Title
Computational Methods for Meiotic Recombination Inference

Author
Yin, Junming

Publication Date
2010

Peer reviewed|Thesis/dissertation

Computational Methods for Meiotic Recombination Inference

by

Junming Yin

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Computer Science
and the Designated Emphasis in
Computational and Genomic Biology
and
Communication, Computation and Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Yun S. Song
Professor Rasmus Nielsen

Fall 2010

Computational Methods for Meiotic Recombination Inference

Abstract

Computational Methods for Meiotic Recombination Inference

by

Junming Yin

Doctor of Philosophy in Computer Science
and the Designated Emphasis in
Computational and Genomic Biology
and
Communication, Computation and Statistics

University of California, Berkeley

Professor Michael I. Jordan, Chair

Meiotic recombination is one of major evolutionary mechanisms responsible for promoting genetic variation in a population, and is important for many problems in evolutionary biology and population genetics. In this thesis, we investigate two computational problems that arise in studying meiotic recombination.

The first problem is concerned with two different type of meiotic recombination: crossovers and gene conversions. Although crossovers and gene conversions have different effects on the evolutionary history of chromosomes and therefore leave behind different footprints in the genome, it is a challenging task to tease apart their relative contributions to the observed genetic variation. In fact, the methods employed in recent studies of recombination rate variation in the human genome actually capture combined effects of crossovers and gene conversions. By explicitly incorporating overlapping gene conversion events, we propose a new statistical model that can jointly estimate the crossover rate, the gene conversion rate and the mean tract length, which is widely regarded as a very difficult problem. Our simulated results show that modeling overlapping gene conversions is crucial for improving the accuracy of the joint estimation of the aforementioned three fundamental parameters. Our analysis of real data from the telomere of the X chromosome of *Drosophila melanogaster* suggests that the ratio of the gene conversion rate to the crossover rate for the region may not be nearly as high as previously claimed.

In the second problem, we investigate the molecular basis of meiotic recombination. In mammalian organisms, recombination events tend to cluster into short 1-2 kb genomic regions known as recombination hotspots. Recent studies have mainly focused on identifying

cis and *trans*-acting elements that can modulate the activity of recombination hotspots in mammals, but most of the work neglects the role of nucleosomes, the basic unit of DNA packaging in eukaryotes. Our analysis on the correlation of H2A.Z nucleosome positions and recombination rates in *Drosophila melanogaster* suggests that nucleosome occupancy could also influence, at least partly, the activity of recombination.

Dedicated to my family

Acknowledgments

I am extremely grateful to my advisors, Michael I. Jordan and Yun S. Song, for their support and guidance from my very first steps all the way to this thesis. Mike's passion and vision for statistics and machine learning has been a great source of inspiration for me during these years, and will always be. I would like to thank him for giving me the freedom and encouragement to pursue my own interests, as well as providing me with priceless academic and professional advice. I am quite fortunate to have Yun as my advisor as well. I am grateful for his thoughtful guidance on how to make progress on vague ideas and push them to the full development. Several parts of this thesis were greatly improved by his insightful criticisms. I would also like to thank the final member of my thesis committee, Rasmus Nielsen, for his great comments and suggestions for revision.

During my stay at Berkeley, I have the fortune to have interacted with a number of excellent professors, in particular Richard Karp, Martin Wainwright, and Bill Noble. I have benefited from interactions with a large number of colleagues, with whom I have had the pleasure of studying and working: Anand Bhaskar, Alexandre Bouchard, Ma'ayan Bresler, Mu Cai, Andrew Chan, Jing Lei, Wei-Chun Kao, Gad Kimmel, Simon Lacoste-Julien, Percy Liang, Kurt Miller, XuanLong Nguyen, Guillaume Obozinski, Sriram Sankararaman, Fei Sha, Alex Simma, Erik Sudderth, Romain Thibaux, Daniel Ting, and Zhihua Zhang. I also thank Jon Kuroda and Jeff Anderson-Lee for their invaluable technical support.

Last but not least I would like to thank my family and friends at Berkeley for their accompany and support along this journey. I especially thank my parents for their lifetime love and for letting me pursue my dream so long and so far away from home. My gratitude to them is beyond the words. Finally, I would like to thank my wife Keling Chen for her love, kindness, and patience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Meiosis is a fundamental cellular division process for sexual reproduction, which produces haploid cells (or gametes) that have only one half of the full set of chromosomes. The reduction of the chromosome numbers is achieved by one round of chromosome duplication followed by two rounds of cell divisions. During the first meiotic division, homologous chromosomes may be broken and joined together, resulting in an exchange of corresponding regions. This process, namely *meiotic recombination*, is one of essential evolutionary factors responsible for promoting genetic diversity within species.

