Exploring Rates and Patterns of Variability in Gene Conversion and Crossover in the Human Genome

Garrett Hellenthal

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2006

Program Authorized to Offer Degree:
Statistics

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Garrett Hellenthal

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____
Matthew Stephens

Reading Committee:

_____
Matthew Stephens

_____
John Storey

_____
Elizabeth Thompson

Date: _____

University of Washington

Abstract

Exploring Rates and Patterns of Variability in Gene Conversion and Crossover in the Human Genome

Garrett Hellenthal

Chair of the Supervisory Committee:
Associate Professor Matthew Stephens
Statistics

Meiotic recombination is a biological process that exchanges the paternal and maternal DNA of chromosomes before they are passed along to offspring. There are two known outcomes of recombination: crossover and gene conversion. Recently, fine-scale human crossover rates have been inferred with some success using population data. However, reliable estimation of gene conversion rates has proven more difficult to come by. We present a new model to jointly estimate crossover and gene conversion rates using an updated version of the PAC Likelihood of Li and Stephens (2003). Furthermore, we incoporate this new model into additional statistical machinery to examine variability in recombination rates across the human genome. We apply our methods to data from the *SeattleSNPs* project to provide insights into the genome-wide relative rate of gene conversion to crossover in humans, among other features of recombination, under various assumptions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express sincere appreciation to his family and friends. In particular I extend a special gratitude to the colleagues who have guided me through this process. In addition, I'd like to thank my reading committee for their patience and insights. My research was generously supported by Genome Training Grant HG00035-09/10/11. My advisor (MS) was supported by NIH Grant 1R01HG/LM02585-01.

Chapter 1
INTRODUCTION AND BACKGROUND

1.1 Crossover and Gene Conversion

The genetic material of humans contains some six billion *basepairs* (bp), each of which is one of four order-dependent paired nucleotide configurations: A/T, G/C, C/G, or T/A. Healthy individuals have 23 pairs of chromosomes that comprise this entire genetic material. Within each pair, one chromosome is inherited from the mother (*maternal* chromosome) and one is inherited from the father (*paternal* chromosome). Chromosomes within a pair are referred to as *homologs*. One chromosome from each pair is passed along to offspring.

This process by which genetic material is prepared for transmission from parent to offspring is called *meiosis*. This process proceeds roughly as follows. Each chromosome within a pair replicates; two identical chromosome replicants are referred to as *sister chromatids*. Then these pairs, each member of the pair now containing two copies of its chromosome, align near a cell's equator so that they can properly segregate into gametes. The final product of a normal round of meiosis is four daughter cells, each containing a chromosome representing each pair. This gives 23 chromosomes in total per daughter cell. However, the chromosomes in each daughter cell are not usually strictly a copy of the maternal or paternal genetic material, but often contain parts of each.

Meiotic recombination between homologous chromosomes, which occurs when the chromosomes are aligned, is the process by which the maternal and paternal copies of a human individual's chromosomes are "mixed-up" before being passed along to

offspring. It is a major force influencing genetic diversity in the human population. Researchers have taken pains in recent years to understand both the causes and effects of recombination, including its application to disease mapping and the biological factors contributing to the process. Current widely-accepted theory suggests recombination events can result in two different outcomes (e.g. Szostak et al. (1983)). One involves the reciprocal exchange of genetic information between two chromosomes, known as a *crossover*, in which material is swapped equally between the two (left column of Fig.1.1). The other is a nonreciprocal transfer of a piece of genetic sequence from one chromosome to another without vice versa exchange, known as a *gene conversion* event (right column of Fig.1.1).

Both crossover and gene conversion can also occur such that a site from one homolog is attached or inserted into an area not directly across from it in the other homolog were the chromosomes properly aligned (e.g. Slightom et al. (1980)). This kind of recombination, known as "intergenic exchange," or "ectopic exchange" or "non-allelic recombination," is not dealt with in this thesis.

1.2 Double-Strand-Break Theory and Predecessors

1.2.1 Holliday (1964) and Meselson and Radding (1975)

Both crossover and conversion events are thought to be distinct results of the same initiating mechanism, the first widely-accepted model for which is that of Holliday (1964), devised out of observations on yeast. The left side of Fig.1.2 illustrates the major concepts of this model. The black and white colors represent distinct homologs, specifically two homologous chromosomes over which recombination in these illustrations will occur (as was the case with the maternal and paternal chromosomes in Fig.1.1). Each chromosome is double-stranded, with a 5' to 3' strand (the top one in the black chromosome) and a 3' to 5' strand (the bottom one in the black chromosome; the reverse is true for the white chromosome). A nick, or small break in the

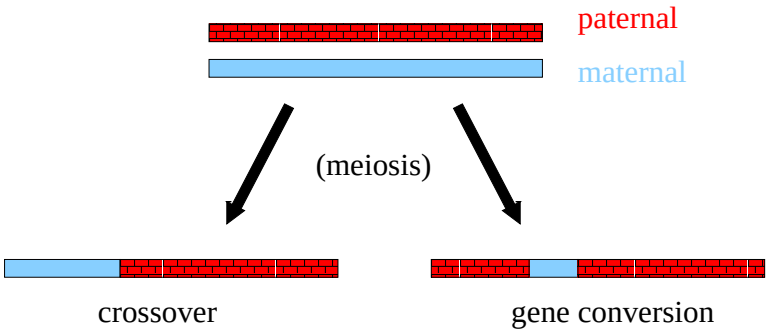


Figure 1.1: The products of meiosis between homologous chromosomes after recombination. The top two rectangles represent paternal and maternal chromosomes within a single chromosome pair. Each chromosome's sister chromatid is omitted from this cartoon. A single gene conversion (right) results in a small segment of DNA from one chromosome (here maternal) replacing the homologous region on the other (here paternal) in the offspring chromosome. A single crossover (left) results in the first part of the offspring chromosome coming from one source (here maternal) and the rest from another (here paternal). The representation on the bottom right is not to scale, in that gene conversions typically cover only a small part of genetic sequence. If each bar above indeed represents a chromosome, the segment of maternal DNA replacing the paternal DNA on the picture at bottom right would be unobservably small.

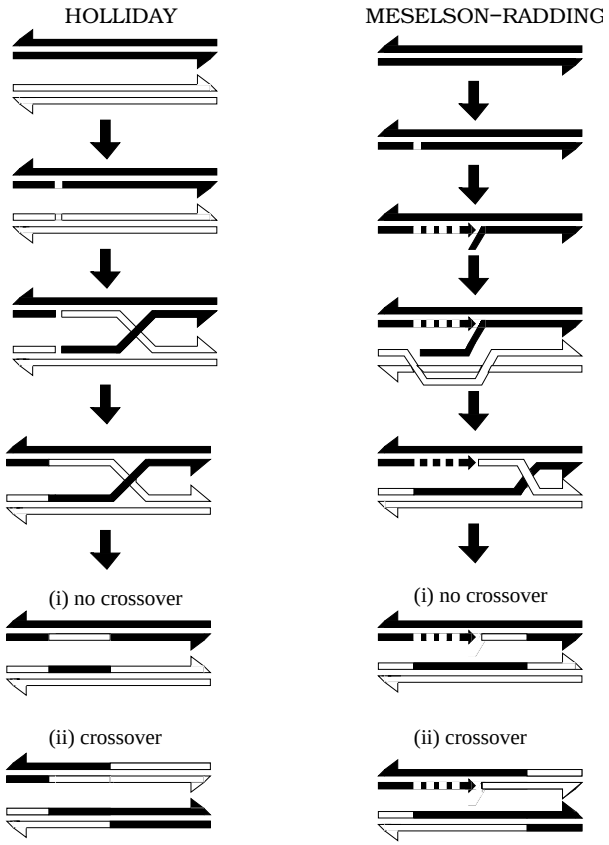


Figure 1.2: The recombination models of Holliday (1964, left) and Meselson-Radding (1975, right) for yeast. Two homologous chromosomes (black and white) align themselves during meiosis. Each chromosome's sister chromatid is omitted here. In the Holliday model, recombination starts with a nick on one strand of each homolog, followed by the formation of a Holliday junction. The junction can be resolved by cutting the inner strands, leading to (i) no crossover, or the outer strands, leading to (ii) crossover. Gene conversion potentially occurs via mismatch correction. In the Meselson-Radding model, a single nicked strand on only one chromosome "invades" its homolog (the white strand), displacing a "D" loop and being corrected by its complementary strand, resulting in asymmetric DNA, which may promote gene conversion. The Holliday junction still forms in this model, and "slides" along the chromosome, allowing for a region of symmetric DNA, before being resolved á la Holliday's model, by cutting the inner (i) or outer (ii) strands. *(images modeled on Stahl (1994))*

DNA, occurs in one strand on each of the two chromosomes at the same spot, and a "Holliday junction", or exchange point between the two strands, is formed. How this junction is resolved, or cut, determines whether crossover occurs in the products or not. In any case, *heteroduplex DNA*, defined as a chromosome segment whose two strands contain DNA material from separate sources – here one strand contains maternal DNA and the other strand paternal DNA – is formed and corrected, either resulting in gene conversion or not. A "correction" refers to synchronizing the two strands of a single chromosome to be appropriately complementary basepairs. For example, for Fig.1.2(i), if all the color mismatches were corrected to black, the result would be a gene conversion onto the white chromosome.

DNA transfer was considered symmetric in this model, so that, for example, the gap left by inserting a part of the paternal DNA into the maternal chromosome is filled by the displaced homologous section of that maternal chromosome (see Fig.1.2 (i) and (ii), left). The net result of this symmetry is that every position along the genetic region has a 2:2 strand ratio of paternal-to-maternal genetic material, before and after the recombination event. A gene conversion can only occur in this model by correction such as that described above. It should be noted that correction does not always occur, resulting in aberrant yeast products. We do not deal with modeling this uncorrected kind of DNA in this thesis.

Meselson and Radding (1975) later revised Holliday's model to account for experimental evidence, initially in yeast, that not all heteroduplex DNA was symmetric. Their model allows for a small region of DNA to be displaced in one homolog by the recombination-initiating *invading strand* of the other homolog. Thus only one strand, the "invading," is nicked, and it forms a Holliday junction with a strand of its homolog after displacing a region of the homolog's DNA. This Holliday junction then "slides" along the chromosome away from the initiating site. The gap left on the invading strand from its nick site part way to the site where the Holliday junction is resolved (cut) is repaired by its own complementary strand, resulting in a 3:1

strand ratio in the region, biased in favor of the invading strand (asymmetric DNA). The model also allows for an area of symmetric heteroduplex DNA as in Holliday's model, here in between the end of this repaired region and the continuing slide of the Holliday junction (see Fig.1.2, right). Finally, analagous to Holliday, correction can occur on mismatches between strands on each of the product chromosomes, resulting in 4:0 bias of the invading strand (gene conversion) or correction back to 2:2 (no gene conversion).

Note that in the case of 4:0 correction, the resulting meiosis chromosome products are 3 to 1 in favor of the invading chromosome. The sister chromatid of the invading chromosome, in addition to each of the two chromosomes involved in the recombination, have the DNA of the invading chromosome. In contrast, the sister chromatid of the non-invading chromosome retains its own DNA.

1.2.2 The Double-Strand-Break Model

Still further studies presented evidence that didn't quite fit the Meselson and Radding (1975) model. For example, it fails to explain experiments that suggest it is the invading strand, the strand that appears to initiate the recombination, that seems to more often get replaced with gene conversion and not the other way around (Jeffreys and Neumann (2002), Nicolas et al. (1989)). The "presently most favoured" (Moens, 2003) model to date seems to be the double-strand break (DSB) model of Szostak et al. (1983), also formulated out of observations on yeast but theorized to extend to higher-order eukaryotes. Fig.1.3 illustrates the basic concepts of this model. Again two separate, double-stranded homologs (paternal and maternal) are depicted, over which recombination is occuring (Fig.1.3-A; sister chromatids are again omitted). In this model, breaks are created on both strands of the recombination initiating chromosome, and a gap is enlarged surrounding the break (Fig.1.3-B). This gap is repaired using the complementary strand of its homolog, resulting in guaranteed gene conversion, and two "Holliday junctions" are formed (Fig.1.3-C). As in the earlier

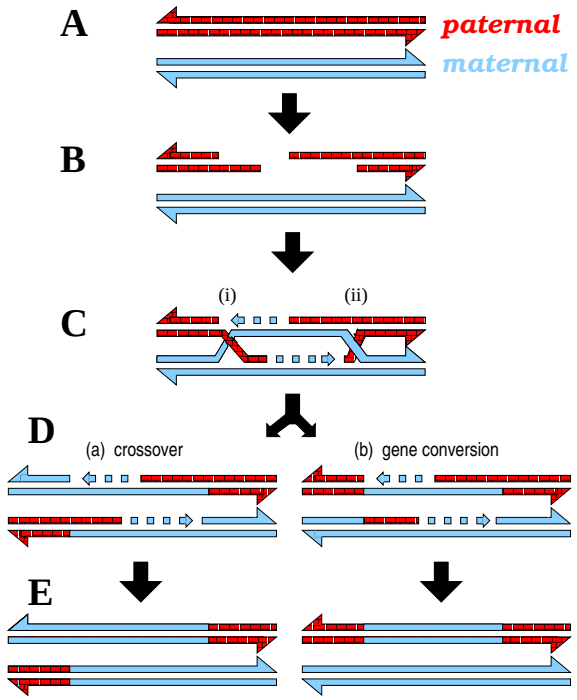


Figure 1.3: The double-strand-break (DSB) repair model for recombination as explored in yeast. During meosis, the two homologous chromosomes (paternal and maternal) align (**A**). Each chromosome's sister chromatid is omitted from this cartoon. A break occurs here on the paternal chromosome (**B**); the resulting gap is repaired using its maternal homolog (**C**). If the two resulting Holliday junctions (i) and (ii) (**C**) are cut in an opposite manner (e.g., such that (i) is cut on the outer strands and (ii) on the inner strands) there will be a crossover (**D**-(a)). If they are cut in the same manner (e.g. (i) and (ii) both cut on the inner strands), there will be no crossover (**D**-(b)). The mismatches between strands on the same chromosome are repaired, resulting in less or more gene conversion, or they are not, resulting in other types of aberrant products. Here all repairs are in favor of the maternal homolog (**E**). *(thanks to the webpage of Franklin Stahl as a model for this image)*

models, the manner these junctions are resolved determines whether crossover occurs in the products or not (Fig.1.3-D). The formation of *heteroduplex DNA* in the regions flanking this area, i.e. regions with a gap on one strand of the recombination initiating chromosome but not on the other strand, can result in mismatches on the two strands. The repair of these mismatches after Holliday junction resolution affects the size of gene conversion events and gives the final products (Fig.1.3-E), one of which might be passed along to an offspring in the next generation.

1.3 Statistical Modeling of Crossover and Gene Conversion

1.3.1 Introduction

Though a model has been proposed by Szostak et al. (1983) to explain the process behind gene conversion and crossover, a model that has survived over twenty years with few major changes, many questions remain. For example, the rates at which these DSBs occur, and how often they are resolved as gene conversions versus as crossovers, has not yet been thoroughly explored. Furthermore, little is known about the variability across the genome of the rates of these processes.

Recent statistical and experimental analyses have shed some light on properties of human crossover rates, e.g. the detection of crossover "hotspots" in humans by Jeffreys et al. (2001), Crawford et al. (2004), McVean et al. (2004). However, much less is known about gene conversion. In fact, joint estimation of the two has proven quite difficult in recent studies (e.g. Frisse et al. (2001), Wall (2004)). Still it seems probable from some recent experimentation that conversions occur more often than crossovers in at least some areas of the genome (Carpenter (1984), Jeffreys and May (2004)). Frisse et al. (2001) conclude in their studies that certains kinds of "short range ... data may lead to invalid inferences about population histories if gene conversion is not taken into account." Indeed Andolfatto and Nordborg (1998) claim that "at intragenic distances, gene conversion, rather than crossing-over, is likely to be the

dominant force that breaks up associations among sites." Przeworski and Wall (2001) found that the data likelihood for ten human loci under their model "increases 44-fold with the inclusion of gene conversion," and Carpenter (1984) notes there is an "evolutionary persistence of high frequencies of gene conversion without crossing over" in eukaryotes. Of particular interest is the relative rate of gene conversion to crossover, denoted as f here and in other literature, which provides an idea of how often Holliday junctions are resolved as a gene conversion relative to as a crossover.

Below we briefly summarize some of the key contributions from pedigree and sperm analysis to our understanding of the recombination process. We then summarize the methods and results of some of the statistical approaches that utilize *population data* to study rates of crossover and gene conversion. Population data consists of chromosomes collected from unrelated individuals randomly sampled from a population. In Chapters 2-5, we propose a new means of analyzing population data to explore the recombination process.

Analysis Using Pedigrees

Large-scale crossover activity over megabases (Mb), or millions of basepairs, has been studied extensively by several methods using data from related individuals in families, i.e. *pedigree data*. Studying chromosomes from members of a pedigree allows one to infer crossovers that have occured across a region as genetic material is passed from one generation to the next. Using such data, crossover rates have been seen to vary substantially over megabase scales. For example, Kong et al. (2002) found crossover rates to vary from 0.1 to 3.5 centimorgans (cM) per Mb along chromosome 3, where 1 *centimorgan* per Mb refers to, approximately, a 1% probability of crossover per meiosis between two chromosomal locations a megabase apart. However, crossovers are relatively rare: regions with substantial crossover activity in humans might only have a crossover frequency of about 0.00001/bp. Therefore, pedigree data, which considers only a few meiosis events per family, is currently only useful in studying

crossover over these megabase scales.

Patterns of gene conversion are considerably more difficult to study with pedigree data than patterns of crossover are. As with crossover, "gene conversion events occur extremely rarely per base pair so that direct measurements often require the examination of a prohibitive number of meioses" (Ptak et al., 2004). In addition, the limited available information suggests that gene conversions typically cover a small area in humans, perhaps only 55-290bp per event (Jeffreys and May, 2004). Therefore, as pedigree data currently has resolution only on the scale of megabases, the study of gene conversion using pedigree data is infeasible.

Analysis Using Sperm Typing

The laboratory technique of *sperm typing* gets around such difficulties by considering the products of thousands of meiosis events at a time, allowing for fine-scale study over hundereds of bp. A sperm sample contains tens of thousands of *haploid* meiosis products from a single individual, where a haploid is one chromosome from a homologous pair. The sperm are analyzed using allele-specific Polymerase Chain Reaction (PCR) techniques to directly observe both crossover and gene conversion events. Using such analysis, extensive variation in crossover rates have been observed on very fine scales in some regions of the genome, such as the Major Histocompatibility Complex (MHC), an area thought to be involved in the lymphatic system (Jeffreys et al., 2001). The majority of the crossover events in this region appear to be clustered into narrow regions of about 1-2 kilobases (kb). Such areas are termed crossover *hotspots*. Hotspots have highly variable intensities; the five hotspots in the MHC found via sperm analysis of eight UK individuals have anywhere from 10 to 1000 times the estimated crossover activity of regions outside the hotspots.

Furthermore, analysis of one such MHC hotspot, called *DNA3*, revealed the hotspot to have a substantial amount of gene conversion activity as well, probably because *DNA3* is really a hotspot for double-strand-breaks. In fact, within the *DNA3* hotspot,

gene conversion was observed to occur 4-15 times more often than crossover events, i.e. $f \approx$4-15, in the two men studied (Jeffreys and May, 2004). A similar relative rate of gene conversion to crossover was observed in another MHC hotspot examined in one man, called *DMB2* (Jeffreys and May, 2004). However, in another known crossover hotspot, called *NID1*, also thought to be a hotspot for gene conversion, gene conversion occurred about four times *less* often than crossovers, i.e. $f \approx 1/4$, in the one man studied for both (Jeffreys and Neumann, 2005). Similarly, f was $\approx 1/3$ in a hotspot, called *SHOX*, in the pseudoautosomal pairing region on the sex chromosomes (Jeffreys and May, 2004). In yet another hotspot in the β-globin gene region on chromosome 11, f was estimated to be $<1/12$ in one man (Holloway et al., 2006). This disparity in f across these four hotspots brings up a question as to what extent f varies genome-wide, in and out of hotspots.

The sparse information we have on the basepair coverage of gene conversion events in humans, known as the *tract length*, comes from sperm typing, which has observed average lengths to be around 55-290bp (Jeffreys and May, 2004). Studying tract length is greatly affected by the density of genetic markers, which for this thesis will be assumed to be *Single Nucleotide Polymorphims (SNPs)*, in the region being studied. SNPs are sites that have at least two different nucleotides present in the chromosomes of individuals sampled from a population. High frequency SNPs, ones in which both nucleotides occur with appreciable frequency in a population of chromosomes, occur roughly 1 every kb in humans on average. Jeffreys and May (2004) estimated tract lengths in the MHC, which has a considerably higher density of SNPs than the $\approx$1 per kb genome-wide average and therefore the potential for better precision in tract length estimation. However, some caution is necessary, as the MHC is thought to be under selective pressure and may have features, recombination or otherwise, that are not typical of the genome as a whole. Still Jeffreys and May (2004) suggests tract lengths are generally small and that "the only reasonably clear evidence for long conversion tracts in humans comes from rare germline reversions of triplet repeat expansions."

Despite the advances in our understanding of recombination that sperm analysis has provided, this kind of "direct (laboratory-based) estimation of gene conversion rates [and crossover ones over fine-scales] is technically challenging and extremely laborious, even for a single locus" (Wall, 2004). Therefore it cannot be realistically performed over large regions and is therefore currently infeasible to use for studying genome-wide rates of recombination. Another limitation of sperm analysis is that it is naturally only representative of rates in males. There is evidence from pedigree studies that crossover rates vary between males and females (Kong et al., 2002). In addition, such data is free from selective pressures that may be acting on a chromosome's fitness; perhaps some gametes observed in sperm typing are not reproductively viable. For such reasons, there has been substantial interest in developing methods for recombination estimation at the population level.

1.3.2 Analysis Using Linkage Disequilibrium Approaches

Population data consists of chromosomes collected from unrelated individuals randomly sampled from a population. Such data can be used to study *Linkage Disequilibrium* (LD), or patterns of nonrandom associations among alleles on a chromosome. These associations can reflect historical recombination events that have occured over the many meioses events of the ancestral tree relating the chromosomes, i.e. the *coalescent* (see the subsection on *Coalescent Approaches* below for a more detailed description of the coalescent). Relative rates of the recombination process over this history can be studied on a fine-scale, i.e. kilobases, as is true with sperm typing (though they are measuring slightly different things). In contrast to sperm typing, however, large amounts of fine-scale biological data spanning the entire genome have become recently available, and all that remains is to find efficient computational means of extracting the information of recombination rates out of this data. We will initially focus on some of the efforts to estimate gene conversion rates alone using LD methods in *Early Approaches*, as a contribution to this goal is one of the primary

focuses of this thesis. A few of the most popular crossover rate estimation methods that utilize LD will be mentioned again in *Coalescent Approaches*.

Early Approaches

Some of the initial attempts at estimating rates of gene conversion are rather *ad hoc*. One of the first is that of Stephens (1985), who looked at the clustering of SNPs to test if gene conversion is likely to have occured between two groups of sequences. Essentially, a set of sequences is partitioned into two groups, and all the variable sites between them are found. Variable sites are sites for which all members of each group have the same nucleotide, but the nucleotide is different between each group. It is then checked to see if these variable sites are significantly clustered. If so, it may be thought that there was a conversion between some members of one of the groups at these sites, as they have similar sequences in this area compared to the other group.

Of course, one does not know which partitions will reveal such clustering, so all must be tested, which can reduce the power to detect an anamoly if it is present. Sawyer (1989) attempted to avoid this issue while still working with numbers and patterns of variable sites between sequences. Roughly, his test for detecting gene conversion consists of taking all pairs of individuals from a sample of n and, for each pair, dividing the sequence into fragments split by the variable sites between the two. He then counts the number of segregating sites per each particular fragment that occur in the entire sample of n chromosomes but not in this pair. A high maximum value of all these pairwise/fragmentwise values would indicate a potential gene conversion event, assuming the mutation rate is similar across the sequence. Thus you have $\binom{n}{2}$ tests to perform, rather than $2^{n-1} - 1$ as with Stephens (1985). Still both of these approaches are essentially testing for some potential presence of gene conversion without explicitly estimating rates or basepair lengths of the events, though you could maybe get some rough estimate of each. The approaches do not differentiate at all between crossover and gene conversion, a much more difficult problem.

Hilliker et al. (1994) showed that gene conversion tract lengths, determined by crosses of *Drosophila melanogaster* in the *rosy* locus, a region that could easily be screened for non-crossover recombination, seem to closely follow a geometric distribution.

Betran et al. (1997) used the assumption of geometric tract lengths in developing their own statistical methods, which involves initially dividing a sample into two subpopulations based on diversity. The most relevant estimated parameter is then a nucleotide's "informativeness," or the ability to find the first and last sites of a gene conversion event. Their criterion for "informativeness" states (a) the alleles at the polymorphic site must have $\leq 20\%$ frequency in the subpopulation presumed to be the recipient of a gene conversion event, and (b) the same allele must be at least three times greater in the other "donor" subpopulation. The number and tract lengths of gene conversion events can be estimated in this manner, after a correction for the bias expected due to "uninformativeness" (e.g. monomorphic sites contained at either end of a gene conversion tract would result in an observed tract length too small). The authors admit, however, that they don't expect to observe many of the true gene conversion events in this manner, and in fact estimate they are only observing 26% in one reported data set (*rp49* locus data in *Drosophila subobscura*). This in addition to the potential difficulty of choosing proper subpopulations and the lack of joint crossover/conversion estimation ensured further efforts to come.

Coalescent Approaches

A powerful new means of exploring how chromosomes sampled from a population relate to one another, and hence how patterns of recombination might affect such relations, arrived via the *coalescent* (Kingman, 1982). The idea is that all "unrelated" humans share common ancestors if you go back far enough in time. The coalescent refers to the ancestral tree that relates the chromosomes of different individuals. If you go back far enough in time, the homologous chromosomes of n individuals

sampled from a population will have a common ancestor. Thus a certain region of a particular individual's genetic material will look more or less similar to the same region of another individual, based on how far back in time these regions of those individuals "coalesced," or had a common ancestor. All other things equal, the less time it takes for two chromsome segments to coalesce, the more alike they are expected to be. Biological phenomena that might create differences between any two chromosomes that eventually share a common ancestor include mutation and recombination. Therefore, one might be able to estimate rates of these phenomena by examining the degree of relatedness among a sample of chromosomes, i.e. patterns of LD, under various assumptons regarding the behavior of the coalesent tree. We outline two of the more noteworthy models that utilize coalescent theory in such a manner: a "Composite Likelihood" method developed by Hudson (2001), and a "Product of Approximate Conditionals" model created by Li and Stephens (2003).

Composite Likelihood One of the more interesting approaches to joint estimation of crossover and gene conversion rates is that of Frisse et al. (2001). Their method uses the "composite likelihood" approach developed by Hudson (2001) initially to estimate crossover rates alone. His method uses pairs of sites and simulations generated via coalescent theory to achieve a likelihood for the crossover rate. This rate is represented by $\rho, = 4N_e r$, where N_e is the effective population size and r is defined here as the probability per unit physical distance of a crossover in a single transmission from parent to offspring. The use of coalescent theory on population data allows for estimation of ρ, but not of r itself. This theory allows us to relate independent individual data in an informative manner, but we are considering an averaged recombination rate over many generations, which depends on the population size N_e in addition to the recombination probability r. For a pair of sites, a coalescent-based *ancestral recombination graph (ARG)* is constructed for each value of ρ, from the present day until all sampled chromosomes have a common ancestor at each site. For every

pair of branches between the trees at each site, a single mutation is placed on each branch, weighted by its probability, and checked for compatibility with the observed sample. The fraction of such compatibilities over all branch pairs gives an idea of the probability with which the simulated ARG might have generated the observed data. Many such ARGs are constructed to get an expected probability for each value of ρ. Therefore, a likelihood is formed over a grid of ρ values, and appropriate statistical inference, e.g. maximum likelihood estimation (MLE), can be used to estimate ρ.

Actually a single pair of sites gives little information towards providing accurate estimation of ρ. Therefore the likelihoods for all pairs of segregating sites in the sample are multiplied together, creating Hudson's "composite likelihood" under assumptions of constant effective population size N_e, random mating, infinite-sites, and no selection. (All of these are assumptions of the coalescent simulations.) Some of these assumptions can be relaxed; for example, McVean et al. (2002) altered Hudson (2001)'s model to relax the infinite sites assumption. As noted by Hudson (2001), the composite likelihood is not estimating a true likelihood in that it assumes all pairs of sites are independent, which is blatantly untrue. In particular, this assumption generally gives a more peaked likelihood than is appropriate, so that assessing variation in estimates and confidence intervals can be troublesome. For example, confidence intervals are usually found through simulation, e.g. Frisse et al. (2001), or permutation, e.g. Ptak et al. (2004). Nevertheless, Hudson's method does provide reasonably accurate point estimates of ρ (see Hudson (2001), Li and Stephens (2003)).

All of these assumptions carry over to Frisse et al. (2001)'s additional amendment to the "composite likelihood" to allow for estimation of γ, the assumed-constant rate of gene conversion in a region. Here $\gamma = 4N_e g$, where g is the probability of a gene conversion event per unit physical distance in a single transmission from parent to offspring. Specifically, the authors jointly estimate ρ and the relative rate of gene conversions to crossovers, $f = \gamma/\rho$. To improve the precision of their estimates, Frisse et al. (2001) combine the likelihood information of ten "locus pairs," or genetic

regions of size ≈2kb from various locations of the genome, assuming f and ρ to be constant across all ten. Despite combining the information across multiple loci in such a manner, Frisse et al. (2001) find that their method gives very large confidence intervals, constructed via simulations, for ρ and f (see Fig.2 of Frisse et al. (2001)). They also note that their model has difficulty distinguishing between large values of ρ coupled with small values of f and large values of f coupled with small values of ρ. This complication is not unexpected: a gene conversion event in the ancestral history of a sample of chromosomes potentially leaves a LD pattern similar to the pattern left by two crossover events that occur at different times in the ancestral history but near to each other in terms of chromosome location. Therefore, a model might potentially interpret a gene conversion event as a single gene conversion or as two distinct crossovers.

Using data from *SeattleSNPs*, described in Chapter 2, and assuming both f and crossover to be constant across the genome, Frisse et al. (2001) estimate genome-wide average f to be around 4 to 25; that is, double-strand-breaks are resolved as gene conversions some 4-25 times more often than as crossovers. It is important to point out that gene conversion rates from population data are confounded with the *tract length*, or basepair length of genetic material replaced by a gene conversion event. While sperm analysis techniques can observe gene conversions and their tracts directly, with precision depending on SNP density, Frisse et al. (2001) must assume a tract length for these estimates. As little was known at the time about tract lengths in humans, they based their assumed tract length on studies of yeast and fruit flies. For the estimates presented here, they used a mean tract length of 500bp.

In addition to rates of crossover, another potential confounding factor that can affect gene conversion rate estimation is genotyping error, as Ptak et al. (2004) note. The method of Ptak et al. (2004) is similar to that of Frisse et al. (2001) but allows for crossover rates to vary among loci (or genes or "independent" regions). To jointly estimate f and $\vec{\rho}$, the vector of ρ across loci, they use a *profile likelihood* on f: for each

value of f, they find $\hat{\rho}_j$, the MLE of ρ per each locus j, and multiple the likelihood values $L(f, \hat{\rho}_j)$ together for all j. They then maximize f over these product likelihoods. To improve inference on f, the authors assume f to be constant genome-wide and combine the likelihood information across ≈70-80 assumed-independent loci, some of which have more than 200 segregating sites. They estimate genome-wide average f to be ≈0.3-1.0 using the *SeattleSNPs* data, assuming the same mean tract length of 500bp as Frisse et al. (2001) above. However, when they applied their model to simulated datasets where f=0 and genotype error occurs at a rate of 0.5%, they found that their method often estimates f to be in this range. This highlights the difficulty in extracting the subtle effect of gene conversions on the data: gene conversion rates on the order of human crossover rates perhaps have as faint an effect on LD as genotyping error as small as 0.5%.

