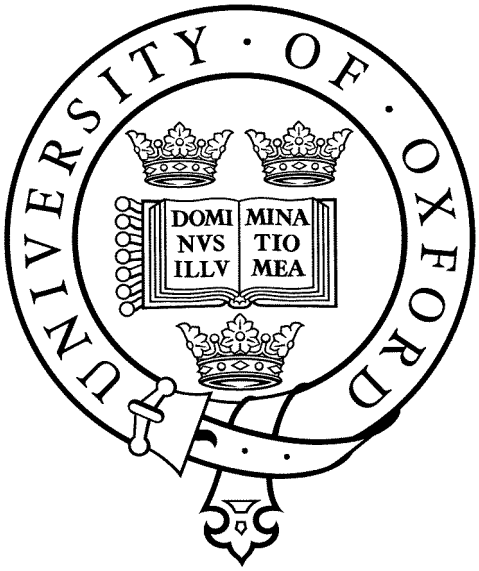
The Estimation of Recombination Rates from Population Genetic Data



A Thesis submitted for the Degree of Doctor of Philosophy

Adam Auton

Hertford College, University of Oxford

Trinity 2007

The Estimation of Recombination Rates from Population Genetic Data

Adam Auton, Hertford College, University of Oxford

DPhil Thesis; Trinity 2007

Abstract

Genetic recombination is an important process that generates new combinations of genes on which natural selection can operate. As such, an understanding of recombination in the human genome will provide insight into the evolutionary processes that have shaped our genetic history. The aim of this thesis is to use samples of population genetic data to explore the patterns of variation in the rate of recombination in the human genome. To do this I introduce a novel means of estimating recombination rates from population genetic data. The new, computationally efficient method incorporates a model of recombination hotspots that was absent in existing methods.

I use samples from the International HapMap Project to obtain recombination rate estimates for the autosomal portion of the genome. Using these estimates, I demonstrate that recombination has a number of interesting relationships with other genome features such as genes, DNA repeats, and sequence motifs. Furthermore, I show that genes of differing function have significantly different rates of recombination. I explore the relationship between recombination and specific sequence motifs and argue that while sequence motifs are an important factor in determining the location of recombination hotspots, the factor that controls motif activity is unknown.

The observation of many relationships between recombination and other genome features motivates an attempt to quantify the contributions to the recombination rate from specific features. I employ a wavelet analysis to investigate scale-specific patterns of recombination. In doing so, I reveal a number of highly significant correlations between recombination and other features of the genome at both the fine and broad scales, but find that relatively little of the variation in recombination rates can be explained. I conclude with a discussion of the results contained in the body of the thesis, and suggest a number of areas for future research.

Acknowledgements

My interest in population genetics emerged during my first year at Oxford as a member of the Life Sciences Interface Doctoral Training Centre. I would like to extend thanks to all members of the LSI DTC and especially to those that worked so hard in setting up such an innovative program. In particular, I would like to thank David Gavaghan, James Wakefield and Maureen York. Thank you also to the EPSRC, who funded the LSI DTC program.

The majority of my DPhil has been spent working with my supervisor, Gil McVean, to whom I am extremely grateful for the many hours of advice, support and inspiration he has given me.

I would also like to thank all members of the Mathematical Genetics and Bioinformatics Group. I would especially like to thank Peter Donnelly, Daniel Falush, Colin Freeman, Bob Griffiths, and Chris Spencer, who have provided insight and advice during my DPhil. Former members of the group that deserve mention include Simon Myers and Daniel Wilson, both of whom continue to be of assistance.

Many thanks go to my office mates over the years, including Niall Cardin, Jed Francis, Jo Gay, and Chris Hallsworth. I apologise if I have been difficult to tolerate during the last few weeks of thesis writing.

Thank you to both the staff and student members of Hertford College for both support and friendship.

To my examiners, Paul Fearnhead and Jonathan Marchini, thank you for taking the time to consider this thesis.

I would like to thank all members of my family for the enormous amount of support they have given me. Finally, I would like to thank Sarah, whose companionship has been so important.

Table of Contents

Chapter 1 Introduction

Genetic recombination is of crucial importance in the process of evolution and occurs in most known organisms, including eukaryotes, bacteria, and viruses. While mutation generates new gene variants for natural selection to work on, recombination ensures that new combinations of genes are generated. A useful source of information regarding historical recombination events can be found in population genetic data, and recent advances in sequencing and genotyping technologies have greatly increased the availability of such data. Given a sample of population data, patterns of similarities between different sequences can provide information regarding the genealogical history of the sample. However, interpretation of the observed patterns can be problematic without an understanding of the process that generated the patterns in the first place. Statistical modelling of the evolutionary process by which the data was generated can therefore provide a useful tool by which patterns of variation can be understood.

In this thesis, I aim to use population genetic methods to gain an insight into recombination rates with a specific focus on the human genome. In the remainder of this chapter, I introduce the process of recombination and the various methods by which it can be detected from genetic data. I describe our current knowledge of recombination rate variation in the human genome as obtained from experimental techniques. I then introduce methods by which recombination may be detected in population genetic samples, and a commonly used model of population evolution known as the Coalescent. I outline how this model may be used to understand patterns

of variation in population genetic samples. Finally, I describe a number of existing methods for detecting the locations of regions with highly elevated recombination rates, known as recombination hotspots.

In Chapter 2, I describe a new method for the estimation of recombination rates from population genetic data. This method, which can be applied on a genome-wide scale, improves on existing methods by incorporating a model of recombination hotspots. In Chapter 3, I discuss the performance of this method using both simulated and real datasets and find the method to be superior to an existing method for rate estimation, but inferior to existing methods for hotspot detection. In Chapter 4, I apply the new rate estimation method to genome-wide data taken from the International HapMap project (THE INTERNATIONAL HAPMAP CONSORTIUM 2007). I consider the relationship between recombination and various genome features such as genes, repeat elements, and sequence motifs. I also investigate whether epigenetic features of the sequence motifs can explain the locations of recombination hotspots. In Chapter 5, a wavelet analysis is used to explore influences on recombination on a scale-by-scale basis, and I identify a number of scale specific relationships. The conclusion of this thesis is contained in Chapter 6, which outlines the successes and failures of this analysis and suggests avenues for further research.

The Process of Recombination

Homologous recombination is the process by which a pair of homologous DNA sequences exchange some portion of their DNA. These are usually located on two copies of the same chromosome, although other similar DNA molecules may also participate. Most current knowledge of the recombination mechanism was originally

derived from studies in bacteria and yeasts as the short generation times and relatively small genomes of these organisms allowed mutants with defective recombination processes to be isolated and the associated proteins to be characterised. More recently, homologues of these proteins have been discovered in *Drosophila*, mice and humans (LICHTEN 2001).

A model for the process by which recombination occurs in eukaryotes was proposed in 1964 by Robin Holliday (HOLLIDAY 1964), and has subsequently formed the basis of the generally accepted model of recombination. Holliday's model suggested a number of intermediate stages of recombination, as shown in Figure 1.

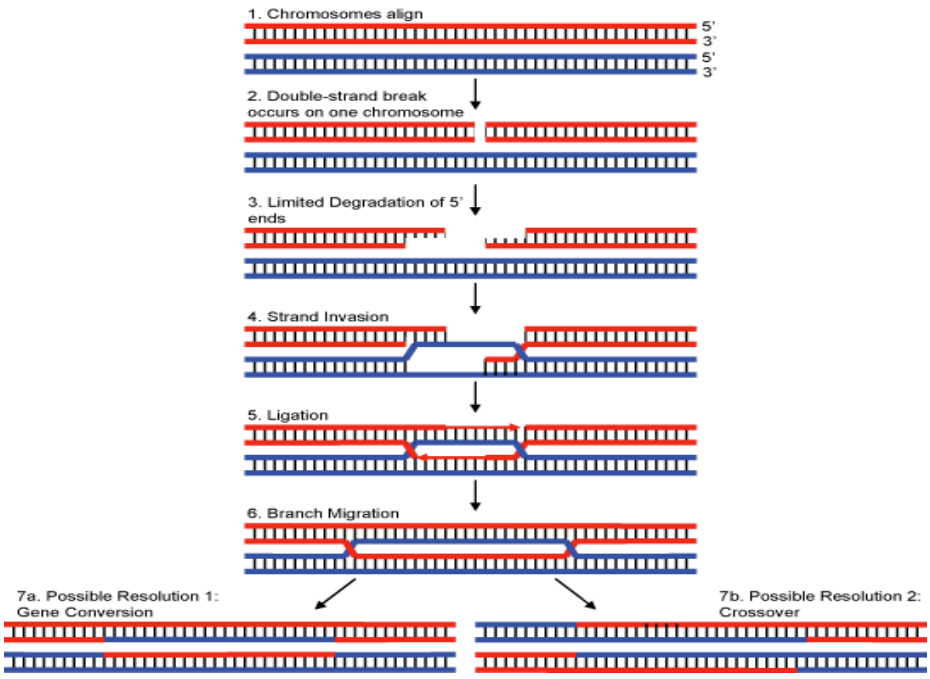


Figure 1. The Holliday model of recombination. As indicated, the process begins with a double strand break (DSB) occurring on one chromosome. Sections of the DNA material immediately surrounding the DSB are degraded by an exonuclease enzyme, which creates small portions of single-strand DNA that can invade and bind to the homologous DNA in the other chromosome. The damage is then repaired using the other chromosome as a template and a structure known as the Holliday junction is formed. The Holliday junction can resolve in two possible ways; Gene Conversion (7a) and the focus of this thesis, Crossover (7b).

The recombination process has two possible resolutions, known as Gene Conversion and Crossover. In gene conversion, only a small amount of genetic information is exchanged between chromosomes. In humans, gene conversion is estimated to account for approximately 90% of recombination events (JEFFREYS and MAY 2004). However, the amount of material exchanged between the two chromosomes (known as the tract length) is as low as 300 base pairs (JEFFREYS and MAY 2004). Accordingly, gene conversion can leave very little, if any, trace in the DNA sequence of the generated gametes. Detection of gene conversion events is therefore a challenging problem (GAY and MCVEAN 2007; HELLENTHAL and STEPHENS 2006). The alternative resolution, crossover, leads to large tracts of genetic material being exchanged and hence leaves a much larger signal in the genome. In this case, all genetic material beyond the recombination points is exchanged.

Throughout this thesis, I will be concerned only with the process of crossover. The reader should note that when used in this thesis, the term 'recombination' should be understood to mean 'crossover' unless otherwise stated.

Experimental Techniques for Detecting Recombination

The rate at which recombination occurs is usual measured in terms of the expected number of recombination events between two loci per generation. A commonly used unit of measurement is the centimorgan (cM), which is defined as a 1% chance that two loci will be separated by a recombination event in one generation. Two loci are said to be one centimorgan apart if a recombination event occurs between them in 1% of meioses on average. The unit is additive, so that if two loci are

separated by, say, 200 centimorgans, one would expect to observe 2 recombination events between the two loci per generation on average.

The standard approach to studying rates of recombination across a genome is to build a genetic map by genotyping a large number of individuals in families. Given a high enough density of markers, it is possible to observe chromosomes in later generations that are recombinant forms of those in earlier generations. By calculating the number of recombination events between markers, it is possible to obtain a measure for the distance between the loci (expressed in centimorgans). The genetic map is created by finding the distances between a set of markers on the same chromosome, which have ideally been chosen to avoid significant gaps between markers so as to avoid the inaccuracies that can occur as a result of multiple recombination events.

Genetic maps have been constructed for a number of organisms. In eukaryotes, there is a notable uniformity in the total recombination map length regardless of genome size (AWADALLA 2003). The average recombination rate is therefore negatively correlated with genome size (Table 1), a pattern that extends over four orders of magnitude in both recombination rate and genome size.

Organism	Recombination Rate (cM/Mb)	Genome Size (Mb)	Reference
Human	1.19	3000	(KONG *et al.* 2002)
Rat	0.6	2750	(JENSEN-SEAMAN *et al.* 2004)
Mouse	0.56	2632	(JENSEN-SEAMAN *et al.* 2004)
Maize	0.7	2500	(FU *et al.* 2002)
Drosophila	1.5	123	(NACHMAN 2002)
Arabidopsis Thaliana	4.6	115	(MEZARD 2006)
Caenorhabditis elegans	3.06	100	(BARNES *et al.* 1995)
Yeast	370	16	(PETES 2001)
HIV	30000	0.01	(JETZT *et al.* 2000)

Table 1 - Genome Average Recombination Rates in Various Organisms. On a log-log scale, the above data shows a strong negative correlation, with a coefficient of determination of 0.91.

The deCODE Map

Of particular interest in this thesis is the human genetic map constructed by the deCODE company based in Iceland (KONG *et al.* 2002). The deCODE map was constructed from a large pedigree study of 869 individuals in 146 Icelandic families. The study used a total of 5,136 microsatellite markers to observe the product of 1,257 meiosis (of which 628 were paternal and 629 maternal).

The deCODE map revealed a high degree of fine-scale structure in recombination rates. The shorter chromosomes tended to have higher recombination rates than the larger chromosomes, with the average recombination rates of chromosomes 21 and 22 (2.06 and 2.11 cM/Mb respectively) being twice those of chromosomes 1 and 2 (0.96 and 1.02 cM/Mb respectively). There was also a high degree of rate heterogeneity within chromosomes (Figure 2). The total map length of

the genome was also found to be an average of 1.65 times longer in females than in males.

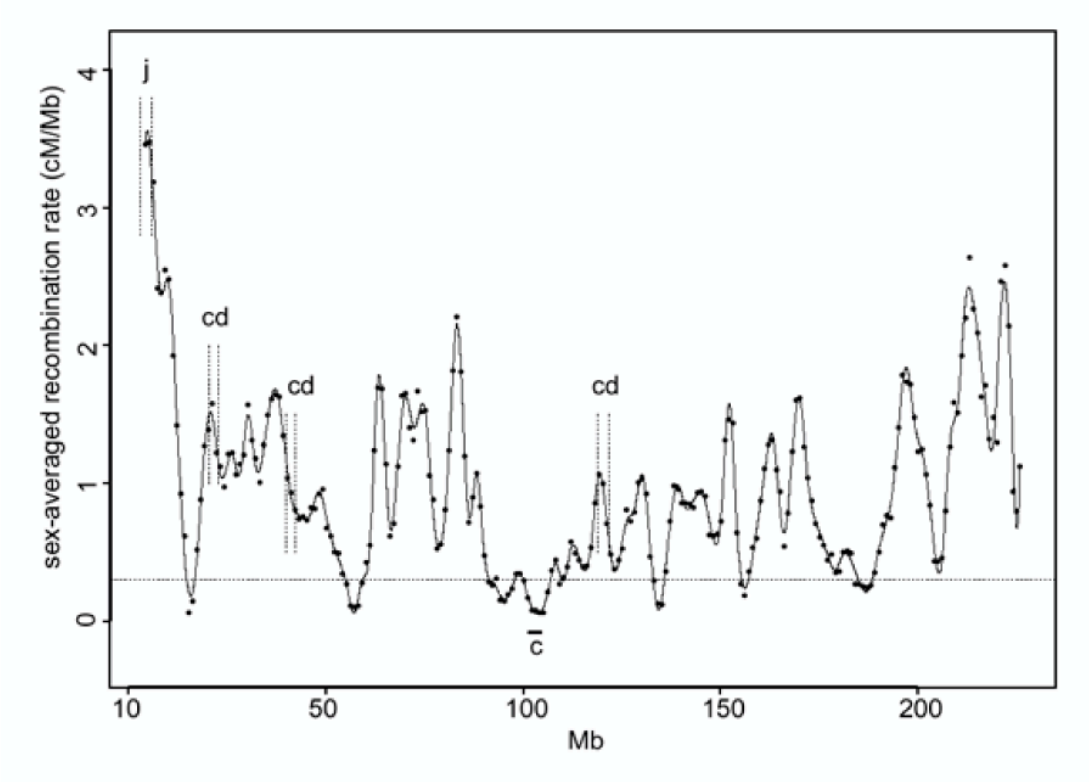


Figure 2. The sex-averaged recombination rate for chromosome 3. The recombination rate was calculated in a moving window of 3Mb in width. The centromere is represented by 'c'. The letters 'cd' and 'j' represent recombination 'deserts' and 'jungles' respectively, as defined by the authors of the original paper. Adapted from Kong et al. 2002.

In short, the deCODE map provided a remarkable insight into the properties of recombination in the human genome. However, many questions remained. While the resolution of the map (of the order of 1 to 3Mb) was higher than had previously been achieved, it was not high enough to identify anything more than the broadest features of recombination rate variation. As will be seen in the following section, other studies have identified recombination rate variation at a much finer scale.

Sperm Typing

The fundamental limitation of pedigree studies is the resolution is determined by both the marker density and the number of observable meioses. Even the largest pedigree studies (such as the deCODE study) do not contain more than a few hundred individuals, and therefore cannot achieve a recombination detection resolution much below a megabase.

An alternative method, known as sperm typing, avoids this problem by searching for recombinant sequences in the sperm product of male individuals. As recombination happens during spermatogenesis, it is possible to obtain many thousand recombinant sequences from a single individual. In one highly successful application of the sperm typing method (JEFFREYS *et al.* 2001), Single Nucleotide Polymorphisms (SNPs) are used as markers. SNPs are locations in a DNA sequence where a single nucleotide - A, T, C, or G - differs between members of a species (or between paired chromosomes in an individual). For example, if two sequenced homologous DNA fragments from different individuals were to read GGAA**C**TC and GGAA**T**TC, then we would call a SNP at the fifth base. In this case, we say that there is a SNP with two alleles: C and T.

Once a region of interest has been identified, batches of DNA from the sperm of a man showing a high degree of SNP heterozygosity are amplified using the Polymerase Chain Reaction (PCR). By carefully choosing PCR primers, it is possible to selectively amplify recombinant sequences above the level of non-recombinant sequences. The location of a crossover event can then be determined by typing the SNPs and comparing to the ancestral chromosomes.

The achievable resolution is much higher than that achieved by pedigree studies, as it is possible to observe the outcome of literally thousands of meioses from a single individual. The resolution at which recombination events can be detected is therefore no longer limited by the number of observable meioses. The limiting factor is now the density of available markers (SNPs), and resolutions of less than 0.5kb are feasible (KAUPPI *et al.* 2004).

There are however at least two major drawbacks to the sperm typing method. First, it can only be applied to males. Second, the method is technically challenging and cannot easily be scaled up to cover large regions of the genome. Generally regions of no more than approximately 200kb have been studied at the highest resolution (e.g. JEFFREYS *et al.* 2001), and a region of 2.5Mb has been studied at a lower resolution (GREENAWALT *et al.* 2006).

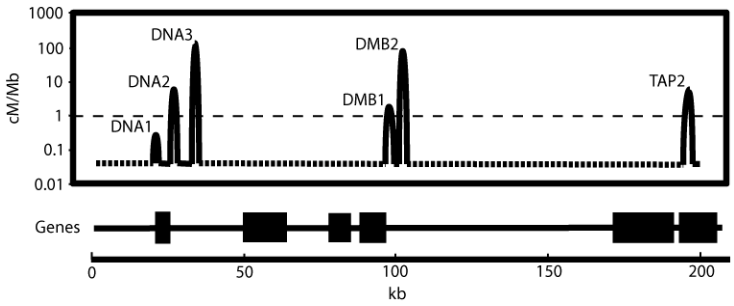
Nevertheless, sperm typing has provided valuable insights into the patterns of recombination in the human genome. Two of the studied regions will be visited repeatedly in this thesis, so I take the opportunity to describe these regions in more depth here.

The MHC Region

A 216kb region of the major histocompatibility complex (MHC) on chromosome 6 has been extensively studied by sperm typing (JEFFREYS *et al.* 2001). This study typed 274 SNPs in sperm donated by eight unrelated men of North European ancestry. A further 50 men were genotyped so that the correlation between alleles at separate loci could be estimated. Analysis of the recombinant DNA from sperm revealed that recombination events tend to cluster in highly localised regions.

The rate of recombination in these regions could be hundred or thousands of times that of the surrounding regions, and hence these areas were dubbed recombination hotspots.

A total of six hotspots were visible in the MHC analysis (Figure 3). The vast majority (but not all) of recombination events observed in the region occurred in these hotspots. The rare events outside of the hotspots suggest a recombination rate of approximately 0.04cM/Mb (JEFFREYS *et al.* 2001) which, if correct, would indicate that 95% of crossovers occur within hotspots. Furthermore, the recombination rate over the whole region is 0.9cM/Mb, which is very close to the male genome average of 0.89cM/Mb and suggests that this region is not atypical in terms of recombination (although the MHC is highly atypical in other respects). Despite showing large differences in activity, all hotspots show similar morphology, with an apparently symmetric distribution having a 95% width of approximately 1 to 2kb. The similar width of the hotspots suggests a common process operating at each hotspot.



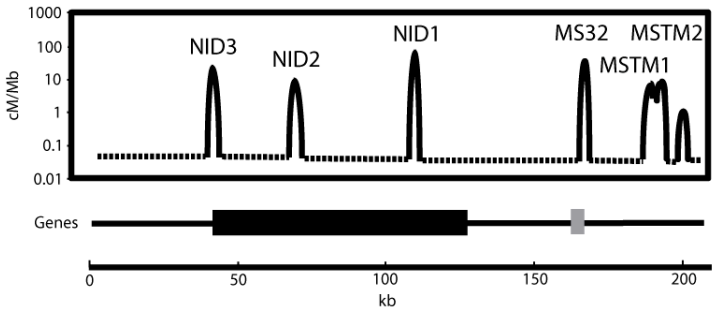
Hotspot Name	*95% Width (kb)*	*Peak Rate (cM/Mb)*
DNA1	1.9	0.4
DNA2	1.3	8
DNA3	1.2	100
DMB1	1.8	1
DMB2	1.2	20
TAP2	1.0	8

Figure 3. Recombination hotspots in the MHC region. Also shown as a thin dashed line is the male genome average recombination rate.

The MS32 Region

The MHC region described above provided an insight into the patterns of recombination in the human genome. However, the MHC is a very unusual part of the genome, being gene rich, and under intense selection pressures (MEYER *et al.* 2006). It was therefore not known if the observed crossover patterns were indeed typical of the human genome. A second region unremarkable 206kb region of chromosome 1 was therefore selected for study by Jeffreys *et al* (2005). I will refer to this as the MS32 region due to a highly variable minisatellite located towards the 3' end of the region, which shares the same name. The study genotyped 200 SNPs in 80 unrelated men of North European ancestry. Crossovers were detected via sperm typing in seven men.

Again, the analysis revealed a number of recombination hotspots with properties similar to those observed in the MHC (Figure 4). The high resolution of this study allowed the identification of two so-called 'double' hotspots with centres separated by less than 2kb (the hotspots in question are NID2 and MSTM2). Furthermore, at least one hotspot (MS32) has apparently left little signal in the patterns of genetic diversity of the men sampled, despite being one of the hottest regions in the sperm analysis. This was cited as evidence that hotspots are transient features of the genome. If hotspots such as MS32 have evolved recently, then they may not have had sufficient time to leave their mark on haplotype diversity (JEFFREYS *et al.* 2005). This hypothesis is still debatable, but is consistent with the observation that recombination rates estimated in humans and chimps show poor correlation (PTAK *et al.* 2005; WINCKLER *et al.* 2005).



Hotspot Name	*95% Width (kb)*	*Peak Rate (cM/Mb)*
NID3	2.0	70
NID2a	1.4	10
NID2b	1.1	4
NID1	1.5	70
MS32	1.5	40
MSTM1a	1.6	9
MSTM1b	2.1	16
MSTM2	1.3	0.9

Figure 4. Recombination hotspots in the MS32 region. Note that the double hotspot in MSTM1 is clearly visible, whereas the one at NID2 is difficult to detect even in sperm. Also note that the location of the MS32 minisatellite is shown as a small grey rectangle.

A final observation of note from this study is that three of the hotspots (NID1, MS32, and MSTM2) showed significant rate variation in different men. This observation has been used to support the hypothesis that the location of hotspots is sequence dependent, with differing alleles being associated with very different levels of hotspot activity (MYERS *et al.* 2005).

Detecting Recombination from Samples of Genetic Variation

The experimental methods considered so far have provided excellent evidence for extensive amounts of rate variation in the human genome. However, both methods have serious limitations. Pedigree studies lack the resolution to determine more than crudest features of rate variation. Conversely, while sperm studies provide excellent rate estimates at high resolution, they cannot be easily scaled up to provide genome wide estimates.

An alternative source of information regarding recombination can be found in samples of population genetic data (Figure 5a). In such data, the non-independence of alleles allows the allele at one locus to be informative of the allele at another locus.

This non-random association of alleles is known as Linkage Disequilibrium (LD). Summary statistics can be used to investigate patterns in LD. Two commonly used statistics, *D'* and r^2, are shown in Figure 5b. If f_{AB} is the frequency of haplotypes with allele A at the first locus and allele B at the second locus, f_{A-} is the frequency of haplotypes with the A allele at the first locus, and f_{-B} is the frequency of haplotypes with the B allele at the second locus, then these statistics can be calculated using equations (1.1) to (1.3). The *D'* statistic (LEWONTIN 1964) is a measure linkage disequilibrium defined as the difference between the frequency of a two-locus haplotype and the product of the component alleles, divided by the most extreme possible value given the marginal allele frequencies. Alternatively, the r^2 statistic (HILL and ROBERTSON 1968) is the square of the coefficient of association of gene frequencies between two loci.

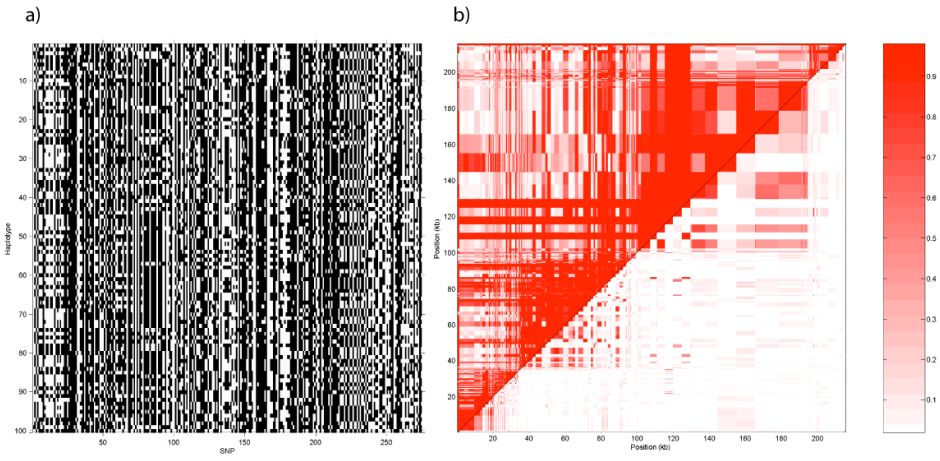


Figure 5. The MHC data, and two commonly used statistics. a) The MHC SNP data from the Jeffreys et al. (2005) study discussed in the previous section. Shown here are 100 haplotypes. Each row represents an individual haplotype, with black and white representing the two alleles at each locus. The original data is in the form of genotypes, so the data displayed here has been phased using the program, *PHASE* (STEPHENS and SCHEET 2005). b) Two non-parametric statistics of the same data (augmented with SNP positions), *D'* (top left) and r^2 (bottom right). Each pairwise SNP comparison is represented by a region on the grid, with bright red colours indicating high values of the statistic and faded colour representing low values.

$$D = f_{AB} - f_{A-}f_{-B} \tag{1.1}$$

$$r^2 = \frac{D^2}{f_{A-}f_{-B}f_{a-}f_{-b}} \tag{1.2}$$

$$D' = \begin{cases} \dfrac{D}{\min\left(f_{A-}f_{-b}, f_{a-}f_{-B}\right)} & \text{if } D > 0 \\ \dfrac{-D}{\min\left(f_{A-}f_{-B}, f_{a-}f_{-b}\right)} & \text{if } D < 0 \end{cases} \tag{1.3}$$

Recombination events break down the amount of LD between loci. However, it is unclear how such statistics can be directly related to the underlying recombination rate (DEVLIN and RISCH 1995). Therefore, perhaps a more sophisticated approach is to look for patterns in the data that could only have been

caused by recombination (and making certain assumptions). The simplest such method is known as the four-gamete test (WEIR 1979). Given two bi-allelic loci with alleles A/B and a/b respectively, there are four possible haplotypes: AB, Ab, aB and ab. If all four haplotypes are observed in a sample, then either a recurrent mutation or a recombination event has occurred. If we assume an infinite sites mutation model (KIMURA 1969), then only recombination could have generated the observed pattern.

In certain organisms (such as humans), the genome is large and the mutation rate is sufficiently low that the infinite sites model is not an unreasonable approximation in most cases. Therefore, applying the four-gamete test to all pairs of loci in a sample identifies regions where recombination must have occurred. However, the test has very low power, as very specific conditions are needed for a recombination to leave a mark in a sample (WIUF *et al.* 2001). An illustrative example is given by considering how the power of the method increases with the size of a sample. A large sample is preferable, as there is a greater chance of sampling rare haplotypes that inform about recombination. However, the number of detectable recombination events increases with the log of the log of the sample size (MYERS 2002).

Aside from the four-gamete test, there are other techniques for detecting recombination from population genetic data, many of which are more powerful and / or sophisticated (for example, MYERS and GRIFFITHS 2003; SONG and HEIN 2005; WIUF 2002). However, none is able to detect the majority of recombination events. Furthermore, even if we did know how many recombination events have occurred in the history of a sample, we still would not be able to infer per-generation recombination rate without knowing how many generations the events span.

We can learn more about the data by modelling the underlying process. Once a suitable model of the underlying process has been constructed, it is possible to use so-called 'likelihood methods' to perform inference on the recombination rate. In the next section, I introduce a probabilistic model known as the coalescent. In this model, the process by which the sample data was generated is described.

Introduction to the Coalescent

In order to build tractable models of natural populations, it is necessary to make simplifying assumptions that are almost always unrealistic. One of the simplest models of populations is known as the Wright-Fisher Model (FISHER 1930; WRIGHT 1931) which makes the following assumptions:

- Constant Population Size
- Random mating with the possibility of selfing
- No migration
- No selection
- Non-overlapping generations

In this model, given a population of N haploid individuals, the next generation is formed by sampling (with replacement) from the current population (Figure 6).

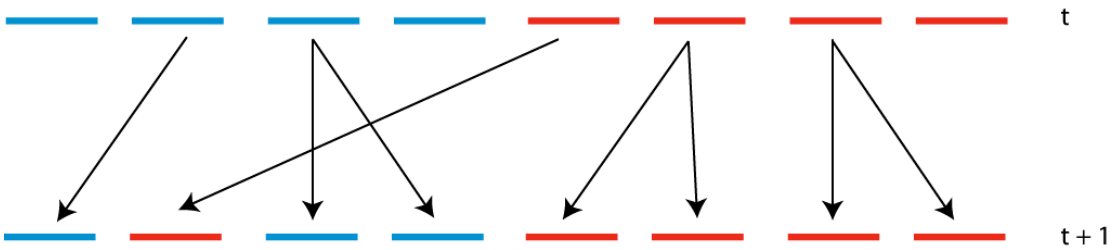


Figure 6. The Wright-Fisher population model. The genes making up the next generation are drawn randomly with replacement from the current generation.

In the simplest case where there are only two (neutral) segregating alleles (say, B and b) in a haploid population, and no mutation between alleles, the probability that there are j B alleles in generation $t + 1$ given that there were i in generation t is binomially distributed:

$$\Pr\left(\#\text{B alleles} = j\right) = \binom{N}{j}\left(\frac{i}{N}\right)^{j}\left(1-\frac{i}{N}\right)^{N-j}. \qquad (1.4)$$

The number of B alleles in generation $t + 1$ depends only on the number of alleles in the current generation t. The change in frequency of the allele is therefore an example of a Markov chain. Also, note that a diploid population containing N individuals is modelled as a haploid population of $2N$ individuals.

It is possible to impose a process of mutation on top of the Wright-Fisher model by simply allowing each chromosome to mutate between generations with a small probability, μ. That is, a chromosome is passed unchanged to the next generation with probability $(1 - \mu)$, and with probability μ a mutation occurs. If the mutations are selectively neutral (that is, they do not influence the probability of survival into the next generation), then the remainder of the Wright-Fisher model is unchanged (Figure 7a).

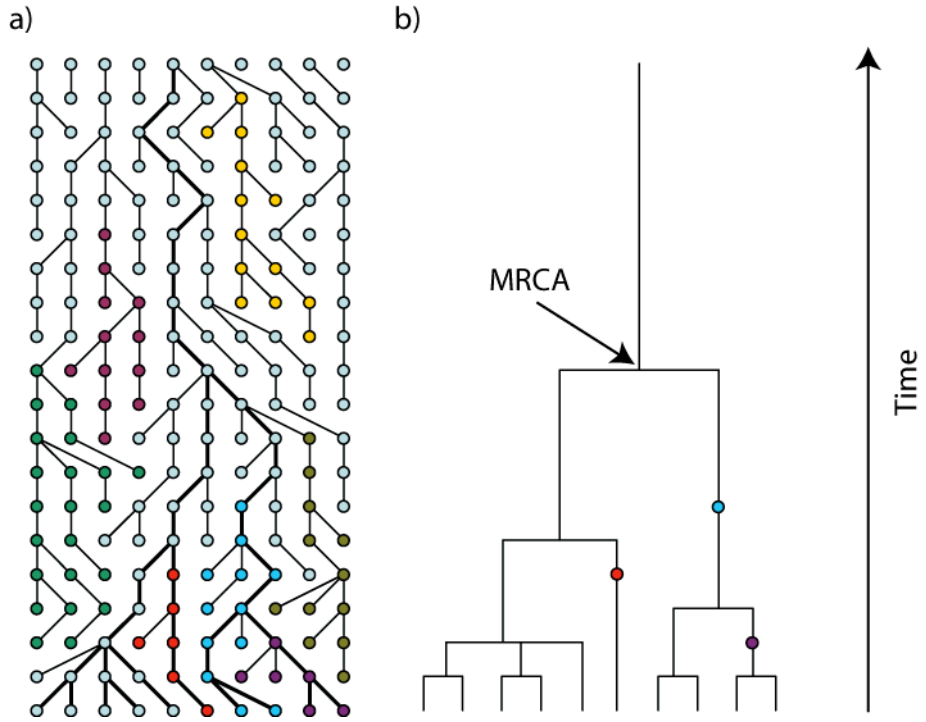


Figure 7. An example Wright-Fisher population and the corresponding coalescent tree. a) A sample Wright-Fisher population with mutation. The original generation is represented at the top of the figure. Each subsequent generation is generated by sampling from the previous generation. Mutations occur with some probability, and are represented by changes in colour. b) The corresponding coalescent tree. As we move back in time, lineages coalesce until a common ancestor is reached. Mutations are represented by coloured circles. In the coalescent, time is usually measured backwards in units of $2N_e$ generations.

Although true populations may violate many of the assumptions of the Wright-Fisher model, it is often possible to approximate the behaviour of the true population using a Wright-Fisher population with an effective population size, N_e. The effective population size gives the size of the Wright-Fisher population that in some sense best approximates the true population.

The Wright-Fisher model does however contain a large amount of redundancy. If we consider a population evolving over a sustained period of time, many lineages will die out and therefore contribute nothing to the final population.

Indeed, a current-day neutrally-evolving population is expected to have a single common ancestor at $2N$ generations in the past (e.g. NORDBORG 2000). Furthermore, if we obtain a genetic sample, we would like to be able to model the history of this particular sample without considering the whole population.

The coalescent follows from the Wright-Fisher model, and provides a simple stochastic model for the history of a genetic sample. The coalescent was originally devised by Kingman (1982), but was also discovered independently on at least two other occasions (HUDSON 1983a; TAJIMA 1983). It has since become the dominant model for population genetic analysis.

Consider two chromosomes chosen randomly from a Wright-Fisher population of size $2N_e$. The probability that both chromosomes share a common ancestor in the previous generation is simply $(2N_e)^{-1}$. If such an event occurs, the two chromosomes are said to have *coalesced*. If, however, they did not coalesce (something that happens with probability $1-(2N_e)^{-1}$), then the probability that they do so at the second generation given they did not coalesce in the first, is still $(2N_e)^{-1}$ due to the Markovian property of the Wright-Fisher model. The same is also true for all previous generations, so the probability distribution of the time until the two chromosomes coalesce is geometric.

$$\Pr\left(\text{coalesce at time } t\right) = \frac{1}{2N_e}\left(1-\frac{1}{2N_e}\right)^{t-1} \tag{1.5}$$

The same reasoning can be applied for a sample of size n. The probability of no coalescent events occurring in the previous generation is

$$\begin{aligned}\left(1-\frac{1}{2N_e}\right)\left(1-\frac{2}{2N_e}\right)\ldots\left(1-\frac{n-1}{2N_e}\right) &= \prod_{i=1}^{n-1}\left(1-\frac{i}{2N_e}\right) \\ &= 1-\sum_{i=1}^{n-1}\frac{i}{2N_e}+\mathrm{O}\left(\frac{1}{(2N_e)^2}\right) \\ &\approx 1-\binom{n}{2}\frac{1}{2N_e}\end{aligned} \tag{1.6}$$

If we assume that $2N_e$ is large relative to the sample size, then two approximations can be made. First, terms of the order of $(2N_e)^{-2}$ can be ignored. Second, the probability of a coalescence occurring in any given generation is small enough that a continuous approximation to the geometric distribution can be made. By rescaling time in units of $2N_e$ generations (i.e. $t = j / 2N_e$, where j is the number of generations), the probability that no chromosomes have coalesced by time t is

$$\begin{aligned}\Pr(\tau > t) &= \lim_{N_e\to\infty}\prod_{i=1}^{n-1}\left(1-\frac{i}{2N_e}\right)^{2N_e t} \\ &= \exp\left(-\binom{n}{2}t\right).\end{aligned} \tag{1.7}$$

Hence the waiting time until the first coalescent event is exponentially distributed with rate $\binom{n}{2}$.

The advantage of the coalescent over the Wright-Fisher model is that only the lineages of the sample need be considered. The Markovian nature of the process allows realisations of the coalescent process to be efficiently generated using the following algorithm.

Algorithm 1

1. Start with $k = n$ lineages, where n is the sample size.
2. If $k = 1$, finish.
3. Draw a random waiting time τ, which is exponentially distributed with rate $\binom{k}{2}$.
4. Choose two lineages uniformly and coalesce them, leaving k - 1 edges. Decrease k by one and go to 2.

Once there is a single lineage, the Most Recent Common Ancestor (MCRA) of the sample is said to have been reached, and the process terminates. The specific combination of coalescent events and times can then be represented using a tree, with the samples at the tips, and the MCRA at the root (Figure 7b).

As with the original Wright-Fisher population, mutations are considered to be selectively neutral. Neutral mutations do not influence the structure of the coalescent tree, and can therefore be considered separately. However, recall that the probability of a single chromosome experiencing a mutation between generations in the Wright-Fisher model was μ. We have since rescaled time in units of $2N_e$ generations, so we introduce a more appropriate measure of the mutation rate: the population mutation rate, $\theta := 4N_e\mu$. The population mutation rate can be interpreted as the expected number of mutations separating a sample of two sequences, since the expected time until coalescence is $2N_e$ and μl mutations are expected on each branch where l is the length of the branch.

Recall that the rate of coalescence is $\binom{k}{2}$ where k is the number of lineages. Likewise, the rate at which mutations occur along a single lineage is θ / 2, and hence

the rate at which mutations occur on all lineages is θk / 2. Thus, as we move backwards in time (and up the genealogy), the probability that the next event we encounter is a coalescence event is

$$\Pr(\text{Coalescence}) = \frac{\binom{k}{2}}{\binom{k}{2} + \frac{k\theta}{2}} = \frac{k-1}{k-1+\theta}. \tag{1.8}$$

Likewise, the probability that the next event is a mutation is:

$$\Pr(\text{Mutation}) = \frac{\frac{k\theta}{2}}{\binom{k}{2} + \frac{k\theta}{2}} = \frac{\theta}{k-1+\theta} \tag{1.9}$$

As mutations do not alter the shape of the genealogy, it is possible to add mutations after a genealogy has been generated. Algorithm 1 proceeds with the following two additional steps, which execute after the genealogy has been generated.

5. For each branch, draw a number, M_l, from a Poisson distribution with intensity $l\theta$ / 2 where l is the length of the branch.
6. For each branch, scatter M_l mutations uniformly on the branch.

The simulation algorithm has complexity that is linear in n, making simulation efficient even for large sample sizes (HUDSON 1983b).

The Coalescent with Recombination

So far, the coalescent process we have considered contains no model of recombination. To incorporate recombination, we return to the Wright-Fisher model.

Whereas in the original model each chromosome had a single parent, in the Wright-Fisher model with recombination each chromosome can have two parents (Figure 8). A recombination event occurs at an individual locus with rate r per generation. If a recombination event occurs then a location along the paternal chromosome is chosen at random, and the two parental chromosomes recombine at this point.

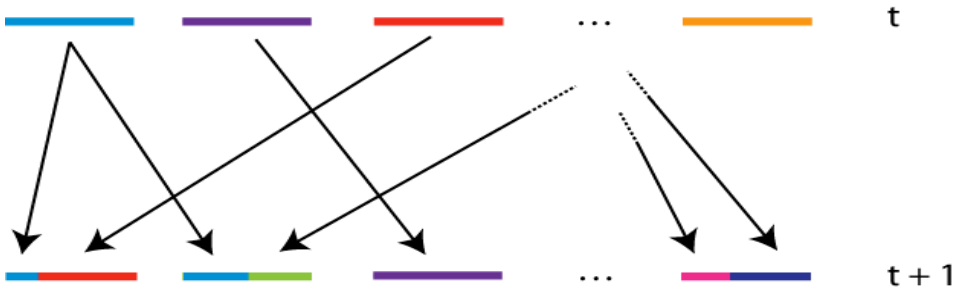


Figure 8. The Wright-Fisher model with recombination. Recombination occurs with some probability. If a recombination event occurs, the recombinant individual chooses two parents and the recombinant chromosome is generated by recombining at a random location.

Viewed backwards in time, recombination allows each separate locus to have a distinct coalescent history. When a recombination occurs, it is therefore necessary to trace the ancestry of both the maternal and paternal chromosomes. Furthermore, the locations of recombination events on the chromosome need to be recorded, as they are required to determine the ancestry of each specific locus.

In the coalescent model of recombination, the rate at which recombination events occur is determined by the population recombination rate, ρ, which is dependent on the rate of recombination per locus per generation, r, and the effective population size N_e:

$$\rho := 4N_e r. \tag{1.10}$$

There are now three possible events that can occur back in time: coalescent events (combining of lineages), recombination events (splitting of lineages) and mutation events. The rate at which these three events occur is given by λ_C for coalescent events, λ_R for recombination events and λ_M for mutation. Given k lineages, these rates are given by:

$$\begin{aligned} \lambda_C(k) &= \binom{k}{2} \\ \lambda_R(k) &= \frac{\rho k}{2} \\ \lambda_M(k) &= \frac{\theta k}{2} \end{aligned} \tag{1.11}$$

The probabilities that the next event is a coalescence, recombination or mutation are given by:

$$\begin{aligned} \Pr(\text{Coalescence}) &= \frac{k-1}{k-1+\theta+\rho} \\ \Pr(\text{Recombination}) &= \frac{\rho}{k-1+\theta+\rho} \\ \Pr(\text{Mutation}) &= \frac{\theta}{k-1+\theta+\rho} \end{aligned} \tag{1.12}$$

It is no longer possible to represent an ancestry as a tree as in Figure 7. Instead, a more complex graph known as an Ancestral Recombination Graph (ARG; GRIFFITHS and MARJORAM 1996) is used (Figure 9).