There are two main classes of meiotic recombination products, depending on the configuration of chromosome arms flanking the heteroduplex region, a short region of duplex DNA that contains one paternal strand and one maternal strand. If chromosome arms on one side of the heteroduplex region have been replaced by its homolog, this event is called a *crossover* (CO). On the other hand, if the original configuration of chromosome arms is maintained, it is called a *non-crossover* (NCO) event. Both CO and NCO events can cause *gene conversion*, the non-reciprocal transfer of genetic materials from a "donor" chromosome to an "acceptor" chromosome, as a consequence of mismatch repair of the heteroduplex region. No matter whether gene conversion occurs or not with a CO event, the descendant chromosome consists of some prefix of one parental chromosome, followed by a suffix of the other parental chromosome. But if gene conversion is associated with a NCO event, then the descendant chromosome has an alternating pattern between two parental genomes: a short segment (called a "conversion tract") from one parental chromosome is copied to the same position in the other parental chromosome. See Figure 2.1 and Figure 2.2 for more detail. In what follows in this thesis, we will use the term "gene conversion" exclusively for gene conversion event accompanied by NCO (i.e., gene conversion without concurrent CO), and will use the term "recombination" to refer to either crossover or gene conversion.

An understanding of recombination has implications for several important problems in population genetics and evolutionary biology. The most relevant one for medical practice is

association mapping, in which the primary goal is to identify genotyped markers that are in strong LD with untyped genetic variants responsible for susceptibility to complex diseases. Linkage disequilibrium (LD) refers to the non-random association of alleles at different loci at a population level. It is quantified by comparing the proportion of an observed haplotype with the proportion that is predicted based on the population frequencies of the alleles at each site. There are many different measures of LD, and most of them only capture the strength of pairwise association between two biallelic loci (Wall and Pritchard, 2003). Understanding the structure of LD is crucial for the design of large-scale disease association studies (Carlson *et al.*, 2004). In particular, the haplotype-block structure of the genome, within each block the markers are in strong LD with each other, can provide valuable guidance on how to choose optimal single-nucleotide polymorphisms (SNPs) density for association mapping.

The pattern of LD in a population is shaped by many genetic factors, including mutation, recombination, natural selection, population history, and etc. Recombination is a major mechanism that can reduce LD. Crossovers and gene conversions have different ranges of effect: LD is almost exclusively affected by crossovers if two markers are far away from each other, while over short ranges both crossovers and gene conversions can have impact. Ignoring gene conversions in population genetics models may cause serious problems in association studies (Wall, 2004a). If the local gene conversion rate is high but the marker density is low in this genomic region, the strength of LD over short ranges will tend to be overestimated. This may affect the definition of haplotype blocks and hence reduce the efficiency of a study. Therefore, it is desirable but challenging to tease apart the relative contributions of crossovers and gene conversions to the observed pattern of LD. In particular, for a given population SNP dataset, the joint estimation of the crossover rate, the gene conversion rate and the mean conversion tract length is widely viewed as a very difficult problem.

In Chapter 2, we devise a likelihood-based method using an interleaved hidden Markov model (HMM) that can jointly estimate the aforementioned three parameters fundamental to recombination. Our method significantly improves upon a recently proposed method based on a factorial HMM. We show that modeling overlapping gene conversions is crucial for improving the joint estimation of the gene conversion rate and the mean conversion tract length. We test the performance of our method on simulated data. We then apply our method to analyze real biological data from the telomere of the X chromosome of *Drosophila melanogaster*, and show that the ratio of the gene conversion rate to the crossover rate for the region may not be nearly as high as previously claimed when all the parameters of interest are estimated from the data.

In mammals, recombination events are not randomly distributed along the chromosome, but are concentrated in highly localized regions, known as recombination hotspots. The precise mechanisms of how hotspot activity is modulated are yet unknown. Although several *cis* and *trans* regulatory elements for recombination in mammals have been identified, there

is relatively little work that treats the role of nucleosome. Each nucleosome consists of a 147-bp segment of DNA wrapped twice around a histone octamer, and is a basic unit for the first level of compaction of eukaryotic DNA into the cell nucleus. In Chapter 3, we study the correlation of nucleosome positions and recombination rates in *Drosophila melanogaster*. Our results indicate that the majority of regions that exhibit highly elevated rates of recombination are nucleosomes-depleted. But among the blocks where recombination is suppressed, the blocks with few nucleosomes are also observed frequently. Collectively, this suggests that nucleosome-depleted regions might be more favored by recombination process, but by themselves are not sufficient to confer recombination.