To attempt to alleviate some of these difficulties, Wall (2004) extended the composite-likelihood approach to looking at three-site likelihoods rather than two. Three sites might be able to capture the effects of gene conversion more effectively than two, as the result of a gene conversion event is to produce a segment of genetic material from one source (i.e. the middle SNP) flanked by genetic material from another source (i.e. the outer two SNPs). However, while the likelihoods for every pair of sites in a region can be relatively efficiently calculated for a large number of chromosomes, all 3-site likelihoods can be calculated for only ten or so chromosomes at a time. For a dataset with more than ten chromosomes, Wall (2004) randomly samples several subsets of ten chromosomes, calculates all 3-site likelihoods for each subset, and multiples the likelihood values across subsets. Within each subset, simulations analagous to Hudson (2001) are used to check for compatibility, but now using three sites instead of two. Therefore the same assumptions – a neutral model with constant, randomly mating population N_e across generations, biallelic sites, infinites sites, constant ρ within loci – currently apply, though some could probably be relaxed in a relatively straight-forward manner. Despite the use of three loci at a time rather than two, estimates

of f are not very accurate. Again assuming f to be constant across multiple loci and combining likelihood information across these loci appears to be necessary for reliable inference. While it takes only a single locus to estimate ρ with accuracy in simulated datasets, it requires at least 20 loci to have MLE estimates of f within a factor of two of the true f even 50% of the time. This again illustrates the complexities of jointly estimating crossover and gene conversion rates. In addition, it is unknown whether the assumption that f is constant across different loci is even close to biologically true.

As a final note, the composite likelihood method of Hudson (2001) has been updated by McVean et al. (2004) to look at variable crossover rates across the genome on fine-scales, e.g. within a gene or ≈20-kb region. No similar work has yet been done for gene conversion, which is a substantially more difficult problem due to the rarity and short tract length of events. McVean et al. (2004) found that crossover hotspots occur throughout the genome rather than just in the few regions sperm typing has considered. Subsequently, using McVean et al. (2004)'s method, Myers et al. (2005) found that crossover hotspots occur perhaps one every 50kb across the genome. Genomewide analysis by another composite-likelihood method, one that considers regions of 5-10 sites rather than pairs or triplets of sites, predicts that hotspots occur every 30kb (Fearnhead and Smith, 2005). Crawford et al. (2004), using the PAC model described below, also concluded that substantial numbers of crossover hotspots appear to be occuring genome-wide. They found at least one hotspot, with estimated intensity ten times larger than crossover rates outside the hotspot, in 35 of 74 genes from the *SeattleSNPs* dataset.

The PAC Likelihood of Li and Stephens (2003) Another method that has shown success in estimating fine-scale rates of crossover is the PAC Likelihood of Li and Stephens (2003), which Wall (2004) notes has a smaller root mean square error for estimation of ρ than both the three-site and two-site methods when estimating

crossover rates alone. This model formulates an approximation to the likelihood $L(\rho)$ given a collection of haplotypes, considering all segregating sites in a sample region simultaneously rather than as pairs or triplets (see Chapter 2).

This approximation is subject to conditions similar to those of the previously mentioned methods, namely in the coalescent model it implicitly assumes. Specifically, the PAC model assumes that the chromosomes in the present-day sample have evolved under a constant-sized, randomly mating population with no selection. The infinite-sites assumption, as is shown later, is relaxed.

Their method has also been shown to estimate non-uniform ρ within loci with reasonable accuracy, a topic currently of hot interest in genetics. For example, applied to LD data in the MHC, the locations and intensities of the PAC model's estimated crossover hotspots closely corroborate with those found directly via sperm analysis (Crawford et al., 2004).

As simulations have shown that their method has a lower root mean square error for estimation of ρ than any of the other methods reviewed in this paper, perhaps the PAC model might be able to better estimate gene conversion rates as well. In this thesis, we incorporate gene conversion into the PAC model of Li and Stephens (2003) and investigate the accuracy of the resulting model for inference on rates of gene conversion and crossover. We furthermore aim to examine both the average relative rate of gene conversion to crossover, f, in the human genome and the variability in rates of these recombination processes across the human genome.

Towards this end, Chapter 2 considers estimating rates of gene conversion and crossover in a revised version of the PAC model, assuming constant rates of gene conversion and crossover as well as known haplotype information. Chapter 3 explores a bias in gene conversion estimation assessed via empirical studies and considers a means of correcting this observed bias. Chapter 4 considers estimating rates of gene conversion and crossover using phase-unknown genotype information. Chapter 5 explores variability in rates of recombination across the genome, using additional

statistical methodology. Chapter 6 presents a new coalescent-based, chromosome simulating program that incorporates both crossover and gene conversion hotspots. Chapter 7 summarizes our main findings and conclusions.

Chapter 2

MODELING CONSTANT RATES OF GENE CONVERSION WITH THE PAC LIKELIHOOD

In this chapter, we consider estimating constant rates of crossover and gene conversion over kilobase scales using population data. Population data consists of data collected on chromosomes of individuals randomly sampled from a population. Here we use SNP data. Each individual has two chromosomes representing a genetic region (paternal and maternal copies). A combination of SNPs along the same chromosome is referred to as a *haplotype*. In most current data collection protocols, one observes only the pair of SNP configurations at each site on a chromosome for an individual, i.e. the *genotype* information. Haplotypes typically have to be estimated from this information after data collection, via LD analysis or similar means. For this chapter, we will assume haplotypes are known without error. We will relax this assumption in Chapter 4.

First we update the model of Li and Stephens (2003) to jointly estimate rates of gene conversion and crossover using LD data. Next we explore some properties of this updated model and assess its accuracy in recombination rate estimation when applied to simulated datasets. Finally, we apply our model to the *SeattleSNPs* dataset to infer genome-wide rates of recombination in humans under various assumptions.

2.1 The PAC Likelihood with Crossover and Mutation

The PAC Likelihood of Li and Stephens (2003) is a model for estimating crossover rates using biallelic Single Nucelotide Polymorphism (SNP) population data. Let ρ be the population-scaled rate of crossover per unit physical distance. That is, $\rho = 4N_e r$,

where N_e is the effective population size, and r is the average probability of a crossover per unit physical distance per chromosome in a single transmission from parent to offspring. Since r is typically small, on the order of 1cM/Mb in humans, dividing this parameter by the physical distance in basepairs gives the crossover probability per basepair, r_{bp}, which can be appropriately scaled by $4N_e$ to the crossover rate per basepair, ρ_{bp}. Consider n haplotypes $h_1, ..., h_n$. The likelihood $\text{L}(\rho)$ can be broken down in the following manner:

$$\text{L}(\rho) = \Pr(h_1, ..., h_n \mid \rho) = \Pr(h_1 \mid \rho) \Pr(h_2 \mid h_1; \rho) ... \Pr(h_n \mid h_1, ..., h_{n-1}; \rho). \quad (2.1)$$

Thus if one can compute $\Pr(h_j \mid h_1, ..., h_{j-1}; \rho)$, for every $j = 1, ..., n$, (2.1) can be computed. This conditional probability will depend on assumptions regarding the processes affecting the evolution of the region under study, including selection, mutation, recombination, and population evolutionary history. Unfortunately, even under unrealistic simplifying assumptions on these processes, these probabilities cannot be computed. Therefore Li and Stephens proceed to *approximate* $\Pr(h_j \mid h_1, ..., h_{j-1}; \rho)$ and multiply these conditionals together to get an approximation of (2.1), called the Product of Approximate Conditionals (PAC). Their formulation assumes that $h_1, ..., h_n$ is a random sample from a population that has evolved with constant size, random mating, and no selection.

Li and Stephens (2003) assumes each new haplotype h_j is a mosaic of the previously observed haplotypes $h_1, ..., h_{j-1}$ as in Fig.2.1, which considers two distinct examples of h_4 conditional on observing $h_1, ..., h_3$. In this illustration, the first two SNPs (represented by white and black circles) of h_{4A} are considered most closely related in their ancestral history to h_1 and the next three SNPs to h_2. An important trait this model captures is that as the crossover rate increases, we expect the new haplotype to switch which previously observed haplotype it "copies" from more often (we will term these switches "jumps"), as is the case with h_{4B}. Thus if h_{4B} is observed,

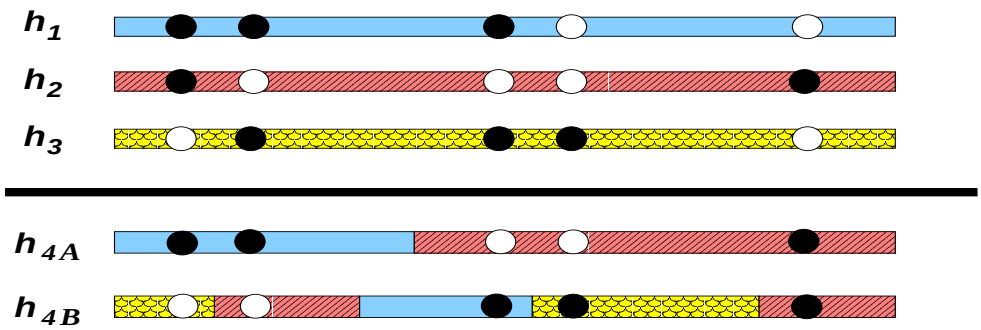


Figure 2.1: Two examples of newly observed haplotype h_4 as a mosiac of previously observed haplotypes h_1,h_2,h_3. Here there are five SNPs, represented by black/white circles, in the region. The expected number of "jumps" along h_4 from copying one haplotype to copying another increases as the rate of crossover in the region, ρ, increases. For this reason, h_{4B} might be representative of a higher underlying ρ than h_{4A} would be.

we might suspect the crossover rate is higher in a region than if we had observed h_{4A}. For the next conditional probability in the PAC model, $\Pr(h_5 \mid h_1, ..., h_4)$, h_4 is put into the collection of previously observed haplotypes, assigned its own distinct pattern or color, and h_5 is formed out of pieces of h_1,...,h_4.

This mosaic idea is easily captured by a Hidden Markov Model, shown in Fig.2.2. Here the hidden state sequence (the X_l's in Fig.2.2) is an indicator (color) of which haplotype $h_1, ..., h_j$ the new haplotype h_{j+1} copies at each site, $l = 1$,...,L, where L is the number of SNPs. The observed data are the SNP alleles (the S_l's in Fig.2.2), which in this example are of h_{j+1}. A key question is how to mathematically relate the notion of a "jump" to the underlying crossover rate ρ in a region. Based on heuristic arguments, Li and Stephens (2003) suggest that, conditional on observing j haplotypes, jumping from copying one previously observed haplotype to another in $h_1, ..., h_j$ while forming the $(j+1)^{\text{st}}$ haplotype occurs as a Poisson process along

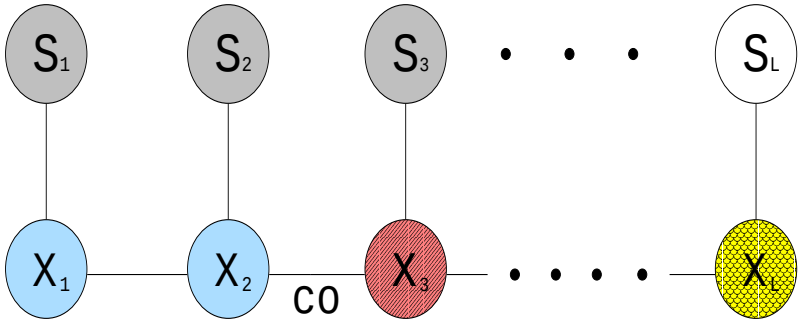


Figure 2.2: Illustration of the Hidden Markov Model of Li and Stephens (2003). The black and white $S_l, l = 1, ..., L$, represent the observed SNPs of new haplotype h_{j+1}. The colored and patterned X_l represent the unobserved indicator of which previously seen haplotype $h_1, ..., h_j$ is being "copied" at SNP l in generating h_{j+1}. The jump from copying the solid-color haplotype to the dash-line patterned haplotype inbetween X_2 and X_3 is indicative of crossover-like ("co") behaviour.

the chromosome with rate ρ/j per unit physical distance. Thus, conditioning on j observed haplotypes, the transition probabilities of the Hidden Markov Chain are as follows, based on the probability of a jump occuring or not between sites l and $l+1$:

$$\Pr(X_{l+1} = x_{l+1} | X_l = x_l) = \begin{cases} \exp(-\rho_{\text{bp}} \text{d}_l/j) + \Big(1 - \exp(-\rho_{\text{bp}} \text{d}_l/j)\Big)(1/j) & \text{if } x_{l+1} = x_l; \\ \Big(1 - \exp(-\rho_{\text{bp}} \text{d}_l/j)\Big)(1/j) & \text{otherwise,} \end{cases} \tag{2.2}$$

where d_l is the basepair distance between SNPs l and $l+1$. There is an equal probability of jumping to any given haplotype, including itself, given a jump occurs. We note that ρ_{bp} need not be constant across a genetic region but can represent a vector of values denoting fine-scale variablity in crossover rates in the region (i.e. $\vec{\rho} = (\rho_1, \rho_2, ..., \rho_{L-1})$ for L SNPs, where ρ_l is the crossover rate between SNPs l and $l+1$). For this chapter, we consider a constant rate of ρ_{bp} in a genetic region.

Using simulations, Li and Stephens (2003) show that this formulation works quite well but produces systematic biases in estimated values of ρ (a bias we similarly found after incorporating gene conversion in Sec.2.2, results omitted). They use these simulation results to suggest a slightly alternative formulation of equation (2.2), termed "PAC-B" in Li and Stephens (2003). These simulations assumed a constant-sized, randomly mating population free of selection.

In the original Li & Stephens implementation, the observed state sequence component of the Hidden Markov Chain, the probability of observing a SNP configuration given the haplotype you are copying, allows for "imperfect" copying that depends on a mutation parameter $\theta = 4N_e\mu$, with μ the probability of mutation per meiosis per site:

$$\Pr(h_{j+1,l} = a | X_l = x, h_1, ..., h_j) = \begin{cases} (1/2)\left[\tilde{\theta}/(j+\tilde{\theta})\right] + j/(j+\tilde{\theta}) & h_{x,l} = a; \\ (1/2)\left[\tilde{\theta}/(j+\tilde{\theta})\right] & h_{x,l} \neq a. \end{cases} \tag{2.3}$$

Here X_l refers to the previously observed haplotype h_1,...,h_j that h_{j+1} is copying at SNP l, $\tilde{\theta}$ refers to the Li and Stephens (2003) fixed choice to represent θ, and $h_{j,l}$ refers to the SNP allele of haplotype j at SNP l, of which there are two possibilities (e.g. "black" or "white") as the SNPs are biallelic.

Li and Stephens (2003) use $\tilde{\theta} = (\sum_{m=1}^{n-1} 1/m)^{-1}$, based on Watterson's estimate (Watterson, 1975), with n equal to the total number of haplotypes in the sample. This choice of $\tilde{\theta}$ gives *a priori* an expectation of one mutation event per site in the history of the sample but allows for repeat mutation. In the event of a mutation, the site mutates to either SNP allele (e.g. "black", "white") with equal probability.

In reality, of course, for a new haplotype h_{j+1}, it is uncertain which previously observed haplotype h_1,...,h_j, is copied at each site 1,...,L or whether the copying was imperfect. Thus the authors sum over all possibilities to achieve $\Pr(h_{j+1} \mid h_1, ..., h_j)$:

$$\Pr(h_{j+1} \mid h_1, ..., h_j) = \sum_{\vec{x}} \Big[\prod_{l=1}^{L} \Pr(h_{j+1,l} \mid X_l = x_l) \prod_{l=1}^{L-1} \Pr(X_{l+1} = x_{l+1} \mid X_l = x_l)(1/j) \Big], \tag{2.4}$$

where x_l is the lth element of the vector $\vec{x}$ that indicates the haplotype $h_1, ..., h_j$ copied at each SNP 1,...,L in haplotype h_{j+1}. This summation can be accomplished efficiently using the Baum-Welch algorithm. This is calculated for each j=1,...,$(n-1)$, and the resulting $\Pr(h_{j+1} \mid h_1, ..., h_j)$ are multiplied together to give $L(\rho)$. Li and Stephens (2003) use $\Pr(h_1) = \frac{1}{2^L}$.

A final comment is that the PAC likelihood depends on the order in which the haplotypes $h_1, ..., h_n$ are considered in the conditional probability. To deal with this, Li and Stephens (2003) suggest averaging over likelihood values from several randomly selected orders, and this is also the strategy we adopt here. For all analyses presented in this chapter, we averaged over estimates for ten random orders of the haplotypes.

2.2 Incorporating Gene Conversion into the PAC Likelihood

Gene conversion will also leave distinct marks on LD patterns amongst haplotypes that the mosaic idea of the PAC likelihood can ideally pick up. Figure 2.3 gives an example of what these patterns might look like. Again consider two distinct examples of h_4 conditional on observing $h_1, ..., h_3$. In this illustration, the majority of h_{4A} copies from h_1 (thus h_{4A} might be thought to be most closely related in its ancestral history to h_1) except for the middle SNP, which has been "replaced" by a SNP from h_2 at this location. Such a copying scheme resembles the pattern that might be expected following a gene conversion event. We will term such events – signatures of potential gene conversion – "replacements." Analagous to the crossover case, as the gene conversion rate increases in a region we would expect more such replacements to occur when generating the new haplotype from previously observed ones, as is the case with h_{4B}. Thus if h_{4B} is observed, we might suspect that the gene conversion rate is higher in the region than if we had observed h_{4A}.

Note that the gene conversion pattern looks similar to the pattern expected from two crossovers. However, a key difference is that with a "replacement" event, the haplotype copied before and after the event is forced to be the same, while with two "jumps," this need not be the case. Thus perhaps the most plausible explanation of such a pattern is a single "replacement" rather than two "jumps." When considering only the four haplotypes of Fig.2.3, it may be difficult to elucidate whether a relatively higher gene conversion rate or a relatively higher crossover rate best fits the data under our model. However, adding more haplotypes, one at a time, helps identify the most likely scenario. In particular, the remaining unique haplotypes, ones whose SNP patterns have not been seen in the previously observed haplotypes, provide further information. For example, if the majority of the remaining unique haplotypes can be pieced together in a manner similar to that of h_{4A} in Fig.2.3, it seems likely that the rate of gene conversion in the region is higher than the rate of crossover. Otherwise, if

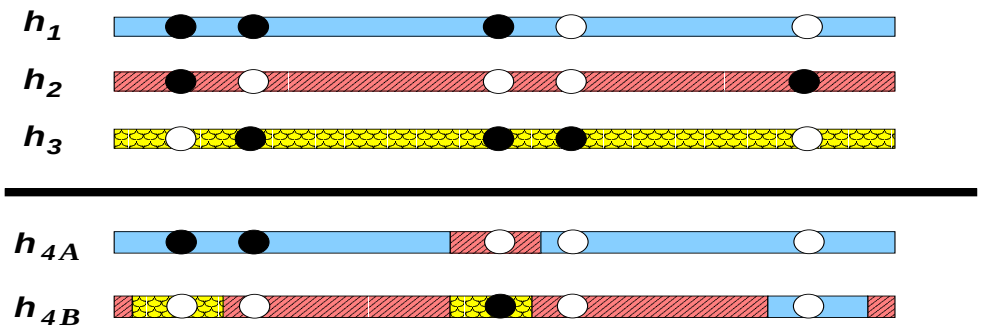


Figure 2.3: Two examples of newly observed haplotype h_4 as a mosiac of previously observed haplotypes h_1,h_2,h_3. The expected number of "replacements" along h_4, where copying one haplotype is interupted by briefly copying another, increases as the rate of gene conversion in the region, γ, increases. For this reason, h_{4B} might be representative of a higher underlying γ than h_{4A} would be.

the opposite were true, you would expect to see the majority of the remaining unique haplotypes pieced together in a manner similar to that of h_{4A} in Fig.2.1. However, we caution that identifiability does become a concern if either the rate of crossover or gene conversion is infinite. In practice, this could be an issue if either rate is "too large," though we have yet to find a "threshold" for which identifiability is a problem in all simulations we have used to test our model. The results for some of these simulations will be presented later in this chapter and the remainder of the thesis.

Perhaps the most obvious way to incorporate gene conversion into the PAC model would be to incorporate these "replacements" into the hidden state sequence $X_1, ...X_L$, as in Fig.2.4. However, capturing the features of these events, specifically recalling the haplotype we were copying from before the replacement began in addition to the haplotype we are currently copying, requires a second-order Markov Chain, which substantially increases the computional complexity. A naive implementation would

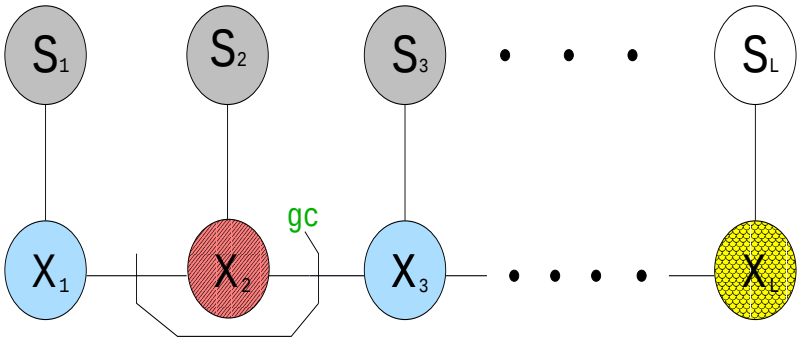


Figure 2.4: Incorporating gene conversion into the hidden state sequence. Here we copy from the solid-color haplotype until the gene conversion event occurs, between X_1 and X_2, at which point we copy a new dash-line patterned haplotype. When the gene conversion event ends, between X_2 and X_3, we return to copying the solid-color haplotype. Unlike our approach, this model allows a gene conversion event to span more than one marker but increases the computational complexity of the algorithm.

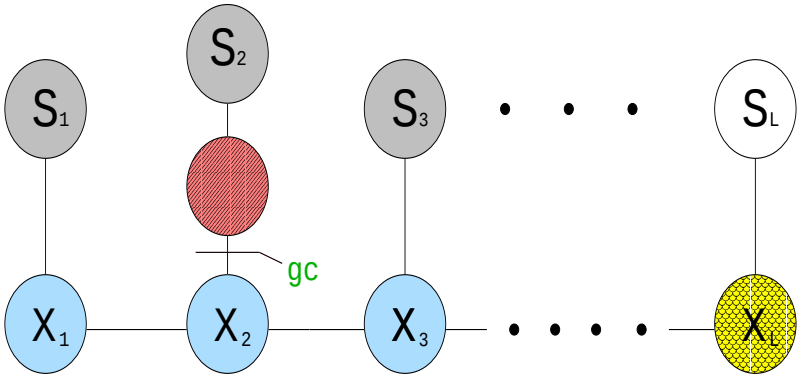


Figure 2.5: Incorporating gene conversion into the observed state sequence. Here gene conversion is assumed to affect at most one site, so that the hidden state sequence remains unchanged. For example, in this figure h_{j+1} copies from the solid-color haplotype at X_2, but does so "imperfectly," copying the dash-line patterned haplotype instead, signifying the replacement event. This formulation supposes the tract length of gene conversion events is less than the spacing between consecutive SNPs.

require computational complexity $o(j^4 L)$ for j haplotypes and L SNPs, compared to $o(jL)$ for the original PAC with no gene conversion.

We therefore took a different approach, based on assuming that replacement events affect at most one SNP. Under this assumption we can model replacements in the observed state sequence as in Fig.2.5, similar to the way mutation was modeled in Li and Stephens (2003). This assumption seems reasonable if the tract length of a gene conversion event is sufficiently smaller than the spacing between the SNPs of our data. This appears to be the case in humans based on the modest evidence on human tract lengths currently available. For example, Jeffreys and May (2004) suggest mean tract lengths in the MHC are perhaps 55-290bp, while fine-scale marker data are, on average, 300-1000bp apart, even for resequencing data. This alteration of PAC, which retains the same first-order Markov Chain as the no gene conversion model, requires only a trivial increase in computation time. Therefore we alter equation (2.3) of the HMM, the probability of observed SNP l in haplotype h_{j+1} given the previous haplotype $h_1, ..., h_j$ copied at l, to incorporate gene conversion:

$$\Pr(h_{j+1,l} = a | X_l = x, h_1, .., h_j) = \begin{cases} (1/2)\,[\tilde{\theta}/(j+\tilde{\theta})] \\ +[j/(j+\tilde{\theta})]\left((1-GC_l) + GC_l\, f_a\right) & h_{x,l} = a; \\ \\ (1/2)\,[\tilde{\theta}/(j+\tilde{\theta})] + [j/(j+\tilde{\theta})]\, GC_l\, f_a & h_{x,l} \neq a. \end{cases} \tag{2.5}$$

The probability of a replacement event affecting locus l is denoted by GC_l and $f_a = f_w 1_{[a=w]} + (1-f_w) 1_{[a \neq w]}$, where f_w is the proportion of SNP alleles that are "white" in $h_1, ..., h_j$ at locus l. This equates to allowing equal probability to copying each of the existing haplotypes at the given locus given a replacement has occured.

Replacement events, analagous to the "jumps" used to model crossover events in Li

and Stephens (2003), are assumed to occur as a Poisson process along the genome with rate γ/j per unit physical distance. Wherever a replacement initiates, the mechanism spreads over $t/2$ basepairs to the left and right, where t represents the basepair *tract length* of a gene conversion event. Thus a replacement event initiating to either the left or right will affect the locus l if the initiation occurs within $t/2$ basepairs of locus l. This gives GC_l the following form:

$$\begin{aligned} GC_l &= 1 - \exp\big(-\gamma_{\text{bp}} t/(2j)\big) \exp\big(-\gamma_{\text{bp}} t/(2j)\big) \\ &= 1 - \exp(-\gamma_{\text{bp}} t/j). \end{aligned} \tag{2.6}$$

We let $\gamma = 4N_e g$, where g represents the average gene conversion probability per unit physical distance per chromosome in a single transmission from parent to offspring. Since g is typically small, perhaps on the order of ≈1-10cM/Mb in humans, dividing it by the physical distance in basepairs gives the gene conversion probability per basepair, g_{bp}, which can be appropriately scaled by $4N_e$ to give the gene conversion rate per basepair, γ_{bp}, of equation (2.6). Analagous to ρ_{bp}, we note that γ_{bp} need not be constant across a genetic region but can represent a vector of values denoting fine-scale variablity in gene conversion rates in the region. For this chapter, we consider a constant rate of γ_{bp} in a genetic region.

2.3 Preliminary Investigations

The performance of the modified PAC Likelihood was tested against simulated data, using the ms simulator (Hudson, 2002). In the ms simulator, coalescent theory is used to generate a random sample of haplotypes from a population in the absense of selection. The ms simulator assumes an infinite sites mutation model; that is, each SNP in the sample is the result of a single mutation in the history of the sample. The exact simulation parameters used in ms are reported for each plot. In some instances, the likelihoods of several simulations were combined in order to get more accurate estimates of (ρ, γ). Each independent simulation of a dataset generated in such a

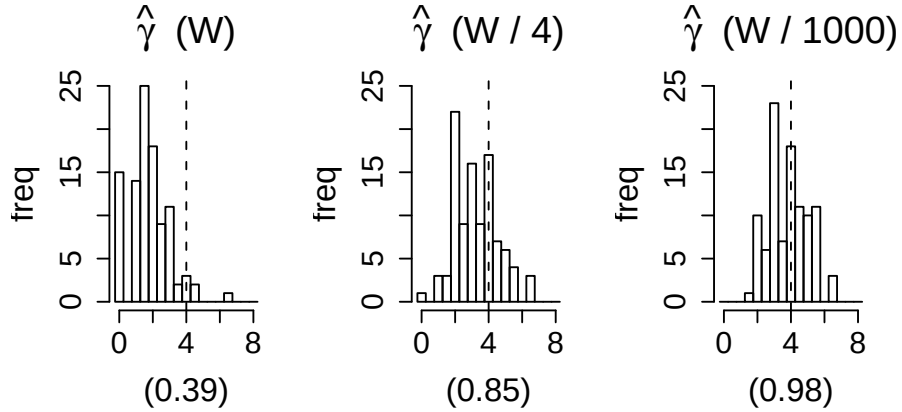


Figure 2.6: Example of mutation rate and gene conversion estimation confounding. The fixed estimate of $\tilde{\theta}$, the mutation parameter, decreases across columns as W,$W/4$, and $W/1000$. The true value of the gene conversion parameter γ (4.0/kb) is depicted by the dotted vertical line. The parentheses at the bottom of the plots represent the factor of times $\hat{\gamma}$ is within two of the truth, i.e. $2.0 \leq \hat{\gamma}/\text{kb} \leq 8.0$. Note how decreasing $\tilde{\theta}$ increases the accuracy of γ estimation.

manner is referred to here as a "locus." In past studies, accurate joint estimation of (ρ, γ) has proven difficult, and the combination of several independent loci has been necessary to achieve reasonable accuracy (e.g. Frisse et al. (2001), Ptak et al. (2004), Wall (2004)). In this section we combine information across multiple loci, assuming that rates of gene conversion are the same across loci and that the rates of crossover are the same across loci. For comparisons in which two different versions of the model consider the same simulated datasets, the same random orderings of haplotypes were used for each version.

2.3.1 Choice of $\tilde{\theta}$: Effect on Gene Conversion Rate Estimation

We found that the choice of mutation rate, $\tilde{\theta}$, influences the MLE estimate of gene conversion, $\hat{\gamma}$. This is not surprising, as both components attempt to capture patterns of "imperfect copying" in the observed state sequence of the HMM and are thus both

incorporated into equation (2.5). In particular, we found that the Li and Stephens (2003) choice of $\tilde{\theta}$ often resulted in a substantial underestimation of γ.

As the mutation parameter estimate $\tilde{\theta}$ of Li and Stephens (2003) is somewhat arbitrary, we decided to see how altering it might change the accuracy of γ estimates. We will refer to their fixed estimate of $\tilde{\theta}$ as W. We divided W by several factors and re-tested the revised PAC Likelihood model in simulated datasets. These simulations, which match those of Wall (2004), consist of 100 datasets, each dataset containing five independent loci with 50 haplotypes over a 5kb region, with $\rho = 1.0$/kb and $\gamma = 4.0$/kb, each constant across the region. A mutation parameter of 1.0/kb was used. Under an infinite sites model with a mutation parameter per physical distance θ, sequence length D, and haplotype number n, the expected number of sites can be calculated as $\theta D \sum_{i=1}^{n-1} 1/i$, which gives $\approx$22 here, or about 1 SNP every 225bp. The mean conversion tract was set to 125bp; this value was used in the model as well. The MLEs of ρ and γ were estimated over a grid of values based on Wall (2004) (see Sec.2.4).

Histograms of $\hat{\gamma}$ estimated in these simulated datasets are shown in Fig.2.6, for W, $W/4$, and $W/1000$, where $W = 0.223$ for these simulations. Note that the number of datasets where $\hat{\gamma} = 0$ drastically decreases as the fixed estimate $\tilde{\theta}$ decreases (from left to right in Fig.2.6). The right-most figure depicts $W/1000$, which is approximately equal to 0.0002 here. This is the value we now fix $\tilde{\theta}$ at for subsequent use of this model; i.e. we use $\tilde{\theta} = 0.0002$. Effectively, this very small value of $\tilde{\theta}$ can be thought of as approximating $\tilde{\theta} \to 0$, or assuming infinite sites. Thus any recurrent mutation will most likely be explained as gene conversion by our model, though recurrent mutation appears to occur substantially less often than gene conversion, rendering such an assumption reasonable.

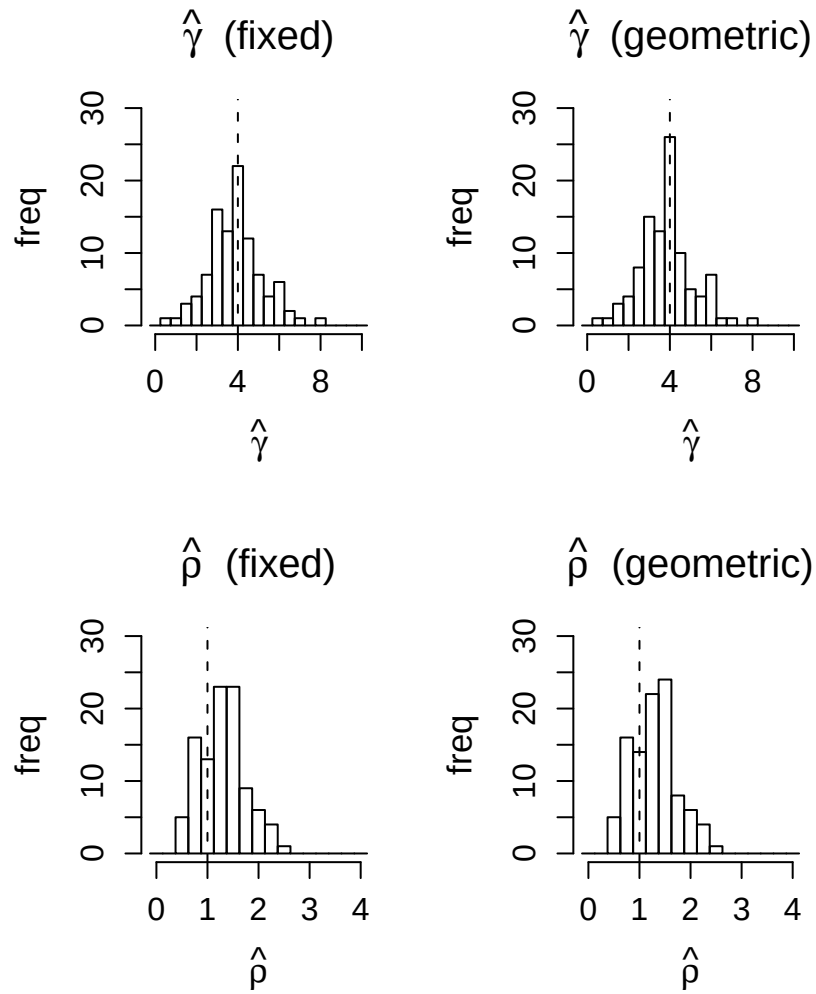


Figure 2.7: Comparing the subtle differences of changing the fixed tract length to a geometrically distributed one. We considered some of the simulated datasets described in Sec.2.3.1. We estimated γ and ρ over a 21-point, equally spaced ρ/kb grid on [0.0,4.0] and a 51-point, equally spaced γ/kb grid on [0.0,10.0]. The vertical dotted lines represent true values. The top row depicts a histogram of MLE estimates of γ; the bottom row depicts a histogram of MLE estimates of ρ. The columns depict estimation using a fixed tract length, left, and a geometric tract length, right. Note the inference is nearly identical for each of γ and ρ across columns.