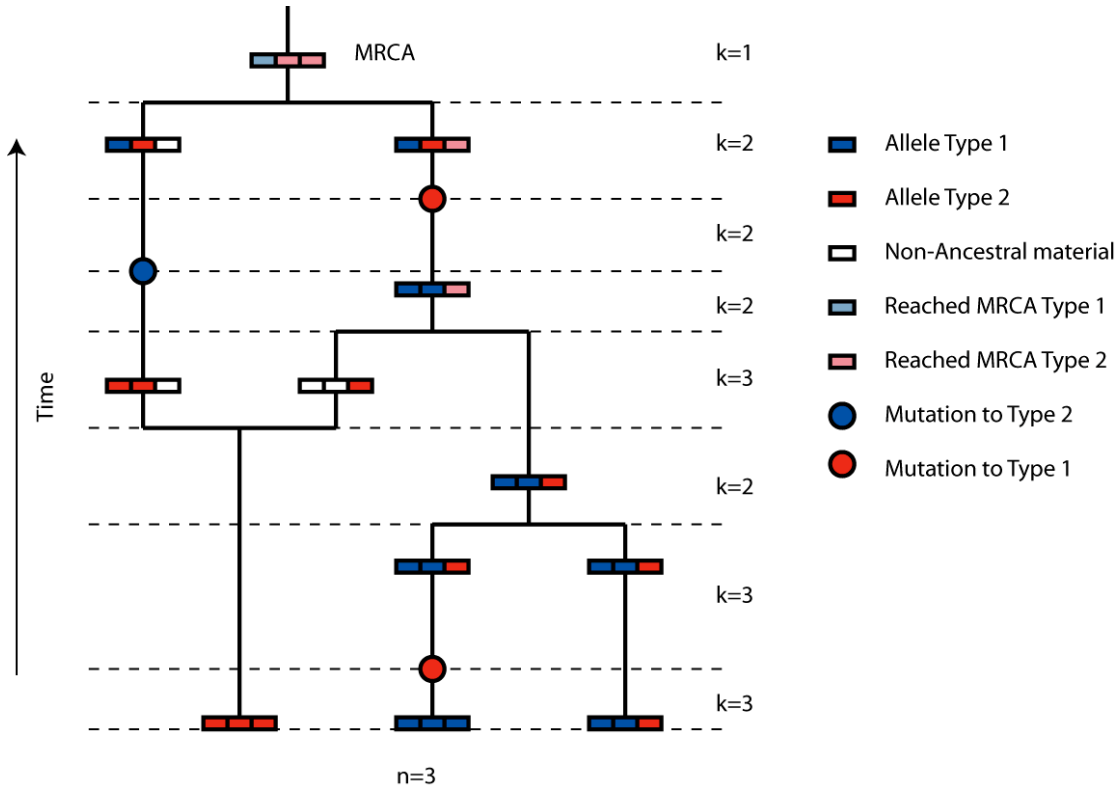


Figure 9. The Ancestral Recombination Graph. This ARG contains three coalescent events, one recombination event, and three mutations. The first event that occurs back in time is a coalescent, which is represented by the joining of lineages. The next event that occurs is a recombination event, which is represented by the splitting of a lineage. The process continues until all sites have reached their MRCA. Lighter colours indicate that the MRCA has been reached at that site.

Simulation of ARGs is more complex than for coalescent trees, especially in the case of high recombination, as the number of lineages can become quite large. The basic algorithm proceeds as shown in Algorithm 2.

Algorithm 2

1. Start with $k = n$ lineages, where n is the sample size.
2. If $k = 1$, finish.
3. For k sequences with ancestral material, draw the time until the next event from an exponential distribution with rate $k(k - 1)/2 + k\rho/2$.
4. With probability $(k\text{ -}1) / (k - 1 + \rho)$, choose a coalescent event. Otherwise choose a recombination event.
5. If a recombination event, draw a random sequence and a random location in the sequence. Create two ancestor sequences by splitting the sequence at the randomly chosen location. Increase k by one and return to 2.
6. If a coalescent event, choose two sequences and merge into one sequence at a randomly chosen location. Decrease k by one and return to two.

As for the coalescent without recombination, neutral mutations can be added after the generation of the genealogy. The method is the same as that described in Algorithm 1. The times and locations of all coalescent, recombination and mutation events need to be stored in order to fully describe the ARG.

A possible improvement to the above algorithm can be made by keeping track of ancestral material. In doing so, the efficiency of the algorithm can be improved, especially for high recombination rates. If a lineage contains no material that is ancestral to the final population sample, it is no longer necessary to simulate that lineage. Furthermore, any recombination event which occurs outside of the absolute

boundaries of ancestral material (i.e. in the non-ancestral material at the chromosome ends) has no effect on the final sample and therefore need not be modelled. This can be achieved by altering the rate of recombination of a chromosome with non-ancestral material to proportional to the sum of material between the left and right endpoints of ancestral material. When a recombination event occurs, it can only occur within the left and right endpoints of ancestral material. The algorithm that incorporates this model is known as Hudson's algorithm (HUDSON 1983a).

Calculating the Probability of a Dataset

Given a dataset consisting of genetic samples, ***D***, we would like to perform inference of the population recombination rate. In order to achieve this, we need to calculate the probability of obtaining our data under an assumed model. In the unrealistic situation in which the genealogy that generated the observed data, *G*, is known, it is possible to calculate the likelihood of the data using the relative rate at which events occur. Given the order in which events occurred in the genealogy, the likelihood of the mutation and recombination rates is given by:

$$L(\theta,\rho \mid G) = \prod_{j:k_{j+1}>k_j} \frac{\lambda_R(k_j)}{\lambda_R(k_j)+\lambda_C(k_j)+\lambda_M(k_j)} \cdots \prod_{j:k_{j+1}<k_j} \frac{\lambda_C(k_j)}{\lambda_R(k_j)+\lambda_C(k_j)+\lambda_M(k_j)} \cdots \prod_{j:k_{j+1}=k_j} \frac{\lambda_M(k_j)}{\lambda_R(k_j)+\lambda_C(k_j)+\lambda_M(k_j)}. \quad (1.13)$$

where j indexes the events and k_j is the number of lineages at event j (MYERS 2002). Intuitively, the above equation expresses the likelihood as the product of the probabilities of a recombination events, coalescent events, and mutations.

However, in practice, the genealogy is almost always unknown. The likelihood can therefore only be calculated via integration over all possible genealogies.

$$L(\theta,\rho \mid \boldsymbol{D}) = P(\boldsymbol{D} \mid \theta,\rho) = \int P(\boldsymbol{D} \mid G,\theta,\rho) P(G) dG \tag{1.14}$$

Direct calculation of likelihood using the above equations fail in all but the simplest scenarios (see, for example, GRIFFITHS and MARJORAM 1996) due to the unfeasible number of genealogies which need to be summed over.

An alternative approach is to use Monte Carlo methods to obtain approximations to the likelihood (see, for example, GIVENS and HOETING 2005). The most direct approach uses direct Monte Carlo integration, and draws genealogies by simulating directly from the coalescent prior. The likelihood is now calculated via:

$$\int P(\boldsymbol{D} \mid G,\theta,\rho) P(G) dG \approx \frac{1}{M} \sum_{i=1}^{M} P(\boldsymbol{D} \mid G_i,\theta,\rho) \tag{1.15}$$

where G_i is a genealogy drawn from the prior.

However, in practice even this approach gives extremely poor estimates of the likelihood, as the vast majority of the generated genealogies are incompatible with the data, and hence contribute nothing to the likelihood. Therefore, in order to obtain useful estimates of the likelihood, further approximations are required. Approximations can be made to the likelihood calculation or to the coalescent itself (or both). Some of the more popular methods are discussed in the following section, with a focus on the application of recombination rate estimation.

Existing Methods of Recombination Rate Estimation

Importance Sampling Methods

An alternative to the vanilla Monte Carlo integration method is known as Importance Sampling. This method attempts to make the process of simulating genealogies more efficient by only simulating those genealogies that contribute something to the likelihood. The likelihood may then be calculated by weighting the contribution to the likelihood by the probability of obtaining the simulated genealogy from the coalescent prior. More formally, I rewrite equation (1.15) as:

$$L(\theta,\rho \mid \boldsymbol{D}) \propto \frac{1}{M}\sum_{i=1}^{M} P\left(\boldsymbol{D} \middle| G_i,\theta,\rho\right)\frac{P(G_i \mid \theta,\rho)}{Q(G_i \mid \theta,\rho)}. \qquad (1.16)$$

The ratio $P(.)$ / $Q(.)$ in the above equation gives the ratio of the probability of the generated genealogy under the coalescent to the probability of the genealogy under the generating procedure. The beauty of this method is that we may use almost any distribution to generate genealogies, subject to certain criteria being met (for example, the sampling distribution, $Q(.)$, must share the same support as the posterior distribution).

Genealogies can now be generated via a stochastic heuristic that ensures that all generated genealogies are compatible with the data. Such Importance Sampling techniques were first applied to the sampling of genealogies without recombination (GRIFFITHS and TAVARE 1994; KUHNER *et al.* 1998; STEPHENS and DONNELLY 2000), and later with recombination (FEARNHEAD and DONNELLY 2002; GRIFFITHS and MARJORAM 1996).

To perform inference on the recombination rate, a likelihood surface is calculated for a range of θ and ρ values. This surface may be used to obtain an

estimate of ρ either in the Bayesian sense, or a point-estimate may be obtained via the maximum likelihood estimator.

These methods, whilst being a significant improvement over the simple Monte Carlo integration techniques, are computationally intractable on all but the smallest datasets. Furthermore, the recombination rate is generally assumed to be constant across the region. Nevertheless, they have been applied with great success to some datasets (FEARNHEAD and DONNELLY 2002), and can be considered in some sense as the gold-standard at least in the constant rate case.

Approximate Likelihood Methods

As the likelihood of the complete dataset is generally difficult to obtain, attempts have been made to approximate the likelihood surface and hence reduce the computational complexity. This can be achieved by either removing sections of the data which are largely uninformative of the recombination rate (such as low frequency SNPs), or by splitting the data into smaller sections. In the latter case, likelihoods may be calculated for subsets of the data, and then combined to form a *composite likelihood*. If L_j is the likelihood of region j, then the composite likelihood, CL, is given by:

$$CL(\rho) = \prod_j L_j(\rho). \tag{1.17}$$

In the most extreme case, the data is split into all pairs of segregating sites and the likelihood is calculated for each pair. Non-segregating sites are not considered. If *Seg* denotes the set of segregating sites, then the pairwise composite likelihood is defined as:

$$CL(\rho) = \prod_{i,j \in Seg} L_{ij}(\rho). \quad (1.18)$$

This method was first proposed by Hudson (HUDSON 2001), and subsequently extended to allow for complex mutation models (MCVEAN *et al.* 2002).

Despite the apparent *ad hoc* nature of this approach, it performs surprisingly well. It has been informally demonstrated that the maximum composite-likelihood estimate is correlated with the maximum full-likelihood estimate (MCVEAN *et al.* 2002), and is a more accurate estimator than many other methods (SMITH and FEARNHEAD 2005; WALL 2000). However, the composite likelihood has a number of undesirable properties, and is usually sharply peaked in comparison to the full-likelihood. There is also no easily interpretable meaning for the composite likelihood surface. This in turn makes obtaining estimates of uncertainty difficult.

Despite the problems of the composite likelihood, a major benefit of this method is that likelihoods may be pre-calculated and stored for any reasonably sized dataset (about 200 sequences is not unreasonable). Subsequent calculation of the likelihood for a given recombination rate simply involves a repeated table look-up and multiplication operation (or summation for log-likelihoods), and hence can be calculated very quickly on modern computers. This speed of likelihood calculation allows more complex models of recombination rate variation to be considered. By fitting a piecewise constant model, McVean *et al.* were able to obtain variable rate estimates (MCVEAN *et al.* 2004). Furthermore, the method was sufficiently fast that it could be applied on a genome-wide scale.

The composite likelihood method forms a major part of this thesis, and will therefore be considered in greater detail in Chapters 2 and 3.

Approximate Genealogy Methods

The previous two methods have concentrated on approximating the likelihood of a dataset assuming a coalescent model. An alternative approach is to approximate the coalescent process itself. In one particularly successful method of this type, the probability of observing a new haplotype is conditioned on those previously observed (LI and STEPHENS 2003). This model is based on the notion that the new haplotype may be constructed as an imperfect mosaic of those previously observed. A fast algorithm can be used to approximate the likelihood of a set of population data. This likelihood has generated a great deal of interest, and is commonly referred to as the 'Product of Approximate Conditionals' likelihood, or simply the PAC likelihood for short.

To describe the PAC likelihood we start by noting that, given a recombination map $\boldsymbol{\rho}$, the probability of n sampled haplotypes, $H_1,\ldots,H_n$, can be written as:

$$P\left(H_1,...,H_n\middle|\boldsymbol{\rho}\right)=P\left(H_1\middle|\boldsymbol{\rho}\right)P\left(H_2\middle|H_1;\boldsymbol{\rho}\right)...P\left(H_n\middle|H_1,...,H_{n-1};\boldsymbol{\rho}\right). \qquad (1.19)$$

The conditional distributions on the right of this equation are generally unknown, so we substitute a set of approximate distributions, which we denote by π:

$$P\left(H_1,...,H_n\middle|\boldsymbol{\rho}\right)\approx\pi\left(H_1\middle|\boldsymbol{\rho}\right)\pi\left(H_2\middle|H_1;\boldsymbol{\rho}\right)...\pi\left(H_n\middle|H_1,...,H_{n-1};\boldsymbol{\rho}\right) \qquad (1.20)$$

We define the PAC likelihood as:

$$L_{PAC}\left(\boldsymbol{\rho}\right)=\pi\left(H_1\middle|\boldsymbol{\rho}\right)\pi\left(H_2\middle|H_1;\boldsymbol{\rho}\right)...\pi\left(H_n\middle|H_1,...,H_{n-1};\boldsymbol{\rho}\right) \qquad (1.21)$$

To calculate π, we assume that each haplotype is made up of an imperfect mosaic of previously observed haplotypes (Figure 10).

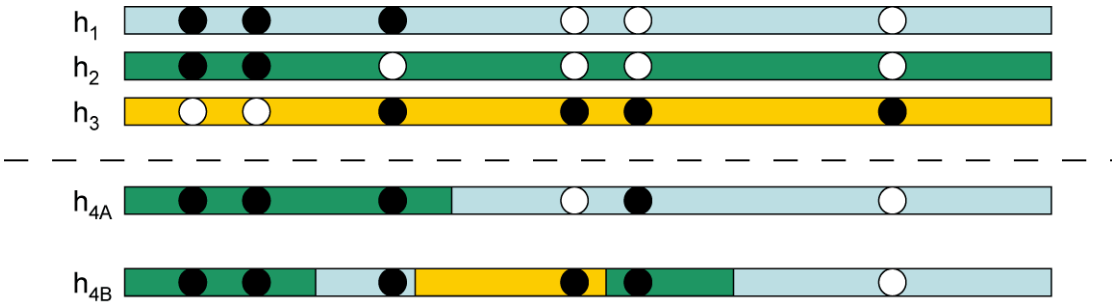


Figure 10. The PAC scheme. Observed SNPs are shown as black or white circles. Given that the first three haplotypes have been observed, we construct the fourth haplotype as an imperfect mosaic of the previous haplotypes, with two possibilities shown here. For example, h_{4A} is copied from h_2 and h_1 with a mutation at the fifth SNP. The likelihood of a given mosaic is a function of the number of recombination and mutation events.

For the initiation of the algorithm, we require the probability of the first observed haplotype. This is calculated by assuming that each SNP allele occurs with probability ½ and is independent of $\boldsymbol{\rho}$. Therefore, for S SNPs, the probability of the first haplotype is given by:

$$\pi\left(H_1\right)=1/2^S\,. \tag{1.22}$$

The conditional distribution of H_{k+1}, given $H_1,\ldots,H_k$, is now calculated using the imperfect mosaic structure. Let X_j be the haplotype that H_{k+1} copies at site j. To mimic the effects of recombination, we model X_j as a Markov chain, with $\Pr(X_1 = x) = 1/k$. Recombination is modelled as a transition from the haplotype currently being copied to another. Taking d_j as the distance between markers j and $j + 1$, the transition of X_j to X_{j+1} occurs with probabilities given by:

$$\Pr\left(X_{j+1}=x'\middle|X_j=x\right)=\begin{cases}\mathrm{e}^{-\rho_j d_j/k}+\left(\dfrac{1}{k}\right)\left(1-\mathrm{e}^{-\rho_j d_j/k}\right) & \text{if } x'=x;\\ \left(\dfrac{1}{k}\right)\left(1-\mathrm{e}^{-\rho_j d_j/k}\right) & \text{otherwise.}\end{cases} \tag{1.23}$$

The above equation attempts to capture the idea that SNPs with only a small distance between them are unlikely to have a recombination event between them. To model the effects of mutation, the haplotypes are copied imperfectly. With probability $k/\left(k+\tilde{\theta}\right)$ the copy is exact, while with probability $\tilde{\theta}/\left(k+\tilde{\theta}\right)$ a 'mutation' occurs. If $H_{i,j}$ denotes the allele of haplotype i at site j then the probability of an allele on a given haplotype and site is:

$$\Pr\left(H_{k+1,j}=a \middle| X_j=x, H_1,\ldots,H_k\right)=\begin{cases} k/\left(k+\tilde{\theta}\right)+\frac{1}{2}\left(\tilde{\theta}/\left(k+\tilde{\theta}\right)\right), & H_{x,j}=a \\ \frac{1}{2}\left(\tilde{\theta}/\left(k+\tilde{\theta}\right)\right), & H_{x,j}\neq a. \end{cases}$$

(1.24)

The mutation rate (per site) used in the PAC model, $\tilde{\theta}$, is fixed to be:

$$\tilde{\theta}=\left(\sum_{i=1}^{n-1}\frac{1}{i}\right)^{-1}. \qquad (1.25)$$

However, we do not know the actual underlying mosaic for each haplotype. Therefore, we have to sum over all possible mosaics, or more formally, over all possible values of X_j. The number of possible paths grows exponentially with the length of the sequence, so brute force evaluation of the equation is not sensible. An alternative method, known as the Forward algorithm, has been developed to sum over all possible states in Hidden Markov Models (see, for example, RABINER 1989). The algorithm is applicable to the summation we wish to achieve. The details of the algorithm do not add anything to this discussion, so we omit the details. Nevertheless, the reader should note that this algorithm allows us to calculate L_{PAC} in time that is linear with the number of SNPs, and quadratic with the number of chromosomes.

A major issue is that the PAC likelihood is dependent on the order in which haplotypes are introduced. In principle, this can be corrected by averaging over all

possible orderings, but this is generally not feasible. The original authors suggest that little is gained by averaging over all possible orderings, and recommend that the likelihood be calculated by averaging over only a few orderings. It is suggested that 20 orderings is sufficient (LI and STEPHENS 2003) as the variability in position of the likelihood maximum is small compared to the uncertainty in the recombination estimates.

Once L_{PAC} has been calculated, it is possible to estimate the recombination rate by taking the maximum likelihood estimate. However, a disadvantage of the basic PAC scheme is that the rate estimates exhibit a systematic bias dependent on SNP density. Although an *ad hoc* correction for this bias which is dependent on the dataset being analysed was suggested by the original authors (LI and STEPHENS 2003), this is a broadly unsatisfying solution, as the correction is entirely empirical.

Despite this, the PAC scheme is known to perform well in many situations including recombination rate estimation (CRAWFORD *et al.* 2004; JEFFREYS *et al.* 2005; LI and STEPHENS 2003; WILSON and MCVEAN 2006). The PAC likelihood also has the advantage over the composite likelihood scheme of being a 'true' likelihood. Therefore, it can be used not only to estimate recombination rates, but also to give estimates of the uncertainty of the estimate. Unfortunately, as the PAC scheme is an approximation of the coalescent model, the uncertainty estimates could themselves be inaccurate.

Hotspot Detection Methods

So far, I have only considered the methods for recombination rate estimation. As was outlined earlier, the majority of recombination events are concentrated in

hotspots. Much of the information regarding recombination rate variation may therefore be obtained by identifying regions that contain hotspots and a number of methods have been developed with this aim. The most successful methods use the same likelihood schemes described in the previous section. As hotspots are of interest in this thesis, I briefly describe some of these methods here.

Fearnhead's Method and *sequenceLDhot*

The most computationally intensive method considered here is Fearnhead's method (FM; FEARNHEAD and SMITH 2005). The method starts by dividing the region under analysis into overlapping windows containing 6 SNPs. For each region, a likelihood surface is calculated using the importance sampling method of Fearnhead and Donnelly described earlier (FEARNHEAD and DONNELLY 2002). The likelihoods are combined in a composite likelihood manner, but with an additional term that penalises the number of hotspots. If $l_i(\rho_i)$ is the estimated likelihood of the i^{th} window, then the pseudo-log-likelihood of the whole region is given by:

$$l_{FM}(\rho) = \sum_{i=1}^{S-5} l_i(\rho_i) - C \cdot h \tag{1.26}$$

where h is the number of hotspots in the region, and C is a constant (generally chosen to be 16; FEARNHEAD and SMITH 2005).

The method determines the location of hotspots by maximisation of $l_{FM}(\rho)$. This is achieved by a recursively adding the single hotspot that causes the largest increase in $l_{FM}(\rho)$. The method stops when adding an additional hotspot does not cause an increase in $l_{FM}(\rho)$.

Due to the complexity of the importance sampling step, this method is relatively computationally expensive. Each six SNP sub region can take between 10-20 minutes to evaluate on a modern computer. However, it is one of the more powerful hotspot detection methods with power to detect average sized hotspots in the region of 65% and a false discovery rate of approximately 2.5% (FEARNHEAD and SMITH 2005).

A related, but more efficient, method has been implemented in the program *sequenceLDhot* (FEARNHEAD 2006). This method considers a small number (~7) of informative SNPs around each possible hotspot location. Using the same importance sampling method, a likelihood ratio is calculated for the presence of a hotspot (with a rate at least 10 times the background rate) at each location. A hotspot is called if the likelihood ratio is greater than some arbitrary threshold (the original paper used a threshold of either 10 or 12 depending on the situation). This method, which I will be using in Chapter 3, appears to have comparable performance to the original penalised likelihood method, but with a substantial reduction in computational cost.

LDhot

The *LDhot* method (MCVEAN *et al.* 2004) uses the pairwise composite likelihood outlined earlier. To detect recombination hotspots, the data is analysed in a window of 200kb, which is moved 1kb at each step. For each window, the maximum composite likelihood is calculated for two models: one in which there is no rate variation in the window, and one in which the central 2kb is allowed a rate greater or equal to the surrounding rate. A test statistic, which is the log of the ratio of the maximum composite likelihoods from the two models, is then calculated.

The distribution of the test statistic under the null hypothesis of no rate variation is calculated by simulation of data using the standard coalescent model. A total of 1000 replicates are used, conditioned on the observed number of samples, the number of segregating sites, the empirical estimate of the recombination rate, and the SNP ascertainment strategy. This distribution is used to obtain the significance of the observed test statistic. A hotspot is called in the central region if there is at least a fivefold increase in the local rate, and the test statistic is statistically significant ($p < 0.001$). When adjacent tests are significant, the hotspot location is chosen to be the point of highest recombination.

Simulations suggest that *LDhot* has reasonably high power (~50-60%), and a very low false positive rate (MYERS *et al.* 2005). It is also computationally efficient and may be applied on a genome wide scale. As such, *LDhot* has been used to identify 25,000 hotspots in the human genome (MYERS *et al.* 2005). However, due to the complexity of the method in terms of parameterisation (particularly the simulation step), *LDhot* has never been released as a stand-alone program.

Hotspotter

The next hotspot detection method that we consider is known as *Hotspotter* (LI and STEPHENS 2003). In this method, successive SNP intervals are tested for hotspots using a likelihood ratio test. In each test, the maximum likelihood of the SNP interval is assessed under the null hypothesis (H_0) of no rate variation, and under the alternative hypothesis (H_1) of a hotspot located in the SNP interval with rate greater than the background rate. The likelihoods in the ratio are calculated using the PAC scheme, but with a bias correction to account for the dependency on SNP density.

Under standard asymptotic theory, twice the log likelihood ratio is (asymptotically) distributed as a chi-square distribution with one degree of freedom.

$$2\ln\left(\frac{L_{PAC}(H_1)}{L_{PAC}(H_0)}\right) \sim \chi^2_1 \tag{1.27}$$

The likelihood ratio test can therefore be used to assess the significance of detected hotspots. If the asymptotic assumption held, then a likelihood ratio greater than 1.92 would give a false positive rate of 5%. Although it seems unlikely that asymptotic theory actually applies, the authors state that using the 1.92 ratio threshold does indeed give false positive rates close to 5%, and hence provides some guidance as to what may be considered 'large' values of the likelihood ratio. Furthermore, simulations in simple settings suggest that with the 1.92 threshold *Hotspotter* has high power (~80%) to detect large hotspots (LI and STEPHENS 2003). However, independent usage of *Hotspotter* in more realistic settings has suggested that while it has good power to detect hotspots, it is also prone to a high false positive rate (JEFFREYS *et al.* 2005).

Li's Method

The final method under consideration I will refer to as Li's Method (LI *et al.* 2006). This method also uses a pairwise composite likelihood, but with an additional weighting. The authors refer to this as the Truncated, Weighted Pairwise Likelihood (TWPL). Originally given in a logarithmic form, the non-logarithmic version of this likelihood is defined as:

$$TWPL(\rho) = \prod_{i,j \in Seg} \left(L_{ij}(\rho)\right)^{w_{j-i}} \tag{1.28}$$

where w_k is a weight dependent on the separation between the i^{th} and j^{th} SNP. The optimal choice of w_k is unclear (FEARNHEAD 2003), so the authors choose to set w_k as follows.

$$w_k = \begin{cases} 1/k & \text{if } k \leq 7 \\ 0 & \text{otherwise} \end{cases} \tag{1.29}$$

The choice of k ≤ 7 is arbitrary and is made for convenience.

Having defined the likelihood, the method detects hotspots in the same manner as Fearnhead's Method; that is by recursively adding non-overlapping hotspots that allow give the largest increase in the likelihood.

Simulations suggest that this method achieves power in the region of 60-70% with a false positive rate of 0.4 per Mb. Furthermore, the method uses the composite likelihood and is therefore fast enough to be applied to large datasets. In the original paper, the method was applied to 5Mb of data from the ENCODE project (LI *et al.* 2006). This method is computationally efficient and publicly available (although only in precompiled executable form). In Chapter 3, I compare the performance of this method to that of *sequenceLDhot* and my new method.

Discussion

The aim of this chapter was to introduce recombination from a population genetic perspective. I introduced the process of meiotic recombination before proceeding to describe the experimental methods by which recombination rates can be estimated. I noted that recombination rates vary over both the broad and the fine scale, and described the properties of recombination hotspots. The signature that is left by

recombination in patterns of genetic diversity was discussed, along with some commonly used statistics. However, relating these patterns to the underlying recombination rate is complex and requires a description of the evolutionary process that generated them. This leads to the introduction of a commonly-used model of the evolutionary process; the coalescent. I described how the coalescent could be used to calculate the probability of observing a given dataset, and how this may be used to determine the underlying recombination rate. I then outlined some successful methods for both recombination rate estimation and hotspot detection.

In the following chapter, I describe a new method which both estimates recombination rates, and estimates the location and properties of hotspots.

Chapter 2 A New Method for Recombination Rate Estimation

As described in the previous chapter, statistical analysis of population genetic data provides an alternative to experimental methods for estimating recombination rates. A number of methods have been proposed for estimating the population genetic recombination rate (FEARNHEAD and DONNELLY 2001; LI and STEPHENS 2003; MCVEAN *et al.* 2004; WALL 2000). However, the majority of available methods either assume a constant recombination rate across the region, or cannot be applied on a genome-wide scale.

In an attempt to address these issues, a fast method was developed by McVean *et al*. for the estimation of variation in recombination rates at the fine scale (MCVEAN *et al.* 2004). This method, distributed in the *LDhat* program, used a coalescent model to obtain an approximation to the likelihood of the population genetic data; specifically, the pairwise composite likelihood was used. Despite the likelihood approximation, simulations have shown that the *LDhat* produces robust and largely unbiased rate estimates (MCVEAN 2007; MCVEAN *et al.* 2002; MCVEAN *et al.* 2004). A further advantage of *LDhat* is that it is currently one of only a few available population based methods for recombination rate estimation that can be applied to genome-wide samples containing large numbers of chromosomes. The application of *LDhat* to large datasets has established that hotspots are a ubiquitous feature of the human genome, with between 25,000 and 50,000 expected to exist (MCVEAN *et al.* 2004; MYERS *et al.* 2005), and has provided a number of insights into the relationship

between recombination and other genome features (MYERS *et al.* 2005; MYERS *et al.* 2007; MYERS *et al.* 2006).

However, no model of recombination hotspots was included in the prior model of *LDhat*, and hence the true level of heterogeneity implied by the presence of recombination hotspots was not well captured. In this chapter, I describe a replacement of the *LDhat* prior with one that includes a description of recombination hotspots. By incorporating a hotspot model, it is expected that the accuracy of rate estimates can be improved. Furthermore, the new method can be used to simultaneously estimate the locations of recombination hotspots as part of the rate estimation procedure.

The Composite Likelihood Revisited

In this thesis, the parameter of primary interest is the population recombination rate $\boldsymbol{\rho} = 4N_e\boldsymbol{r}$, where N_e is the effective population size, and $\boldsymbol{r}$ is the map of the sex-averaged recombination rate (expressed in terms of expected cross-over events per generation per kilobase between adjacent SNPs). Given a genetic sample of a population, we would like to make inferences about $\boldsymbol{\rho}$. To do so, we need to calculate the likelihood of the data, $P(D \mid \zeta)$, where D is the data (the haplotypes or genotypes in our sample) and ζ represents our model parameters. However, calculating the full likelihood of the data under the coalescent model is computationally prohibitive on all but the smallest of datasets (FEARNHEAD and DONNELLY 2001). I therefore have adopted a method for calculating an approximation to the full likelihood, known as the composite likelihood, which was described briefly in the previous chapter, and which I describe in more detail now.

The composite likelihood scheme (HUDSON 2001; MCVEAN *et al.* 2002) considers only pairs of segregating sites, or SNPs, in the data. For each pair of SNPs, a coalescent model is used to calculate a likelihood surface over a range of recombination rates. A pseudo-likelihood is then constructed as the product of the likelihood over all pairs of SNPs in the region under consideration. Compared to full-likelihood approaches, the required computation is reduced by many orders of magnitude, making the composite scheme suitable for much larger datasets.

As the first stage of the composite scheme, a population mutation rate is estimated using an approximate finite-sites version of the Watterson estimate (MCVEAN *et al.* 2002; WATTERSON 1975). Given n sampled gene sequences of length L, with S segregating sites, the population mutation rate per site is estimated using:

$$\hat{\theta}_W^* = \left(\sum_{i=1}^{n-1} \frac{1}{i} \right)^{-1} \ln\left(\frac{L}{L-S} \right). \qquad (1.30)$$

In the second stage, every pair of sites with only two alleles are grouped into equivalent sets. As an example, suppose we have five sequences. For one pair of SNPs, the haplotypes are (AA, AT, TA, TA, AA) which have the minor allele 'T' at both sites. For a separate pair of SNPs, the haplotypes are (GG, CC, CG, GG, CG), which have the minor allele 'G' at the first site, and 'C' at the second site. These sets are both equivalent to the unordered set (11, 10, 01, 01, 11), where 0 represents the minor allele *at each site*. The number of sets is clearly dependent on the number of sequences and the variability in the dataset. Assuming that every possible combination occurs in a dataset, the number of uniquely identifiable sets scales with an order of n^3 (MCVEAN *et al.* 2002).

The third stage is to estimate the likelihood of each set. This is achieved using the Importance Sampling method of Fearnhead and Donnelly (FEARNHEAD and

DONNELLY 2001). Informally, a large number of genealogies are generated for each set at the assumed mutation rate using a stochastic process (and allowing for reverse mutation), and over a range of recombination rates (a typical range would be $0 \leq \rho \leq 100$). The likelihood at each recombination rate is calculated by averaging over the importance weights of the sampled genealogies. This method is not usually tractable for large datasets due to the large number of genealogies that need to be generated; however, by only considering sets of equivalent pairs of SNPs, the method becomes practical for datasets containing hundreds of sequences and thousands of SNPs. In this way, it is possible to pre-calculate and store likelihood tables for any dataset of a given number of haplotypes.

Finally, given the pre-calculated likelihood surface, we can calculate a pseudo-likelihood of the data using an assumed recombination map. To construct the pseudo-likelihood, I assume that pairs of SNPs are independent of each other (which in reality they are not). In the original *LDhat* implementation, given a vector of recombination rates, $\boldsymbol{\rho}$, in which the i^{th} element gives the recombination rate between the i^{th} and $(i+1)^{\text{th}}$ SNP, the composite likelihood is given by:

$$P_C(D \mid \boldsymbol{\rho}) = \prod_{i,j} P(D_{ij} \mid \rho_{ij}) \qquad (1.31)$$

where $P(D_{ij} \mid \boldsymbol{\rho}_{ij})$ is the likelihood of the data at segregating sites i and j given a recombination rate of ρ_{ij} between them (HUDSON 2001). This approximation to the true likelihood surface is required to keep the computational cost down. Nevertheless, the vast majority of the computational cost of the composite scheme is contained in the Importance Sampling section. Fortunately, likelihood tables have been pre-calculated for a variety of possible datasets of up to 192 chromosomes, and are available for download (http://www.stats.ox.ac.uk/~mcvean/LDhat/).

A major advantage of the composite scheme is the ability to use genotype data. As only pairs of SNPs are considered, genotype data can be considered by summing over all possible phases of each SNP pair. In a similar manner, the scheme can incorporate missing data – although the efficiency of the algorithm does not scale well with increasing amounts of missing data and loci with more than approximately 10% of missing data should generally be discarded.

Compared to full-likelihood methods, the likelihood surface of the composite scheme tends to be sharply peaked. However, the maximum-likelihood estimates of the two methods are well correlated (McVEAN *et al.* 2002). Unfortunately, the peaked nature of the composite scheme can be unsuitable for use with rjMCMC, as the chain may become stuck in a local maximum. To compensate for this, the original *LDhat* implementation introduced user-specified 'block penalties', which in some sense increased the strength of the prior relative to the composite likelihood.

I have taken an alternative approach. I also found that the original composite likelihood severely limited the mixing of the method. In a similar fashion to *LDhat*, I explored the possibility of using penalties to strengthen the prior relative to the composite likelihood. However, choosing suitable penalties was troublesome, as the suitability of a given set of penalties was dependent on the dataset under analysis. I therefore informally investigated adaptations of the composite likelihood that would in some sense 'flatten' the likelihood surface and hence allow the method to mix well. Given S SNPs, a suitable alternative to equation (1.31) is given by:

$$P_C\left(D \mid \boldsymbol{\rho}\right) = \sqrt[S-1]{\prod_{i,j} P\left(D_{ij} \mid \rho_{ij}\right)}. \qquad (1.32)$$

Intuitively, the correction can be thought of as a correction for the inherent double counting in the composite likelihood. In the case of $\rho = \infty$, the original composite likelihood is equal to the $(S\text{-}1)^{\text{th}}$ power of the full likelihood, due to each SNP interval

being considered multiple times. The $(S\text{-}1)^{th}$ root was therefore chosen as a suitable correction, although it will tend to over-flatten the likelihood for small recombination rates.

In order to maintain the computational feasibility of the method, I do not consider the contribution to the composite likelihood from SNPs separated by more than 50 intermediate SNPs. That is I assume $P(D_{ij} \mid \rho_{ij}) = 1$ if $|i\text{-}j| > 50$ and adjust the root in equation (1.32) accordingly. The choice of 50 SNPs is arbitrary, but it was found that using larger subsets did not significantly improve the results (data not shown). Furthermore, there are both theoretical and empirical studies which suggest that limiting the number of SNPs may actually improve the performance of the estimator (FEARNHEAD 2003; SMITH and FEARNHEAD 2005).

Obtaining Estimates of the Recombination Rate

So far, I have only considered how to calculate a pseudo-likelihood of a given dataset. I have said little about how to use that likelihood to perform inference about the recombination rate. Given a likelihood curve, there are a number of possibilities to obtain an estimate of the parameter of interest. For example, in classical statistics, we may obtain a point-estimate of our parameter of interest by finding the value of the parameter that maximises the likelihood function; the so-called Maximum Likelihood Estimate (MLE). Alternatively, in Bayesian statistics, we may wish to incorporate information regarding our prior belief about the parameter of interest. In this thesis, the parameter of interest is the underlying recombination rate $\boldsymbol{\rho}$ (which may or may not be constant over the region of study). If $P(\boldsymbol{\rho})$ describes our prior belief about $\boldsymbol{\rho}$,

and $P(D|\boldsymbol{\rho})$ is the conditional probability of our data given $\boldsymbol{\rho}$, we obtain a posterior distribution on $\boldsymbol{\rho}$ via Bayes' theorem:

$$P(\boldsymbol{\rho} \mid D) = \frac{P(D \mid \boldsymbol{\rho}) P(\boldsymbol{\rho})}{\int_{\rho} P(D \mid \boldsymbol{\rho}) P(\boldsymbol{\rho}) d\boldsymbol{\rho}}. \tag{1.33}$$

The posterior describes our updated belief about $\boldsymbol{\rho}$ having observed the data. It is worth noting that the denominator of (1.33) is constant, and therefore:

$$P(\boldsymbol{\rho} \mid D) \propto P(D \mid \boldsymbol{\rho}) P(\boldsymbol{\rho}) = \text{likelihood x prior} \tag{1.34}$$

While in simple situations we may be able to calculate the posterior distribution directly, in many practical situations we are unable to do so (for example, the integral in (1.33) may be difficult to evaluate). In such cases, we may resort to a popular method known as Markov Chain Monte Carlo (MCMC). In brief, a Markov chain is initiated using values drawn at random. At each iteration of the method, one or more of the chain parameters are updated according to a proposal distribution. The proposed parameter values are accepted with certain acceptance probabilities, which are chosen so that the Markov chain explores the target (posterior) distribution. In the popular Metropolis-Hastings algorithm (HASTINGS 1970; METROPOLIS *et al.* 1953), the acceptance probabilities of a move from the current state, $\boldsymbol{\rho}$, to a new state, $\boldsymbol{\rho}'$, are given by:

$$\alpha(\boldsymbol{\rho} \to \boldsymbol{\rho}') = \min\left\{1, \frac{P(D \mid \boldsymbol{\rho}')}{P(D \mid \boldsymbol{\rho})} \frac{P(\boldsymbol{\rho}')}{P(\boldsymbol{\rho})} \frac{q(\boldsymbol{\rho}' \to \boldsymbol{\rho})}{q(\boldsymbol{\rho} \to \boldsymbol{\rho}')}\right\} \tag{1.35}$$

where $q(\boldsymbol{\rho}' \to \boldsymbol{\rho})$ is the proposal kernel density. The three ratio terms in (1.35) are referred to as the likelihood, prior and proposal ratios respectively.

The Metropolis-Hasting algorithm has been applied with great success to many complex problems in a wide range of fields (GILKS *et al.* 1996). However, the algorithm assumes that the parameter of interest is of fixed dimensionality.

Consequently, the Metropolis-Hasting algorithm can only be used in situations where the dimensionality of the parameter is known in advance.

However, in many situations (including our own), the dimensionality of the parameter of interest may not be known *a priori*. This situation has been addressed with the development of Reversible Jump Markov Chain Monte Carlo (rjMCMC; GREEN 1995). In rjMCMC, transitions are allowed between models with differing dimensionality. If ρ_1 is the current state of the Markov chain in parameter space 1, a new state ρ_2' may be proposed in parameter space 2, using a number of random deviates, u. The move is accepted with probability:

$$\alpha(\rho_1 \to \rho_2') = \min\left\{1, \frac{P(D \mid \rho_2')}{P(D \mid \rho_1)} \frac{P(\rho_2')}{P(\rho_1)} \frac{q(\rho_2' \to \rho_1)}{q(\rho_1 \to \rho_2')} \left| \frac{\partial(\rho_2')}{\partial(\rho_1, u)} \right| \right\}. \qquad (1.36)$$

The final term of (1.36) is known as the Jacobian determinant, which relates the parameters in space 2 to those in space 1 and the random deviates. For many types of move the Jacobian determinant reduces to unity, and the acceptance probability therefore reverts to the Metropolis-Hastings case.

In *LDhat*, the rjMCMC was then used in conjunction with the pairwise composite likelihood to obtain a pseudo-posterior estimate of the recombination rate. The recombination rate was assumed to vary in a piecewise-constant fashion with an exponential prior on the rate within a block. Simple Metropolis-Hastings moves were used to explore the rate estimates within piecewise blocks, and move change-points. Reversible jump moves were used to vary the number of piecewise blocks.

However, the prior on recombination rate variation used in *LDhat* (i.e. recombination rates varied in a piecewise constant fashion) is a poor model of the true levels of variation. In the following section, I describe a much more sophisticated

prior model (and hence rjMCMC scheme), which I will subsequently use to obtain recombination rate estimates.

Priors on Recombination Rate Variation

To obtain a pseudo-posterior distribution on $\boldsymbol{\rho}$, the *LDhat* method imposed a prior of piecewise-constant structure with constant recombination rate over SNP intervals and change-points located only at SNPs. In the new scheme, I maintain a similar structure for the estimation of background recombination rates, with the exception that change-points are no longer restricted to SNP locations. The major novelty of the method comes from the incorporation of a hotspot model. I model hotspots as sharp peaks in the recombination rate with a double exponential shape. Under my prior model, hotspots are uniformly scattered along the analyzed region with the number of hotspots and their properties (such as position, magnitude and width) determined as part of the rjMCMC scheme. To illustrate the differences between the *LDhat* prior and the new prior, I have generated individual realizations of each (Figure 11). I encourage the reader to compare these realizations of the prior with the sperm-typing rate estimates of the MHC and MS32 regions described in Chapter 1.

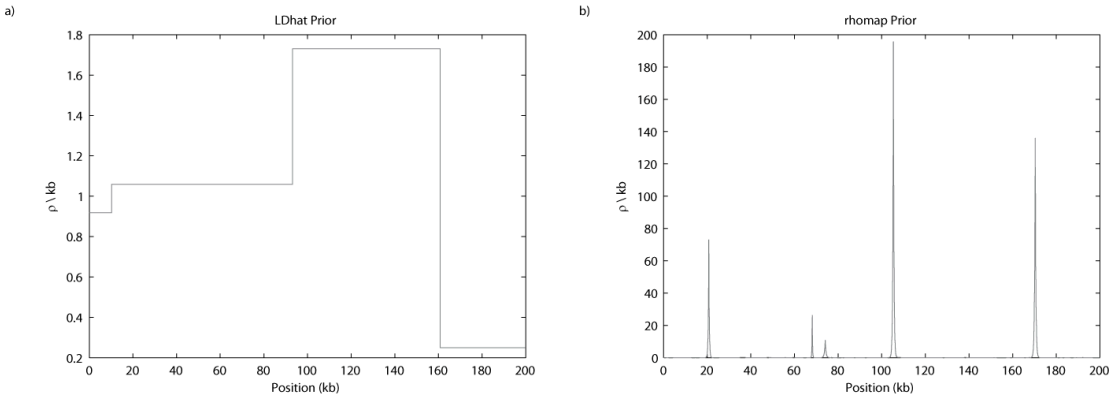


Figure 11. Illustration of the priors of *LDhat* (a) and *rhomap* (b). Shown here are individual realisations of a 200kb region. Note the difference in the y-axis scale.

Prior on Background Rate Variation

I aim to estimate a piecewise constant rate function, over the data range [0, L], where 0 and L are the positions of the first and last SNPs in our data respectively. The prior for background rate variation is very similar to that used in *LDhat*, but includes a small change to encourage spacing between change-points. I suppose there are k change-points in the piecewise function, with change-points at locations s_j where $0 < s_1 < s_2 < \ldots < s_k < L$, and that the recombination rate takes the value h_j on the subinterval, or block, $[s_j, s_{j+1}]$ for $j = 0, 1, 2, \ldots, k$, with $s_0 = 0$ and $s_{k+1} = L$. Given k, the change-point locations are distributed as the even-number order statistics from $2k+1$ points uniformly distributed on [0, L]. The probability density function is therefore given by:

$$p(s_1,\ldots,s_k) = \frac{(2k+1)!}{L^{2k+1}}(s_1 - 0)(s_2 - s_1)\ldots(s_k - s_{k-1})(L - s_k) \qquad (1.37)$$

The advantage of using this prior over a uniform distribution is that small intervals are penalised, and hence the background rate change-points are separated probabilistically. I define the prior on the number of change-points, k, as a Poisson distribution with mean $\gamma = \min(L, N\text{-}2)$, where L is measured in kilobases. The initial value of k is set equal to the number of SNPs in the data minus two. Furthermore, the initial change-points are set to be the internal SNP positions, but are subsequently allowed to vary between SNPs.

I define the prior on the block heights as an exponential distribution with mean ϕ:

$$p\left(h_j\right) = \frac{1}{\phi}\exp\left(-\frac{h_j}{\phi}\right). \tag{1.38}$$

Again note that this is very similar to the prior in *LDhat*. However, whereas *LDhat* used this prior to describe all recombination rate variation, I am only using to explain background variation. Therefore, the mean rate ϕ is generally much lower in the new method compared to *LDhat*.