Chapter 2

Crossovers and Gene Conversions

2.1 Introduction

A major evolutionary mechanism responsible for generating genetic variation in a population is meiotic recombination, which creates a chimeric genome from the two homologous genomes of an individual. Figure 2.1 summarizes the current thinking about mechanisms of mammalian recombination during meiosis. After a round of DNA duplication, trans-acting factors (such as PRDM9) bind to DNA consensus motif to activate chromatin, allowing DNA double-strand breaks (DSBs) created on one of the four chromatids by a topoisomerase-like enzyme (SPO11). Next, the 5′ ends of the DSB are resected to form two 3′-single-stranded DNA tails; one tail invades a non-sister chromatid to form a displacement (D)-loop, which is then extended by DNA synthesis using the intact strand as template. It is then followed by different recombination pathways[1], which yield either crossovers (COs) or non-crossover (NCO) products. In the double-strand-break-repair (DSBR) model, the other 3′ end of the DSB is captured (second-end capture) and paired with the extended D-loop, which leads to the formation of two Holliday junctions (HJs) (Stahl, 1994). The random resolution of double HJs results in either CO (indicated by a vertical cut at one HJ and a horizontal cut at the other HJ) and NCO (indicated by two horizontal cuts or two vertical cuts at both HJs) products. In the synthesis-dependent strand-annealing (SDSA) model (McMahill *et al.*, 2007), the synthesized strand is displaced from the template and anneals to the other 3′ end of the DSB, yielding mainly NCO products. See Figure 6.23 in Hartwell *et al.* (2006) for more detail.

The final forms of recombinants, after correction of the mismatch in the heteroduplex region, can be categorized into two main types (Figure 2.2): *crossovers* and *gene conversions*. Both recombinants involve taking two equal-length parental sequences to produce a descendant

[1]There is another mechanism, known as double-HJ dissolution, which is not shown here. See, for example, Chen *et al.* (2007) and references therein.

