pyrimidine motif can be tolerated [19-22]. Intramolecular quadruplex formation requires rows of consecutive G/C base pairs, with the fold-back interaction in these structures occurring exclusively between guanine residues, and several variations of quadruplex have recently been described (reviewed in [23]). Moreover, non-B-DNA structures can be context dependent, i.e. they can be destabilized by distal sequence changes [24]. The question of whether or not recombination hotspot sequences form them can therefore only be confidently answered using wet laboratory experiments.

There are already at least two cases in which non-B-DNA structure-forming potential has been shown for DNA from chromosomal regions that frequently recombine during meiosis. The PKD1 locus on human chromosome 16, implicated in kidney disease arising from genetic mutations, shows evidence for past meiotic recombination events and contains a triplex-forming PPT (reviewed in [25]), and the insulin promoter region, which has the potential to form an intramolecular quadruplex [26] is a recombination hotspot judging by local patterns of linkage disequilibrium [27]. One reason why there have been no reports of a general exploration of the link between meiotic recombination hotspots and non-B-DNA structures could be the fact that it is not straightforward to infer the functionality of structures formed in model systems such as plasmids, and the biological significance of non-B-DNA structures in general is still uncertain. All of the major described variants on B-DNA are favoured by supercoiling, and all have been studied almost exclusively in supercoiled plasmids and/or short synthetic oligonucleotides (reviewed in [2]). However, evidence has shown that intramolecular triplexes present in supercoiled plasmid DNA molecules do form, though much less frequently, in protein-free chromosomal DNA [21]. Moreover, structure-specific antibodies for intramolecular triplexes [30] and Z-DNA [4] have been used to demonstrate their occurrence on chromosomal DNA *in vivo*. Non-B-DNA structure formation on chromosomes presumably requires some unpacking of DNA from its normal tight packaging in chromatin, since a degree of denaturation of normal Watson-Crick base pairing is needed for secondary structures to form, and some evidence suggests that they could exist transiently during the processes of transcription [28] and replication [20, 29]. This could be favoured by increased supercoiling tension resulting from the unwinding of the DNA duplex that takes place during these processes, and also by the increased opportunity they afford for intrastrand interactions resulting from an unzipped DNA duplex [29]. The prevalence of these phenomena has not yet been determined however.

A second problem for the investigation of the possible function in meiotic recombination of non-B-DNA structures is that the extent to which they are formed by PPTs, or in poly-purine-rich regions in general, is not yet known. Structure formation in recombination hotspots would therefore not constitute compelling evidence for functional importance, unless it could be shown that they do not occur in poly-purine-rich cold regions. Structure formation in hotspot central regions not seen in hotspot non-central regions would, however, suggest functional significance, because meiotic recombination events are most concentrated in narrow regions within 200-300 base pairs of hotspot mid points [31, 32] . Another approach to investigating the possible functional importance of non-B-DNA structures formed by recombination hotspot sequences takes advantage of the fact that humans and chimpanzees do not have a substantial proportion of hotspot locations in common despite sharing 98% DNA sequence identity [33, 34]. The reason for this is presently unknown. At first sight, the possibility that divergence in intramolecular structures could be responsible looks unlikely, since while the structures are preferentially formed by GC-rich, poly-purine-rich sequences, they can tolerate a substantial proportion of mismatches to the poly-purine motif [19-23]. However, intramolecular triplex formation can be disrupted by single base substitutions [24, 35, 36]. Furthermore, disruption of intramolecular triplexes by distal sequence changes has also been demonstrated, by a study that showed two such structures cannot form within 1500 bp of each other on the same plasmid, due to their requirement for supercoiling energy, which is limited in any naturally occurring DNA duplex [24]. The effect of sequence variation on non-B-DNA structure formation is therefore reminiscent of its effect on meiotic recombination hotspot activity (see Section 1.2.2), further motivating a test of whether sequence changes associated with recombination activity are also linked to changes in DNA secondary structure.

Ideal for addressing this question is a previously described sodium bisulphite modification assay, which enables visualization of the location and extent of single-strandedness in individual DNA molecules, allowing the effects of particular sequence variants to be evaluated [21, 37-39]. The technique involves incubation of DNA fragments of interest in a near-saturated solution of sodium bisulphite, which deaminates cytosine bases to deoxyuracil in single-stranded DNA such as occurs in intramolecular triplexes [21], quadruplexes [40] and cruciforms [41]. Deoxyuracil can be converted to thymidine by PCR, allowing visualization of single-stranded regions following ordinary sequencing. I have used this assay to probe non-B-DNA structure formation in the 500 bp regions spanning the mid points of six human meiotic recombination hotspots found not to have orthologous activity in chimpanzees [34]. These were the Beta-Globin [42], Tap2, DMB1, DMB2, DNA2 and DNA3 [43] hotspots I studied in Chapters 3 and 4 of this thesis (see Section 3.2.1). To date, these are the only six human hotspots for which orthologous recombination rates have been measured in chimpanzees.

7.2 Methods

7.2.1 DNA extraction and preparation

I extracted human genomic DNA from cheek cells obtained with a saline mouthwash using a standard chelex preparation. Genomic DNA from a Western chimpanzee (*Pan Troglodytes verus*) was donated by Victor Wiebe and Svante Paabo from the Max Planck Institute for Evolutionary Anthropology in Liepzig, Germany. I initially amplified genomic DNA from both species using PCR with primers designed to cover 500-600 bp of the genomic regions of interest using the Primer3 software [44], with standard PCR conditions. PCR primers used are listed in Table 7.1.

Table 7.1: PCR primers used in this study
I designed primers to cover 500-600 bp centred on the Bcl-2 major breakpoint region, six recombination hotspot mid points and three non-hot regions (see text). For the six recombination hotspots, the primers were chosen in regions of no difference between the human and chimpanzee reference sequences.

Region	Forward primer	Reverse primer
Bcl-2	cacgtggagcatactgcaaa	tctgttgtccctttgaccttg
non-hot 1	agtggcccacacctgtactc	tgtacttaacacaacttcgtttcaca
non-hot 2	tgacacagagatggtgctgt	tgaacttcttctaactaataggggaaa
non-hot 3	ttgcaatgaacatggagcat	caggcaacaaaagcaaaaat
Beta-globin	tgaagatcgttttcccaatttt	aagtcacagaggctttttgttc
DMB1	ttgagaggccccactgtatt	attggacccaggaagaggag
DMB2	ggatgctgcatgaggagaat	cctggaacctaggaacatgc
DNA2	cggttttcaaaccagaatgc	gcaggagaatggcttgaact
DNA3	ttcaggaacatgccaccata	aattcagctacttttacttgcttttt
Tap2	acctaacactgtgggcgact	ctgcctcctacctcctaccc

I isolated PCR products of the desired molecular weight by electrophoresis through 1% agarose containing 2 µl/75ml of 5 mg/ml ethidium bromide, visualization under ultraviolet light, limiting exposure to a few seconds, and excision from the agarose using a razor blade. I then purified the PCR products using the MiniElute gel purification kit (Qiagen) and ligated them into the ampicillin resistant pGEM®-T Easy vector (Promega), using the pGEM®-T cloning kit according to the manufacturer's instructions. With the ligation mixture I transformed tetracycline-resistant *E.coli* X blue cells, which I had made chemically competent by the Inoue method, following the protocol for the transformation of these cells [45]. I selected well isolated colonies for blue/white screening as described [45] on agar plates which consisted of LB medium containing 15 g/ml agar, 50 µg/ml ampicilin and 15 µg/ml tetracycline, with 40 µl of 2% X-gal and 10 µl of 100 mM IPTG spread onto the plates' surfaces and allowed to absorb overnight (16 hours) at room temperature.

I grew cells from colonies that tested positive in the blue/white screen for 12 hours in 1 ml of liquid LB medium containing ampicillin and tetracycline at the above concentrations at 37 ºC with shaking at 200 rpm. I then purified plasmid DNA by alkaline lysis mini-preparation as described [45] with the exception that molecular biology grade dd H_2O was used to dissolve the DNA instead of TE buffer. I checked the purified plasmid preparation for the desired insert by sequencing using a capillary ABI3100 Genetic Analyzer from Applied Biosystems Inc. Where the correct insert was present, I grew the colony at large scale for preparation of plasmid DNA by lysis with SDS, which I performed as described [45]. Lysis with SDS was used instead of alkaline lysis at this stage to avoid denaturation of any secondary structure in the plasmid, as recommended in the literature [39]. I isolated supercoiled plasmid from this raw preparation by equilibrium centrifugation in a continuous CsCl-ethidium bromide gradient as described [45], with the ultracentrifugation step done at 62,000 rpm for 6 hours in a Beckman NVT 65 rotor (366,000g). I checked supercoiled plasmid preparations by gel electrophoresis as above, and ascertained that each preparation contained only one species of plasmid. This was based on the property of supercoiled, nicked and linear plasmid DNA to migrate at different rates through agarose.

7.2.2 Sodium bisulphite modification assay

I carried out the sodium bisulphite modification assay essentially as described in the papers that reported the intramolecular triplex-forming properties of my positive control fragment, amplified from the human Bcl-2 major breakpoint region [21, 37]. The rationale for this method is that sodium bisulphite ($NaHSO_3$) deaminates single-stranded, non-methylated cytidines to uridines, which can be detected as thymidines after subsequent PCR and sequencing [39]. It can thus be used to detect regions of single-stranded DNA, which occur in non-B-DNA structures including intramolecular triplexes, intramolecular quadruplexes, and cruciforms (reviewed in [2]).

The published protocols prescribe 2.5 M sodium bisulphite [37], which could be misleading because the raw material exists as an equilibrium with inactive sodium metabisulphite, so I carried out preparation of sodium bisulphite according to the instructions given to me by the authors [21, 37] (Michael Lieber, personal communication). I firstly weighed about 0.3 g of sodium bisulphite per reaction and dissolved it in 1,050 µl dd H_2O/g sodium bisulphite. To this I added 525.2 µl of 2 M NaOH/g sodium bisulphite and I dissolved this mixture at 37 ºC with shaking at 200 rpm for 1 hour. In a 2 mL microcentrifuge tube, I mixed 457.5 µl of the sodium bisulphite solution with 12 µl of 20 mM hydroquinone and then added 1 µg of supercoiled plasmid (concentration measured using a Nanodrop spectrophotometer) re-constituted in 30 µl of dd H_2O. I then incubated the mixture for 16 hours at 37 ºC in the dark.

I purified the bisulphite-treated DNA using the Wizard DNA cleanup system (Promega), according to the manufacturer's instructions, desulfonated the purified DNA with 0.3 M NaOH at 37 ºC for 15 minutes, and then precipitated it with two volumes of absolute ethanol at -20 ºC for 1-24 hours. I collected the precipitate by centrifugation at 14,000 g for 15 minutes, washed the pellet with ice cold 70 % ethanol and dissolved it in 30 µl dd H_2O. I then PCR amplified, cloned and sequenced the DNA as described above (Section 7.2.1). I included in my sequence dataset for a analysis all sequences for which a read of at least 300 bp was obtained, and I compared sequences using the BLAST algorithm [46]. The reference sequences I refer to in Sections 7.3 and 7.4 are those reported by the papers that originally characterized the hotspots in humans [42, 43] (see Section 3.2.1). The chimpanzee reference sequence was Build 2.1.

7.3 Results

As a positive control, I ran the sodium bisulphite modification assay on a fragment from the Bcl-2 major breakpoint region shown previously to test positively for non-B-DNA structure with this and other assays [21, 37]. I also ran three negative controls in order to gain some idea of background levels of modification. These were chosen from recombinationally inactive DNA between hotspot clusters in the MHC Class II region. I initially chose two high PPT density and two low PPT density regions. Two of the controls, which I refer to as non-hot 1 and non-hot 2, spanned the highest point of the PPT density peak between the Tap2 and DMB1 hotspots (based on a 10 kb sliding window; see Figure 4.9). These were the low PPT-density products, since this peak is the lowest in the region. One of the high PPT-density products, from the Ring3 gene promoter region, could not be cloned. Therefore, I only ran one further negative control fragment, which was amplified from highest peak in PPT density occurring between the DNA3 and DPA hotspots (10 kb sliding window; see Figure 4.9), and I refer to this fragment as non-hot 3. I also tested 500-600 bp products spanning the mid points of the six human meiotic recombination hotspots listed above, and the orthologous chimpanzee regions shown not to contain a hotspot [34].

The results of the sodium bisulphite modification assay are summarized in Table 7.2. The table shows the amount of modification in each tested molecule, and also the number of molecules containing secondary structure. The rationale for my definition of secondary structure was as follows. Non-B-DNA structures contain single-stranded DNA, revealed after sodium bisulphite modification as continuous stretches of converted cytosines manifested as G/C→A/T transition mutations occurring on the same strand [21, 37, 41]. In the Bcl-2 positive control fragment I tested, the maximum number of contiguous converted cytosines was six, so I defined substantial evidence for secondary structure as six or more contiguous converted cytosines (with one mismatch allowed) occurring in an area of at least 12 bp. I used 12 bp as a minimum based on the fact that it was the length of the shortest reported secondary structure I could find in the literature [47]. I found that about half of the sequenced molecules from the Bcl-2 major breakpoint region contained substantial evidence for secondary structure by this definition (Table 7.1), which was the proportion of structure-forming supercoiled plasmid molecules reported in the original studies [21, 37]. The Bcl-2 structure, which has been well characterized, has at least 30 bp of single-strandedness, and not all cytosines in this region are converted in any given reaction, either in my hands, or in those of the original authors [21]. I therefore measured, for each test fragment, the total amount of modification as well as the number of molecules containing continuous stretches of converted cytosines. I also mapped the locations of hotspots of conversion, and in the Bcl-2 fragment the locations of these are similar to those described previously [21, 37] (Figure 7.1).

Table 7.2 Summary of results from the sodium bisulphite modification assay
Listed for each type of molecule are the total number of fragments subjected to sodium bisulphite modification and sequenced, the number showing evidence for secondary structure as defined (see text), the total number of G/C bases sequenced, and the total number of modifications, defined as G/C→A/T transition substitutions relative to the untreated molecule. Statistical comparisons of the proportion of sequenced cytosine bases showing modification in humans compared with chimpanzees were by Mann-Whitney U Test (Bonferroni-corrected alpha = 0.004).

Region	Species	Molecules tested	Molecules with structure	Cytosines modified	Total G/C bases sequenced	Percentage modification	P value (human v chimp.)
Bcl-2 (+ cntl)	Human	15	8	127	2998	4.23	n/a
Non-hot 1 (- cntl)	Human	16	1	200	3616	5.53	n/a
Non-hot 2 (-cntl)	Human	15	0	67	2397	2.80	n/a
Non-hot 3 (-cntl)	Human	14	0	80	2352	3.40	n/a
Beta-globin	Chimp.	15	0	81	1777	4.56	n/s
	Human	15	0	56	1694	3.31	
DMB1	Chimp.	20	6	295	4930	5.98	0.0037
	Human	36	9	641	8800	7.28	
DMB2	Chimp.	18	7	197	3538	5.57	n/s
	Human	17	7	206	3500	5.89	
DNA2	Chimp.	21	2	307	4947	6.21	$<10^{-4}$
	Human	18	10	350	4099	8.54	
DNA3	Chimp.	33	2	95	6374	1.5	$<10^{-43}$
	Human	22	4	278	4218	6.59	
Tap2	Chimp.	15	2	154	3384	4.55	n/s
	Human	18	2	186	3786	4.91	

There was no substantial evidence for secondary structure in negative controls. Non-hot region 1 showed a relatively high 5.53 % level of cytosine modification, but only one out of 16 sequenced molecules from this region contained evidence for a non-B-DNA structure as defined above. Between 25 and 50% of the DMB1, DMB2 and human DNA2 hotspot molecules tested showed evidence for structure according to my definition. The DMB1, DNA2 and DNA3 hotspot central regions were more sensitive to sodium bisulphite, in terms of the proportion of total cytosines modified, in humans than chimpanzees. However, numbers of molecules with long strings of consecutive modified cytosines did not clearly differ between the two species in either the DNA3 or DMB1 fragments. The biological significance of the statistical differences found for these regions is therefore not clear. A relatively low proportion of molecules from the Tap2 hotspot also formed structures as defined in both the human and chimpanzee fragments, but the significance of this is also unclear in view of the relatively small number of molecules tested from this region.