Prior on Hotspots

Hotspots are uniformly scattered over the interval [0, L]. I define the total contribution to the recombination rate by the hotspot as λ, which I call the hotspot heat. I define the morphology of the hotspot to be a truncated double-exponential curve with scale μ and I define the width of a hotspot to be the region in which 95% of the hotspot mass is contained. While I accept that the double-exponential curve may not reflect the true hotspot morphology, the resolution of SNPs in most datasets make the determination of the true morphology impossible. The choice of the double-

exponential curve is made for convenience and is consistent with current experimental data - the true morphology of hotspots is currently undetermined (JEFFREYS *et al.* 2001; JEFFREYS and NEUMANN 2002).

For efficient implementation, it is important that hotspots only contribute to the recombination map over a finite region. For this reason, I truncate the tails of the hotspot either at a change-point, or some arbitrary distance m from the hotspot peak (at which point the contribution from the hotspot in negligible), as shown in Figure 12. The mass that would be lost by the truncation of the tails is redistributed uniformly in the body of the hotspot via a function ψ. Note that the function ψ allows hotspots to be non-symmetric. If the maximum allowed width of a hotspot is $2m$, the recombination rate at position X ($s_j \le X < s_{j+1}$) is given by:

$$\begin{cases} h_j + \dfrac{\lambda}{2\mu}\exp\left(-\dfrac{|X-t|}{\mu}\right) + \psi(X); & \max(s_j, t-m) \le X < \min(s_{j+1}, t+m) \\ h_j; & \text{otherwise} \end{cases} \quad (1.39)$$

The first term of equation (1.39) is the contribution from the background rate. The second term gives the morphology of the hotspot. The correction function, ψ, due to the maximum allowed width of the hotspot is given by:

$$\psi(X) = \begin{cases} \dfrac{\lambda}{4\mu\left(t - \max(s_j, t-m)\right)} \displaystyle\int_{-\infty}^{\max(s_j, t-m)} \exp\left(-\frac{|y-t|}{\mu}\right) dy & X < t \\ \dfrac{\lambda}{4\mu\left(\min(s_{j+1}, t+m) - t\right)} \displaystyle\int_{\min(s_{j+1}, t+m)}^{+\infty} \exp\left(-\frac{|y-t|}{\mu}\right) dy & X \ge t \end{cases} \quad (1.40)$$

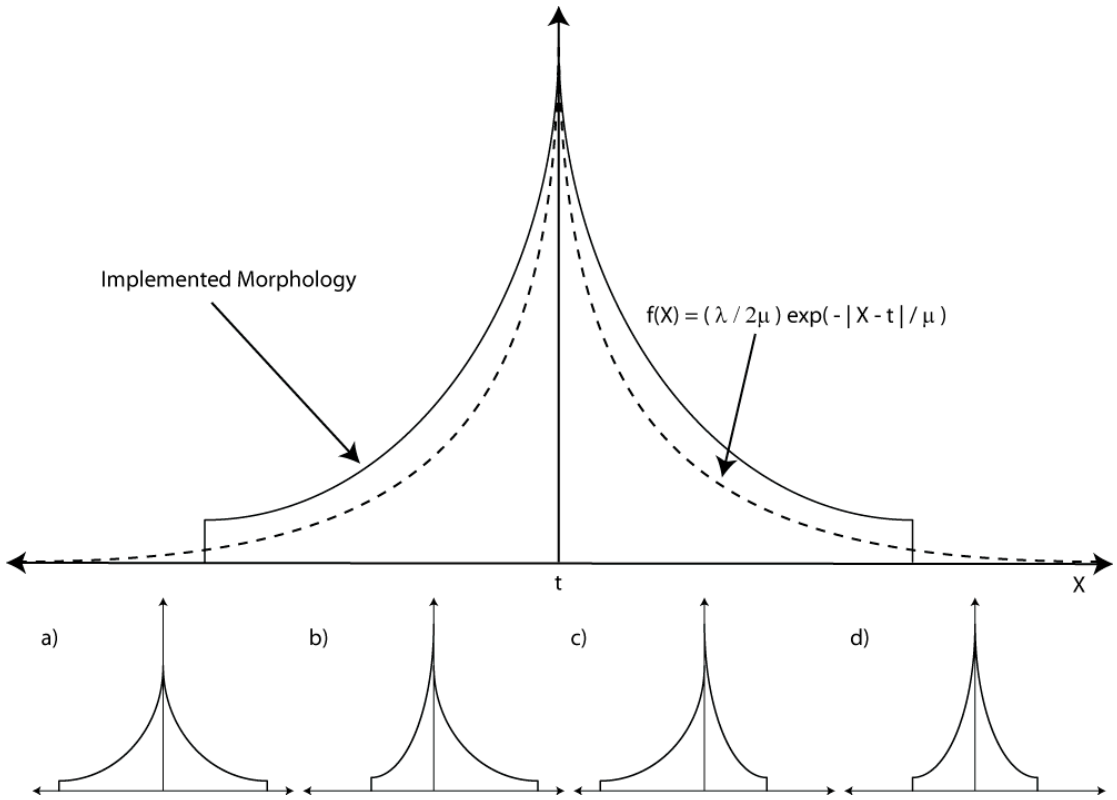


Figure 12. Implemented hotspot morphology. This diagram shows how the implemented morphology relates to the double-exponential curve. Note that the mass of the truncated tails is added to the main body of the distribution. Also shown are four scenarios demonstrating how the morphology changes with respect to background rate change points. a) No background rate change points in near vicinity – hotspot extends to maximum allowed width. b) Background rate change point in near vicinity to left of hotspot. c) Background rate change point in near vicinity to right of hotspot. d) Hotspot bounded by two change points in near vicinity. Note that the integral of each hotspot over the total width is the same.

The reader should note that m is a fixed parameter (which can be altered by the user) and is not estimated by the MCMC scheme. It is included so that the contribution to the map from each hotspot need only be calculated over a small range, and hence can be calculated efficiently. In general, the contribution to the map at the distance m from the hotspot centre is not significant and can therefore be ignored. The shape of the hotspot is, however, controlled by the scale parameter, μ, which is estimated by the MCMC scheme. A small μ corresponds to a hotspot with a highly

concentrated peak, whereas a large μ corresponds to a hotspot with a less concentrated peak (Figure 13). In practical situations with m sufficiently large, the width of the hotspot is determined either by the surrounding change-points, or by the scale parameter, μ, and not by m.

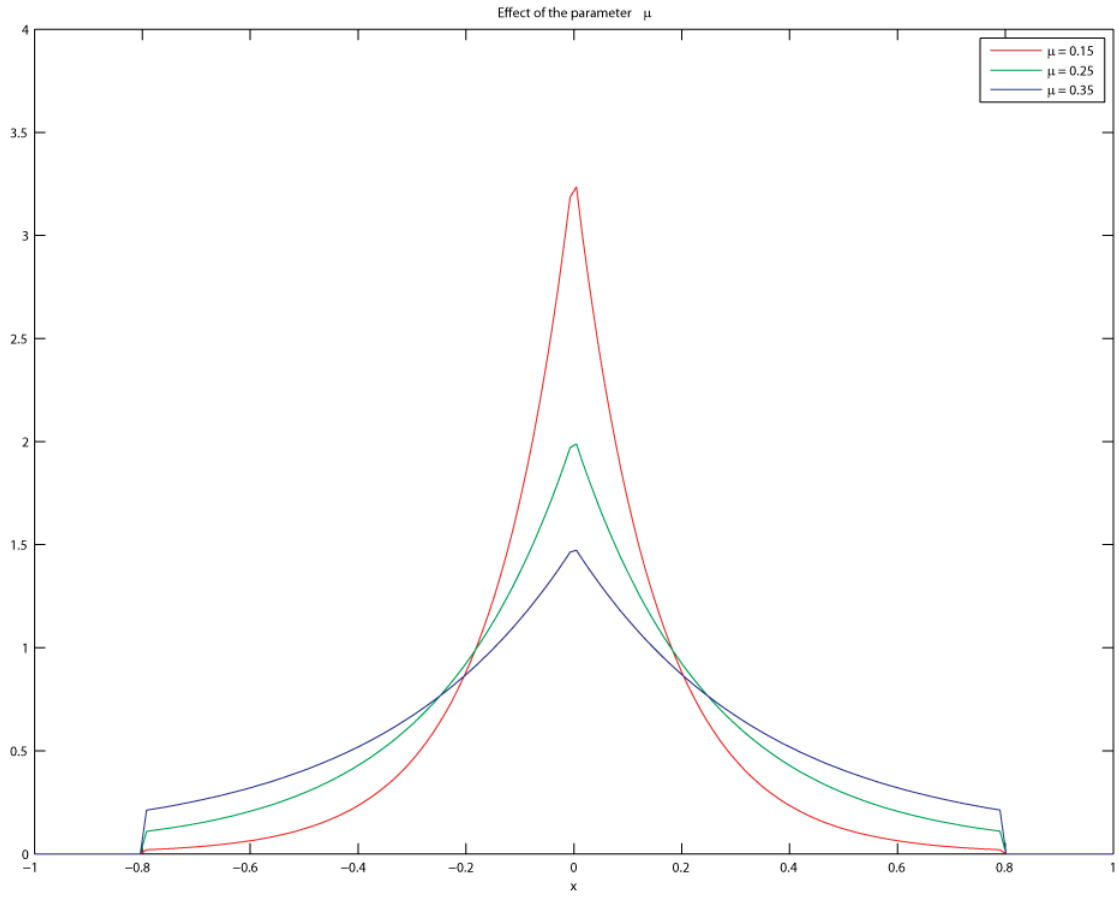


Figure 13. Illustration demonstrating the change in the μ scale parameter. Shown here is the implemented morphology without interference from background blocks. For illustration, I have used λ = 1, and m = 0.8 (which is artificially small as way of demonstration). The reader should note that the area under each curve is the same – the μ parameter has no effect on the total contribution to the recombination map from the hotspot.

The prior on the hotspot scale parameter, μ_i, is a gamma distribution with parameters α_1 and β_1:

$$\Pr(\mu_i) = \frac{1}{\beta_1^{\alpha_1}\Gamma(\alpha_1)}(\mu_i)^{\alpha_1 - 1} e^{-\frac{\mu_i}{\beta_1}}. \quad (1.41)$$

The prior on the hotspot heats, λ_i, is defined as a gamma distribution with parameters α_2 and β_2:

$$\Pr(\lambda_i) = \frac{1}{\beta_2^{\alpha_2}\Gamma(\alpha_2)}(\lambda_i)^{\alpha_2 - 1} e^{-\frac{\lambda_i}{\beta_2}}. \quad (1.42)$$

In the initialization, the number of hotspots is zero. Hotspot positions are drawn uniformly on the range [0, L]. The prior number of hotspots is given by a Poisson distribution with mean $\omega = \min(L/T, N\text{-}2)$, where T is the expected distance between hotspots.

rjMCMC Move Definitions

In developing the rjMCMC scheme, I was guided by intuition in the design of appropriate moves. I therefore do not claim that the choices made are optimal. The possible rjMCMC moves are:

A. Change the rate of a background block

B. Move a background rate change-point

C. Split a background block

D. Merge two background blocks

E. Change a hotspot heat

F. Change a hotspot scale parameter

G. Move a hotspot

H. Insert a hotspot

I. Delete a hotspot

I now consider each move in more detail in the following section. For each move I refer to the 'likelihood ratio', which is defined as the likelihood of the data given the proposed rate function divided by the likelihood of the data given the current rate function. As discussed in the main text, the likelihoods are calculated using an approximation to the coalescent likelihoods.

Move A: Change the rate of a background block

Block j is chosen uniformly from the $(k + 1)$ background blocks. A new height is proposed so that $h_j' = h_j \exp(u)$ where $u \sim$ Uniform(-½, ½). The acceptance probability is given by:

$$\min\left[1, \frac{P(D \mid \Theta')}{P(D \mid \Theta)} \frac{h_j'}{h_j} \exp\left(\frac{h_j - h_j'}{\phi}\right)\right]. \qquad (1.43)$$

Move B: Move a background rate change-point

Change-point j is chosen uniformly from the k change-points. A new position is chosen so that $s'_j = u$ where $u \sim \text{Uniform}(s_{j-1}, s_{j+1})$. Either h_{j-1} or h_j is altered to h' with equal probability so that the total recombination map over the region $[s_{j-1}, s_{j+1}]$ is unchanged. Moves that imply a rate less than zero are rejected. Given h is the original block height, the acceptance probability of this move is given by

$$\min\left[1, \frac{P(D\mid\Theta')}{P(D\mid\Theta)}\exp\left(\frac{h-h'}{\phi}\right)\frac{(s_{j+1}-s'_j)(s'_j-s_{j-1})}{(s_{j+1}-s_j)(s_j-s_{j-1})}\right]. \qquad (1.44)$$

Move C and D: Split a background block / Merge two background blocks

Moves C and D are proposed with relative probabilities $P(C_k)$ and $P(D_k)$, where

$$\frac{P(C_k)}{P(D_k)} = \frac{\min\{1, P(k+1)/P(k)\}}{\min\{1, P(k-1)/P(k)\}} = \frac{\min\{1, \gamma/(k+1)\}}{\min\{1, k/\gamma\}} \qquad (1.45)$$

When proposing a new change-point, a position, s^*, is chosen uniformly over the region $[0, L]$. This must lie within an interval of an existing region, say $[s_j, s_{j+1}]$, which I refer to as block j. If the move is accepted, then the change-points s^*, s_{j+1}, s_{j+2}, …, s_k are relabelled as s_{j+1}, s_{j+2}, s_{j+3} , …, s_{k+1}, with corresponding changes to the block height labels. New block heights, h'_j and h'_{j+1} are proposed so that the total contribution to the map from the background rate over the original interval $[s_j, s_{j+1}]$ is unchanged, as the current estimate of recombination rate is likely to be respectable. The condition is therefore:

$$(s^* - s_j)h'_j + (s_{j+1} - s^*)h'_{j+1} = (s_{j+1} - s_j)h_j \tag{1.46}$$

A perturbation is defined so that

$$\frac{h'_{j+1}}{h'_j} = \frac{1-u}{u} \tag{1.47}$$

where $u \sim$ Uniform(0,1). Substituting gives

$$\begin{aligned} h'_j &= \frac{uh_j(s_{j+1} - s_j)}{u(s^* - s_j) + (1-u)(s_{j+1} - s^*)} \\ h'_{j+1} &= \frac{(1-u)h_j(s_{j+1} - s_j)}{u(s^* - s_j) + (1-u)(s_{j+1} - s^*)}. \end{aligned} \tag{1.48}$$

Note that the hotspots remain unchanged. The acceptance probability of the Split move is given by:

$$\min\left[1, \frac{P(D|\Theta')}{P(D|\Theta)} \frac{\gamma(4k+6)}{\phi L(k+1)} \frac{(s^* - s_j)(s_{j+1} - s^*)}{(s_{j+1} - s_j)} \frac{e^{(-h'_j - h'_{j+1})/\phi}}{e^{-h_j/\phi}} \frac{P(D_{k+1})}{P(C_k)} \frac{(h'_j + h'_{j+1})^2}{h_j}\right] \tag{1.49}$$

.

For the reverse move of removing a change-point, change-point j is chosen uniformly from the k change-points. We merge the block to the left of the change-point, block j-1, with the block to the right, block j. To do this, block j is removed, and a new rate, h'_{j-1}, is proposed for block j-1 over the region $[s_{j-1}, s_{j+1}]$. In order to achieve detailed balance, the calculations of the Split move must be reversed to obtain h'_{j-1}:

$$(s_j - s_{j-1})h_{j-1} + (s_{j+1} - s_j)h_j = (s_{j+1} - s_{j-1})h'_{j-1}. \tag{1.50}$$

The acceptance probability of the Merge move is given by:

$$\min\left[1, \frac{P(D|\Theta')}{P(D|\Theta)} \frac{\phi Lk}{\gamma(4k+2)} \frac{(s_{j+1} - s_{j-1})}{(s_j - s_{j-1})(s_{j+1} - s_j)} \frac{e^{-h'_{j-1}/\phi}}{e^{(-h_{j-1} - h_j)/\phi}} \frac{P(C_{k-1})}{P(D_k)} \frac{h'_{j-1}}{(h_{j-1} + h_j)^2}\right] \tag{1.51}$$

.

Move E: Change a hotspot heat

Hotspot i is chosen uniformly from the K hotspots. A new heat is proposed so that $\lambda_i' = \lambda_i \exp(u)$ where $u \sim$ Uniform(-½ , ½). The move is accepted with probability

$$\min\left[1, \frac{P(D|\Theta')}{P(D|\Theta)}\left(\frac{\lambda_i'}{\lambda_i}\right)^{\alpha_2} \exp\left(\frac{\lambda_i - \lambda_i'}{\beta_2}\right)\right]. \quad (1.52)$$

Move F: Move a hotspot

Hotspot i is chosen uniformly from the K hotspots. A new position is proposed so that $t'_i = t_i + u$ where $u \sim$ Normal(0, σ_F^2). The move is rejected if the proposed hotspot position is outside the range [0, L]. Otherwise, the acceptance probability is given by

$$\min\left[1, \frac{P(D|\Theta')}{P(D|\Theta)}\right]. \quad (1.53)$$

Move G: Change a hotspot scale parameter

Hotspot i is chosen uniformly from the K hotspots. A new scale parameter is proposed so that $\mu'_i = \mu_i \exp(u)$ where $u \sim$ Uniform(-½, ½). This move is accepted with probability

$$\min\left[1, \frac{P(D|\Theta')}{P(D|\Theta)}\left(\frac{\mu_i'}{\mu_i}\right)^{\alpha_1}\exp\left(\frac{\mu_i-\mu_i'}{\beta_1}\right)\right]. \quad (1.54)$$

Moves H and I: Insert a hotspot / Delete a hotspot

Moves H and I are proposed with relative probabilities $P(H_K)$ and $P(I_K)$, where

$$\frac{P(H_k)}{P(I_k)} = \frac{\min\{1, P(K+1)/P(K)\}}{\min\{1, P(K-1)/P(K)\}} = \frac{\min\{1, \omega/(K+1)\}}{\min\{1, K/\omega\}} \quad (1.55)$$

The position of a new hotspot is chosen uniformly on the region [0, L]. The heat and scale is drawn from the prior. The acceptance probability is given by:

$$\min\left[1, \frac{P(D|\Theta')}{P(D|\Theta)}\frac{\omega}{(K+1)}\frac{P(I_{K+1})}{P(H_k)}\right]. \quad (1.56)$$

For the corresponding delete move, hotspot i is chosen uniformly from the K hotspots. The acceptance probability is given by

$$\min\left[1, \frac{P(D|\Theta')}{P(D|\Theta)}\frac{K}{\omega}\frac{P(H_{K-1})}{P(I_K)}\right]. \quad (1.57)$$

Prior Parameter Choices

In total, there a seven prior parameters and one rjMCMC move parameter (Table 2). I have attempted to parameterise the model for human data by using the sperm analysis of the MHC and MS32 regions as guidance (Figure 14). However, as the available data from sperm studies is relatively sparse, I make no claim that parameter values are optimal. The hotspot prior parameters λ and μ were obtained by

using a maximum likelihood method to fit a gamma distribution to the hotspot properties estimated via sperm typing (see Figures 3 and 4 in Chapter 1). The expected contribution to the recombination map from a hotspot using these parameters is ρ = 32.1, and an expected width of 1.5kb. The expected distance between hotspots, *T*, was set to 50kb, which would give a total number of hotspots in the human genome as 60,000 – in line with previous predictions (MYERS *et al.* 2005). The background rate parameter was selected to give an average recombination rate approximately equal to the genome-wide average. The *m* parameter was chosen to be sufficiently large (5kb) that the remaining contribution to the recombination map from the tails of the hotspot is generally negligible. The random deviate parameter, σ_F, was selected to provide adequate mixing of the MCMC. In this thesis, the parameters in Table 2 should be assumed unless otherwise stated.

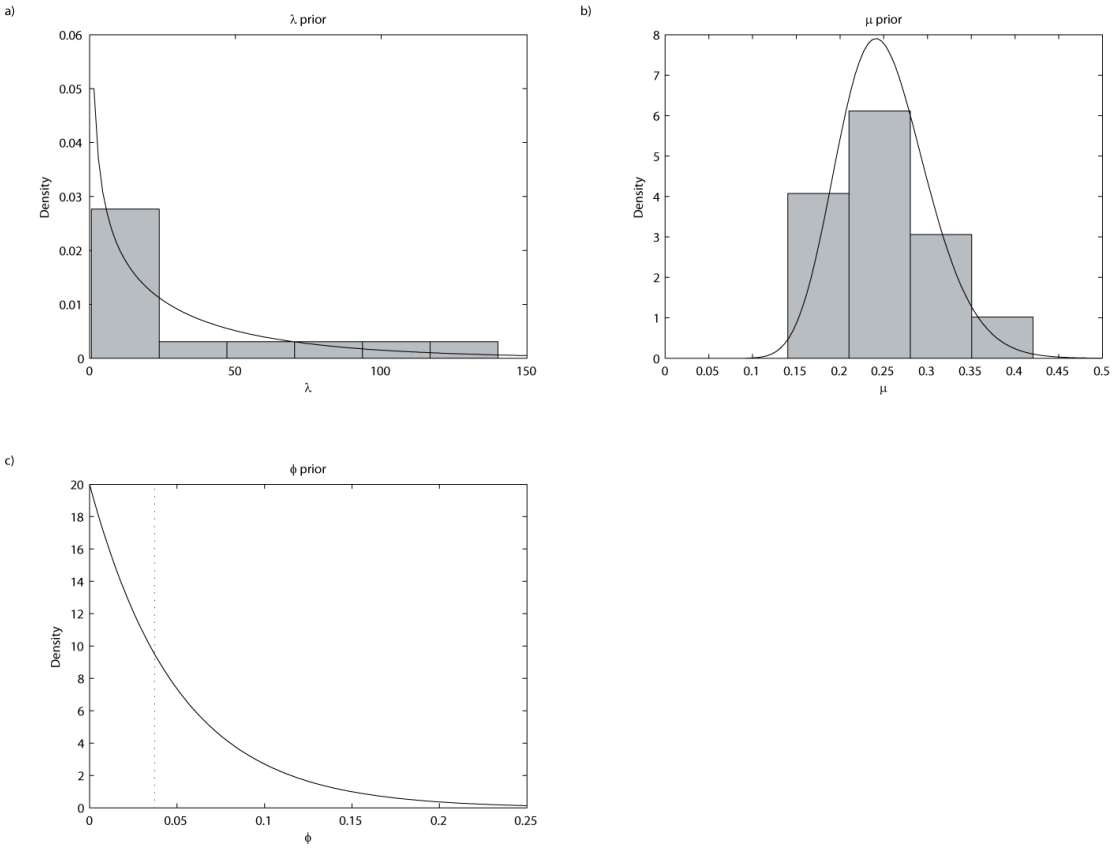


Figure 14. Prior Distributions. a) Prior distribution (black) and empirical distribution estimated from sperm typing data (grey) for the hotspot mass parameter, λ. b) Prior distribution and empirical distribution estimated from sperm data for the hotspot scale parameter, μ. c) Prior distribution (solid line) for the mean background rate parameter, ϕ. Also shown is the maximum background rate estimated in sperm data from the MHC and MS32 regions (dotted line).

Prior Parameters	Description	Default Value
ϕ	Mean background rate.	0.05 / kb
α_1, β_1	Parameters for hotspot scale prior.	23.984, 0.010488
α_2, β_2	Parameters for hotspot heat prior.	0. 61248, 52.470
m	Maximum distance allowed for hotspot map contribution.	5kb
T	Expected distance between hotspots.	50kb
σ_F	Standard deviation of Normal distribution used to generate random deviates.	1.0kb

Table 2 - Default parameters of the new method. All of these parameters may be altered by the user if required.

Properties of Mixing and Convergence

Having defined the rjMCMC scheme, I now informally consider the convergence and mixing properties of the chain. It was not expected that mixing would be an issue with this scheme, as the composite-likelihood is so weak (indeed – the original motivation for correcting the composite-likelihood was that it would improve the mixing properties). This weakness of the likelihood leads to inflated move acceptance rates (30-90%), which would generally be considered problematic for true likelihood MCMC schemes, as it would be indicative of the chain not full exploring the posterior distribution. However, it would appear that this is not the case

here and the chain appears to both mix and converge well. To demonstrate this, I simulated a dataset containing three moderately sized hotspots towards the centre of the region. Figure 15 shows two chains with different starting conditions run for a total 25,000 iterations using this simulated dataset. The first chain was started from a very low recombination rate of $\rho = 0.001$ / kb, whereas the second chain was started with a very high rate of $\rho = 10$ / kb. It can be seen that the chains rapidly converge towards a common distribution and subsequently mix well. This small simulation suggests that the chains have converged after approximately 5,000 to 10,000 iterations.

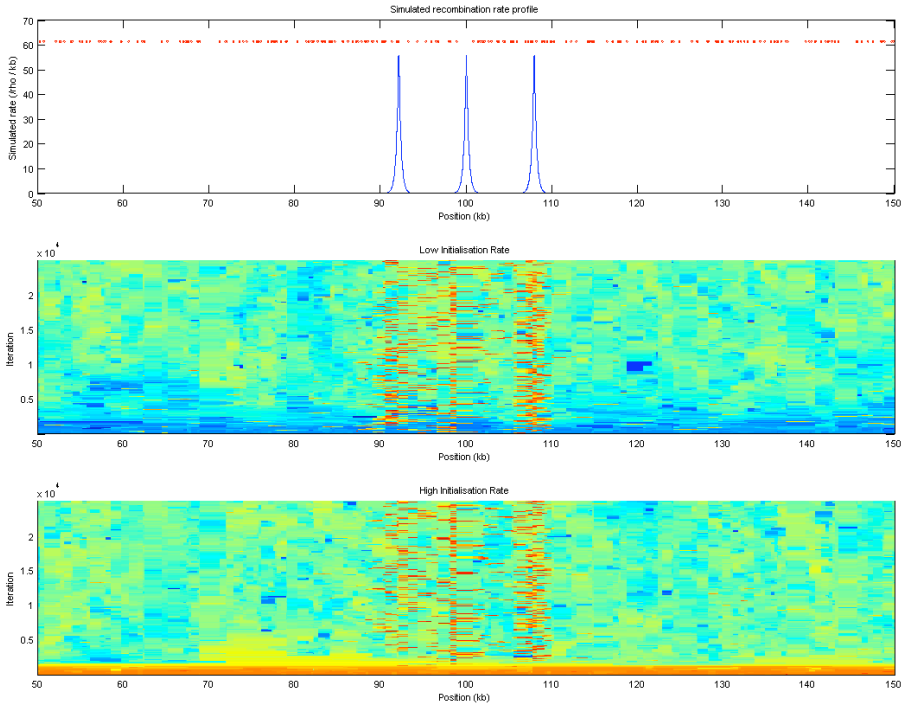


Figure 15. Convergence and mixing of the Markov chain. Top plot indicates the true simulated recombination rate (blue), with SNP positions shown as red marks. The central and lower panels show individual samples of the chain as heat plots over 25,000 iterations, with earlier samples being towards the bottom of the plot. Red indicates a high recombination rate estimate, and blue indicates a low recombination rate estimate. The central panel shows a chain starting at a very low recombination rate, and the lower panel shows a chain starting at a very high recombination rate.

As a further demonstration of convergence, Figure 16 shows two chains running on the same dataset (again started from low and high recombination rates respectively) that have been allowed to burn-in for 50,000 iterations before samples were taken. Each plot shows 500 samples of the cumulative recombination map taken from the following 50,000 iterations. The chains appear to have converged, and are mixing around a common value. While these two examples suggest the chain

converges relatively quickly, I generally discard at least 100,000 iterations when using the method to analyse real datasets.

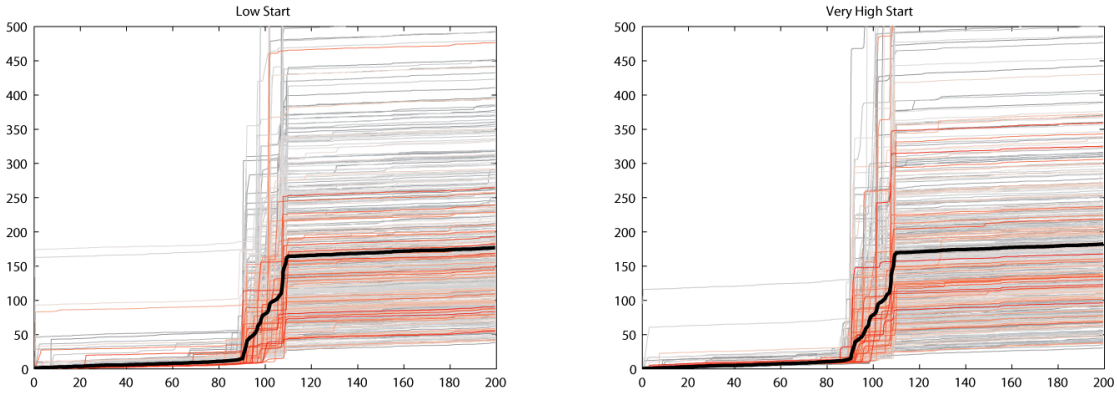


Figure 16. Mixing of the chain after convergence. Shown here are samples of the cumulative genetic map from two chains started with very different starting points. The left hand plot shows samples taken from a chain started with a very low recombination rate, whereas the chain in the right hand plot was started with an extremely high recombination rate. Both chains were allowed to burn in for 50,000 iterations. Samples were then taken every 100 iterations for a further 50,000 iterations. Earlier samples are shown in grey with later samples shown in shades of red. The average of the samples taken from each chain is shown as a black line on each plot.

Discussion

In this chapter, I have described the composite likelihood in further detail. I have also introduced a modified version of the composite likelihood, which corrects the overly peaked nature of the original composite likelihood (but leaves the maximum estimate unchanged). I then described the detail of a new method by which recombination rates may be estimated using the modified composite likelihood. Unlike previous methods, the new method includes a model of recombination

hotspots, the parameters of which were estimated from hotspots characterised in sperm studies. The method uses a rjMCMC scheme to explore the multidimensional space that describes the underlying recombination rate variation. I demonstrated that, when using the parameters estimated from sperm typing data, the chain both mixes and converges well.

The new method has been implemented in the program *rhomap*, which has subsequently been included in the *LDhat* package (AUTON and MCVEAN 2007). It is expected that the inclusion of this hotspot model will allow the method to provide more accurate estimates of the recombination rate than those obtained from the original *LDhat* implementation. In the next chapter, I assess the performance of *rhomap* as both a rate estimation tool and a hotspot detection tool.

Chapter 3 The Performance of *rhomap*

In this chapter, I assess the performance of the new method, which has been implemented in the program known as *rhomap*. The method is assessed using simulation studies that aim to test the method both as a recombination rate estimation tool, and as a hotspot detection tool. After the simulation studies, the method is demonstrated using human datasets from the MHC and MS32 regions.

The Performance of rhomap *on Simulated Data*

To investigate the performance of *rhomap*, I carried out a number of separate simulation studies designed to measure the performance in different situations. In the first study, I simulated data with a constant recombination rate. The second study consisted of simulated data generated with a randomly chosen and variable recombination rate. In the third study, I simulated data using three fixed recombination maps with individual hotspots of differing magnitude. The remaining studies were designed to test the performance of *rhomap* using data with a low SNP density or unknown haplotypes. For these studies, I generated data using a fixed recombination map with three hotspots clustered at the centre of the region.

Each study simulated datasets containing 100 haplotypes of 200kb in length. Data was simulated using the *fin* program (AUTON and MCVEAN 2007; MCVEAN *et al.* 2002), which is based on Hudson algorithm (HUDSON 1983a). The simulation population-scaled mutation rate per base was chosen to be 3.86×10^{-4}, which gives 400 expected segregating sites (see equation (1.30)).

In all simulation studies, *rhomap* was run for a total of 1,100,000 iterations which included a burn-in of 100,000 iterations. Samples of the chain were taken every 100 iterations after the burn-in. For comparison, the datasets were also analysed with the *LDhat* (version 2.0) method using 10 million iterations and a block penalty of 5 (as used by JEFFREYS *et al.* 2005; MYERS *et al.* 2005). With these parameters, the computational cost of the two methods was approximately equal. Using a 1.8Ghz personal computer, both methods took about 17 minutes to analyse a typical dataset from the simulation studies (although this is using pre-calculated lookup tables, which can take many hours of computer time to generate). However, it should be noted that *rhomap* scales less favourably with the number of SNPs than *LDhat*.

Simulation Study A

I simulated 100 datasets using a fixed recombination rate of ρ = 0.5 / kb, giving a total recombination map length for the region of ρ = 100. In this study, *rhomap* tended to slightly overestimate the total map length, with *LDhat* estimates being less biased (Figure 17a, b). The average estimates of ρ / kb were 0.58 for *LDhat* and 0.65 for *rhomap* (Figure 18a). The upwards bias in the *rhomap* estimates is caused by the weakness of the flattened composite likelihood relative to the prior allowing the method to insert spurious hotspots. However, as will be seen in the next simulation study, the upwards bias primarily affects estimates of background rate variation and is less of a problem in the presence of hotspots.

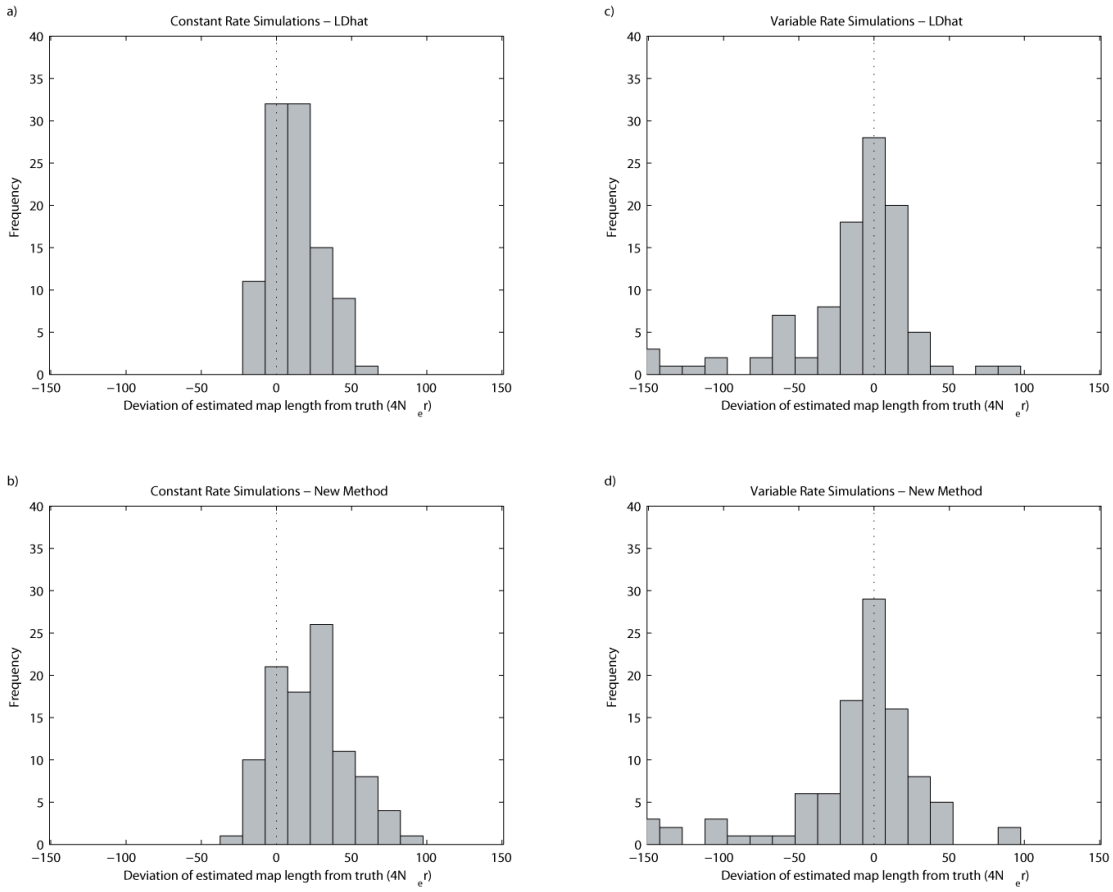


Figure 17. Deviation of the estimated total ρ from the simulated value. Rate estimates from the constant rate simulations (Simulation Study A) using *LDhat* and *rhomap* are shown in (a) and (b) respectively. Rate estimates from the variable rate simulations (Simulation Study B) using *LDhat* and *rhomap* are shown in (c) and (d) respectively.

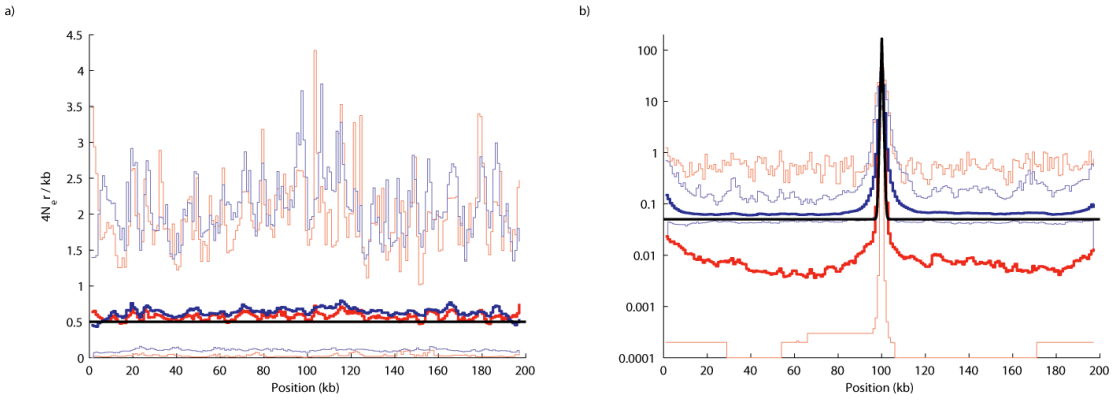


Figure 18. Average recombination rate estimates from 100 simulated datasets. (a) Results from Simulation Study A with a constant recombination rate. (b) Results from Simulation Study C with an active central hotspot. Rate estimates from *LDhat* and *rhomap* are shown as thick red and blue lines respectively. The simulated recombination profile is shown in black. The 2.5th and the 97.5th percentile of the estimated rates are shown in faded colours. Note that for clarity, the constant rate simulation estimates are shown on a linear scale, whereas the hotspot simulation estimates are shown on a logarithmic scale.

Despite the upwards bias of the mean estimates, the coverage of the *rhomap* estimate is better than that of *LDhat*. Considering the rate estimates between SNPs, the 2.5 to 97.5 percentiles of *LDhat* estimate contain the true rate 52% of the time, whereas those of *rhomap* contain the true value 83% of the time.

Simulation Study B

This study was designed to assess the performance of *rhomap* using randomly simulated variable recombination maps that included hotspots. I simulated 100 datasets using recombination maps generated from our prior on recombination rate variation. The expected number of hotspots per simulation was four, each with an expected width of 1.5kb (where the width is defined as the region in which 95% of the

hotspot mass is contained) and an expected contribution to ρ of 32.1. Thus the expected total recombination distance for the region of $\rho = 138.6$.

To assess the performance of two methods on the variable rate datasets, I again considered the total ρ estimate over the region (Figure 17c, d). By this measure, both methods showed similar performance, with *LDhat* estimating an average ρ over the region of 115.9, and *rhomap* estimating an average of 121.85. However, the two methods behaved differently as the simulated rate varied (Figure 19). *LDhat* produced relatively unbiased estimates at both high and low rates, but exhibited more bias at intermediate rates. Furthermore, the *LDhat* estimates showed a high amount of variance, which was due to the high level of noise in the estimates at the fine scale. Conversely, *rhomap* tended to overestimate at low rates (in a similar manner to the constant rate simulation study), with performance improving at intermediate to high rates. The *rhomap* estimates also showed significantly less variance than those from *LDhat*. The corresponding reduction in noise relative to the *LDhat* estimates improves the correlation coefficient between the estimated rate and the simulated rate over each SNP interval (Figure 20). Compared to *LDhat*, the *rhomap* estimates were almost universally better correlated with the simulated rate.

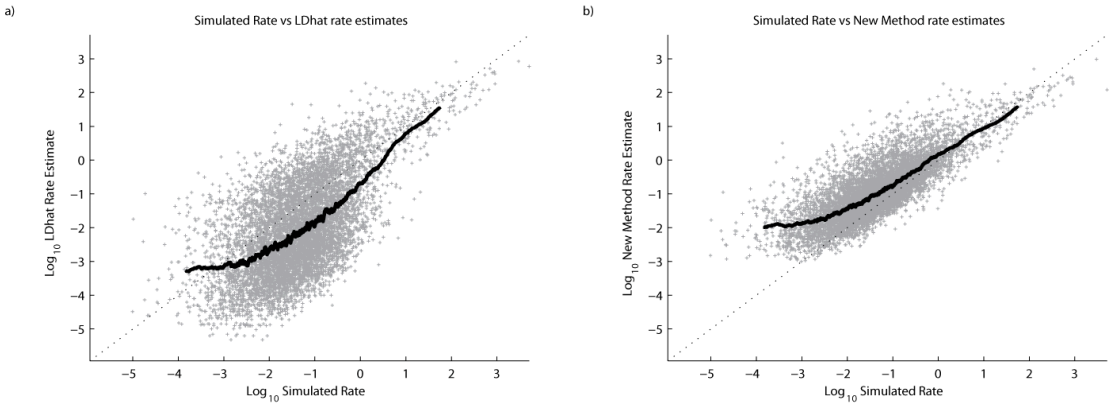


Figure 19. Results from Simulation Study B. Scatter plot of simulated rate versus estimated rate for *LDhat* (a) and *rhomap* (b). Each point represents an estimate of recombination rate between two adjacent SNPs. A 250 point moving average is also shown.

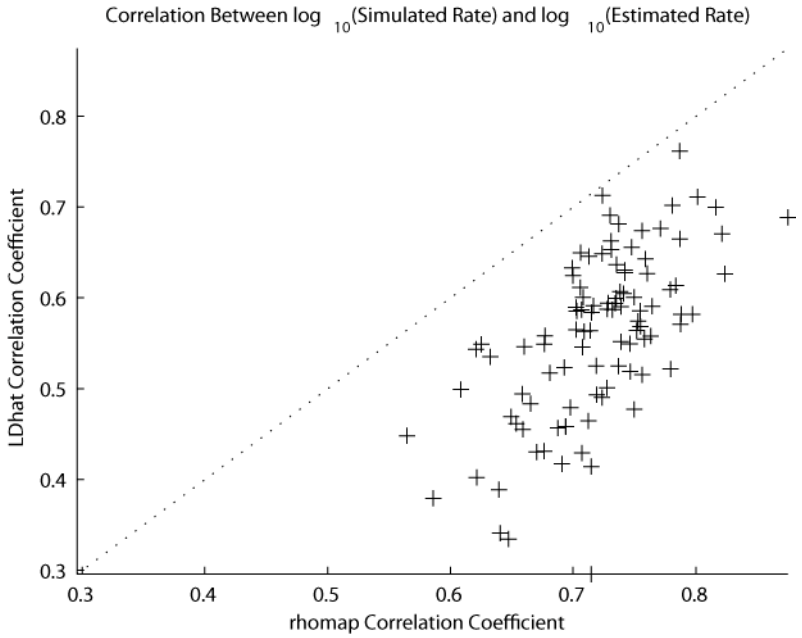


Figure 20. Results from Simulation Study B. Correlation coefficient between the log10 estimated rate and the $\log_{10}$ simulated rate for 100 datasets, as measured over SNP intervals. The correlation coefficients obtained using rate estimates from *LDhat* are shown on the vertical axis, and the coefficients obtained using *rhomap* are shown on the horizontal axis.

As with the constant rate simulations, the sample distribution of the *rhomap* estimate was more likely to contain the true rate than that of *LDhat*. Again considering the rate estimates between SNPs, the 2.5 to 97.5 percentiles of *LDhat*

estimate contained the true rate 32% of the time, whereas those of *rhomap* contained the true value 93% of the time.

A useful benefit of *rhomap* is that it may be used as a hotspot detection tool. The inclusion of a hotspot model in the rate estimation procedure allows the locations of hotspots to be sampled from the Markov Chain. To determine the location of hotspots, I calculated the average number of hotspots per sample between each adjacent pair of SNPs and divided by the inter-SNP distance (measured in kilobases). I call this statistic the posterior hotspot density (although technically it should be called the pseudo-posterior hotspot density to emphasize the use of the composite likelihood). I then identified hotspots as regions where the local maxima in this statistic were greater than some arbitrary threshold (Figure 21). In this simulation study, I called a 'detected' hotspot as correct if the estimated peak in posterior hotspot density is within 1.5kb of a true hotspot peak. Otherwise, the hotspot was considered to be a false positive. This study suggested that a suitable threshold was 0.25, which gives a detection power of approximately 50% and a false discovery rate of 4%. I have therefore used this threshold in subsequent analyses. As I will show later, using *rhomap* as a hotspot detection tool is not as powerful as other methods (FEARNHEAD 2006; LI *et al.* 2006; LI and STEPHENS 2003; MCVEAN *et al.* 2004). However, it is capable of identifying candidate hotspots with a low false discovery rate as part of the rate estimation procedure, and therefore is useful for identifying potential hotspot locations.