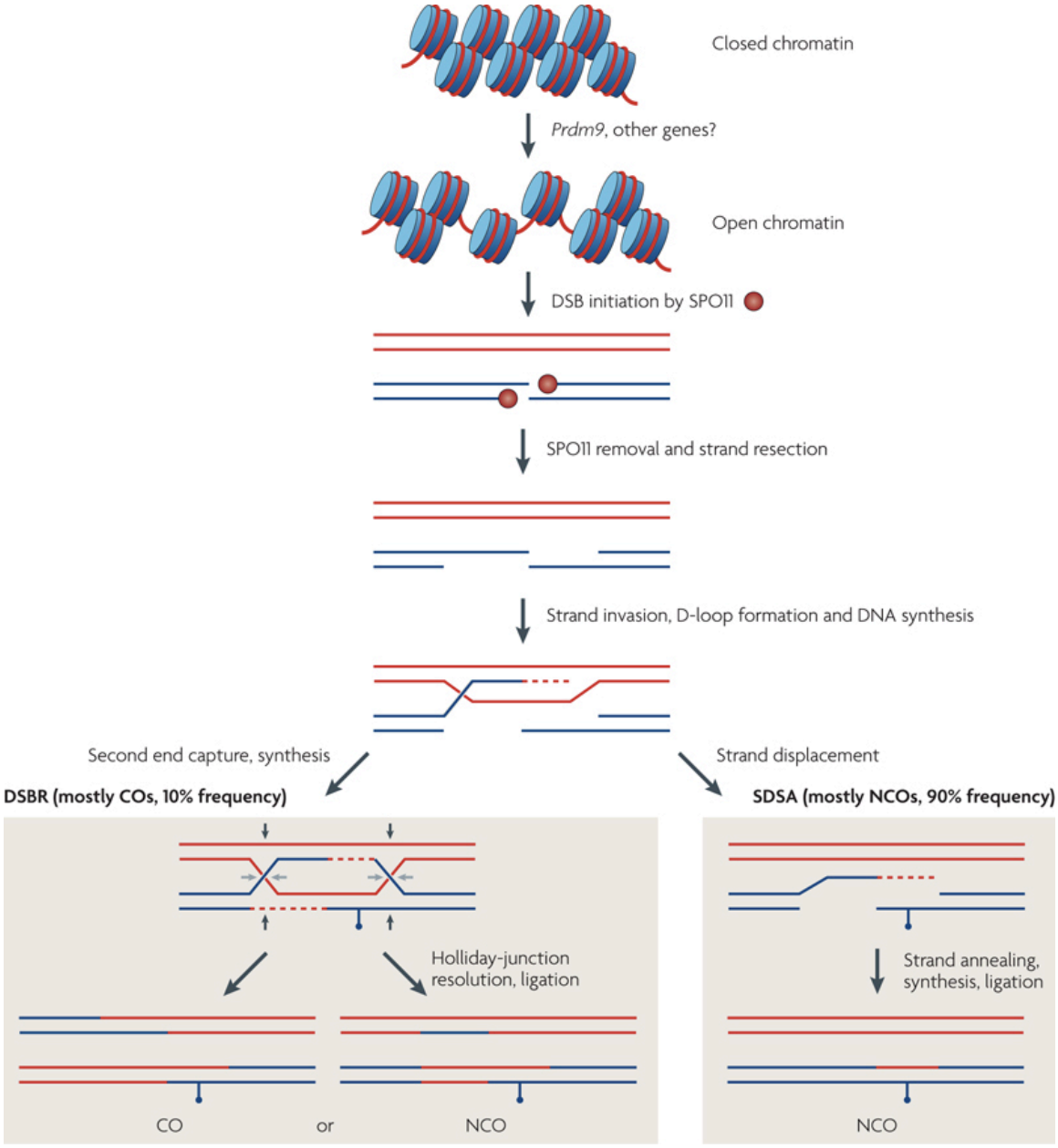


Figure 2.1: Mechanisms of mammalian recombination. Blue lines are strands of one parental chromatid and red ones are strands of the other parental chromatid (after DNA replication). Heteroduplex is a region of chromatid that contains one parental strand (red) and one maternal strand (blue). See main text for a description of processes. Image source: Paigen and Petkov (2010).

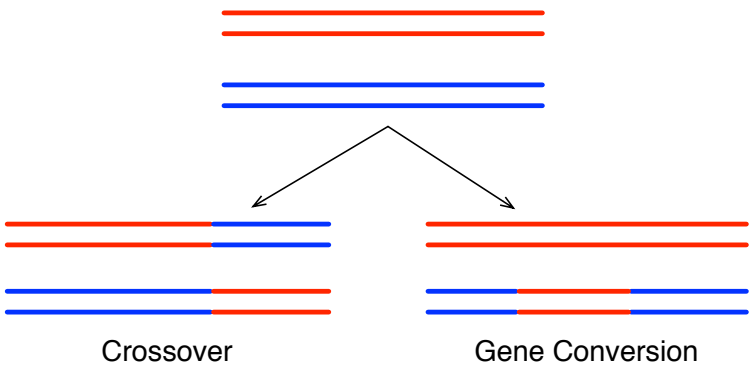


Figure 2.2: The difference between crossover and gene conversion recombinant. Blue lines are strands of one parental chromatid and red ones are strands of the other parental chromatid.

sequence of the same length. The crossover recombinant consists of some prefix of one of the parental sequences, followed by a suffix of the other parental sequence, whereas the gene conversion recombinant is formed by copying a short segment (called a "conversion tract") starting at a particular position in one of the parental sequences to the same position in the other parental sequence. Hence, the typical pattern created by gene conversion is: a prefix of sequence h followed by a short internal fragment of a sequence h', which is then followed by a suffix of the first sequence h. It is believed that the conversion tract typically ranges between 50 and 2000 bp (Jeffreys and May, 2004; Hilliker *et al.*, 1994).

Although crossovers and gene conversions have different effects on the evolutionary history of chromosomes and therefore leave behind different footprints in the genome, it is a challenging task to tease apart their relative contributions to the observed genetic variation. For example, the methods employed in recent studies (Crawford *et al.*, 2004; Myers *et al.*, 2005; International HapMap Consortium, 2005) of recombination rate variation in the human genome actually capture combined effects of crossovers and gene conversions (Figure 2.3).