The most noteworthy result from this study therefore appears to be the observation that ten out of 18 molecules from the human DNA2 hotspot central region contained non-B-DNA structure as defined, and only two out of 21 chimpanzee molecules showed such evidence. Figure 7.1 shows that the proportion of cytosines modified was highest in four areas of the human hotspot sequence (bases highlighted in red indicating at least 40% modification), all of which occurred on the same strand. The fourth of these, reading left to right, overlaps with a region of three single nucleotide differences between the human and chimpanzee sequences. The chimpanzee fragment was modified to some extent in the areas of the first three peaks most active in humans, but not at all in the area of the fourth peak. Combined with the fact that modification in the first three peak regions was much less frequent in the chimpanzee than in the human fragment, this result suggests that the chimpanzee haploytype disrupts the formation of a non-B-DNA structure. The nature of this structure is uncertain based on these data, but there are no obvious inverted repeats in the sequence, suggesting that it is most likely an intramolecular triplex or quadruplex rather than a cruciform.

Figure 7.1 Sequence contexts of bisulphite-modified bases

Sequences from the Bcl-2 major breakpoint region (A). and the DNA2 hotspot central region amplified from human (B) and chimpanzee (C) genomic DNA. Modified cytosines are highlighted in blue (0-10% modification), green (11-20% modification), yellow (21-30% modification), pink (31-40% modification) or red (more than 40% modification). The mid point of the Bcl-2 major breakpoint region, and the hotspot mid point, are underlined. Nucleotides differing from the reference sequence in the original, unmodified molecules are shown in red coloured text. Human/chimpanzee sequence differences are highlighted in grey in the chimpanzee sequence. Total numbers of molecules of each fragment type sequenced are shown in brackets.

A) Bcl-2 major breakpoint region (n = 16)

```
CACGTGGAGCATACTGCAAACTGACTCCATTAAAATGATTTTGGCAGGATAGCAGCACAG
GTGCACCTCGTATGACGTTTGACTGAGGTAATTTTACTAAAACCGTCCTATCGTCGTGTC

GATTGGATATTCCATATTCATCACTTTGACAATGTAAACCTTTCATAAAATAATATTTTG
CTAACCTATAAGGTATAAGTAGTGAAACTGTTACATTTGGAAAGTATTTTATTATAAAAC

CTTAAAAATTAGAATCATTCAAAGGTCTGATCATTCTGTTCCCTGAGGCCCGCCGGGGAG
GAATTTTTAATCTTAGTAAGTTTCCAGACTAGTAAGACAAGGGACTCCGGGCGGCCCCTC

GTCTGGCTTCATACCACAGGTTTCCTGCTTTCTTGGTGGAGCGTAAGCACCACTGCATTT
CAGACCGAAGTATGGTGTCCAAAGGACGAAAGAACCACCTCGCATTCGTGGTGACGTAAA

CAGGAAGACCCTGAAGGACAGCCATGAGAAAGCCCCTGCGGAAGGAGGGCAGGAGGGCTC
GTCCTTCTGGGACTTCCTGTCGGTACTCTTTCGGGGACGCCTTCCTCCCGTGGTCCCGAG

TGGGTGGGTCTGTGTTGAAACAGGCCACGTAAAGCAACTCTCTAAAGGTCAAACCACCAT
ACCCACCCAGACACAACTTTGTCCGGTGCATTTCGTTGAGAGATTTCCAGTTTGGTGGTA

AGATTTGAATCTGCTGGTCATTCGCCATCTGGATTTTTAACTGAATGAATCTCATGGGTT
TCTAAACTTAGACGACCAGTAAGCGGTAGACCTAAAAATTGACTTACTTAGAGTACCCAA

TAACCAAACATGCATGTAATCCTGAATACCGTGAATTAAATGCGGAATTGCCCAGGGACG
ATTGGTTTGTACGTACATTAGGACTTATGGCACTTAATTTACGCCTTAACGGGTCCCTGC
```

Figure 7.1 continued from previous page

B) DNA2 hotspot central region – chimpanzee (n = 21)

```
CATAAGAACTGCTTGGGATCCTTTTAAAAGTACAGGCATTGGCCTGGTGCAGTGGCTCAT
GTATTCTTGACGAACCCTAGGAAAATTTTCATGTCCGTAACCGGACCACGTCACCGAGTA

TCCTGTAATCCCAGCACTTTGGGAGGCCAAGGGGACAGGACTGCTTGAGGCCAAGAGGTG
AGGACATTAGGGTCGTGAAACCCTCCGGTTCCCCTGTCCTGACGAACTCCGGTTCTCCAC

GAAACCATCTTGGGCTACATAGAGAGACCCCATCTCTAAAAAGAAAGATTTAAAAATTAA
CTTTGGTAGAACCCGATGTATCTCTCTGGGGTAGAGATTTTTCTTTCTAAATTTTTAATT

CCAGGCATGGTGGCTCGCACCTGTATTCCCAGCCATTCGAGAGGCTGAGGCTGGAGGAGT
GGTCCGTACCACCGAGCGTGGACATAAGGGTCGGTAAGCTCTCCGACTCCGACCTCCTCA

GCTTGAGCCCAGGAGTTCAAGGCTGCAGTGAGCCAAGATTGCGCCACTGCACTCCAGCCT
CGAACTCGGGTCCTCAAGTTCCGACGTCACTCGGTTCTAACGCGGTGACGTGAGGTCGGA

AGGTGACAGAGTGAGACCCTGTCTC-------ATAAATAAATAAAATATAAAAATAACAG
TCCACTGTCTCACTCTGGGACAGAG-------TATTTATTTATTTTATATTTTTATTGTC

TCATCACCCAGACCTACTGAATTAGAATCTCGGGGTGCAAGGGGCAGCAACAGGGAAGC
AGTAGTGGGTCTGGATGACTTAATCTTAGAGCCCCACGTTCCCCGTCGTTGTCCCTTCG

TGTCTTTTTTGGGATGGGGTCTCACTCTGTCACCAGGCTGGAGTGCCGTGGCATGATCTC
ACAGAAAAAACCCTACCCCAGAGTGAGACAGTGGTCCGACCTCACGGCACCGTACTAGAG

AGCTCACTGCAACCTCCACC
TCGAGTGACGTTGGAGGTGG
```

C) DNA2 hotspot central region – human (n = 18)

```
CATAAGAACTGCTTGGGATCCTTTTAAAAGTACAGGCATTGGCCTGGTGCAGTGGCTCAT
GTATTCTTGACGAACCCTAGGAAAATTTTCATGTCCGTAACCGGACCACGTCACCGAGTA

TCCTGTAATCCCAGCACTTTGGGAGGCCAAGGGGACAGGACTGCTTGAGGCCAAGAGGTG
AGGACATTAGGGTCGTGAAACCCTCCGGTTCCCCTGTCCTGACGAACTCCGGTTCTCCAC

GAAACCATCTTGGGCTACATAGAGAGACCCCATCTCTACAAAGAAAGATTTAAAAACTAA
CTTTGGTAGAACCCGATGTATCTCTCTGGGGTAGAGATGTTTCTTTCTAAATTTTTGATT

CCAGGCATGGTGGCTCGCACCTGTATTCCCAGCCACTGGGGAGGCTGAGGCCGGAGGAGT
GGTCCGTACCACCGAGCGTGGACATAAGGGTCGGTGACCCCTCCGACTCCGGCCTCCTCA

GCTTGAGCCCAGGAGTTCAAGGCTGCAGTGAGCCAAGATTGCGCCACTGCACTCCAGCCT
CGAACTCGGGTCCTCAAGTTCCGACGTCACTCGGTTCTAACGCGGTGACGTGAGGTCGGA

AGGTGACAGAGTGAGACCCTGTCTCTAAATAAATAAATAAATAAAATATAAAAATAACAG
TCCACTGTCTCACTCTGGGACAGAGATTTATTTATTTATTTATTTTATATTTTTATTGTC

TCATCACCCAGACCTACTGAATTAGAATCTCGGGAGTGCAGGGGGCAGCAACAGGTGGGT
AGTAGTGGGTCTGGATGACTTAATCTTAGAGCCCTCACGTCCCCCGTCGTTGTCCACCCA

GTCTTTTCTGAGATGGGGTCTCACTCTGTCACCAGGCTGGAGTGCCATGGCATGATCTCA
CAGAAAAGACTCTACCCCAGAGTGAGACAGTGGTCCGACCTCACGGTACCGTACTAGAGT

GCTCACTGCAACCTCCACC
CGAGTGACGTTGGAGGTGG
```

7.4 Discussion

With the exception of the DNA2 hotspot, there appears to be no clear difference between humans and chimpanzees in the non-B-DNA structure-forming potential of the fragments assayed. This should be tested further, however. In view of the potential dependence of structure-forming sequences on flanking regions [36], the assay should be repeated on whole hotspot fragments. It is also difficult to make firm conclusions based on my data because the numbers of sequenced molecules are low at the time of writing. The DNA2 hotspot is an exception to this because of the clearly significant difference between the two species, and the presence of base substitutions in a highly bisulphite-sensitive area of the human hotspot fragment not sensitive in the chimpanzee. These three substitutions were expected based on the reference sequences, but there are allelic differences from the reference sequences in the human and chimpanzee individuals tested (Figure 7.1). This complication is also present in some of the other tested hotspots, which suggests, considering the sensitivity of hotspots to single nucleotide changes (see Section 1.2.2), that the assays should be repeated in several individuals with different haplotypes.

What can be concluded from the data presented in this chapter is that there is clear evidence for non-B-DNA structure formation in supercoiled plasmid DNA by the sequences from the central regions of the DNA2 and DMB2 human meiotic recombination hotspots. Arguably, evidence is also compelling for a third hotspot, DMB1, and the Tap2 fragment showed some elevation above background in terms of the number of molecules with substantial numbers of consecutive converted cytosines, though numbers of Tap2 molecules tested are relatively low at the time of writing. The Beta-Globin hotspot central region showed no such evidence, but in the context of this thesis it should be pointed out that the region does contain a 50 bp microsatellite consisting of alternating purines and pyrimidines, a motif with the potential to form cruciform and hairpin structures (reviewed in [48]) and Z-DNA (reviewed in [14]), none of which would be expected to produce extended areas of bisulphite modification with the assay I used.

Modification with sodium bisulphite is a highly regarded method for detecting non-B-DNA structures today, based on the amount of emphasis placed on it in recent published papers [21, 37, 38]. These papers also utilized several other techniques in conjunction with the assay, to confirm the existence of a structure, and to characterize its form. Two of these are quite inexpensive, and should be considered promising avenues for research into the questions

I have raised here. Firstly, the formation of structures on linear DNA can be confirmed using a gel shift assay, which takes advantage of the differing mobility in poly-acrylamide gel of molecules with non-B-DNA structure [21]. Secondly, spectrophotometric methods have commonly been used to look for specific light absorbance patterns characteristic of particular structural forms. These include circular dichroism, ultra-violet and nuclear magnetic resonance spectroscopy [21, 49-52]. Another relatively inexpensive way to explore intramolecular structure formation is to test for replication pausing *in vitro* with and without the addition of ions that stabilize specific types of secondary structure, namely Mg^+ for triplexes and K^+ or Na^+ for quadruplexes [20, 53].

References

1. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**:737-738.
2. Mirkin S: **Discovery of alternative DNA structures: a heroic decade (1979-1989)**. *Frontiers in Bioscience* 2008, **13**:1064-1071.
3. Wang: **Molecular structure of a left-handed double helical DNA fragment at atomic resolution**. *Nature* 1979, **282**(5740):680-686.
4. Lipps HJ, Nordheim A, Lafer EM, Ammermann D, Stollar BD, Rich A: **Antibodies against Z DNA react with the macronucleus but not the micronucleus of the hypotrichous ciliate stylonychia mytilus**. *Cell* 1983, **32**(2):435-441.
5. Panayotatos N, Wells RD: **Cruciform structures in supercoiled DNA**. *Nature* 1981, **289**(5797):466-470.
6. Haniford DB, Pulleyblank DE: **Transition of a cloned d(AT)n-d(AT)n tract to a cruciform in vivo**. *Nucleic Acids Res* 1985, **13**(12):4343-4363.
7. Larsen A, Weintraub H: **An altered DNA conformation detected by S1 nuclease occurs at specific regions in active chick globin chromatin**. *Cell* 1982, **29**(2):609-622.
8. Hentschel CC: **Homocopolymer sequences in the spacer of a sea urchin histone gene repeat are sensitive to S1 nuclease**. *Nature* 1982, **295**(5851):714-716.
9. Felsenfeld G, Rich A: **Studies on the formation of two- and three-stranded polyribonucleotides**. *Biochim Biophys Acta* 1957, **26**(3):457-468.
10. Sen D, Gilbert W: **Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis**. *Nature* 1988, **334**(6180):364-366.
11. Zain R, Sun JS: **Do natural DNA triple-helical structures occur and function in vivo?** *Cell Mol Life Sci* 2003, **60**(5):862-870.
12. Qin Y, Hurley LH: **Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions**. *Biochimie* 2008.
13. Wells RD, Collier DA, Hanvey JC, Shimizu M, Wohlrab F: **The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences**. *Faseb J* 1988, **2**(14):2939-2949.

14. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM *et al*: **A second generation human haplotype map of over 3.1 million SNPs**. *Nature* 2007, **449**(7164):851-861.
15. Biet E, Sun JS, Dutreix M: **Stimulation of D-loop formation by polypurine/polypyrimidine sequences**. *Nucleic Acids Res* 2003, **31**(3):1006-1012.
16. Sheflin LG, Kowalski D: **Altered DNA conformations detected by mung bean nuclease occur in promoter and terminator regions of supercoiled pBR322 DNA**. *Nucleic Acids Res* 1985, **13**(17):6137-6154.
17. Umek RM, Kowalski D: **The ease of DNA unwinding as a determinant of initiation at yeast replication origins**. *Cell* 1988, **52**(4):559-567.
18. Htun H, Dahlberg JE: **Topology and formation of triple-stranded H-DNA**. *Science* 1989, **243**(4898):1571-1576.
19. Hanvey JC, Shimizu M, Wells RD: **Intramolecular DNA triplexes in supercoiled plasmids. II. Effect of base composition and noncentral interruptions on formation and stability**. *J Biol Chem* 1989, **264**(10):5950-5956.
20. Dayn A, Samadashwily GM, Mirkin SM: **Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization**. *Proc Natl Acad Sci U S A* 1992, **89**(23):11406-11410.
21. Raghavan SC, Chastain P, Lee JS, Hegde BG, Houston S, Langen R, Hsieh CL, Haworth IS, Lieber MR: **Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation**. *J Biol Chem* 2005, **280**(24):22749-22760.
22. Klysik J: **An intramolecular triplex structure from non-mirror repeated sequence containing both Py:Pu.Py and Pu:Pu.Py triads**. *J Mol Biol* 1995, **245**(5):499-507.
23. Burge S, Parkinson GN, Hazel P, Todd AK, Neidle S: **Quadruplex DNA: sequence, topology and structure**. *Nucleic Acids Res* 2006, **34**(19):5402-5415.
24. Boles TC, Hogan ME: **DNA structure equilibria in the human c-myc gene**. *Biochemistry* 1987, **26**(2):367-376.
25. Bissler JD: **Triplex DNA and human disease**. *Frontiers in Bioscience* 2007, **12**:4536-4546.
26. Hammond-Kosack MC, Dobrinski B, Lurz R, Docherty K, Kilpatrick MW: **The human insulin gene linked polymorphic region exhibits an altered DNA structure**. *Nucleic Acids Res* 1992, **20**(2):231-236.
27. Chakravarti A, Elbein SC, Permutt MA: **Evidence for increased recombination near the human insulin gene: implication for disease association studies**. *Proc Natl Acad Sci U S A* 1986, **83**(4):1045-1049.
28. Kohwi Y, Panchenko Y: **Transcription-dependent recombination induced by triple-helix formation**. *Genes Dev* 1993, **7**(9):1766-1778.
29. Samadashwily GM, Dayn A, Mirkin SM: **Suicidal nucleotide sequences for DNA polymerization**. *Embo J* 1993, **12**(13):4975-4983.
30. Ohno M, Fukagawa T, Lee JS, Ikemura T: **Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies**. *Chromosoma* 2002, **111**(3):201-213.
31. Baudat F, de Massy B: **Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot**. *PLoS Genet* 2007, **3**(6):e100.
32. Kauppi L, Jeffreys AJ, Keeney S: **Where the crossovers are: recombination distributions in mammals**. *Nat Rev Genet* 2004, **5**(6):413-424.

33. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**(4):429-434.

34. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P *et al*: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**(5718):107-111.

35. Ulrich MJ, Gray WJ, Ley TJ: **An intramolecular triplex is disrputed by point mutations associated with hereditary persistence of fetal hemoglobin**. *Journal of Biological Chemistry* 1992, **267**:18649-18658.

36. Bacolla A, Ulrich MJ, Larson JE, Ley TJ, Wells RD: **An intramolecular triplex in the human gamma-globin 5'-flanking region is altered by point mutations associated with hereditary persistence of fetal hemoglobin**. *J Biol Chem* 1995, **270**(41):24556-24563.

37. Raghavan S, Swanson PC, Wu X, Hsieh CL, Lieber MR: **A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex**. *Nature* 2004, **4**(428):88-93.

38. Yu K, Chedin F, Hsieh CL, Wilson TE, Lieber MR: **R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells**. *Nat Immunol* 2003, **4**(5):442-451.

39. Raghavan SC, Tsai A, Hsieh CL, Lieber MR: **Analysis of non-B DNA structure at chromosomal sites in the mammalian genome**. *Methods in Enzymology* 2006, **409**:301-316.

40. Sun D, Guo K, Rusche JJ, Hurley LH: **Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents**. *Nucleic Acids Res* 2005, **33**(18):6070-6080.

41. Gough GW, Sullivan KM, Lilley DM: **The structure of cruciforms in supercoiled DNA: probing the single-stranded character of nucleotide bases with bisulphite**. *Embo J* 1986, **5**(1):191-196.

42. Schneider JA, Peto TE, Boone RA, Boyce AJ, Clegg JB: **Direct measurement of the male recombination fraction in the human beta-globin hot spot**. *Hum Mol Genet* 2002, **11**(3):207-215.

43. Jeffreys AJ, Kauppi L, Neumann R: **Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex**. *Nat Genet* 2001, **29**(2):217-222.

44. Rozen S, Skaletsky H: **Primer3 on the www for general users and for biologist programmers**. In: *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Edited by Krawetz S, Misener S. Totowa, N.J.: Humana Press; 2000: 365-386.

45. Sambrook J, Russell DW: **Molecular Cloning**, vol. 1, Third edn. Cold Spring Harbour, New York: Cold Spring Harbour Laboratory Press; 2001.

46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.

47. Matsugami A, Ouhashi K, Kanagawa M, Liu H, Kanagawa S, Uesugi S, Katahira M: **An intramolecular quadruplex of (GGA)(4) triplet repeat DNA with a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad, and its dimeric interaction**. *J Mol Biol* 2001, **313**(2):255-269.

48. Lobachev KS, Rattray A, Narayanan V: **Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells**. *Front Biosci* 2007, **12**:4208-4220.

49. Fernando H, Reszka AP, Huppert J, Ladame S, Rankin S, Venkitaraman AR, Neidle S, Balasubramanian S: **A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene**. *Biochemistry* 2006, **45**(25):7854-7860.

50. Sun XG, Cao EH, He YJ, Qin JF: **Spectroscopic comparison of different DNA structures formed by oligonucleotides**. *J Biomol Struct Dyn* 1999, **16**(4):863-872.

51. Mergny JL, Li J, Lacroix L, Amrane S, Chaires JB: **Thermal difference spectra: a specific signature for nucleic acid structures**. *Nucleic Acids Res* 2005, **33**(16):e138.

52. Bishop GR, Chaires JB: **Characterization of DNA structures by circular dichroism**. *Curr Protoc Nucleic Acid Chem* 2003, **Chapter 7**:Unit 7 11.

53. Han H, Hurley LH, Salazar M: **A DNA polymerase stop assay for G-quadruplex-interactive compounds**. *Nucleic Acids Res* 1999, **27**(2):537-542.

Chapter 8

General Discussion

8.1 Links between microsatellites, poly-purine/poly-pyrimidine tracts and meiotic recombination hotspots

In Chapter 2 I described an association between microsatellite abundance and meiotic double-strand break hotspots in the genome of the yeast *Saccharomyces cerevisiae*, and in Chapter 3 I showed that there is no association between microsatellites in general and meiotic recombination hotspots in humans. In Chapter 4 I repeated these investigations for poly-purine/poly-pyrimidine tracts (PPTs), and I found that these elements are over-represented in recombination hotspots in both humans and yeast. I was able show that the associations I found are not caused by genomic factors shown previously, or theoretically expected, to have the potential to correlate with both simple sequences and recombination. In yeast, I showed that transposable elements, GC-content, promoter regions and transcriptional frequency have no substantial effect on the correlation between simple sequence density and recombination intensity in intergenic regions. In the human genome the situation is slightly different because most hotspots do not occur in promoter regions [1], and intergenic regions are many-fold larger than in yeast. This and other evidence pointed to the possibility that simple sequences could co-vary with general genetic diversity in non-functional areas of the genome, and recombination also correlates with genetic diversity [2] while tending to avoid genes [1]. I therefore controlled for single nucleotide polymorphism and gene density using generalized linear models, and the analysis showed that these factors have little effect on the correlation between recombination rate and simple sequences of the types I studied. Although the influence of GC-content on the correlation between PPTs and recombination is quite substantial in humans, other evidence indicates that it is not the explanation for the enrichment of PPTs in human recombination hotspots. Most notably, PPT frequency increases markedly toward hotspot mid points, while GC-content shows no such tendency.

Variation in GC-content, and its higher level in the human genome relative to yeast, might, however, be related to the lack of conservation between humans and yeast of the association between recombination hotspots and microsatellites. This is because the discrepancy can be explained to a great extent by the fact that poly-A and poly AT are by far the predominant species of microsatellite in yeast hotspots (see Section 2.3), but are negatively associated with frequently recombining regions in the human genome [1, 3]. However, PPT density, including poly-A considered separately, correlates quite strongly with GC-content in humans, so increasing GC-content, in evolutionary time, has not always been associated with an increased frequency of GC-biased mutations of poly-A. One other possible explanation of the patterns I observed is that G/C→A/T transitions in recombination hotspots occurring as a result of biased gene conversion (see Section 1.3) cause poly-A to convert to poly-purine with some GC-content and poly-AT to gain in complexity thereby rendering it undetectable by microsatellite search algorithms. Similar to the idea that recombination could drive high frequencies of simple sequences by a mutation bias, however, this explanation must assume greater constraint on hotspot locations in evolutionary time than is evident from their lack of conservation between humans and chimpanzees [4, 5], and their sensitivity to single nucleotide [6] and non-sequence [7] changes.

The patterns I observed could also reflect a selective constraint on AT-rich microsatellites in yeast recombination hotspots that is not present in the human genome. This is suggested by evidence for a functional role of poly-A in the regulation of gene expression [8, 9], since yeast intergenic regions average only about 500 bp, so most or all yeast recombination hotspots occur in relatively close proximity to gene promoters. A notable difference between the two types of sequence is that poly-A is stiff and cannot form fold-back structures [10, 11], while GC-rich PPTs can form the fold-back structures intramolecular triplexes and quadruplexes (see Section 7.1). A unique way in which poly-A might function, therefore, is to help modulate large-scale chromosome structure by providing regions that cannot bend, potentially helping to bring about interactions between regulatory sequences [9].

It is also possible that speculation on the function of simple sequences in recombination hotspots is meaningless, since their link with recombination could be driven solely by a mutation bias, but the results presented in Chapters 5 and 6 suggest that this is unlikely. Both types of sequence are less complex than normal DNA, suggesting that they could mutate by strand or chromosome misalignment, the occurrence of which might plausibly be increased in recombination hotspots (see Section 5.1.1). However, I could find no consistent association between recombination and length polymorphism in PPTs or microsatellites in the human genome. I found marginal evidence for a possible mutation bias of recombination on PPTs in terms of single nucleotide changes, and while the possibility that this could drive the association between PPTs and hotspots, particularly the strong enrichment of PPTs in hotspot central regions, seems remote, it is consistent with the possibility that single nucleotide changes over evolutionary time have converted hotspot-associated poly-A tracts into GC-rich PPTs. This assumes GC-biased mutations, however, and I have not investigated whether such a bias exists. The situation with regard to a mutation bias might of course be different in yeast, and this possibility has not yet been tested either, but in view of the fact that all evidence points to a high degree of conservation in the recombination machinery, even between species as diverse as humans and yeast (reviewed in [12]), it seems likely that the strong associations between microsatellites, PPTs and recombination hotspots in yeast are not driven solely by a mutation bias. I also found that neither two-copy microsatellites, nor short PPTs, are over-represented in meiotic DSB hotspots in *S. cerevisiae*, suggesting that recombination is not involved in the initial formation of these simple sequences.

A functional association between microsatellites, PPTs and recombination hotspots is therefore suggested by my data, and this possibility is supported by a considerable amount of previous evidence (see Sections 2.4 and 4.4). Some common features of microsatellites and PPTs could relate to this putative function. As noted (see Sections 2.4 and 4.4) sequences of both types can bind transcription factors [13, 14] and both types of sequence have the potential to modulate chromatin structure [8, 15-21], and to promote replication pausing [22, 23]. All three of these processes could regulate recombination hotspots by potentiating alterations of local chromatin structure to allow access to the recombination machinery [24]. Replication pausing and modulation of chromatin structure have been linked to formation of non-B-DNA structures by simple sequences in the presence of supercoiling, and these structures have also been implicated in the regulation of gene expression (see Section 7.1), so secondary structure could commonly mediate functions of simple sequences in general [25-30].

In view of previous reports over a period of decades, it is obvious that the potential of simple sequences to stimulate recombination is well known. The generality of findings that microsatellites can function in meiotic recombination hotspots [21, 31, 32] has apparently not been tested, however, and there are two likely reasons for this. Firstly, evidence has emerged that the initiation of meiotic recombination is not regulated simply at the level of local sequence (see Section 1.2). The location and usage of recombination hotspots is governed by a combination of local sequence [21, 32] and distal sequence factors [33], and the involvement of non-sequence (epigenetic factors) is indicated by sex-specific recombination rates [6]. These may result at least in part from differential expression of proteins involved in the recombination machinery [34], but they are primarily driven by sex-specific frequency of hotspot use rather than sex-specific hotspot locations [6]. It is clear that all of these factors somehow work together to regulate hotspot recombination, but the manner in which each of them operates remains largely unknown. The report of a presence/absence polymorphism at a human hotspot without a change in sequence [7] suggests that local sequence is not always involved, and this has recently led authors away from the idea that hotspots are predominantly regulated by local sequences in general [35]. A large number of studies have, however, reported direct experimental evidence for the importance of local sequences [21, 32, 33, 36-42]. In view of currently available data, a plausible hypothesis is that hotspots are potentiated by a combination of local and distal sequences, but that hotspot use at any meiotic cell division is governed to a degree by epigenetic factors. Simple sequences within hotspots, and in hotspot flanking regions, could have a functional role within this framework, in the ways described above.

A second probable reason for the lack of recent attention given to the possible functional role of simple sequences in recombination hotspots is that obvious simple sequence patterns common to large proportions of hotspots have not been found. The data I presented in Chapter 7, along with previous evidence for unexpectedly loose sequence restraints on the formation of intramolecular triplexes [28] and quadruplexes [43], suggest that obvious sequence patterns such as the canonical mirror repeat H-DNA motif, inverted repeats, closely spaced guanine runs, or microsatellites are not always required for the formation of non-B-DNA structures. The hypothesis of a widespread causal link between intramolecular structures and recombination hotspots is therefore plausible, though clearly the link must be context-dependent (see Section 1.2).

Another point worth considering in relation to the reasons underlying the link between simple sequences and meiotic recombination sites is that it may have evolved to provide a way in which the potential of these sequences to damage chromosomes can be regulated. This potential is considerable, since non-B-DNA structures formed by PPTs (reviewed in [44]), trinucleotide repeats (reviewed in [45]) and other microsatellites (reviewed in [46, 47]) have been implicated in many genetic diseases including cancer-causing chromosomal translocations (for a recent general review see ref [48]). However, a contrasting report showed that mutations associated with hereditary persistence of foetal haemoglobin destabilize the ability of their surrounding sequence to form an intramolecular triplex [49], and the common occurrence of non-B-DNA structures shows that they do not always cause genomic instability.

Much of the recorded disease-causing instability of simple sequences occurs in somatic cells, and the regulation of mitotic recombination is not yet well understood in comparison to heritable meiotic recombination, but a link with meiotic recombination has been demonstrated in some cases (see the aforementioned reviews). If there is a commonly occurring link between non-B-DNA structures and meiotic recombination, the plausibility of which is suggested by this thesis, however, some localized abnormality in the regulation of this link must occur in cases in which it is involved in genetic instability causing disease. Many studies using model systems have implicated the axis of intramolecular structure formation, replication blockade and recombination in disease-causing simple sequence instability, but its potential to cause damage is presumably repressed in the normal replication of chromosomal DNA. This could occur through the prevention of non-homologous recombination by other processes involved in the regulation of recombination hotspots. Another possibility is that replication blockade by non-B-DNA structures is normally suppressed by protein binding, and the heterogeneous nuclear ribonucleoprotein family of proteins have been shown to possess such an activity *in vitro* [50].

8.2 Promising directions for future work

The emergence of high-definition genome-wide recombination maps based on direct observation of crossover events (as in, for example, ref [6]) will provide an opportunity to test further some of the questions raised in this thesis. Firstly, it will enable analysis of a greater number of larger regions by wavelet analysis. The scale of this analysis was limited for the recombination map I used [1] because of breaks in the data caused by the need to convert annotations between genome builds. As a result, I could not analyse a substantial amount of the human genome using a uniform region size of more than 32 mega bases. If this could be increased to 64 or 128 mega bases, the hypothesis of a very broad scale correlation between recombination and simple sequences at the level of recombination “jungles” and “deserts” [51] could be tested. This could resolve the question of why I did not see a correlation between microsatellite abundance and recombination rate in the human genome while one had been found previously with window sizes of five and ten mega bases [52]. In relation to this avenue for future work, it should be noted that my data suggest that the correlation between simple sequences and recombination may be quite variable across regions, so the analysis of a few very large areas, for example the long arm of chromosome 1, considered in isolation, may not be able to provide a complete picture of scale-specific and third factor influences on the correlation genome-wide.

A second important feature of high-definition recombination hotspot maps based on direct observation of crossovers is that they include data about sequence polymorphisms associated with localized changes in recombination rate [6]. These data will enable a test of the generality of my observation that polymorphisms affecting recombination occur in GC-rich poly-purine-rich sequence contexts. If they do, the possibility that secondary structure functions in meiotic recombination will be supported, and targets for testing this possibility will be presented. Another possibility is that specific sequence motifs regulating recombination hotspots will be discovered using this new data, and the extent to which the 5-9 bp poly-purine-rich motifs, including CCTCTCCC and CCCCACCCC, previously found to be enriched in human hotspots [1], are linked to recombination-associated haplotypes can be tested.