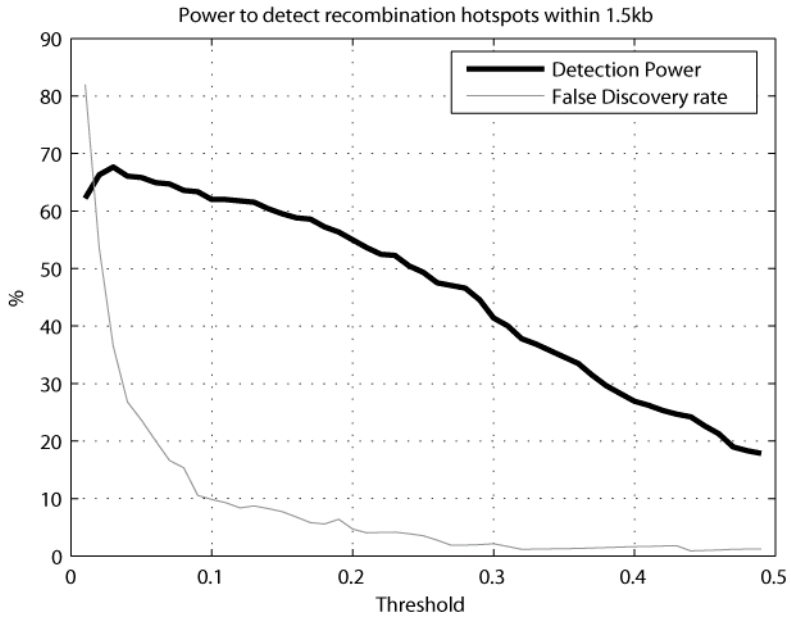


Figure 21. Results from Simulation Study B. Using *rhomap* as a hotspot detection tool in the variable rate simulation study. This plot shows the power of *rhomap* to detect recombination hotspots (thick line) and the false discovery rate (thin line). Hotspots were called if the average number of hotspots per sample per kb at a local maxima was above the threshold shown on the horizontal axis. The hotspot was considered to be correctly detected if it was within 1.5kb of the location of a simulated hotspot. Otherwise, the hotspot was considered a false positive.

Simulation Study C

In this study, I generated 100 datasets for each of three recombination maps. Each recombination map contained a single recombination hotspot of differing magnitude at the centre of the region. The three hotspots used contributed ρ = 80, 22.13 and 6.07 to the recombination map, and I subsequently refer to these hotspots as the strong, moderate and weak hotspots respectively. The hotspots all had a width of 1.5kb and fixed background rate of ρ = 0.05 / kb.

The results of the strong hotspot simulation study are shown in Figure 18b. As in Simulation Study B, it is clear that *rhomap* tended to overestimate the background

rate (and again this is most likely due to the weakness of the composite likelihood relative to the prior allowing the insertion of spurious hotspots). However, *rhomap* produced a significantly smoother signal than *LDhat* as can be seen from the range of the estimates. Both methods are consistently able to resolve the hotspots in all three cases. Using *rhomap* as a hotspot detection method, and applying the 0.25 threshold from the previous simulation study, I found that 61% of the hotspots were detected in the weak hotspot study, 69% were detected in the moderate hotspot study, and 91% were detected in the strong hotspot study. Out of the 300 simulations, I counted a total of 11 false positive detections (4, 6 and 1 false detections in the weak, moderate, and strong hotspot simulations respectively), which equates to a false positive rate of approximately one per 5Mb. However, neither method performed well at estimating the peak rate of the hotspot (Table 3). This is perhaps not surprising, as once a hotspot becomes sufficiently large, the data either side of the hotspot becomes (essentially) independent; hence distinguishing between hotspots of different sizes will be difficult. Despite this inaccuracy, both methods generally estimated a total map length within a factor of two of the truth.

	Region Map Length (ρ)	Hotspot Contribution to Map (ρ)	Hotspot Peak Rate (ρ / kb)	Estimated Peak Rate (**mean**, lower quartile, upper quartile)		Estimated Map Length (**mean**, lower quartile, upper quartile)	
				LDhat	*rhomap*	*LDhat*	*rhomap*
Strong Hotspot	100	80	179.7	**27.0**, 17.8, 34.2	**30.4**, 19.5, 38.7	**76.8**, 65.6, 87.9	**77.0**, 66.6, 85.6
Moderate Hotspot	42.13	22.13	64.2	**15.5**, 7.8, 22.0	**16.6**, 7.3, 24.2	**49.4**, 35.0, 58.7	**55.4**, 42.4, 66.5
Weak Hotspot	26.07	6.07	32.1	**9.3**, 5.2, 12.6	**8.7**, 3.5, 11.0	**30.3**, 22.9, 34.7	**36.5**, 28.2, 40.6

Table 3 – Summary of method performance in Simulation Study C.

Simulation Study D

This simulation study was designed to assess the resolution of *rhomap*, and investigate how this affected by SNP density. Specifically, I was interested in the ability of *rhomap* to distinguish closely spaced hotspots. I generated 100 datasets with three hotspots contained within a 20kb region at the centre of the simulated map. The contribution to the map from each hotspot was $\rho = 26.7$ and the background rate was $\rho = 0.05$ / kb, giving a total map length of approximately $\rho = 100$. As before, the hotspots had a width of 1.5kb.

To assess how SNP density affects the performance of *rhomap*, I artificially thinned the data using two separate methods. In the first method, I removed a proportion of SNPs in a uniformly random manner to give an average SNP density of 1 SNP / kb. In the second method, I randomly removed SNPs in a frequency dependent manner. The probability that a SNP was not deleted from the data was

$1 - e^{-Bf}$, where f is the minor allele frequency, and B is a constant. The constant B was chosen as 20.ln(2), so that the SNPs with a minor allele frequency of 5% had a 50% chance of being retained in the dataset. In practice, this scheme reduced the average SNP density to approximately 1.2 SNPs per kilobase, which is similar to that obtained by the International HapMap Project (2007).

I first consider the cumulative map estimates of *rhomap*, compared to those from *LDhat* (Figure 22). For all three datasets, the average estimated map length from *LDhat* is more accurate than that from *rhomap*. However, as with the previous simulation studies, the variance in the *rhomap* estimate is smaller than the *LDhat* estimates.

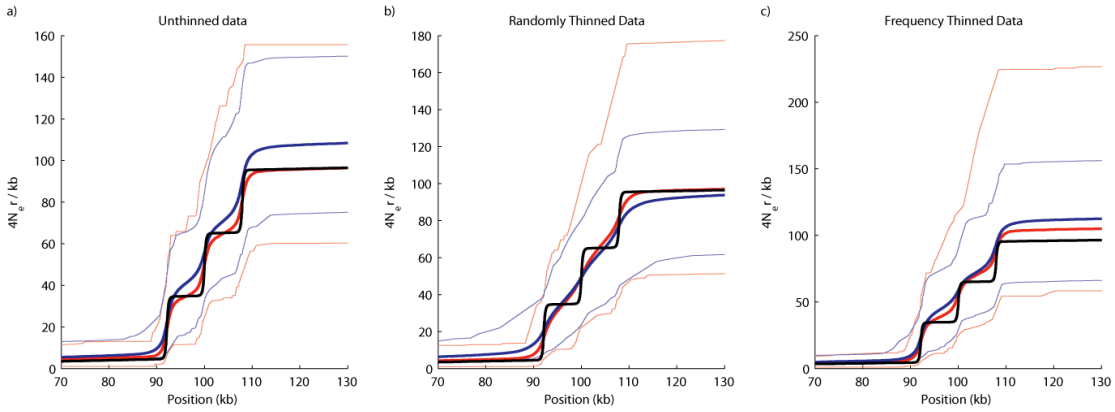


Figure 22. Simulation Study D. *Rhomap* cumulative map estimates around a cluster of hotspots, averaged over 100 replications. a) Unthinned dataset, b) Dataset randomly thinned to average SNP density of 1 / SNP per kb, c) Dataset randomly thinned depending on Minor Allele Frequency. The estimates from *rhomap* are shown in blue, whereas those from *LDhat* are shown in red.

I assessed the performance of *rhomap* via its ability to detect the three hotspots (Figure 23). In the unthinned datasets, *rhomap* was generally able to detect the hotspots on the edges of the cluster, but had lower power to detect the hotspot in

the centre of the cluster. Applying the 0.25 threshold from Simulation Study B would give a detection power of 60%, 34% and 59% for the left-hand, central and right-hand hotspots respectively, and 5 false positives. At least one hotspot was correctly detected in the region 89% of the time.

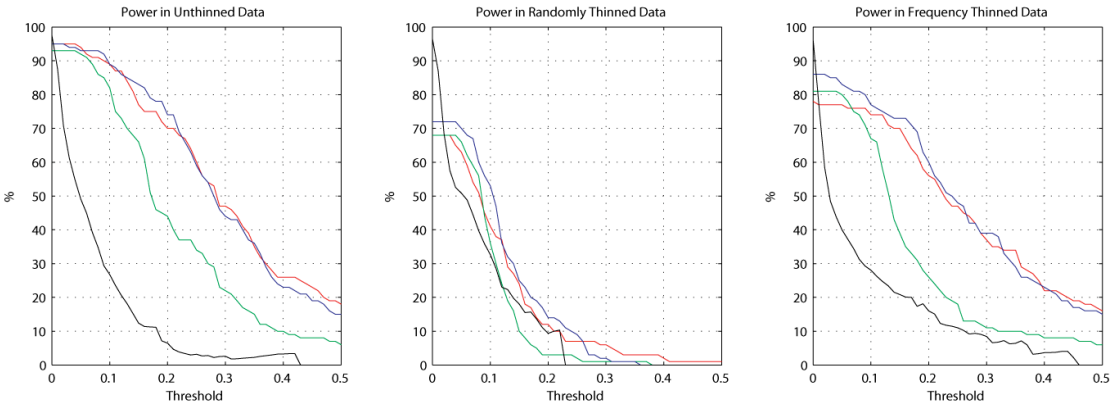


Figure 23. Simulation Study D. Power of *rhomap* to detect hotspots within a cluster. a) Unthinned dataset, b) Dataset randomly thinned to average SNP density of 1 / SNP per kb, c) Dataset randomly thinned depending on Minor Allele Frequency. Hotspots were deemed to be correctly detected if the estimated peak was within 1.5kb of the true peak. Estimated hotspots not within 1.5kb of a true hotspot were deemed to be false positives. The power to detect the left, central and right hotspots is shown in red, green and blue respectively. The False Discovery Rate is shown in black.

By comparison, *rhomap* performed poorly when the uniformly randomly thinned dataset was used. The power to detect the hotspots was heavily reduced. Using the 0.25 threshold gave a detection power below 10% for all hotspots, with at least one hotspot being correctly detected within the region 17% of the time. However, no false positives were recorded.

For the dataset thinned depending on the minor allele frequency, the performance of *rhomap* was the intermediate of the previous two cases. The power to detect the two exterior hotspots was approximately 48%, and the power to detect the

central hotspot was 18%. However, there were 14 false positives. These seemed to be largely a result of the lower SNP density not allowing *rhomap* to resolve the hotspot peak within 1.5kb of the truth. If we account for the low SNP density by calling correct detection if a hotspot is called within 2.5kb a true hotspot peak (as apposed to the 1.5kb used in the previous studies), then the power to detect the three hotspots was 53%, 18% and 51% respectively, with 5 false positives. At least one hotspot was correctly detected within the region 79% of the time.

A Comparison of rhomap *to Other Hotspot Detection Methods*

I wanted to compare the performance of *rhomap* as a hotspot detection tool to that of other methods specifically designed to detect hotspots. I originally planned to compare *rhomap* to four other methods, namely *LDhot* (McVean et al. 2004), *Hotspotter* (Li and Stephens 2003), *HotspotFisher* (Li *et al.* 2006), and *sequenceLDhot* (Fearnhead 2006).

It was not possible to use the first method, *LDhot*, as the program has not been released either as source-code or in a precompiled form. This is unfortunate, as published results would suggest that *LDhot* is one of the more powerful methods.

The second method, *Hotspotter*, is publicly available both as precompiled binaries and original source code. However, I found the running time of *Hotspotter* to be surprisingly long. While the program could analyse small datasets within a few minutes, the time taken did not scale well with the size of the dataset. For datasets of the size of those in the simulation studies outlined earlier, *Hotspotter* had not

completed the first iteration of the analysis after approximately 8 hours on a 2.0 GHz computer, and the analysis was therefore abandoned.

The third method was that of *HotspotFisher*. This method is only available in a precompiled form. However, I have found the program to be quite troublesome, with the program often crashing. The problem seemed to be that *HotspotFisher* is extremely sensitive to deviations in the required input file format. As the source code is unavailable, it was very difficult to diagnose the exact problem. Nevertheless, it was possible to run *HotspotFisher* (albeit with some difficulty).

I was able to run the final method, *sequenceLDhot*, without problems and found it to complete in reasonable time. This method appears to be quick enough to be used on a genome-wide scale. Indeed, as *LDhot* has not been made publicly available, it would seem to be sensible to consider using *sequenceLDhot* in future large-scale studies of recombination.

I have therefore compared the performance of *HotspotFisher* and *sequenceLDhot* to *rhomap*. To do this, I have used the datasets from the hotspot cluster study (Simulation Study D).

For *HotspotFisher*, I used the default parameters as recommended by the authors. I ran *sequenceLDhot* using the same parameters as the original paper (FEARNHEAD 2006). Specifically, the number of runs was 15000, with a minimum of 300 iterations per hotspot. Three ρ driving values were used. A constant background of $\rho = 0.05$ / kb and per-site $\theta = 0.001$ were assumed (both of which approximately match the simulated values).

I assessed the power of the two hotspot detection methods both by considering their ability to detect the individual hotspots in the cluster, and by their ability to detect any hotspot within the region. Both methods output the location of a hotspot

within some window that covers a number of SNPs. For comparison with *rhomap*, I took the location of the hotspot to be the centre of the window as the hotspot location. However, using this scheme, the two hotspot detection methods were generally unable to resolve hotspots within 1.5kb of the truth and would therefore give an exaggerated false positive rate. I was therefore more generous to all of the methods and called a hotspot as correctly detected if it was called within 2.5kb of a true hotspot location in all three datasets.

The results of this study are shown in Table 4. It is clear that *sequenceLDhot* and *HotspotFisher* outperformed *rhomap* in terms of detection power. Of the three methods, *sequenceLDhot* appears to be the most powerful. However, *sequenceLDhot* does appear to have a higher false positive rate, which appears to be due to the inability of *sequenceLDhot* to resolve the location of the hotspots beyond a seven SNP window. The majority of the *sequenceLDhot* false positives were within 5kb of a true hotspot. There were a total of eight false positives which were more than 5kb away from a true hotspot, which compares favourably with *rhomap's* 10.

The disappointing performance of *rhomap* relative to the other two methods suggests that *rhomap* is not a particularly good hotspot detection method due to low power. While the posterior hotspot density statistic provides a useful summary, it is not suitable for the testing of the existence of a hotspot. The issue seems to be that it is difficult to determine a sensible threshold of the posterior hotspot density for a given dataset. For example, in the case of the randomly thinned data, *rhomap* performed vary badly in terms of power. However, if one visually inspects the posterior hotspot density, one can see clear peaks in the density at the hotspot locations despite the density rarely reaching the 0.25 threshold (Figure 24). In such a situation a lower threshold of say 0.15 may be more appropriate. Indeed, applying this

threshold increases the power of *rhomap* to detect at least one hotspot in the randomly thinned data to 53% (compared to 17% originally) with three false positives.

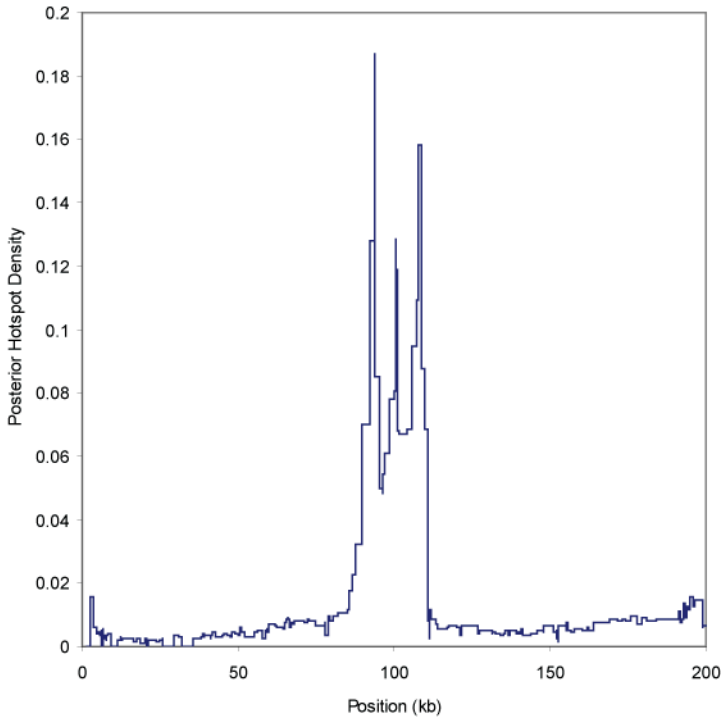


Figure 24. Example of the Posterior Hotspot Density from a Randomly Thinned Dataset. While clear peaks are visible, none achieve the required 0.25 threshold.

Dataset	*rhomap*					*sequenceLDhot*				
	Power to detect hotspot (%)			Detected > 0 hotspots (%)	False +ves	Power to detect hotspot (%)			Detected > 0 hotspots (%)	False +ves
	Left	Mid	Right			Left	Mid	Right		
Unthinned	61	35	59	89	5	69	73	72	97	21
Randomly Thinned	7	2	9	17	0	59	58	59	95	32
Frequency Thinned	53	18	51	79	5	66	61	72	97	19

Dataset	*HotspotFisher*				
	Power to detect hotspot (%)			Detected > 0 hotspots (%)	False +ves
	Left	Mid	Right		
Unthinned	62	69	65	97	4
Randomly Thinned	35	35	46	81	9
Frequency Thinned	68	70	66	97	5

Table 4 – Power to detect hotspots within a cluster of three hotspots for *rhomap*, s*equenceLDhot* and *HotspotFisher*.

Ideally, we would like to have a formal process by which this threshold could be determined. If we consider the distribution of posterior hotspot densities from a given dataset, then we may expect the majority of SNP intervals to show small densities with some fluctuation due to noise. Hotspot regions would be expected to show higher densities, again with some noise. In general, we would expect the number of 'hotspot' regions to be considerably less than the number of background regions. We are essentially left with an outlier identification problem, with the hotspot

regions providing the outliers. Unfortunately, the distribution of the posterior hotspot density is unknown.

A sensible scheme may be to simulate data using SNP densities similar to the dataset under analysis but with a constant recombination rate. By repeating the rate estimation on various simulated datasets, it would be possible to obtain a distribution of posterior hotspot densities for a given SNP density, and hence inform what would be a suitable threshold. Such scheme would be computationally intensive and in any case, it seems unlikely that such a scheme would improve on the power of methods that are already available. I therefore recommend that *rhomap* be used primarily as a rate estimation tool, and not as a hotspot detection tool. In the case where a user wishes to obtain both rate estimates and hotspot locations, it would be sensible to use *rhomap* to obtain the rate estimates, but use a separate method such as *sequenceLDhot* to detect the hotspots. Given that *sequenceLDhot* provides the locations of hotspots within a broad window, it would then be possible to use the *rhomap* rate to further localise the hotspot locations.

The Effect of Phasing Genotype Datasets

I have so far only considered simulated haplotype datasets. However, in many real-life situations, only genotype information is available. Both *LDhat* and *rhomap* can make use of genotype data by averaging over all possible haplotypes in each pairwise likelihood calculation. An alternative method would be to use statistical methods to infer the underlying haplotypes; a process that is known as phasing. This is commonly achieved using the publicly available programs such as *PHASE* (STEPHENS and SCHEET 2005) and *fastPHASE* (SCHEET and STEPHENS 2006).

However, such methods can rarely infer the haplotypes with total accuracy, and also make underlying assumptions about the structure of recombination. A recent study found that many estimators of the recombination rate are robust to the use of phasing (SMITH and FEARNHEAD 2005), but only considered the simple case of constant recombination rate. It is therefore interesting to ask what affect the phasing of the data has on the variable rate estimates obtained by *rhomap* and *LDhat*.

To address this question, I returned to the unthinned hotspot cluster datasets from Simulation Study D. For each of the 100 datasets I used a random ordering of the 100 haplotypes to obtain 50 genotypes. Using these genotype datasets, I used both *PHASE* (version 2.1.1) and *fastPHASE* (version 1.2.3) to infer haplotypes. For *PHASE*, I used the default parameters, but restarted the algorithm 10 times (using the '-x' flag). The haplotypes in the 'best' reconstruction (as defined by *PHASE*) were stored for subsequent use. For *fastPHASE*, the default parameters where used (as suggested in the documentation), with 25 random starts to the algorithm. Two files are outputted by *fastPHASE*, one that minimises the individual error and one that minimises the switch error (as defined in STEPHENS and DONNELLY 2003). I stored the haplotypes that minimised the switch error, as I have found that these gave better results in the recombination rate estimation. It is worth noting that the computational cost of the two methods differs significantly with *PHASE* taking much longer than *fastPHASE* (as implied by the name). Whereas *fastPHASE* would take minutes to complete the analysis of a dataset, *PHASE* would take hours.

Using the three new datasets (genotypes, *PHASE* haplotypes, *fastPHASE* haplotypes), I obtained recombination rate estimates from *rhomap* and *LDhat* using the same parameters as before. The resulting map and rate estimates can be seen in

Figure 25. For comparison, the estimates obtained from the original haplotype datasets are also shown.

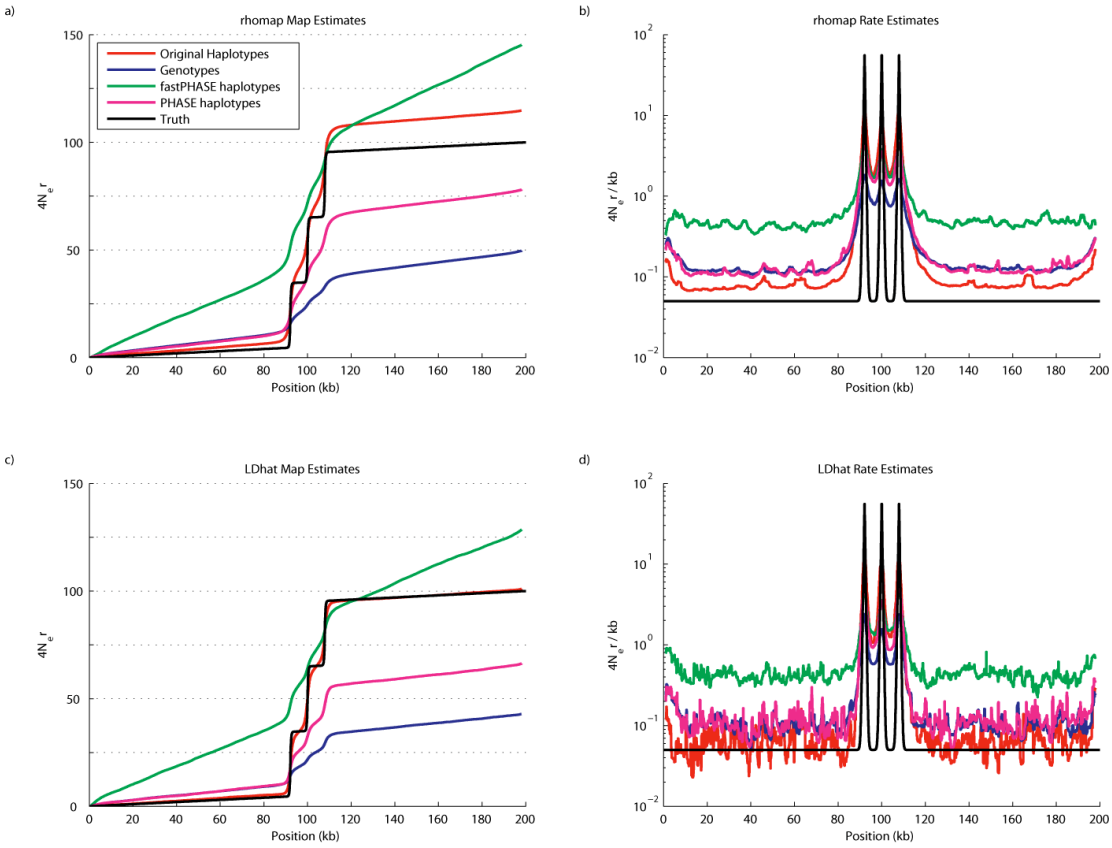


Figure 25. The effect of using genotype data on recombination rate estimates. These plots show the estimated cumulative recombination maps from *rhomap* (a) and *LDhat* (c), averaged over 100 hotspot cluster datasets. Also shown are mean recombination rate estimates for the two methods (b and d). The simulated recombination maps are shown in black, and the rate estimates from the original haplotypes are shown in red. The estimates from genotype data are shown in blue. The estimates obtained using the *PHASE* and *fastPHASE* inferred haplotypes are shown in green and magenta respectively.

Both *LDhat* and *rhomap* give largely similar results for these datasets. Starting with the genotype datasets, we can see that both *rhomap* and *LDhat* severely underestimate the magnitude of the hotspots and consequently underestimates the total map length over the region. However, the background rate estimates are only

slightly higher than those obtained from the original haplotypes. It therefore appears that using raw genotype data severely decreases the ability of both methods to obtain information regarding the magnitude of hotspots. On average, the map lengths estimated from genotype data were approximately 40-50% of the truth, with the *rhomap* estimates being marginally more accurate than those from *LDhat*.

Conversely, estimates obtained from the haplotypes inferred by *fastPHASE* show a large and consistent overestimate of background recombination rates. The underestimation of peak hotspot rates remains, although the estimate is closer to the truth than the genotype case. On average, the total map length of the region was overestimated by approximately 25-50%, the majority of which was contributed from the overestimation of the background rate. In this instance, *LDhat* produces map estimates closer to the truth.

Finally, the estimates obtained from the *PHASE* haplotypes are the intermediate of the two previous cases. The background rate estimates are close to those obtained by from the genotype datasets, although are marginally higher than those obtained from the original haplotypes. Again, the hotspot magnitudes are underestimated, although as with *fastPHASE* the estimates are closer to the truth than the genotype case. Of the three methods, the *PHASE* haplotypes provide the best estimates of the total map length over the region with an underestimate of approximately 25-35%. On average, the map length estimates from *rhomap* are closer to the simulated value than those obtained from *LDhat* in this case.

These results suggest that phasing of genotype data can introduce a number of biases in the resulting recombination rate estimates. However, the accuracy of the phasing (and hence the recombination rate estimates) can be improved by the use of external information. For example, in the genome-wide datasets that I analyse in

Chapter 4 and 5, the phase could be inferred with very high accuracy as data was available from family trios (i.e. from both parents and the offspring; MARCHINI *et al.* 2006).

Using rhomap *with Human Datasets*

We now compare rate estimates obtained by *rhomap* to those obtained by sperm typing. These two datasets were outlined in an earlier section – one from the MHC region (JEFFREYS *et al.* 2001) and the other from the MS32 region (JEFFREYS *et al.* 2005) - both of which consist of genotype data. Both datasets are of comparable size, with the MHC dataset containing 50 genotype sequences with 274 segregating sites in 216kb and the MS32 dataset containing 80 genotype sequences with 199 segregating sites in 209kb.

For both datasets, we ran *rhomap* for a total of 1,100,000 iterations including a burn-in of 100,000 iterations and taking a sample every 100 iterations. For each dataset, the estimation procedure took approximately 20 minutes on a 2.0Ghz computer.

The MHC Dataset

The MHC dataset contains six clearly defined hotspots visible in sperm. I obtained rate estimates that are well correlated with those obtained via sperm typing (Figure 26), although *rhomap* tended to estimate the hotspots to be larger than they appeared in the sperm estimates. Using *rhomap* as a hotspot detection tool (as

explained in the earlier simulation study), we see that *rhomap* was able to identify five of the six hotspots clearly visible in sperm. While there is also some evidence for the undetected hotspot (DMB1), the posterior hotspot density statistic does not reach the required threshold. Furthermore, the leftmost hotspot (DNA1) is apparently displaced by approximately 2kb relative to the location in sperm.

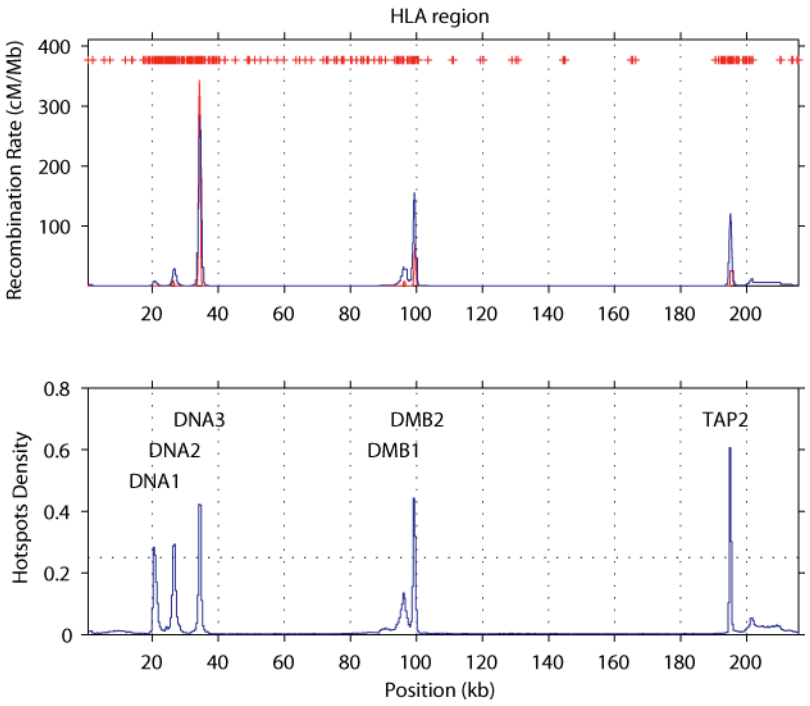


Figure 26. Output of *rhomap* for the MHC region. The top plots shows the recombination rate estimates, with the estimated rate in blue, and (sex-averaged) sperm typing rate in red. SNP locations are shown as red marks. Estimates from *rhomap* were converted to cM/Mb by assuming Ne = 10,000. The bottom plot shows the average number of hotspots per sample per kb for the same regions.

The MS32 Dataset

This dataset contains at least six hotspots found by sperm typing. There is also evidence of two apparent 'double' hotspots with the edges of the hotspots overlapping, yet still retaining two identifiable peaks (these hotspots are known as NID2a / b and MSTM1a / b; JEFFREYS *et al.* 2005). As with the MHC region, *rhomap* again obtains rate estimates that are well correlated with those obtained via sperm typing (Figure 27), although the two methods differ in the estimates of the magnitude of the hotspots.

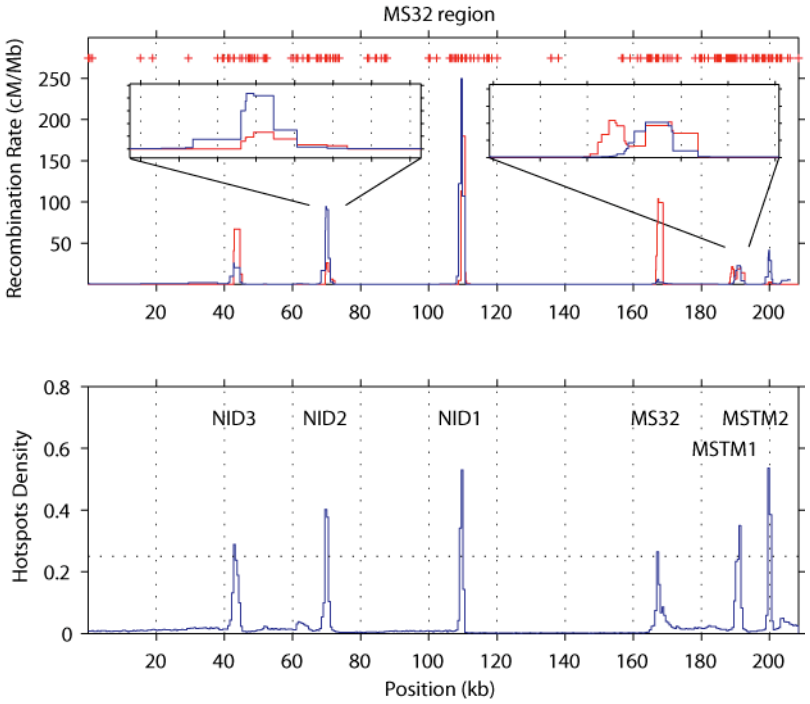


Figure 27. Output of *rhomap* for the MS32 region. The top plots shows the recombination rate estimates, with the estimated rate in blue, and (sex-averaged) sperm typing rate in red. SNP locations are shown as red marks. Estimates from *rhomap* were converted to cM/Mb by assuming Ne = 10,000. The bottom plot shows the average number of hotspots per sample per kb for the same regions.

We can identify six hotspots that reached the 0.25 hotspot posterior density threshold. Notably, *rhomap* was able to detect the fourth hotspot from the left (known as MS32), despite the relative increase in recombination rate being very small. This hotspot has previously been reported as being extremely weak in coalescent analysis despite being strong in sperm analysis, which possibly indicates that the hotspot has only recently become active and therefore has not yet left a signature in LD (JEFFREYS *et al.* 2005). For the 'double' hotspots, *rhomap* was able to detect hotspots in the vicinity, but was unable to resolve the separation between the hotspots. Interestingly, it appears that the MSTM1b hotspot is well resolved by *rhomap*, but the MSTM1a hotspot is not detected. This is, however, likely to be due to lack of resolution to resolve two hotspots which are so close. Other methods have also had difficulty in distinguishing these hotspots (JEFFREYS *et al.* 2005; LI *et al.* 2006).

Discussion

In this chapter, I have assessed the performance of the new method as implemented in the program, *rhomap*. I have found that *rhomap* offers improved rate estimates relative to *LDhat*. I have also assessed *rhomap* as a hotspot detection tool. In this case, I found that *rhomap* is not as successful as existing methods. I therefore recommend that rate estimates be obtained using *rhomap*, but hotspot locations are determined using a separate program. I also demonstrated the performance of *rhomap* using human datasets from the MHC and MS32 regions. While these datasets have an unusually high SNP density (especially around hotspots), the excellent estimates obtained from *rhomap* demonstrate the abilities of the method in 'ideal' circumstances.

Like *LDhat*, the new method can be applied on large-scale datasets of up to approximately 200 chromosomes. It is perhaps worth noting that even larger datasets containing thousands of chromosomes have recently become available (e.g. WELLCOME TRUST CASE CONTROL CONSORTIUM 2007). However, it appears unlikely that using more than approximately 200 chromosomes will significantly alter recombination rate estimates. The reason is that the rate at which adding chromosomes increases the number of detectable recombination events is extremely low (of the order of the log of the log sample size; MYERS 2002).

In the next chapter, I apply *rhomap* on a genome-wide scale using the data from Phase II of the International HapMap Project (2007). I validate the rate estimates by comparison to those obtained the deCODE pedigree study (KONG *et al.* 2002). I then use the resulting rate estimates to study the distribution of recombination in the human genome and identify relationships between recombination and various genome annotations.

Chapter 4 The Distribution of Recombination in the Human Genome

In this chapter, I use *rhomap* to obtain recombination rate estimates on a genome wide scale. These estimates provide a number of novel insights into patterns of recombination in the human genome, notably in relation to genic regions and DNA repeats. I demonstrate that the certain elements show significantly elevated recombination rates if they include a hotspot-associated motif. As a number of such motifs have been identified, I attempt to unify these motifs into a single degenerate motif using a genetic algorithm. Finally, I investigate whether the activity of the hotspot-associated motifs can be explained by epigenetic factors.

Introduction to the HapMap Project

It is thought that about 9-10 million SNPs with a minor allele frequency of at least 5% exist in the non-repetitive sequence of the human genome (THE INTERNATIONAL HAPMAP CONSORTIUM 2007). As discussed in the introduction of this thesis, linkage ensures that alleles of nearby SNPs will tend to be inherited together. This leads to a non-random association, or linkage disequilibrium (LD), between SNPs at separate locations. In the human genome, the level of LD is such that many regions of the genome can be described using a few commonly occurring haplotypes. A chromosomal region may contain many SNPs, but the correlations between SNPs ensure that most of the information about specific alleles in a region can be captured

using a few "tag" SNPs (REICH *et al.* 2001). However, the level of LD in the genome is highly heterogeneous, and hence while some regions can be typed with only a few tag SNPs, other regions require many more tags to adequately describe the level of variation.

The International HapMap Project, or HapMap for short, was launched in 2002 with the goal of characterising these common patterns of human DNA sequence variation and thereby identify a suitable set of tag SNPs (THE INTERNATIONAL HAPMAP CONSORTIUM 2005; THE INTERNATIONAL HAPMAP CONSORTIUM 2007). The resulting publicly available resource was to be of use in the design and analysis of genetic association studies. The project was divided between scientists in Japan, the U.K., Canada, China, Nigeria, and the U.S.

The original project genotyped approximately 1.2 million polymorphic SNPs with the aim of genotyping at least one common SNP every 5kb (with minor allele frequency, MAF, > 5%) across the non-repetitive portions of the autosomes and the X chromosome. The genotyping was performed in 269 individuals from four geographically diverse populations. The 269 DNA samples were divided as follows: 30 parent-child trios from the Yoruba people in Ibadan, Nigeria (abbreviated as YRI), 44 unrelated Japanese individuals in Tokyo (abbreviated as JPT), 45 unrelated Han Chinese individuals from Beijing (abbreviated as CHB), and 30 parent-child trios from Utah with ancestry in Northern and Western Europe (abbreviated as CEU). For the purposes of analysis, the CHB and JPT populations were combined, and I refer to this combined population by the abbreviation CHB+JPT.

Despite the primary aim of the project being to advance medical genetics, the HapMap also provided a valuable resource for the analysis of evolutionary processes such as selection (VOIGHT *et al.* 2006), and molecular processes such as

recombination (MYERS *et al.* 2007; MYERS *et al.* 2006). It was possible to estimate recombination rates on a genome-wide scale using the *LDhat* method, and detect evidence for recombination hotspots using *LDhot* with 21,617 hotspots identified in Phase I.

In Phase II of the HapMap (THE INTERNATIONAL HAPMAP CONSORTIUM 2007), an additional 2.1 million SNPs were added to the original map, and the requirement of MAF > 5% was relaxed. Once quality control has been taken in account, the Phase II HapMap contained a total of 3.1 million SNPs that were polymorphic in at least one panel giving an average SNP density of approximately one SNP per kilobase across the autosomes and X chromosome. It is therefore thought that approximately 25-35% of all common SNPs in the covered regions of the genome are captured by the Phase II HapMap. However, it should be noted that there is a large amount of heterogeneity in the local HapMap SNP density (Figure 28A), including interesting patterns around local genomic features such as genes (Figure 28B).

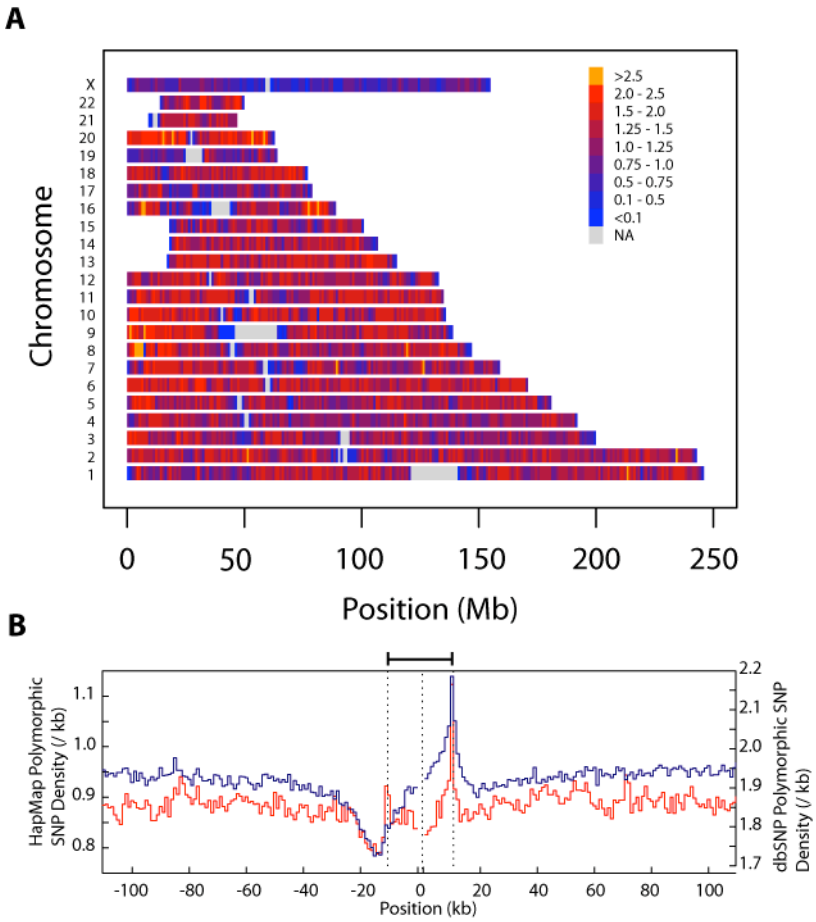


Figure 28. Local Heterogeneity in Phase II HapMap SNP density. A) SNP density across chromosomes. Colours indicate the number of polymorphic SNPs per kilobase, with brighter colours indicating higher density (Adapted from The International HapMap Consortium, 2007). B) SNP density around genes. Densities estimated from both Phase II HapMap (red, left axis) and dbSNP release 125 (blue, right axis) are shown. Densities were calculated separately around the transcription start point (left dotted line) and the transcription end point (right dotted line). The two plots were joined at the median midpoint of the transcription tract.

The increased density of SNPs in the Phase II HapMap provides greatly increased resolution to the estimation of fine-scale recombination rate estimates. In this chapter, I describe how I have used this increased resolution to provide new insights into the distribution of recombination in the human genome.

Genome-Wide Recombination Rate Estimation

The HapMap Phase II data was provided in genotype format for the three analysis panels. The data was phased using the program *PHASE (STEPHENS and SCHEET 2005)*, making use of trio information where available. For recombination rate estimation, the data was divided into regions of 2,000 SNPs with an overlap of 200 SNPs between regions. Recombination maps for the three populations were estimated for the autosomal chromosomes using the *rhomap* program. A total of 4,100,000 iterations were performed (the first 100,000 being discarded as burn-in iterations) with a sample taken every 400 iterations. This computation took approximately 72 hours on a computer cluster consisting of 150 nodes 2.0 GHz processors. The resulting estimates of pseudo-posterior distribution consisted of approximately 32 gigabytes of data.

Estimated recombination maps were obtained by taking the mean of the pseudo-posterior distribution at each SNP interval. The regions between the SNPs bounding the centromeric regions were set to have a recombination rate of zero. Contiguous rate estimates were obtained splicing the regions back together by removing 100 SNPs from both ends of the regions.

Comparison of HapMap with deCODE

To convert the population genetic estimates to centimorgans via the $\rho = 4N_e r$ relation, effective population sizes were estimated by matching the total map length to that estimated by the pedigree method used by deCODE (KONG *et al.* 2002).

Estimated effective population sizes are given in Table 5, with the estimates obtained from *LDhat* shown for comparison. A combined map was obtained by simply averaging over the three populations and interpolating where necessary.

Population	***LDhat* N_e**	***rhomap* N_e**
CEU	10,930	12,062
YRI	17,428	18,905
CHB+JPT	13,612	14,524

Table 5 – Phase II estimated autosomal effective population sizes for the three HapMap panels. *LDhat* estimates are shown for comparison.

As validation of the population genetic approach, scatter plots of the *rhomap* estimated rates were compared to rates obtained from the deCODE pedigree study (KONG *et al.* 2002). Rate estimates were binned at five megabase intervals, which is towards the lower limit of the resolution of the pedigree-based estimates. We see that the genome-wide correlation between HapMap and deCODE is extremely good with a Pearson correlation coefficient of 0.95 (Figure 29).