2.3.2 Geometrically Distributed Tract Lengths

Equation (2.6) assumes that the tract length t is fixed. There is some evidence that the tract length of a gene conversion is geometrically distributed with some fixed mean length t^* (Hilliker et al., 1994). Such a distribution has thus been subsequently used by several researchers (e.g. Betran et al. (1997), Wiuf and Hein (2000), Frisse et al. (2001), Ptak et al. (2004), Wall (2004)). Following this precedent, we modified (2.6) to allow for a geometrically distributed tract length in the following manner:

$$\begin{aligned} GC_l &= 1 - \Pr(\text{no } GC_{\text{left}}) \Pr(\text{no } GC_{\text{right}}) \\ &= 1 - \Pr(\text{no } GC_{\text{left}}) \textstyle\sum_{t=1}^{\infty} \Pr(\text{no } GC_{\text{right}}|t) \Pr(t) \\ &= 1 - \Pr(\text{no } GC_{\text{left}}) \big[q \exp(-\gamma_{\text{bp}}/(2j))\big] / \big[1 - \exp(-\gamma_{\text{bp}}/(2j)) + q \exp(-\gamma_{\text{bp}}/(2j))\big] \\ &= 1 - \Big(\big[q \exp(-\gamma_{\text{bp}}/(2j))\big] / \big[1 - \exp(-\gamma_{\text{bp}}/(2j)) + q \exp(-\gamma_{\text{bp}}/(2j))\big] \Big)^2 . \end{aligned} \tag{2.7}$$

Here q $\equiv 1/t^*$.

However, simulation results (e.g. Fig.2.7) showed that there was little difference between using (2.6) versus (2.7), i.e. assuming a fixed tract length versus assuming a geometrically distributed tract. Therefore, for reasons of simplicity and interpretability, we recommend using the version in (2.6). One nice feature of using a fixed tract is that doubling the tract length gives exactly half the estimated gene conversion rate.

2.3.3 Conversion Tract Length Restrictions

One disadvantage to the way gene conversion has been incorporated into this model is that the method, not surprisingly, performs poorly when the tract length t is larger than the average spacing between markers. The model outlined implicitly assumes, for the computational convenience of avoiding a second order Markov chain, that a "replacement" event will affect at most one site. The effect of a simulated conversion

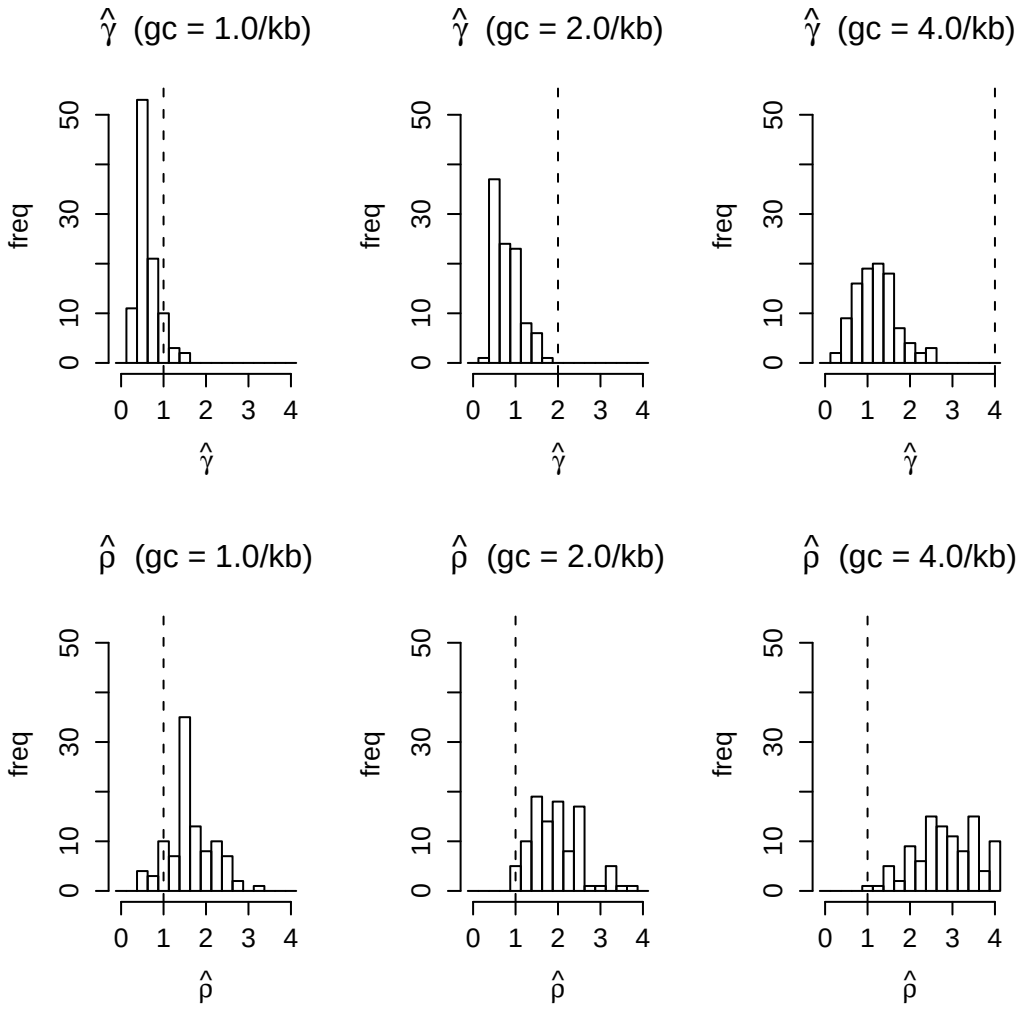


Figure 2.8: The effect of having a tract length considerably larger than marker spacing. The top row depicts a histogram of MLE estimates of γ; the bottom row depicts a histogram of MLE estimates of ρ. The columns depict true values of γ=1.0, 2.0, and 4.0/kb, respectively. The vertical dashed lines represent the true simulated values of γ and ρ. Note that γ is underestimated and ρ overestimated, the magnitude of these effects increases as the true value of γ increases.

tract length that does not follow this assumption is displayed in Fig.2.8. These simulations consisted of 50 datasets, each of which contained five independent loci of 50 haplotypes over a 5kb region, with $\rho = 1.0$/kb and $\gamma = 1.0$, 2.0, and 4.0/kb. The mutation parameter was set to 1.0/kb, giving about 1 SNP per 225bp. The mean tract length was considerably larger at 500bp, which is in accordance to what Ardlie et al. (2001) assumed for their mean tract when they found evidence of gene conversion in data with 47 human subjects (otherwise the simulations mimic those of Wall (2004)). The tract length was fixed at 500bp in the model. The MLEs of ρ and γ were estimated over a 21-point, equally spaced ρ/kb grid on [0.0,4.0] and a 51-point, equally spaced γ/kb grid on [0.0,10.0].

Under these parameters, many of the gene conversion events will affect more than one SNP. As our model assumes gene conversions can affect only one SNP at a time, such events will likely be interpreted as two crossovers. Thus the MLE estimate of γ is biased downwards and the MLE estimate of ρ is biased upwards, this effect increasing under this scenario as γ increases.

However, this assumption that a gene conversion event affects at most one site should not prove the model too restrictive in most data scenarios. Even for resequencing, data marker densities are generally on the order of one per 500 bp for common SNPs, and, despite what Ardlie et al. (2001) assumed for a tract length, current evidence of gene conversion tract lengths seem to suggest tracts are shorter, particularly in humans. For example, Zangenberg et al. (1995) found in certain loci of the MHC that many gene conversion events "encompassed only one polymorphic region," suggesting potentially very short tracts perhaps as small as 12bp. Of course, tracts as small as this would render the effects of gene conversion negligible in most population data sets. On average, tract lengths are probably longer. Jeffreys and May (2004), conducting sperm analysis studies on two subjects, found conversion tracts to have a mean length of about 55-290 bp. They suggest that "the only reasonably clear evidence for long conversion tracts [e.g. 1 kilobase or so] in humans comes from

rare germline reversions of triplet repeat expansions." However, application of our model to some regions of the genome, such as the MHC, which is known to have an atypically high SNP density, may lead to dubious inference, unless SNPs are thinned in some random manner.

2.3.4 Simulations with Population Expansion and Structure

The previous simulations were conducted assuming a constant population size throughout time. In truth, the demographic histories of human populations are likely quite a bit more complex (e.g. Kruglyak (1999), Pritchard and Przeworski (2001)). To get an idea of how demography and structure that depart from our assumptions might affect our estimation, we tried a few different simulation scenarios. Specifically, we present two expansion and two migration scenarios modeled after those of Pritchard and Przeworski (2001). Each of these demographic scenarios was previously considered by Pritchard and Przeworski (2001) and Li and Stephens (2003). We shift our primary focus in this section to estimating $f(\equiv \gamma/\rho)$, in accordance with previous literature (e.g. Frisse et al. (2001), Ptak et al. (2004), Wall (2004)). As each of ρ $(= 4N_e r)$ and γ $(= 4N_e g)$ are defined assuming a constant, single population size N_e, it is unclear how to define the true value of each under these alternative scenarios. However, one might suspect the ratio f to be somewhat robust to changes in N_e, as it should divide out to g/r.

We performed ms (Hudson, 2002) simulations with values of $f = 0$, 1, and 10 for each population model. These values represent no gene conversion, an equal amount of gene conversion and crossover, and a substantially larger amount of gene conversion compared to crossover, respectively. For all scenarios, ρ and f were estimated over a 51-point, equally spaced ρ/kb grid on [0,2] and a 21-point, equally spaced f grid on [0,20]. These simulations consist of 100 datasets (each comprised of 5 independent loci). Each locus had 50 haplotypes and 50 segregating sites over 25kb. These values were chosen to somewhat mimic the Li and Stephens (2003) robustness simulations.

We used $\rho = 0.4$/kb; this value of ρ coincides with a genome-wide average estimate for humans, with an estimated sex-averaged crossover probability of 1%/Mb (i.e. 1cM/Mb) and an estimated N_e of 10^4 (see, e.g., Hudson (2001), Przeworski and Wall (2001), Ardlie et al. (2001), or Pritchard and Przeworski (2001) and references therein). For simulations where $\gamma \neq 0$, the mean tract length was simulated and fixed in our model as 200bp. The SNP density in these simulations is ≈1 SNP per 500bp, comparable to typical resequencing data.

We chose two expansion scenarios, both of which were done to match the simulations of Pritchard and Przeworski (2001), which assumes a current population size, N_0, of 10^5 expanded some time T ago. We take T in units of generations, where a single generation is ≈20 years. They determine via simulation a growth rate α that matches, on average, the number of segregating sites under a constant-size population growth scenario with $N_e = 10^4$, the estimated N_e for humans. For one case, we chose $T = 500$ generations, giving $\tilde{\alpha} = 1960$. This corresponds to an expansive population growth from approximately 8,600 to 100,000 over 500 generations. For the second case, we chose $T = 5000$ generations, giving $\hat{\alpha} = 350$, suggesting growth from ≈ 1,200 to 100,000 over 5000 generations. It should be noted that $\rho = 4.0$/kb was used for these particular simulations, as ms takes $\rho = 4N_0r$ as input, where $N_0 = 10^5$ in the expansion models, while PAC estimates the effective-size rate of crossover, $\rho = 4N_er$, with $N_e = 10^4$.

Two migration procedures were chosen to mimic Li and Stephens (2003), the parameters again coming from Pritchard and Przeworski (2001). In one, the "two-island even" migration scenario, we sample equally from two distinct subpopulations, 25 haplotypes from each, that have a migration rate between them of 4.0, $= 4N_em$, where m is the fraction of migrants in each subpopulation in each generation. In the other migration scenario, the "two-island uneven" migration scenario, we sample from only one of the two subpopulations, but the migration parameter remains the same. The "two-island even" scenario corresponds to equal sampling from two distinct

lineages; for example, pooling the African-American and European individuals in the *SeattleSNP* data. The "two-island uneven" scenario corresponds to the homogeneous population sampled experiencing periodic effects of intermixing from another distinct population, i.e. *admixture*. It should be noted that $\rho = 0.2/\text{kb}$ was used for these particular simulations, as ms takes $\rho = 4N_s r$ as input for migration scenarios, where N_s is the sub-population size, while PAC estimates the effective-size rate of crossover, $\rho = 4N_e r$, where N_e is the effective size of both populations combined.

Table 2.1: Population growth and structure simulation summaries. The parentheses in the Median column represent 2.5% and 97.5% quantiles.

Scenario	Median	Mean	$\%(\hat{f} \leq 5)$
STANDARD			
f = 0	4.0(2.0-8.0)	4.06	0.82
f = 1	4.0(2.0-9.0)	4.68	0.72
f = 10	9.0(6.0 - 16.5)	9.92	0.02
EXPANSION, T = 500			
f = 0	3.0(2.0-6.5)	3.66	0.92
f = 1	4.0(2.0-7.0)	4.48	0.77
f = 10	9.0(5.0 - 18.0)	9.84	0.06
EXPANSION, T = 5000			
f = 0	7.0(1.0-20.0)	8.52	0.35
f = 1	8.0(2.0-20.0)	9.92	0.26
f = 10	14.0(5.0 - 20.0)	14.11	0.06
MIGRATION, 2 IS EVEN			
f = 0	4.0(2.0-8.5)	4.41	0.81
f = 1	4.0(2.0-10.0)	4.84	0.67
f = 10	10.0(5.5 - 19.0)	10.93	0.03
MIGRATION, 2 IS UNEVEN			
f = 0	5.0(2.0-10.5)	5.20	0.63
f = 1	6.0(3.0-13.0)	6.45	0.45
f = 10	12.0(5.5 - 19.0)	11.83	0.03

The results of these simulations are shown in Table 2.1. Standard coalescent simulations (a constant-sized, randomly mating population, the same assumptions assumed in our model) were run to provide a "gold standard." Distinguishing between $f = 0$ and $f = 1$ is not possible with only five loci, though either can perhaps be distinguished from $f = 10$. The 95% confidence intervals from simulations overlap for $f = 0$, 1 and $f = 10$, but if one chooses an "interesting" cut-point, e.g. $\hat{f} = 5$ as we use here, appreciable differences are noticeable. For instance, the proportion of $\hat{f} \leq 5$ is 82% for $f = 0$ and 72% for $f = 1$, compared to only 2% for $f = 10$. Thus while we are not likely to gauge f estimates very precisely, we do have some ability to distinguish "high" from "low" levels of gene conversion relative to crossover with five independent loci.

Our conclusions appear to not alter under the $T = 500$ expansion and the even migration scenario. Median values of $\hat{f}$ across simulated datasets where f=0 and f=1 are rather distinct from median values of $\hat{f}$ across simulated datasets where $f = 10$. This again suggests that we have power to distinguish between high and low relative rates of gene conversion to crossover under these demographic models.

For the $T = 5000$ expansion, estimation of f was considerably less precise than under the standard model. Population expansions are thought to decrease LD, which would result in artificially high recombination rate estimates, which is certainly the case here. Both $\hat{\rho}$ and $\hat{\gamma}$ are biased upwards; $\hat{\gamma}$ is affected more, resulting in higher $\hat{f}$ overall (results omitted). However, this appears to occur only in the slower, longer, $T = 5000$ expansion case. Thus accuracy of our method perhaps depends on the true underlying expansion scenario. Still there is some power to see a difference between f=0,1 and $f = 10$, as evidenced by the different medians. Confidence intervals have a large overlap, but perhaps this wouldn't be as evident if our f grid had extended beyond $f = 20$.

Under the uneven migration scenario, ρ estimates are biased slightly downwards, though the majority are still within a factor 2 of the truth (results omitted). As

$\hat{\gamma}$ remained similar to that of the standard model (results omitted), this results in slightly less accurate estimation of f than under the standard model. This effect on $\hat{\rho}$ is expected under such demography, as subpopulation stratification tends to increase LD, though it is unclear why the effect appears stronger here than in the even migration scenario. Regardless, median values of $\hat{f}$ across simulated datasets where f=0 and f=1 are quite different from median values of $\hat{f}$ across simulated datasets where $f = 10$, suggesting we have some power to distinguish between high and low relative rates of gene conversion to crossover in this demographic model as well.

2.4 Comparisons to Other Methods

Frisse et al. (2001) developed a method for jointly estimating (ρ, f) using the composite 2-site likelihood approach of Hudson (2001). For each pair of sites in a region, a likelihood is calculated using coalescent simulations. These coalescent simulations assume that the haplotypes in the data are samples from a constant-sized, randomly mating population of infinite sites, i.e. the *standard coalescent* model. The likelihoods of all pairs of sites in a region are multiplied together, creating a "composite likelihood." Wall (2004) extended this to using composite 3-site likelihoods instead of 2-site likelihoods. Both Frisse et al. (2001) and Wall (2004) assume ρ and f to be constant across genes, though their methods have been extended to allow ρ to vary among genes by Ptak et al. (2004) and Wall (2004), respectively. In this section, we will compare our method to those of Frisse et al. (2001) and Wall (2004) using simulations from a standard coalescent model.

Wall (2004) generated simulations using the standard coalescent framework, with 50 haplotypes, 5kb of sequence, $f = \gamma/\rho = 4.0$, $\rho = 1.0$/kb, and mean tract length = 125 bp. He used a mutation rate of 1.0/kb in the region, which gives ≈1 SNP per 225bp. We created simulations using Hudson (2002)'s ms with the same parameters in order to compare our method to his, specifically Wall's Figure 3, which also contains

the results of the 2-site likelihood method of Frisse et al. (2001). We have attempted to tabulate the histogram results from this figure for Frisse et al. (2001) and Wall (2004) in Table 2.2. Following Wall (2004), we obtain approximate MLEs $\hat{\rho}$ and $\hat{f}$ by computing likelihoods on a 2-D grid, with ρ values = (0.0,0.2,...,3.8,4.0/kb) and f values = (0,1,1.4,2,2.8,4,5.6,8,11.2,16). While a direct comparison might not be appropriate because we do not use the exact simulated datasets of Wall (2004), this should provide a rough idea of how the methods compare. For all three methods, the tract length was fixed to the "true" value of 125bp.

Table 2.2: Tabulation of revised PAC results and Fig.3 of Wall (2004). The values represent the percentage of $\hat{f}$ within a factor of 2 of the truth (i.e. $2 \leq \hat{f} \leq 8$). Simulations for the Frisse et al. (2001) and Wall (2004) estimates were done by Wall (2004) and simulations for the PAC estimates were run by the authors.

Method	1 locus	5 loci	20 loci
2-site (Frisse et al., 2001)	0.36	0.67	0.92
3-site (Wall, 2004)	0.50	0.66	0.80
revised PAC	0.42	0.83	0.94

In these simulations, all three methods appear to have similar accuracy for estimating f. For a single locus, all three methods perform poorly. For comparison, note that with the f grid used here, there is a 50% chance of getting within a factor of two of the true f via random guessing. All three methods do only as well or worse.

Accuracy for all three methods raises appreciably when data from five independent loci are combined. The revised PAC provides estimates of f within a factor two of the truth 83% of the time, compared to 67% and 66% for Frisse et al. (2001) and Wall (2004), respectively. With 20 independent loci, we're within a factor of two of the truth 94% of the time, similar to the other two methods.

We also examined accuracy of joint estimation of (ρ,f). To directly compare to Wall (2004), we counted the fraction of datasets in which *both* the f and ρ estimates

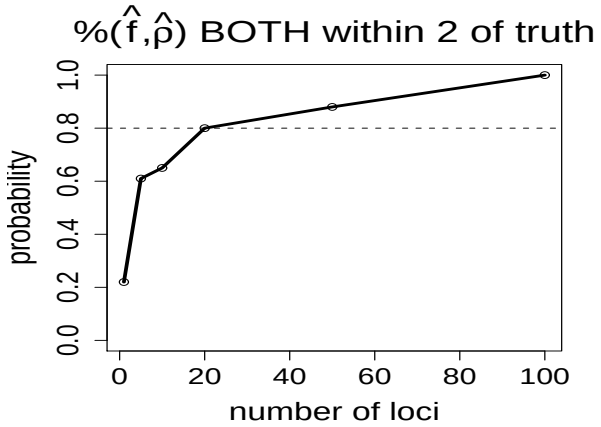


Figure 2.9: Testing the probability of jointly estimating (ρ, f) each within (and *not* including) a factor two of their true values. It takes the combined likelihoods of 20 independent loci to achieve this aim for 80% of the simulations. Comparable to Wall (2004) Fig.5. (Note: we used 100 datasets for the combined likelihoods of 1, 5, or 10 loci, and 50 datasets for the combined likelihoods of 20, 50, or 100 loci.)

were within, and *not* including, a factor of two of the truth, i.e. $2.0 < \hat{f} < 8.0$ and $2.5 < \hat{\rho} < 10.0$. Considering Figure 5 in Wall (2004), we see the method of Wall (2004) takes ≈10-20 loci to have a 50% chance of estimating ρ and f within these guidelines. Comparing this to our identically plotted Fig.2.9, we see we meet the same restrictions 80% of the time with 20 loci. In fact, we are over 50% after only 5 loci.

2.5 Application to *SeattleSNPs* Dataset

2.5.1 Analysis

We applied our model to data from the *SeattleSNPs* project (website: *http://pga.gs.washington.edu*) to estimate average genome-wide f in humans. *SeattleSNPs* is an on-going project sequencing genes in 24 African-American and 23 European (CEPH) subjects. We have a dataset of 184 of these genes, in which haplotypes

have been previously estimated by another lab using PHASE (Stephens et al. (2001), Stephens and Donnelly (2003)) on the combined data. For the preliminary analysis we present here, we make a number of simplifying assumptions: (1) the haplotype estimates are correct, (2) f is constant within and among genes, and (3) ρ is constant within genes, though it is allowed to vary among them. To reduce computational burden, singleton SNPs were removed. The mean tract length of gene conversion events was set to 100bp, based on the Jeffreys and May (2004) MHC sperm analysis results. We estimated our parameters using a 100-point, equally spaced $\log_{10} f$ grid on [-0.5,3.5] and a 100-point, equally spaced $\log_{10}(\rho/\text{kb})$ grid on [-2.0,2.0]. For each value of f and vector of ρ values across genes, $\vec{\rho}$, we calculate the likelihood $L(f, \vec{\rho})$ as:

$$L(f, \vec{\rho}) = \prod_i L_i(f, \rho_i), \tag{2.8}$$

where $L_i(f, \rho_i)$ is the likelihood calculated at f and ρ_i for gene i. We maximize (2.8) over f and $\vec{\rho}$. We denote the value of f that maximizes (2.8) as $\hat{f}$. This is the same "profile" method used in Ptak et al. (2004) and Wall (2004).

The scaled profile log-likelihood for f is shown in Fig.2.10, for both African-Americans and Europeans. Our estimates of f are quite close for African-Americans and Europeans, at $\hat{f} = 6.8$ and 7.4, respectively. Our African-American f estimate coincides closely to what Ptak et al. (2004) found. Though they estimated $\hat{f}$ to be 1 and 0.25 for the African-Americans and Europeans, respectively, they used a tract length of 500bp. Since their analysis, experimentation has suggested 100bp to be a perhaps more reasonable value for humans (e.g. Jeffreys and May (2004)). At any rate, we expect our estimates to be roughly five times larger than that of Ptak et al. (2004). Dividing our estimates by five gives estimated f of 1.4 and 1.5 for the African-Americans and Europeans, respectively. The European estimates of f are discrepant by a factor of six between our method and that of Ptak et al. (2004). In some sense, it

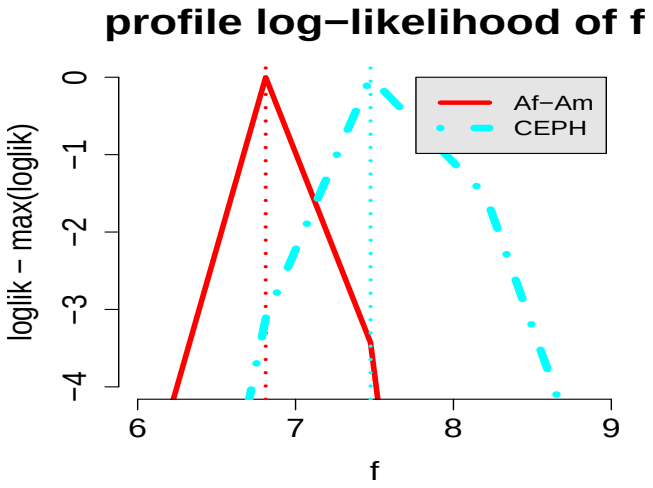


Figure 2.10: The scaled log-likelihood of f, based on maximizing ρ per locus for each grid point of f, for the *SeattleSNPs* dataset, for African-American (red solid line) and CEPH (blue dot-and-dash line) individuals. The vertical lines represent the MLEs. This likelihood is calculated over a grid of $\log_{10} f$ values, resulting in the observed lack of smoothness.

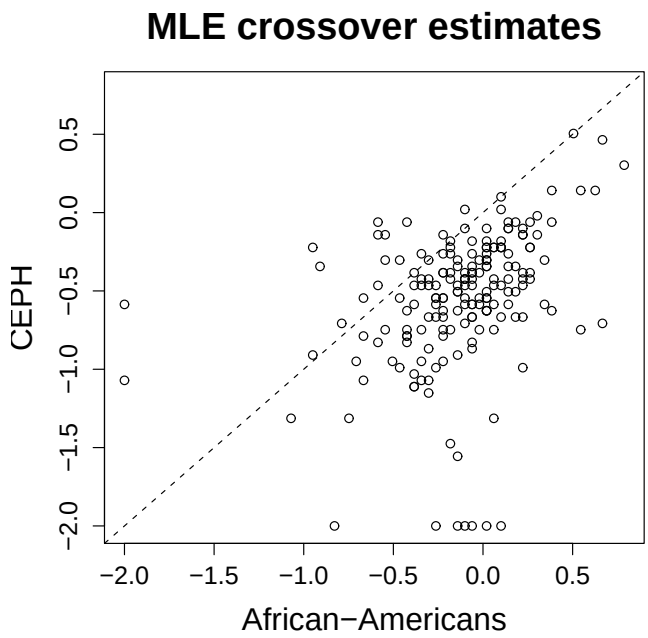


Figure 2.11: Estimated ρ for CEPH (y-axis) vs African-Americans (x-axis) across 184 genes of *SeattleSNPs*, based on the profile likelihood approach of Sec.2.5. Values of $\hat{\rho}$/kb are on the $\log_{10}$ scale. The dotted line represents y=x. Though each population was examined separately, their estimates are correlated, with African-American $\hat{\rho}$ generally higher than that for CEPH, due to a larger N_e in the African-Americans.

is comforting that our f estimates are close between the African-American and CEPH populations, as we discuss below. Frisse et al. (2001), assuming ρ to be constant across the genome, which recent evidence suggests is unlikely (e.g. Crawford et al. (2004), McVean et al. (2004)), had somewhat larger estimates of f, at 4-25 for a tract length of 500bp. Though they deliberately selected loci in a manner to fit the assumption of constant ρ as closely as possible, perhaps their results are still influenced by variable ρ among loci. In addition, Frisse et al. (2001) used information from only 15 Hausa individuals genotyped at ten 2-kb loci, which is less data than either us or Ptak et al. (2004).

It is comforting that our estimates of f overlap for the two populations. African-

Americans and Europeans share much of their evolutionary history and – as f should not depend on the different effective population sizes N_e between populations – we would expect similar results. Also encouragingly, estimated ρ across genes were correlated between the two populations (Fig.2.11), with African-Americans typically having higher values than the Europeans. The typically larger values of ρ for the African-Americans likely reflects a larger effective population size than European populations.

2.5.2 Discussion

Of course, our analysis makes several simplifying assumptions. One we know to be unrealistic is fixing ρ within genes. In truth, there is substantial evidence of fine-scale crossover rate varation within small regions (e.g. McVean et al. (2004), Crawford et al. (2004), Jeffreys et al. (2001)). In addition, Ptak et al. (2004) noted a substantial bias in their gene conversion estimation in the presence of genotyping error. Genotyping error may be around 0.5% in the *SeattleSNPs* dataset (Ptak et al., 2004). Furthermore, the effect of repeat mutation on LD patterns can mimic gene conversion and impact the accuracy of f estimation. As methylated CpG groups are known mutational hotspots (Frisse et al. (2001), and references therein), perhaps removing all CpG regions and reanalyzing our data is a reasonable thing to do. We will address each of these possibilities in a future application of our model to the *SeattleSNPs* dataset, discussed in Chapter 5, but we do not do so here.

2.6 Summary

In this chapter, we have presented an extension of the PAC Likelihood of Li and Stephens (2003) to jointly estimate crossover and gene conversion, much like Frisse et al. (2001), Ptak et al. (2004), and Wall (2004) extended the composite likelihood approach of Hudson (2001) to jointly estimate the two. We found that our estimation of γ, the gene conversion rate, is influenced by our choice of $\tilde{\theta}$, the mutation rate

estimate. We have thus fixed $\tilde{\theta}$ at 0.0002, which appears to provide the most accurate estimates of γ. When the tract length in a region is larger than the average spacing between markers, simulations suggest $\hat{\gamma}$ will underestimate rates of gene conversion and $\hat{\rho}$ will overestimate rates of crossover. For regions in which the tract length is less than the average spacing between markers, our simulations suggest our model performs as well as or better than composite-likelihood methods that consider joint crossover and gene conversion estimation (Frisse et al. (2001), Wall (2004)). We have shown by simulations that our model is robust to a variety of demographic scenarios that depart from our model's assumptions that our sample is derived from a constant-sized, randomly mating population.

One potential advantage of our method over those that use composite likelihoods is that our method provides a true likelihood. Therefore we might get a rough estimate of the likelihood curvature, perhaps providing better estimates of uncertainty. In addition, our likelihood allows for the potential advantage of using direct Bayesian inference that does not involve likelihood penalizing techniques necessary for using such inference with composite likelihoods (e.g. McVean et al. (2004)). In Chapter 5, we will outline such a Bayesian method. In addition, the methods of Frisse et al. (2001), Ptak et al. (2004), and Wall (2004) calculate likelihood values over grids of ρ and f. Our method, which does not require the use of coalescent simulations to calculate likelihoods for ρ and f, can calculate likelihood values over a continuous scale of ρ and f values, e.g. in an MCMC framework as discussed in Chapter 4.

Chapter 3

EXPLORING BIAS IN GENE CONVERSION ESTIMATION

Via simulation studies, Li and Stephens (2003) found some systematic bias in estimation of ρ in their PAC model. Specifically they found that the bias decreased with SNP spacing and depended on the number of SNPs and haplotypes. They corrected this bias based on the results for simulated datasets. Noting that our estimator for gene conversion performs poorly for analysis of single $\approx$20 SNP regions (see Sec.2.4), we decided to investigate single locus analysis for potential bias as well.

First we use simulations with a wide range of parameters to characterize the bias in gene conversion rate estimation. Next we correct this bias based on the simulation results and assess the bias-corrected model's performance on independent simulations. Finally, we re-analyze the *SeattleSNPs* dataset with the bias-corrected model, noting the key differences between this new analysis and the analysis of Chapter 2. In addition, we explore some issues that may be affecting our conclusions about recombination rates in the genes of *SeattleSNPs*.

As with Chapter 2, we deal only with phase-known haplotype data in this chapter. We use ten random orderings of the haplotypes for all analyses presented in this chapter. A mean tract length of 100bp was used for all simulations, and the fixed value of 100bp was used in the model for γ estimation.