Apart from the sodium bisulphite modification assay I utilized (see Chapter 7), there are several available methods to test the potential of sequence polymorphism to affect secondary structure formation (see Section 7.4). These tests could be applied to the extension of my work into the comparison of human hotspot central regions and orthologous non-hot regions in chimpanzees. Elucidating the sequence requirements for non-B-DNA structures to form is a goal of this work, but it seems likely that this will be extremely complex because numerous forms of intramolecular triplexes and quadruplexes are possible, they probably do not require particular motifs, and they can be destabilized by distal sequence changes (see Section 7.1). In view of the role of epigenetic factors in regulating meiotic recombination events, it seems unlikely in any case that the locations of interchromosomal crossovers at any given meiosis will prove to be predictable from sequence alone. Even if secondary structures are universally involved and can be predicted, their involvement is presumably context dependent, or probabilistic, given the evidence discussed above and in Section 1.2.2.

The necessity of context in the putative functionality of non-B-DNA structures was shown by similar structure-forming potential in humans and chimpanzees in some regions in which there is a hotspot in humans but none in chimpanzees (Chapter 7). As noted, however, this should be investigated further (see Section 7.4), and the question of whether structures form only in the central regions of hotspots should be addressed by performing sodium bisulphite modification assays on fragments covering entire hotspots. It is of course possible, based on current knowledge, that non-B-DNA structures could themselves be involved in the context dependence of recombination hotspots, perhaps by constituting enhancer-like elements in hotspot flanking regions. Intramolecular quadruplexes can bind the nuclear matrix-associated type III intermediate filament proteins [53], suggesting that they could function in mediating higher-order chromosome structure, which could also be involved in hotspot regulation, so intramolecular structure-forming potential in hotspot flanking as well as central regions should be investigated. The ideal first target for this investigation is the *HIS2* hotspot in *S. cerevisiae*, because it has been shown that moving the hotspot sequence, including 5.2 kb of DNA surrounding it, to a different chromosomal location does not preserve hotspot activity [54], but moving a fragment extending 11.5 kb downstream from the *HIS2* gene is sufficient for hotspot activity at different chromosomal locations [33]. Other hotspot flanking regions shown to affect recombination frequency have been found in the mouse genome [40, 55].

It is also possible that simple sequences could function in recombination hotspots without forming non-B-DNA structures (see Sections 2.4 and 4.4). This possibility could be further explored using deletion studies similar to that which showed a poly-A tract to be a functional component of the yeast *ARG4* hotspot [32]. The motivation for such studies, and for further investigating the links between simple sequences and recombination hotspots in general, is to contribute to the elucidation of meiotic recombination and the mechanisms of its regulation, and to the understanding of its role in the stability, function and evolution of genomes.

References

1. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**(5746):321-324.
2. Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G: **The influence of recombination on human genetic diversity**. *PLoS Genet* 2006, **2**(9):e148.
3. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G *et al*: **A high-resolution recombination map of the human genome**. *Nat Genet* 2002, **31**(3):241-247.
4. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P *et al*: **Comparison of fine-scale recombination rates in humans and chimpanzees**. *Science* 2005, **308**(5718):107-111.
5. Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S: **Fine-scale recombination patterns differ between chimpanzees and humans**. *Nat Genet* 2005, **37**(4):429-434.
6. Coop G, Wen X, Ober C, Pritchard JK, Przeworski M: **High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans**. *Science* 2008, **319**(5868):1395-1398.
7. Neumann R, Jeffreys AJ: **Polymorphism in the activity of human crossover hotspots independent of local DNA sequence variation**. *Hum Mol Genet* 2006, **15**(9):1401-1411.
8. Iyer V, Struhl K: **Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure**. *Embo J* 1995, **14**(11):2570-2579.
9. Suter B, Schnappauf G, Thoma F: **Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo**. *Nucleic Acids Res* 2000, **28**(21):4083-4089.
10. Drew HR, Travers AA: **DNA bending and its relation to nucleosome positioning**. *J Mol Biol* 1985, **186**(4):773-790.
11. Hanvey JC, Klysik J, Wells RD: **Influence of DNA sequence on the formation of non-B right-handed helices in oligopurine.oligopyrimidine inserts in plasmids**. *J Biol Chem* 1988, **263**(15):7386-7396.
12. Nishant KT, Rao MR: **Molecular features of meiotic recombination hot spots**. *Bioessays* 2006, **28**(1):45-56.
13. Sandaltzopoulos R, Mitchelmore C, Bonte E, Wall G, Becker PB: **Dual regulation of the Drosophila hsp26 promoter in vitro**. *Nucleic Acids Res* 1995, **23**(13):2479-2487.
14. Lu Q, Teare JM, Granok H, Swede MJ, Xu J, Elgin SC: **The capacity to form H-DNA cannot substitute for GAGA factor binding to a (CT)n*(GA)n regulatory site**. *Nucleic Acids Res* 2003, **31**(10):2483-2494.
15. Kunkel GR, Martinson HG: **Nucleosomes will not form on double-stranded RNa or over poly(dA).poly(dT) tracts in recombinant DNA**. *Nucleic Acids Res* 1981, **9**(24):6869-6888.
16. Prunell A: **Nucleosome reconstitution on plasmid-inserted poly(dA) . poly(dT)**. *Embo J* 1982, **1**(2):173-179.
17. Elgin SC: **The formation and function of DNase I hypersensitive sites in the process of gene activation**. *J Biol Chem* 1988, **263**(36):19259-19262.

18. Gross DS, Garrard WT: **Nuclease hypersensitive sites in chromatin**. *Annu Rev Biochem* 1988, **57**:159-197.

19. Satchwell SC, Drew HR, Travers AA: **Sequence periodicities in chicken nucleosome core DNA**. *J Mol Biol* 1986, **191**(4):659-675.

20. Otten AD, Tapscott SJ: **Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure**. *Proc Natl Acad Sci U S A* 1995, **92**(12):5465-5469.

21. Kirkpatrick DT, Wang YH, Dominska M, Griffith JD, Petes TD: **Control of meiotic recombination and gene expression in yeast by a simple repetitive DNA sequence that excludes nucleosomes**. *Mol Cell Biol* 1999, **19**(11):7661-7671.

22. Hile SE, Eckert KA: **Positive correlation between DNA polymerase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences.** *J Mol Biol* 2004, **335**(3):745-759.

23. Hile SE, Eckert KA: **DNA polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellites sequences**. *Nucleic Acids Res* 2008, **36**(2):688-696.

24. Petes TD: **Meiotic recombination hot spots and cold spots**. *Nat Rev Genet* 2001, **2**(5):360-369.

25. Rooney SM, Moore PD: **Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells**. *Proc Natl Acad Sci U S A* 1995, **92**(6):2141-2144.

26. Wells RD, Collier DA, Hanvey JC, Shimizu M, Wohlrab F: **The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences**. *Faseb J* 1988, **2**(14):2939-2949.

27. Radhakrishnan I, Patel DJ: **DNA triplexes: solution structures, hydration sites, energetics, interactions, and function**. *Biochemistry* 1994, **33**(38):11405-11416.

28. Raghavan SC, Chastain P, Lee JS, Hegde BG, Houston S, Langen R, Hsieh CL, Haworth IS, Lieber MR: **Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation**. *J Biol Chem* 2005, **280**(24):22749-22760.

29. Dayn A, Samadashwily GM, Mirkin SM: **Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization**. *Proc Natl Acad Sci U S A* 1992, **89**(23):11406-11410.

30. Kohwi Y, Kohwi-Shigematsu T: **Altered gene expression correlates with DNA structure**. *Genes Dev* 1991, **5**(12B):2547-2554.

31. Gendrel CG, Boulet A, Dutreix M: **(CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis**. *Genes Dev* 2000, **14**(10):1261-1268.

32. Schultes NP, Szostak JW: **A poly(dA.dT) tract is a component of the recombination initiation site at the ARG4 locus in Saccharomyces cerevisiae**. *Mol Cell Biol* 1991, **11**(1):322-328.

33. Haring SJ, Halley GR, Jones AJ, Malone RE: **Properties of natural double-strand-break sites at a recombination hotspot in Saccharomyces cerevisiae**. *Genetics* 2003, **165**(1):101-114.

34. Kong A, al. E: **Sequence variants in the RNF212 Gene Associate with Genome-wide recombination rate**. *Science* 2008, **319**:1398-1401.

35. Arnheim N, Calabrese P, Tiemann-Boege I: **Mammalian Meiotic Recombination Hot Spots**. *Annu Rev Genet* 2007, **41**:369-399.

36. Jeffreys AJ, Neumann R: **Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot**. *Nat Genet* 2002, **31**(3):267-271.

37. Schuchert P, Langsford M, Kaslin E, Kohli J: **A specific DNA sequence is required for high frequency of recombination in the ade6 gene of fission yeast**. *Embo J* 1991, **10**(8):2157-2163.

38. Jeffreys AJ, Neumann R: **Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot**. *Hum Mol Genet* 2005, **14**(15):2277-2287.

39. Jeffreys AJ, Murray J, Neumann R: **High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot**. *Mol Cell* 1998, **2**(2):267-273.

40. Baudat F, de Massy B: **Cis- and trans-acting elements regulate the mouse Psmb9 meiotic recombination hotspot**. *PLoS Genet* 2007, **3**(6):e100.

41. Kon N, Krawchuk MD, Warren BG, Smith GR, Wahls WP: **Transcription factor Mts1/Mts2 (Atf1/Pcr1, Gad7/Pcr1) activates the M26 meiotic recombination hotspot in Schizosaccharomyces pombe**. *Proc Natl Acad Sci U S A* 1997, **94**(25):13765-13770.

42. Wahls WP, Smith GR: **A heteromeric protein that binds to a meiotic homologous recombination hot spot: correlation of binding and hot spot activity**. *Genes Dev* 1994, **8**(14):1693-1702.

43. Guedin A, De Cian A, Gros J, Lacroix L, Mergny JL: **Sequence effects in single-base loops for quadruplexes**. *Biochimie* 2008.

44. Bissler JD: **Triplex DNA and human disease**. *Frontiers in Bioscience* 2007, **12**:4536-4546.

45. Kovtun IV, McMurray CT: **Features of trinucleotide repeat instability in vivo**. *Cell Res* 2008, **18**(1):198-213.

46. Mirkin SM: **DNA structures, repeat expansions and human hereditary disorders**. *Curr Opin Struct Biol* 2006, **16**(3):351-358.

47. Lobachev KS, Rattray A, Narayanan V: **Hairpin- and cruciform-mediated chromosome breakage: causes and consequences in eukaryotic cells**. *Front Biosci* 2007, **12**:4208-4220.

48. Wang G, Vasquez KM: **Non-B DNA structure-induced genetic instability**. *Mutat Res* 2006, **598**(1-2):103-119.

49. Ulrich MJ, Gray WJ, Ley TJ: **An intramolecular triplex is disrputed by point mutations associated with hereditary persistence of fetal hemoglobin**. *Journal of Biological Chemistry* 1992, **267**:18649-18658.

50. Nakagama H, Higuchi K, Tanaka E, Tsuchiya N, Nakashima K, Katahira M, Fukuda H: **Molecular mechanisms for maintenance of G-rich short tandem repeats capable of adopting G4 DNA structures**. *Mutat Res* 2006, **598**(1-2):120-131.

51. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW *et al*: **Comparison of human genetic and sequence-based physical maps**. *Nature* 2001, **409**(6822):951-953.

52. Jensen-Seaman MI, Furey TS, Payseur BA, Lu Y, Roskin KM, Chen CF, Thomas MA, Haussler D, Jacob HJ: **Comparative recombination rates in the rat, mouse, and human genomes**. *Genome Res* 2004, **14**(4):528-538.

53. Li G, Tolstonog GV, Traub P: **Interaction in vitro of type III intermediate filament proteins with triplex DNA**. *DNA Cell Biol* 2002, **21**(3):163-188.

54. Malone RE, Kim S, Bullard SA, Lundquist S, Hutchings-Crow L, Cramton S, Lutfiyya L, Lee J: **Analysis of a recombination hotspot for gene conversion occurring at the HIS2 gene of Saccharomyces cerevisiae**. *Genetics* 1994, **137**(1):5-18.

55. Shiroishi T, Sagai T, Hanzawa N, Gotoh H, Moriwaki K: **Genetic control of sex-dependent meiotic recombination in the major histocompatibility complex of the mouse**. *Embo J* 1991, **10**(3):681-686.

Appendix A: Supporting information for Chapter 2

Table A1: Mean microsatellite repeat frequencies in intergenic regions (IGRs) throughout the *S. cerevisiae* genome. IGRs were divided by recombination (double-strand break) intensity as reported by Gerton and co-workers [1] into 473 hot, 89 cold and 5431 other regions, which were all IGRs not categorized as either hot or cold. The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus repeated motif. All p values < 0.01 are shown (no Bonferroni correction).

Repeat type			Mean repeat frequency			P value	
Motif length	Copy number	Mismatches allowed	Hot	Other	Cold	Hot v non hot	Cold v other
1 (A)	3 to 5	perfect	35.0	39.9	39.5	< 0.0001	n/s
		e=10	34.3	39.4	39.0	< 0.0001	n/s
		e=6	31.8	36.7	36.6	< 0.0001	n/s
	6+	perfect	5.42	4.63	3.90	< 0.0001	n/s
		e=10	5.24	4.51	3.87	< 0.0001	n/s
		e=6	6.12	5.54	4.86	0.00173	n/s
	14+	perfect	0.418	0.172	0.0663	< 0.0001	n/s
		e=10	0.733	0.315	0.0841	< 0.0001	n/s
		e=6	0.854	0.382	0.0841	< 0.0001	0.0447
1 (G)	3 to 5	perfect	9.18	7.26	6.47	< 0.0001	n/s
		e=10	9.16	7.25	6.49	< 0.0001	n/s
		e=6	8.89	7.14	6.48	< 0.0001	n/s
	6+	perfect	0.118	0.0744	0.0934	0.0109	n/s
		e=10	0.114	0.0731	0.0934	0.0185	n/s
		e=6	0.160	0.0931	0.0934	0.00124	n/s
	14+	perfect	0.00350	0.000737	0	0.00179	n/s
		e=10	0.00350	0.000737	0	0.00179	n/s
		e=6	0.00350	0.000737	0	0.00179	n/s
2 (AT)	2	perfect	7.687	9.243	9.82	< 0.0001	n/s
		e=10	7.568	9.102	9.68	< 0.0001	n/s
		e=6	6.261	7.593	8.39	< 0.0001	n/s
	3 to 5	perfect	2.685	2.644	2.09	n/s	n/s
		e=10	2.420	2.481	2.03	n/s	n/s
		e=6	3.028	3.272	3.22	n/s	n/s
	6+	perfect	0.308	0.178	0.0398	0.00172	n/s
		e=10	0.450	0.242	0.0989	< 0.0001	n/s
		e=6	0.627	0.460	0.174	0.00155	n/s
	10+	perfect	0.142	0.0575	0	0.0124	n/s
		e=10	0.197	0.0898	0.00404	0.00586	n/s
		e=6	0.221	0.102	0.00404	0.00202	n/s
2 (AC)	2	perfect	6.80	6.60	6.48	n/s	n/s
		e=10	6.69	6.55	6.40	n/s	n/s
		e=6	6.13	6.01	5.95	n/s	n/s
	3 to 5	perfect	0.908	0.584	0.772	< 0.0001	n/s
		e=10	0.924	0.569	0.772	< 0.0001	n/s
		e=6	1.32	0.979	1.04	0.000335	n/s
	6+	perfect	0.0518	0.0155	0	0.0110	n/s
		e=10	0.0772	0.0211	0	0.00653	n/s
		e=6	0.134	0.0453	0.0499	0.000973	n/s