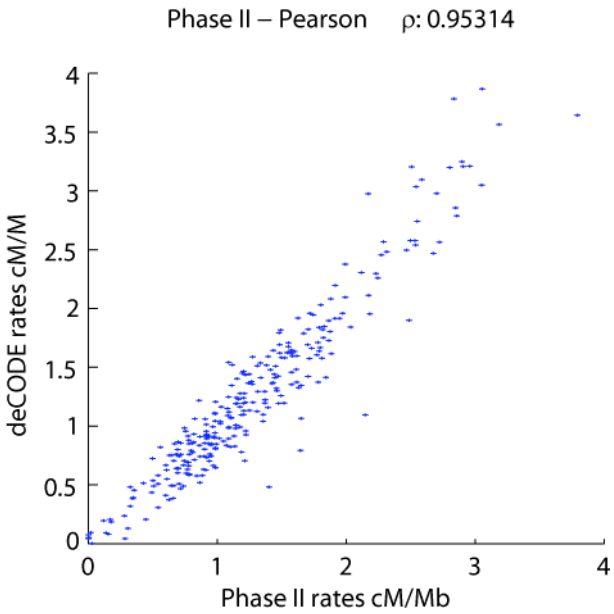


Figure 29. Comparison of deCODE map to that estimated by *rhomap* for all autosomes. Rate estimates from both methods were binned at the 5Mb scale.

The MHC and MS32 Regions Revisited

To give the reader a better understanding of the quality of the rate estimates from Phase II HapMap, I reconsider the MHC and MS32 regions that were described and analysed in the previous chapters of this thesis. In Phase II of the HapMap, the MHC and MS32 regions contained 444 and 228 SNPs respectively, averaged over the three populations. As with all the HapMap datasets, I used *rhomap* to obtain rate estimates from each population separately, using 4,100,000 iterations with a burn-in of 100,000 iterations and a sample taken every 400 iterations. The resulting rate estimates for the MHC and MS32 regions can be seen in Figure 30 and Figure 31 respectively. Also shown in the bottom panels of these figures are the rate estimates from *LDhat* and hotspots detected by *LDhot*.

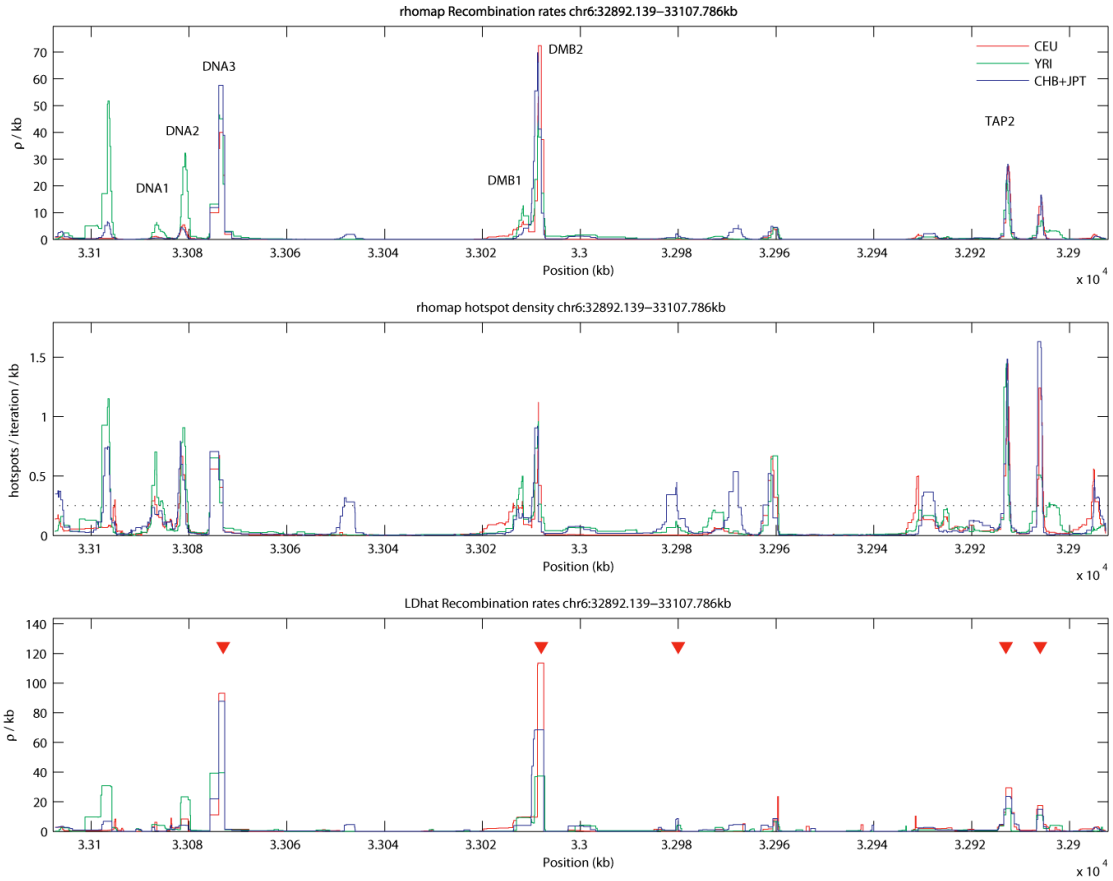


Figure 30. Rate estimates in the MHC region using HapMap Phase II data. The three populations, CEU, YRI and CHB+JPT, are shown in red, green and blue respectively. Top: *rhomap* rate estimates for the three populations. Central: *rhomap* hotspot densities, with the 0.25 posterior hotspot density threshold indicated by a dotted line. Bottom: *LDhat* rate estimates (shown for comparison), with *LDhot* hotspot positions indicated by red triangles.

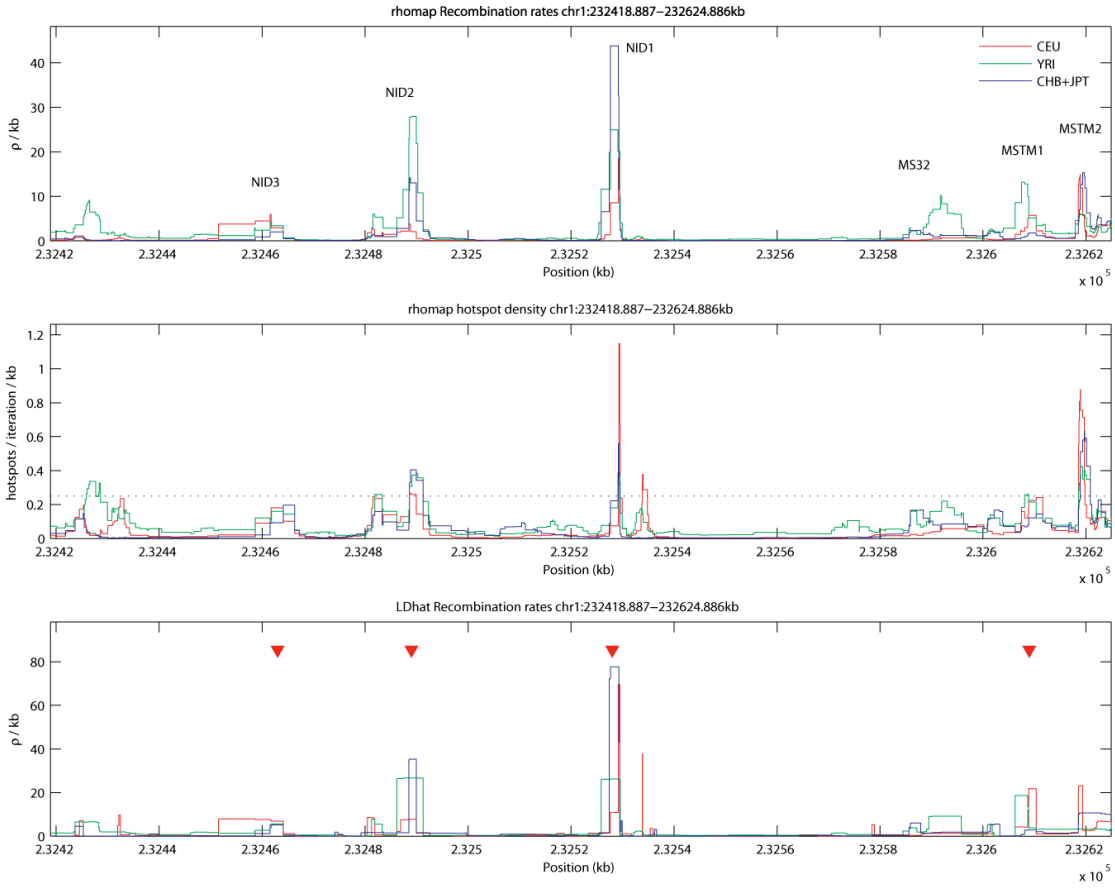


Figure 31. Rate estimates in the MS32 region using HapMap Phase II data. The three populations, CEU, YRI and CHB+JPT, are shown in red, green and blue respectively. Top: *rhomap* rate estimates for the populations. Central: *rhomap* hotspot densities, with the 0.25 posterior hotspot density threshold indicated by a dotted line. Bottom: *LDhat* rate estimates (shown for comparison), with *LDhot* hotspot positions indicated by red triangles.

In the MHC region, we see that all of the hotspots detected in the sperm analysis are also detected by *rhomap*, including the leftmost hotspot cluster, which is clearly resolved. However, *rhomap* also detects a number of previously undescribed hotspots, at least three of which are visible in all three populations. The two largest of these novel hotspots occur towards the edges of the analysed region, which may indicate why they were not visible in the original sperm dataset (JEFFREYS *et al.* 2001). The remaining novel hotspots are all either very small or do not appear in more than one population, which could suggest that they are spurious.

In the MS32 region, there are visible peaks in the estimated rates for all of the hotspots previously described. However, only three of these hotspots clearly achieve hotspot densities over 0.25 in more than one population. There is a notable feature around the MS32 hotspot itself. While the hotspot density statistic in this region does not cross the 0.25 threshold in any population, there is a large and broad region of elevated recombination rate in the YRI estimates, which at least superficially resembles a hotspot. If this is indeed the MS32 hotspot, then it would be contrary to the hypothesis that this is a newly emerged hotspot (JEFFREYS *et al.* 2005), as its existence would have to predate the divergence of the human populations. A similar feature can be seen to the left-hand side of the region.

90% of Recombination Occurs Within 30% of Sequence

A plot showing the proportion of recombination versus the proportion of sequence for all autosomes is shown in Figure 32. For each SNP interval, the contributions to the genetic map and to the physical map of each chromosome were calculated as proportions of the totals. SNP intervals were then reordered by the recombination rate. If recombination were evenly distributed throughout the chromosome then the lines would run along the diagonal. We therefore note that recombination is highly concentrated with approximately 90% of recombination occurring within 30% of sequence. This pattern is largely consistent between the chromosomes with the exception of chromosome 19, which exhibits a less concentrated pattern of recombination (although still highly non-uniform). This pattern has been noted before (MYERS *et al.* 2005), and may be related to

chromosome 19 having the highest gene density (LANDER *et al.* 2001) and proportion of open chromatin (GILBERT *et al.* 2004) of all the human chromosomes.

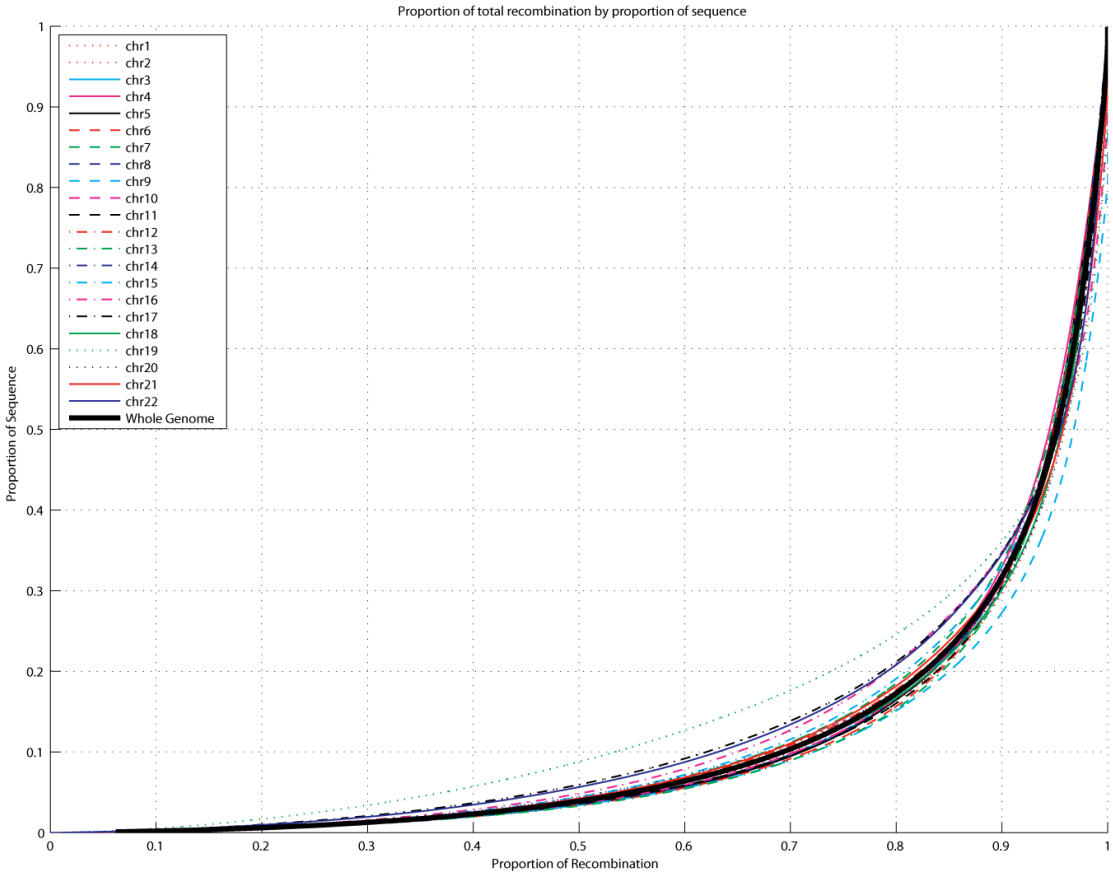


Figure 32. The distribution of recombination in the autosomal chromosomes. Shown here is the cumulative proportion of the genetic map verses the proportion of physical chromosome length. Each autosome is shown as a separate line, with details given in the legend. Also shown in this plot is an average for all autosomes, weighted by total genetic map length (thick black line).

The Distribution of Hotspots

To detect hotspots in HapMap using *rhomap*, I averaged the posterior hotspot density statistic between the three populations. I then called hotspots in regions where

this statistic is above the 0.25 threshold as usual. Using this method, *rhomap* detects at total of 28,995 hotspots. This compares well with the 34,142 hotspots detected using *LDhot* with the same data (MYERS *et al.* 2007). There is also good agreement between the locations of hotspots, with 64% of the hotspots called by *rhomap* being contained within the boundaries of the *LDhot* hotspots. Given the expected power of *rhomap* and *LDhot* (in the region of 50-60%; MCVEAN *et al.* 2004), the expected number of recombination hotspots in the human genome is in the region of 50,000 – 70,000. This in turn suggests an average density across the genome of approximately one hotspot per 50kb.

The *LDhot* hotspots show near-uniform density across chromosomes, whereas *rhomap* shows a more complicated pattern (Figure 33). Despite this, there is a large degree of correlation between the densities of called hotspots from the two methods (Figure 33d). However, as noted in the previous chapter, *rhomap* is not particularly well suited to hotspot detection. While the majority of my analysis in the following section focuses on patterns of rate variation, there are occasions when I will be referring to hotspots. In such sections, I will be using the *LDhot* hotspots and the reader therefore should assume that I am using this set of hotspots, unless otherwise stated.

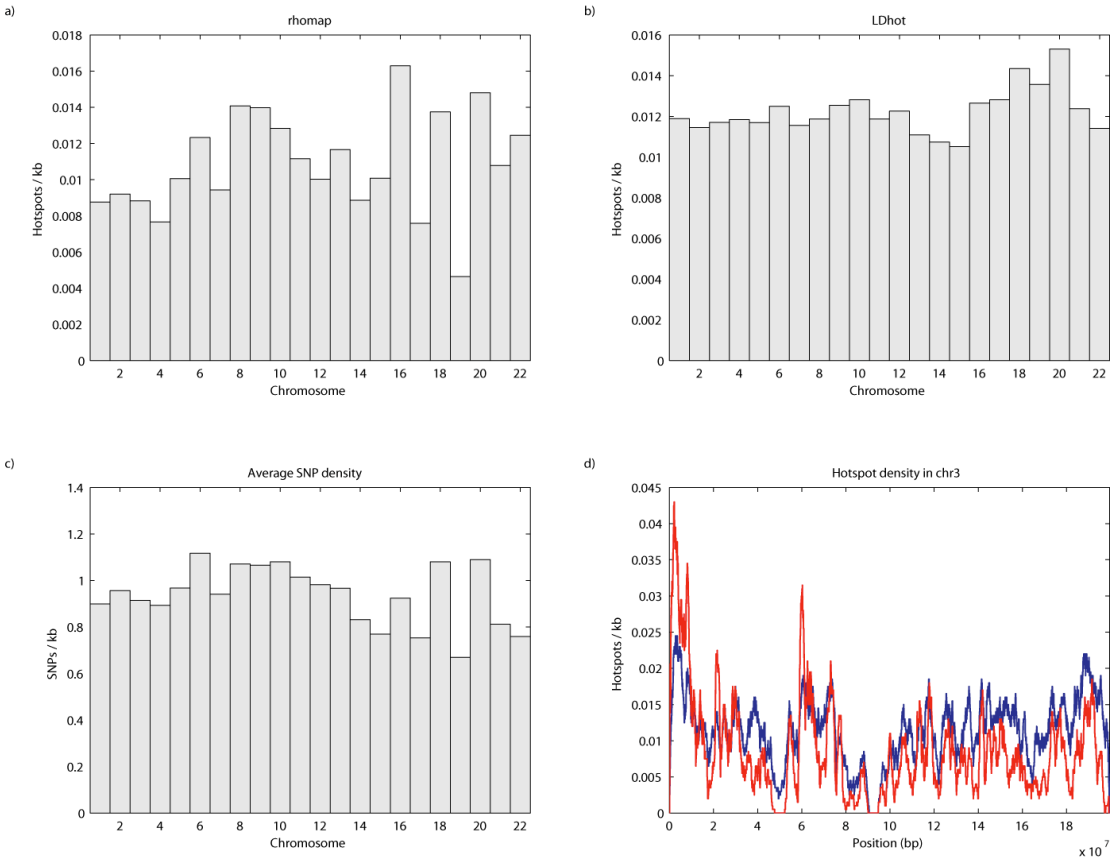


Figure 33. Hotspot density by chromosome. a) Density of hotspots detected using *rhomap*. b) Density of hotspots detected by *LDhot*. c) Average SNP density (averaged over the three populations). d) Hotspot density in chromosome 3. The lines represent 1Mb moving average of the density of *LDhot* hotspots (blue) and *rhomap* (red).

A Hotspot-related Motif

Earlier work using genome-wide population data has attempted to identify DNA sequence features that are related to recombination hotspots. I will be repeatedly referring to these sequence motifs later in this chapter, so provide a brief summary here. In 2005, a study investigated the relationship between recombination hotspots and DNA sequence (MYERS *et al.* 2005). This study identified a number of short

sequence motifs as being significantly over represented in hotspots (MYERS *et al.* 2005). For each hotspot in this study (as identified by *LDhot*), a corresponding coldspot (a region that shows no evidence of recombination) was also identified in the local vicinity. Each coldspot was chosen to match the corresponding hotspot for size, SNP density and GC content.

The two motifs most over-represented in hotspots were the 7-mer CCTCCCT, and the 9-mer CCCCACCCC. Further evidence for the activity of these motifs comes directly from the sperm typing studies of the MHC and MS32 regions (JEFFREYS *et al.* 2001; JEFFREYS *et al.* 2005). The CCTCCCT motif is found in the central region of the DNA2 hotspot in the MHC dataset. Furthermore, a SNP at the third base of the motif allowed a change from a T to a C. It was found that individuals with the C allele reduced the activity of the hotspot by a factor of three relative to individuals with the T allele (JEFFREYS and NEUMANN 2002). The second most over-represented motif, the 9-mer CCCCACCCC, was found in the NID1 hotspot of the MS32 region. Again, a SNP within the motif (which altered the first base from a C to a T) disrupted the activity of the hotspot (JEFFREYS and NEUMANN 2002).

These results clearly demonstrate that there is some sequence dependency to the location and activity of hotspots. However, the 7-mer motif can account for no more than 11% of hotspots (MYERS *et al.* 2005), and the 9-mer is less predictive still. This suggested that there were yet to be discovered factors involved.

With the increased number and improved localisation of hotspots in Phase II of the HapMap, the search for hotspot-related sequence motifs was repeated (MYERS *et al.* 2007). Again, a large number of motifs were identified. Notably however, the majority of motifs identified by this study showed strong homology to the 13-mer CCTCCCTNNCCAC.

I will be referring to these motifs in later sections of this thesis. I will generally specify which motif I am referring to, but may also refer to the motifs by length. Therefore, CCTCCCT may be referred to as the 7-mer, CCCCACCCC as the 9-mer and CCTCCCTNNCCAC as the 13-mer.

It is perhaps worth noting at this point that while these motifs are highly over-represented in hotspots, they are poor predictors of hotspots (MYERS *et al.* 2006). For example, there are 6,655 occurrences of CCTCCCTNNCCAC in the genome, 1359 of which occur in *LDhot* hotspots, and 344 of which occur in the matched coldspots. Therefore, given a randomly selected occurrence of the motif, there is a 20% chance of the motif being in a hotspot, and a 5% chance of being in a coldspot. However, only 4% of detected hotspots contain this motif. In an attempt to explain a greater proportion of hotspots, a set of over-represented motifs within 2 substitutions of CCTCCCTNNCCAC has been considered, and it has been hypothesised that up to 40% of hotspots may contain an 'active' motif (MYERS *et al.* 2007). However, the number of motifs within this set is large and the subset of 'active' motifs is currently unknown (an issue I will attempt to address in a later section). Given that a large amount of the human DNA sequence may be within two substitutions of the 13-mer, the predictive power of these motifs is small.

Therefore, while it appears that sequence motifs have an important influence in determining the location of hotspots, they appear to be neither necessary nor sufficient. Why a motif should be 'hot' in one instance and 'cold' in another remains unknown. I will be returning to this issue at a later stage of this chapter.

Patterns of Recombination Associated with Genomic Features

Having obtained recombination rate estimates for the whole genome (excluding the sex chromosomes), the relationship between recombination and various genome features becomes of interest. Previous work has identified a number of associations between recombination and various features of the human genome. These include specific DNA motifs influencing hotspot position (MYERS *et al.* 2005; MYERS *et al.* 2006), genes (MYERS *et al.* 2005; SMITH *et al.* 2005), base composition (JENSEN-SEAMAN *et al.* 2004; KONG *et al.* 2002) and DNA hypersensitivity (LI *et al.* 2006). The resolution of the Phase II HapMap rate estimates allows these influences to be studied in more detail.

Recombination is Suppressed Within Genes

Figure 34 shows the average recombination rate around 14,979 non-overlapping autosomal genes taken from the NCBI RefSeq annotation (PRUITT *et al.* 2005). The left hand sections of the plots were constructed by aligning recombination rate estimates at the transcription start points. Likewise, the right hand sections of these plots were constructed by aligning at the transcription end points. The two sections were joined at the median gene centre (as measured from the transcription start and end points).

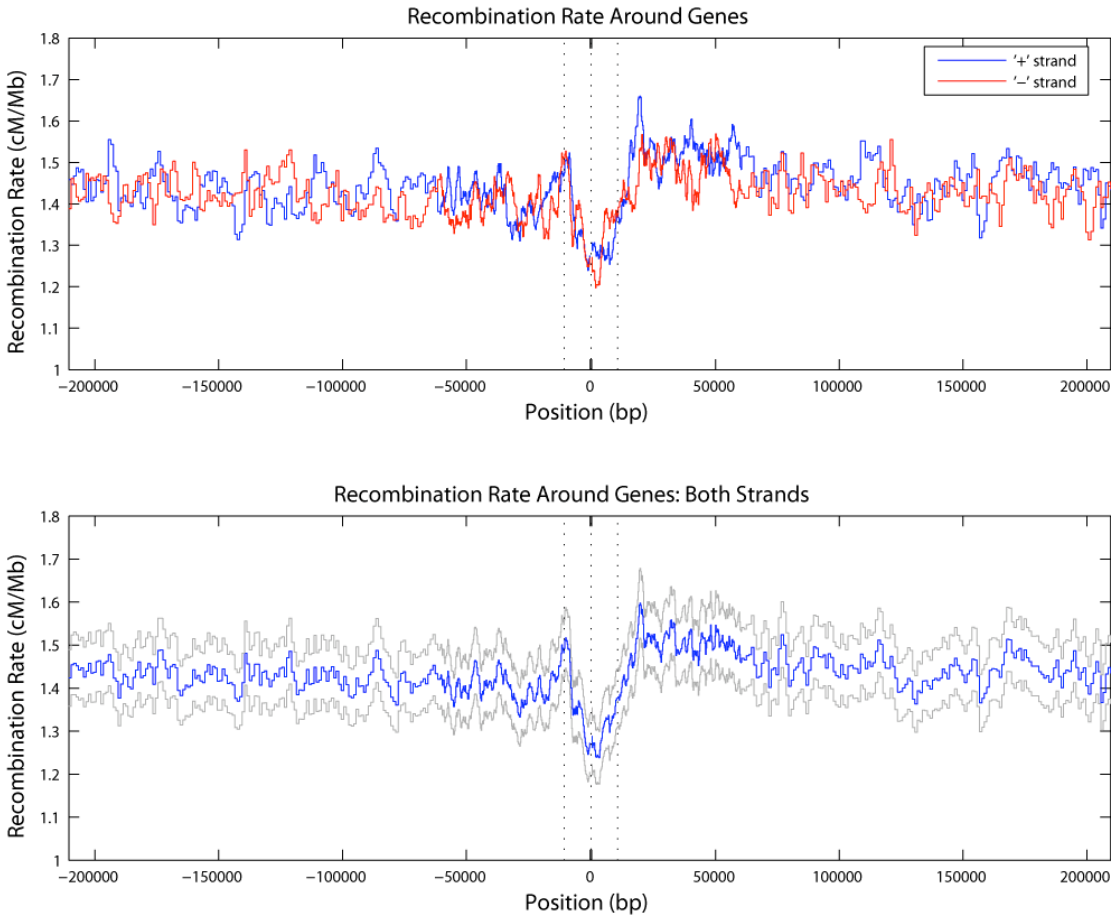


Figure 34. The recombination rate around genes. The top plot shows the average recombination rate averaged over genes with plus strand transcription (blue) and minus strand transcription (blue). The bottom plot shows the average over both strands (blue), and ± 2 standard errors (grey). The vertical dotted lines indicate the transcription start point, the median transcription point (at which the left and right hand plots have been joined), and the transcription end point respectively.

The plots clearly show that the recombination rate is suppressed within the transcription region of genes. There is a small peak in recombination rate around the transcription start point before a sharp drop. Towards the transcription end point, we observe a recovery in recombination rate, albeit a more gradual change than that observed around the transcription start point. There is also a slight asymmetry in recombination rate outside the transcribed region, with the region downstream of transcription being slightly elevated relative to the region before transcription.

One possible explanation of the observed dip in recombination is the presence of selection in this region. It is known that selection can affect patterns of LD and hence estimates of recombination rates (McVean 2007; Reed and Tishkoff 2006; Spencer 2006). Previous results have shown that, while *LDhat* is largely robust to the presence of selection (McVean 2007), a positive selected sweep leading to complete fixation of the selected allele can cause a small decrease in estimated rates (Spencer 2006). Given that both *LDhat* and *rhomap* are based on the composite-likelihood, I would expect that *rhomap* to show similar performance to *LDhat* in these situations. However, as Figure 34 shows the average over thousands of genes, each of which will be under differing selection pressures, it seems unlikely that selection can account for the observed pattern.

We are able to further analyse this pattern by considering other annotations that are known to affect recombination. Figure 35 again shows the distribution of recombination around genes, but also shows the GC content and the density of a hotspot-related motif (specifically CCTCCCTNNCCAC). We see that the peak in recombination rate at the transcription start site corresponds to peaks in the other two annotations. The peak in GC largely reflects the presence of CpG islands in the promoter regions (Cross and Bird 1995). Likewise, the peak in motif density is a reflection of the high GC content; the motif mostly consisting of Cytosine bases. However, it is interesting to note that the peak in recombination is much smaller than the peak in the other two annotations, as this indicates that high local GC content and the presence of the motif are not sufficient to cause a hotspot alone. Indeed, it may be that the observed peak is not associated with the motif at all, and may instead be caused by the accessible chromatin in these regions allowing increased rates cross-over. Studies in yeast have suggested an association between promoter regions and

recombination hotspots (PETES 2001; WU and LICHTEN 1994). The results presented here suggest there is a significant, albeit weak, relationship between promoters and recombination in humans. It should however be noted that the vast majority of human hotspots occur outside of gene promoter regions.

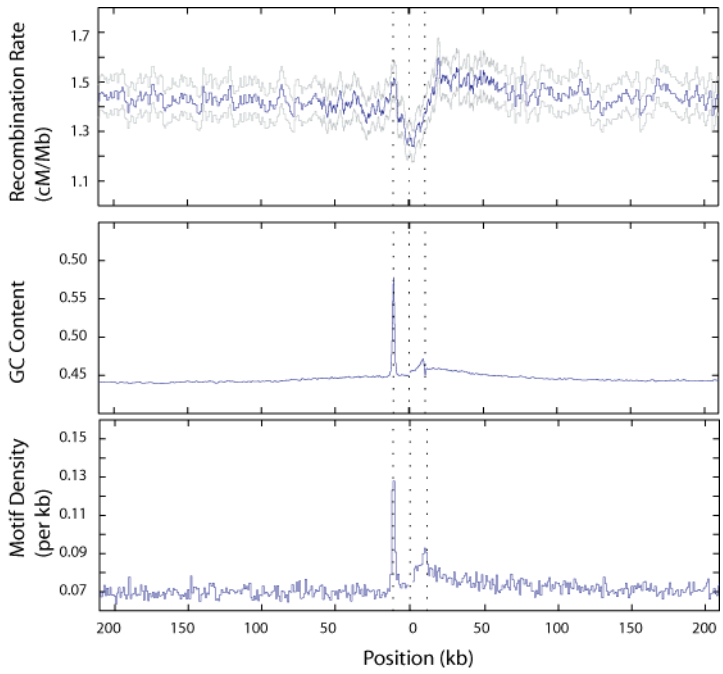


Figure 35. The distribution of recombination and associated features around genes. Top: The distribution of recombination around genes (shown again for clarity). Central: GC content around genes. Bottom: Density of hotspot-associated motifs within one substitution from the consensus CCTCCCTNNCCAC.

Levels of Recombination Vary Between Gene Ontology Groups

Given that rates of recombination are affected by the presence of genes, it is interesting to ask if there are systematic differences in the rates of recombination

between classes of genes. Previous work has demonstrated variation in the magnitude of linkage disequilibrium between genes of differing function (SMITH *et al.* 2005). I have therefore used the recombination rate estimated from Phase II of the HapMap to assess the evidence of rate variation between separate gene classes.

The Panther Database (THOMAS *et al.* 2003) is a gene ontology database containing predicted details of the gene molecular function (MF) and biological process. The MF ontology groups together genes of similar activity, such as enzymes and ion channels, each of which may be part of separate pathways. Conversely, the biological process ontology attempts to group genes that are required to achieve a process (such as signal transduction or electron transport), each of which may have a separate function. For the analysis presented here, I have focused on the MF categorisation as a similar study of the biological process categorisation did not reveal significant differences between gene classes.

In the MF ontology, genes are grouped into 28 top-level groups, with each gene allowed to exist in more than one group. I collected 14,979 non-overlapping autosomal genes from the RefSeq Annotation (PRUITT *et al.* 2005) for which recombination rates could be obtained. Of these, 9,735 had at least one assigned MF and genes without a MF were removed from the corresponding analysis. To control for gene size, I estimated the mean recombination rate for each gene over a 20kb region centred on the mid-point of the gene transcription region.

Genes were grouped by MF and a mean recombination rate was calculated for each group. The significance of the result from each group was calculated via a permutation test involving one-hundred thousand random groupings of genes. No correction was made for multiple testing. Furthermore, the permutation test assumes that gene recombination rate estimates are independent of each other, which may not

be an appropriate assumption due to broad scale autocorrelation in recombination rate estimates.

However, with due regard of the above caveats, my results indicate that there are significant differences between gene ontology groups. The mean recombination rate for genes with a MF was 1.34 cM/Mb. Recombination rates vary more than four-fold between gene groups (Figure 36), with Defence / Immunity genes showing the highest average rate (2.03 cM/Mb) and Chaperone genes showing the lowest (0.47 cM/Mb). Genes with molecular functions relating to external regions of the cell (such as Defence/Immunity, Cell Adhesion, Extra-cellular Matrix, Ion channels and Signalling) tend to show higher levels of recombination, while those with internal cell functions (such as Chaperones, Ligase, Isomerase, and Nucleic Acid Binding) tend to show lower rates of recombination. For the Chaperone, Defence / Immunity, Ligase, Nucleic Acid Binding, Receptor, and Signalling groups, all 100,000 permutations showed less extreme values, indicating that the results would remain significant after correction for multiple testing.

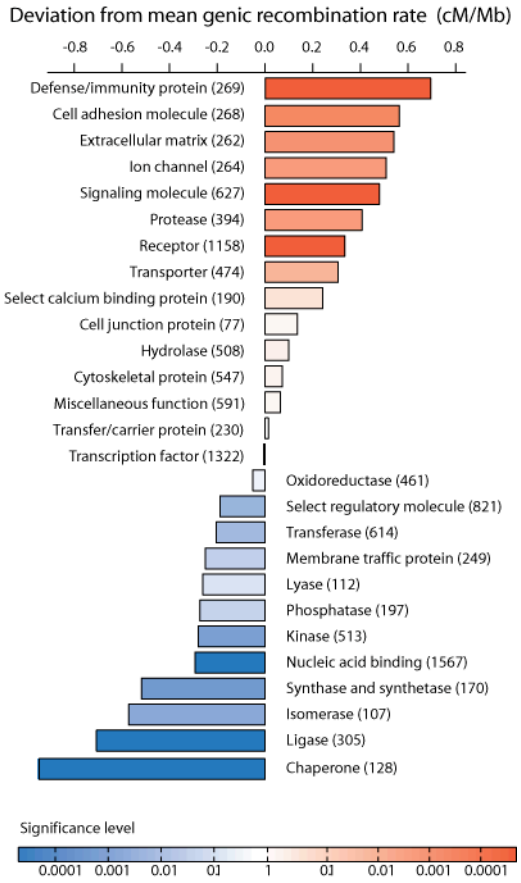


Figure 36. Average recombination rates for top level MF classes. The chart shows the relative increase or decrease relative to the genic genome average of 1.34 cM/Mb. Significance (as assessed using a permutation test) is indicated by the strength of the colour, with red indicating an increased recombination rate relative to the genic mean and blue indicating a decreased rate. The number of genes in each category is shown in brackets.

I wanted to explore if the observed pattern could be explained by the GC content of the genes. The raw correlation between GC content and recombination rate in the gene set explains approximately 5% of the observed variance. Nevertheless, GC content is known to correlate with recombination rates at least at the broad scale (KONG *et al.* 2002), and therefore should be accounted for. I performed a linear regression between the GC content and recombination rate of all the genes in each sample. Using the estimated regression parameters, the proportion of recombination explained by GC content was subtracted from each gene. Using the 'GC corrected'

recombination rates, the permutation test procedure was repeated. The resulting significance levels are very similar to those observed earlier (Figure 37, which can be compared to Figure 36), which confirms that the observed pattern in gene ontology rate estimates cannot be explained by GC content.

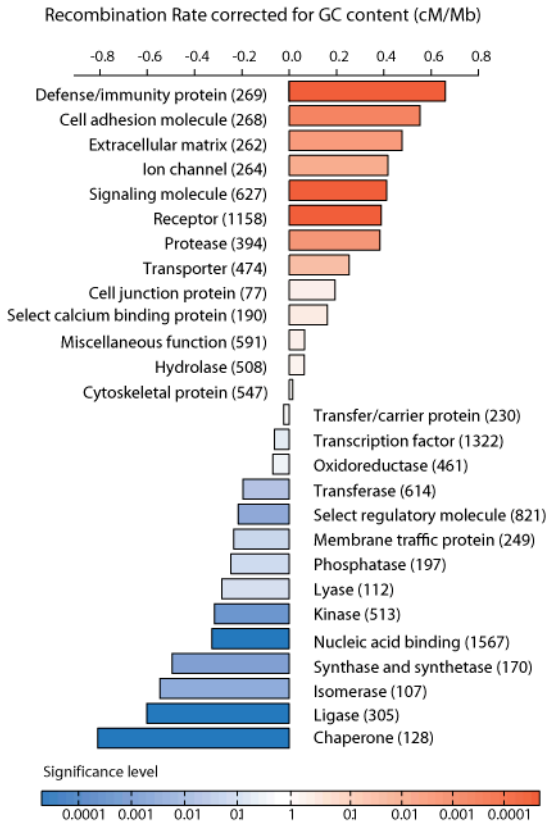


Figure 37. Average recombination rates for top level MF classes with recombination rates having been corrected for GC content. Significance (as assessed using a permutation test) is indicated by the strength of the colour, with red indicating an increased recombination rate relative to the genic mean and blue indicating a decreased rate. The number of genes in each category is shown in brackets.

That GC does not account for the observed patterns in recombination rates is perhaps not surprising when one notes that the correlation between GC content and recombination for the separate ontology groups is weak (Figure 38; top plot). However, the hotspot-related motif, CCTCCCTNNCCAC, does show systematic

variation between the gene classes. This in turn supports the claim that the observed pattern is not an artefact in the recombination rate estimates caused by selection in genic regions.

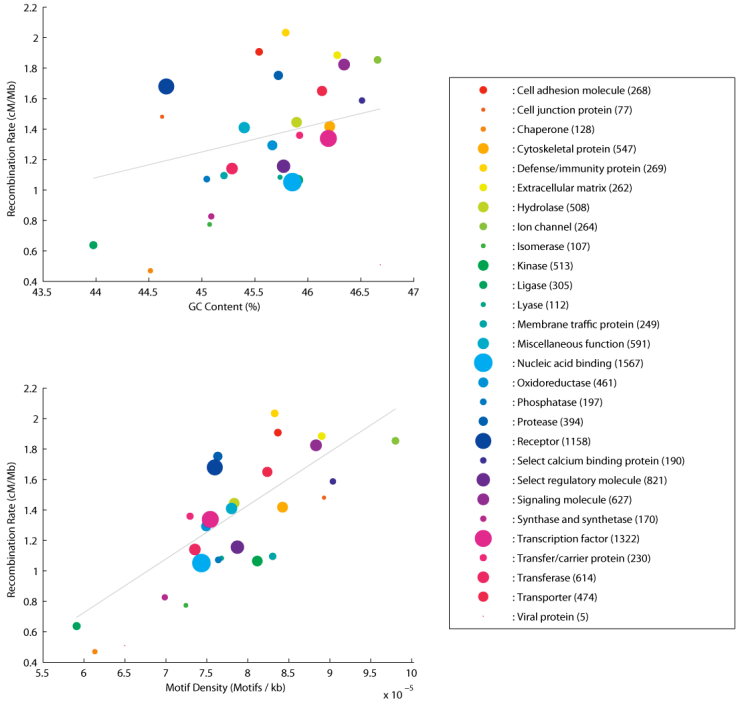


Figure 38. Top: GC content verses recombination rate for the Molecular Function Ontology groups. Bottom: Motif density (where a motif is defined as being no more than one substitution from the consensus CCTCCCTNNCCAC) verse recombination rate for the same ontology groups. The number of genes in each group is indicated by the size of the colour-coded points, and is also shown in brackets in the legend. Also shown on both plots is a weighted linear regression (grey line).

Finally, to check that the results are not an artefact caused by the chance clustering of genes, I repeated the analysis after locally shuffling the positions of genes along the chromosome. If clustering had occurred by chance then this process should leave much of the clustering of genes intact, but remove much of the

correlation between genes and regions of high or low recombination rate. In this shuffled set, none of the ontology categories obtained significance levels as high as those observed in the unshuffled set. This suggests that any grouping of genes in regions of high or low recombination rate is not something that occurred by chance.

Taken with the earlier results regarding the general suppression of recombination in genic regions, these gene ontology results pose interesting evolutionary questions. It is plausible that recombination is selected against in conserved genic regions, as double-strand breaks inherently involve some form of DNA damage, and such damage will result in deleterious haplotypes that are removed from the population via natural selection. Conversely, high levels of recombination may be provide a selective advantage for genes in which there is a benefit to a high level of allelic diversity due to changes in selection pressures (for example, resulting from changes in the environment or by the emergence of new pathogens). A similar pattern has been observed before (SMITH *et al.* 2005) in patterns of LD at the broad scale, and using a different ontology grouping. That similar patterns are observed in fine scale estimates using separate ontology definitions, suggests that this pattern is not an artefact. It will therefore be interesting to learn if similar patterns are observed in other species.

Local Patterns in Recombination around DNA Repeats

Having shown that genes can alter the local recombination landscape, I now consider another common genome feature, specifically DNA repeats. It has previously been shown that there is a large degree of heterogeneity in recombination rates between repeat families (MYERS *et al.* 2005). Using the Phase II HapMap, I have

explored patterns of recombination around a number of repeat families. To do this, I have used the RepeatMasker annotation (SMIT *et al.* 2004) to identify and classify regions of repeat DNA.

I start by noting that there are significant differences in the recombination rates of various repeat classes, a selection of which are shown in Figure 39. In this plot, I have considered the recombination rate estimated across 1kb windows centred on the midpoint of each repeat. We can see that certain repeats show significant deviations from the distribution of rates associated with randomly selected regions of the same size. Notably, the ALR/Alpha satellites (which are associated with centromeric regions) are significantly colder than average, whereas THE1B repeat elements containing the 13-mer motif are hotter.

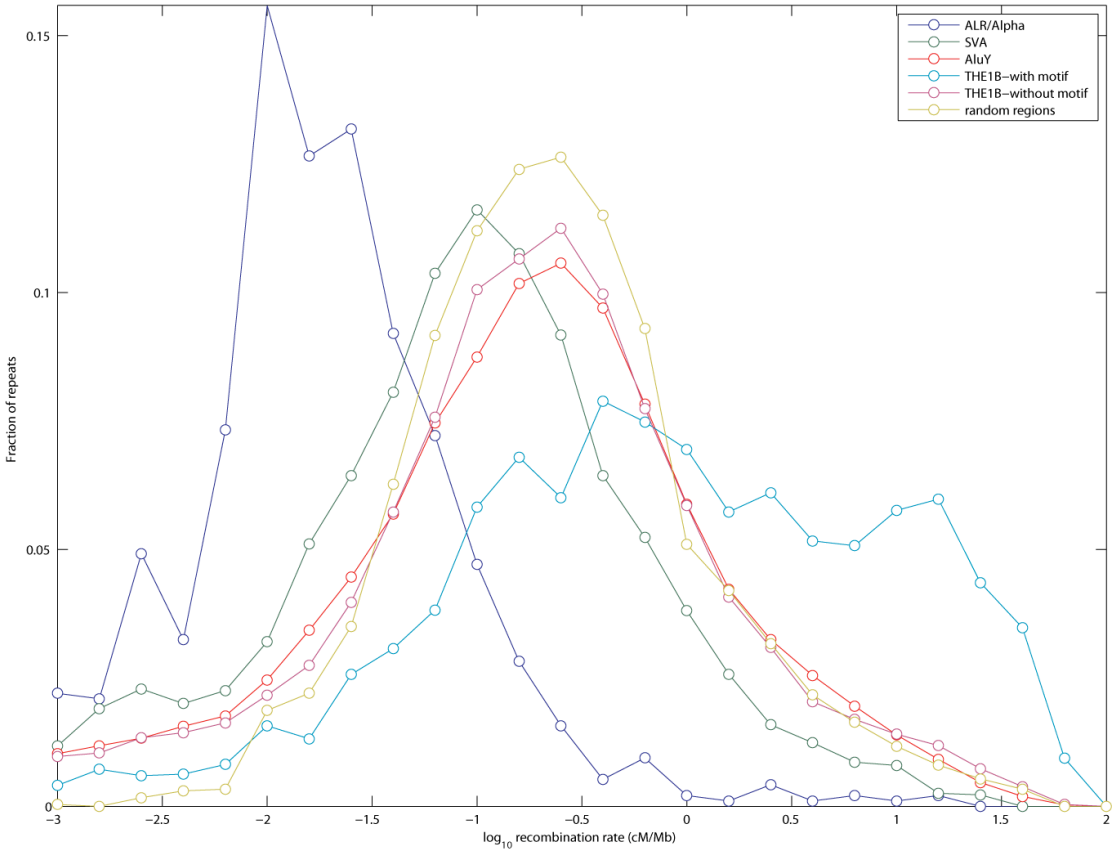


Figure 39. Distribution of recombination in a selection of repeat types. Shown here are the distributions recombination of the central 1kb of ALR/Alpha Satellites which are generally found near centromeric regions (blue), SVA retrotransposons (green), AluY elements (red), and the THE1B elements both with and without the motif (cyan and magenta respectively). Also shown is the distribution for 3,000 randomly selected 1kb regions from chromosome 2 (yellowish green). Note the logarithmic scale on the horizontal axis. Only repeats with coverage in the HapMap are considered. THE1B elements were said to have the motif if they contained a sequence within one substitution of the CCTCCCTNNCCAC consensus.