3.1 Characterization of Bias via Standard Coalescent Simulation

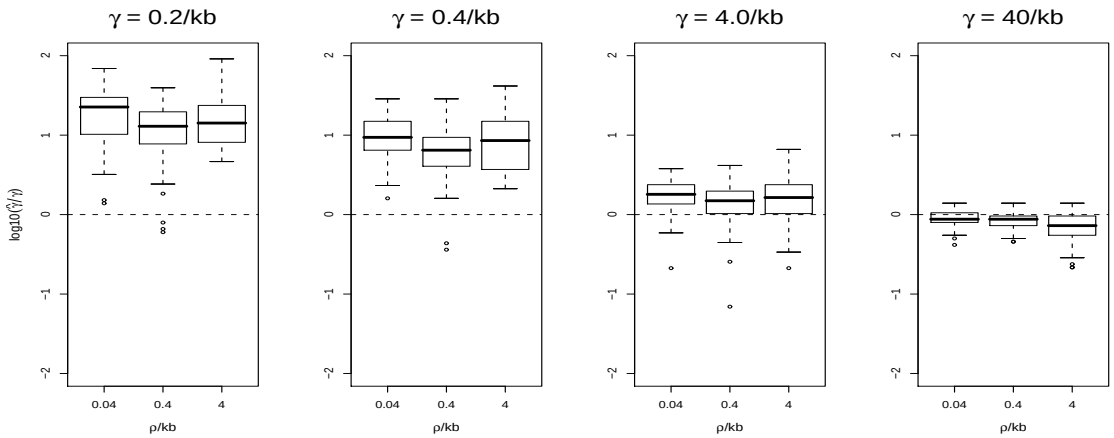
As with the Li and Stephens (2003) bias-correction procedure, we investigated bias in estimates by simulating chromosomes from a constant-sized, randomly mating popu-

lation with an infinite sites mutation model undergoing no selection. We used the ms simulator (Hudson, 2002) to do so.

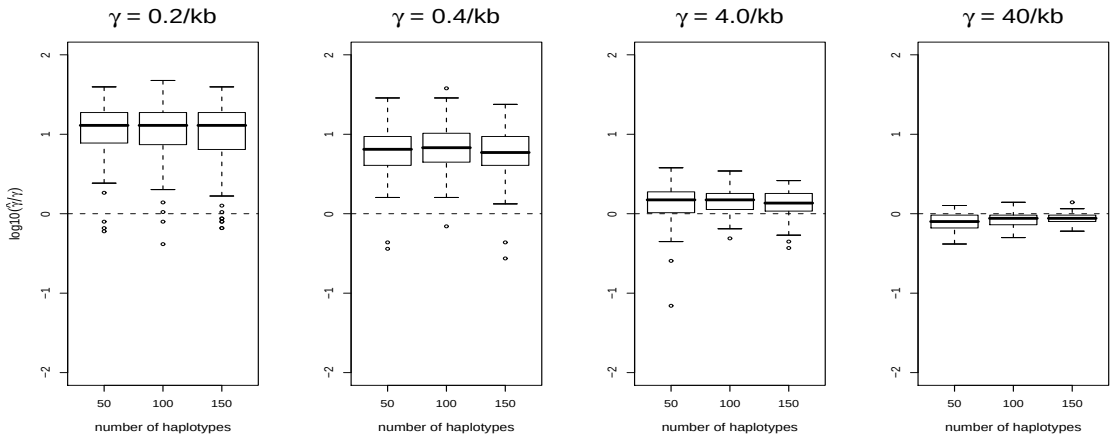
We ran some preliminary simulations to determine which parameters seem to most affect γ estimation. First we examine how bias in estimates of γ depend on the true value of γ, the true value of ρ, and the number of sampled haplotypes. These simulations consisted of 50 segregating sites over 50kb of sequence. The MLEs of ρ and γ were found over a 100-point equally-spaced grid of estimates for $\log_{10}(\rho/\text{kb})$ on [-2.4,1.6], and a 100-point equally-spaced grid of estimates for $\log_{10} f$ on [-0.5,...,3.5]. As in Li and Stephens (2003), we examined bias on the log scale, by considering boxplots of $\log_{10}(\hat{\gamma}/\gamma)$.

Results of some of these initial simulations are shown in Figure 3.1. The most notable trend is that our model tends to consistently overestimate γ when the true γ is small. For large values of gene conversion, i.e. $\gamma \geq 4$/kb or so, assuming a tract of 100bp, our model appears to provide reliable estimation. These trends, and the magnitude of the bias, seem largely independent of both the true value of ρ (Fig.3.1a) and the number of haplotypes in the sample, n (Fig.3.1b). For subsequent study of the bias, we focused on $\rho = 0.4$/kb and samples of $n = 50$ haplotypes.

We considered the possibility that our choice of $\tilde{\theta}$, the mutation rate, might be contributing to the observed bias. We tried three different choices of $\tilde{\theta}$, to see how each affects single locus estimation of γ across a more comprehensive range of true values than was considered in Sec.2.3.1. Specifically, we considered the original Li and Stephens (2003) $\tilde{\theta}$, equal to Watterson's estimate, $W = (\sum_{m=1}^{n-1} 1/m)^{-1}$ for n haplotypes, in addition to $\tilde{\theta} = W/5$ and $\tilde{\theta} = W/1000 \approx 0.0002$, the latter being our previous choice of $\tilde{\theta}$ from Sec. 2.3.1. We simulated with $\rho = 0.4$/kb, $n = 50$ haplotypes, and $\gamma = 0.2$,0.4,4.0,16.0 and 40.0/kb, running each simulated dataset in PAC with $\tilde{\theta}$ fixed at W/k for $k = 1$, 5, and 1000. We used the same $\log_{10} f$,$\log_{10} \rho$ grid as before for estimation. The results are shown in Fig.3.2. From Fig.3.2, it is unclear which option is "best" to use, as there are different patterns of bias across



(a) $\log_{10}(\hat{\gamma}/\gamma)$ for varying ρ



(b) $\log_{10}(\hat{\gamma}/\gamma)$ for varying number of haplotypes

Figure 3.1: Investigating the bias in $\hat{\gamma}$ for simulations with (a) varying ρ and (b) varying number of haplotypes. Shown are $\log_{10}(\hat{\gamma}/\gamma)$ for $\gamma = 0.2, 0.4, 4.0$, and 40.0/kb, from left to right. Each plot in (a) has results for $\rho = 0.04, 0.4$, and 40.0/kb (n=50). Each plot in (b) has results for $n = 50$, 100, and 150 (ρ=0.4/kb). Note that there appears to be a substantial positive bias in $\hat{\gamma}$ for lower values of true γ and that this bias appears independent of true values of ρ and the number of haplotypes n.

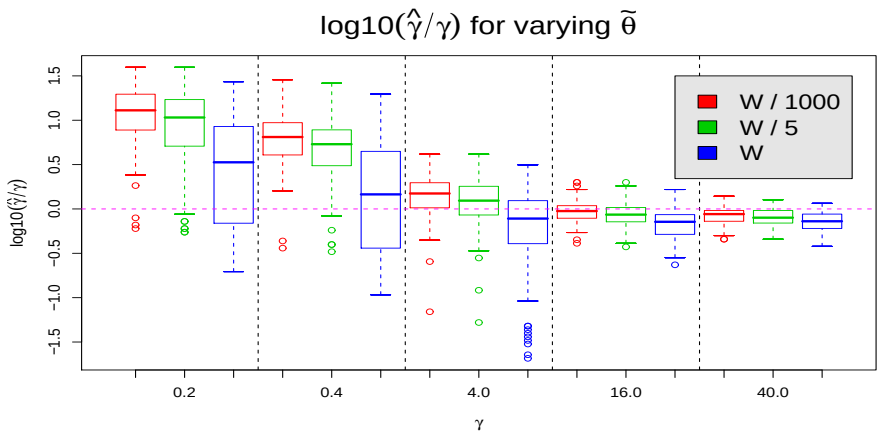


Figure 3.2: Investigating the bias in $\hat{\gamma}$ for simulations with varying $\tilde{\theta}$. Shown are $\log_{10}(\hat{\gamma}/\gamma)$ for γ = 0.2,0.4,4.0,16.0 and 40.0/kb, from left to right. Between each consecutive vertical line are the results for $\tilde{\theta} = W/1000$ (red), $W/5$ (green), and W (blue), where W is equal to Watterson's estimate, from left to right. Note that while the choice of $\tilde{\theta}$ somewhat affects $\hat{\gamma}$ bias, it is not clear which $\tilde{\theta}$ might be the "best" to use.

the different choices of $\tilde{\theta}$. However, in the region where estimates are least variable, i.e. large γ, the use of $\tilde{\theta} = W/1000 \approx 0.0002$ is least biased. In addition, using $\tilde{\theta} = 0.0002$ has a nice interpretability as an assumption of essentially no repeat mutation. For these reasons, we continue to use $\tilde{\theta} = 0.0002$.

The magnitude of the bias in estimates of small γ appears to depend on the number of SNPs, L, in the sample. We ran a large number of simulations with $\rho = 0.4$/kb, $n = 50$, and $L = 20$, 50, 100, and 150 in a 50kb region. (20 is the minimum number of SNPs we suspect is necessary for reliable inference, and we note the trend does not differ much between $L = 100$ and $L = 150$, suggesting it probably will not change much beyond 150 SNPs.) For each simulated dataset, we estimated ρ and γ using the same $\log_{10}\rho$, $\log_{10} f$ grid as before. In Figure 3.3, we plot the median of $\hat{\gamma}$ across 100 datasets for each true value of simulated γ. This plot highlights the tendancy for smaller values of γ to be substantially over-estimated, regardless of SNP number.

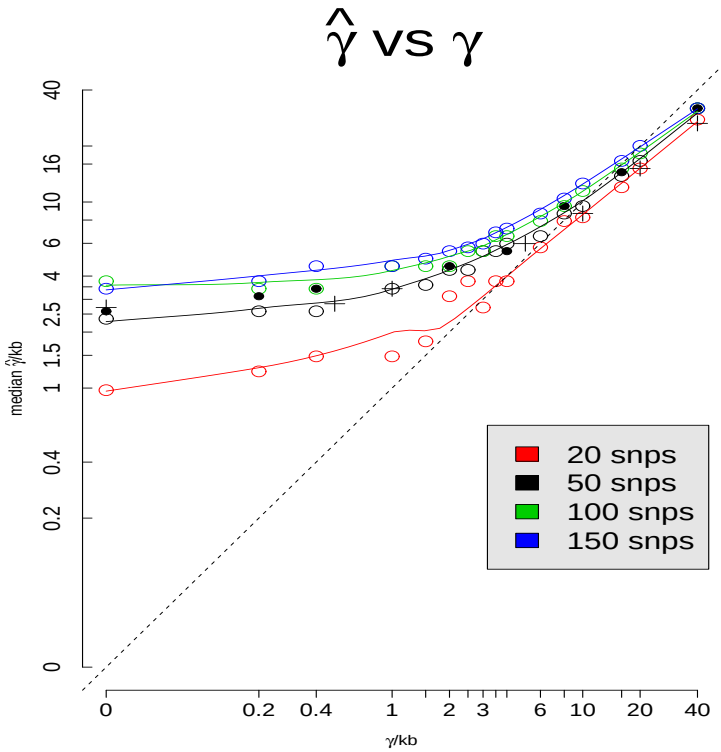


Figure 3.3: Investigating bias in $\hat{\gamma}$ for simulations with varying SNP size. Shown are median values of $\hat{\gamma}$ over 100 datasets of several SNP sizes plotted against the true values of γ. Estimates are spaced on the $\log_{10}$ scale. As $\log_{10} 0$ is undefined, we arbitrarily chose $\log_{10} \gamma$ = -1.5 to represent $\gamma = 0$. Note that the principal trend of the bias is to substantially over-estimate small values of γ. The lines are loess-smooth curves for each SNP size. For 50 SNPs, we tried various SNP densities, of ≈1 per 200bp ("+"), ≈1 per 500bp ("•"), and ≈1 per 1000bp ("∘"). While there may be an interaction with SNP density and SNP number, it does not appear to be too substantial in these simulations.

Furthermore, the magnitude of over-estimation becomes larger as L increases.

As simulated regions for each SNP size were 50kb in length, we tried to distinguish if the main determinant of the magnitude of overestimation of small values of γ might be SNP density and not SNP size. In particular the 20, 50, 100, and 150 SNPs represent $\approx$1 SNP per 2500, 1000, 500, and 333bp, respectively. Using 50 SNPs, we simulated 100 regions of size 10kb, which have $\approx$1 SNP per 200bp, and 100 regions of size 25kb, which have $\approx$1 SNP per 500bp, for various values of γ to compare with the original 50kb simulations, which have $\approx$1 SNP per 1000bp. The results of $\hat{\gamma}$ do not appear to depend much on SNP density (see Fig.3.3). Thus we consider only SNP number, and not density, in our bias correction of Sec.3.2 below.

3.2 Correction of Bias

To deal with this bias, we used the loess-smoothed curves illustrated in Fig.3.3. Specifically, we match a proposed value of γ with the appropriate median $\hat{\gamma}$ output by our model from Fig.3.3, according to four different SNP number bins. For regions with number of SNPs, L, such that $L \leq 35$, $35 < L \leq 75$, $75 < L \leq 125$, and $L > 125$, we use the curves in Fig.3.3 for "20 SNPs," "50 SNPs," "100 SNPs," and "150 SNPs," respectively. (Actually this correction is applied to the compound parameter γt, where t is the assumed tract length of the gene conversion events.)

This loess-correction, characterized using the simulations of Section 3.1, was applied to independent simulations of the same parameters, using the same $\log_{10} \rho$,$\log_{10} f$ grid as before. The results are shown in Fig.3.4. Again the median values of $\hat{\gamma}$ across 100 simulated datasets are plotted for each true value of γ. Essentially the correction flattens the likelihood curve of the PAC model to reflect the uncertainty of our model in estimating mid to low-range values of γ/kb. The effect of flattening the curve is illustrated in Fig.3.5, which shows boxplots of $\log_{10}(\hat{\gamma}/\gamma)$ across 100 simulated datasets of size 50 SNPs, for each true value of γ, for simulations before and after the loess-correction. Clearly there is substantially more variance in $\hat{\gamma}$ across simulated datasets

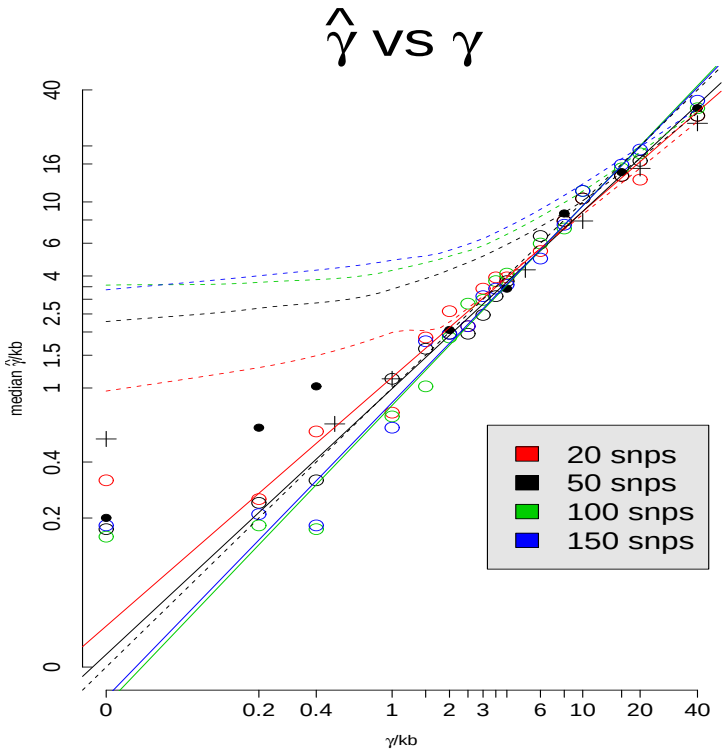


Figure 3.4: Using the loess-correction for $\hat{\gamma}$ on simulations with varying SNP size. Shown are median values of $\hat{\gamma}$ over 100 datasets for several SNP sizes plotted against the true values of γ. Estimates are spaced on the $\log_{10}$ scale, with the exception of $\gamma = 0$. The solid lines represent best-fit linear regression lines for estimating median $\hat{\gamma}$ on γ. As $\log_{10} 0$ is undefined, we omitted it from the regressions. The loess-correction appears to work quite well on eliminating the bias in the remaining values of γ. The dotted lines are the original, pre loess-corrected curves for each SNP size. For 50 SNPs, the correction applied to datasets with variable SNP densities (see Fig.3.3) of ≈1 per 200bp ("+"), ≈1 per 500bp ("•"), and ≈1 per 1000bp ("∘") are shown. For $\gamma = 0.2, 0.4$/kb in the ≈1 per 500bp simulations ("•"), the bias is not fully corrected, suggesting a more complicated interaction with SNP size and SNP density might be appropriate. However, this remaining bias does not appear to be too substantial, as (1) for, e.g., $\gamma = 0.4$/kb, the remaining absolute bias for the ≈1 per 500bp, 50 SNP simulations ("•") is not much larger than that of 100 and 150 SNPs (green/blue "∘"), and (2) the variance was large across simulations for $\gamma = 0.2, 0.4$/kb in the ≈1 per 500bp simulations ("•") and do include the true γ within the interquartile range (results omitted).

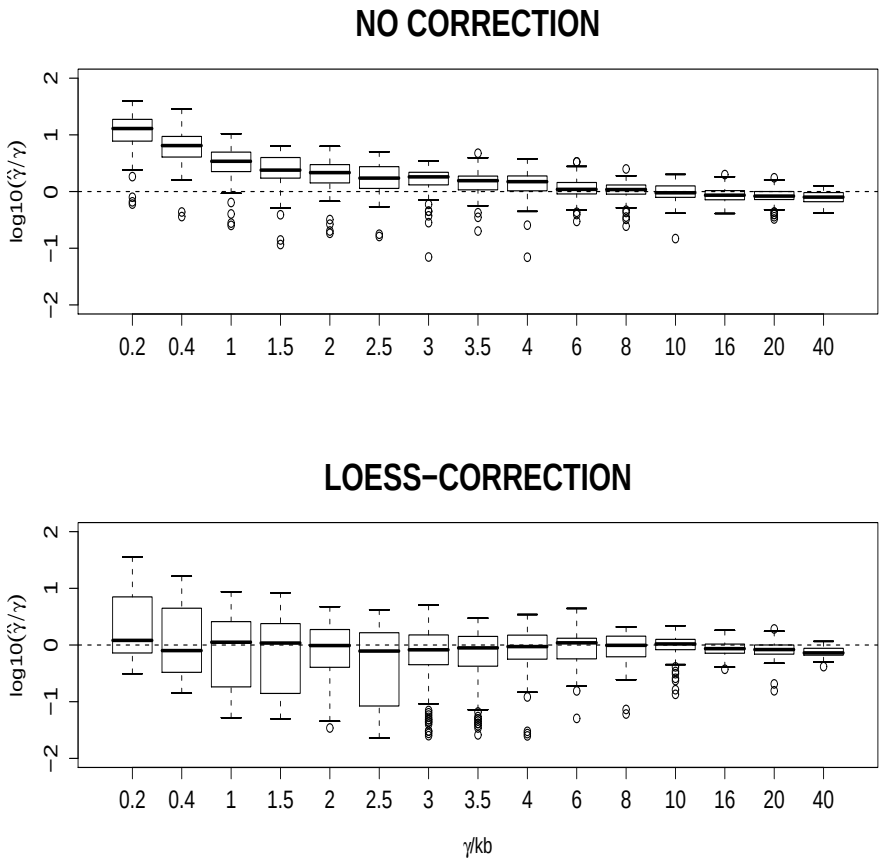


Figure 3.5: Simulation results of $\log_{10}(\hat{\gamma}/\gamma)$ for 100 datasets for several true values of γ, using PAC before (top) and after (bottom) the loess-correction, for 50 SNPs. The loess-correction increases the model's uncertainty in estimation for small to mid-range values of γ, which helps to eliminate the bias apparent in the top picture.

after the loess-correction for smaller to medium values of γ, helping to eliminate the bias.

3.3 Assessment of Loess-corrected PAC Model

3.3.1 Simulations Under Standard Coalescent

We applied the loess-corrected version to simulations described in Sec.2.3.4. These simulations included datasets with $f = 0$, 1, and 10. We tabulate how well the loess-corrected model distinguishes between these different simulated values of f in Table 3.1.

As one might hope, estimation of f appears to be more accurate here than in the pre-corrected model for the standard coalescent simulations, comparing Table 3.1 to Table 2.1. For instance, median and mean values of $\hat{f}$ are closer to the truth for $f =$ 0 and 1, corresponding to $\gamma = 0.0$ and 0.4/kb, respectively, here than they were in the Chapter 2 simulations. There is slightly more of a tendancy to underestimate $f =$ 10 than was the case with the pre-corrected model; this is likely a result of increasing uncertainty. Still the difference for $f = 10$ between Table 3.1 and Table 2.1 appears small considering the general difficulty in obtaining precise estimates of f.

3.3.2 Simulations with Population Expansion and Structure

As this empirical bias was corrected using simulations under the standard coalescent model, we wanted to see how it holds up against departures from this specific model. The pre-corrected PAC model was found to be fairly robust to a variety of demographic scenarios (see Sec.2.3.4). We applied the loess-corrected version to these same demographic scenarios, which include two population expansion and two admixture models based on Pritchard and Przeworski (2001) (see Sec.2.3.4 for simulation details). The results for the loess-corrected model are shown in Table 3.1.

Table 3.1: Population growth and structure simulation summaries for 100 datasets – after loess-correction. The parentheses in the Median column represent 2.5% and 97.5% quantiles. Each dataset uses the combined likelihoods of five independent simulated loci.

Scenario	Median	Mean	$\%(\hat{f} \leq 5)$
STANDARD			
f = 0	1.0(1.0-3.5)	1.33	1.00
f = 1	1.0(1.0-5.0)	1.93	0.98
f = 10	8.0(4.0 - 15.0)	8.55	0.12
EXPANSION, T = 500			
f = 0	1.0(1.0-4.0)	1.42	1.00
f = 1	4.0(1.0-5.0)	4.85	0.99
f = 10	9.0(3.5 - 16.0)	9.27	0.11
EXPANSION, T = 5000			
f = 0	6.0(1.0-20.0)	7.68	0.39
f = 1	8.0(1.0-20.0)	9.15	0.33
f = 10	13.0(4.5 - 20.0)	13.42	0.06
MIGRATION, 2 IS EVEN			
f = 0	1.0(1.0-3.5)	1.32	0.99
f = 1	1.0(1.0-4.0)	1.58	1.00
f = 10	9.0(4.0 - 18.5)	9.62	0.07
MIGRATION, 2 IS UNEVEN			
f = 0	1.0(1.0-4.0)	1.29	0.99
f = 1	1.0(1.0-5.5)	1.79	0.97
f = 10	8.0(2.5 - 18.1)	8.41	0.19

As was the case prior to the bias correction, the results of our loess-corrected model appear to be quite robust to the rapid expansion, i.e. $T = 500$, and even migration scenarios. Indeed the loess-corrected model appears to also be robust to the uneven migration scenario, which did not appear to be true for the pre-corrected PAC model. For both versions of the PAC model, performance under the slow expansion scenario, i.e. $T = 5000$, is less accurate, though with both versions there is still modest ability to distinguish high rates of f from low ones.

Therefore, performing a loess correction explicitly based on the standard coalescent model does not seem to make the model any less useful for application to data sets formed under different demographic models.

3.4 Re-analysis of *SeattleSNPs* Dataset

We applied the loess-corrected version of our PAC model to 183 genes in the African-American and CEPH individuals of the *SeattleSNPs* dataset in the same manner as described in Sec.2.5, to see if this revised model results in different inference about genome-wide average f in humans. The primary concern is that estimates of f in both poplations may be biased upwards, as our simulation studies here using the pre-corrected model suggest $\hat{\gamma}$ can show substantial positive bias for small values of γ.

The scaled profile log-likelihood for f is shown in Fig.3.6, for both African-Americans and Europeans, comparable to Fig. 2.10 before. Our estimates of f are substantially reduced for both the African-American and European populations from before, with $\hat{f} = 3.6$ (previously 6.8) and 1.7 (previously 7.4), respectively. Though these new estimates of f are still modestly similar between populations, the profile log-likelihoods somewhat surprisingly no longer overlap within 2 log-likelihood units. Given the assumptions of our model, some of which we know to be untrue, this probably more illustrates the limitations of our approach than provides evidence of a real variation in f across populations. (On the other hand, some features of recombina-

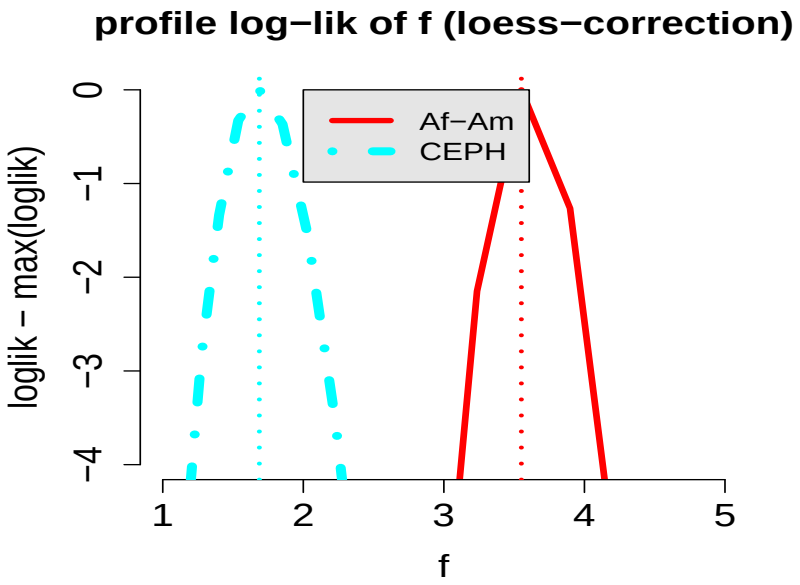


Figure 3.6: The scaled profile log-likelihood of f, based on maximizing ρ per locus for each grid point of f, for the *SeattleSNPs* dataset, for African-American (red solid line) and CEPH (blue dot-and-dash line) individuals, using the loess-corrected version of PAC described in this chapter. The vertical lines represent the MLEs. This likelihood is over a grid of $\log_{10} f$ values, resulting in the observed lack of smoothness.

tion such as crossover hotspots appear to evolve quickly (e.g. Wall et al. (2003), Ptak et al. (2004), Winckler et al. (2005)) and perhaps vary across human populations (Fearnhead and Smith, 2005), so that it is not inconceivable this might also be the case with f.)

Despite this lack of consistency between populations after applying the loess-corrected PAC model, the results are more similar than the pre-corrected PAC model to the composite-likelihood estimates of Ptak et al. (2004). Dividing our estimates by five to adjust for tract length discrepancies as in Sec. 2.5, we get $\hat{f} = 0.72$ and 0.34 for the African-Americans and Europeans, respectively. These $\hat{f}$ estimates are less than a factor of two different from the Ptak et al. (2004) estimates of 1 and 0.25, respectively. Again our estimates of genome-wide f considerably differ from those of Frisse et al. (2001), whom estimate f to be 4-25, using less data and some different assumptions (see Sec. 2.5).

As was the case prior to the loess-correction, estimated ρ across genes were correlated between the two populations (results omitted).

3.4.1 Robustness to Outlying Genes

A potential problem in interpreting the results of a profile likelihood is that one or a few genes may be dominating our inference on f. In theory this is not a problem if the model is correct, but the PAC model is an approximation and we know some of its assumptions to be false. Indeed f may not be constant across the genome. If f does vary, we might want to avoid outlying genes that are atypical of the majority and driving inference, as we are intending to estimate a genome-wide average f that reflects the majority of genes. To see if a relative few genes are indeed driving our profile-likelihood estimate of f, we looked at the scaled profile log-likelihoods of $\log_{10} f$ for 50 randomly selected genes from each population, shown in Figure 3.7. That is, we plot $l(f, \hat{\rho}_i^f)$ for gene i, i = 1,...,50, where $l(f, \rho)$ is the log-PAC value for (f,ρ) and $\hat{\rho}_i^f$ is the ρ that maximizes $l(f, \rho)$ in gene i for a particular grid value of f. For

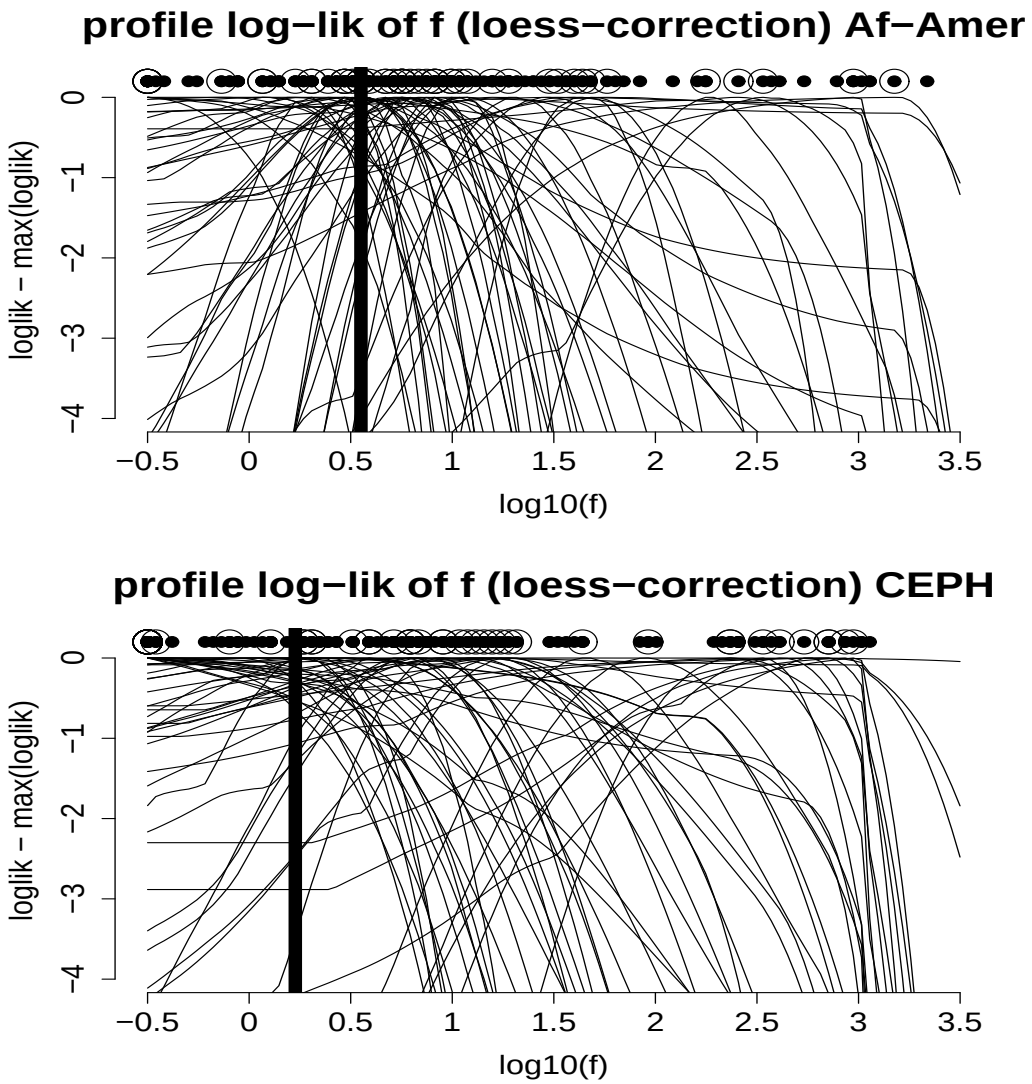


Figure 3.7: Scaled profile log-likelihoods of $\log_{10} f$ for 50 randomly selected genes for the African-American (top row) and CEPH (bottom row) individuals of *SeattleSNPs*, using the loess-corrected version of PAC. Open circles ("◯") represent the MLEs of these 50 genes for each population; closed circles ("•") represent the MLEs of all genes for each population. The vertical lines represent the overall profile likelihood MLE of f. Note that the overall MLE seems small for both populations considering the number of genes with $\hat{f}$ greater than this overall MLE.

each population, it seems the majority of genes have $\hat{f}$ greater than the overall profile likelihood MLE. This suggests that some or several genes with smaller $\hat{f}$ might have very peaked likelihoods.