Table A1 continued

Repeat type			Mean repeat frequency			P value	
Motif length	Copy number	Mismatches allowed	Hot	Other	Cold	Hot v non hot	Cold v other
2 (AC)	10+	perfect	0.0159	0.00339	0	n/s	n/s
		e=10	0.0218	0.00375	0	n/s	n/s
		e=6	0.0283	0.00758	0	n/s	n/s
2 (AG)	2	perfect	7.57	7.03	7.21	0.0286	n/s
		e=10	7.53	7.01	7.19	0.0332	n/s
		e=6	6.73	6.32	6.76	n/s	n/s
	3 to 5	perfect	0.940	0.653	0.635	0.000249	n/s
		e=10	0.918	0.641	0.635	0.000261	n/s
		e=6	1.61	1.15	0.826	< 0.0001	n/s
	6+	perfect	0.00828	0.00381	0.0207	n/s	0.0351
		e=10	0.0196	0.00960	0.0207	n/s	n/s
		e=6	0.0354	0.0267	0.0914	n/s	0.00391
	10+	perfect	0	0.00065	0	n/s	n/s
		e=10	0.00307	0.00108	0	n/s	n/s
		e=6	0.00307	0.00120	0	n/s	n/s
2 (CG)	2	perfect	1.76	1.43	1.12	0.000294	n/s
		e=10	1.76	1.43	1.12	0.000263	n/s
		e=6	1.64	1.38	1.06	0.00149	n/s
	3 to 5	perfect	0.132	0.0810	0.112	0.0129	n/s
		e=10	0.132	0.0809	0.112	0.0121	n/s
		e=6	0.213	0.122	0.112	< 0.0001	n/s
	6+	perfect	0	0	0	n/s	n/s
		e=10	0	< 0.0001	0	n/s	n/s
		e=6	0	< 0.0001	0	n/s	n/s
	10+	perfect	0	0	0	n/s	n/s
		e=10	0	0	0	n/s	n/s
		e=6	0	0	0	n/s	n/s
2 (all)	2	perfect	23.8	24.3	24.6	n/s	n/s
		e=10	23.5	24.1	24.4	n/s	n/s
		e=6	20.8	21.3	22.2	n/s	n/s
	3 to 5	perfect	4.67	3.96	3.61	< 0.0001	n/s
		e=10	4.34	3.69	3.55	0.000266	n/s
		e=6	6.17	5.52	5.20	0.000234	n/s
	6+	perfect	0.368	0.198	0.0605	0.000248	n/s
		e=10	0.599	0.360	0.125	< 0.0001	n/s
		e=6	0.797	0.532	0.316	< 0.0001	n/s
	10+	perfect	0.158	0.0615	0	0.014492	n/s
		e=10	0.221	0.0946	0.00404	0.00164	n/s
		e=6	0.252	0.110	0.00404	0.00132	n/s

Table A1 continued

Repeat type			Mean repeat frequency			P value	
Motif length	Copy number	Mismatches allowed	Hot	Other	Cold	Hot v non hot	Cold v other
3 (all)		e=6	1.97	1.93	1.32	n/s	0.0365
	6+	perfect	0.0460	0.0126	0	0.0134	n/s
		e=10	0.0627	0.0291	0	n/s	n/s
		e=6	0.109	0.0533	0	0.00270	n/s
	10+	perfect	0.00970	0.00559	0	n/s	n/s
		e=10	0.0215	0.0105	0	n/s	n/s
		e=6	0.0215	0.0116	0	n/s	n/s
4 (all)	2	perfect	4.36	3.97	4.17	0.0462	n/s
		e=10	4.17	3.77	3.95	0.0483	n/s
		e=6	3.44	3.04	3.28	0.0485	n/s
	3 to 5	perfect	0.151	0.107	0.118	0.00311	n/s
		e=10	0.274	0.286	0.248	n/s	n/s
		e=6	0.329	0.368	0.402	n/s	n/s
	6+	perfect	0	0.00509	0	n/s	n/s
		e=10	0	0.00589	0	n/s	n/s
		e=6	0	0.00865	0	n/s	n/s
	10+	perfect	0	0.00157	0	n/s	n/s
		e=10	0	0.00157	0	n/s	n/s
		e=6	0	0.00157	0	n/s	n/s
5 (all)	2	perfect	1.72	1.58	1.38	n/s	n/s
		e=10	1.63	1.45	1.34	0.102	n/s
		e=6	1.28	1.12	1.04	n/s	n/s
	3 to 5	perfect	0.0482	0.0356	0.0205	n/s	n/s
		e=10	0.0999	0.103	0.0732	n/s	n/s
		e=6	0.0867	0.0992	0.0732	n/s	n/s
	6+	perfect	0	0.00082	0	n/s	n/s
		e=10	0.00137	0.00106	0	n/s	n/s
		e=6	0.00137	0.00267	0	n/s	n/s
	10+	perfect	0	0	0	n/s	n/s
		e=10	0	0	0	n/s	n/s
		e=6	0	0	0	n/s	n/s
6 (all)	2	perfect	0.811	0.654	0.564	0.0291	n/s
		e=10	0.727	0.575	0.534	0.0399	n/s
		e=6	0.518	0.394	0.363	0.0302	n/s
	3 to 5	perfect	0.0491	0.0233	0.0109	n/s	n/s
		e=10	0.0509	0.0440	0.0412	n/s	n/s
		e=6	0.0409	0.0290	0.0412	n/s	n/s
	6+	perfect	0.00552	0.000628	0	n/s	n/s
		e=10	0.00552	0.00459	0	n/s	n/s
		e=6	0.00552	0.00347	0	n/s	n/s
	10+	perfect	0	0	0	n/s	n/s
		e=10	0	0.00181	0	n/s	n/s
		e=6	0	0.00089	0	n/s	n/s

Table A2: Correlations between DSB intensity and microsatellite repeat frequencies for all IGRs in the *S. cerevisiae* genome. Showing non-parametric correlation coefficients and p values (Spearman's rho) correlating microsatellite frequency (MF) and DSB intensity. Partial correlations are also shown, controlling for GC content (GC), and transcriptional frequency (TF), which was the mean of the two adjacent ORFs. The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus repeated motif. All p values <0.01 are shown (no Bonferroni correction).

Repeat type			**MF v DSB intensity**		**Controlling for GC**		**Controlling for TF**	
Motif length	Copy numb.	Mismatch type	Coeff.	P value	Coeff.	P value	Coeff.	P value
1 (A)	3 to 5	perfect	-0.137	<.0001	-0.0948	<.0001	-0.14	<.0001
		e=10	-0.142	<.0001	-0.102	<.0001	-0.144	<.0001
		e=6	-0.152	<.0001	-0.12	<.0001	-0.152	<.0001
	6+	perfect	0.0743	<.0001	0.122	<.0001	0.0615	<.0001
		e=10	0.0726	<.0001	0.119	<.0001	0.0601	<.0001
		e=6	0.0662	<.0001	0.119	<.0001	0.0537	<.0001
	14+	perfect	0.0716	<.0001	0.0804	<.0001	0.0705	<.0001
		e=10	0.111	<.0001	0.123	<.0001	0.107	<.0001
		e=6	0.114	<.0001	0.127	<.0001	0.11	<.0001
1 (G)	3 to 5	perfect	0.132	<.0001	0.0638	<.0001	0.132	<.0001
		e=10	0.131	<.0001	0.0623	<.0001	0.131	<.0001
		e=6	0.125	<.0001	0.057	<.0001	0.125	<.0001
	6+	perfect	0.038	0.0034	0.016	n/s	0.0471	0.0004
		e=10	0.0389	0.0027	0.0169	n/s	0.0478	0.0003
		e=6	0.0586	<.0001	0.0325	n/s	0.0652	<.0001
	14+	perfect	0.0214	n/s	0.016	n/s	0.0197	n/s
		e=10	0.0214	n/s	0.016	n/s	0.0197	n/s
		e=6	0.0214	n/s	0.016	n/s	0.0197	n/s
2 (AT)	2	perfect	-0.118	<.0001	-0.0778	<.0001	-0.125	<.0001
		e=10	-0.124	<.0001	-0.0861	<.0001	-0.13	<.0001
		e=6	-0.115	<.0001	-0.0883	<.0001	-0.117	<.0001
	3 to 5	perfect	-0.00831	n/s	0.0212	n/s	-0.0136	n/s
		e=10	-0.0133	n/s	0.0112	n/s	-0.0171	n/s
		e=6	-0.034	0.0086	-0.0005	n/s	-0.0401	0.0024
	6+	perfect	0.0437	0.0007	0.0555	<.0001	0.0439	0.0009
		e=10	0.0395	0.0023	0.0511	<.0001	0.037	0.0052
		e=6	0.0257	n/s	0.0442	0.0007	0.0217	n/s
	10+	perfect	0.0278	n/s	0.0378	0.0035	0.029	n/s
		e=10	0.0393	0.0024	0.0501	0.0001	0.0417	0.0016
		e=6	0.0447	0.0006	0.0547	<.0001	0.047	0.0004
2 (AC)	2	perfect	0.0067	n/s	-0.0227	n/s	0.0124	n/s
		e=10	0.00313	n/s	-0.0261	n/s	0.00894	n/s
		e=6	-0.00982	n/s	-0.0369	0.0044	-0.00587	n/s
	3 to 5	perfect	0.0678	<.0001	0.0378	0.0035	0.0671	<.0001
		e=10	0.0624	<.0001	0.033	n/s	0.0615	<.0001
		e=6	0.0739	<.0001	0.0414	0.0014	0.0745	<.0001
	6+	perfect	0.029	n/s	0.0225	n/s	0.0327	n/s
		e=10	0.0339	0.0089	0.026	n/s	0.0427	0.0012
		e=6	0.0579	<.0001	0.0443	0.0006	0.0633	<.0001

Table A2 continued

Repeat type			**MF v DSB intensity**		**Controlling for GC**		**Controlling for TF**	
Motif length	Copy numb.	Mismatch type	Coeff.	P value	Coeff.	P value	Coeff.	P value
2 (AC)	10+	perfect	0.0154	n/s	0.0139	n/s	0.0196	n/s
		e=10	0.0171	n/s	0.0151	n/s	0.0204	n/s
		e=6	0.045	0.0005	0.0356	0.006	0.0512	0.0001
2 (AG)	2	perfect	0.0264	n/s	-0.0138	n/s	0.026	n/s
		e=10	0.026	n/s	-0.014	n/s	0.0255	n/s
		e=6	0.0188	n/s	-0.0187	n/s	0.0185	n/s
	3 to 5	perfect	0.051	<.0001	0.0213	n/s	0.0536	<.0001
		e=10	0.0497	0.0001	0.0201	n/s	0.0519	<.0001
		e=6	0.0677	<.0001	0.0335	0.0097	0.0723	<.0001
	6+	perfect	0.0388	0.0028	0.0332	n/s	0.0369	0.0053
		e=10	0.0554	<.0001	0.0478	0.0002	0.0528	<.0001
		e=6	0.0472	0.0003	0.038	0.0033	0.0476	0.0003
	10+	perfect	0.0239	n/s	0.0226	n/s	0.0237	n/s
		e=10	0.0355	0.0061	0.0317	n/s	0.0363	0.0061
		e=6	0.0383	0.0031	0.0335	0.0097	0.0387	0.0034
2 (CG)	2	perfect	0.0615	<.0001	0.00827	n/s	0.0613	<.0001
		e=10	0.0609	<.0001	0.0077	n/s	0.0607	<.0001
		e=6	0.0533	<.0001	0.00138	n/s	0.0549	<.0001
	3 to 5	perfect	0.0537	<.0001	0.0282	n/s	0.0606	<.0001
		e=10	0.0529	<.0001	0.0275	n/s	0.0606	<.0001
		e=6	0.0787	<.0001	0.0466	0.0003	0.0823	<.0001
	6+	perfect	n/a	n/a	n/a	n/a	n/a	n/a
		e=10	0.0153	n/s	0.0129	n/s	n/a	n/a
		e=6	0.0153	n/s	0.0129	n/s	n/a	n/a
	10+	perfect	n/a	n/a	n/a	n/a	n/a	n/a
		e=10	n/a	n/a	n/a	n/a	n/a	n/a
		e=6	n/a	n/a	n/a	n/a	n/a	n/a
2 (all)	2	perfect	-0.0375	0.0038	-0.0532	<.0001	-0.0372	0.005
		e=10	-0.0428	0.001	-0.0595	<.0001	-0.0425	0.0013
		e=6	-0.0484	0.0002	-0.0723	<.0001	-0.0459	0.0005
	3 to 5	perfect	0.0426	0.001	0.0555	<.0001	0.039	0.0032
		e=10	0.0375	0.0038	0.0459	0.0004	0.0346	0.0089
		e=6	0.0405	0.0018	0.0508	<.0001	0.0393	0.003
	6+	perfect	0.0559	<.0001	0.0643	<.0001	0.0561	<.0001
		e=10	0.0565	<.0001	0.0614	<.0001	0.0532	<.0001
		e=6	0.0522	<.0001	0.0633	<.0001	0.0485	0.0002
	10+	perfect	0.0341	0.0085	0.0431	0.0009	0.0361	0.0064
		e=10	0.0468	0.0003	0.0562	<.0001	0.0497	0.0002
		e=6	0.0634	<.0001	0.0687	<.0001	0.067	<.0001
3 (all)	2	perfect	-0.0105	n/s	-0.0173	n/s	-0.0123	n/s
		e=10	-0.0176	n/s	-0.0261	n/s	-0.0192	n/s
		e=6	-0.0282	n/s	-0.0445	0.0006	-0.0282	n/s

Table A2 continued

Repeat type			MF v DSB intensity		Controlling for GC		Controlling for TF	
Motif length	Copy numb.	Mismatch type	Coeff.	P value	Coeff.	P value	Coeff.	P value
3 (all)	3 to 5	perfect	0.0484	0.0002	0.0417	0.0013	0.0464	0.0005
		e=10	0.046	0.0004	0.0377	0.0037	0.0439	0.0009
		e=6	0.0342	0.0083	0.0353	0.0065	0.0253	n/s
	6+	perfect	0.0212	n/s	0.0269	n/s	0.0238	n/s
		e=10	0.0312	n/s	0.0364	0.005	0.0306	n/s
		e=6	0.0489	0.0002	0.0539	<.0001	0.0462	0.0005
	10+	perfect	-0.00896	n/s	-0.00192	n/s	-0.00829	n/s
		e=10	0.0237	n/s	0.0336	0.0096	0.0218	n/s
		e=6	0.0292	n/s	0.039	0.0026	0.0268	n/s
4 (all)	2	perfect	0.0306	n/s	0.0376	0.0037	0.0232	n/s
		e=10	0.0298	n/s	0.0344	0.0079	0.023	n/s
		e=6	0.017	n/s	0.0132	n/s	0.00953	n/s
	3 to 5	perfect	0.0397	0.0022	0.04	0.002	0.0422	0.0014
		e=10	0.0439	0.0007	0.044	0.0007	0.0417	0.0016
		e=6	0.0463	0.0003	0.0473	0.0003	0.0424	0.0014
	6+	perfect	-0.00019	n/s	0.00273	n/s	0.00078	n/s
		e=10	0.00675	n/s	0.0113	n/s	-0.00017	n/s
		e=6	0.0231	n/s	0.0257	n/s	0.0178	n/s
	10+	perfect	0.00595	n/s	0.00966	n/s	0.00481	n/s
		e=10	0.00595	n/s	0.00966	n/s	0.00481	n/s
		e=6	0.00595	n/s	0.00966	n/s	0.00481	n/s
5 (all)	2	perfect	0.0357	0.0059	0.0354	0.0062	0.0242	n/s
		e=10	0.031	n/s	0.0279	n/s	0.0194	n/s
		e=6	0.024	n/s	0.0157	n/s	0.0148	n/s
	3 to 5	perfect	0.0294	n/s	0.0215	n/s	0.0324	n/s
		e=10	0.0366	0.0047	0.0309	n/s	0.0333	n/s
		e=6	0.0331	n/s	0.028	n/s	0.0328	n/s
	6+	perfect	-0.0273	n/s	-0.0274	n/s	-0.0292	n/s
		e=10	-0.0171	n/s	-0.0215	n/s	-0.0146	n/s
		e=6	0.00077	n/s	0.00032	n/s	-0.00104	n/s
	10+	perfect	n/a	n/a	n/a	n/a	n/a	n/a
		e=10	n/a	n/a	n/a	n/a	n/a	n/a
		e=6	n/a	n/a	n/a	n/a	n/a	n/a
6 (all)	2	perfect	0.044	0.0007	0.0386	0.0029	0.0457	0.0005
		e=10	0.0348	0.0072	0.0289	n/s	0.0377	0.0044
		e=6	0.0282	n/s	0.0157	n/s	0.0344	0.0093
	3 to 5	perfect	0.0299	n/s	0.0233	n/s	0.0337	n/s
		e=10	0.0278	n/s	0.0234	n/s	0.0305	n/s
		e=6	0.0232	n/s	0.0218	n/s	0.0309	n/s
	6+	perfect	0.0197	n/s	0.0166	n/s	0.0206	n/s
		e=10	0.0248	n/s	0.0131	n/s	0.0303	n/s
		e=6	0.00238	n/s	-0.00224	n/s	-0.00122	n/s
	10+	perfect	n/a	n/a	n/a	n/a	n/a	n/a
		e=10	0.013	n/s	0.0042	n/s	0.0234	n/s
		e=6	0.00502	n/s	-0.00078	n/s	0.00154	n/s

Table A3: Mean lengths of microsatellites of at least six copies in IGRs throughout the yeast genome. IGRs were divided by recombination (double-strand break) intensity as reported by Gerton and co-workers [1] into 473 hot, 89 cold and 5431 other regions, which were all IGRs not categorized as either hot or cold. The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus repeated motif. All p values <0.01 are shown, but caution is recommended in view of the multiple hypotheses being tested.