When a repeat family exhibits a deviation in recombination rates, then the deviations often appear to be local features extending no more than a couple of kilobases from the repeat as can be seen in Figure 40. These plots were constructed by averaging the local recombination rate over thousands of repeats, while ensuring that repeats are thinned sufficiently as to not interfere with each other. As the patterns are so localised, one is tempted to speculate that the increase (or decrease) in

recombination is caused by the presence of the repeats, as apposed to the repeats localising themselves in regions of high or low recombination. However, as we will see in Chapter 6, broad scale patterns also exist.

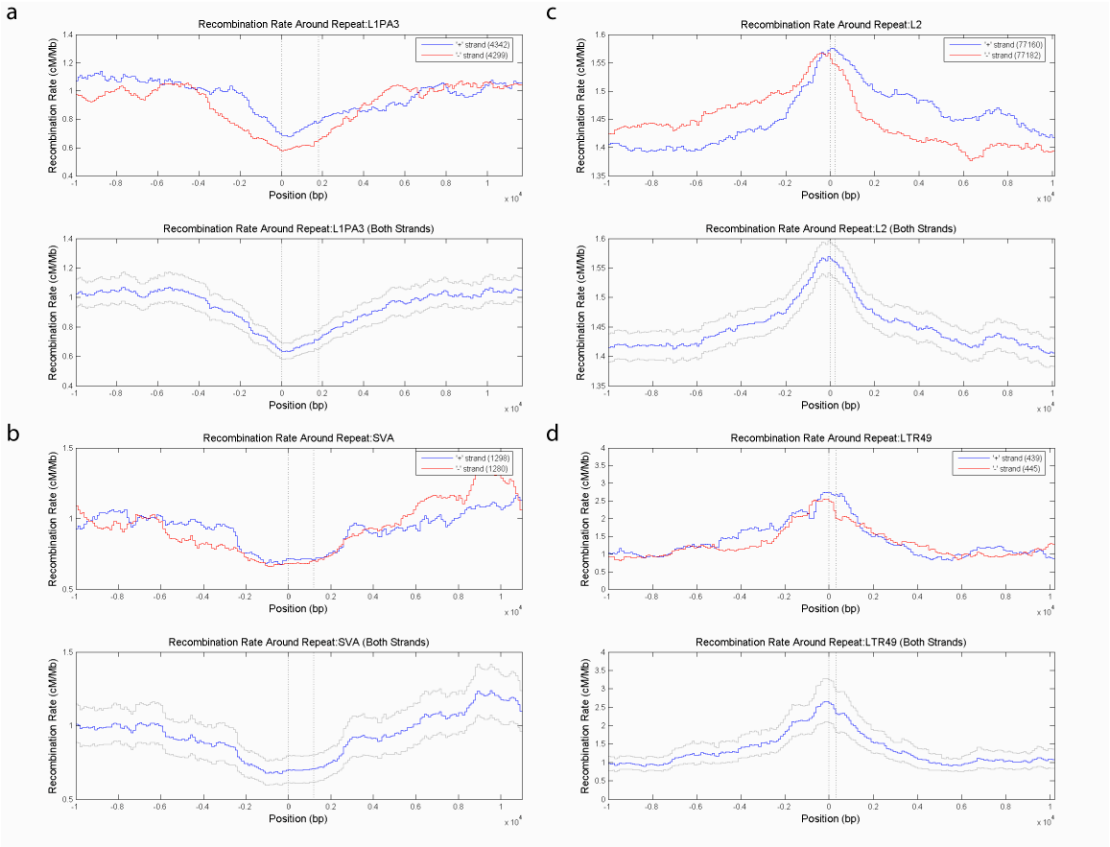


Figure 40. Local patterns in recombination rate variation. Shown here are four repeat types. For each repeat type, two plots are shown. The upper plot shows the mean recombination rate for repeats on the plus (blue) and minus strand (red). The lower plot shows the recombination rate averaged over both strands (blue), with 95% confidence intervals estimated by bootstrapping in grey. The four repeats shown are (a) L1PA3, (b) SVA, (c) L2, and (d) LTR49. The number of repeats in each group can be seen in the legend.

In Phase II HapMap, the hottest and coldest common repeats (with more than 1,000 occurrences) are shown in Table 5. It is notable that the hottest repeats all show some relation to the hotspot motifs discussed earlier (the consensus sequences of both the THE1 elements and LTR49 contain the 7-mer motif). It has been found that THE1

elements are found approximately twice as often in hotspots as in coldspots (MYERS *et al.* 2007), and my results include THE1A and THE1B within the six hottest repeat classes.

The coldest elements in the table are all part of the L1 family of retrotransposons, with the exception of the SVA elements. The SVA element is interesting in itself, as despite being one of the coldest repeat elements, it is almost entirely composed of C and G bases and the consensus contains at least seven copies of the 7-mer motif. This is highly suggestive of other factors controlling the activity of the motif, be it other sequence features or epigenetic factors.

Repeat Name	*# Occurrences*	*Mean Rate (cM/Mb)*	*95% Confidence Interval*	
(TGG)n	1237	3.77	3.24	4.31
THE1A	3843	2.60	2.37	2.85
(CCA)n	1207	2.56	2.22	2.91
(TCCC)n	1364	2.45	2.11	2.80
LTR49	1205	2.44	1.99	2.88
THE1B	20447	2.36	2.27	2.46
L1MEa	1391	0.66	0.55	0.76
L1HS	1098	0.65	0.48	0.82
L1PA15-16	1017	0.64	0.48	0.81
SVA	3152	0.64	0.57	0.72
L1PA2	4248	0.63	0.55	0.72
L1PA3	9104	0.58	0.54	0.63

Table 6 – The hottest and coldest common repeats elements in the human genome. The table shows the six hottest and coldest repeats with more than 1,000 occurrences.

With the exception of the SVA repeats, the repeat elements generally provide further evidence that the motif has a hotspot-causal role. The THE1B elements that contain the CCTCCCT hotspot-related motif are at least 5 times more frequent in hotspots than in THE1Bs without the motif. If we consider motifs within one substitution of the 13-mer, we can see that the activity of the THE1B is almost completely controlled by the presence or absence of this motif (Figure 41, and also Figure 39). THE1B elements with the motif are almost synonymous with hotspots, whereas those without the motif show very little in terms of rate elevation. Very similar patterns are also visible in the THE1A, THE1C and THE1D elements, although the magnitude of the elevation differs.

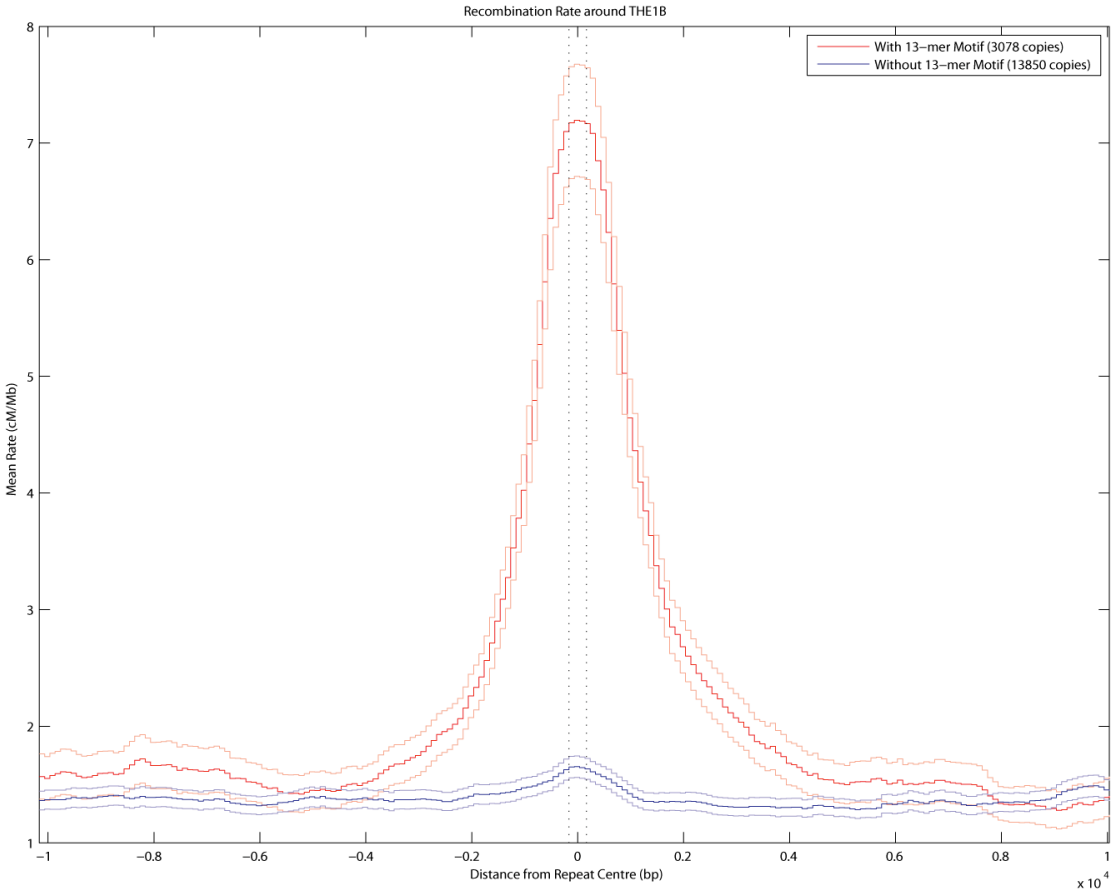


Figure 41. Recombination Rates around THE1B elements. Shown here are average rate estimates around elements containing motifs within one substitution of the 13-mer consensus (red), and those without such a motif (blue). Faded colours indicate the 95% confidence intervals of the average rate estimates based on ± 2 standard errors. Vertical dashed lines indicate the average extent of the THE1B elements.

Furthermore, while the activity of the motif is clearest in these THE1 elements, there also appears to be active motifs in other repeats. In Alu elements, there is evidence of an active 13-mer (specifically CCGCCTTGGCCTC), which shows homology to CCTCCCTNNCCAC (MYERS *et al.* 2007) but with three substitutions. As mentioned earlier, it remains unclear what conditions must be met for a motif to cause a hotspot and why the motif should differ between the THE1 elements and the Alu elements.

A Degenerate Motif?

While it is clear that sequence motifs play a role in determining the location of hotspots, the actual conditions that the motif must meet are unclear. For example, there are many instances of CCTCCCTNNCCAC that are not associated with hotspots, and in SVA elements the motif seems to be entirely inactive. Furthermore, there are other motifs that seem to be causal of hotspots, but are a full three substitutions away from the 13-mer consensus (for example, Alu elements containing the motif CCGCCTTGGCCTC show evidence of being hotspots; MYERS *et al.* 2007).

It is therefore highly unlikely that any single motif will account for recombination hotspots. However, it remains possible that a related family of motifs exist, each of which can cause a hotspot with some probability. With this in mind, I have attempted to search for motifs associated with hotspots in which bases are allowed to be more than one nucleotide. For example, such a motif may allow a certain base to be, say, a C *or* a T. However, searching for such degenerate motifs by brute force is not feasible. The alphabet I use to represent degenerate motifs contains 16 letters (not counting indels – see Table 7). For a motif of length L, the number of possible motifs is therefore 16^L, and it therefore quickly becomes impossible to test all motifs for all but small L (not to mention issues relating to multiple testing). Furthermore, we do not know how large L should be – it is quite feasible that hotspots are conditioned on two separate motifs separated by some distance, and hence L could be large. For this reason, we also wish to incorporate gaps, or indels, into the motifs for which we are searching.

A natural solution to searching such a large search space is a Genetic Algorithm (GA; see, for example, MITCHELL 1998). I created such a GA to search for

degenerate motifs, and the algorithm proceeded as follows. The algorithm took two sets of DNA sequence data as input. One dataset consisted of sequences from hotspots and the other consisted of sequences from coldspots. A population of motifs was randomly generated. At each iteration, the population was evaluated to find the motifs which best differentiated the two sets of data (that is, motifs which appeared more frequently in one dataset than the other). Motifs with low p-values (as assessed by Fisher's Exact Test) were considered to have a higher 'fitness' than those with higher p-values. The algorithm also regarded less degenerate motifs as fitter (the degeneracy penalties are shown in Table 7). Specifically, the algorithm attempted to minimise the log of the product of the p-value and total motif degeneracy. The fittest motifs survived to the next iteration, with the lower half being removed from the population. These surviving motifs were randomly selected (with a bias towards fitter motifs) to produce a new set of motifs to replenish the population. The new motifs were generated by a process of 'recombination', which combined motifs at a randomly selected location, and 'mutation', which randomly altered a single base. This process was repeated for a number of iterations.

Code	Base Name	Meaning	Degeneracy Penalty
A	Adenine	A	1
C	Cytosine	C	1
G	Guanine	G	1
T	Thymine	T	1
R	Purine	G / A	2
Y	Pyrimidine	T / C	2
K	Keto	G /T	2
M	Amino	A / C	2
S	Strong Interaction (3 H bonds)	G / C	2
W	Weak Interaction (2 H bonds)	A / T	2
B	Not A	G / T / C	3
D	Not C	G / A / T	3
H	Not G	A / T / C	3
V	Not T	G / A / C	3
N	Any nucleotide	A / C / G / T	4
*	Gap	Gap of any length	5

Table 7 - DNA alphabet and meanings. Also shown is the Degeneracy, which describes the penalty imposed on the motif fitness by including a degenerate base.

Using a dataset consisting of sequence data from 9292 *LDhot* hotspots less than 5kb wide, with a set of an equal number coldspots matched for size, SNP density and GC content (within 1%), I employed the GA to search for motifs. The single motif with the highest fitness found by the GA was the 13-mer CCNCCNYNVCCMY (hereafter referred to as Motif A). This motif was found in 4304 hotspot sequences, against 2669 coldspot sequences – a relative ratio of 1.61. The motif is compatible with many of those identified as significant by Myers *et al*, including those identified in the THE1A/B elements (MYERS *et al.* 2007). If one considers the relative position

of the motif with the DNA sequence, one observes that the motif tends to be located towards the centre of the hotspot sequences (Figure 42).

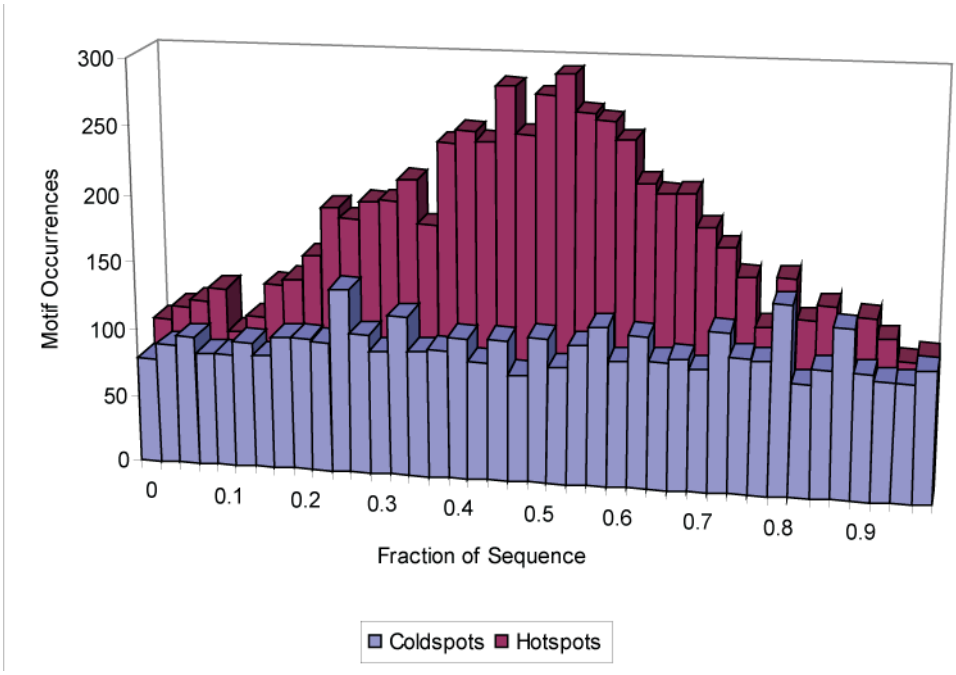


Figure 42. The frequency distribution of Motif A in hotspots (magenta) and coldspots (cyan). Each hotspot and coldspot region was rescaled to be of a unit interval in width. This plot shows the frequency of the motif at each position on the unit interval. We see that motifs occur more frequently towards the centre of a hotspot interval, whereas have a near uniform distribution in coldspots.

Assessing the significance of this finding is problematic. In an informal sense, the genetic algorithm is the equivalent of performing a vast number of statistical tests, and hence the p-values reported by the algorithm cannot be trusted without further corroboration. Therefore, to assess the significance of the Motif A, I re-ran the algorithm on a set of data with the 'hotspot' and 'coldspot' labels randomly permuted. I repeated this procedure 100 times (which was the maximum achievable due to the high computational cost of running the GA many times). Of all 101 runs, the non-

permuted run achieved the highest fitness, suggesting that the found motif is indeed 'significant'.

Motif A is consistent with the 7, 9 and 13-mer hotspot related motifs that were discussed earlier, and suggests that certain bases of the motif are more important than others. The most important feature seems to be the spacing of the Cytosine bases at positions 1, 2, 4, 5, 10 and 11. However, this is hardly a rigid definition. In total, Motif A corresponds to a family of 1536 unique non-degenerate motifs. To further examine the family of motifs contained with Motif A, I tested all of these non-degenerate motifs to see how they differentiated the hotspot and coldspot sequence sets. Of these, 308 did not occur at all in the dataset, and 111 occurred only in coldspots. Of the remaining motifs, 303 only occurred in hotspots and 814 had relative ratios greater than one. However, only 20 motifs had p-values below 0.001 and these are shown in Table 8. Interestingly, only one of these 20 motifs was enhanced in coldspots (Motif 6). This result is unique in the 100 motifs with the lowest p-values. This motif appears to be associated with L1 elements (SMIT *et al.* 1995), which are known to occur more frequently outside of hotspots (see Table 6 and Myers *et al*., 2006).

The top five motifs in Table 8 are all within a few mutations of the original (non-degenerate) 13-mer motif, which is enriched within THE1 elements. Despite this, we found Motif A to be further enhanced outside of repeat sequence. Again using RepeatMasker (SMIT *et al.* 2004), I identified and masked all repetitive sequence in the dataset. I found that Motif A was strongly enhanced outside repeat sequence (relative ratios of 1.90 for non-repeat sequence compared to 1.42 in repeat sequence). However, almost paradoxically, no single non-degenerate motif was enhanced with p-values less than 10^{-5} in the non-repeat sequences (Table 9 and Table 10). This is

consistent with a family of motifs being associated with hotspots. In repetitive sequence, the probability of each motif in the family is skewed by the requirement of being a repetitive element. In non-repetitive sequence, no such constraint exists, and hence each motif in the family may occur at low frequency.

Finally, it is also interesting to consider what the GA did not find. Despite being able to extend the motif to contain more bases, the GA did not consider this to be a useful move and settled quite stubbornly on a motif of 13 bases. Furthermore, it was also possible for the GA to look for secondary motifs operating at some distance from the core motif (i.e. upstream or downstream of the core motif). Again, the GA did not consider this to be beneficial. Although not conclusive, this suggests to this author that there is a limit to the relationship between sequence and hotspot location, and that the deciding factor between 'hot' and 'cold' motifs is not related to the DNA sequence. Possible other factors include epigenetic factors, which I will consider towards the end of this chapter.

	Motif	# Hotspots	# Coldspots	p-value	Relative Ratio
1	CCTCCCCAGCCAC	160	25	1.66E-25	6.4
2	CCTCCCTAGCCAT	100	4	3.69E-25	25
3	CCTCCCTAGCCAC	31	0	9.08E-10	n/a
4	CCTCCCTGACCAC	30	1	2.92E-08	30
5	CCTCCCTGACCCT	19	0	3.78E-06	n/a
6	CCCCCCTCCCCCC	68	131	8.45E-06	0.519084
7	CCTCCCTTCCCAC	26	4	5.87E-05	6.5
8	CCTCCCTGGCCAC	20	2	0.00012	10
9	CCACCCTGACCAC	14	0	0.000121	n/a
10	CCACCCTACCCCC	13	0	0.000243	n/a
11	CCACCCTTGCCCC	13	0	0.000243	n/a
12	CCTCCCTACCCCC	13	0	0.000243	n/a
13	CCCCCCCAACCCC	21	3	0.000275	7
14	CCTCCCTGACCCC	20	3	0.000485	6.66667
15	CCTCCCTCCCCAC	40	14	0.000526	2.85714
16	CCTCCCTGCCCCC	35	11	0.000527	3.18182
17	CCCCCCCACCCCC	55	24	0.000623	2.29167
18	CCCCCCCGCCCCC	29	8	0.000744	3.625
19	CCTCCCTTGCCCC	14	1	0.000972	14
20	CCTCCCTAGCCCC	11	0	0.000974	n/a

Table 8 - Motifs compatible with CCNCCNYNVCCMY showing the most significance in the complete dataset (Fisher Exact Test p < 0.001).

Motif Number	Motif	# Hotspots	# Coldspots	p-value	RR
17	CCCCCCCACCCCC	44	13	4.59E-05	3.38462
18	CCCCCCCGCCCCC	21	2	6.54E-05	10.5
7	CCTCCCTTCCCAC	17	1	0.000144	17
6	CCTCCCTTCCCCC	17	1	0.000144	17
-	CCCCCACCCCCAC	51	19	0.000162	2.68421
5	CCTCCCTGACCCT	12	0	0.000487	n/a

Table 9 - Significant (Fisher Exact Test p < 0.001) non-degenerate motifs in non-repeat sequence. Motif Number refers to the motif numbers in Table 8.

Motif Number	Motif	# Hotspots	# Coldspots	p-value	RR
2	CCTCCCTAGCCAT	98	3	1.06E-25	32.66667
1	CCTCCCCAGCCAC	151	23	1.82E-24	6.56522
3	CCTCCCTAGCCAC	28	0	7.30E-09	n/a
6	CCCCCCTCCCCCC	55	126	1.19E-07	0.436508
4	CCTCCCTGACCAC	18	0	7.57E-06	n/a

Table 10 – Significant (Fisher Exact Test p < 0.001) non-degenerate motifs in repeat sequence. Motif Number refers to the motif numbers in Table 8.

No Evidence of Unrelated Secondary Motifs

To investigate the possibility of an additional motif, I removed all sequences from the hotspot and coldspot datasets that contained the CCNCCNYNVCCMY motif. I then used the GA on this subset of data. This analysis was repeated for the dataset with masked repeat sequence. In all cases, the motifs found by the GA showed homology to the CCNCCNYNVCCMY motif, and no motif achieved a fitness remotely near that of the original motif. This would suggest that there are no

secondary hotspot-related motifs that cannot be related to the motifs already discovered. It would therefore appear that a significant fraction of hotspots do not contain sequence motifs with easily identifiable common features.

The Motif in Relation to Epigenetic Factors

I return now to the original CCTCCCTNNCCAC non-degenerate motif. As at least one study has reported a relationship between DNA hypersensitivity and recombination hotspots (LI *et al.* 2006), I wanted to investigate the possibility that the activity of this motif is controlled by an epigenetic factor. Of 3395 occurrences of the motif in the human genome, 685 occur in (detected) hotspots, 164 in (matched) coldspots and 2546 occur elsewhere. Furthermore, the majority of hotspots do not appear to contain a motif of this form. It is therefore clear that while the motif is a major factor, it is neither necessary nor sufficient to cause a recombination hotspot.

Further sequence analysis has failed to identify the conditions by which the motif becomes active. It is therefore tempting to consider so-called epigenetic factors. Epigenetic factors affect a cell, organ or individual without directly affecting its DNA sequence. For example, an epigenetic change may indirectly influence the expression of genes in the genome. In the context of this thesis, an epigenetic factor is a quantity related to, and varying along, a DNA sequence that cannot be assessed from the sequence directly. However, quantitative study of such epigenetic factors is complicated by the lack of common standards.

In an attempt to develop suitable technologies for the quantification of epigenetic factors, the National Human Genome Research Institute (NHGRI) launched a public research consortium named ENCODE, the Encyclopaedia Of DNA

Elements, in September 2003 (THE ENCODE PROJECT CONSORTIUM 2007). This project aimed to identify all functional elements in approximately 1% (30MB) of the human genome.

The regions of study selected by the ENCODE project were grouped into two types: manually selected and randomly selected. The manually selected regions were chosen for being in some sense 'interesting' – that is they contained well studied genes or other known sequence elements. A total of 14.82Mb of sequence was selected in this way, divided into 14 regions ranging from 500kb to 2Mb. The so-called random regions comprised of 30 regions all of 500kb in size. Despite the name, the random regions were not in fact selected uniformly at randomly, but by a method that ensured a selection of regions which varied widely in terms of gene content and other functional elements.

The ENCODE project provides a dataset for the analysis of epigenetic factors that can be downloaded from the UCSC website (http://genome.ucsc.edu/ENCODE/). Using this data, I have undertaken an exploratory investigation into the epigenetic properties of the motif. However, working with ENCODE data presents a number of practical challenges. Firstly, the number of annotations is large - as of September 2006, the ENCODE database contained over 500 tables (UCSC build hg17) with at least four different table formats. Secondly, the annotations cover a wide range of experimental techniques, each of which exhibit peculiarities of their own. As it was unclear what signal to look for, and to avoid making assumptions about the properties of the data, I have explored the data using a graphical technique.

I identified all instances of the motif within the ENCODE sequence which were within one substitution of the motif. In total, there were 1801 examples. Of these 217 occurred within hotspots, and 124 occurred within the matched set of coldspots.

Using only the motifs in hotspots and coldspots, I removed bias due to clustering by thinning motifs so that no two were within 250bp of each other. This procedure left 205 motifs contained within hotspots, and 122 contained within coldspots. I further divided these groups into motifs on the plus strand, and motifs on the minus strand of the DNA (i.e. motifs with the sequence CCTCCCTNNCCAC were deemed to be on the plus strand, and motifs with the complementary sequence GTGGNNAGGGAGG were deemed to be on the minus strand).

Using the ENCODE database (build hg17), I selected 216 annotations that provided information which could not be assessed using the sequence alone (as described in Table 11). A window was constructed around each motif, and annotations were retrieved for each of these windows. Windowed annotations were aligned at the motif start base, and averaged over the set of hot and cold motifs separately. Confidence intervals were calculated by bootstrapping the annotation values at each site. The procedure described here is similar to that used for displaying the average recombination rate over genes (e.g. Figure 34) or DNA repeats (e.g. Figure 41).

The averaged annotation could then be compared visually and assessed for significant differences between hot and cold motifs. To guard against issues due to sequence dependence (which many annotations exhibited – data not shown), I treated the plus and minus strands separately and required that any observed signal be consistent between the plots of the plus and minus strands. If a signal were visible on only one strand, then the most probable explanation is that the annotation is in some way dependent on base composition.

Variations on the above method were also tried, including only considering the exact motif (with no substitutions). However, in these situations, it was often the case that there was insufficient data to gain a reliable annotation signal.

Annotation Type	Description	Experimental Groups
Chromatin Immunoprecipitation	Uses antibody binding to identify regions of DNA attached to protein. Provides a measure of the accessibility of the DNA.	Affy[1], GIS[2], Sanger[3], Stanford[4], UC Davis[5], UCSD[6], Uppsala[7], UT-Austin[8], Yale[1]
Methylation	Methyl-sensitive restriction enzymes used to assess the methylation status of CpG regions.	Stanford[3]
DNA Structure	Uses various methods to assess the structure of DNA.	BU[9], NHGRI[10], UNC[11], UT-Austin[12], UVa[13], UW[14]
Transcript Levels	Estimate of RNA abundance (transcription) for several cell types.	Affy[1], Riken[15], Yale[16]

Table 11 - Considered ENCODE annotations. References: 1 (CAWLEY *et al.* 2004) 2. (WEI *et al.* 2006) 3. No Reference 4. (TRINKLEIN *et al.* 2004) 5. (BIEDA *et al.* 2006) 6. (KIM *et al.* 2005b) 7. (RADA-IGLESIAS *et al.* 2005) 8. (KIM *et al.* 2005a) 9. (BALASUBRAMANIAN *et al.* 1998) 10. (CRAWFORD *et al.* 2006) 11. (NAGY *et al.* 2003) 12. (BHINGE *et al.* 2007) 13. (JEON *et al.* 2005) 14. (SABO *et al.* 2006) 15. (SHIRAKI *et al.* 2003) 16. (CHENG *et al.* 2005).

The results of this study were largely negative. While a number of annotations showed differences between the hot and cold motif sets on a single strand, no single annotation showed an obvious difference on both DNA strands. In fact, only a single

annotation showed any statistically significant difference between the two sets on both DNA strands, namely the Sanger H4ac Molt4 ChIP annotation (Figure 43). However, the Sanger ChIP signals are relatively noisy and the difference in signal between the two groups only just achieves significance. The difference is therefore unconvincing, and is most likely a false positive, which would be expected due to the large number of annotations that have been considered. The conclusion that this is a false positive is further supported by the fact that a similar pattern is not observed in other ChIP annotations.

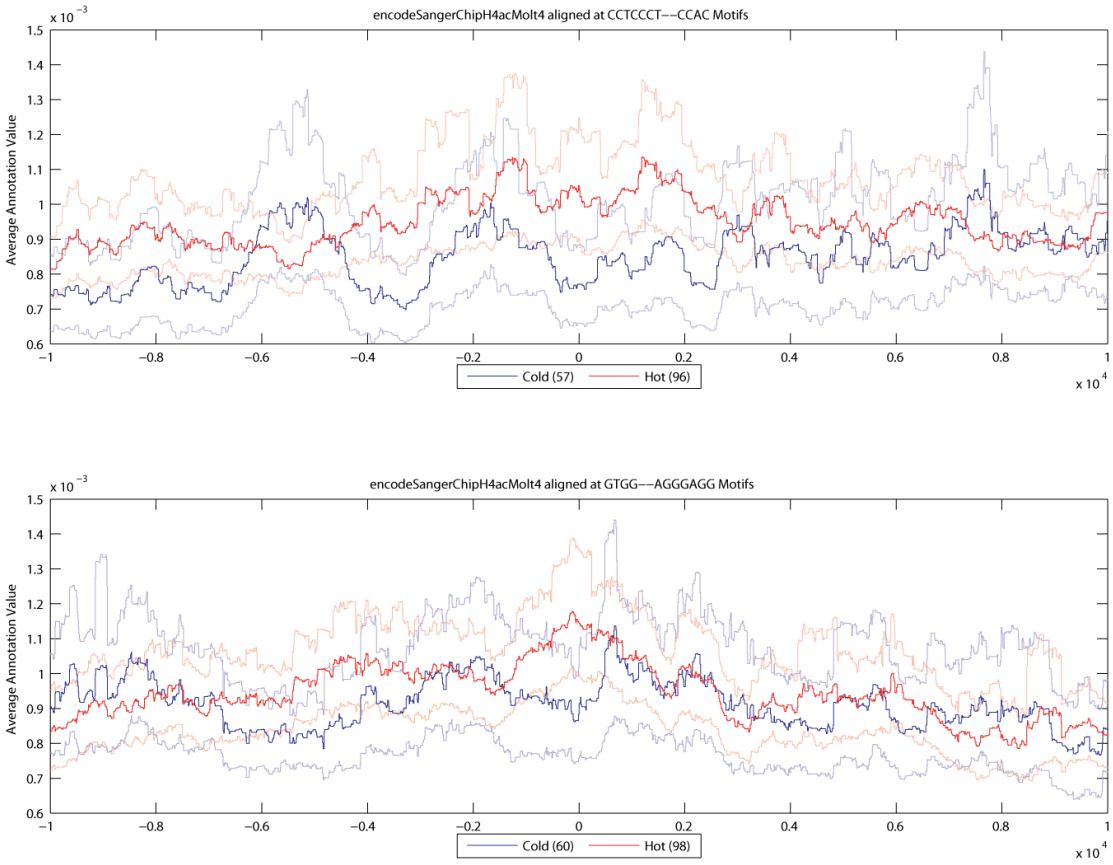


Figure 43. The ENCODE Sanger ChIP H4ac Molt4 Annotation. Shown here is the annotation aligned at motifs on the + strand (top) and the - strand (bottom) over a 20kb region. The average annotation value for the hot motifs is shown in red, and the same value is shown for the cold motifs in blue. 95% confidence intervals are shown in faded colours (as assessed by bootstrapping). Note the small region around the motif (position 0 on the horizontal axis) where the annotations diverge and just achieve significance.

The hypothesis that the activity of the motif is controlled by epigenetic factors is therefore not supported by this work. There are at least three possible explanations for the negative result. The first possibility is that I may not have sufficient power to detect differences between hot and cold motifs via the visual method. Given the relatively small number of motifs in our dataset and the high degree of noise in many of the annotations, it is plausible that the epigenetic signal was too faint to be detected via the methodology used. Unfortunately, given the wide range of technologies

considered, it is difficult to assess our expected power as the level of noise and resolution associated with each signal is unclear.

A second possibility is that the epigenetic factor that is responsible for motif activity may not contained within ENCODE. Alternatively, the relevant epigenetic factor may be included within the ENCODE data, but has not been applied to the correct cell line. As we are primarily interested in meiotic recombination, the ideal cell lines for study would be those involved in the generation of gamete cells (so-called gametogonia and gametocytes). However, no such experimental cell lines are currently in existence. As it is unclear to what level each annotation is conserved between cell types, this is a plausible explanation.

A third possibility is that no single epigenetic factor controls that activity of a motif, but specific a combination of factors does. Such a possibility was not covered by my analysis, and it is difficult to envisage a study being able to demonstrate such a relationship without prior knowledge of the relationship given the large number of possible combinations and relatively small amounts of data available (the ENCODE project covers approximately 1% of the genome).

A final possibility is that the activity of the motif is not controlled by epigenetic factors. If this is indeed the case, then the controlling factor must be some as-yet unknown sequence feature. While it is very difficult to rule this possibility out, it seems unlikely given the extensive search for sequence features that has been undertaken by myself and others (JEFFREYS and NEUMANN 2005; KONG *et al.* 2002; MCVEAN *et al.* 2004; MYERS *et al.* 2005; MYERS *et al.* 2007).

Discussion

In this chapter, I have described the features of recombination rate variation in the human genome. The recombination rate estimates were obtained using *rhomap* using data from Phase II of the HapMap project. Analysis of these rates showed extensive variation over a wide range of scales. However, recombination in the human genome is dominated by hotspots with more than 25,000 detected.

The rate of recombination also appears to be affected by genome features such as genes. Primarily, recombination is suppressed within genes. However, I have shown that there is extensive variation between gene ontology groups. The observed pattern has an evolutionary interpretation that requires testing in other species.

I also observed local patterns of recombination around DNA repeats. These patterns are repeat family specific and local in nature. As such, it appears probable that these patterns are caused by the properties of the repeats themselves. Furthermore, the activity of at least one repeat family appears to be controlled by the presence or absence of a sequence motif.

I have attempted to isolate the features of hotspot-associated motifs by using a genetic algorithm to construct a degenerate motif. The algorithm discovered a motif that is consistent with the majority of motifs discovered previously (MYERS *et al.* 2005; MYERS *et al.* 2007). However, despite this motif being highly enriched within hotspots, the majority of hotspots remain unexplained. I therefore investigated the possibility that there exist epigenetic factors that control the activity of hotspot-associated motifs. However, this investigation did not provide positive results.

In the next chapter, I continue to investigate recombination in the human genome using the rate estimates described in the chapter. Whereas my investigation in

this chapter has largely focused on local patterns of variation, I will use wavelet analysis in the following chapter to study variation at a wide range of scales.

Chapter 5 A Wavelet Analysis of Recombination

In this thesis we have seen that recombination rates in humans vary significantly both at the fine and broad spatial scales (JEFFREYS and NEUMANN 2005; KONG *et al.* 2002). At the fine scale, the recombination landscape is dominated by recombination hotspots of 1 to 2kb in width where the recombination rate can be hundreds of times that of the surrounding region. At the broad scale, recombination 'jungles' and 'deserts' of megabase scale have been identified (KONG *et al.* 2002), with the recombination rate in jungles being ten-fold greater than that of deserts. Furthermore, recombination rate in a given region is related to a number of genome features that also vary over a number of scales. For example, the hotspot motif may be expected to affect recombination at the fine scale, whereas GC content has been shown to have an effect at a much broader scale (JENSEN-SEAMAN *et al.* 2004; KONG *et al.* 2002). In this chapter, I use a Wavelet analysis to study the *rhomap* recombination rate estimates from the Phase II HapMap discussed in the previous chapter, and to relate changes in the recombination rate to those in other genome features on a scale-by-scale basis.

Introduction to Wavelets

Wavelets are a mathematical tool commonly used in signal analysis. While they are most often applied to time series or images, they can be applied to a wide range of signals. They are particularly useful for the analysis of signals that show

variation over a range of scales. For this reason, they are a very useful tool for the analysis of genome features. In the following discussion, I introduce the basic principles of wavelet analysis. For a more complete introduction and discussion of wavelet analysis, I direct the reader the book of Percival and Walden (2000).

There are two types of wavelet analysis. The first, known as the Continuous Wavelet Transform (CWT), is intended to work (as the name suggests) with signals defined over an entire axis. The second, the Discrete Wavelet Transform (DWT), is conversely intended to analyse a signal defined at a finite range of integer points. To motivate the use of wavelets, I begin with a discussion of the CWT.

A wavelet can be thought of as a waveform with finite range. However, to be a wavelet, the following two properties must also be satisfied.

The integral of the wavelet function, $\psi(.)$, is zero:

$$\int_{-\infty}^{\infty} \psi(u) du = 0 \tag{5.1}$$

The integral of the square of the wavelet function is unity:

$$\int_{-\infty}^{\infty} \psi^2(u) du = 1. \tag{5.2}$$

Note that an infinite wave such as the sine function may satisfy the first property, but does not satisfy the second property. These two conditions ensure that while a wavelet may make local deviations from zero, these deviations must be cancelled out by deviations elsewhere in the wavelet. Furthermore, the wavelet must be of finite size.

Examples of three wavelet functions are shown in Figure 44. The simplest wavelet is the Haar wavelet, which can be seen in the left-hand plot and has the function:

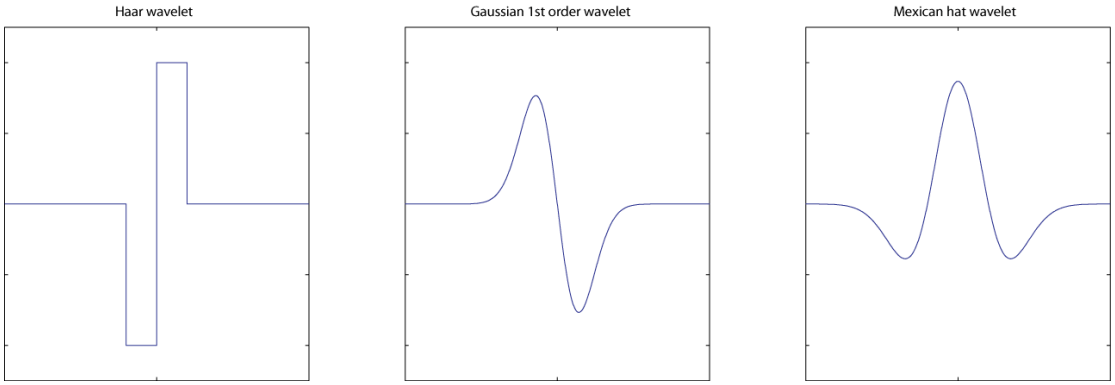


Figure 44. Three wavelet functions. From left to right: The Haar wavelet; a wavelet related to the derivative of the Gaussian PDF; the Mexican hat wavelet. The function defining the wavelet is often referred to as the 'mother wavelet'.

$$\psi^{(H)}(u) \equiv \begin{cases} -1/\sqrt{2}, & -1 < u \le 0; \\ 1/\sqrt{2}, & 0 < u \le 1; \\ 0, & \text{otherwise} \end{cases} \tag{5.3}$$

The Haar is the oldest wavelet, and was named after Alfred Haar who developed it in early 20th century (HARR 1910), although the term 'wavelet' was not used in this context until the 1980s. I will be making extensive use of the Haar wavelet in this chapter.

Given a wavelet function, it is possible to obtain a representation of a signal in terms of a combination of wavelets and hence break a signal down into component parts. Analogous to the Fourier Transform for sine and cosine waveforms, the CWT can be used to deconstruct the signal in terms of the wavelet function. The CWT returns a number of coefficients that relate the contribution of a wavelet at a given position and of a given size (or scale) to the original signal. In an informal sense, the original signal is replaced by a collection of wavelets that have been scaled, stretched and translated (Figure 45).

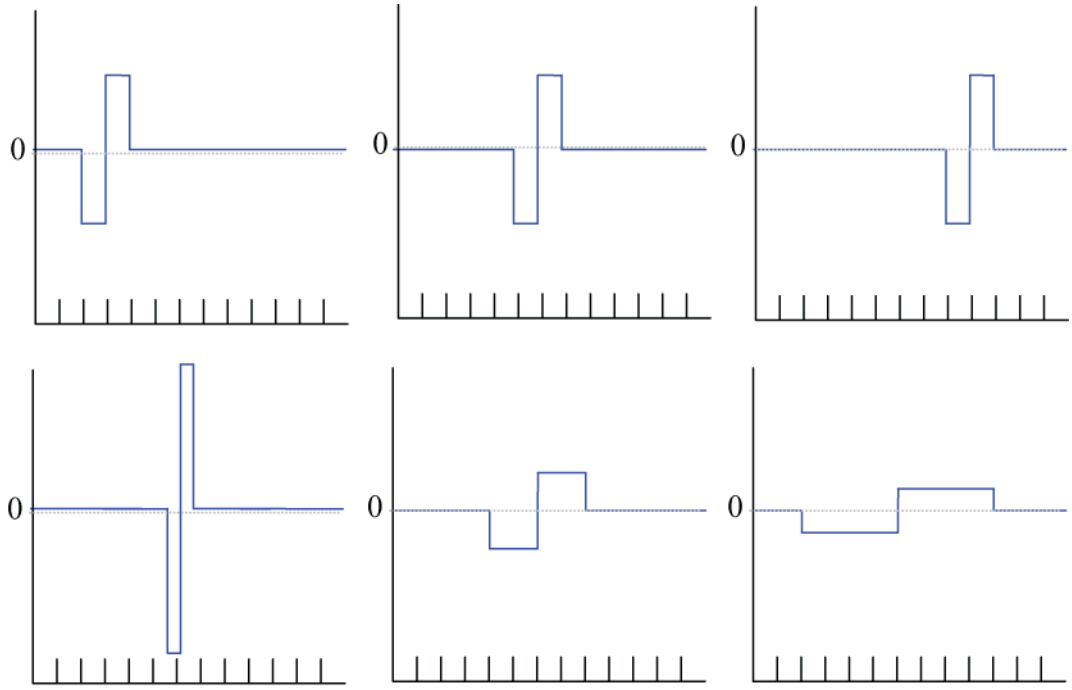


Figure 45. Examples of transformations applied to the Haar wavelet. The CWT expresses the original signal in terms of wavelet functions that have undergone some transformation, as demonstrated here. The top row shows three Haar wavelets from a single scale translated to differing locations, whereas the lower row shows three Haar wavelets at the same location but from differing scales. The coefficients resulting from a wavelet decomposition describe the contribution to the signal from the set of all possible transformed wavelets.