Figure 3.8 considers the scaled profile log-likelihoods of $\log_{10} f$ for all genes whose $\hat{f}$=0.32, the lowest f grid value. Note that for some of these genes, the log-likelihood values rapidly decrease for larger values of f, particularly in the CEPH individuals. This suggests genes with small $\hat{f}$ may indeed be strongly affecting inference on the profile-likelihood estimate of f by substantially lowering the average likelihood value over all genes at large grid values of f. Therefore, as an alternative, we removed genes whose $\hat{f}$ were "atypically" large or small in Table 3.2 to see how inference in f changes. It appears that the inference for the CEPH population is particularly sensitive. Figure 3.9 shows the profile log-likelihood for genes whose $\hat{f}$ were in the middle 80% only, i.e. removing genes whose $\hat{f}$ were in the highest 10% of all genes and genes whose $\hat{f}$ were in the lowest 10% of all genes. Note that the European estimate of f has changed substantially, while the African-American estimate has not. Note also that the profile log-likelihoods now overlap for the two populations, as was the case in Sec.2.5.

Table 3.2: Profile log-likelihood estimates of f for genes whose profile log-likelihood $\hat{f}$ are in the middle P percent.

P	Af-Amer	CEPH
100	3.55	1.69
90	3.90	3.55
80	3.90	3.55
70	4.69	2.95
60	4.69	2.45
50	5.15	2.45

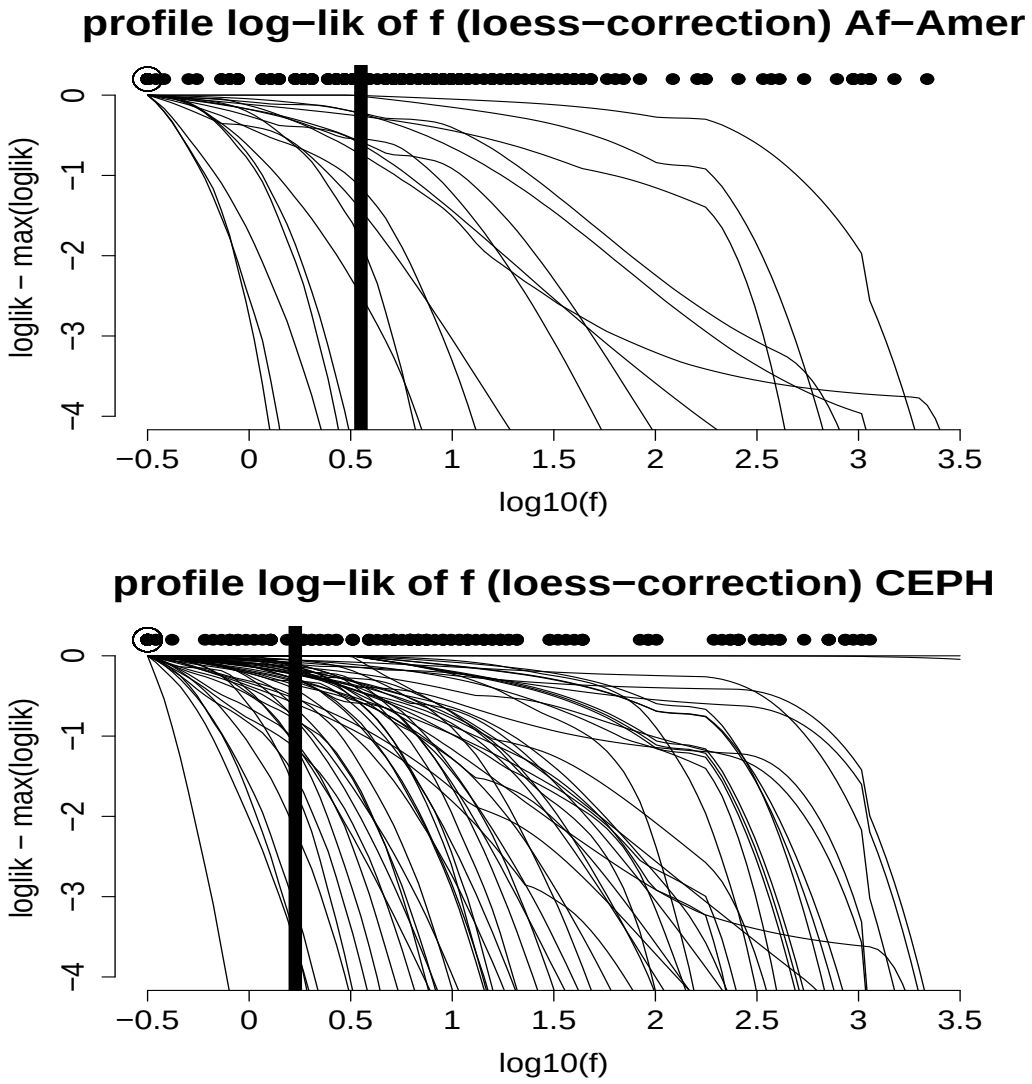


Figure 3.8: Scaled profile log-likelihoods of $\log_{10} f$ for genes with $\hat{f}$=0.32, the lowest allowed value, for the African-American (top row) and CEPH (bottom row) individuals of *SeattleSNPs*, using the loess-corrected version of PAC. Open circles ("◯") represents f=0.32; closed circles ("•") represent the MLEs of all genes for each population. The vertical lines represent the overall profile likelihood MLE of f. Note that some of these genes have very peaked profile log-likelihoods. For example, the left-most log-likelihood of the CEPH individuals has a very sharp decline, which might act to radically influence the overall profile likelihood f MLE.

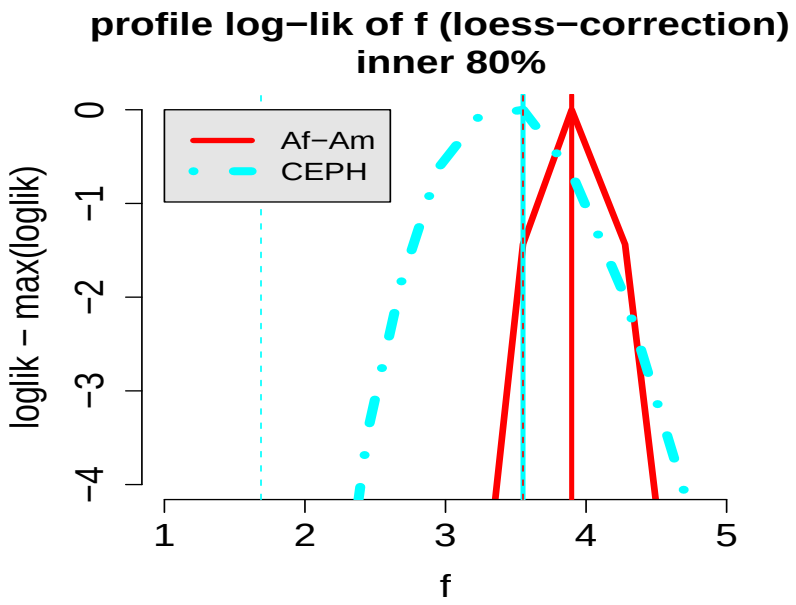


Figure 3.9: The scaled profile log-likelihood of f for African-American (red solid line) and CEPH (blue dot-and-dash line) individuals, using the loess-corrected version of PAC described in this chapter, after removing genes whose $\hat{f}$ was in the highest 10% or lowest 10% of all genes' $\hat{f}$. The solid vertical lines represent the profile likelihood f MLEs for these genes. The dotted vertical lines represent the profile likelihood f MLEs using all genes. Note the considerable change in f MLE for the CEPH data. This likelihood is over a grid of $\log_{10} f$ values, resulting in the observed lack of smoothness.

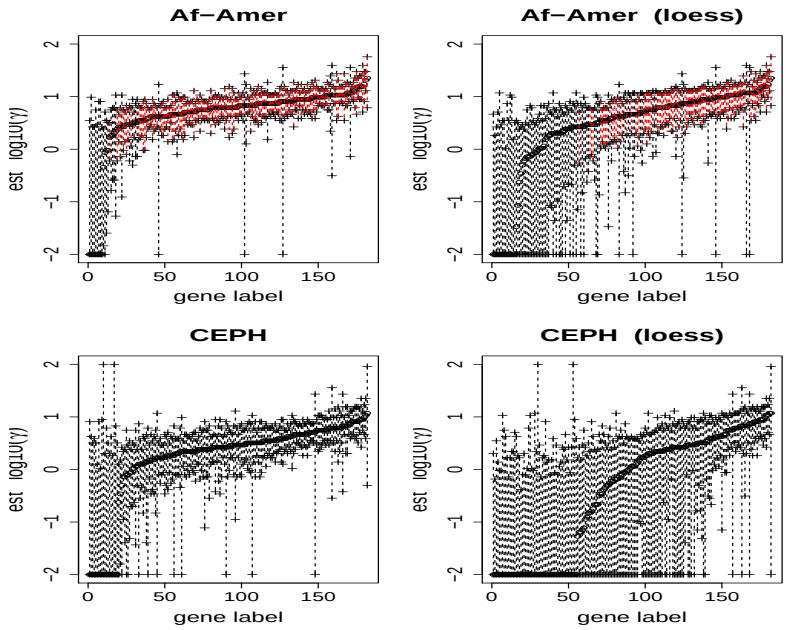


Figure 3.10: Estimated $\log_{10}(\gamma/\text{kb})$ for African-Americans (top row) and CEPH (bottom row) for 183 genes of *SeattleSNPs*, using the version of PAC described in Chapter 2 (left) and the loess-corrected version described in this chapter (right). The genes are labeled in order of increasing $\hat{\gamma}$. The dotted lines extend from the closest γ grid values such that $l(\gamma, \hat{\rho})$ is two log-likelihood units below $l(\hat{\gamma}, \hat{\rho})$, for each of $\gamma < \hat{\gamma}$ and $\gamma > \hat{\gamma}$ (these γ values are denoted by '+'). Note that the lengths of these "confidence regions" for genes with low $\hat{\gamma}$ increase after using the loess-corrected version of PAC compared to the original version, i.e. from left to right. For example, for the African-American dataset, we have highlighted in red the genes for which the length of this region is $\leq$1, of which there are noticeably fewer in the loess-corrected plot.

3.4.2 Effect of Loess Correction on Shape of Likelihood

A key point of using the loess-correction is that we have altered the likelihood to more accurately reflect our uncertainty in estimating low-to-mid-range values of γ. That is, the loess correction not only corrects a bias in point estimates, but it actually changes the shape of the likelihood curve. To illustrate this, we found the MLEs of ρ and γ per gene over a 100-point, equally spaced $\log_{10}(\rho/\text{kb})$ grid on [-2.40,1.60] and a 100-point, equally spaced $\log_{10}(\gamma/\text{kb})$ grid on [-2.00,2.00] using the PAC model before and after

the loess-correction. In Fig.3.10, we plot $\hat{\gamma}$ and the closest γ grid values within two log-likelihood units either side of $\hat{\gamma}$ for each gene. That is, for each gene, we plot $\hat{\gamma}$ and its closest grid value $\gamma < \hat{\gamma}$ such that $l(\gamma, \hat{\rho}) \leq l(\hat{\gamma}, \hat{\rho}) - 2$ and its closest grid value $\gamma > \hat{\gamma}$ such that $l(\gamma, \hat{\rho}) \leq l(\hat{\gamma}, \hat{\rho}) - 2$, where $l(\gamma, \rho)$ is the log-PAC value for γ,ρ. These "confidence regions" narrow as $\hat{\gamma}$ increases, reflecting the model's increasing certainty in estimated γ for large $\hat{\gamma}$. However, prior to loess-correction, perhaps too few of the genes with small $\hat{\gamma}$ have wide enough such regions to appropriately reflect our model's difficulty in accurately estimating small values of γ. After loess-correction, we see the number of genes with small $\hat{\gamma}$ and wide confidence regions increases considerably, for both populations.

3.5 Summary

In this chapter we have characterized a bias in γ estimation, which appears to be most strongly influenced by the true rate of gene conversion in the region. In particular, our model appears to substantially overestimate low to medium amounts of gene conversion. The magnitude of this overestimation appears to depend on the number of SNPs in the region, with a larger number of SNPs typically leading to an increased bias in estimated γ. The bias was assessed and corrected using loess curves on data simulated under a standard coalescent model. However, we found that inference based on the loess-corrected model appears robust to simulations that include population expansion and some degree of admixture. We re-analyzed the *SeattleSNPs* dataset using the loess-corrected version of our model to examine genome-wide f and found our estimates decrease by a factor of $\approx$2 from the original version of the model. Essentially the correction increases our uncertainty in accurately gauging small values of γ by flattening the PAC likelihood at these values. This may make the loess-corrected version of our likelihood more appropriate for use in a Bayesian model, which will be explored in Chapter 4.

Chapter 4

ESTIMATING RATES OF CROSSOVER AND GENE CONVERSION USING GENOTYPE DATA

Up to now we have assumed known haplotype information. In most current data collection protocols, one observes only the pair of SNP configurations at each site on a chromosome for an individual, i.e. the *genotype* information. The Li and Stephens (2003) PAC model has been previously implemented in a Bayesian Markov Chain Monte Carlo (MCMC) program called PHASE that iteratively updates (1) estimates of haplotypes conditional on the genotype data and current realization of ρ and (2) estimates of ρ conditional on the current realization of the haplotypes. We similarly incorporated our revised PAC model into PHASE to use genotype data to iteratively generate (1) estimates of haplotypes conditional on the current realization of ρ and γ and (2) estimates of ρ and γ conditional on the current realization of the haplotypes.

First we briefly outline the updated PHASE algorithm and describe our choice of prior parameters in the Bayesian model. Next we apply the PHASE algorithm to genotype data in the *SeattleSNPs* dataset. Finally, we briefly compare our results of this analysis to those of simulated data for which we know the truth.

For all analyses in this chapter, we use the γ loess-correction version of the PAC model described in Chapter 3.

4.1 Description of PHASE

4.1.1 The PHASE Core Algorithm

The program PHASE (Stephens et al. (2001), Stephens and Donnelly (2003)) uses a Gibbs update scheme to estimate each individual's haplotype configuration condi-

tional on his genotype information and the other haplotypes. For each step m,=1,...,M, of the chain, the potential haplotype configuration of individual i at this step, $\vec{h}_i^m$, which contains two chromosomes in diploid organisms, is sampled according to its relative probability, $\Pr(\vec{h}_i^m \mid H_{\{-i\}}^{(m-1)}, \vec{g}_i)$. Here $\vec{g}_i$ represents the observed genotype configuration of individual i and $H_{\{-i\}}^{(m-1)}$ represents the estimated haplotypes from step $(m-1)$ of the chain for all individuals in the sample excluding individual i. This is done for all n individuals in the sample at each step m for some large number of runs M. This algorithm has been updated by the authors of Li and Stephens (2003) to jointly estimate the haplotypes, H, and the crossover rate in the region, ρ, using the PAC model. We have incorporated our revised PAC model into the the most recently released version of PHASE, version 2.1.1, to jointly estimate H, ρ, and γ, the gene conversion rate in the region, as described below. (This updated version of PHASE incorporating gene conversion is not yet released.)

Begin the chain with $\rho^{(0)}$ and $f^{(0)}$, the starting values of ρ and $f = \gamma/\rho$, respectively. Then for each step m = 1,...,M of the MCMC algorithm of PHASE, we:

1. Sample $\vec{h}_i^{(m)}$ from $\Pr(\vec{h}_i^{(m)} \mid \vec{g}_i, H_{\{-i\}}^{(m-1)}, \rho^{(m-1)}, f^{(m-1)})$. $\qquad i = 1,...,n$

2. Update $\rho^{(m)}$ given $H^{(m)}$ and $f^{(m-1)}$.

3. Update $f^{(m)}$ given $H^{(m)}$ and $\rho^{(m)}$.

Steps (1) and (2) are achieved via Gibbs and Metropolis-Hasting steps, as in Stephens and Scheet (2005). Step (3) is achieved by a random-walk Metropolis-Hasting step on $\log_{10} f$.

New values of ρ and f are accepted based on their posterior probability, which is proportional to their likelihood value from the revised PAC model and the prior probability of ρ and f. A sampled pair of (ρ,f) gives γ. We note that ρ and γ can be constant across the region or vectors that allow the rates of crossover and gene conversion to vary across the region, as in Sec.4.2.2 below.

4.1.2 Priors on ρ,f

We use independent priors on ρ and f to define $\Pr(\rho, \gamma)$, the prior probability of ρ and γ. The assumption of independence of ρ and f is chiefly for simplicity. Specifically we assume each of $\log_{10} \rho$ and $\log_{10} f$ are normally distributed with mean and variance parameters (μ_ρ,σ_ρ^2) and (μ_f,σ_f^2), respectively; i.e.:

$$\log_{10} f \sim N(\mu_f, \sigma_f^2)$$
$$\log_{10} \rho \sim N(\mu_\rho, \sigma_\rho^2).$$

The normality assumption is motivated primarily by convenience, while working on the log scale is motivated by a desire to allow that these parameters might vary by an order of magnitude across different genomic regions. In addition, this formulation restricts each of ρ and f to be positive.

4.2 Application to Genotype Data of *SeattleSNPs* Dataset

4.2.1 Choice of Prior Parameters

Before analyzing the genotype data of each of the African-American and CEPH populations of the genes of *SeattleSNPs*, i.e. performing a new analysis that relaxes the assumption that haplotypes are known without error and uses informative priors on f and ρ, it remains to find appropriate values for μ_f, μ_ρ, σ_f, and σ_ρ. In principle one might like to treat these parameters as themselves unknown, or estimate them from the data. We pursue this in Chapter 5, but, for simplicity, we here take a less principled two-stage approach, by using our previous analysis to inform our priors. Specifically, we use values based on our results of Sec.3.4. In that section, the loess-corrected PAC model was applied to haplotypes from *SeattleSNPs* that were estimated using a different version of PHASE that did not include γ.

We consider four different sets of prior parameters, labeled I-IV, which span a range of scenarios based on the previous analysis of Sec.3.4. We stress that this is not

a comprehensive range of all plausible scenarios one might be able to think of, but rather a small subset that nonetheless may help to improve inference. The parameters of the priors are as follows, with μ_ρ corresponding to per kb values:

I. μ_f = 1.00, σ_f = 0.85, μ_ρ = -0.15 (AA) -0.70 (EA), σ_ρ = 0.80

II. μ_f = 1.00, σ_f = 0.50, μ_ρ = -0.15 (AA) -0.52 (EA), σ_ρ = 0.50

III. μ_f = 0.60, σ_f = 0.85, μ_ρ = -0.15 (AA) -0.70 (EA), σ_ρ = 0.80

IV. μ_f = 0.60, σ_f = 0.50, μ_ρ = -0.15 (AA) -0.70 (EA), σ_ρ = 0.50

For each prior, different values of μ_ρ are used for each population of *SeattleSNPs* (AA = African-Americans, EA = CEPH). We expect μ_ρ to be different between the two populations as a result of them having different effective population sizes, if not different average rates of crossover.

In contrast, in each prior we use the same values of μ_f and σ_f for both populations. Unlike ρ, the parameter f is not scaled by effective population size, and there is no strong reason *a priori* to suspect f will differ greatly between the two populations. As our model shows considerable more confidence in its estimates of γ for the genes of the African-American population than for the European population, we use the results for the African-American data to guide our choice of μ_f and σ_f. The increased precision for the African-American data is not surprising as these data have more SNPs compared to the CEPH data.

We also *a priori* do not necessarily expect the variation in ρ to differ across populations. Therefore we use the same σ_ρ value for each population as well, again using the African-American population to guide our choices.

The remainder of this section provides insight into our choice of these particular values for priors I-IV.

The parameters of prior I are based on the medians and standard deviations of $\log_{10}\hat{\rho}$ and $\log_{10}\hat{f}$ across genes with small γ "confidence regions," as defined in Sec.3.4.2. Tables 4.1 and 4.2 show the values for f and ρ, respectively, obtained using different definitions of "small." For ρ we ended up using all the genes; these values are highlighted in red in Table 4.2. For f, which is directly affected by estimates of γ, we ended up (somewhat arbitrarily) looking only at the 47% of genes with the smallest γ "confidence regions" in the African-American data. These values are highlighted in red in Table 4.1.

Table 4.1: Summary statistics for $\log_{10}\hat{f}$ in *SeattleSNPs*, estimated using PAC (Chapter 3), for various lengths of γ "confidence region." The "cut-off" refers to the length of the confidence region on the $\log_{10}$ scale (e.g. cut-off = 1 corresponds to a factor of 10 difference between the upper and lower bounds of the interval).

	Af-Amer				CEPH			
cut-off	% genes	median	mean	std dev	% genes	median	mean	std dev
–	100	0.76	0.76	1.15	100	0.46	0.44	1.45
3.0	94	0.76	0.77	1.15	92	0.46	0.43	1.43
2.5	75	0.93	1.00	1.04	57	0.40	0.41	1.60
2.0	67	0.93	1.17	0.86	34	1.09	1.04	1.55
1.5	62	0.93	1.19	0.87	22	1.49	1.75	1.03
1.0	53	1.00	1.36	0.88	19	1.57	1.81	1.06
0.9	47	1.01	1.22	0.85	14	1.67	1.96	1.02
0.7	34	1.01	1.21	0.86	9	2.26	2.18	0.98
0.6	21	1.05	1.24	0.77	3	2.32	2.40	0.85
0.5	13	1.05	1.32	0.77	0.5	3.35	3.35	–

For prior II, we use the same or similar values of μ_f and μ_ρ as in prior I. However, we use a smaller value for each of σ_f and σ_ρ, equal to 0.5 for both. The use of these alternative σ_f and σ_ρ is based on the fact that the standard deviation of the estimates above may overestimate variability in the true values.

In Sec.3.4, we note that our model is more confident in its estimation of genes with large $\hat{\gamma}$. Therefore, one obvious problem with both prior I and prior II is that

Table 4.2: Summary statistics for $\log_{10} \hat{\rho}$ in *SeattleSNPs*, estimated using PAC (Chapter 3), for various lengths of γ "confidence region." The "cut-off" refers to the length of the confidence region on the $\log_{10}$ scale.

	Af-Amer				CEPH			
cut-off	% genes	median	mean	std dev	% genes	median	mean	std dev
–	100	-0.16	-0.35	0.80	100	-0.66	-0.86	0.87
3.0	94	-0.18	-0.34	0.78	92	-0.70	-0.85	0.84
2.5	75	-0.16	-0.35	0.80	57	-0.74	-0.91	0.89
2.0	67	-0.14	-0.36	0.82	34	-0.66	-0.91	0.98
1.5	62	-0.14	-0.36	0.82	22	-0.72	-1.00	0.99
1.0	53	-0.18	-0.38	0.84	19	-0.78	-1.03	1.00
0.9	47	-0.14	-0.33	0.83	14	-0.98	-1.14	0.99
0.7	34	-0.02	-0.28	0.86	9	-1.35	-1.33	0.96
0.6	21	-0.08	-0.25	0.79	3	-1.45	-1.50	0.79
0.5	13	-0.06	-0.28	0.81	0.5	-2.40	-2.40	–

our choice of μ_f may be biased upwards because it likely consists primarily of genes with large $\hat{\gamma}$. Therefore, for priors III and IV, we use our profile-likelihood estimate of f as an alternative choice of μ_f, as this estimate of f combines information in the likelihoods of several genes. We found that the profile likelihoods of only a few genes substantially affected the profile-likelihood estimate of f for all genes; therefore we base our μ_f on the profile likelihoods of genes whose $\hat{f}$ were in the middle 80% of all the genes (see Sec.3.4). All other parameter choices of prior III and prior IV match or are similar to those of prior I and prior II, respectively.

We reiterate that we do not expect any of the above priors to be "correct" or "most appropriate" but rather hope to use them for comparison and as starting points for more sophisticated inference.

4.2.2 Analysis of SeattleSNPs Using Genotype Data

We applied the revised PHASE to genes from the African-American and CEPH populations separately, using the genotypes of 24 individuals in the former population

and the genotypes of 23 individuals in the latter. For each population, we considered only biallelic SNPs and removed *singletons*, i.e. SNPs with only one copy of the rare allele across all n individuals of the population ignoring missing data. We considered only genes with at least 20 SNPs remaining after this step; this left 204 genes in the African-American dataset and 173 genes in the CEPH dataset, out of an original 225 genes at the time the data were downloaded.

We ran this data through the revised PHASE described in Sec.4.1.1, using each of the priors I-IV described in Sec.4.1.2. We used μ_f and μ_ρ as the starting values in the MCMC chain for $\log_{10} f$ and $\log_{10} \rho$, respectively, for each gene. We ran the chain for the number of runs the PHASE authors recommend for recombination rate estimation, i.e. 10 times the default settings. We specified mutation rate $\tilde{\theta} = 0.0002$, and a tract length of 100bp for gene conversion events, as in Chapters 2 and 3.

Recent analyses of linkage disequilibrium data suggest crossover hotspots of ≈1-2kb in length are a common feature of the genome, perhaps occuring as frequently as one every 30-50kb (Kauppi et al. (2004), Fearnhead and Smith (2005), Myers et al. (2005)). Genes in our *SeattleSNPs* dataset are, on average, ≈28kb in length. We therefore allow for a maximum of one recombination hotspot per gene, in which f remains constant and equivalent to regions outside the hotspot. Thus any hotspot is assumed to be a hotspot for gene conversion as well, which was the case with the majority of hotspots examined via sperm analysis in Jeffreys and May (2004) and Jeffreys and Neumann (2005). The prior in PHASE assumes hotspots occur as a Poisson process of one every 40kb and have a minimum length of 200bp, with $\log_{10}$ of the hotspot intensity distributed as a Uniform(0,3). This allows for hotspots as intense as 1000 times the background rate in a region.

Variability in f

When we consider $\hat{f}$, the PAC MLE estimate of f, per gene, we see considerable variability in estimates across the genome. However, it is difficult to determine from

the PAC estimates alone whether the variability in $\hat{f}$ is primarily the result of noise in estimation or if there is indeed signal in the data that f varies considerably. For many genes, the likelihood of f may be relatively flat, so that there may be small differences in likelihood values between very large f and very small f. This effect can be seen in the abundance of very wide γ "confidence regions" for most genes (see Table 4.1). The use of a prior for inference hopefully somewhat eliminates this problem, by shrinking the f estimates of genes with flat likelihoods towards more "typical" values.

In Fig.4.1 we plot histograms of $\log_{10} f$ estimates across genes from PAC and from PHASE for each of priors I-IV, for each population of *SeattleSNPs*. For the PAC results, we find the MLE of f over a grid of values for each gene using the model and grid described in Sec.3.4. For the PHASE results, we take the median of 1000 posterior samples as our f estimate per gene. Note that the histograms for PHASE are more peaked than PAC, presumably from use of a prior to eliminate noise. This suggests that extreme f estimates in PAC indeed appear to be maxima of flat likelihoods rather than strong signal in the data towards large or small f. We expect this beneficial effect of noise reduction from the use of our priors, though at the expense of some degree of bias.

All histograms show considerable variability in estimated $\log_{10} f$, although it is difficult to know whether this reflects continuous noise in estimates or genuine biological variability, a topic we revisit in the next chapter. The level of variability depends on the prior, with priors II and IV, which use a smaller σ_f than priors I and III, having the shortest range of values. Median values of f among the four PHASE priors range from $\approx$1-4, perhaps slightly lower than the range of $\approx$3-5 in PAC. Surprisingly the medians for the PHASE results of some priors are below that of both PAC and the prior; we currently have no explanation for this result.

While it is tempting to interpret the PHASE results as more accurate genome-wide representations of f than the PAC ones, it is first necessary to explore what we

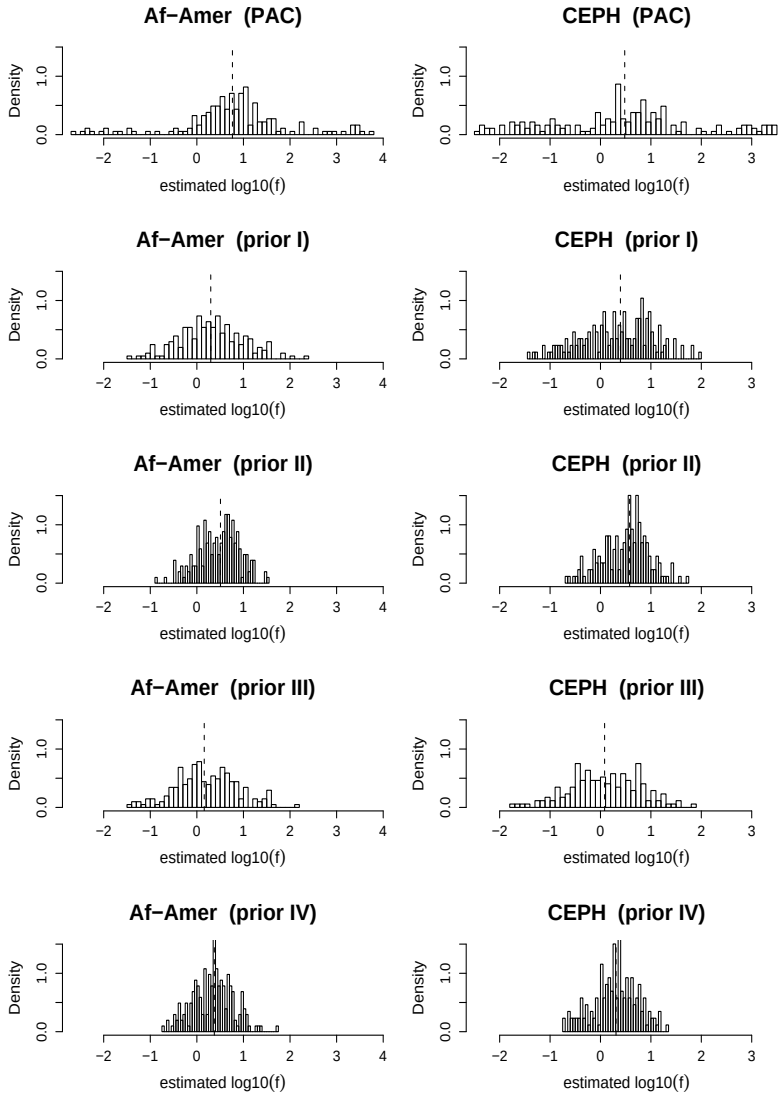


Figure 4.1: Histograms of $\log_{10} f$ estimates across all genes, for the African-American (left) and CEPH (right) data of *SeattleSNPs*. The estimates were obtained using the PAC model (top) and PHASE (rows 2-5, corresponding to use of priors I-IV in PHASE, respectively). The dashed vertical lines in each plot represent the median across estimates. Note that the histogram of PAC estimates has heavier tales, which may be a reflection of noise, which is eliminated via use of any of the priors.

might be gaining by using PHASE. There are two main aspects of PHASE that set it apart from the PAC model. One is that PHASE uses prior information. From Fig.4.1, it appears the benefit of doing so is the elimination of some noise in estimation. The second is that PHASE uses genotype information rather than assuming haplotypes are known correctly. We use simulations to explore the effect of this in the next section.

4.3 Simulations

All simulations in this section contained 204 genes matched to the sequence lengths and SNP numbers of those in the African-American *SeattleSNPs* dataset of this chapter. Each gene had 24 individuals, also matching the African-Americans. To further match our observed data, a large number of SNPs was generated for each gene, singleton SNPs were removed, and SNPs were randomly chosen from the remaining sites until the SNP number of the simulated gene equaled that of the real gene being matched. We ran each simulated dataset through PHASE using each of priors I-IV with the African-American value of μ_ρ. They were run in PHASE in the same manner as described in Sec.4.2.2, with one exception: to decrease computation time, all simulations had no DSB hotspots, and PHASE was run assuming this to be the case. The effect of hotspots in simulations will be explored in Chapter 5. The expected tract length of a gene conversion event in the simulations was 100bp; this value was fixed as the tract length in our model. As before, simulations were generated using ms (Hudson, 2002).

We simulated using three distinct sets of true values, which we label (A), (B), and (C). For simulation (A), each gene i was simulated with $f = 2.0$ and $\rho = \bar{\rho}_i$, where $\bar{\rho}_i$ equals the median of 1000 posterior samples of ρ from PHASE for gene i in the African-American *SeattleSNPs* dataset using prior I. For simulation (B), each gene i was simulated with $f = 2.0$ and $\gamma = \bar{\gamma}_i$, where $\bar{\gamma}_i$ equals the median of 1000 posterior samples of γ from PHASE for gene i in the African-American *SeattleSNPs* dataset

using prior I. For simulation (C), each gene i was simulated with $\gamma = \bar{\gamma}_i$ and $\rho = \bar{\rho}_j$, where j is randomly chosen from the 204 genes of the African-American *SeattleSNPs* dataset, with each gene used only once. The f of (A) and (B) matches the median of PHASE estimates for *SeattleSNPs* using prior I (see Fig. 4.1).