Repeat type		Mean repeat copy number (6-copy repeats and longer) and total number of microsatellties by IGR type (N)						P value	
Motif length	Mismatch type	Hot		Other		Cold		Hot v non-hot	Cold v other
		Mean	N	mean	N	mean	N		
1 (A)	perfect	8.24	1174	7.63	11388	7.19	240	< 0.0001	n/s
	e=10	8.67	1236	7.87	12025	7.37	237	< 0.0001	n/s
	e=6	9.26	1473	8.53	14870	8.05	294	< 0.0001	n/s
1 (G)	perfect	6.52	31	6.37	232	6.43	7	n/s	n/s
	e=10	7.16	32	6.44	233	6.43	7	0.0059	n/s
	e=6	8.09	46	7.13	298	7	7	0.0038	n/s
2 (AT)	perfect	8.7	46	8.89	308	8.5	2	n/s	n/s
	e=10	9.65	66	9.74	429	8.13	4	n/s	n/s
	e=6	8.8	99	8.41	836	7.25	8	n/s	n/s
2 (AC)	perfect	8.13	8	9.19	37	n/a	0	n/s	n/a
	e=10	8.59	11	9.13	47	n/a	0	n/s	n/a
	e=6	7.73	22	15.93	108	6.25	2	n/s	n/s
2 (AG)	perfect	6.67	3	10.67	9	8	1	n/s	n/s
	e=10	8.7	5	8.32	25		0	n/s	n/s
	e=6	7.5	9	7.24	63	6.75	4	n/s	n/s
2 (CG)	e=10	n/a	0	6.5	1	n/a	0	n/a	n/a
	e=6	n/a	0	6.5	1	n/a	0	n/a	n/a
2 (all)	perfect	8.51	57	8.96	354	8.33	3	n/s	n/s
	e=10	9.45	82	9.6	502	8.3	5	n/s	n/s
	e=6	8.53	130	9.14	1008	6.96	14	n/s	n/s
3 (all)	perfect	7.86	7	9.89	27	n/a	0	n/s	n/a
	e=10	9.24	11	9.36	66	n/a	0	n/s	n/a
	e=6	8.3	21	8.41	118	n/a	0	n/s	n/a
4 (all)	perfect	n/a	0	7.8	5	n/a	0	n/a	n/a
	e=10	n/a	0	8	12	n/a	0	n/a	n/a
	e=6	n/a	0	7.88	19	n/a	0	n/a	n/a
5 (all)	perfect	n/a	0	6.5	2	n/a	0	n/a	n/a
	e=10	6.6	1	6.95	4	n/a	0	n/a	n/a
	e=6	6.6	1	6.8	5	n/a	0	n/a	n/a
6 (all)	perfect	6	1	7	3	n/a	0	n/s	n/a
	e=10	6	1	9.65	21	n/a	0	n/s	n/a
	e=6	6	1	9.67	10	n/a	0	n/s	n/a

Table A4: The five most common multiply represented trinucleotide repeat motifs in each type of region. Poly-purine/poly-pyrimidine motifs are emboldened. Perfect repeats only were considered for this analysis. Division of regions was as for Table A1, i.e. regions not classed as hot or cold are denoted “other”.

2 copy repeats			**3 to 5 copy repeats**			**6+ copy repeats**		
Region type	Motif	N	Region type	Motif	N	Region type	Motif	N
hot ORFs	**TTC**	170	hot ORFs	**TCT**	14	hot ORFs	TTG	2
	GAA	169		**AAG**	13			
	CAA	156		AGC	13			
	CTT	150		**GAA**	13			
	TTG	146		AAC	12			
				AGA	12			
				TCA	12			
hot IGRs	AAT	139	hot IGRs	TAT	15	hot IGRs	TAT	4
	TAT	131		ATA	11		AAT	1
	AAG	124		ATT	10		ATT	1
	TTC	123		AAT	9		TTA	1
	GAA	122		**TTC**	9			
other ORFs	**TTC**	4787	other ORFs	**TTC**	353	other ORFs	CAG	12
	GAA	4778		**GAA**	351		TCA	12
	AAG	4017		TCA	264		**TTC**	12
	AAT	3992		**AAG**	256		CAA	10
	ATT	3743		**TCT**	251		TGT	10
other IGRs	AAT	1896	other IGRs	TAT	153	other IGRs	TAT	6
	TAT	1697		ATA	138		ATA	3
	ATT	1604		AAT	127		TAA	3
	ATA	1562		TAA	110		AAT	2
	TTA	1408		TTA	100		CAA	2
							GAA	2
							TAG	2
cold ORFs	ATT	47	cold ORFs	TGC	6			
	TTC	47		GAT	5			
	TTA	43		**TCT**	5			
	TGA	42		**AAG**	4			
	TCT	41		**GAA**	4			
				TTG	4			
cold IGRs	ATT	48	cold IGRs	TAT	4			
	AAT	46		TAA	3			
	TTC	38		ATA	3			
	TTA	35		ATT	2			
	AAG	33		**GAA**	2			

Table A5: The five most common multiply represented tetranucleotide repeat motifs in each type of region. Poly purine/poly pyrimidine motifs are emboldened. Perfect repeats only were considered for this analysis. Division of regions was as for Table A1, i.e. regions not classed as hot or cold are denoted “other”.

2 copy repeats			**3 to 5 copy repeats**			**6+ copy repeats**		
Region type	Motif	N	Region type	Motif	N	Region type	Motif	N
hot ORFs	**AAAG**	19						
	TCTT	16						
	ATTT	15						
	TTTG	15						
	TTGT	14						
hot IGRs	**TTTC**	31	hot IGRs	AAAT	4			
	AAAG	30		**TTTC**	4			
	GAAA	27		**AAGA**	3			
	TATG	23		ATAC	2			
	AAAT	21		ATTT	2			
	CTTT	21		GTAT	2			
	TTAT	21		TTTA	2			
other ORFs	**TTTC**	506	other ORFs	**CTTT**	11			
	AAAG	445		**AAAG**	9			
	AAGA	416		**TTTC**	8			
	GAAA	407		**AGAA**	6			
	AAAT	385		**GAAA**	6			
other IGRs	**TTTC**	308	other IGRs	TTTA	22	other IGRs	AATA	2
	AAAT	282		ATAA	16			
	AAAG	279		AAAT	15			
	TTTA	272		AATA	15			
	TATT	270		TATT	12			
cold ORFs	**AGAA**	9						
	TTTG	8						
	TCTT	7						
	AAAG	6						
	TTTC	5						
cold IGRs	AAAT	8	cold IGRs	ATAA	2			
	ATTT	8						
	TTAT	8						
	AAAG	7						
	AATA	7						

Table A6: The five most common multiply represented pentanucleotide repeat motifs in each type of region. Poly purine/poly pyrimidine motifs are emboldened. Perfect repeats only were considered for this analysis. Division of regions was as for Table A1, i.e. regions not classed as hot or cold are denoted "other".

2 copy repeats			**3 to 5 copy repeats**			**6+ copy repeats**		
Region type	Motif	N	Region type	Motif	N	Region type	Motif	N
hot ORFs	**GAAAA**	6						
	TTCTT	4						
	AAAGA	3						
	AAGAA	3						
	AAGCA	3						
	CAGAG	3						
	CATTC	3						
	TCTTC	3						
hot IGRs	**TTTTC**	15						
	AGAAA	10						
	GAAAA	10						
	AAAAG	8						
	AAAAT	7						
	TTTCT	7						
other ORFs	**TTTTC**	104	other ORFs	**CTTTT**	3			
	AAGAA	84		GGTGT	2			
	TTCTT	84		TTTGT	2			
	GAAAA	82						
	AAAAT	79						
other IGRs	**TTTTC**	91	other IGRs	CACAC	6			
	AAAAG	86		**GAAAA**	5			
	AAAAT	82		GATGA	5			
	ATATA	58		ATAAT	4			
	AAGAA	57		GGTGT	4			
	ATTTT	57		**TTTTC**	4			
cold ORFs	**TTTCT**	3						
	ACCAA	2						
	AGAAT	2						
	GAAAA	2						
	TCAAA	2						
	TGAAT	2						
	TTCTG	2						
	TTTCC	2						
	TTTTC	2						
	TTTTG	2						
cold IGRs	ATTTT	3						
	AAAAG	2						
	AAAGG	2						
	ATAAA	2						
	ATCTT	2						
	CTAAA	2						
	CTTTT	2						
	TTATA	2						

Table A7: The five most common multiply represented hexanucleotide repeat motifs in each type of region. Poly purine/poly pyrimidine motifs are emboldened. Perfect repeats only were considered for this analysis. Division of regions was as for Table A1, i.e. regions not classed as hot or cold are denoted "other".

2 copy repeats			**3 to 5 copy repeats**			**6+ copy repeats**		
Region type	Motif	N	Region type	Motif	N	Region type	Motif	N
hot ORFs	ACCACT	3						
	CAACAG	3						
	GAAGAT	3						
	CTTTTT	2						
	GACGAA	2						
	TATACA	2						
	TCTTCG	2						
	TTCAGT	2						
	TTCGTC	2						
hot IGRs	**AAGAAA**	5						
	AGAAAA	4						
	TATACA	4						
	TTTTTC	4						
	AAAAGA	3						
	TTTCTT	3						
	TTTTCT	3						
	TTTTTA	3						
other ORFs	AGATGA	31	other ORFs	CAGCAA	6			
	AAAGAA	26		TGTTGC	5			
	TTCATC	24		GTTGCT	4			
	AAGAAA	22		TGCTGT	4			
	TTTTTC	21		GATGAA	3			
				TGTGCT	3			
other IGRs	**TTTTTC**	34	other IGRs	CCACAC	14	other IGRs	GGTGTG	2
	GTGTGG	33		GTGTGG	10			
	CCACAC	30		AAAACA	3			
	TTTCTT	30		**AAAGAA**	3			
	AAAAAG	29		**CTTTTT**	2			
				GCGGAA	2			
				GGTGTG	2			
				TATATG	2			
cold ORFs	ACCGAG	5						
cold IGRs	**GAAAAA**	2						
	TGTTTT	2						

Table A8: Mean GC-content of microsatellites with at least six copies for all IGRs in the *S. cerevisiae* genome. IGRs were divided by recombination (double-strand break) intensity as reported by Gerton and co-workers [1] into 473 hot, 89 cold and 5431 other regions, which were all IGRs not categorized as either hot or cold. The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus repeated motif.

Repeat type		**Mean repeat GC content and total number of repeats (n)**						**P value**	
Motif length	Mismatch type	Hot		Other		Cold		Hot v non-hot	Cold v other
		Mean	n	Mean	n	Mean	n		
3 (all)	perfect	0	7	0.148	27	n/a	0	n/s	n/a
	e=10	0.094	11	0.191	66	n/a	0	n/s	n/a
	e=6	0.192	21	0.17	118	n/a	0	n/s	n/a
4 (all)	perfect	n/a	0	0.05	5	n/a	0	n/a	n/a
	e=10	n/a	0	0.082	12	n/a	0	n/a	n/a
	e=6	n/a	0	0.11	19	n/a	0	n/a	n/a
5 (all)	perfect	n/a	0	0.4	2	n/a	0	n/a	n/a
	e=10	0.457	1	0.508	4	n/a	0	n/a	n/a
	e=6	0.457	1	0.222	5	n/a	0	n/a	n/a
6 (all)	perfect	0.167	1	0.5	3	n/a	0	n/s	n/a
	e=10	0.189	1	0.582	21	n/a	0	n/s	n/a
	e=6	0.189	1	0.361	10	n/a	0	n/s	n/a

Table A9: The influence of promoter regions. Showing mean per kb frequencies of short tandem repeats for all IGRs in the *S. cerevisiae* genome divided according to the number of promoters they contain into 1537 with no promoters, 2894 with one and 1530 with two. P values are for Kruskal Wallis non- parametric ANOVA. The e value denotes the number of bases in any part of a repeat within which no more than one mismatch was allowed with respect to the consensus repeated motif. All p values <0.01 are shown, but caution is recommended in view of the multiple hypotheses being tested.