Mathematically, given a signal, $x(.)$, the CWT coefficients of the wavelets at each scale, λ, and position, t are given by the following integral:

$$W(\lambda,t) = \int_{-\infty}^{\infty} \psi_{\lambda,t}(u)x(u)du \ , \quad \text{where } \psi_{\lambda,t}(u) \equiv \frac{1}{\sqrt{\lambda}}\psi\left(\frac{u-t}{\lambda}\right) \tag{5.4}$$

Importantly, all the information in the original signal is preserved by the CWT. We can recover $x(.)$ via:

$$x(t) = \frac{1}{C_\psi}\int_0^{\infty}\left[\int_{-\infty}^{\infty} W(\lambda,t)\frac{1}{\sqrt{\lambda}}\psi\left(\frac{t-u}{\lambda}\right)du\right]\frac{d\lambda}{\lambda^2} \tag{5.5}$$

The inner integral of equation (5.5) can be considered as the summation of wavelets at all possible positions while the outer integral can be considered as the summation of wavelets at all possible scales. The success of the signal reconstruction depends on the scaling constant C_ψ, known as the admissibility constant (MALLAT 1999), which can be calculated using the Fourier Transform of the wavelet function;

$$\Psi(f) = \int_{-\infty}^{\infty} \psi(u) e^{-i2\pi fu} du$$

and must satisfy the admissibility condition:

$$C_\psi \equiv \int_0^\infty \frac{|\Psi(f)|^2}{f} df \text{ satisfies } 0 < C_\psi < \infty \qquad (5.6)$$

Depending on the wavelet used, the CWT can be used to build a picture of how the signal averages (in the case of the Haar wavelet) or weighted averages (in the case of more complicated wavelets) are changing from one region to the next. For example, Figure 46 shows a CWT and DWT applied to the GC content annotation of a 1Mb region of chromosome 6. The wavelet used in these decompositions comes from the family of wavelets known as the symlets (due to their near-symmetrical nature; DAUBECHIES *et al.* 1992). The wavelet used here (specifically the 6th symlet) largely resembles the Mexican Hat wavelet in Figure 44. Apparent from the form of this wavelet, we think of the coefficients of this wavelet as taking the difference between the mean of the signal at the wavelet centre and the mean of the flanking regions. The coefficients are therefore large when signal peaks are surrounded by signal troughs.

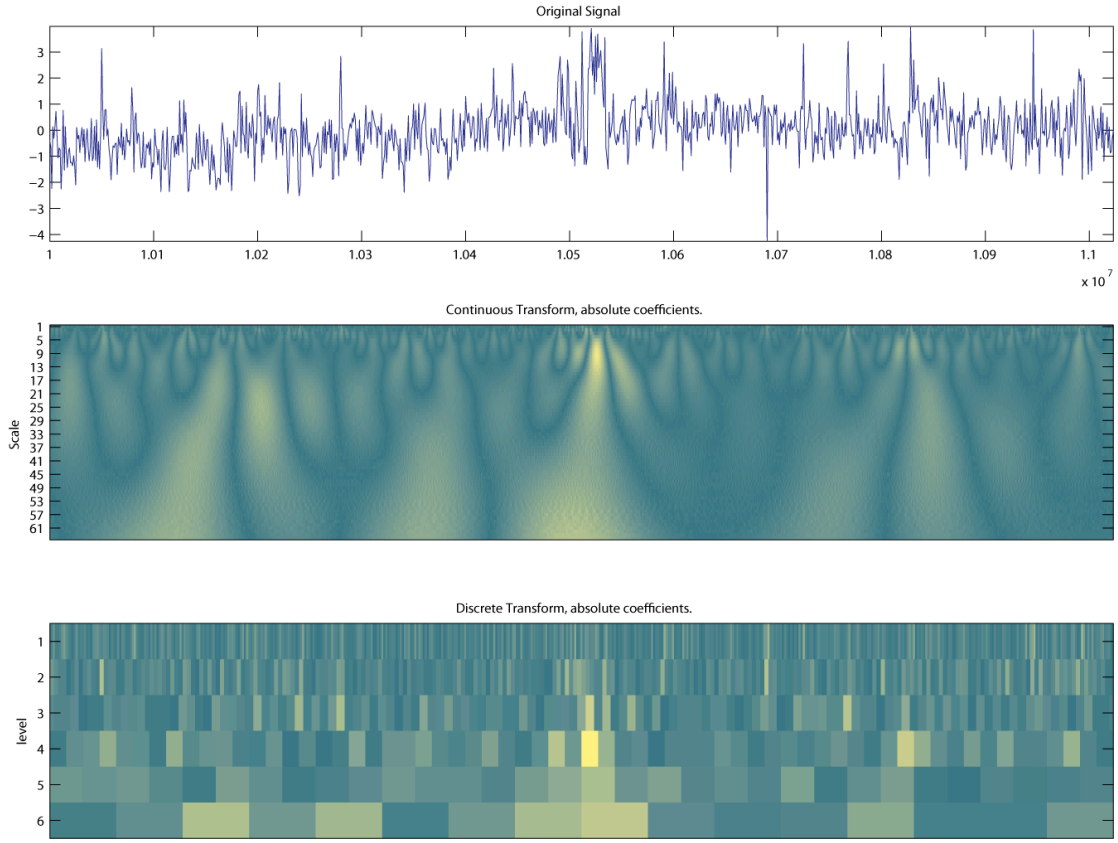


Figure 46. Example decomposition of a signal. The top plot shows the GC content along a 1Mb region of chromosome 6. The signal has been normalised prior to the wavelet decomposition. The central plot shows the CWT of this signal, and the lower plot shows the DWT. Lighter colours indicate larger absolute wavelet coefficients.

When presented with a CWT plot as shown in Figure 46, the CWT is essentially being used as a visual data analysis tool. Visual inspection of such plots reveals regions of interest. However, since the CWT converts a one-dimensional signal into a two-dimensional image, it seems clear that there is a large amount of redundancy. While there is a lot of information at the lower scales, at higher scales the signal varies at a much lower rate. We may therefore decide to retain much of the information at lower scales, but keep less information as the scales increase.

The DWT can be considered as an attempt to preserve the important features captured by the CWT in a more efficient manner. Instead of considering all scales,

only scales of a power of two are considered (that is $\lambda = 2^{j-1}, j = 1,2,3,...$ where j is known as the 'level'). This is best illustrated with an example.

Let the original data vector of length N be $X = (x_1, x_2, x_3, \ldots x_N)$, where x_i have been sampled at regular intervals and N is an even number. We obtain the first level of Haar wavelet DWT coefficients by taking the difference between successive values in the data vector:

$$d_i^{(1)}(x) = \frac{x_{2i-1} - x_{2i}}{\sqrt{2}}, \tag{5.7}$$

where $i = 1,2,3,...,N/2$. These coefficients are known as the detail coefficients at the first level. As these coefficients are calculated using only adjoining data points, they contain information on the finest scale of variation. To obtain the original data vector from the vector of detail coefficients would require knowledge of the mean of the two data points from which the difference was taken. This knowledge is stored in what are known as 'approximation', or 'smooth', coefficients. The first level of approximation coefficients is given by:

$$s_i^{(1)}(x) = \frac{x_{2i-1} + x_{2i}}{\sqrt{2}}. \tag{5.8}$$

Note that in both the detail and the smooth coefficients, the division by the square root of 2 ensures that the conditions (5.1) and (5.2) are satisfied.

At this point, we have converted the original data vector of length N into two vectors of length $N/2$. To obtain the detail coefficients at the next level, the process describe above is repeated *using the approximation coefficients as the data vector*. At each level, the number of coefficients in each vector is halved, and the process can be repeated until single approximation and detail coefficients are obtained. At each level, the data is smoothed to half the resolution of the previous level, and hence the nth level contains information about variation at the 2^n scale. Note that the original signal

can be completely reconstructed from the complete set of detail coefficients and the final set of approximation coefficients.

The recursive algorithm described above outlines the pyramid algorithm (MALLET 1989). We have used the Haar wavelet to demonstrate the algorithm. However, the reader should be aware that any wavelet could be used by replacing the coefficient calculations.

Figure 47 shows an example of decomposing a signal by applying a DWT using the Haar wavelet. Note that large features persist in the approximation coefficients, but smaller details are gradually removed.

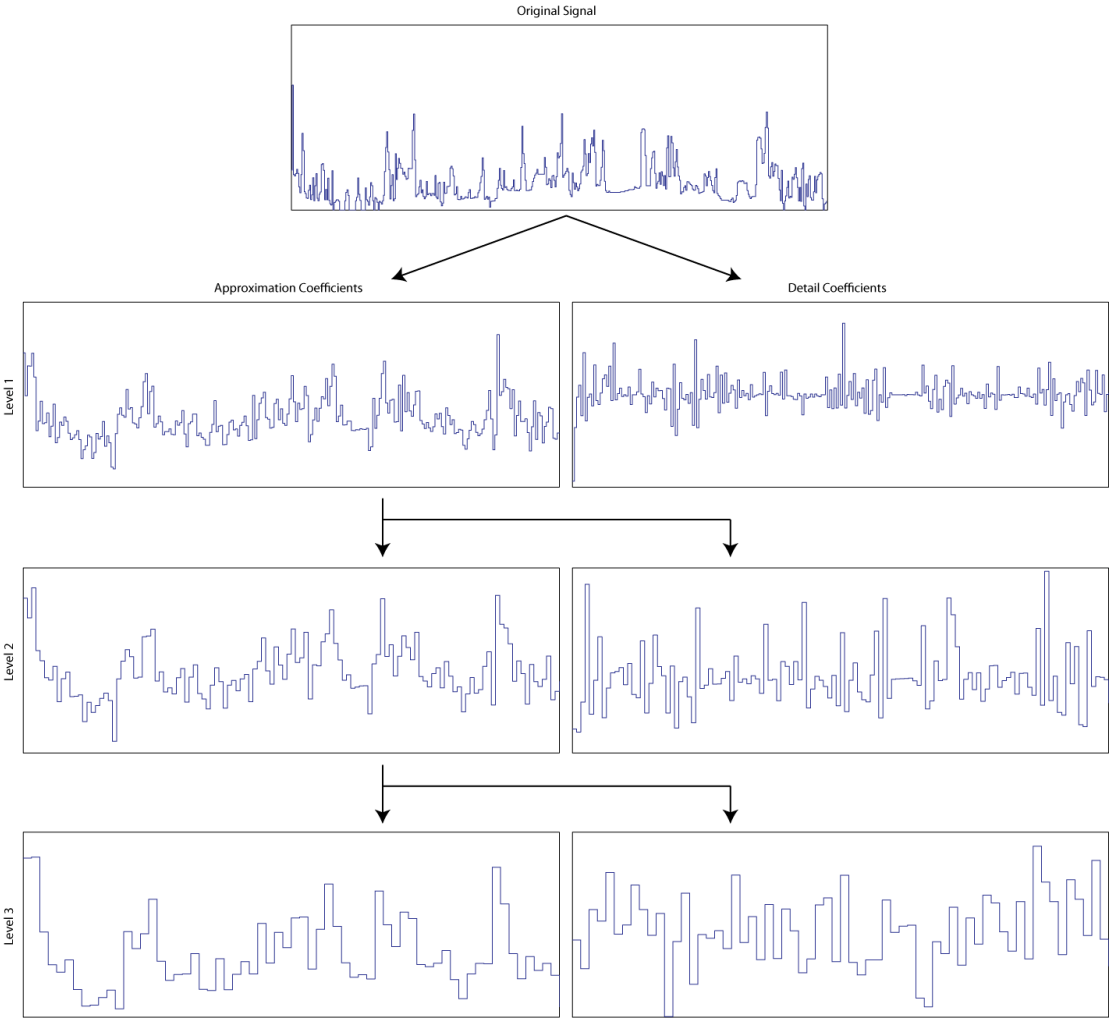


Figure 47. Level by level example of the DWT using the Haar wavelet. The original signal (top), is decomposed into a set of approximation coefficients (left), and detail coefficients (right). At each successive level, the approximation coefficients are further decomposed with more detail being removed each time.

If the original signal has $N = 2^j$ data-points, where j is an integer, then the pyramid algorithm can continue until a single detail and approximation coefficient are obtained. Given such a decomposition, it is possible to analyse the signal on a scale-by-scale basis. For example, given n levels of coefficients with N_i coefficients at the i^{th} level, it is possible to write the variance of the original signal, x, in terms of the detail coefficients (PERCIVAL and WALDEN 2000):

$$\mathrm{var}(x) = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left(d_j^{(i)}\right)^2 \tag{5.9}$$

Note that the above allows us to write down the proportion of variance contributed at each scale. We call this the Wavelet Power Spectrum:

$$Pw(k) = \frac{\sum_{j=1}^{N_k}\left(d_j^{(k)}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{N_i}\left(d_j^{(i)}\right)^2}. \tag{5.10}$$

Although the transformed signal contains no more information than the original, there are several benefits to analysing wavelet-transformed data. Firstly, analysis of the signal at multiple scales removes the need for an arbitrary choice of window size. Second, as detail coefficients at one scale are orthogonal to those at a different scale, correlations observed at one scale cannot be attributed to variation at other scales. If two sets of detail coefficients are correlated, then this can be interpreted as a *change* in the first signal being correlated to a *change* in the other signal.

However, there are a number of disadvantages to using the DWT of which the reader should be aware. Perhaps the most serious issue arises if the number of data-points in the original signal is not 2^j. In this case, the pyramid algorithm will have to terminate prematurely as the number of approximation coefficients will be non-even at some level, and hence the detail coefficients of the next level cannot be calculated. It is therefore common practice to pad the data so that it contains 2^j points. For example, the data may be padded with zeros, the signal mean, or some other set of values (such as the reflection of the signal). However, all such methods suffer from

'edge effects' in which the coefficients representing regions near the edge of the signal become increasing affected.

Another issue of the DWT is that it is dependent on the starting point of the signal. If we were to perform a unitary circular shift of the signal being analysed, we would obtain a different set of DWT coefficients and power spectra. In an attempt to overcome this issue Percival and Walden suggest using a modified version of the DWT known as the Maximal Overlap DWT (or MODWT; PERCIVAL and WALDEN 2000). However, the disadvantage of this method is that the orthogonality of detail coefficients from separate scales is lost.

Therefore, for the purposes of the analysis presented in this chapter, I will be using the simple DWT. I account for edge effects by excluding coefficients that may be affected. I also checked that my results are robust to changes in the signal start point and are consistent between chromosomes.

Wavelets Applied to Recombination

Having introduced wavelet analysis, I now apply such an analysis to our feature of choice: recombination rates. For the purpose of this analysis, I will be working with the recombination rates estimated from the Phase II HapMap data using *rhomap* as described in the previous chapter. Figure 48 shows the decomposition of the recombination rate along a 16Mb region of on the p-arm of chromosome 2. For clarity, the figure shows the decomposition of the logarithm (base 10) of the recombination rate. As all of the detail coefficients are shown on the same scale, one can see that the detail coefficient variance from the finest scales is less than that of broader scales. To quantify this, I use equation (5.10) to describe the proportion of

variance contributed at each scale (i.e. the wavelet power spectrum). I considered the recombination rates along a 65Mb gapless region of chromosome 2 (that also contains the region shown in Figure 48). The recombination rate signal was decomposed using the DWT with the Haar wavelet. I repeated the analysis using both the original recombination rates and the logarithm of the recombination rates. The resulting power spectra are shown in Figure 49.

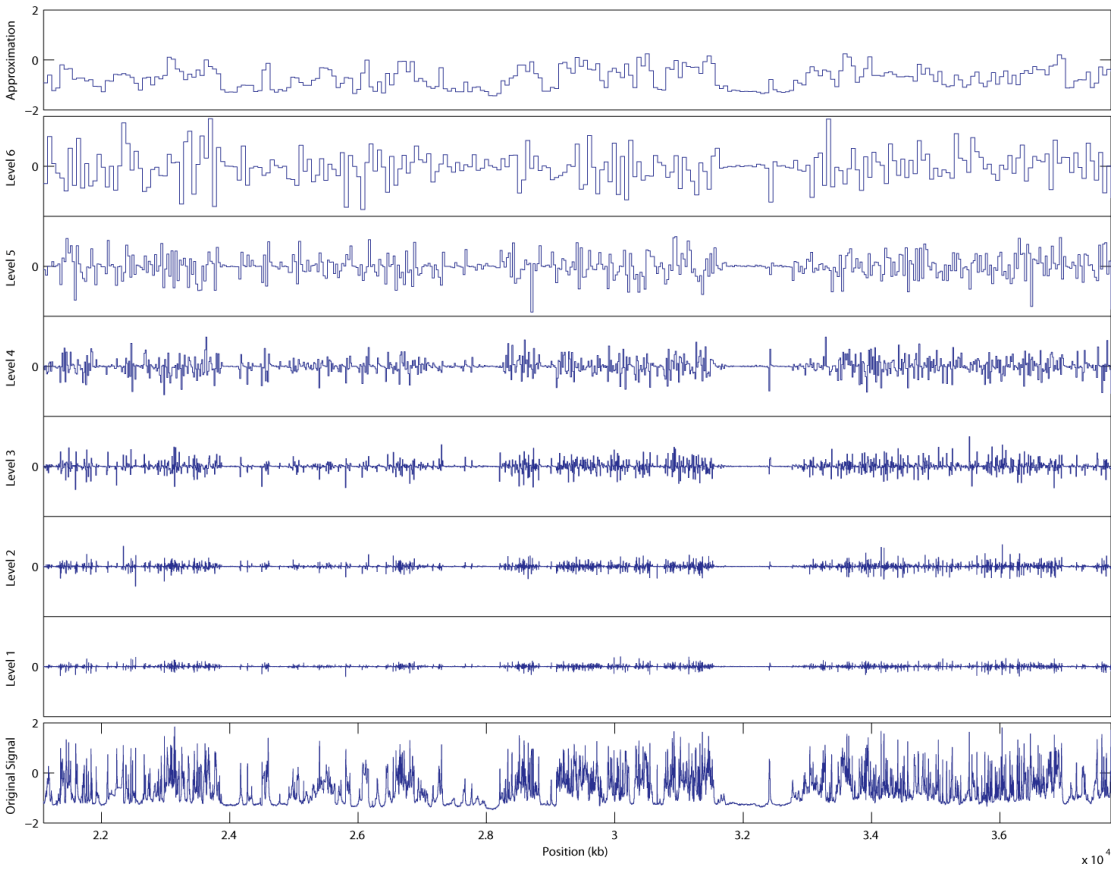


Figure 48. Multi-resolution Analysis of recombination rates along a 16Mb region of chromosome 2. The bottom plot shows the logarithm of the recombination rate. The six plots above show the detail coefficients from the first six decomposition levels, with the first level nearest the bottom of the figure. All detail coefficient plots are shown on the same scale (from -5 to +5 units). The top plot shows the remaining approximation coefficients at the sixth level.

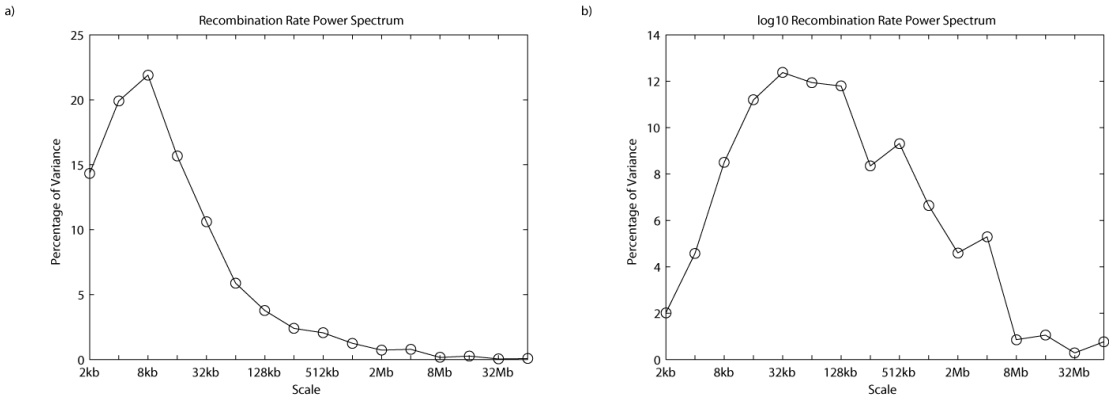


Figure 49. The Wavelet Power Spectrum for the recombination rates (a) and $\log_{10}$ recombination (b) as estimated by *rhomap* along a contiguous 65Mb region of chromosome 2.

If we consider the raw (i.e. unlogged) recombination rate power spectrum first (Figure 49a), we see that the majority of the greatest contributions to the variance come from scales below 32kb. Given the abundance of hotspots this is perhaps to be expected. If we consider the logarithm of the recombination rate (Figure 49b), we see that, as expected from Figure 48, there are significant contributions to the signal variance from a wide range of scales, with the strongest contribution between 16kb and 128kb, although there are significant contributions up to 4Mb. For the remainder of this analysis, I will in the main be working with the logarithm of the recombination rate, as I have found that this gives clearer results in the later sections of this chapter. The reader should therefore assume that I am working with the logarithm of the recombination rate unless otherwise stated.

Wavelets as a Tool for Decomposing Correlation Contributions at Differing Scales

Where multiple annotations are available, each may be decomposed independently and it is then possible to use the detail coefficients to determine correlations between signals at each scale (KEITT and URBAN 2005). I have therefore used this form of wavelet analysis to assess the relationship between recombination rate and a number of other annotations at scales from 2kb to 1024kb on all 22 human autosomes. This was achieved by fitting a linear model to the detail coefficients from a number of annotations at each scale. To avoid issues with edge effects, any coefficients that are potentially affected by such effects (including those from gaps in the data) are discarded prior to the fitting of the linear model (see Figure 50 for an example). If $d^{(j)}(RR)$ is the set of remaining detail coefficients of the recombination rate annotation at level *j*, and $d^{(j)}(A_i)$ is the set of remaining detail coefficients of annotation *i* at level *j*, then the fitted model is described by equation (5.11).

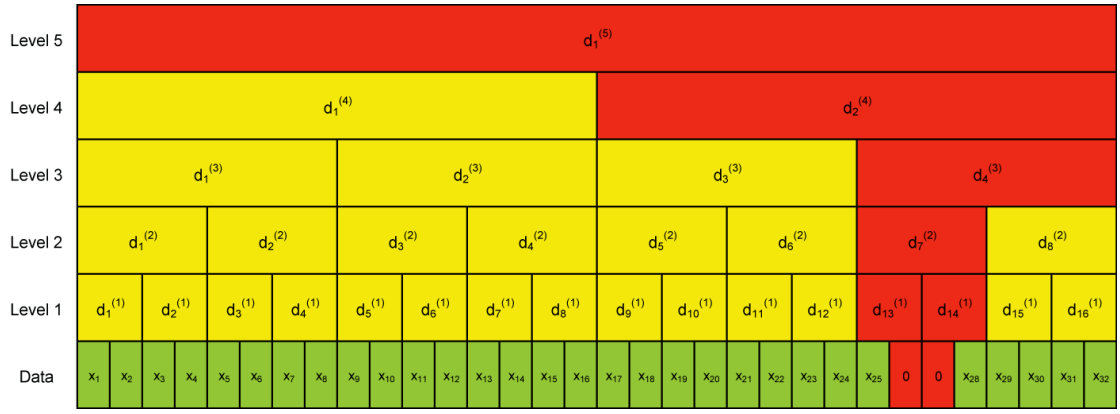


Figure 50. A simple example of wavelet coefficients affected by gaps in the data. In this simple example, the original data is shown in green and has a gap between x_{25} and x_{28} (shown in red). The decomposition would be performed for the whole region by padding the gap with an arbitrary value. The unaffected detail coefficients at each level are shown in yellow, and the potentially affected coefficients are shown in red. All coefficients shown in red would be discarded prior to the fitting of the linear model.

$$d^{(j)}(RR) = \sum_i \beta_i d^{(j)}(A_i) \tag{5.11}$$

In the above equation, β_i gives the regression coefficients for annotation *i*.

Using such a linear model, we are also able to separate the total explained variance into contributions from each scale. It can be shown (SPENCER *et al.* 2006) that the covariance between two signals, *x* and *y*, of length *N* can be written as:

$$\begin{aligned} Cov(x,y) &= \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)\left(\frac{1}{N}\sum_{i=1}^{N} y_i\right) \\ &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{N_i} d_j^{(i)}(x) d_j^{(i)}(y) \end{aligned} \tag{5.12}$$

That is the covariance of the two signals can be written in terms of the sum of the detail coefficient dot product. The correlation between the two signals can therefore be written:

$$Cor(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \sum_{i=1}^{n} \left(\alpha_x^{(i)}\alpha_y^{(i)}\right)^{1/2} \rho\left(d^{(i)}(x), d^{(i)}(y)\right) \tag{5.13}$$

where

$$\alpha_x^{(i)} = \frac{\sum_{j=1}^{N_i}\left(d_j^{(i)}(x)\right)^2}{Var(x)} \text{ and } \alpha_y^{(i)} = \frac{\sum_{j=1}^{N_i}\left(d_j^{(i)}(y)\right)^2}{Var(y)} \tag{5.14}$$

and

$$\rho\left(d^{(i)}(x), d^{(i)}(y)\right) = \frac{\sum_{j=1}^{N_i} d_j^{(i)}(x) d_j^{(i)}(y)}{\sqrt{\left(\sum_{j=1}^{N_i}\left(d_j^{(i)}(x)\right)^2\right)\left(\sum_{j=1}^{N_i}\left(d_j^{(i)}(y)\right)^2\right)}} \tag{5.15}$$

Note that the square of equation (5.15) is the proportion of the variance in $d^{(i)}(y)$ explained by a linear model with predictor $d^{(i)}(x)$ and intercept at zero. If both sets of detail coefficients have zero mean, then $\rho\left(d^{(i)}(x), d^{(i)}(y)\right)$ is simply the Pearson correlation coefficient between the two sets. Furthermore, if the detail coefficients at all levels have zero mean, then it is possible to write the correlation coefficient of the original signals as a weighted sum of correlations between the signal detail coefficients at each scale. Given that the coefficient of determination for the regression between the detail coefficients at level *k* is $\rho_k^2\left(d^{(k)}(x), d^{(k)}(y)\right)$, then we can use equation (5.10) to write down the contributions to the variance explained by the linear model on a scale-by-scale basis. The coefficient of determination for the two signals is given by:

$$r^2 = \frac{\text{Sum of Explained Variance at Each Level}}{\text{Total Variance}} = \frac{\sum_{k=1}^{n} \rho_k^2 \left(d^{(k)}(x), d^{(k)}(y)\right)\left(\sum_{j=1}^{N_k} \left(d_j^{(k)}(y)\right)^2\right)}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} \left(d_j^{(i)}(y)\right)^2} \quad . \tag{5.16}$$

The above formula depends on the detail coefficients having zero mean at all levels. In practice, this will never be true. However, for all the genome features considered in this chapter have detail coefficients with means close to zero.

In my analysis, I will be using a multiple regression. As I will be including a large number of predictor variables, I have used an adjusted coefficient of determination that has been corrected for the increased degrees of freedom in a multiple regression (DRAPER and SMITH 1998). For N data-points and K predictor variables, the adjusted coefficient of determination is defined as:

$$R^2 = \frac{(N-1)r^2 - K}{N-K-1}. \tag{5.17}$$

In the following analysis, the response variable is the recombination rate. As I will be including a large number of predictor variables in the regression, it is sensible to attempt to identify the subset of predictors that best explain the response variable. Therefore, I performed a model selection process at each scale. I used a greedy local search algorithm to minimise the Bayesian Information Criterion (BIC: for example, MCQUARRIE and TSAI 1998), allowing the addition or removal of one predictor variable at each step. If RSS is the residual sum of squares of a regression, then the BIC is given by the following formula.

$$BIC = N \ln\left(\frac{RSS}{N}\right) - K \ln(N) \tag{5.18}$$

To guard against local minima the local search was repeated ten times from randomly chosen starting points (i.e. the starting model consisted of a set of randomly chosen predictor variables). Only the model with the lowest discovered BIC was recorded.

I will now outline the predictor variables that were included in the regression. The majority of the genome annotations used in this analysis are from the UCSC database (build hg17; available from http://genome.ucsc.edu/), and those that are not (specifically GC content and motif density) were estimated directly from the DNA sequence (build 35). When selecting annotations, I only included those with genome-wide coverage – a requirement that excluded the majority of epigenetic annotations. I also attempted to restrict the selection to annotations that in some sense describe local properties of the genome. For this reason, annotations describing, say, conservation of sequence between species were excluded. The reason for this is that while correlations between recombination and such annotations would be interesting, they do not offer any direct insight into the causes of recombination rate variation. Finally, I also attempted to exclude annotations that showed correlations with other annotations (with the exception of GC content, which was accounted for separately and will be described later). For example, annotations relating to transcription were excluded due to strong correlation with exon density. For completeness, I describe the various included annotations here.

The 'GC content' annotation is the percentage GC coverage of each 1kb bin. The 'Exons' annotation is the proportion of each 1kb bin covered by exons, as defined by the UCSC knownGene database table (Hsu *et al.* 2006). The '13mer Motif' annotation is the number of CCTCCCTNNCCAC motifs in each bin respectively. The '13mer Motif (one subs)' gives the number of occurrences of motifs within one substitution of the 13mer motif, not counting the actual CCTCCCTNNCCAC motif

itself (in order to avoid problems with collinearity). The 'SegDups' annotation is the proportion of each bin covered by a segmental duplication, as defined by UCSC genomicSuperDups database table (BAILEY *et al.* 2001). The 'THE1A/B/C/D' annotation is the number of THE1 elements in each bin, taking the repeat centre as the repeat location. The 'L1 Family', 'L2 Family', 'Alu Family', 'MIR Family' annotations are the number of occurrences of repeats from each repeat family in each bin, taking the repeat centre as the repeat location. The 'Polypurine' and 'Polypyrimidine' annotations are the number of occurrences in each bin of simple repeats that are solely composed of Guanine and Adenine in the case of polypurine, and Thymine and Cytosine in the case of polypyrimidine. The 'Other repeats' annotation is the number of repeats that do not fall into the above categories contained within each bin. All repeats are taken from the UCSC rmsk database table (SMIT *et al.* 2004). The 'Microsatellites' annotation is the number of microsatellites in each bin, taking the microsatellite centre as the microsatellite location (BENSON 1999). The 'GIS Chip Pet' annotation is the number of binding sites of the p53 and c-Myc transcription factors (as assessed by chromatin immunoprecipitation, ChIP; WEI *et al.* 2006). This ChIP annotation was selected as it is genome-wide, and uncorrelated with exon density (data not shown). The exact location of the antibody binding site is generally uncertain, so the centre of the possible region was used in each case.

Annotations were binned at a 1kb resolution for use in the wavelet analysis. All annotations were subsequently standardised by subtracting the mean value and dividing by the standard deviation.

Recombination Rates Correlate with GC Content over a Wide Range of Scales

The correlation between GC content and recombination rate has been well documented (JENSEN-SEAMAN *et al.* 2004; KONG *et al.* 2002; SPENCER 2006). Primarily this has been seen as a broad-scale association. The wavelet analysis reveals that there are also small but significant correlations at the fine scale, as shown by the wavelet regression of recombination rates along chromosome 1 (Figure 51). Despite the significance of the coefficients, very little of the observed variance is explained at the fine scale (adjusted $R^2 < 0.001$). However, much more of the variance is explained at the broader scale (adjusted $R^2 = 0.3$ for scales above 128kb).

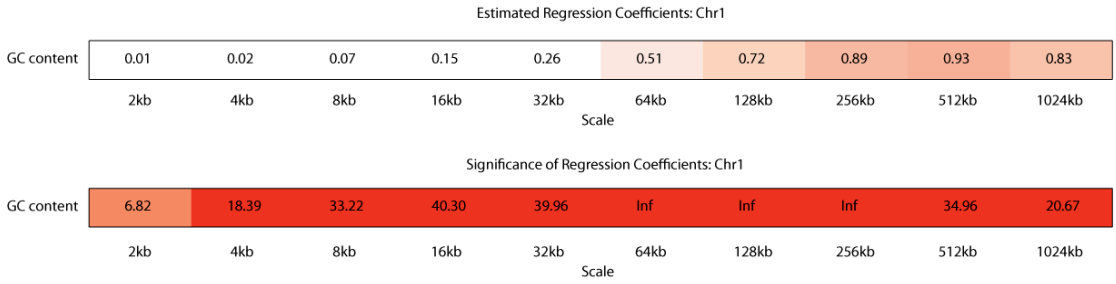


Figure 51. Scale specific regression between recombination and GC content for chromosome 1. The top table shows the estimated regression between the detail coefficients at each scale, with larger values shown in red shades. The lower table shows the significance of the estimated regression coefficients with the numbers indicating the -log p-value (base 10). Values shown as 'Inf' are beyond the precision of the statistical package used (MATLAB). Very similar patterns are observed for the other chromosomes.

Accounting for Correlations with GC Content

It is well known that GC content correlates with a number of genome annotations (VENTER *et al.* 2001). It is therefore desirable to pre-process annotations to remove correlations with GC content. To remove correlations with GC from each annotation, I performed a linear regression at each scale between the GC content detail coefficients and the corresponding annotation coefficients. If the resulting regression was significant (t-test $p < 0.05$) at a given level, I replaced the detail coefficients of the annotation with the residuals of the regression. In other words, before the multiple regression analysis, all annotations included in the recombination rate linear model were corrected for GC content by removing the linear component attributable to GC content (with the obvious exception of GC content itself). Therefore, if an annotation shows a positive correlation with recombination, then this should be interpreted as a correlation over and above that expected from GC (assuming linearity).

The Association between Motif Density and Recombination Rates is Greater than that Expected from GC Content Alone

A suitable check of the method is to include other annotations that are known to influence recombination. The hotspot-associated motif described earlier is one such annotation. I have therefore included two motif annotations, which I refer to as the motif density annotations. These annotations are simply the number of motifs (either the exact 13-mer or those motifs within one substitution) that occur in each 1kb bin.

These annotations were corrected for GC content using the method outlined in the previous section.

As would be expected, the wavelet regression shows that there are significant contributions from the motif at the fine scale, with the most significant contributions coming from the 8kb scale (Figure 52). However, despite the significance of the motif annotations, this regression does not explain much more of the recombination rate variance than GC content alone ($R^2 < 0.001$ at the finest scale, increasing to $R^2 = 0.36$ at the broadest scale for chromosome 1). Furthermore, the significance of the GC content coefficients is largely unaffected by the inclusion of the motif annotations.

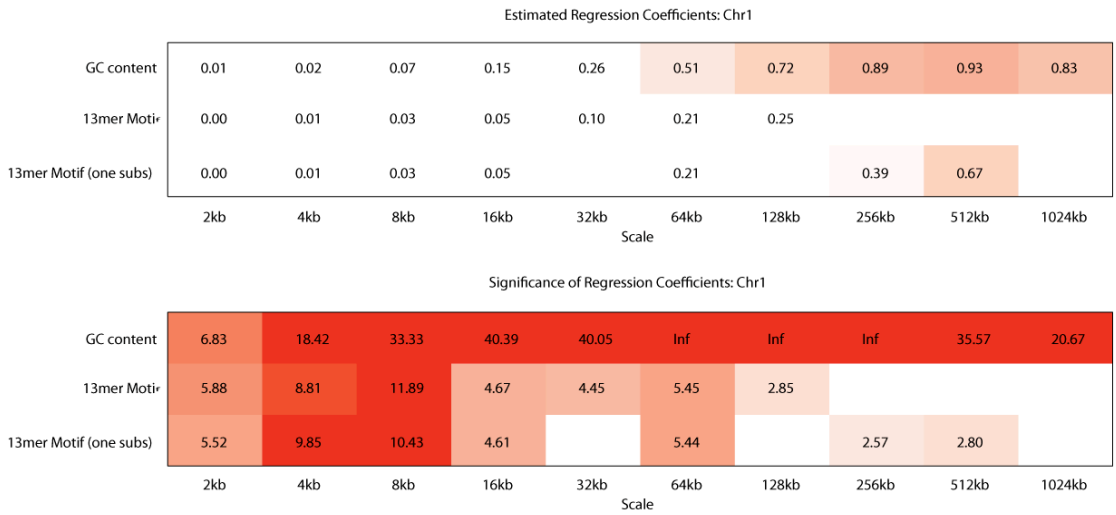


Figure 52. Wavelet coefficient regression of recombination rates with GC content and Motif density for chromosome 1. The top table shows the estimated regression between the detail coefficients at each scale, with larger values shown in red shades. The lower table shows the significance of the estimated regression coefficients with the numbers indicating the -log p-value (base 10). Values shown as 'Inf' are beyond the precision of the statistical package used. Blank cells indicate that the annotation was not included in the linear model at that scale.

Recombination Shows Scale-Specific Correlations with Many Annotations

It is now possible to include a number of other annotations in the linear model. However, as more annotations are included in the regression it becomes difficult to discern clear patterns as many of the relationships are weak and often excluded from the model. While the patterns appear to be largely consistent across chromosomes, there is also a large amount of noise so that the patterns in any one chromosome can be difficult to interpret. I therefore combined results across chromosomes by simply recording if an annotation is included in the linear model at each scale (Figure 53a). Also shown is the linear regression using the approximation coefficients (Figure 53b), which is equivalent to assessing correlations between annotations in windows of increasing size. While the approximation coefficients can reveal interesting patterns, but are harder to interpret, as they are sensitive to confounding factors and correlations between the annotations. Once the results are combined in this manner, it is possible to discern some clear results.

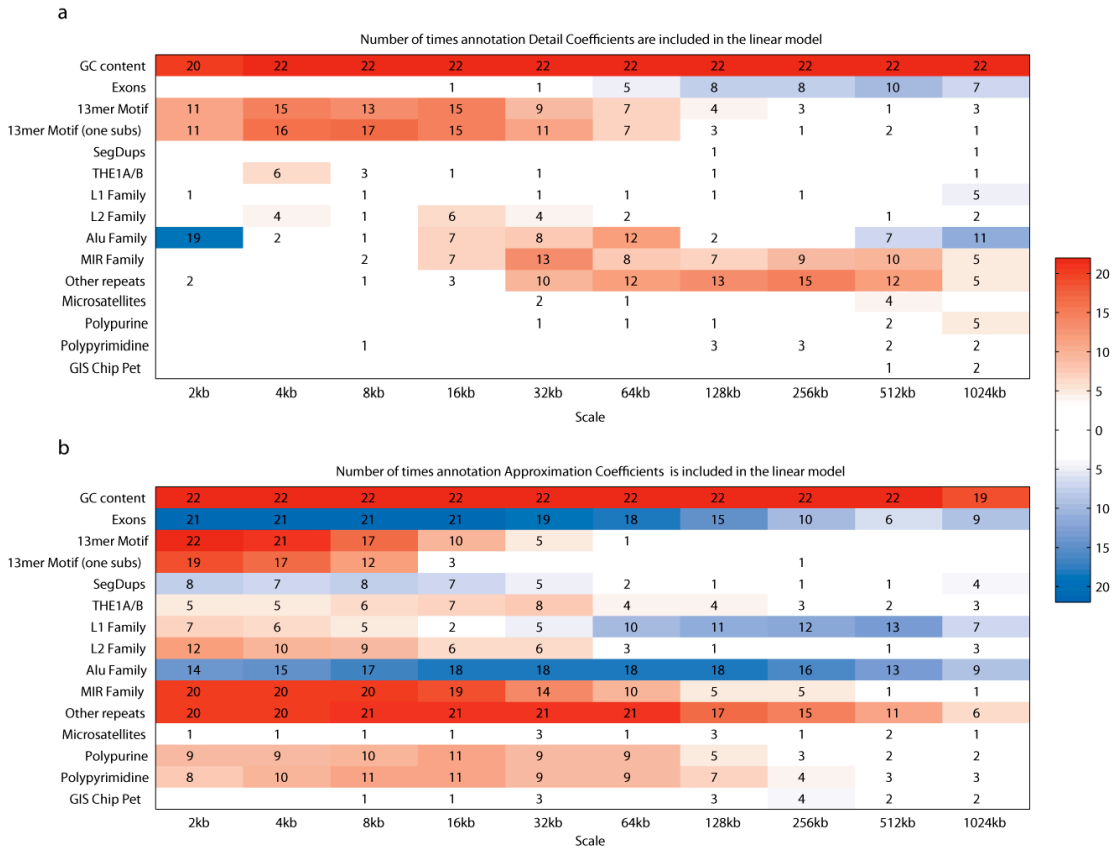


Figure 53. Number of autosomes in which each annotation is included in the linear model at each scale. Numbers indicate the number of autosomes for which the annotation was included in the model at each scale. Red colours indicate a predominately positive relationship with recombination, whereas blue colours indicate a predominately negative relationship. The regression was performed using the detail coefficients (a) and the approximation coefficients (b).

One striking observation from Figure 53 is that the majority of annotations have associations at certain scales with recombination. The most consistent signal remains the positive association between recombination and GC content at all scales. As seen with the chromosome 1 analysis, the hotspot-related motifs are also consistently associated with recombination at the fine scale.

As expected from the results in Chapter 4, exon density suppresses recombination. Furthermore, this effect is visible at relatively broad scales, which is

consistent with the average size of transcribed regions in the human genome being 59kb.

There are many interesting signals associated with the DNA repeat elements, with MIR and 'Other' repeat elements showing a strong positive association at the medium to broad scales, and L2 elements show a weak positive association at the fine to medium scale. The L1 elements do not show associations in the detail coefficients except for a negative correlation at the broad scale, which is consistent with these elements being relatively cold.

The Alu elements show a particularly unusual pattern. At the finest scale, the detail coefficients indicate a negative association, whereas the medium scales indicate a positive association, before returning to a negative association at the broad scale. This is perhaps unexpected as Alu elements do not in general exhibit obvious changes in recombination at the fine scale (Figure 54a). In order to interpret this result, one must first recall that the annotations have been corrected for GC content. Alu elements are locally high in GC content (with some elements being up to 60% GC; JURKA 2004), while being surrounded by regions of low GC (Figure 54b). If we perform a wavelet regression between GC content and Alu density, we therefore observe an unusual pattern (Figure 54c).

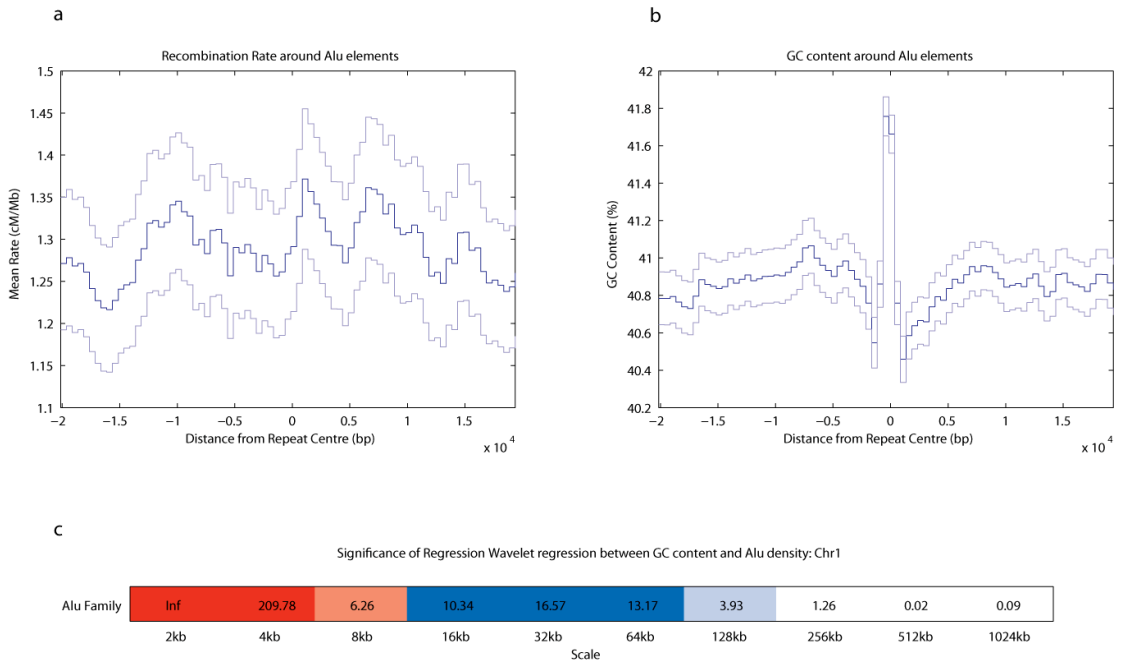


Figure 54. The average recombination rate and GC content around Alu elements. Plots (a) and (b) show a 40kb window, centred on Alu locations. (a) Recombination rates (no logarithm) averaged over 12,000 Alu elements on chromosome 1. (b) GC content around the same elements. Faded colours show the mean ±2 standard errors. The pattern in GC content around Alu elements leads to a complex pattern in the wavelet regression of GC content and Alu density (c). Numbers indicate the significance of the regression ($-\log_{10}$ p-value). Red colours indicate a positive relationship, whereas blue colours indicate a negative relationship.