We chose these simulation parameters primarily to examine two things. The first is to explore properties of our model's estimation for recombination rates that match our inferred ones. This includes, for example, how accuracy in estimating these recombination rates are affected by using genotype information rather than correct haplotypes, which will be examined in the next section. The second is to examine the reliability of our model's inference on variability in f, which will be addressed in Chapter 5.

Haplotypes vs Genotypes

For simulations (A), (B), and (C), we ran PHASE on both the unphased genotype data and the correct phase-known haplotype data. The primary aim here is to see what loss is incurred in the accuracy of our inference when having to estimate haplotypes. Having complete and accurate haplotype information clearly represents a best case scenario; in typical real-life scenarios haplotypes must be estimated. Errors in this estimation can affect inference, so that using genotype information rather than assuming haplotypes are known correctly might be preferable in such cases. However, for this section, we examine only the potential loss incurred by not having complete haplotype information.

Table 4.3 shows the mean-squared-error (MSE) in estimation, for $\log_{10} \bar{\gamma}$, $\log_{10} \bar{\rho}$, $\log_{10} \bar{f}$, and $\log_{10} \bar{\delta}$, when using prior I. Here $\delta = 4N_e d$, with d the average rate of double-strand-breaks per unit physical distance per chromosome in a single transmission from parent to offspring. We calculate δ as $\gamma + \rho$. As before, we let $\bar{x}$ denote the median of 1000 posterior samples from PHASE per gene. Somewhat surprisingly, the MSE is not that much different for the genotype data compared to using correct

haplotype data and, in fact, is often smaller. This also appears to be the case when using prior II; the MSE results when using this prior are shown in Table 4.4.

This lack of a difference in MSE between the haplotype-known and haplotype-estimated scenarios is probably a reflection of the fact that the considerable uncertainty in estimating rates of γ overwhelms any uncertainty in estimating haplotypes. In support of this point, in most applications where earlier versions of PHASE have been tested, PHASE has shown relatively small rates of error when estimating haplotypes for datasets of this size (Smith and Fearnhead (2005), Marchini et al. (2006), Scheet and Stephens (2006)).

Table 4.3: Mean-Squared-Error for simulations based on African-American *SeattleSNPs* data results, using the correct haplotypes ("Haplotypes") or simultaneously estimating haplotypes ("Genotypes"), using prior I.

		$\log_{10}\bar{\gamma}$	$\log_{10}\bar{\rho}$	$\log_{10}\bar{f}$	$\log_{10}\bar{\delta}$
(A) f=2.0, $\bar{\rho}$					
	Haplotypes	0.372	0.067	0.341	0.132
	Genotypes	0.445	0.058	0.462	0.111
(B) f=2.0, $\bar{\gamma}$					
	Haplotypes	0.510	0.119	0.392	0.232
	Genotypes	0.480	0.065	0.435	0.164
(C) $\bar{\gamma}$, $\bar{\rho}$ random					
	Haplotypes	0.425	0.061	0.399	0.173
	Genotypes	0.379	0.055	0.440	0.118

Table 4.4: Mean-Squared-Error for simulations based on African-American *SeattleSNPs* data results, using the correct haplotypes ("Haplotypes") or simultaneously estimating haplotypes ("Genotypes"), using prior II.

		$\log_{10}\bar{\gamma}$	$\log_{10}\bar{\rho}$	$\log_{10}\bar{f}$	$\log_{10}\bar{\delta}$
(A) f=2.0, $\bar{\rho}$					
	Haplotypes	0.145	0.047	0.116	0.081
	Genotypes	0.156	0.041	0.161	0.067
(B) f=2.0, $\bar{\gamma}$					
	Haplotypes	0.259	0.102	0.144	0.181
	Genotypes	0.217	0.064	0.160	0.128
(C) $\bar{\gamma}$, $\bar{\rho}$ random					
	Haplotypes	0.288	0.059	0.320	0.124
	Genotypes	0.251	0.055	0.319	0.094

4.4 Summary

In this chapter, we have incorporated our revised PAC likelihood of Chapters 2 and 3 into the Bayesian MCMC algorithm PHASE (Stephens et al. (2001), Stephens and Donnelly (2003)) to jointly estimate haplotypes and (γ,ρ) using genotype information. We found that the loss in accuracy of inference incurred by estimating haplotypes rather than having them known correctly appears negligible. In addition, we found that the use of prior information, based on the results of the application of our revised PAC to *SeattleSNPs* in Chapter 3, appears to reduce noise in estimation of f compared to using PAC alone. Application to the *SeattleSNPs* dataset shows evidence for substantial variability in genome-wide f. However, our inference shows some dependence on which prior parameters are used in PHASE. This suggests we perhaps do not want to fix these prior parameters but rather allow for some flexibility. This idea will be explored in Chapter 5.

Chapter 5

JOINTLY ESTIMATING GENOME-WIDE VARIATION IN RECOMBINATION IN ADDITION TO RATES OF CROSSOVER AND GENE CONVERSION

Though the use of an informative prior appears to improve accuracy in estimating f compared to using the PAC model alone, we found in Chapter 4 that our inference depends on the parameters used in the prior. For this reason, instead of fixing our prior parameters, we extend our model to allow for more flexibility. In particular, we incorporate a hierarchical model that uses importance sampling to sample these parameters conditional on the observed data.

First we describe the hierarchical model and our choice of prior parameters. Next we apply this new model to the simulated datasets described in Chapter 4 and assess the model's properties and accuracy. We then apply our model to the genotype data of *SeattleSNPs* and describe the results. In particular, we focus on exploring the genome-wide variability in rates of f and δ in this dataset, as well as how recombination rates correlate with the sequence features SNP density and %G+C content. Finally, we apply our model to simulated datasets designed to evaluate the reliability of our main conclusions from the model's application to *SeattleSNPs*.

As in Chapter 4, for all analyses in this chapter, we consider only the γ loess-correction version of the PAC model described in Chapter 3.

5.1 The Hierarchical Model

5.1.1 Introduction: Exploring Variability in Recombination Rates

Rates of crossover appear to vary substantially across the genome at megabase (Kong et al., 2002) and kilobase scales (Crawford et al. (2004), McVean et al. (2004), Myers et al. (2005)). Though much of this variability may be due to the phenomenon of crossover hotspots, the *background* crossover rates, i.e. the rates of crossover activity outside of any fine-scale hotspots, may still vary by an order of magnitude between two genes or regions of similar size (Crawford et al., 2004). According to the double-strand break (DSB) model (Szostak et al., 1983), outlined in Sec.1.2.2, the rate of crossover in a region is determined by (1) the rate of double-strand-breaks occuring in the region and (2) the proportion of double-strand-breaks that are resolved as crossovers. Therefore, if this model is correct, the observed genome-wide variability in rates of crossover could be due to either genome-wide variability in (1) or variability in (2), or variability in both. But which of these two features of the DSB model contributes most?

It seems probable that variability in double-strand-break rates is the dominant feature contributing to the crossover rate variation between a region inside a hotspot and a region just outside. The amount of crossover activity in most human hotspots is very large compared to regions just outside the hotspot, suggesting there must be increased DSB activity in the hotspot. Otherwise, if the crossover rate difference was all due to variability in the proportion of resolution, then regions outside of crossover hotspots – which account for the vast majority of the genome – would have extreme amounts of gene conversion occuring, as in factors of tens to hundreds that of crossover. Based on the relatively small signal for gene conversion in LD data, this does not appear to be the case. Consistent with the theory that crossover hotspots reflect heightened DSB activity, most hotspots examined by sperm analysis for both crossovers and gene conversions were found to be highly active for both (Jeffreys

and May (2004), Jeffreys and Neumann (2005)). It seems that LD data are not very informative for rates of gene conversion on scales as small as hotspots, which are typically 1-2kb in length. Therefore, it appears that sperm analysis, despite its technical difficulties, will likely remain the preferred means of examining variability in gene conversion rates within a single multiple-kilobase region, and hence DSB rates in such regions. A possible exception might be regions with both hotspots and high SNP density, such as the MHC; LD patterns in such regions may contain adequate enough information on gene conversion events at very fine scales to allow for the study of DSB rates in hotspots.

But what about comparing rates outside of hotspots between two, e.g., 20 kilobase – or genic – regions? Estimated background crossover rates across 74 *SeattleSNPs* genes in Crawford et al. (2004) showed substantial variability beyond that which is expected by error in estimation. This variability could conceivably be due to either variability in DSB rates, variability in the proportion of DSBs resolved as crossovers, or both. If we can reliably estimate rates of crossover and gene conversion on the scale of genes, we might hope to gain insight into this question. In particular, as crossovers and gene conversions, as we have defined them, are mutually exclusive results of a double-strand-break, we can estimate the double-strand-break rate, δ, for a given region by summing the rates of each for that region, i.e.:

$$\delta = \rho + \gamma. \tag{5.1}$$

Here $\delta = 4N_e d$, where d is the average rate of double-strand-breaks per unit physical distance per chromosome in a single transmission from parent to offspring. Furthermore, we can get at the variability in the proportion of DSB resolution by considering the variability in the relative rate of gene conversion to crossover, $f = \gamma/\rho$.

Therefore we wish to compare the variabilities of δ and f among genes to see which contributes most to the observed variability in ρ among genes. Because our estimates

for each of δ and f for individual genes can have significant uncertainty, we wish to avoid simply using the standard deviations of our estimates of δ and f across genes to address this question. As a proposed improvement, we present a hierarchical model framework that simultaneously "borrows strength" across genes to improve individual gene estimation and provides a means for estimating the genome-wide variability in f and δ.

5.1.2 The Hierarchical Model

The structure of our hierarchical model is as follows in this section. The PHASE priors in Chapter 4 used independent priors on $\log_{10} f$ and $\log_{10} \rho$. As we are here primarily interested in comparing rates of δ and f across the genome, we re-parameterize in terms of $\log_{10} \delta$ and $\log_{10} f$. We will assume that $\log_{10} \delta$ and $\log_{10} f$ values of genes are multivariate-normal distributed. That is, for gene i:

$$\begin{pmatrix} \log_{10} f_i \\ \log_{10} \delta_i \end{pmatrix} \sim N\Big(\begin{pmatrix} \mu_f \\ \mu_\delta \end{pmatrix} \equiv \vec{\mu}, \begin{pmatrix} \sigma_f^2 & \sigma_{f,\delta} \\ \sigma_{f,\delta} & \sigma_\delta^2 \end{pmatrix} \equiv \Sigma \Big),$$

where (μ_f,σ_f) and (μ_δ,σ_δ) represent the genome-wide mean, standard deviation of $\log_{10} f$ and $\log_{10} \delta$, respectively, and $\sigma_{f,\delta}$ the covariance, under this normality assumption. We note that these regions i need not be genes, but could be regions of size 20-100kb, for example. To avoid confusion, in this chapter we will denote values of μ_f, σ_f, μ_ρ, and σ_ρ used in the PHASE priors of Chapter 4 as $\mu_f^{(0)}$, $\sigma_f^{(0)}$, $\mu_\rho^{(0)}$, and $\sigma_\rho^{(0)}$, respectively.

Instead of fixing μ_f, σ_f, μ_δ, σ_δ, and $\sigma_{f,\delta}$ at specific values, we wish to estimate them from the data. Initially we tried using maximum likelihood by computing the likelihood for a grid of values of the parameters and approximating the integrals necessary to compute the likelihood by Guassian quadature. However, the results we

obtained by this approach did not appear to make sense. We suspect this was due to the coarseness of the grid we were using to make computations feasible and possibly the quality of the numerical integration we were performing. We therefore tried a Bayesian implementation, specifying priors for the parameters and attempting to sample approximately from their posterior distributions conditional on the observed data.

To make sampling from the posterior distribution computationally efficient, we make use of conjugate priors for the multivariate-normal distribution; i.e.:

$$\Sigma \sim \text{Inv-W}(\upsilon_0, S_0)$$

$$\vec{\mu} \mid \Sigma \sim \text{MVN}(\vec{\mu_0}, \Sigma/\kappa_0),$$

where inv-W(υ_0, S_0) is an inverse-Wishart distribution with scale matrix S_0 and degrees of freedom υ_0, and MVN$(\vec{\mu_0}, \Sigma/\kappa_0)$ is a multivariate-normal distribution with mean vector $\vec{\mu_0}$ and variance matrix Σ/κ_0. The values we use for υ_0, S_0, $\vec{\mu_0}$, and κ_0 are given in Sec.5.1.3.

With this formulation, the posterior distributions of $\vec{\mu}$ and Σ take the following form, with $l\vec{F}$, $l\vec{D}$ denoting the values of $\log_{10} f$ and $\log_{10} \delta$ across genes:

$$\Sigma \mid l\vec{F}, l\vec{D} \sim \text{Inv-W}(\upsilon_0\text{+c}, (S_0 + S + \tfrac{\kappa_0 c}{\kappa_0 c + c}(\bar{y} - \vec{\mu_0})(\bar{y} - \vec{\mu_0})^T)^{-1})$$

$$\vec{\mu} \mid l\vec{F}, l\vec{D}, \Sigma \sim \text{MVN}(\tfrac{\kappa_0}{\kappa_0 + c}\vec{\mu_0} + \tfrac{c}{\kappa_0 + c}\bar{y}, \Sigma/(\kappa_0 + c)).$$

Here $\bar{y} = \begin{pmatrix} l\bar{F} \\ l\bar{D} \end{pmatrix}$, with $l\bar{F}$ and $l\bar{D}$ representing the means of $\log_{10} f$ and $\log_{10} \delta$ across genes, respectively, c is the number of genes, and $S = \sum_{i=1}^{c}(\vec{y_i} - \bar{y})(\vec{y_i} - \bar{y})^T$, with $\vec{y_i} = \begin{pmatrix} lF_i \\ lD_i \end{pmatrix}$, the $\log_{10} f$, $\log_{10} \delta$ values for gene i. We primarily wish to compare sampled values of σ_δ and σ_f, each components of Σ, to address our question of which feature of ρ appears to contribute most to genome-wide variability. As we consider δ

and f on the log scale, σ_δ and σ_f represent a measure of magnitude variability in δ and f, respectively.

The model is implemented into the following MCMC framework. We begin the chain with initial starting values of $\vec{\mu}$ and Σ. Then for each step $m = 1,...,M$ of the MCMC algorithm of our hierarchical model, letting G represent the genotype data, the outline of our algorithm is as follows:

1. sample $(lF_i^{(m)}, lD_i^{(m)})$ from $\Pr(lF_i^{(m)}, lD_i^{(m)} \mid \vec{\mu}^{(m-1)}, \Sigma^{(m-1)}, G)$ $\quad i = 1, ..., c =$ genes
2. sample $\Sigma^{(m)}$ from $\Pr(\Sigma^{(m)} \mid l\vec{F}^{(m)}, l\vec{D}^{(m)}, \vec{\mu}^{(m-1)})$
3. sample $\vec{\mu}^{(m)}$ from $\Pr(\vec{\mu}^{(m)} \mid l\vec{F}^{(m)}, l\vec{D}^{(m)}, \Sigma^{(m)})$.

Steps (2) and (3) are straight-forward, involving sampling from the closed form inverse-Wishart and multi-variate normal distributions outlined above. Step (1) is precisely the problem solved by PHASE in Chapter 4. However, it is computationally time-consuming to use PHASE to obtain samples from this distribution for even one single value of $\vec{\mu}$ and Σ. Therefore, to avoid the need to perform a run of PHASE for many values of $\vec{\mu}$ and Σ, we take an approach analagous to *importance sampling*. We use PHASE to provide samples from $\Pr(lF_i, lD_i \mid \vec{\mu}^{(0)}, \Sigma^{(0)}, G)$, where $\vec{\mu}^{(0)}$ and $\Sigma^{(0)}$ are fixed parameter choices we will refer to as *driving values*, and then re-weight these samples to provide an approximate sample from $\Pr(lF_i, lD_i \mid \vec{\mu}^{(m)}, \Sigma^{(m)}, G)$. It intuitively works as follows. Clearly

$$\Pr(lF_i, lD_i \mid \vec{\mu}^{(m)}, \Sigma^{(m)}, G) = \frac{\Pr(lF_i, lD_i \mid \vec{\mu}^{(m)}, \Sigma^{(m)}, G)}{\Pr(lF_i, lD_i \mid \vec{\mu}^{(0)}, \Sigma^{(0)}, G)} \Pr(lF_i, lD_i \mid \vec{\mu}^{(0)}, \Sigma^{(0)}, G). \tag{5.2}$$

Sampling from (5.2) can be approximated by sampling values of $(\log_{10} f_i, \log_{10} \delta_i)$ from $\Pr(lF_i, lD_i \mid \vec{\mu}^{(0)}, \Sigma^{(0)}, G)$ – we will denote such samples (lF_i^*, lD_i^*) in (5.3) – and then sampling from among these values according to

$$\Pr(lF_i^*, lD_i^* \mid \vec{\mu}^{(m)}, \Sigma^{(m)}, G) \quad \propto \quad \frac{\Pr(lF_i^*, lD_i^* \mid \vec{\mu}^{(m)}, \Sigma^{(m)}, G)}{\Pr(lF_i^*, lD_i^* \mid \vec{\mu}^{(0)}, \Sigma^{(0)}, G)}. \tag{5.3}$$

Furthermore, we note that the right-hand side of (5.3) can be broken down:

$$\begin{aligned} \frac{\Pr(lF_i, lD_i \mid \vec{\mu}^{(m)}, \Sigma^{(m)}, G)}{\Pr(lF_i, lD_i \mid \vec{\mu}^{(0)}, \Sigma^{(0)}, G)} \quad &\propto \quad \frac{\Pr(G \mid lF_i, lD_i)\Pr(lF_i, lD_i \mid \vec{\mu}^{(m)}, \Sigma^{(m)})}{\Pr(G \mid lF_i, lD_i)\Pr(lF_i, lD_i \mid \vec{\mu}^{(0)}, \Sigma^{(0)})} \\ &\propto \quad \frac{\Pr(lF_i, lD_i \mid \vec{\mu}^{(m)}, \Sigma^{(m)})}{\Pr(lF_i, lD_i \mid \vec{\mu}^{(0)}, \Sigma^{(0)})} \end{aligned} \tag{5.4}$$

The final step in (5.4) replaces step (1) above. As this simply involves a quotient of two multivariate-normal distributions, this step is now computationally trivial.

In principle, if we were to use PHASE to obtain an infinite sample, the results should not depend on the driving values ($\vec{\mu}^{(0)}$,$\Sigma^{(0)}$). In practice, we obtain relatively small samples from PHASE, and so to assess dependence on ($\vec{\mu}^{(0)}$,$\Sigma^{(0)}$), we use multiple sets of driving values.

For this chapter we will use four distinct sets of values for ($\vec{\mu}^{(0)}$,$\Sigma^{(0)}$), corresponding to priors I-IV described in Sec.4.2.1. The MCMC chain is started at m=1 by randomly sampling ($\log_{10} f_i$,$\log_{10} \delta_i$) for each gene i=1,...,c as a substitute for step (1).

5.1.3 Prior and Prior Parameters

Our choices for the prior parameters of $\vec{\mu}$ and Σ are:

- $\upsilon_0 = 3$
- $S_0 = \begin{pmatrix} 0.1 & 0.001 \\ 0.001 & 1.0 \end{pmatrix}$
- $\vec{\mu_0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
- $\kappa_0 = 1/20.$

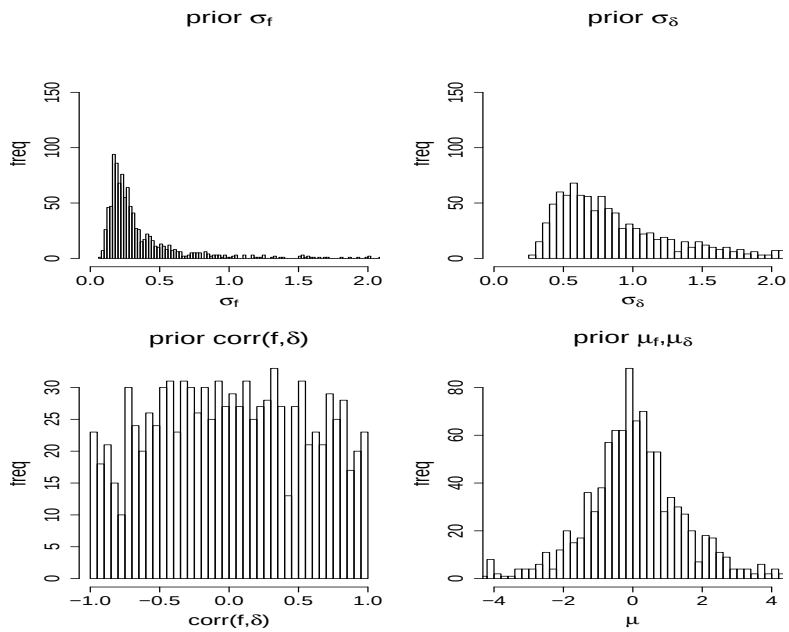


Figure 5.1: Histograms of 1000 prior samples of $\vec{\mu}$ and Σ based on the parameter values of Sec.5.1.3. Samples of σ_f (top left), σ_δ (top right), $corr(f,\delta)$ (bottom left), and μ_f or μ_δ (bottom right) are shown.

Histograms for samples from the prior distributions for σ_δ, σ_f, and the correlation between $\log_{10} f$ and $\log_{10} \delta$ ($\equiv corr(f,\delta)$) are shown in Fig.5.1. Some properties of these values are that they allow for a large range of possible μ_f and μ_δ, are to some degree skewed towards small σ_δ and σ_f, take σ_δ presumed greater than σ_f *a priori*, and give roughly uniform mass to all possible $corr(f,\delta)$ values.

5.2 Application to Simulated Datasets

We applied the algorithm of Sec.5.1.2 to simulations (A), (B) and (C), described in Sec.4.3, using each of priors I-IV as our driving values ($\vec{\mu}^{(0)}$,$\Sigma^{(0)}$) above. For each simulation and driving value, we used 1000 posterior values of ($\log_{10} f$,$\log_{10} \delta$) from one seed of PHASE as our sampling values. Using only one seed from PHASE per driving value was done for computational reasons; results in which we used poste-

rior samples from multiple seeds in our hierarchical model showed little difference in inference than just using samples from one seed (results omitted). We applied our hierarchical model to each dataset, sampling from the posterior distribution every 50 iterations of the MCMC chain after a 5,000 run burn-in.

5.2.1 Mean-Squared-Error of Estimates

The mean-squared-error (MSE) of the hierarchical model results for estimates of $\log_{10}\gamma$, $\log_{10}\rho$, $\log_{10}f$, and $\log_{10}\delta$, when using the median across posterior samples as a point estimate of each parameter per gene, are shown in Table 5.1. Regardless of which of the prior parameters I-IV is used in $(\vec{\mu}^{(0)},\Sigma^{(0)})$, estimates of $\log_{10}\rho$ and $\log_{10}\delta$ have lower MSE than either of $\log_{10}\gamma$ or $\log_{10}f$. This shows that percentage error in estimates of ρ and δ are smaller than for γ and f. Using the parameters of priors II and IV as driving values gives a lower MSE than using priors I and III. This is probably because priors II and IV have values of $\vec{\mu}^{(0)}$ and $\Sigma^{(0)}$ closer to the true values used to generate the data, and the importance sampling approach we use might be expected to bias estimates towards these $\vec{\mu}^{(0)}$, $\Sigma^{(0)}$.

Table 5.1: Mean-Squared-Error for simulated genotypes based on African-American *SeattleSNPs* data results, using four different priors in PHASE as driving values in the hierarchical model sampling scheme.

		$\log_{10}\bar{\gamma}$	$\log_{10}\bar{\rho}$	$\log_{10}\bar{f}$	$\log_{10}\bar{\delta}$
(A) f=2.0, $\bar{\rho}$					
	prior I	0.585	0.049	0.636	0.097
	prior II	0.299	0.050	0.319	0.080
	prior III	0.630	0.045	0.640	0.114
	prior IV	0.416	0.040	0.427	0.094
(B) f=2.0, $\bar{\gamma}$					
	prior I	0.682	0.056	0.668	0.147
	prior II	0.397	0.067	0.352	0.142
	prior III	0.809	0.073	0.813	0.151
	prior IV	0.451	0.066	0.395	0.139
(C) $\bar{\gamma}$, $\bar{\rho}$ random					
	prior I	0.504	0.056	0.623	0.114
	prior II	0.331	0.057	0.426	0.110
	prior III	0.547	0.057	0.697	0.126
	prior IV	0.353	0.061	0.474	0.115

5.2.2 Variability and Mean Estimates for f, δ

The densities of posterior samples of σ_f and σ_δ for simulations (A)-(C) are shown in Fig.5.2. Posterior distributions of σ_δ are in strong concordance across driving values for each simulation, flanking the true standard deviation of the simulated values. The distributions for σ_f are less precise. The posterior distribution flanks the true value for simulation (C), for which simulated $\sigma_f = 0.77$, only when priors I and III, for which $\sigma_f^{(0)}$=0.85, are used as driving values. For priors II and IV, for which $\sigma_f^{(0)}$=0.5, the vast majority of the posterior distribution for simulation (C) falls below the true value. The posterior distributions of σ_f for simulations (A) and (B) are considerably higher than the truth, σ_f=0, for all four driving values. While the prior in our hierarchical model does not allow for σ_f=0, it does allow for values as small as ≈0.1 (see Fig.5.1).

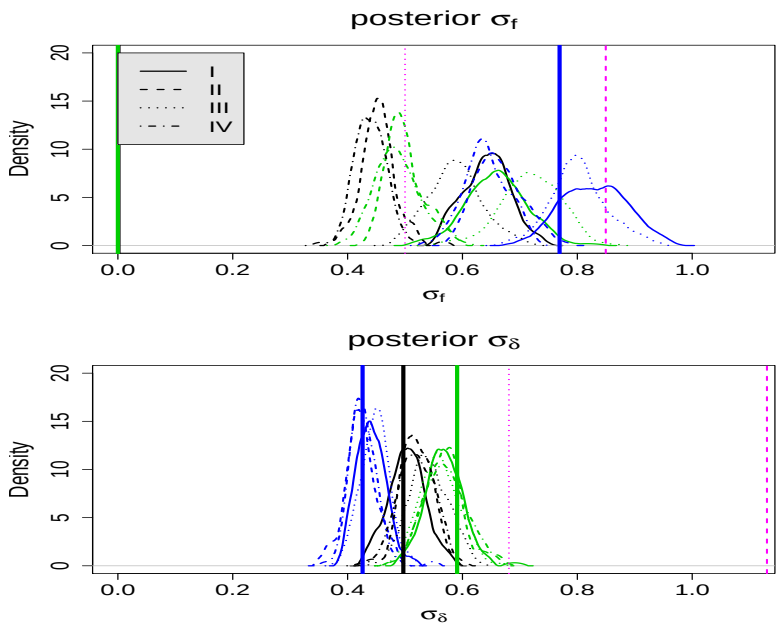


Figure 5.2: Hierarchical-model posterior samples of σ_f (top) and σ_D (bottom), for simulations (A) (black), (B) (green), and (C) (blue), using each of priors I-IV as driving values (see plot legend). The solid vertical lines represent the true σ_f and σ_δ for each simulation ($\sigma_f = 0$ for simulations (A) and (B)). The dashed vertical lines represent the values of $\sigma_f^{(0)}$ or $\sigma_\delta^{(0)}$ used in priors I and III. The dotted vertical lines represent the values and $\sigma_f^{(0)}$ or $\sigma_\delta^{(0)}$ used in priors II and IV. Note that though there is considerable noise in σ_f estimates, σ_δ appears to be estimated quite well for these simulations. In addition, posterior samples of σ_f appear to be between the true σ_f and the $\sigma_f^{(0)}$ used in the driving value, suggesting there is some moderate signal in the data for estimating variability in f.

However, none of the distributions of posterior σ_f for simulations (A) and (B) come very close to this lower bound. In contrast, when we used an alternative driving value with $\sigma_f^{(0)}$=0.2, the posterior densities of σ_f did reach ≈0.1 for each of simulations (A) and (B) (results omitted).

It makes some intuitive sense that our model would have a decent ability to pick up patterns of historical recombination but some difficulty in correctly crediting these patterns to crossover and gene conversion, given some of the similarities crossover and gene conversion have on patterns of LD. This ability would result in more reliable inference on aspects of δ, such as σ_δ, than on aspects of f, such as σ_f, as the accurate estimation of f critically depends on correctly differentiating LD patterns as gene conversions or crossovers. However, we do note that the posterior densities of σ_f for each simulation and driving value predominantly lie between the true σ_f and the driving value $\sigma_f^{(0)}$, suggesting there is nonetheless some moderate signal in the data about variability in f.

The densities of posterior samples of μ_f and μ_δ for simulations (A)-(C) are shown in Fig.5.3. Both μ_f and μ_δ are underestimated for each of simulations (A)-(C), regardless of the driving value used. For each of these simulations, rates of crossover are estimated quite well (e.g. see Table 5.1), while rates of gene conversion are consistently underestimated, a result we currently have no explanation for. Unlike the posterior densities of σ_f, posterior densities of μ_f and μ_δ are not between the truth and the respective μ used in the driving value. Using alternative priors in the hierarchical model, such as increasing κ_0 so that the priors on μ_f and μ_δ are nearly uniform, did not have much of an effect on posterior densities of $\vec{\mu}$.

5.3 Application to SeattleSNPs

We applied our hierarchical model to the African-American and CEPH populations of the *SeattleSNP* dataset using the parameter values of priors I-IV of Sec.5.1.3 as driving values. We sampled posterior values every 50 iterations of the MCMC chain after a

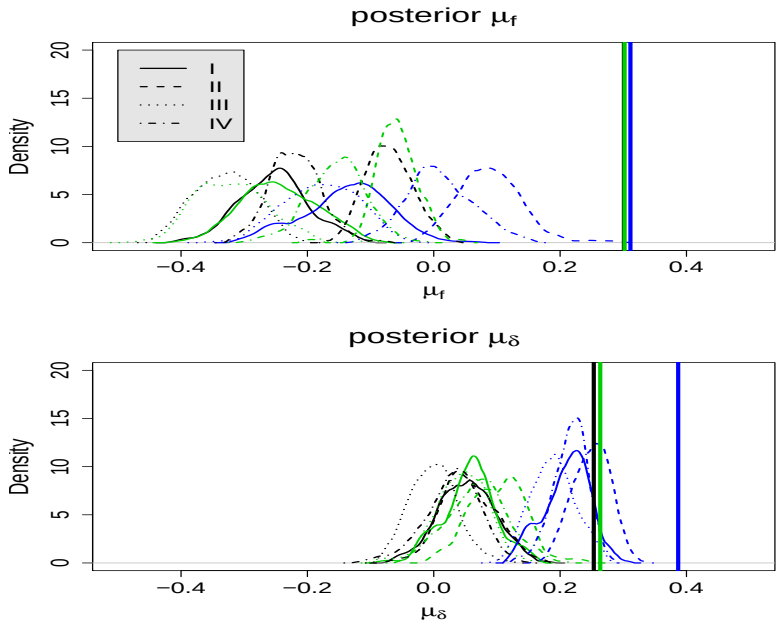


Figure 5.3: Hierarchical-model posterior samples of μ_f (top) and μ_δ (bottom), for simulations (A) (black), (B) (green), and (C) (blue), using each of priors I-IV as driving values (see plot legend). The solid vertical lines represent the true μ_f and μ_δ for each simulation. Priors I and II use $\mu_f^{(0)} = 1.0$, $\mu_\delta^{(0)} \approx$ -1.1; priors III and IV use $\mu_f^{(0)} = 0.6$, $\mu_\delta^{(0)} \approx$ -1.4 (all values out of range of the plot above). Both μ_f and μ_δ are underestimated for each simulated dataset regardless of the driving value used; both are a result of the model underestimating rates of gene conversion.

5000 run burn-in. We present the results in this section. We note that results were very similar when we ran the chain 10 times longer, 10 times shorter, or sampled 10 times as many values (results omitted). However, for each driving value, inference was subtly different between runs when the posterior samples from different PHASE seeds were used in the hierarchical model. Therefore, for each of the African-American and CEPH populations and for each ($\vec{\mu}^{(0)}$,$\Sigma^{(0)}$), we used the combined posterior samples of five different PHASE seeds, using 1000 samples per seed per gene, as our sampling values of $\log_{10} f$ and $\log_{10} \delta$ in the hierarchical model. We note, however, that the overall inference and conclusions we make would be the same using only one seed of PHASE for each driving value ($\vec{\mu}^{(0)}$,$\Sigma^{(0)}$).