Repeat type			**Mean per kb repeat freq. in hot IGRs**				**Mean per kb repeat freq. in non-hot IGRs**			
Motif length	Copy numb.	Mismatch type	Number of promoters			P value	Number of promoters			P value
			None	One	Two		None	One	Two	
1 (A)	3 to 5	perfect	35	35.5	34.2	n/s	41	39.8	39.1	< 0.0001
		e=10	33.6	34.9	33.8	n/s	40.5	39.4	38.7	0.00028
		e=6	30.9	32.3	31.6	n/s	37.1	36.7	36.3	n/s
	6+	perfect	6.71	5.51	4.36	0.00289	5.12	4.61	4.15	0.00037
		e=10	6.49	5.33	4.21	0.00371	4.99	4.49	4.07	0.00144
		e=6	7.96	6.04	4.98	0.00037	6.09	5.53	5.01	< 0.0001
	14+	perfect	0.503	0.473	0.261	n/s	0.226	0.174	0.111	n/s
		e=10	1.08	0.692	0.564	n/s	0.437	0.294	0.22	n/s
		e=6	1.25	0.821	0.632	n/s	0.495	0.37	0.277	n/s
1 (G)	3 to 5	perfect	6.62	9.71	10	< 0.0001	5.24	7.58	8.42	< 0.0001
		e=10	6.62	9.68	10	< 0.0001	5.22	7.57	8.43	< 0.0001
		e=6	6.36	9.4	9.76	< 0.0001	5.16	7.44	8.29	< 0.0001
	6+	perfect	0.0788	0.133	0.117	n/s	0.0618	0.0744	0.0904	< 0.0001
		e=10	0.0788	0.133	0.103	n/s	0.0571	0.0744	0.0904	< 0.0001
		e=6	0.156	0.175	0.138	n/s	0.0652	0.0959	0.119	< 0.0001
	14+	perfect	0	0	0.0121	n/s	0	0.00151	0	n/s
		e=10	0	0	0.0121	n/s	0	0.00151	0	n/s
		e=6	0	0	0.0121	n/s	0	0.00151	0	n/s
2 (AT)	2	perfect	9.8	7.27	6.94	n/s	11.3	9.02	7.64	< 0.0001
		e=10	9.46	7.22	6.85	n/s	11	8.91	7.58	< 0.0001
		e=6	7.14	6.16	5.83	n/s	8.82	7.5	6.59	< 0.0001
	3 to 5	perfect	4.3	2.48	1.91	n/s	4.21	2.41	1.47	< 0.0001
		e=10	3.87	2.22	1.76	n/s	3.9	2.29	1.39	< 0.0001
		e=6	4.6	2.75	2.41	n/s	4.94	3.07	1.98	< 0.0001
	6+	perfect	0.882	0.153	0.177	n/s	0.377	0.137	0.0478	< 0.0001
		e=10	1.11	0.286	0.276	n/s	0.497	0.183	0.0871	< 0.0001
		e=6	1.45	0.485	0.3	n/s	0.956	0.352	0.145	< 0.0001
	10+	perfect	0.546	0.0378	0.043	n/s	0.137	0.0391	0.00826	0.00027
		e=10	0.678	0.0639	0.0915	n/s	0.208	0.0591	0.0224	0.00012
		e=6	0.684	0.0879	0.129	n/s	0.222	0.0737	0.026	0.00103
2 (AC)	2	perfect	6.14	6.81	7.26	n/s	6.33	6.6	6.86	< 0.0001
		e=10	5.92	6.69	7.21	n/s	6.28	6.57	6.84	< 0.0001
		e=6	5.46	6.04	6.78	n/s	5.8	6.03	6.22	< 0.0001
	3 to 5	perfect	0.964	0.922	0.844	n/s	0.574	0.573	0.611	< 0.0001
		e=10	1.03	0.926	0.844	n/s	0.559	0.563	0.603	< 0.0001
		e=6	1.45	1.31	1.27	n/s	0.934	0.966	1.06	< 0.0001
	6+	perfect	0.0755	0.0461	0.0452	n/s	0.0183	0.0138	0.00924	n/s
		e=10	0.151	0.0659	0.0452	n/s	0.0276	0.021	0.0118	n/s
		e=6	0.332	0.0937	0.0671	n/s	0.046	0.0387	0.04	n/s

Table A9 continued

Repeat type			Mean per kb repeat freq. in hot IGRs				Mean per kb repeat freq. in non-hot IGRs			
Motif length	Copy numb.	Mismatch type	Number of promoters			P value	Number of promoters			P value
			None	One	Two		None	One	Two	
2 (AC)	10+	perfect	0.0755	0	0.00203	n/s	0.00458	0.00401	0.00083	n/s
		e=10	0.0755	0.0117	0.00203	n/s	0.00574	0.00401	0.00083	n/s
		e=6	0.0755	0.0244	0.00203	n/s	0.00402	0.00415	0.00179	n/s
2 (AG)	2	perfect	6.78	7.65	7.98	0.00116	5.98	7.04	8.12	< 0.0001
		e=10	6.81	7.57	7.97	0.00221	5.97	7.02	8.08	< 0.0001
		e=6	6.23	6.71	7.13	0.00479	5.44	6.31	7.3	< 0.0001
	3 to 5	perfect	1	0.916	0.941	n/s	0.54	0.645	0.75	< 0.0001
		e=10	0.916	0.911	0.931	n/s	0.536	0.634	0.736	< 0.0001
		e=6	1.33	1.72	1.61	0.00237	0.9	1.17	1.31	< 0.0001
	6+	perfect	0	0.0121	0.0074	n/s	0.00772	0.00293	0.00262	n/s
		e=10	0.0572	0.00605	0.0171	n/s	0.00772	0.00837	0.0118	n/s
		e=6	0.0572	0.0371	0.0171	n/s	0.0282	0.0219	0.0361	0.00402
	10+	perfect	0	0	0	n/a	0	0.00133	0	n/s
		e=10	0	0.00605	0	n/s	0	0.00221	0	n/s
		e=6	0	0.00605	0	n/s	0	0.00221	0.00046	n/s
2 (CG)	2	perfect	1.38	1.78	1.98	0.00155	0.89	1.59	1.7	< 0.0001
		e=10	1.38	1.78	1.98	0.00155	0.89	1.59	1.7	< 0.0001
		e=6	1.19	1.68	1.9	0.00032	0.87	1.54	1.63	< 0.0001
	3 to 5	perfect	0.00507	0.16	0.18	n/s	0.0542	0.0938	0.0864	< 0.0001
		e=10	0.00507	0.16	0.18	n/s	0.0542	0.0937	0.0864	< 0.0001
		e=6	0.0894	0.23	0.26	n/s	0.0719	0.138	0.145	< 0.0001
	6+	perfect	0	0	0	n/a	0	0	0	n/a
		e=10	0	0	0	n/a	0	0.00011	0	n/s
		e=6	0	0	0	n/a	0	0.00011	0	n/s
	10+	perfect	0	0	0	n/a	0	0	0	n/a
		e=10	0	0	0	n/a	0	0	0	n/a
		e=6	0	0	0	n/a	0	0	0	n/a
2 (all)	2	perfect	24.1	23.5	24.2	n/s	24.5	24.3	24.3	n/s
		e=10	23.6	23.3	24	n/s	24.2	24.1	24.2	n/s
		e=6	20	20.6	21.6	n/s	20.9	21.4	21.7	0.00031
	3 to 5	perfect	6.27	4.48	3.87	n/s	5.38	3.72	2.92	< 0.0001
		e=10	5.76	4.18	3.63	n/s	4.91	3.51	2.76	< 0.0001
		e=6	7.47	6.01	5.56	n/s	6.85	5.34	4.5	< 0.0001
	6+	perfect	0.957	0.211	0.23	n/s	0.403	0.153	0.0597	< 0.0001
		e=10	1.38	0.393	0.413	n/s	0.675	0.281	0.169	< 0.0001
		e=6	1.84	0.616	0.385	n/s	1.03	0.413	0.221	< 0.0001
	10+	perfect	0.622	0.0378	0.045	n/s	0.141	0.0444	0.00908	0.00019
		e=10	0.754	0.0816	0.0935	n/s	0.214	0.0653	0.0232	< 0.0001
		e=6	0.76	0.118	0.131	n/s	0.226	0.0801	0.0283	0.00227
3 (all)	2	perfect	9.45	10.9	12	n/s	11.2	11.2	11.5	n/s
		e=10	9.2	10.7	11.9	0.00491	11	11.1	11.4	0.00262
		e=6	7.23	9.25	10.8	< 0.0001	9.4	9.73	10.1	< 0.0001

Table A9 continued

Repeat type			Mean per kb repeat freq. in hot IGRs				Mean per kb repeat freq. in non-hot IGRs			
Motif length	Copy numb.	Mismatch type	Number of promoters			P value	Number of promoters			P value
			None	One	Two		None	One	Two	
3 (all)	3 to 5	perfect	1	0.559	0.614	n/s	0.643	0.539	0.446	0.00108
		e=10	0.871	0.527	0.569	n/s	0.586	0.525	0.438	0.00025
		e=6	2.58	1.96	1.57	n/s	2.1	1.84	1.78	n/s
	6+	perfect	0.0875	0.0149	0.0714	n/s	0.0269	0.00821	0.00573	n/s
		e=10	0.111	0.044	0.0619	n/s	0.0644	0.0175	0.0127	n/s
		e=6	0.138	0.102	0.1	n/s	0.109	0.0361	0.0258	n/s
	10+	perfect	0.0478	0	0	n/s	0.0159	0.00215	0.0013	n/s
		e=10	0.0875	0.00744	0	n/s	0.0265	0.00641	0.0013	n/s
		e=6	0.0875	0.00744	0	n/s	0.0275	0.00811	0.0013	n/s
4 (all)	2	perfect	5.36	4.33	3.72	n/s	4.45	3.93	3.56	n/s
		e=10	5.14	4.13	3.58	n/s	4.13	3.75	3.43	n/s
		e=6	4.04	3.47	2.98	n/s	3.14	3.1	2.86	0.00017
	3 to 5	perfect	0.0653	0.209	0.111	n/s	0.15	0.105	0.0664	n/s
		e=10	0.119	0.374	0.207	n/s	0.454	0.245	0.19	n/s
		e=6	0.233	0.419	0.24	n/s	0.568	0.319	0.258	n/s
	6+	perfect	0	0	0	n/a	0.0152	0.00167	0.00096	n/s
		e=10	0	0	0	n/a	0.0116	0.00523	0.00096	n/s
		e=6	0	0	0	n/a	0.0155	0.00825	0.00199	n/s
	10+	perfect	0	0	0	n/a	0.00593	0	0	n/s
		e=10	0	0	0	n/a	0.00593	0	0	n/s
		e=6	0	0	0	n/a	0.00593	0	0	n/s
5 (all)	2	perfect	2.03	1.61	1.72	n/s	1.75	1.52	1.45	0.00066
		e=10	2.05	1.51	1.56	n/s	1.57	1.41	1.33	< 0.0001
		e=6	1.49	1.18	1.28	n/s	1.18	1.13	1.02	< 0.0001
	3 to 5	perfect	0.037	0.0734	0.0119	n/s	0.0326	0.031	0.0351	n/s
		e=10	0.118	0.126	0.0411	n/s	0.118	0.0899	0.0956	n/s
		e=6	0.0809	0.117	0.0378	n/s	0.106	0.0957	0.0981	n/s
	6+	perfect	0	0	0	n/a	0	0.00109	0.00112	n/s
		e=10	0	0	0.00474	n/s	0	0.00109	0.00112	n/s
		e=6	0	0	0.00474	n/s	0.00217	0.00152	0.00527	n/s
	10+	perfect	0	0	0	n/a	0	0	0	n/a
		e=10	0	0	0	n/a	0	0	0	n/a
		e=6	0	0	0	n/a	0	0	0	n/a
6 (all)	2	perfect	1.03	0.705	0.841	n/s	0.825	0.606	0.495	n/s
		e=10	0.861	0.65	0.768	n/s	0.724	0.55	0.457	n/s
		e=6	0.385	0.507	0.63	n/s	0.484	0.375	0.327	n/s
	3 to 5	perfect	0.0557	0.0655	0.0157	n/s	0.0223	0.016	0.02	n/s
		e=10	0.0309	0.0538	0.0596	n/s	0.0366	0.0454	0.0298	n/s
		e=6	0.0031	0.0507	0.0503	n/s	0.0344	0.0306	0.0209	n/s
	6+	perfect	0	0.0109	0	n/s	0	0.00087	0	n/s
		e=10	0	0.0109	0	n/s	0	0.00087	0.00191	n/s
		e=6	0	0.0109	0	n/s	0.00207	0.00324	0.0024	n/s
	10+	perfect	0	0	0	n/a	0	0	0	n/a
		e=10	0	0	0	n/a	0	0	0.00066	n/s
		e=6	0	0	0	n/a	0	0	0.00066	n/s

Table A10: The effect of complex microsatellites. Showing numbers of microsatellites in IGRs located within five or ten bp of other microsatellites of the same or larger size group (compound and degenerate repeats), including the subset of these which had repeat motifs with the same base composition (degenerate repeats). IGRs were divided by recombination (double-strand break) intensity as reported by Gerton and co-workers [1] into 473 hot and and 5520 non-hot regions. Imperfect repeats were allowed, with a maximum of one mismatch per six bp. Degenerate repeats only were considered for microsatellites with less than six copies. This was to avoid inordinate effects on the results caused by the extremely high abundance of short poly-A runs relative to other repeat types.

Repeat type		**Total number of repeats**		**Compound & degenerate repeats**				**Degenerate repeats only**			
				% within 5 bp of another rpt.		% within 10 bp of another rpt.		% within 5 bp of another rpt.		% within 10 bp of another rpt.	
Motif length	Copy numb.	Hot	Non-hot	Hot	Non-hot	Hot	Non-hot	Hot	Non-hot	Hot	Non-hot
1 (A)	3 to 5	8459	106082	n/a	n/a	n/a	n/a	39.1	41.2	61.5	64.3
	6+	1473	15164	12.4	10.1	20.1	17	11.7	9.6	18.6	15.9
	14+	173	919	16.2	15	23.7	24.3	14.5	14.5	20.8	22.9
1 (G)	3 to 5	2428	23094	n/a	n/a	n/a	n/a	13.2	9.44	24.1	19.3
	6+	46	305	8.7	6.89	13	11.8	0	0	0	0.656
	14+	2	2	0	0	0	0	0	0	0	0
2 (all)	2	5586	63536	n/a	n/a	n/a	n/a	8.65	9.11	17.7	17
	3 to 5	1412	14380	n/a	n/a	n/a	n/a	4.67	4.44	7.08	7.48
	6+	130	1022	8.46	11.1	12.3	18.8	0.769	3.33	0.769	4.4
	10+	33	209	3.03	13.4	12.1	19.1	0	5.74	0	6.7
3 (all)	2	2682	29846	n/a	n/a	n/a	n/a	2.16	1.88	3.91	3.81
	3 to 5	520	5524	n/a	n/a	n/a	n/a	0.385	0.597	1.15	1.3
	6+	21	118	9.52	15.3	19	22	0	5.08	0	5.08
	10+	3	20	0	10	0	25	0	5	0	5
4 (all)	2	810	8568	n/a	n/a	n/a	n/a	0.123	0.397	1.6	0.794
	3 to 5	76	828	n/a	n/a	n/a	n/a	0	0.242	0	1.45
	6+	0	19	n/a	0	n/a	0	n/a	0	n/a	0
	10+	0	1	n/a	0	n/a	0	n/a	0	n/a	0
5 (all)	2	298	3175	n/a	n/a	n/a	n/a	0	0.126	0	0.126
	3 to 5	26	272	n/a	n/a	n/a	n/a	0	0	0	0
	6+	1	5	0	0	0	0	0	0	0	0
	10+	0	0	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
6 (all)	2	123	1074	n/a	n/a	n/a	n/a	1.63	0.372	1.63	0.372
	3 to 5	8	69	n/a	n/a	n/a	n/a	0	0	0	0
	6+	1	10	0	20	0	20	0	0	0	0
	10+	0	4	n/a	25	n/a	25	n/a	0	n/a	0

Table A11: Microsatellite frequencies in hotspot flanking regions. Mean microsatellite frequencies in hot IGRs and flanking IGRs one and two ORFs removed from hotspots. Statistical comparisons were made between the flanking IGRs and non-hot IGRs more than four ORFs removed from hotspots**.** All p values <0.01 are shown, but caution is recommended in view of the multiple hypotheses being tested.