As we have corrected the detail coefficients in our wavelet regression of recombination for GC content, we should therefore interpret the fine scale result as Alu elements being not as recombinogenic as would be expected from GC content alone. However, at the medium scale, the local decrease in GC content associated with the presence of Alu elements leads to a positive association with recombination. This is a slightly counterintuitive result, but is consistent with the patterns of GC around Alu elements.

Finally, at the broad scale, the correlation between GC and Alu density is reduced. At this scale, it appears that Alu elements cluster together in regions of low recombination, as can be clearly seen on chromosome 22 (Figure 55). Why Alu

elements should exhibit this pattern is unclear, although it may be speculated that the high number of Alu elements in the genome may be a factor. Recombination occurring in regions with many Alu elements may be disadvantageous due to the possibility of non-homologous recombination. Such events have been linked with disease (e.g. SUMINAGA *et al.* 2000) and hence there may be a strong selective advantage in recombination occurring preferentially away from Alu clusters. The Alu elements therefore provide a good example of wavelet analysis revealing patterns in recombination rate variation occurring at differing scales.

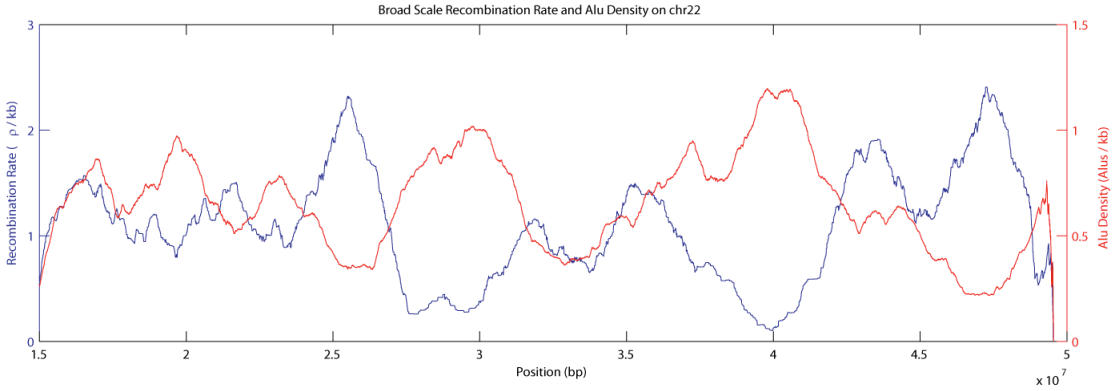


Figure 55. Alu Density and Recombination Rate along the long arm of Chromosome 22. Alu density is shown in red and the recombination rate is shown in blue. Both lines are based on a 1Mb moving average.

The THE1A/B elements show only a weak positive association with recombination. This is worth noting as earlier in the thesis I showed that THE1A/B elements with a motif are highly likely to be within hotspots. It would therefore appear that the contribution to recombination from THE1A/B elements is largely explained by GC content and the motifs.

It is difficult to envisage that the broad-scale patterns associated with DNA repeat elements are causal of recombination (for example, MIR elements appear to promote recombination at scales up to 64kb, despite the average size of these

elements being approximately 140bp in size). I would therefore suggest that this is more indicative of repeats inserting themselves into recombinogenic regions. This is in contrast to the local patterns of recombination observed around repeats, such as the THE1 elements, seen in Chapter 4.

Of the remaining annotations, polypyrimidine (repeats containing predominately TC bases) and polypurine (repeats containing predominately GA bases) do not show associations in the detail coefficients, but do show weak positive associations at the fine to medium scales in the approximation coefficients. Finally, two annotations do not show strong associations with recombination in either set of coefficients. These are Microsatellites and the Genome Institute of Singapore ChIP-PET annotation.

Exploring Interactions between Annotations

As a final analysis of this dataset, I explored the contributions of interactions between the annotations to the recombination rate. In order to explore non-additive interactions between annotations, I generated a number of new annotations that consisted of the dot product of two original annotations. I then repeated the multiple wavelet analysis using both the original annotations and the new interaction annotations.

As the number of possible combinations was large (there were 105 possible combinations) and computational resources limited, it was not practical to include all possible annotation interactions in the regression. To reduce the number of annotations, I firstly removed the two annotations that showed little or no relationship with recombination in the original linear regression (i.e. microsatellites and GIS Chip

Pet). I then performed a non-exhaustive search for interesting interactions, primarily focusing on interactions between the motif and other annotations. This was done in an informal fashion, adding and removing interactions depending on observed correlations with recombination. The full list of interactions that I explored is shown in Table 12.

Annotation 1	Annotation 2
13-mer Motif *	THE1A/B/C/D
13-mer Motif *	L1 Family
13-mer Motif *	L2 Family
13-mer Motif *	Alu Family
13-mer Motif *	MIR Family
13-mer Motif *	GC content
13-mer Motif *	Exons
13-mer Motif *	Polypurine
13-mer Motif *	Polypyrimidine
13-mer Motif (exact)	13-mer Motif (allowing one substitution)
Exons	GC content
GC content	THE1A/B/C/D
GC content	L1 Family
GC content	L2 Family
GC content	Alu Family
GC content	MIR Family

Table 12 - List of explored interactions. Each row in this table gives two annotations that multiplied together to obtain a new annotation for use in the multiple regression. The motif annotations marked with a '*' indicate that both interactions using the exact 13-mer motif, and using motifs within one substitution (but not including the 13-mer itself) were explored.

The majority of interactions did not show significant relationships with recombination. For clarity, I only show interactions that have significant relationships with recombination on four or more chromosomes at any given scale (Figure 56). This

requirement reduced the number of 'interesting' interactions to nine, which can be seen towards the lower half of each subplot.

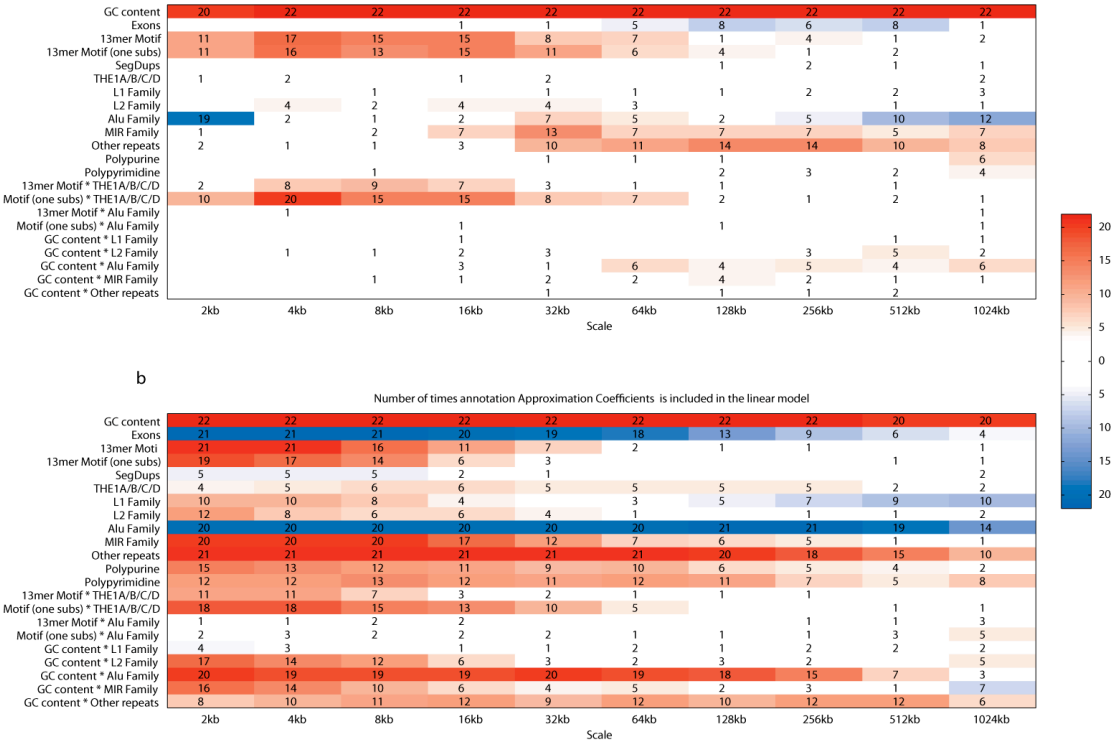


Figure 56. The number of times each annotation was included in the linear model at each scale - including interactions. As previously, numbers indicate the number of autosomes for which the annotation was included in the model at each scale. Red colours indicate a predominately positive relationship with recombination, whereas blue colours indicate a predominately negative relationship. The regression was performed using the detail coefficients (a) and the approximation coefficients (b).

Perhaps unsurprisingly, the interaction showing the strongest relationship with recombination is that of the THE1 elements and the 13-mer motifs. At the 4kb scale, 20 out of 22 autosomes show a significant positive relationship between the detail coefficients – a figure that is only bettered by GC content. This result emphasizes the non-linear relationship between recombination and these annotations. No other repeat element showed strong interactions with the motif.

The strong effect of GC content is again apparent in the interactions, with four of the repeat families showing relationships (at least in the approximation coefficients). However, few significant relationships are visible in the detail coefficients for these interactions. However, the positive association between recombination and Alu elements at the medium scale does seem to have some weak basis in the interaction between these elements and GC content at these scales.

As has been usual in this analysis, while the observed relationships are highly statistically significant, the majority of the variance in recombination is unexplained. Including the interactions, the average R^2 value at the finest scale is below 0.01, rising gradually to ~0.4 at the 1Mb scale.

Linear Model is Unable to explain much of the Variance in Recombination Rates

Despite the strong evidence for correlations between the annotations and the recombination rate, the linear model explains little of the observed variance. The largest contribution to the variance in the recombination rate signal is from the scales below 256kb (Figure 49b). However, it is at these scales that the linear model performs most poorly. Indeed, the adjusted R^2 value, which can be informally interpreted as the proportion of the variance explained by the linear model, is generally less than 0.05 at the fine scale (Figure 57). However, as the scale increases, so does the amount of explained variance. At the megabase scale, between 30 and 50% of the variance is explained by the linear model, the majority of which is

contributed by GC content. This is comparable to the (unadjusted) value of 37% obtained by Kong *et al.* using six predictors at the 3Mb scale (KONG *et al.* 2002).

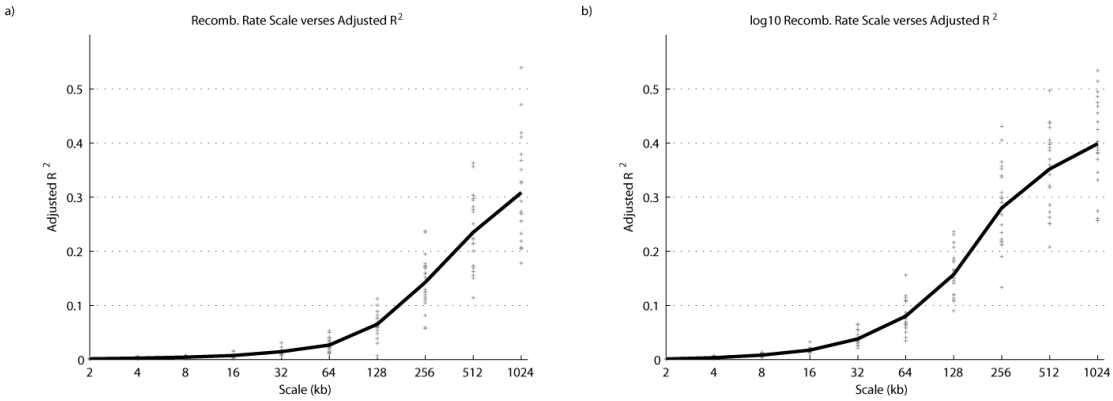


Figure 57. Adjusted R^2 by level for (a) the recombination rate and (b) the logarithm (base 10) of the recombination rate. The adjusted R^2 value can be interpreted as the proportion of variance explained by the linear model. The value computed for each chromosome is shown by grey markers. A weighted average of the R^2 values (weighted by chromosome size) is also shown as a black line.

There are a number of plausible reasons for the poor performance of the linear model at the fine scale. First, there may be an annotation that is associated with the recombination rate which missing from our model. Alternatively, the linear model may itself be inappropriate, as recombination rate variation is strictly non-linear. I have noted in earlier sections of this thesis that recombination shows strong association with motifs and THE1B repeats. If both are present, then a hotspot is expected to exist with high probability. This would appear to be a non-linear mechanism, and may explain why the proportion of variance explained at fine scales is so low. Furthermore, there are many instances of identical motifs appearing in both hot and coldspots, with the factor that determines which are hot and which are cold currently unknown. Finally, the simulation studies in Chapter 3 would suggest that there is a greater deal of uncertainty (and therefore noise) in rate estimates at the fine

scale. All explanations probably contribute. Nevertheless, my results do provide a useful summary of the important relationships between recombination and other genome annotations.

Discussion

In this chapter, I have used a wavelet analysis to investigate the scale-specific features of recombination in the human genome. I have shown that recombination shows complex relationships with a number of annotations, with GC content being the strongest predictor. However, despite the high significance of the relationships, no more than approximately 40% of the variation in recombination rates can be explained at the megabase scale. That so many annotations show scale-specific relationships, and yet explain so little, demonstrates the complexity of the patterns of recombination in the genome. A possible reason of the low proportion of explained variance is that the linear model is incapable of capturing the features of non-linear relationships between the annotations and recombination rate. Despite this inability of the linear model to explain much of the variance, I believe the results are useful in summarising our knowledge of recombination.

This chapter concludes the analysis portion of the thesis. In the next and final chapter of this thesis, I provide a summary of my findings. I also outline some possibilities for future research.

Chapter 6 Conclusion

In this thesis, I have attempted to describe patterns of recombination rate variation in the human genome. In this final chapter, I discuss both the successes and limitations contained within the previous chapters. I also take this opportunity to speculate as to the significance of these findings and suggest areas for future research.

To investigate patterns of recombination in the human genome, I have developed a new method for the estimation of recombination rates from population surveys of genetic data. The new method incorporated a prior model of recombination hotspots, which was lacking from previous rate estimation schemes. As the calculation of the likelihood of population genetic data is computationally prohibitive under the full coalescent model, I used a composite likelihood based on the product of precalculated likelihoods for pairs of SNPs. The composite likelihood can be quickly evaluated for large datasets, and it was therefore possible to use a rjMCMC procedure to explore the pseudo-posterior distribution of rate estimates that varied along the sequence. The new method has been implemented in the C++ program *rhomap*, which is available for public download (http://www.stats.ox.ac.uk/~mcvean/LDhat/; AUTON and MCVEAN 2007).

The simulation studies undertaken in Chapter 3 would suggest that the estimates obtained from the new method compare favourably to those obtained from the commonly used method, *LDhat*. This was further confirmed by comparison with the rate estimates from sperm typing in the MHC and MS32 regions. The new method provides rate estimates with similar accuracy to *LDhat* at the broad scale, but considerably less variance at the fine scale.

As my prior model included a model of recombination hotspots, I investigated the possibility of using the new method as a hotspot detection tool. To do this, I used a simple statistic to describe the average number of hotspots per kilobase in each sample from the rjMCMC. The statistic was calculated between each SNP interval, and I called this statistic the posterior hotspot density. I applied a threshold to this statistic to determine the location of hotspots, with a suitable threshold level being determined via simulation. Despite the crude nature of this method, I found that it is possible to achieve a hotspot detection power of approximately 50% with a low false positive rate. However, the power to detect hotspots was low compared to existing hotspot detection methods, especially when the SNPs in the data were randomly thinned. Due to the lack of power, I suggested that the new method should not be used solely for hotspot detection.

The new method was based on a composite likelihood, which is both a strength and a weakness; a strength because it allows recombination rate estimates to be obtained on a genome-wide scale, but a weakness as the composite likelihood is not in any sense a 'real' likelihood. The composite likelihood is sharply peaked in comparison to likelihoods estimated from the full coalescent model. I have attempted to correct this 'over-peaked' nature of the composite likelihood using a simple transformation that leaves the maximum composite likelihood estimate unchanged. This change was made simply for practical reasons, as efficient use of MCMC was problematic with the original composite likelihood due to slow mixing times. However, as with the original composite likelihood, one cannot directly use the corrected composite likelihood to obtain estimates of uncertainty.

To compensate for the issues introduced by the composite likelihood, I have used a number of *ad hoc* procedures to achieve results that would have been much

easier in a true Bayesian framework. The most obvious example of this is the use of *rhomap* as a hotspot detection tool. With a 'true' likelihood, one may have hoped that the posterior hotspot density would largely reflect the probability of a hotspot in any given location. However, the same cannot be said when using a composite likelihood. Furthermore, thresholding the posterior hotspot density statistic is a quite unsatisfactory decision making procedure, as the chosen threshold is largely arbitrary. While my simulations suggested a suitable threshold, the suitability of this threshold is likely to be influenced by a number of factors including SNP density, population demographics and SNP ascertainment to name just a few. It is therefore perhaps more sensible to simply report the value of the posterior hotspot density along the region being analysed, and allow the user to make a subjective judgement as to the location of hotspots. Future work in this area should investigate alternative methods by which the evidence of hotspots can be assessed. Suitable schemes will almost certainly be heavily reliant on extensive simulations, and hence may suffer in terms of execution time.

Nevertheless, *rhomap* can be used to obtain recombination rate estimates on a genome-wide scale and thereby provide a number of insights into recombination in the human genome. I therefore used the method to obtain estimates for the majority of the human genome using data from Phase II of the HapMap project (THE INTERNATIONAL HAPMAP CONSORTIUM 2007). The accuracy of the rate estimates at the broad scale was demonstrated via the excellent correlation with those obtained from the deCODE pedigree study (KONG *et al.* 2002).

At the fine scale, I found that distribution of recombination is highly non-uniform, with the majority of recombination occurring within recombination hotspots. In total, approximately 90% of recombination occurs within 30% of genome

sequence. This should perhaps be considered a lower bound, as my simulation studies suggest that the background rates are slightly overestimated by *rhomap*, while the peak rate of hotspots are underestimated. It is therefore quite plausible that a greater proportion of recombination occurs in a smaller fraction of sequence.

To give the reader an example of the estimates obtained from the HapMap data, I returned to the MHC and MS32 regions, which have been visited a number of times in this thesis. An intriguing result was found in the MS32 region. In a separate study, the MS32 hotspot was found to be very strong in sperm analysis but weak in estimates obtained from population studies (JEFFREYS *et al.* 2005). In Chapter 3, I was able to confirm this result using the original dataset from that study – the MS32 hotspot, although detectable, appeared to be very weak in the *rhomap* estimates. The original study suggested that the disparity between the sperm estimates and those obtained from population genetic studies was indicative of a newly emerged hotspot, and it was speculated that recombination hotspots are therefore transient features of the genome (JEFFREYS *et al.* 2005). However, the *rhomap* estimates of the region obtained from the Phase II HapMap show a relatively large increase in recombination rate in the same vicinity, but only in the African population. If this peak is the same MS32 hotspot, then it indicates that the emergence of the hotspot actually predates the divergence of the three human populations.

If hotspots differ between populations, and sequence features (specifically motifs) determine the locations of hotspots, then it may be possible to identify alleles associated with hotspots. There is a single motif within one substitution of the 13-mer consensus contained within the MS32 hotspot. It would be of great interest to learn if this motif contains a SNP with significantly different frequencies between the populations. Unfortunately, no such SNP currently exists within the HapMap.

Leaving the MS32 hotspot, I then used the HapMap data to investigate the relationship between recombination and various other genome features. Starting with genes, my results show that recombination is generally suppressed within genic regions relative to the genome average, with small peaks in recombination rate just beyond the transcribed regions. The peaks are at least partially reflected in the patterns of GC content and motif density. However, there is no corresponding dip in these annotations within the transcribed region. This would suggest that recombination has been suppressed in genic regions due to the inherent damage associated with the process. Alternatively, the presence of selection in these regions may bias rate estimates, although other studies have shown that composite likelihood methods are largely robust in the presence of selection (McVean 2007; Spencer 2006).

Despite recombination being generally suppressed within genes, there is a large degree of heterogeneity between gene ontology groups. Strikingly, genes expressed in the outer areas of the cell show significantly higher rates of recombination than genes expressed in the nuclear areas. This is an interesting result, as one may speculate that recombination has been used as a means of generating diversity in genes that experience selection pressures that vary over time. It may be that there is a selective advantage for, say, immunity genes experiencing relatively high rates of recombination as a diverse population has a better chance of combating the emergence of a new pathogen. Conversely, it is plausible that genes such as Chaperones may not experience rapidly changing environments, and hence be under strong purifying selection as the DNA damage associated recombination events would be a selective disadvantage. Although this is speculative, it will be very interesting to

learn if similar patterns are observed once fine-scale genetic maps become available in other species.

I also investigated patterns of recombination in various types of repeat DNA. I showed that certain repeats have significantly higher recombination rates than others. Some repeats types (such as THE1 and L2 elements) show very local increases in recombination, possibly suggesting that they are indeed causal of the elevated rate. Perhaps most interesting of all is that the recombination rate of a specific repeat family (namely the THE1 family) appears be controlled by the existence of a hotspot-associated motif.

The patterns of recombination between repeats could easily be investigated further. For example, the evolutionary history of Alu elements is relatively well known, with a number of subfamilies having been identified (JURKA 2004) and a large number of copies exist in the human genome. An interesting project would be to identify any common patterns of recombination in elements of each family. It would then be possible to investigate if 'old' elements show different patterns of recombination from 'new' elements.

As the majority of hotspot-associated motifs so far identified show some homology (MYERS *et al.* 2007), I attempted to isolate the common features of these motifs. To do this, I employed a Genetic Algorithm that searched for sequence features that differentiated between a set of hotspots and a matched set of coldspots. The algorithm was allowed to include degenerate bases in the motif, although favoured motifs with lower degeneracy. The algorithm identified a single 13-base degenerate motif that is consistent with the majority of hotspot-associated motifs so far identified. While the motif identified by the algorithm had a high degree of degeneracy, it appears the important aspect of the motif seems to be the relative

spacing of the Cytosine bases. Notably, the algorithm was unable to identify any sequence features up or downstream of the motif that are associated with hotspots. Furthermore, the removal sequences containing the degenerate motif did not reveal any secondary motifs suggesting that there are no other sequence motifs beyond those already identified.

The hotspot motifs present a number of interesting questions. If we accept that the motif is in some sense 'causal' of a recombination hotspot (and the evidence would seem to at least partially support this; MYERS *et al.* 2005), then they provide excellent targets for the investigation of the evolution of recombination hotspots. While we do not have genome-wide fine-scale recombination rate estimates from species other than humans, it has been hypothesised that hotspots are transient features of the genome (COOP and MYERS 2007; JEFFREYS *et al.* 2005). Furthermore, it is now generally accepted that recombination rates between humans and chimps vary substantially at the fine scale (PTAK *et al.* 2005; WALL *et al.* 2003; WINCKLER *et al.* 2005). This is perhaps surprising given the high degree of sequence identity between the two species, and becomes even more surprising in light of the evidence of hotspots having at least some sequence dependency. A reasonable explanation for this is that the hotspot-related motif identified in humans is not active in chimps. If the motif in chimp where different from that in human, then one would expect hotspots in chimps to occur in largely different locations relative to those found in humans. Following this line of reasoning further, if the motif is inactive in chimp then one may expect to observe different substitution rates for the motifs in the two species. The reason for this is that crossover asymmetry can cause recombination-suppressing alleles to be transmitted more often than recombination-promoting alleles in a process known as meiotic drive (JEFFREYS and NEUMANN 2002). Computer simulations and

theoretical work have suggested that this can lead to the eventual fixation of the recombination-suppressing allele (BOULTON *et al.* 1997; COOP and MYERS 2007; JEFFREYS and NEUMANN 2002). The hypothesis that substitution patterns in the motif will be different in the two species is testable, and should be investigated in the near future.

However, despite the evidence for an unambiguous relationship between hotspots and the sequence motifs, the motifs are not in themselves good predictors of recombination hotspots, with many of the motifs being found in coldspots. Furthermore, neither my nor other analyses (MYERS *et al.* 2007) have been able to identify further sequence features which control the activity of the motifs. One may be tempted to conclude that the activity is controlled by something other than sequence features. I therefore attempted to isolate an epigenetic factor by which the activity of the motif could be determined. While my investigation was quite exploratory, I did consider a wide range of epigenetic annotations from the ENCODE dataset (THE ENCODE PROJECT CONSORTIUM 2007), which covered many of the features of the genome that cannot be assessed from sequence alone. However, I found no single epigenetic factor that could distinguish between active and inactive motifs. While this analysis does not rule out the possibility of epigenetic factors being involved in motif activity, it does deepen the mystery.

In the final analysis of this thesis, I studied scale-specific associations between recombination and other genome features. To do this, I adopted a Wavelet Analysis that allowed the annotation signals to be broken down into independent contributions from differing scales. By fitting a linear model to the decompositions of the various annotations, I found that there are a number of significant correlations with recombination. The strongest and most consistent correlation was with GC content.

While such correlations had been noted before at the broad scale, the observation of correlations at the fine scale is novel. This perhaps suggests an intrinsic relationship between GC content and recombination. However, the nature of this relationship is unknown.

A benefit of this method is that it allows the contribution from a number of genome annotations to the recombination signal to be simultaneously assessed. In doing so, it was possible to identify a number of scale-specific correlations. Notably, DNA repeats show different relationships at differing scales. The fact that some of these correlations extend to quite broad scales may suggest that the repeats are not causal of recombination, but preferentially locate themselves in broad regions of high recombination.

The weakness of the wavelet analysis is that, despite the regression achieving very high significance levels, only a small proportion of the variance in the recombination rate signal could be explained. Possibly this is due to the inadequacy of the linear model, and a number of improvements could be made here. Nonetheless, the method does provide a useful method by which the relationship between recombination and a number of annotations can be summarised.

While the work described in this thesis has revealed many interesting features of recombination, a number of questions remain unanswered. Perhaps the single most important question relates to the cause of recombination hotspots. While there is evidence that many hotspots contain a hotspot-associated motif, it remains unclear what other factors are required for a hotspot to occur. Nevertheless, the motif provides an excellent opportunity to investigate this issue further.

In conclusion, the recent advent of genome-wide genetic polymorphism data has provided a great amount of insight into questions surrounding recombination in

the human genome. This thesis has demonstrated that recombination shows a number of complex relationships with other genome features, many of which have yet to be explained. As further data becomes available, we are presented with a valuable opportunity to understand the nature of recombination.

References

AUTON, A., and G. MCVEAN, 2007 Recombination rate estimation in the presence of hotspots. Genome Res**:** In Press.

AWADALLA, P., 2003 The evolutionary genomics of pathogen recombination. Nat Rev Genet **4:** 50-60.

BAILEY, J. A., A. M. YAVOR, H. F. MASSA, B. J. TRASK and E. E. EICHLER, 2001 Segmental duplications: organization and impact within the current human genome project assembly. Genome Res **11:** 1005-1017.

BALASUBRAMANIAN, B., W. K. POGOZELSKI and T. D. TULLIUS, 1998 DNA strand breaking by the hydroxyl radical is governed by the accessible surface areas of the hydrogen atoms of the DNA backbone. Proc Natl Acad Sci U S A **95:** 9738-9743.

BARNES, T. M., Y. KOHARA, A. COULSON and S. HEKIMI, 1995 Meiotic recombination, noncoding DNA and genomic organization in Caenorhabditis elegans. Genetics **141:** 159-179.

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

BHINGE, A. A., J. KIM, G. M. EUSKIRCHEN, M. SNYDER and V. R. IYER, 2007 Mapping the chromosomal targets of STAT1 by Sequence Tag Analysis of Genomic Enrichment (STAGE). Genome Res **17:** 910-916.

BIEDA, M., X. XU, M. A. SINGER, R. GREEN and P. J. FARNHAM, 2006 Unbiased location analysis of E2F1-binding sites suggests a widespread role for E2F1 in the human genome. Genome Res **16:** 595-605.

BOULTON, A., R. S. MYERS and R. J. REDFIELD, 1997 The hotspot conversion paradox and the evolution of meiotic recombination. Proc Natl Acad Sci U S A **94:** 8058-8063.

CAWLEY, S., S. BEKIRANOV, H. H. NG, P. KAPRANOV, E. A. SEKINGER *et al.*, 2004 Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell **116:** 499-509.

CHENG, J., P. KAPRANOV, J. DRENKOW, S. DIKE, S. BRUBAKER *et al.*, 2005 Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science **308:** 1149-1154.

COOP, G., and S. R. MYERS, 2007 Live hot, die young: transmission distortion in recombination hotspots. PLoS Genet **3:** e35.

CRAWFORD, D. C., T. BHANGALE, N. LI, G. HELLENTHAL, M. J. RIEDER *et al.*, 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat Genet **36:** 700-706.

CRAWFORD, G. E., S. DAVIS, P. C. SCACHERI, G. RENAUD, M. J. HALAWI *et al.*, 2006 DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods **3:** 503-509.

CROSS, S. H., and A. P. BIRD, 1995 CpG islands and genes. Curr Opin Genet Dev **5:** 309-314.

DAUBECHIES, I., SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS., CONFERENCE BOARD OF THE MATHEMATICAL SCIENCES. and NATIONAL

SCIENCE FOUNDATION (U.S.), 1992 *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia.
DEVLIN, B., and N. RISCH, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311-322.
DRAPER, N. R., and H. SMITH, 1998 *Applied regression analysis*. John Wiley, New York; Chichester.
FEARNHEAD, P., 2003 Consistency of estimators of the population-scaled recombination rate. Theoretical Population Biology **64:** 67-79.
FEARNHEAD, P., 2006 SequenceLDhot: detecting recombination hotspots. Bioinformatics **22:** 3061-3066.
FEARNHEAD, P., and P. DONNELLY, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299-1318.
FEARNHEAD, P., and P. DONNELLY, 2002 Approximate likelihood methods for estimating local recombination rates. Journal of the Royal Statistical Society Series B-Statistical Methodology **64:** 657-680.
FEARNHEAD, P., and N. G. C. SMITH, 2005 A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. Am. J. Hum. Genet. **77:** 781-794.
FISHER, R. A., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press.
FU, H., Z. ZHENG and H. K. DOONER, 2002 Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. Proc Natl Acad Sci U S A **99:** 1082-1087.
GAY, J., and G. MCVEAN, 2007 Estimation meiotic gene conversion rates from population genetic data. Submitted to Genetics.
GILBERT, N., S. BOYLE, H. FIEGLER, K. WOODFINE, N. P. CARTER *et al.*, 2004 Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. Cell **118:** 555-566.
GILKS, W. R., S. RICHARDSON and D. J. SPIEGELHALTER, 1996 *Markov chain Monte Carlo in practice*. Chapman & Hall, London.
GIVENS, G. H., and J. A. HOETING, 2005 *Computational statistics*. Wiley-Interscience, Hoboken, N.J.; Chichester.
GREEN, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711-732.
GREENAWALT, D. M., X. CUI, Y. WU, Y. LIN, H. Y. WANG *et al.*, 2006 Strong correlation between meiotic crossovers and haplotype structure in a 2.5-Mb region on the long arm of chromosome 21. Genome Res **16:** 208-214.
GRIFFITHS, R. C., and P. MARJORAM, 1996 Ancestral inference from samples of DNA sequences with recombination. J Comput Biol **3:** 479-502.
GRIFFITHS, R. C., and S. TAVARE, 1994 Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci **344:** 403-410.
HARR, A., 1910 Zur Theorie der orthogonalen Funktionensysteme. Mathematische Annalen **69:** 331-371.
HASTINGS, W. K., 1970 Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika **57:** 97-109.
HELLENTHAL, G., and M. STEPHENS, 2006 Insights into recombination from population genetic variation. Curr Opin Genet Dev **16:** 565-572.
HILL, W. G., and A. ROBERTSON, 1968 The effects of inbreeding at loci with heterozygote advantage. Genetics **60:** 615-628.
HOLLIDAY, R., 1964 A mechanism for gene conversion in fungi. Genet. Res. **5:** 282−304.

HSU, F., W. J. KENT, H. CLAWSON, R. M. KUHN, M. DIEKHANS *et al.*, 2006 The UCSC Known Genes. Bioinformatics **22:** 1036-1046.
HUDSON, R. R., 1983a Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183-201.
HUDSON, R. R., 1983b Testing the Constant-Rate Neutral Allele Model with Protein Sequence Data. Evolution **37:** 203-217.
HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805-1817.
JEFFREYS, A. J., L. KAUPPI and R. NEUMANN, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet **29:** 217-222.
JEFFREYS, A. J., and C. A. MAY, 2004 Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet **36:** 151-156.
JEFFREYS, A. J., and R. NEUMANN, 2002 Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. Nat Genet **31:** 267-271.
JEFFREYS, A. J., and R. NEUMANN, 2005 Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. Hum Mol Genet **14:** 2277-2287.
JEFFREYS, A. J., R. NEUMANN, M. PANAYI, S. MYERS and P. DONNELLY, 2005 Human recombination hot spots hidden in regions of strong marker association. Nat Genet **37:** 601-606.
JENSEN-SEAMAN, M. I., T. S. FUREY, B. A. PAYSEUR, Y. LU, K. M. ROSKIN *et al.*, 2004 Comparative recombination rates in the rat, mouse, and human genomes. Genome Res **14:** 528-538.
JEON, Y., S. BEKIRANOV, N. KARNANI, P. KAPRANOV, S. GHOSH *et al.*, 2005 Temporal profile of replication of human chromosomes. Proc Natl Acad Sci U S A **102:** 6419-6424.
JETZT, A. E., H. YU, G. J. KLARMANN, Y. RON, B. D. PRESTON *et al.*, 2000 High rate of recombination throughout the human immunodeficiency virus type 1 genome. J Virol **74:** 1234-1240.
JURKA, J., 2004 Evolutionary impact of human Alu repetitive elements. Curr Opin Genet Dev **14:** 603-608.
KAUPPI, L., A. J. JEFFREYS and S. KEENEY, 2004 Where the crossovers are: recombination distributions in mammals. Nat Rev Genet **5:** 413-424.
KEITT, T. H., and D. L. URBAN, 2005 Scale-specific inference using wavelets. Ecology **86:** 2497-2504.
KIM, J., A. A. BHINGE, X. C. MORGAN and V. R. IYER, 2005a Mapping DNA-protein interactions in large genomes by sequence tag analysis of genomic enrichment. Nat Methods **2:** 47-53.
KIM, T. H., L. O. BARRERA, C. QU, S. VAN CALCAR, N. D. TRINKLEIN *et al.*, 2005b Direct isolation and identification of promoters in the human genome. Genome Res **15:** 830-839.
KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893-903.
KINGMAN, J. F. C., 1982 The coalescent. Stoch. Process. Appl. **13:** 235-248.
KONG, A., D. F. GUDBJARTSSON, J. SAINZ, G. M. JONSDOTTIR, S. A. GUDJONSSON *et al.*, 2002 A high-resolution recombination map of the human genome. Nat Genet **31:** 241-247.

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. Genetics **149:** 429-434.
KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393-1401.
LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.
LEWONTIN, R. C., 1964 The Interaction Of Selection And Linkage. Ii. Optimum Models. Genetics **50:** 757-782.
LI, J., M. Q. ZHANG and X. ZHANG, 2006 A new method for detecting human recombination hotspots and its applications to the HapMap ENCODE data. Am J Hum Genet **79:** 628-639.
LI, N., and M. STEPHENS, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165:** 2213-2233.
LICHTEN, M., 2001 Meiotic recombination: breaking the genome to save it. Curr Biol **11:** R253-256.
MALLAT, S. G., 1999 *A wavelet tour of signal processing*. Academic, San Diego, Calif.; London.
MALLET, S. G., 1989 A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **11:** 674-693.
MARCHINI, J., D. CUTLER, N. PATTERSON, M. STEPHENS, E. ESKIN *et al.*, 2006 A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet **78:** 437-450.
MCQUARRIE, A. D. R., and C.-L. TSAI, 1998 *Regression and time series model selection*. World Scientific, Singapore; River Edge, N.J.
MCVEAN, G., 2007 The structure of linkage disequilibrium around a selective sweep. Genetics **175:** 1395-1406.
MCVEAN, G., P. AWADALLA and P. FEARNHEAD, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231-1241.
MCVEAN, G. A., S. R. MYERS, S. HUNT, P. DELOUKAS, D. R. BENTLEY *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581-584.
METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER and E. TELLER, 1953 Equations of state calculations by fast computing machines. Journal of Chem. Phys. **21:** 1087--1092.
MEYER, D., R. M. SINGLE, S. J. MACK, H. A. ERLICH and G. THOMSON, 2006 Signatures of demographic history and natural selection in the human major histocompatibility complex Loci. Genetics **173:** 2121-2142.
MEZARD, C., 2006 Meiotic recombination hotspots in plants. Biochem Soc Trans **34:** 531-534.
MITCHELL, M., 1998 *An introduction to genetic algorithms*. MIT, Cambridge, Mass.; London.
MYERS, S., 2002 The Detection of Recombination Events Using DNA Sequence Data, pp. 243 in *Department of Statistics*. University of Oxford, Oxford.

MYERS, S., L. BOTTOLO, C. FREEMAN, G. MCVEAN and P. DONNELLY, 2005 A fine-scale map of recombination rates and hotspots across the human genome. Science **310:** 321-324.

MYERS, S., C. FREEMAN, A. AUTON, P. DONNELLY and G. MCVEAN, 2007 A single degenerate sequence motif is responsible for 40% of human recombination hotspots and other instances of genome instability. Submitted to Science.

MYERS, S., C. C. SPENCER, A. AUTON, L. BOTTOLO, C. FREEMAN *et al.*, 2006 The distribution and causes of meiotic recombination in the human genome. Biochem Soc Trans **34:** 526-530.

MYERS, S. R., and R. C. GRIFFITHS, 2003 Bounds on the minimum number of recombination events in a sample history. Genetics **163:** 375-394.

NACHMAN, M. W., 2002 Variation in recombination rate across the genome: evidence and implications. Curr Opin Genet Dev **12:** 657-663.

NAGY, P. L., M. L. CLEARY, P. O. BROWN and J. D. LIEB, 2003 Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. Proc Natl Acad Sci U S A **100:** 6364-6369.

NORDBORG, M., 2000 Coalescent theory in *Handbook of statistical genetics*, edited by D. J. BALDING, M. BISHOP and C. CANNINGS. Wiley, Chichester.

PERCIVAL, D. B., and A. T. WALDEN, 2000 *Wavelet methods for time series analysis*. Cambridge University Press, Cambridge.

PETES, T. D., 2001 Meiotic recombination hot spots and cold spots. Nat Rev Genet **2:** 360-369.

PRUITT, K. D., T. TATUSOVA and D. R. MAGLOTT, 2005 NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res **33:** D501-504.

PTAK, S. E., D. A. HINDS, K. KOEHLER, B. NICKEL, N. PATIL *et al.*, 2005 Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet **37:** 429-434.

RABINER, L. R., 1989 A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition. Proceedings of the Ieee **77:** 257-286.

RADA-IGLESIAS, A., O. WALLERMAN, C. KOCH, A. AMEUR, S. ENROTH *et al.*, 2005 Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. Hum Mol Genet **14:** 3435-3447.

REED, F. A., and S. A. TISHKOFF, 2006 Positive selection can create false hotspots of recombination. Genetics **172:** 2011-2014.

REICH, D. E., M. CARGILL, S. BOLK, J. IRELAND, P. C. SABETI *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199-204.

SABO, P. J., M. S. KUEHN, R. THURMAN, B. E. JOHNSON, E. M. JOHNSON *et al.*, 2006 Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat Methods **3:** 511-518.

SCHEET, P., and M. STEPHENS, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet **78:** 629-644.

SHIRAKI, T., S. KONDO, S. KATAYAMA, K. WAKI, T. KASUKAWA *et al.*, 2003 Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A **100:** 15776-15781.

SMIT, A. F., G. TOTH, A. D. RIGGS and J. JURKA, 1995 Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol **246:** 401-417.

SMIT, A. F. A., R. HUBLEY and P. GREEN, 2004 RepeatMasker Open-3.0., pp.
SMITH, A. V., D. J. THOMAS, H. M. MUNRO and G. R. ABECASIS, 2005 Sequence features in regions of weak and strong linkage disequilibrium. Genome Res **15:** 1519-1534.
SMITH, N. G., and P. FEARNHEAD, 2005 A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. Genetics **171:** 2051-2062.
SONG, Y. S., and J. HEIN, 2005 Constructing minimal ancestral recombination graphs. J Comput Biol **12:** 147-169.
SPENCER, C. C., P. DELOUKAS, S. HUNT, J. MULLIKIN, S. MYERS *et al.*, 2006 The influence of recombination on human genetic diversity. PLoS Genet **2:** e148.
SPENCER, C. C. A., 2006 Human Genetic Variation and the Evidence for Natural Selection, pp. 224 in *Statistics*. University of Oxford, Oxford.
STEPHENS, M., and P. DONNELLY, 2000 Inference in molecular population genetics. Journal of the Royal Statistical Society Series B-Statistical Methodology **62:** 605-635.
STEPHENS, M., and P. DONNELLY, 2003 A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet **73:** 1162-1169.
STEPHENS, M., and P. SCHEET, 2005 Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet **76:** 449-462.
SUMINAGA, R., Y. TAKESHIMA, K. YASUDA, N. SHIGA, H. NAKAMURA *et al.*, 2000 Non-homologous recombination between Alu and LINE-1 repeats caused a 430-kb deletion in the dystrophin gene: a novel source of genomic instability. J Hum Genet **45:** 331-336.
TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437-460.
THE ENCODE PROJECT CONSORTIUM, 2007 Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447:** 799-816.
THE INTERNATIONAL HAPMAP CONSORTIUM, 2005 A haplotype map of the human genome. Nature **437:** 1299-1320.
THE INTERNATIONAL HAPMAP CONSORTIUM, 2007 The Phase II HapMap. Submitted to Nature.
THOMAS, P. D., M. J. CAMPBELL, A. KEJARIWAL, H. MI, B. KARLAK *et al.*, 2003 PANTHER: a library of protein families and subfamilies indexed by function. Genome Res **13:** 2129-2141.
TRINKLEIN, N. D., J. I. MURRAY, S. J. HARTMAN, D. BOTSTEIN and R. M. MYERS, 2004 The role of heat shock transcription factor 1 in the genome-wide regulation of the mammalian heat shock response. Mol Biol Cell **15:** 1254-1261.
VENTER, J. C., M. D. ADAMS, E. W. MYERS, P. W. LI, R. J. MURAL *et al.*, 2001 The sequence of the human genome. Science **291:** 1304-1351.
VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. PLoS Biol **4:** e72.
WALL, J. D., 2000 A comparison of estimators of the population recombination rate. Mol Biol Evol **17:** 156-163.

WALL, J. D., L. A. FRISSE, R. R. HUDSON and A. DI RIENZO, 2003 Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. Am J Hum Genet **73:** 1330-1340.
WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7:** 256-276.
WEI, C. L., Q. WU, V. B. VEGA, K. P. CHIU, P. NG *et al.*, 2006 A global map of p53 transcription-factor binding sites in the human genome. Cell **124:** 207-219.
WEIR, B. S., 1979 Inferences about linkage disequilibrium. Biometrics **35:** 235-254.
WELLCOME TRUST CASE CONTROL CONSORTIUM, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447:** 661-678.
WILSON, D. J., and G. MCVEAN, 2006 Estimating diversifying selection and functional constraint in the presence of recombination. Genetics **172:** 1411-1425.
WINCKLER, W., S. R. MYERS, D. J. RICHTER, R. C. ONOFRIO, G. J. MCDONALD *et al.*, 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. Science **308:** 107-111.
WIUF, C., 2002 On the minimum number of topologies explaining a sample of DNA sequences. Theor Popul Biol **62:** 357-363.
WIUF, C., T. CHRISTENSEN and J. HEIN, 2001 A simulation study of the reliability of recombination detection methods. Mol Biol Evol **18:** 1929-1939.
WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97-159.
WU, T. C., and M. LICHTEN, 1994 Meiosis-induced double-strand break sites determined by yeast chromatin structure. Science **263:** 515-518.