5.3.1 Means of f, δ

Posterior densities of μ_f and μ_δ are shown in Fig.5.4. Posterior distributions of μ_δ appear consistent across driving values for each population, with African-Americans having a higher value. This is an expected result and likely reflects the higher effective population size of the African-American population. The median of posterior μ_f samples ranges from about 0.0-0.3, i.e. $f \approx$ 1.0-2.5, for the African-Americans and from about -0.3-0.1, i.e. $f \approx$ 0.5-1.3, for CEPH, assuming a tract length of 100bp. These values of f are similar to what was found via PHASE in Chapter 4 and slightly smaller than what was found via PAC in Chapter 3. We note from the results of the simulations in Sec.5.2.2 that these values likely underestimate the true f, perhaps by a factor of ≈2-3. When multiplied by, e.g., 2.5, the results are more similar to the results of PAC in Chapter 3, which are probably more accurate.

5.3.2 Variability in f, δ

The posterior densities of σ_f and σ_δ for *SeattleSNPs* are shown in Fig.5.5. For each population, the posterior densities appear considerably more consistent across driving values for σ_δ than for σ_f, which is not surprising based on the simulation results. The

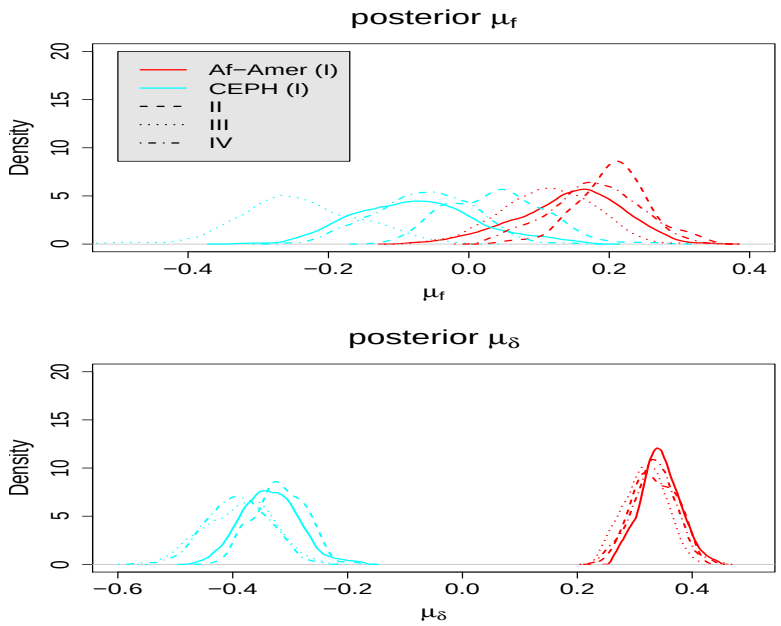


Figure 5.4: Hierarchical-model posterior samples of μ_f (top) and μ_δ (bottom), for the African-American (AA) (red) and CEPH (EA) (cyan) populations of *SeattleSNPs*, using each of priors I-IV as driving values (see plot legend). Prior I uses $\mu_f^{(0)} = 1.0$ and $\mu_\delta^{(0)} \approx$ -1.1 (AA), -1.5 (EA); prior II uses $\mu_f^{(0)} = 1.0$ and $\mu_\delta^{(0)} \approx$ -1.1 (AA), -1.6 (EA); priors III and IV use $\mu_f^{(0)} = 0.6$ and $\mu_\delta^{(0)} \approx$ -1.4 (AA), -2.0 (EA) (all values out of range of the plot above).

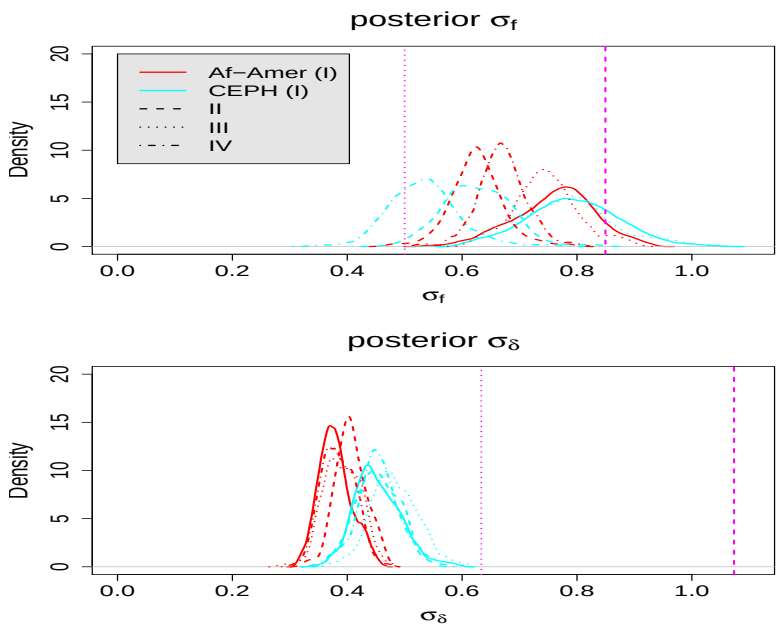


Figure 5.5: Hierarchical-model posterior samples of σ_f (top) and σ_δ (bottom), for the African-American (red) and CEPH (cyan) populations of *SeattleSNPs*, using each of priors I-IV as driving values (see plot legend). The dashed vertical lines represent the values of $\sigma_f^{(0)}$ and $\sigma_\delta^{(0)}$ used in priors I and III. The dotted vertical lines represent the values of $\sigma_f^{(0)}$ and $\sigma_\delta^{(0)}$ used in priors II and IV.

densities of σ_δ overlap substantially between populations across driving values, as a result of shared ancestry between the populaitions and/or a reflection that rates in double-strand-breaks vary at similar magnitudes for both populations. There is also considerable overlap between populations for samples of σ_f, particularly when comparing densities derived using the same $\sigma_f^{(0)}$. It is interesting that the majority of the mass for all σ_f posterior densities lie between the two different $\sigma_f^{(0)}$ values used in the four driving values. As the majority of mass for all σ_f posterior densities in our simulated datasets fell between their corresponding driving value $\sigma_f^{(0)}$ and true σ_f, it is tempting to interpret this to mean the true σ_f for both populations is somewhere in this range between 0.5-0.85. However, given that our estimates of σ_f based on the posterior densities were not very accurate in the simulation study, this remains speculative.

Still there is some support for variability in f in these data. We note that for simulation (C), the simulated dataset where σ_f is most similar to that of our data, posterior densities of σ_f did include the true value. If this variability is correct, and assuming $\log_{10} f$ values across genes are approximately normal, it suggests the upper quartile of f values across the genome is perhaps 4-14 times larger than the lower quartile of f values, compared to a range of perhaps 3-5 for such values of δ.

An alternative means to studying whether f might vary involves studying γ and ρ. If f were constant, then estimates of γ and ρ should be perfectly correlated. In Figure 5.6, we plot the results for hieararchical model estimates of $\log_{10} \gamma$ versus $\log_{10} \rho$, using the median of posterior samples of each parameter as our estimate per gene, for the African-American and CEPH data for each of the four driving values. In addition, we show the best-fit linear regression lines for each *SeattleSNPs* dataset and simulations (A)-(C).

Though estimates of $\log_{10} \gamma$ and $\log_{10} \rho$ are not perfectly correlated for simulations (A) and (B) where f is constant, note that they show a stronger correlation than the African-American data for any particular driving value. In fact, the regression lines of

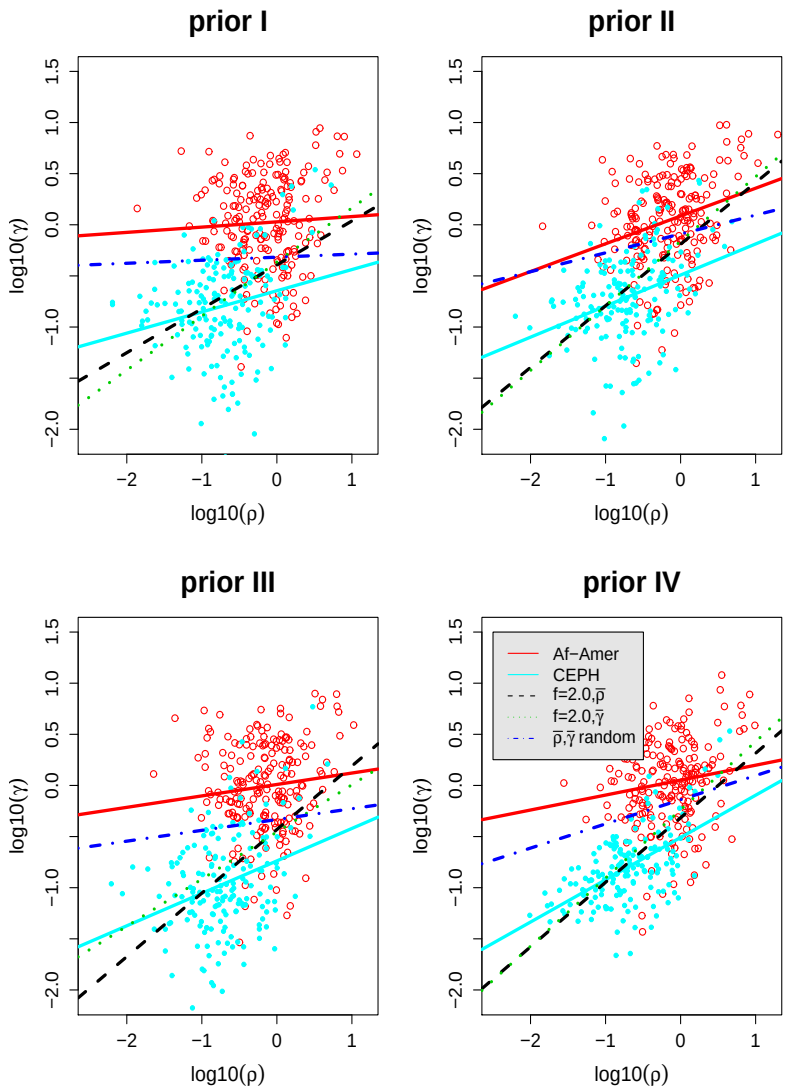


Figure 5.6: Hierarchical-model estimates of $\log_{10}\gamma$ versus $\log_{10}\rho$, using medians across samples as point estimates for each per gene, for the *SeattleSNPs* African-American ("○") and CEPH ("●") data when using priors I (top left), II (top right), III (bottom left), and IV (bottom right) as driving values. Best-fit linear regression lines of these estimates are shown for the African-American (solid red lines) and CEPH (solid blue lines) data, as well as simulations (A) (black dashed lines), (B) (green dotted lines), and (C) (blue dot-and-dashed lines) for each prior. Consistently across driving values, the slopes of these regression lines are smaller for the *SeattleSNPs* data compared to simulations (A) and (B), for which the true γ and ρ are perfectly correlated, and larger than simulation (C), for which the true γ and ρ are independent.

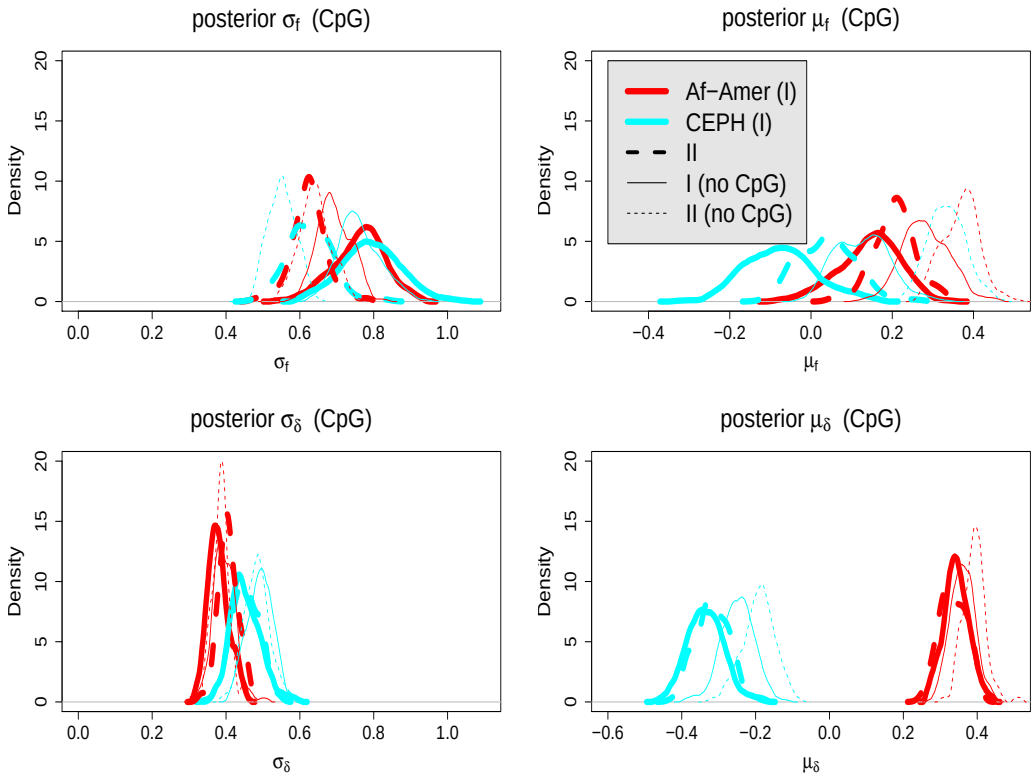
the *SeattleSNPs* data appear to be debatably more comparable to that of simulation (C), for which simulated γ and ρ were independently assigned as described in Sec.4.3. In addition there is considerably more variability around the regression line for the *SeattleSNPs* data than for either of simulations (A) and (B) (results omitted).

Repeat Mutations

Variability in estimated f could actually be an artifact of variability in rates of repeat mutation. Repeat mutation has a very similar effect on linkage disequilibrium (LD) patterns as gene conversion, so we might expect regions with increased rates of mutation to affect our estimates of γ while having little effect on ρ. Our model might therefore overestimate f in such regions. Methylated CpG groups are known mutation hotspots, and – as different genes have different amounts of CpGs – they also have variable rates of mutation that might elevate our σ_f estimates. We removed all CpG regions of our dataset and re-ran our algorithm on the reduced dataset using each of priors I and II as driving values. For this analysis we used only 1000 posterior samples of a single seed of PHASE per driving value as our sampling set in the hiearchical model. The posterior densities of σ_f, σ_δ, μ_f, and μ_δ are shown in Fig. 5.7. None of these densities seem to be greatly affected by the removal of these sites, suggesting that genome-wide variability in rates of mutation is not significantly contributing to our estimated genome-wide variability in f.

Genotyping Error

Analagous to repeat mutation, genotyping error can have a similar effect on patterns of LD as gene conversion. Our estimated variability in f may therefore be inflated if the level of genotyping error varies across the genome. This could well be the case, as some regions of the genome may be more difficult to amplify and identify than other regions.



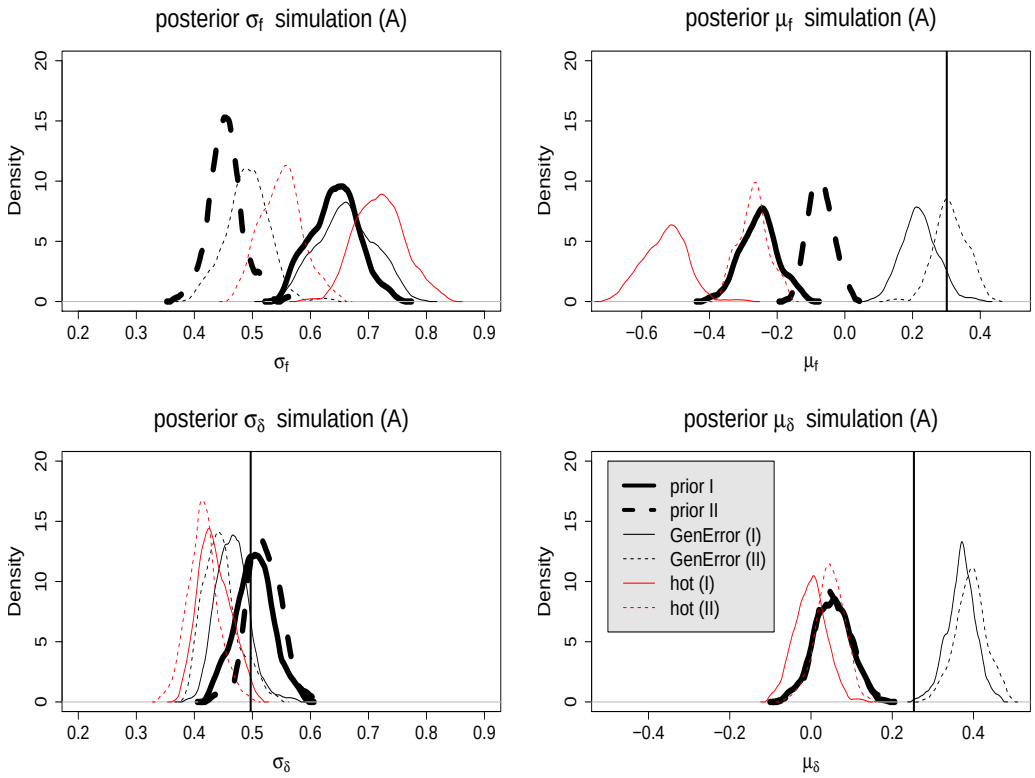
(a) posterior σ_f,σ_δ

(b) posterior μ_f,μ_δ

Figure 5.7: Posterior densities of σ_f (top left), σ_δ (bottom left), μ_f (top right), and μ_δ (bottom right), for the African-American and CEPH populations of *SeattleSNPs*. Results are shown for all data (thick lines) and for data in which all CpG sites have been removed (thin lines), using each of PHASE priors I and II as driving values (see figure legend). Note that inference changes little before and after CpG removal, in particular beyond the variability in using different driving values.

We simulated variable genotyping error in simulated datasets (A), (B), and (C) by changing the SNP configurations of a few selected genotypes per gene. We did this in a manner that attempts to mimic the expected overall genotyping error in the *SeattleSNP* data. For each gene, the proportion of SNPs with error was randomly selected from a Uniform(0,x). For each SNP within a gene selected for error, we chose the number of genotypes, =1,2,3,4,... out of n individuals, to be altered based on probabilities thought to be representative of error rates in *SeattleSNPs*, kindly provided by M.Eberle. Given a number of genotypes to be altered, we preferentially selected homozygous genotypes and miscalled them heterozygous, based on rates provided by M.Eberle. Finally, x was chosen so that the overall genotyping error rate was ≈0.5%, the overall error rate expected in the *SeattleSNPs* data (Ptak et al., 2004). This gives a different rate of genotyping error for each gene, a scenario that might act to mimic variability in f.

We applied this genotyping error scheme to each of simulations (A), (B), and (C), using x=0.2. We ran these altered datasets through the hierarchical model sampling scheme using each of priors I and II as driving values. For each driving value, we used 1000 posterior samples from PHASE as our sampling distribution in the hierarchical model. As the effects of genotyping error on the posterior distributions of Σ and $\vec{\mu}$ are similar for each simulated dataset, we present the results for simulation (A) only in Fig.5.8. This simulated dataset had 2728 total genotyping errors (0.61%), which is slightly higher than what we expect for the *SeattleSNPs* data. The variable genotyping error per gene has very little effect on posterior densities of σ_f and σ_δ (Fig.5.8-(a)). However, the addition of genotyping error significantly alters our posterior densities of μ_f and μ_δ (Fig.5.8-(b)), increasing each by ≈2.5-fold. This increase reflects how such error acts to increase estimates of γ. However, the levels of genotyping error used here do not seem to be high enough to contribute much to our estimated variability of f or δ.



(a) posterior σ_f, σ_δ

(b) posterior μ_f, μ_δ

Figure 5.8: Hierarchical model posterior samples of (a) σ_f (top) and σ_δ (bottom) and (b) μ_f (top) and μ_δ (bottom), simulation A, after incorporating either genotyping error (thin black lines) or multiple DSB hotspots (thin red lines) using PHASE priors I and II as driving values (see figure legend). The posterior densities of each parameter are also shown for the orginal dataset (thick black lines). The vertical lines represent the true simulated values of each parameter (true $\sigma_f = 0$). The inclusion of genotyping error or multiple DSB hotspots has little effect on posterior samples of σ_f and σ_δ. However, both scenarios affect posterior densities of μ_f and μ_δ in predictable ways.

Multiple Hotspots

We also explore the effect of multiple DSB hotspots per gene on our inference. To keep our model somewhat parsimonious, we have assumed that each gene has at most one DSB hotspot. There is some evidence that particular genes of *SeattleSNPs* may have multiple such hotspots (Crawford et al. (2004), Fearnhead and Smith (2005)). As hotspot regions appear to typically be narrow, e.g. 1-2kb, they often contain few if any SNPs, rendering gene conversion estimation in such regions difficult. In contrast, patterns of LD between SNPs flanking a hotspot can be substantially affected by historical crossover events in the hotspot, and estimated crossover rates can pick up this effect. Thus the effect of multiple DSB hotspots per gene will likely increase our estimates of background ρ while having little effect on estimates of γ. (Specifically, under our assumption of one hotspot per gene, we expect PHASE to find the hotspot with the highest intensity and then average rates of crossover over the remaining sequence, including other hotspots, for its estimate of that gene's background ρ.) If the number and intensities of hotspots vary among genes, this could lead to a substantial increase in variability amongst genes of our estimate of background ρ, which could in turn increase our estimate of σ_f.

To explore this possibility, we simulated a Poisson number of hotspots per gene with rate 1 per 30kb, this rate based on results of Fearnhead and Smith (2005) and Jeffreys et al. (2005). Each hotspot was of length 1kb, with $\log_{10}$-intensity sampled from a Uniform(1,2.5). This intensity distribution restricts hotspots to have crossover and gene conversion rates between 10-316 times that of background rates, with 50% of hotspots expected to have intensities between 10 and 32. These simulations were done via the use of the program msHOT described in Chapter 6.

We simulated a dataset of 204 genes that mimics simulation (A) but with the addition of the above hotspot scheme. This gave 77% of the total crossover occuring in 3.3% of the total sequence, these numbers closely matching – if not more extreme

than – current observations (Myers et al., 2005). Hotspots per gene ranged from zero to nine. Using priors I and II as driving values, we ran this dataset through the hierarchical model sampling scheme. For each prior, we used 1000 posterior samples from one seed of PHASE as our sampling distribution in the hierarchical model sampling scheme. The posterior densities of σ_f, σ_δ, μ_f, and μ_δ are shown in Fig.5.8, along with the results for simulated dataset (A) without hotspots. There is a substantial decrease in the values of posterior μ_f in the multiple hotspot scenario compared to the otherwise identical no hotspot scenario (Fig.5.8-(b)). A decrease in estimated μ_f is expected if estimated background ρ for genes with multiple hotspots increases while estimated γ for such genes does not. In contrast, the effect on posterior μ_δ appears negligible (Fig.5.8-(b)). In addition, we note that the addition of multiple hotspots per gene, as with variable genotyping error rates, does not appear to substantially affect our inference on variability in f and δ (Fig.5.8-(a)).

5.3.3 Correlations with Sequence Features

Crawford et al. (2004) found that their estimated background crossover rates appeared to be correlated with sequence features such as SNP density and %G+C content. Other analyses, including pedigree analysis by Kong et al. (2002), support this finding. We also found a correlation between our estimated background crossover rates and each of SNP density and %G+C content. We wanted to investigate whether this observed correlation between crossover and these two sequence features is primarily a result of assocation between δ and these features or f and these features. Analagous to Crawford et al. (2004), we performed a multiple linear regression for our posterior samples of each of $\log_{10}\rho$, $\log_{10}\delta$, and $\log_{10} f$ from the hierarchical model on SNP density and %G+C content, for the *SeattleSNPs* data. (The sequence feature values were kindly provided by T.Bhangale.)

As simulation (A) simulates ρ per gene equivalent to our estimated values of ρ from Chapter 4 for the African-Americans, we would expect to see a correlation between

estimated ρ and each of SNP density and %G+C content for this simulated dataset. Furthermore, as f is constant in (A), this correlation should be entirely attributable to an association between estimated δ and these features. Therefore we performed the same multiple linear regressions as above for this simulated dataset, to see what inference our model makes when we know the truth. For comparison, we use the same procedure on simulated datasets (B) and (C), for which no such strong associations should exist.

To summarize the associations between each of ρ, δ, and f with SNP density and %G+C content, we produced 101 samples of estimates across all genes from our posterior distribution, which produced 101 estimated coefficients for each of SNP density and %G+C content. This approach accounts for uncertainty in estimated values of ρ, δ, and f, but not in uncertainty of the estimated coefficient for each sampled value of ρ, δ, and f. An alternative is to ignore variability in estimated rates of ρ, δ, and f, and examine uncertainty in estimating the coefficients, e.g. via standard confidence intervals of the coefficients from linear regressions of median values of $\log_{10}\rho$, $\log_{10}\delta$, and $\log_{10} f$ on SNP density and %G+C content. We tried this as well, and doing so results in the same conclusions as provided below.

SNP Density

Table 5.2 presents the median and 95% credible intervals of the estimated coefficients for SNP density, from linear regressions using the *SeattleSNPs* data. The coefficient of SNP density for ρ is large and positive, similar to what Crawford et al. (2004) found, with the 95% credible intervals consistently excluding 0. Similarly, the credible intervals for f consistently exclude 0, with coefficients of SNP density large and negative. In contrast, the credible intervals for δ consistently include 0. This suggests that the association between ρ and SNP density is being driven by an association between f and SNP density and not δ and SNP density. This conclusion is consistent among populations and driving values. For the analyses in which CpG's were removed,

this inference is not affected.

Table 5.2: Coefficient for SNP density in Multiple Linear Regression across 101 samples for each of $\log_{10}\rho$, $\log_{10} f$, $\log_{10}\delta$ on SNP-density and %G+C-content, *SeattleSNPs*, using priors I-IV as driving values in the hierarchical model sampling scheme.

		ρ	δ	f
I				
	African-American	79 (66 - 97)	2 (-15 - 25)	-121 (-160 - -80)
	European	55 (29 - 78)	13 (-21 - 37)	-104 (-153 - -52)
	Af-Amer (No CpG)	86 (71 - 100)	9 (-6 - 25)	-123 (-149 - -92)
	Euro (No CpG)	56 (31 - 78)	16 (-13 - 43)	-77 (-116 - -42)
II				
	African-American	74 (62 - 93)	9 (-14 - 29)	-103 (-137 - -70)
	European	57 (30 - 78)	12 (-9 - 41)	-82 (-118 - -41)
	Af-Amer (No CpG)	83 (72 - 100)	7 (-12 - 26)	-105 (-136 - -91)
	Euro (No CpG)	53 (35 - 73)	11 (-12 - 34)	-74 (-107 - -48)
III				
	African-American	80 (64 - 101)	6 (-14 - 25)	-115 (-155 - -75)
	European	57 (30 - 86)	19 (-5 - 54)	-85 (-126 - -20)
IV				
	African-American	76 (63 - 92)	3 (-11 - 30)	-113 (-140 - -77)
	European	57 (32 - 79)	25 (-3 - 51)	-74 (-111 - -41)

Table 5.3 presents the median and 95% credible intervals of the coefficients for SNP density, from linear regressions using simulated datasets (A),(B), and (C). For simulation (A), in which the association between estimated ρ and SNP density should be entirely due to an assocation between δ and SNP density, we still see a strong association between f and SNP density for which the 95% credible interval consistently excludes 0. This is likely due to a strong correlation between estimated ρ and estimated f. The dependence between estimated ρ and estimated f appears to be compounded by the introduction of multiple DSB hotspots in simulations analogous to (A). In contrast, conclusions are not greatly affected by the introduction of genotyping error to (A).

There is the possibility that some of the correlation between SNP density and f may be attributable to our assumption that gene conversion events affect at most one SNP per event. In Sec.2.3.3, we found for simulated regions in which this assumption does not hold, our model tends to underestimate γ and overestimate ρ. This would result in underestimation of f. It seems probable that regions with higher SNP densities would be more likely to violate this assumption. Therefore we thinned each *SeattleSNPs* and simulated dataset by locating all SNPs within 250bp of another SNP and deleting one SNP from each pair, preferentially deleting SNPs with fewer copies of the rare allele. We then generated 1000 posterior samples of each parameter from one seed of PHASE for each of priors I and II and ran these samples in the hierarchical model sampling scheme.

Table 5.3: Coefficient for SNP density in Multiple Linear Regression across 101 samples for each of $\log_{10}\rho$, $\log_{10} f$, $\log_{10}\delta$ on SNP-density and %G+C-content, simulations, using priors I-IV as driving values in the hierarchical model sampling scheme.

		ρ	δ	f
I				
	(A) f=2.0, $\bar{\rho}$	75 (63 - 86)	47 (31 - 66)	-72 (-103 - -36)
	(B) f=2.0, $\bar{\gamma}$	-13 (-30 - 2)	-16 (-31 - 3)	-21 (-61- 6)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	25 (13 - 39)	-14 (-27 - 3)	-75 (-103 - -44)
	(A) geno-error	76 (63 - 93)	45 (29 - 58)	-44 (-73 - -18)
	(A) multi-hot	77 (62 - 91)	47 (30 - 63)	-110 (-128 - -85)
II				
	(A) f=2.0, $\bar{\rho}$	69 (57 - 85)	49 (36 - 65)	-40 (-61 - -22)
	(B) f=2.0, $\bar{\gamma}$	-6 (-21 - 7)	-21 (-38 - -4)	-32 (-55 - -12)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	26 (14 - 37)	-12 (-28 - 2)	-65 (-86 - -43)
	(A) geno-error	82 (71 - 93)	44 (32 - 58)	-54 (-75 - -33)
	(A) multi-hot	74 (63 - 90)	38 (24 - 53)	-109 (-132 - -85)
III				
	(A) f=2.0, $\bar{\rho}$	68 (53 - 86)	52 (38 - 69)	-44 (-71 - -17)
	(B) f=2.0, $\bar{\gamma}$	-9 (-21 - 4)	-18 (-31 - -5)	-23 (-59 - 12)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	28 (14-37)	-2 (-19 - 11)	-68 (-100 - -35)
IV				
	(A) f=2.0, $\bar{\rho}$	72 (58 - 84)	57 (39 - 71)	-42 (-68 - -17)
	(B) f=2.0, $\bar{\gamma}$	-7 (-21 - 3)	-20 (-37 - -4)	-23 (-49 - 6)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	30 (17 - 42)	-3 (-16 - 10)	-74 (-96 - -43)
truth				
	(A) f=2.0, $\bar{\rho}$	71.654	71.654	–
	(B) f=2.0, $\bar{\gamma}$	-22.793	-22.793	–
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	31.541	-12.209	-54.334

Table 5.4: Coefficient for SNP density in Multiple Linear Regression across 101 samples for each of $\log_{10}\rho$, $\log_{10} f$, $\log_{10}\delta$ on SNP-density and %G+C-content, *SeattleSNPs* and simulations (A)-(C), using priors I and II as driving values in the hierarchical model sampling scheme, after thinning SNPs to one every 250bp.

		ρ	δ	f
I				
	African-American	76 (62 - 95)	3 (-20 - 20)	-110 (-145 - -74)
	European	48 (21 - 68)	9 (-16 - 38)	-50 (-96 - -7)
	(A) f=2.0, $\bar{\rho}$	76 (64 - 92)	56 (38 - 76)	-36 (-62 - -4)
	(B) f=2.0, $\bar{\gamma}$	-23 (-42 - -4)	-28 (-47 - -7)	-7 (-40 - 15)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	10 (-6 - 24)	-8 (-26 - 8)	-35 (-74 - -1)
	(A) multi-hot	96 (82 - 110)	53 (39 - 73)	-101 (-124 - -78)
II				
	African-American	71 (53 - 85)	9 (-8 - 30)	-83 (-108 - -59)
	European	45 (26 - 73)	12 (-13 - 37)	-51 (-88 - -22)
	(A) f=2.0, $\bar{\rho}$	77 (60 - 88)	49 (33 - 65)	-40 (-63 - -20)
	(B) f=2.0, $\bar{\gamma}$	-26 (-46 - -14)	-29 (-48 - -8)	0 (-25 - 23)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	15 (-4 - 26)	-18 (-38 - -4)	-51 (-81 - -22)
	(A) multi-hot	68 (53 - 84)	44 (30 - 57)	-43 (-65 - -15)
truth				
	(A) f=2.0, $\bar{\rho}$	82.494	82.494	–
	(B) f=2.0, $\bar{\gamma}$	-22.483	-22.483	–
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	25.999	-14.742	-51.720

The results for the coefficient of SNP density from multiple linear regressions on posterior samples from the hierarchical model, as outlined above, are shown in Table 5.4. Thinning the data indeed reduces the size of the coefficient for simulation (A) by roughly a factor of two while not greatly affecting the coefficient for δ. For the African-American data of *SeattleSNPs*, no such reduction is observed. However, we can not rule out that multiple DSB hotspots per gene is playing a role in the association between estimated f and SNP density, as the coefficient between the two suggests strong association in the multiple hotspots simulation, even though f is constant.