Repeat type			**Mean repeat frequency by IGR type**				**P value**	
Motif length	Copy number	Mismatch type	Hot	1 removed from hot	2 removed from hot	Non-hot	1 removed v non-hot	2 removed v non-hot
1 (A)	3 to 5	perfect	35	40.4	40.7	39.8	n/s	n/s
		e=10	34.3	39.8	40	39.4	n/s	n/s
		e=6	31.8	36.8	37.4	36.7	n/s	n/s
	6+	perfect	5.42	4.93	5.28	4.51	n/s	0.0037
		e=10	5.24	4.79	5.19	4.4	n/s	0.0025
		e=6	6.12	6.06	6.21	5.42	n/s	0.00432
	14+	perfect	0.418	0.288	0.21	0.165	n/s	n/s
		e=10	0.733	0.631	0.372	0.292	0.00027	n/s
		e=6	0.854	0.773	0.412	0.353	0.00029	n/s
1 (G)	3 to 5	perfect	9.18	7.38	6.13	7.32	n/s	0.00022
		e=10	9.16	7.35	6.12	7.31	n/s	0.00024
		e=6	8.89	7.22	6.14	7.18	n/s	0.00094
	6+	perfect	0.118	0.0802	0.0672	0.0739	n/s	n/s
		e=10	0.114	0.0802	0.0672	0.0723	n/s	n/s
		e=6	0.16	0.092	0.0806	0.0914	n/s	n/s
	14+	perfect	0.0035	0	0	0.00093	n/s	n/s
		e=10	0.0035	0	0	0.00093	n/s	n/s
		e=6	0.0035	0	0	0.00093	n/s	n/s
2 (AT)	2	perfect	7.69	8.93	9.61	9.22	n/s	n/s
		e=10	7.57	8.77	9.42	9.1	n/s	n/s
		e=6	6.26	6.97	7.8	7.63	n/s	n/s
	3 to 5	perfect	2.68	2.82	2.76	2.59	n/s	n/s
		e=10	2.42	2.61	2.61	2.43	n/s	n/s
		e=6	3.03	3.61	3.55	3.21	n/s	n/s
	6+	perfect	0.308	0.331	0.148	0.156	n/s	n/s
		e=10	0.45	0.346	0.207	0.224	n/s	n/s
		e=6	0.627	0.611	0.485	0.435	n/s	n/s
	10+	perfect	0.142	0.046	0.0547	0.0505	n/s	n/s
		e=10	0.197	0.205	0.0848	0.0732	n/s	n/s
		e=6	0.221	0.237	0.0948	0.082	n/s	n/s
2 (AC)	2	perfect	6.8	6.37	6.67	6.61	n/s	n/s
		e=10	6.69	6.35	6.65	6.56	n/s	n/s
		e=6	6.13	5.83	6.21	6.01	n/s	n/s
	3 to 5	perfect	0.908	0.709	0.425	0.595	n/s	0.0099
		e=10	0.924	0.689	0.425	0.578	n/s	n/s
		e=6	1.32	1.08	0.774	0.992	n/s	n/s
	6+	perfect	0.0518	0.0064	0.0167	0.0167	n/s	n/s
		e=10	0.0772	0.0064	0.0167	0.0223	n/s	n/s
		e=6	0.134	0.0292	0.0602	0.0469	n/s	n/s
	10+	perfect	0.0159	0	0.0114	0.00291	n/s	n/s
		e=10	0.0218	0	0.0114	0.00336	n/s	n/s
		e=6	0.0283	0	0.00362	0.00876	n/s	n/s

Table A11 continued

Repeat type			Mean repeat frequency by IGR type				P value	
Motif length	Copy number	Mismatch type	Hot	1 removed from hot	2 removed from hot	Non-hot	1 removed v non-hot	2 removed v non-hot
2 (AG)	2	perfect	7.57	6.47	6.52	7.19	0.00752	n/s
		e=10	7.53	6.44	6.53	7.17	0.00688	n/s
		e=6	6.73	5.82	5.78	6.47	n/s	n/s
	3 to 5	perfect	0.94	0.687	0.681	0.656	n/s	n/s
		e=10	0.918	0.655	0.679	0.647	n/s	n/s
		e=6	1.61	1.12	1.2	1.17	n/s	n/s
	6+	perfect	0.00828	0.0164	0.0152	0.00117	0.00027	n/s
		e=10	0.0196	0.0322	0.0152	0.00705	n/s	n/s
		e=6	0.03541	0.08742	0.0222	0.0241	n/s	n/s
	10+	perfect	0	0.00954	0	< 0.0001	< 0.0001	n/s
		e=10	0.00307	0.0164	0	< 0.0001	< 0.0001	n/s
		e=6	0.00307	0.0164	0	0.00022	< 0.0001	n/s
2 (CG)	2	perfect	1.76	1.25	1.4	1.45	0.00978	n/s
		e=10	1.76	1.24	1.4	1.45	0.00761	n/s
		e=6	1.64	1.17	1.3	1.4	0.00375	n/s
	3 to 5	perfect	0.132	0.0979	0.0412	0.0886	n/s	n/s
		e=10	0.132	0.0979	0.0412	0.0886	n/s	n/s
		e=6	0.213	0.148	0.0766	0.131	n/s	n/s
	6+	perfect	0	0	0	0	n/a	n/a
		e=10	0	0	0	< 0.0001	n/s	n/s
		e=6	0	0	0	< 0.0001	n/s	n/s
	10+	perfect	0	0	0	0	n/a	n/a
		e=10	0	0	0	0	n/a	n/a
		e=6	0	0	0	0	n/a	n/a
2 (all)	2	perfect	23.8	23	24.2	24.5	0.00654	n/s
		e=10	23.5	22.8	24	24.3	0.00482	n/s
		e=6	20.8	19.8	21.1	21.5	0.00081	n/s
	3 to 5	perfect	4.67	4.31	3.91	3.93	n/s	n/s
		e=10	4.34	3.93	3.63	3.67	n/s	n/s
		e=6	6.17	5.95	5.6	5.5	n/s	n/s
	6+	perfect	0.368	0.354	0.18	0.174	n/s	n/s
		e=10	0.599	0.514	0.362	0.332	n/s	n/s
		e=6	0.797	0.727	0.567	0.506	n/s	n/s
	10+	perfect	0.158	0.0555	0.0661	0.0535	n/s	n/s
		e=10	0.221	0.222	0.0963	0.0766	n/s	n/s
		e=6	0.252	0.253	0.0984	0.091	n/s	n/s
3 (all)	2	perfect	10.9	11.2	11.4	11.2	n/s	n/s
		e=10	10.8	11	11.2	11.1	n/s	n/s
		e=6	9.29	9.32	9.76	9.7	n/s	n/s
	3 to 5	perfect	0.664	0.448	0.492	0.541	n/s	n/s
		e=10	0.609	0.437	0.392	0.525	n/s	n/s
		e=6	1.97	1.83	1.68	1.91	n/s	n/s
	6+	perfect	0.046	0.0416	0.0156	0.00875	n/s	n/s
		e=10	0.0627	0.0416	0.0495	0.0219	n/s	n/s
		e=6	0.109	0.123	0.0997	0.04	n/s	n/s

Table A11 continued

Repeat type			Mean repeat frequency by IGR type				P value	
Motif length	Copy number	Mismatch type	Hot	1 removed from hot	2 removed from hot	Non-hot	1 removed v non-hot	2 removed v non-hot
3 (all)	10+	perfect	0.0097	0.0378	0.00881	0.00344	n/s	n/s
		e=10	0.0215	0.0416	0.017	0.00781	n/s	n/s
		e=6	0.0215	0.0416	0.0265	0.00816	n/s	n/s
4 (all)	2	perfect	4.36	4.14	3.81	3.97	n/s	n/s
		e=10	4.17	4	3.54	3.77	n/s	n/s
		e=6	3.44	3.21	2.77	3.06	n/s	n/s
	3 to 5	perfect	0.151	0.0986	0.18	0.098	n/s	n/s
		e=10	0.274	0.297	0.431	0.267	n/s	n/s
		e=6	0.329	0.376	0.507	0.35	n/s	n/s
	6+	perfect	0	0.0254	0	0.00443	n/s	n/s
		e=10	0	0.0276	0.00834	0.00276	n/s	n/s
		e=6	0	0.0276	0.0106	0.00461	n/s	n/s
	10+	perfect	0	0.0254	0	0	0.00035	n/a
		e=10	0	0.0254	0	0	0.00035	n/a
		e=6	0	0.0254	0	0	0.00035	n/a
5 (all)	2	perfect	1.72	1.65	1.73	1.56	n/s	n/s
		e=10	1.63	1.48	1.56	1.43	n/s	n/s
		e=6	1.28	1.08	1.22	1.1	n/s	n/s
	3 to 5	perfect	0.0482	0.0418	0.0354	0.0357	n/s	n/s
		e=10	0.0999	0.108	0.144	0.1	n/s	n/s
		e=6	0.0867	0.115	0.126	0.0959	n/s	n/s
	6+	perfect	0	0	0	0.00103	n/s	n/s
		e=10	0.00137	0	0	0.00134	n/s	n/s
		e=6	0.00137	0.0205	0	0.00176	0.0047	n/s
	10+	perfect	0	0	0	0	n/a	n/a
		e=10	0	0	0	0	n/a	n/a
		e=6	0	0	0	0	n/a	n/a
6 (all)	2	perfect	0.811	0.834	0.721	0.64	n/s	n/s
		e=10	0.727	0.718	0.647	0.563	n/s	n/s
		e=6	0.518	0.523	0.424	0.387	n/s	n/s
	3 to 5	perfect	0.0491	0.0534	0.0313	0.0203	n/s	n/s
		e=10	0.0509	0.0916	0.0407	0.0429	n/s	n/s
		e=6	0.0409	0.025	0.0485	0.0297	n/s	n/s
	6+	perfect	0.00552	0	0	0.00079	n/s	n/s
		e=10	0.00552	0	0	0.00578	n/s	n/s
		e=6	0.00552	0	0	0.00291	n/s	n/s
	10+	perfect	0	0	0	0	n/a	n/a
		e=10	0	0	0	0.00228	n/s	n/s
		e=6	0	0	0	0.00112	n/s	n/s

Appendix B: supporting information for Chapter 4

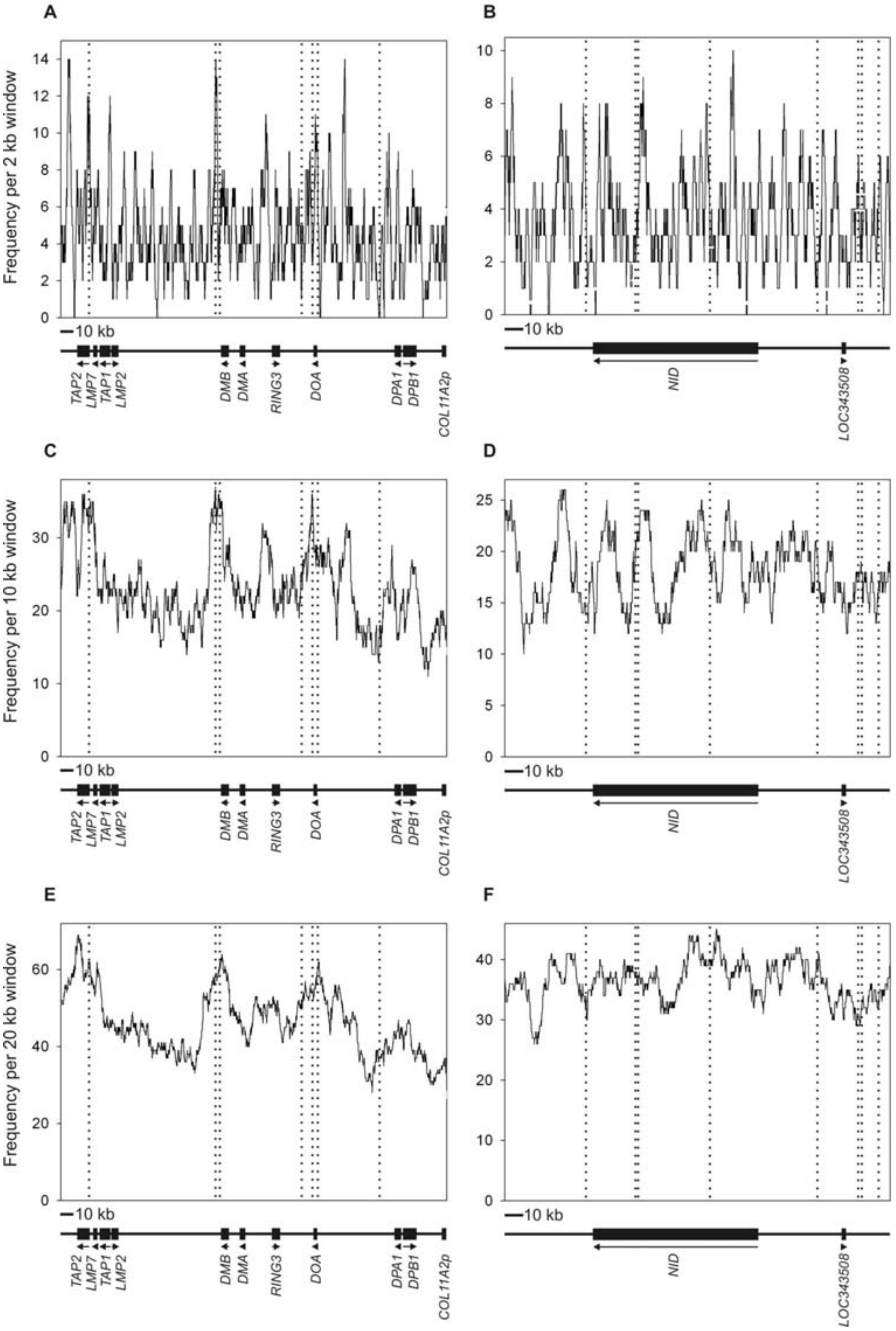


Figure B1: Densities of all PPTs of at least 12 bp relative to human hot spot locations Sliding window plots of the densities of PPTs of at least 12 bp (no GC-content restriction) relative to hot spot locations in the two contiguous areas of the human genome over which multiple hotspots have been well characterized experimentally: A, C and E): a 292 kb region of the human MHC Class II region in which 7 hot spots have been mapped and B, D and F): a 206 kb region of human chromosome 1 in which 8 hot have been mapped. Sliding window plots with different window sizes are shown: 2 kb (A and B), 10 kb (C and D) and 20 kb (E and F). Vertical dotted lines represent hot spot mid point locations. Sliding windows moved in steps of 100 bp. Locations of genes in are shown below the plots with arrows indicating direction of transcription.

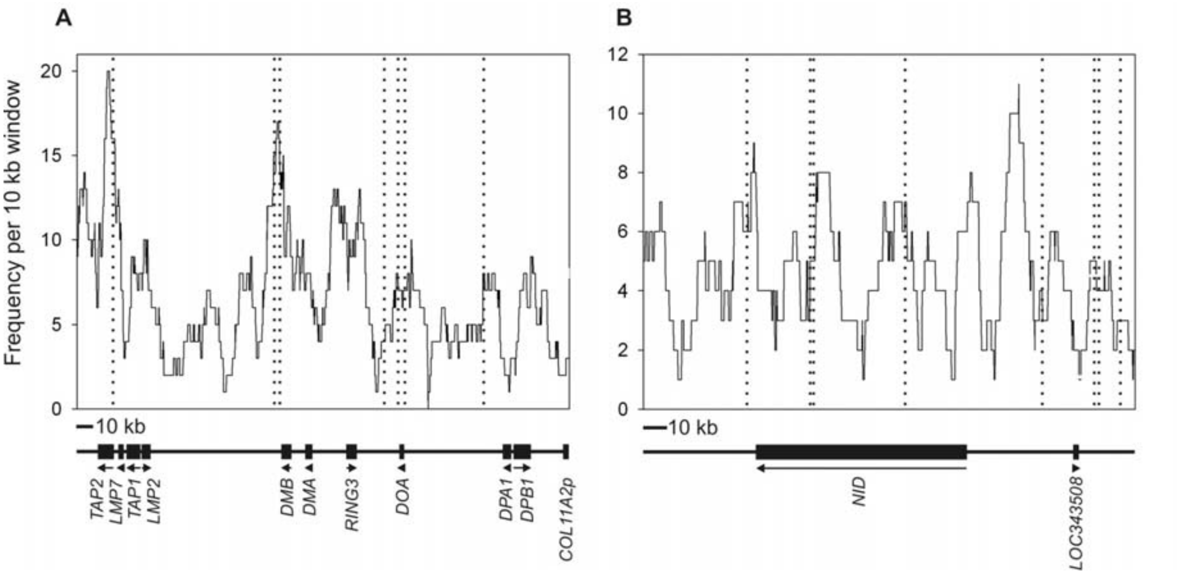


Figure B2: Densities of high GC-content PPTs of at least 20 bp relative to human hot spot locations
Sliding window plots of the densities of PPTs of at least 20 bp (one mismatch allowed per 10 bp), with GC-contents above the mean for PPTs in these regions, relative to hot spot locations in the two contiguous areas of the human genome over which multiple hotspots have been well characterized experimentally: A, C and E): a 292 kb region of the human MHC Class II region in which 7 hot spots have been mapped and B, D and F): a 206 kb region of human chromosome 1 in which 8 hot have been mapped. Pure PPTs of more than 20 bp were relatively rare in these regions, and the patterns shown in these plots only emerged when some mismatches were allowed; for these plots a maximum of one in any 10 bp PPT segment. Vertical dotted lines represent hot spot mid point locations. Sliding windows moved in steps of 100 bp. Locations of genes in are shown below the plots, with arrows indicating direction of transcription.