Despite some irregularities in the association between f and SNP density, we do find a strong lack of association between δ and SNP density in all scenarios, for both the African-American and CEPH data of *SeattleSNPs*. Simulations suggest our model consistently captures associations between δ and SNP density if this association is in reality contributing substantially to the association between ρ and SNP density. Therefore, our results suggest the correlation between ρ and SNP density may indeed be more attributable to a correlation between f and SNP density than to δ and SNP density. Such an association between f and SNP density has been previously reported in a study of *Drosophila* by Langley et al. (2000). In this study, regions with more heterozygosities were found to preferentially resolve DSBs as crossovers, similar to what our results suggest here. To our knowledge, this is the first observation of such a pattern using human data.

Table 5.5: Coefficient for %G+C content in Multiple Linear Regression across 101 samples for each of $\log_{10}\rho$, $\log_{10} f$, $\log_{10}\delta$ on SNP-density and %G+C-content, *SeattleSNPs*, using priors I-IV as driving values in the hierarchical model sampling scheme.

		ρ	δ	f
I				
	African-American	2.1 (1.7 - 2.6)	0.9 (0.4 - 1.3)	-2.0 (-2.7 - -1.1)
	European	2.5 (1.7 - 3.0)	1.7 (1.1 - 2.3)	-1.0 (-2.4 - 0.2)
	Af-Amer (No CpG)	2.1 (1.7 - 2.5)	0.9 (0.5 - 1.3)	-1.6 (-2.3 - -1.0)
	Euro (No CpG)	1.9 (1.4 - 2.4)	1.8 (1.3 - 2.5)	0.0 (-0.9 - 0.8)
II				
	African-American	2.1 (1.7 - 2.4)	1.0 (0.5 - 1.3)	-1.7 (-2.3 - -0.9)
	European	2.3 (1.7 - 3.0)	1 .6 (1.0 - 2.3)	-1.2 (-1.9 - -0.2)
	Af-Amer (No CpG)	2.0 (1.7 - 2.4)	1.2 (0.8 - 1.5)	-1.2 (-1.7 - -0.7)
	Euro (No CpG)	2.2 (1.5 - 2.6)	1.6 (1.0 - 2.3)	-0.7 (-1.5 - 0.0)
III				
	African-American	2.2 (1.7 - 2.6)	0.9 (0.4 - 1.3)	-1.9 (-3.0 - -1.2)
	European	2.3 (1.6 - 2.9)	1.7 (1.0 - 2.5)	-1.3 (-2.5 - 0.3)
IV				
	African-American	2.1 (1.7 - 2.5)	0.9 (0.4 - 1.3)	-1.8 (-2.6 - -1.0)
	European	2.2 (1.7 - 2.8)	1.8 (1.4 - 2.5)	-0.8 (-1.7 - 0.1)

%G+C Content

Table 5.5 presents the median and 95% credible intervals of the estimated coefficients for %G+C content from the linear regressions using the *SeattleSNPs* data. The coefficient of %G+C content for ρ is large and positive, similar to what Crawford et al. (2004) found, with the 95% credible intervals consistently excluding 0. The same is true for δ. For the African-American dataset, credible intervals for f consistently exclude 0, but this is not the case with the CEPH data for some of the driving values. Removing CpG sites does not appear to greatly affect inference.

Table 5.6 presents the median and 95% credible intervals of the coefficients of %G+C content from the linear regressions using simulated datasets (A), (B), and (C). For simulation (A), in which the association between estimated ρ and %G+C content should be entirely attributable to an association between estimated δ and %G+C content, the coefficient for δ is consistently higher than that for f, with 95% credible intervals mutually exclusive. In addition, 95% credible intervals for δ consistently exclude 0, while 95% credible intervals for f include 0 when using priors II-IV as driving values.

The introduction of genotyping error into simulated dataset (A) decreases the coefficient of δ by nearly a factor of two while scarcely changing the coefficient of ρ. This is likely the result of the inclusion of genotyping error introducing noise in γ estimation, which diminishes the correlation between estimated γ and %G+C content and hence estimated δ and %G+C content. Similarly, simulating multiple DSB hotspots for some genes appears to diminish the correlation between δ and %G+C content by introducing noise into estimated ρ. Still neither of these effects alter the coefficients enough that the 95% credible intervals for δ include 0.

Table 5.6: Coefficient for %GC-content in Multiple Linear Regression across 101 samples for each of $\log_{10}\rho$, $\log_{10} f$, $\log_{10}\delta$ on SNP-density and %G+C-content, simulations, using priors I-IV as driving values in the hierarchical model sampling scheme.

		ρ	δ	f
I				
	(A) f=2.0, $\bar{\rho}$	2.2 (1.8 - 2.5)	2.4 (2.1 - 2.8)	0.8 (0.1 - 1.5)
	(B) f=2.0, $\bar{\gamma}$	0.3 (0.0 - 0.6)	0.0 (-0.3 - 0.3)	0.0 (-0.6 - 0.8)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	-0.4 (-0.7 - -0.1)	0.4 (0.1 - 0.7)	1.7 (1.1 - 2.4)
	(A) geno-error	2.1 (1.8 - 2.4)	1.3 (1.0 - 1.7)	-1.2 (-1.7 - -0.6)
	(A) multi-hot	1.1 (0.7 - 1.5)	1.2 (0.8 - 1.5)	0.1 (-0.4 - 0.8)
II				
	(A) f=2.0, $\bar{\rho}$	2.3 (2.0 - 2.7)	2.3 (1.9 - 2.6)	-0.1 (-0.6 - 0.3)
	(B) f=2.0, $\bar{\gamma}$	-0.1 (-0.4 - 0.2)	0.1 (-0.3 - 0.4)	0.4 (-0.1 - 0.9)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	-0.4 (-0.7 - -0.1)	0.4 (0.1 - 0.7)	1.4 (0.9 - 2.0)
	(A) geno-error	2.0 (1.7 - 2.4)	1.5 (1.2 - 1.8)	-0.8 (-1.4 - -0.2)
	(A) multi-hot	1.3 (1.0 - 1.7)	1.1 (0.8 - 1.4)	-0.3 (-0.9 - 0.0)
III				
	(A) f=2.0, $\bar{\rho}$	2.4 (2.0 - 2.8)	2.4 (2.2 - 2.8)	0.3 (-0.1 - 1.0)
	(B) f=2.0, $\bar{\gamma}$	0.2 (-0.2 - 0.5)	0.2 (-0.2 - 0.5)	0.2 (-0.5 - 0.8)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	-0.4 (-0.7 - -0.1)	0.5 (0.2 - 0.8)	1.5 (0.9 - 2.0)
IV				
	(A) f=2.0, $\bar{\rho}$	2.4 (2.1 - 2.8)	2.3 (2.0 - 2.6)	-0.3 (-0.7 - 0.2)
	(B) f=2.0, $\bar{\gamma}$	0.1 (-0.2 - 0.6)	0.2 (-0.3 - 0.5)	0.1 (-0.3 - 0.6)
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	-0.4 (-0.7 - -0.1)	0.4 (0.0 - 0.7)	1.3 (0.8 - 1.8)
truth				
	(A) f=2.0, $\bar{\rho}$	1.889	1.889	–
	(B) f=2.0, $\bar{\gamma}$	0.397	0.397	–
	(C) $\bar{\rho}$,$\bar{\gamma}$ random	-0.374	0.158	0.771

Judging via simulation that our model appears to relatively accurately assess associations between each of ρ, δ, and f with %G+C content, we can perhaps rely on our results for the *SeattleSNPs* data in Table 5.5. Our application to *SeattleSNPs* suggests that the association between ρ and %G+C content is attributable to an association between δ and %G+C content *and* f and %G+C content in the African-American data. For the CEPH population, however, the association between ρ and %G+C content appears to only be attributable to an association between δ and %G+C content.

The association between δ and %G+C content may be the result of a phenomenon known as *biased gene conversion* (BGC) (e.g. Marais (2003)). The BGC hypothesis predicts that for SNP locations on homologous chromosomes with a G or C on one chromosome and an A or T on the other chromosome, double-strand-breaks preferentially occur on the A or T allele and are repaired using the G or C one. Therefore, regions with high DSB rates might create a drive towards higher relative proportions of G+C content. However, some caution should be exercised here, as the majority of our genome content, and hence G+C content, evolved long ago, i.e. before humans diverged from other mammals, while the rates of δ we have measured here have occured in relatively recent biological time, i.e. in the history of humans. Thus we are not comparing these phenomena on the same time scale.

We are unaware of what might cause f and %G+C content to be associated, though our results suggest it is an inverse correlation.

Why there might be this difference between populations is unclear. As the CEPH data contains fewer SNPs than the African-American data, it could simply be that the CEPH results are less accurate and that the descrepancy is noise in estimation. Another possibility is that there is an external factor not accounted for in our model that is contributing to these results. Yet another possibility is a geniune biological difference in recombination machinery between the two populations. In support of this latter possibility, Fearnhead and Smith (2005), similarly using patterns of LD for

inference, found evidence that several hotspots among genes in *SeattleSNPs*, perhaps as many as 30%, are not shared between the African-American and CEPH populations. However, for now it is difficult to say which of these options is the most likely.

5.4 Summary

In this chapter, we described a hierarchical model that incorporates posterior estimates of PHASE into an importance sampling scheme in a manner that allows us to estimate genome-wide variability in f and δ. Via simulation study, we found that our model is able to pinpoint the variability in δ with some accuracy, but has more difficulty in determining variability in f. These difficulties notwithstanding, we found that our estimated f varies substantially, at a similar level as the variability in estimated δ, across the genome in the *SeattleSNPs* dataset. We also attempted to break down the previously observed correlation between estimated rates of crossover and the sequence features SNP density and %G+C content, to see if these correlations were primarily attributable to correlations between δ and these features or between f and these features. We found that the correlation between ρ and SNP density appears to be most attributable to a correlation between f and SNP density and not δ and SNP density. In contrast, we found that the correlation between ρ and %G+C content appears to be roughly equally attributable to each of δ and f. We are unaware of any previous observations of this kind for population data in humans.

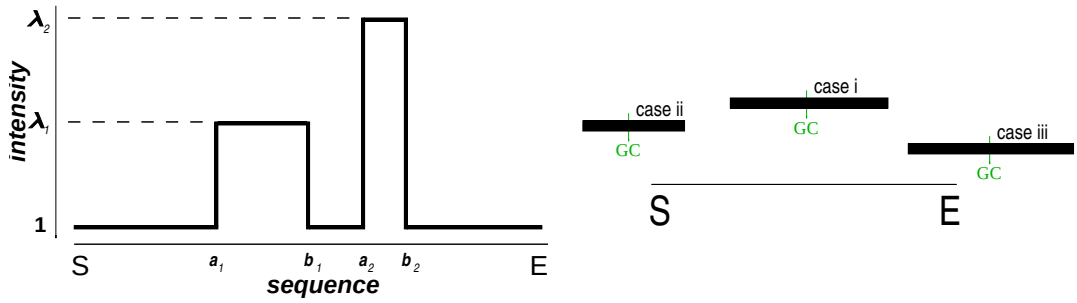
Chapter 6

SIMULATING CROSSOVER AND GENE CONVERSION HOTSPOTS

In Chapter 5, we make use of a simulation program that incorporates crossover and gene conversion hotspots into simulated population haplotype data. To do so, we wrote an extension to Richard R. Hudson's "program for generating samples under neutral models", i.e. the ms simulator (2002), a widely-used program for simulating genetic variation data for randomly-sampled haplotypes from a population. The ms program allows the user to specify various aspects of population demography, including population sizes and migration patterns, and factors governing evolution, including mutation, crossover, and gene conversion rates. However, it presently does not allow for variation in recombination rates. We have incorporated both crossover and gene conversion hotspots into a freely available, updated simulator called msHOT. We describe the principal components of the program in this chapter.

6.1 Model

The current implementation of ms (Hudson, 2002) allows the user to specify the population-scaled rate of crossing-over, ρ, and the relative rate of gene conversion to crossover, f, for the genetic region to be simulated. Here $\rho = 4N_0r$, where N_0 is the current diploid population size and r is the probability of a crossover occuring in the region in a single transmission from parent to offspring, and $f = g/r$, where g denotes the probability of a gene conversion initiating in the region of interest in a single transmission from parent to offspring (Wiuf and Hein, 2000). Since r and g are typically small, on the order of 1-10cM/Mb, dividing these parameters by the



(a) recombination variation in $[S,E]$ (b) the three distinct "types" of gc event

Figure 6.1: (a) Illustration of varying crossover and/or gene conversion intensities in a genetic region $[S,E]$. Here the crossover (respectively gene conversion) probability r_{bp} (respectively g_{bp}) is increased by a multiple λ_1 in $[a_1,b_1]$ and by a multiple λ_2 in $[a_2,b_2]$. (b) Illustration of the three distinct gene conversion types that can influence variation in the genetic region $[S,E]$. The grey vertical lines represent the initiation point of each gc event, and the black horizontal bars represent the *tract length* of each of the gc events.

sequence length of the genetic region gives, respectively, the crossover probability per basepair, r_{bp}, and the gene conversion probability per basepair, g_{bp}.

Our modification msHOT allows the user to insert as many non-overlapping crossover hotspots and non-overlapping gene conversion hotspots along a sequence as they wish by specifying the locations and intensities for each. Specifically, incorporating H crossover hotspots requires the user to specify a left endpoint a_h, right endpoint b_h, and intensity λ_h for each, $h = 1, ..., H$. Inside hotspot h, the probability of a crossover occuring between two adjacent basepairs in a single transmission from parent to offspring is $\lambda_h r_{\text{bp}}$. Outside any hotspot, this probability is r_{bp}. Similarly, incorporating $\tilde{H}$ gene conversion hotspots requires the user to specify a left endpoint $\tilde{a}_h$, right endpoint $\tilde{b}_h$, and intensity $\tilde{\lambda}_h$ for each, $h = 1, ..., \tilde{H}$. Inside gene conversion hotspot h, the probability of a gene conversion intitation between two adjacent basepairs in a single transmission from parent to offspring is $\tilde{\lambda}_h g_{\text{bp}}$. Outside any hotspot, this probability is g_{bp} (see Figure 6.1-a). Gene conversion hotspots may overlap with crossover hotspots.

In Hudson's ms, gene conversion (gc) events initiate at some basepair, which is assumed to form the left-point of the region affected by the gc. The right-point is then determined by the basepair length of the region affected by the gc, i.e. the *tract length*, which is assumed to have a geometric distribution with user-specified mean. This difference in the treatment of the left and right endpoints causes some asymmetry when the rate of gc initiation is allowed to vary along the region. To deal with this, we changed the model to assume that gc events initiate at some point and then spread both right and left independently according to geometric distributions with user-specified mean t^*. Thus, in our model, the tract length is the sum of two independent geometric distributions with mean t^*. Incidentally, this is also perhaps a better representation of the limited knowledge available on the biology underlying gc events than the formulation in the original ms (Szostak et al. (1983), Jeffreys and Neumann (2002)).

6.2 Computation

The basic algorithm of msHOT is as described in Hudson (1983). In brief, ms generates ancestral recombination graphs for a sample of chromsomes by stochastically determining "events" to occur on the ancestral material of the chromosomes going back in time, until all the material has coalesced into a common ancestor. We refer to any individual segment of this ancestral material as an "ancestral segment." Potential "events" include the coalesence of two such segments or a recombination event, i.e. crossover or gene conversion, occuring in a single segment. Incorporating hotspots involves changing the rates at which these recombination events occur, as described below. The consequences of these events, which involve splitting ancestral segments, are not changed by the introduction of hotspots and are already dealt with in Hudson's code. We therefore do not describe this process in detail.

The rate of each possible recombination event, backwards in time, is determined by computing the probability of the event occurring in a single generation *forwards* in time, and multiplying this by $4N_0$. We therefore focus on computing the relevant probabilities forwards in time. In the following we use $[S, E]$ to denote an ancestral segment beginning at S and ending at E.

crossover

Assume $[S, E]$ contains H crossover hotspots, each with left endpoint a_h, right endpoint b_h, and intensity λ_h, $h = 1, ..., H$. Under the model described above, the probability of a crossover initiating at any particular basepair $z \in [S, E]$ is:

$$\Pr(\text{crossover at } z) \quad = \quad \left[1 + \sum_{h=1}^{H} I_{z \in [a_h, b_h]}(\lambda_h - 1)\right] r_{\text{bp}}. \tag{6.1}$$

Here $I_{z \in [a_h, b_h]}$ is an indicator function, taking the value 1 if z is in crossover hotspot h and 0 otherwise. The total probability of a crossover occurring in $[S, E]$ is found by summing over z in (6.1). If a crossover is to occur in $[S, E]$, the location z is

selected with probability proportional to (6.1). (We thank E.C. Anderson for sharing an annotated version of Hudson's code edited to incorporate crossover hotspots.)

gene conversion

Each gene conversion event can be thought of as having an "initiation point" and "right" and "left" endpoints. We distinguish three types of gene conversion event that can influence patterns of genetic variation in $[S, E]$ (see Figure 6.1-b):

1. **Type i**: a gc event initiates within $[S, E]$ and has endpoints that may be either inside or outside this region.

2. **Type ii**: a gc event initiates to the left of S and has a right endpoint within $[S, E]$.

3. **Type iii**: a gc event initiates to the right of E and has a left endpoint within $[S, E]$.

The following subsections give the relative probabilities, and describe how to determine the endpoints, for each of these types of event. Assume $[S, E]$ contains $\tilde{H}$ gene conversion hotspots, each with left endpoint $\tilde{a}_h$, right endpoint $\tilde{b}_h$, and intensity $\tilde{\lambda}_h$, $h = 1, ..., \tilde{H}$.

- "Type i"

The probability of a type i event initiating at $z \in [S, E]$ is

$$\Pr(\text{type i gc at } z) \;=\; \left[1 + \textstyle\sum_{h=1}^{\tilde{H}} I_{z \in [\tilde{a}_h, \tilde{b}_h]} (\tilde{\lambda}_h - 1)\right] g_{\text{bp}}, \tag{6.2}$$

where $I_{z \in [\tilde{a}_h, \tilde{b}_h]}$ is an indicator denoting whether basepair z is in gene conversion hotspot h. If a type i event occurs, its endpoints are determined by first selecting the initiation point, z, with probabilities proportional to (6.2), and then simulating the

left and right endpoints as $z - T_1$ and $z + T_2$, with T_i independently sampled from a geometric(t^*). Note that these endpoints may fall outside $[S, E]$.

- "Type ii"

With $q \equiv \frac{t^*}{1+t^*}$, the probability that a type ii event initiates at a basepair y to the left of S (thus y is outside $[S, E]$, in contrast to z above) and has a right endpoint at $x \in [S, E]$ is given by:

$$\Pr(\text{type ii gc initiating } y\text{, ending } x) \;=\; q^{x-y}(1-q)\big[1 + \textstyle\sum_{h=1}^{\tilde{H}} I_{y\in[\tilde{a}_h,\tilde{b}_h]}(\tilde{\lambda}_h - 1)\big] g_{\text{bp}}, \tag{6.3}$$

where y ranges from $-\infty$ to S, and $I_{y\in[\tilde{a}_h,\tilde{b}_h]}$ is an indicator for whether y is in hotspot h. Note that for simplicity we have assumed, as in ms, that the chromosome has infinite length.

The total probability of a type ii event occurring in $[S, E]$ is then obtained by summing (6.3) over possible values of x and y:

$$\Pr(\text{type ii gc}) \;=\; \textstyle\sum_{x=S+1}^{E}\Big[\sum_{y=-\infty}^{S}\Big(q^{x-y}(1-q)\big[1 + \sum_{h=1}^{\tilde{H}} I_{y\in[\tilde{a}_h,\tilde{b}_h]}(\tilde{\lambda}_h - 1)\big]\Big) g_{\text{bp}}\Big] \tag{6.4}$$

If a type ii event occurs, its right endpoint, x^*, is chosen via a truncated geometric distribution (i.e. $\Pr(X^* = x^*) \propto q^{x^*-S}(1-q)$, for $x^* = S+1, ..., E$).

- "Type iii"

The type iii gc events are similar to the type ii events above, but with links starting from the right end of a ancestral segment and counting from right to left.

The source code for msHOT is available by email from garretth@stat.washington.edu, along with accompanying instructions.

Chapter 7

SUMMARY AND CONCLUSIONS

In Chapter 1, we described the double-strand-break model, as conceived by Szostak et al. (1983), and described early attempts to model gene conversion using LD data, as well as current coalescent-based attempts to model crossover. In Chapter 2, we extended one such coalescent-based LD model for crossover, the PAC likelihood of Li and Stephens (2003), to jointly estimate rates of gene conversion and crossover for haplotype population data. In Chapter 3, we assessed and corrected a bias in gene conversion estimation we found via simulation study. In Chapter 4, we incorporated our model into the PHASE algorithm (Stephens et al. (2001), Stephens and Donnelly (2003)) to jointly estimate haplotypes and rates of gene conversion and crossover using genotype data, using fixed prior parameters on ρ and f. In Chapter 5, we incorporated posterior estimates from PHASE into a hierarchical model that allows us to estimate the genome-wide variability in δ and f. In Chapter 6, we described a freely-available simulating program designed to incorporate crossover and gene conversion hotspots into simulated haplotype population data.

7.1 *Conclusions About Recombination in SeattleSNPs:* Variability and Correlations with Sequence Features

We applied our model and methodology to the *SeattleSNPs* dataset during several stages of our model's development. We estimate genomewide-average f, the relative rate of gene conversion to crossover, to be $\approx$3-4 assuming the tract length of gene conversion events is 100bp, with estimates typically similar between African-American and CEPH populations. That estimated f are similar between the two populations

is not surprising, as these two populations share much of their evolutionary history, and it is the effect of recombinations along this history that our model attempts to capture. Still it would be interesting to try and elucidate differences in f between the populations if they exist, as some studies suggest some features of recombination are rapidly – and perhaps currently – evolving (e.g. recombination hotspots; see Fearnhead and Smith (2005), Jeffreys et al. (2005)). However, it is difficult to tell how much power our method might have to do so.

One question we were particularly interested in was whether there is strong evidence that f varies from location to location in the genome. Estimated rates of crossover have been observed to vary both within and between regions that are tens of kilobases in size. However, it is unknown whether this variability is most attributable to variability in double-strand-break (DSB) rates, δ, or variability in f. While there is reason to think that variation in δ is responsible for the phenomenon of crossover hotspots, it is less clear if variation in δ might be the primary force behind variability between, e.g., the background rates of crossover between two genes. Our studies here suggest that f may vary among genes just as much as δ, outside of hotspots. However, interpretation of our results is made difficult by the lack of concordance between estimates and true values of parameters for simulated data.

We also explored correlations between recombination rates and sequence features. In particular, researchers have previously found a strong correlation between estimated rates of crossover outside of hotspots and each of SNP density and %G+C content. We aimed to see if these correlations were primarily attributable to associations between DSB rates and these features or an association between f and these features. For SNP density, our data and analysis suggest that f is the primary feature of crossover associated with variable SNP density in humans. This association between the relative rate at which DSBs are resolved as gene conversions versus as crossovers and the density of heterozygosities in a region has been previously reported in a study of *Drosophila* (Langley et al., 2000). In contrast, we found similar levels

of association between %G+C content and each of δ and f. The biased gene conversion hypothesis (e.g. Marais (2003)) is a proposed theory that might explain the association between %G+C content and double-strand-break rates. This hypothesis suggests that the DSB and %G+C content association is the result of an evolutionary effect of DSB rates to some degree driving G+C content. The association between f and %G+C content has no previous explanations we are aware of, perhaps due to the difficulty of obtaining estimates of gene conversion across several genomic regions.

7.2 Future Applications of Model

Simulations in Chapter 3 and elsewhere suggest our model is able to capture the effects of gene conversion on patterns of linkage disequilibrium with high precision for large rates of gene conversion (e.g. $\gamma \geq 4.0$/kb, when assuming a tract length of 100bp). However, in the *SeattleSNP* data we examined, our estimates do not suggest that much gene conversion is occuring at this level of approximately ten times the historical genomewide average rate of crossover. However, this dataset consists only of genic regions. As recent studies have found that rates of crossover appear to be lower in genes compared to outside them (Myers et al., 2005), perhaps there will be greater signal for gene conversions in other areas of the genome. Or, perhaps more plausibly, rates of gene conversion of this magnitude may be occuring in other organisms, such as yeast, to which we have yet to apply our model.

We are currently applying our model to the publicly available Phase II data from the *International HapMap Project* (The International HapMap Consortium (2005), www.hapmap.org), in particular to the CEPH population. These data consist of SNPs typed in 30 trios at approximately one every kilobase across the entire genome. We aim to use the genotypes of the parents for these trios in our analysis, as they are presumed independent. As these data contain an extensive amount of noncoding region in addition to genic regions, we might be able to find a greater signal for the higher rates of gene conversion for which our model gives more precise inference. This

may allow us to explore correlations we might find contributing to fluctuations in the rates of recombination features such as f more reliably. Any information that can shed light on the enigmatic double-strand-break process, which has very noted and important implications in disease association studies and the study of our evolutionary history, would be most welcome, as this process – for all its importance – remains largely a mystery.

BIBLIOGRAPHY

Andolfatto, P. and M. Nordborg (1998). The effect of gene conversion on intralocus associations. *Genetics 148*, 1397–1399.

Ardlie, K., S. Liu-Cordero, M. Eberle, M. Daly, J. Barrett, E. Winchester, E. Lander, and L. Kruglyak (2001). Lower-than-expected linkage disequilibrium between tightly linked markers in humans suggests a role for gene conversion. *Am J Hum Genet 69*(3), 582–9.

Betran, E., J. Rozas, A. Navarro, and A. Barbadilla (1997). The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. *Genetics 146*(1), 89–99.

Carpenter, A. (1984). Meiotic roles of crossing-over and of gene conversion. *Cold Spring Harb Symp Quant Biol 49*, 23–9.

Crawford, D., T. Bhangale, N. Li, G. Hellenthal, M. Rieder, D. Nickerson, and M. Stephens (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nat Genet 36*(7), 700–6.

Fearnhead, P. and N. 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*(5), 781–94.

Frisse, L., R. Hudson, A. Bartoszewicz, J. Wall, J. Donfack, and A. Di Rienzo (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet 69*(4), 831–43.

Hilliker, A., G. Harauz, A. Reaume, M. Gray, S. Clark, and A. Chovnick (1994). Meiotic gene conversion tract length distribution within the rosy locus of Drosophila melanogaster. *Genetics 137*(4), 1019–26.

Holliday, R. (1964). A mechanism for gene conversion in fungi. *Genet Res Camb 5*, 282–304.

Holloway, K., V. Lawson, and A. Jeffreys (2006). Allelic recombination and *de novo* deletions in sperm in the human β-globin gene region. *Hum. Mol. Genet. 15*(7), 1099–1111.

Hudson, R. (1983). Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol. 23*, 183–201.

Hudson, R. (2001). Two-locus sampling distributions and their application. *Genetics 159*(4), 1805–17.

Hudson, R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics 18*(2), 337–8.

Jeffreys, A., L. Kauppi, and R. Neumann (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet 29*(2), 217–22.

Jeffreys, A. and C. May (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet 36*(2), 151–6.

Jeffreys, A. and R. Neumann (2002). Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet 31*(3), 267–71.

Jeffreys, A. and R. Neumann (2005). Factors influencing recombination frequency and distribution in a human meiotic crossover hotspot. *Hum Mol Genet 14*(15), 2277–87.

Jeffreys, A., R. Neumann, M. Panayi, S. Myers, and P. Donnelly (2005). Human recombination hot spots hidden in regions of strong marker association. *Nat Genet 37*(6), 601–6.

Kauppi, L., A. Jeffreys, and S. Keeney (2004). Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet 5*(6), 413–24.

Kingman, J. (1982). The coalescent. *Stochastic Processes and Their Applications 13*, 235–248.

Kong, A., D. Gudbjartsson, J. Sainz, G. Jonsdottir, S. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. Palsson, M. Frigge, T. Thorgeirsson, J. Gulcher, and K. Stefansson (2002). A high-resolution recombination map of the human genome. *Nat Genet 31*(3), 241–7.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet 22*(2), 139–44.

Langley, C., B. Lazzaro, W. Phillips, E. Heikkinen, and J. Braverman (2000). Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. *Genetics 156*(4), 1837–52.

Li, N. and M. Stephens (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics 165*(4), 2213–33.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics 19*(6), 330–8.

Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z. Qin, H. Munro, G. Abecasis, and P. Donnelly (2006). A comparison of phasing algorithms for trios and unrelated individuals. *Am J Hum Genet 78*(3), 437–50.

McVean, G., P. Awadalla, and P. Fearnhead (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics 160*, 1231–41.

McVean, G., S. Myers, S. Hunt, P. Deloukas, D. Bentley, and P. Donnelly (2004). The fine-scale structure of recombination rate variation in the human genome. *Science 304*(5670), 581–4.

Meselson, M. S. and C. Radding (1975). A general model for genetic recombination. *Proc Nat Acad Sci USA 72*(1), 358–361.

Moens, P. (2003). The double-stranded DNA helix in recombination at meiosis. *Genome 46*(6), 936–7.

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*(5746), 321–4.

Nicolas, A., D. Treco, N. Schultes, and J. Szostak (1989). An initiation site for meiotic gene conversion in the yeast Saccharomyces cerevisiae. *Nature 338*(6210), 35–9.

Pritchard, J. and M. Przeworski (2001). Linkage disequilibrium in humans: models and data. *Am J Hum Genet 69*(1), 1–14.

Przeworski, M. and J. Wall (2001). Why is there so little intragenic linkage disequilibrium in humans? *Genet Res 77*(2), 143–51.

Ptak, S., A. Roeder, M. Stephens, Y. Gilad, S. Paabo, and M. Przeworski (2004). Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol 2*(6), e155.

Ptak, S., K. Voelpel, and M. Przeworski (2004). Insights into recombination from patterns of linkage disequilibrium in humans. *Genetics 167*(1), 387–97.

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Mol Biol Evol 6*(5), 526–38.

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*(4), 629–44.

Slightom, J., A. Blechl, and O. Smithies (1980). Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell 21*(3), 627–38.

Smith, N. and P. Fearnhead (2005). A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. *Genetics 171*(4), 2051–62.

Stahl, F. (1994). The Holliday junction on its thirtieth anniversary. *Genetics 138*(2), 241–6.

Stephens, J. C. (1985). Statistical methods of dna sequence analysis: Detection of intragenic recombination or gene conversion. *Mol Biol Evol 2*(6), 539–556.

Stephens, M. and P. Donnelly (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet 73*(5), 1162–9.

Stephens, M. and P. Scheet (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet 76*(3), 449–62.

Stephens, M., N. Smith, and P. Donnelly (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet 68*(4), 978–89.

Szostak, J., T. Orr-Weaver, R. Rothstein, and F. Stahl (1983). The double-strand-break repair model for recombination. *Cell 33*(1), 25–35.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature 437*, 1299–1320.

Wall, J. (2004). Estimating recombination rates using three-site likelihoods. *Genetics 167*(3), 1461–73.

Wall, J., L. Frisse, R. Hudson, and A. Di Rienzo (2003). Comparative linkage-disequilibrium analysis of the beta-globin hotspot in primates. *Am J Hum Genet 73*(6), 1330–40.

Watterson, G. (1975). On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol. 7*, 256–276.

Winckler, W., S. Myers, D. Richter, R. Onofrio, G. McDonald, R. Bontrop, G. McVean, S. Gabriel, D. Reich, P. Donnelly, and D. Altshuler (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science 308*(5718), 107–11.

Wiuf, C. and J. Hein (2000). The coalescent with gene conversion. *Genetics 155*(1), 451–62.

Zangenberg, G., M. Huang, N. Arnheim, and H. Erlich (1995). New HLA-DPB1 alleles generated by interallelic gene conversion detected by analysis of sperm. *Nat Genet 10*(4), 407–14.

VITA

Garrett Hellenthal was born in Anchorage, AK, the son of an oilman. He lived and worked there for six years before shipping his camp to Cody, WY, at the promise of greater prosperity. Three impressionable years of mirth passed by without so much a forethought of the days ahead. But time has a knack for making people move some more, and Garrett's time was no exception. His life resumed in Sugarland, TX, for four more years of toil and growth. Accurately sensing a change in tides, he moved along to Billings, MT, to test his fortunes again in the north. This was an unforgiving country, and after five years he again packed his modest possessions to make home in Santa Clara, CA. Four years later he was awarded a Bachelor's of Science in Mathematics at Santa Clara University. Immediately thereafter, he continued on to Seattle, WA, where he wrote this Vita five years later.