Studying gene conversion is important for a number of reasons, a few of which we mention below. First, in several organisms—e.g, humans (Frisse *et al.*, 2001; Pritchard and Przeworski, 2001) and *Drosophila melanogaster* (Langley *et al.*, 2000)—gene conversion has been shown to be necessary to explain the observed pattern of linkage disequilibrium (LD), i.e., the statistical non-independence of alleles at different loci. Second, it has been argued that ignoring gene conversion may cause problems in association studies (Wall, 2004a) and linkage analysis (Mancera *et al.*, 2008). Third, methods for detecting signatures of natural selection usually require estimates of fine-scale recombination rates (see, for example, Voight *et al.* (2006)), and their success may hinge on having reliable estimates of crossover and gene conversion rates, as well as the distribution of the conversion tract length. Lastly, gene

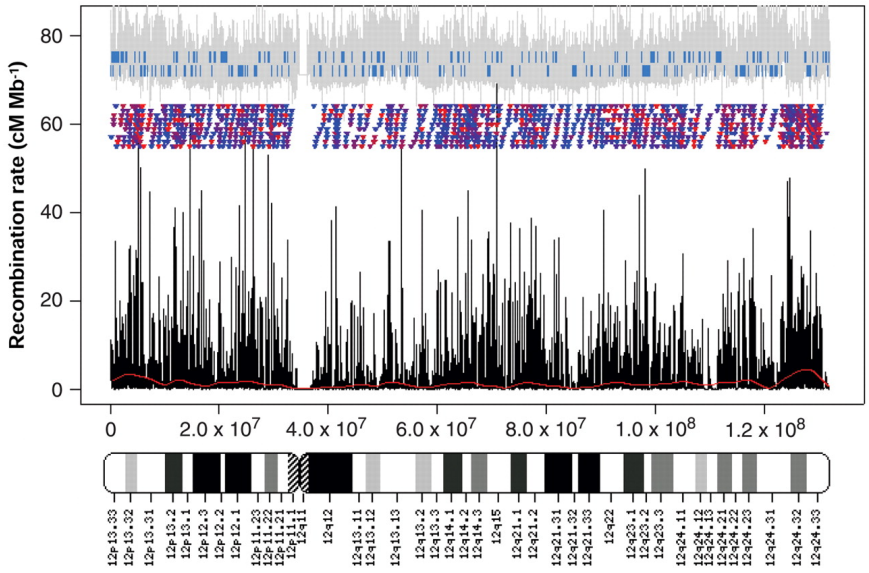


Figure 2.3: Recombination rate variation along chromosome 12 of the human genome. The black curve corresponds to the variation of combined effects of crossovers and gene conversions by using the composite likelihood method of Hudson (2001), adapted to finite-sites models. Image source: (Myers *et al.*, 2005).

conversion also plays an important role in molecular evolution. Biased gene conversion is believed to be a significant source of biases in substitution, and variation in biased gene conversion effects appears to be partially responsible for variation in substitution patterns across the mammalian phylogeny (Hwang and Green, 2004).

Gene conversion rate variation in the human genome is currently not well understood, though a recent sperm-typing study (Jeffreys and May, 2004) of the major histocompatibility complex region suggests that the rate of gene conversion can be about 5 to 15 times higher than that of crossover. Gene conversion has been hard to study in populations because of the lack of fine-scale data. However, the genomic resequencing data to be produced over the next several years will allow us to quantify the fundamental parameters of gene conversion. Therefore, algorithmic and statistical tools to study gene conversion are becoming increasingly more important.

Song *et al.* (2007) recently developed algorithms to distinguish the role of gene conversion from crossover in the derivation of SNP sequences in a population. Their method can produce an explicit evolutionary history of the input sequences using mutations and recombinations (crossovers and gene conversions), but it cannot produce estimates of recombination parameters. The parameters fundamental to recombination are the crossover rate, the gene

conversion rate, and the mean conversion tract length—the conversion tract length is often assumed to follow a geometric distribution (Wiuf and Hein, 2000; Wiuf, 2000), in which case the mean completely specifies the distribution. Joint estimation of all three parameters is widely viewed as a very difficult problem. There currently exist several coalescent-based methods (reviewed in Section 2.2) that can jointly estimate crossover and gene conversion rates, but all existing methods, with the only exception being the recent work of Gay *et al.* (2007), cannot estimate the mean conversion tract length at the same time.

To obtain accurate parameter estimates, it is crucial to make full use of data, and that is exactly what Gay *et al.* (2007) aimed to achieve in their work. Specifically, they constructed a likelihood-based method by incorporating gene conversion into a popular framework called the "Product of Approximate Conditionals" (PAC), first proposed by Li and Stephens (2003) to estimate crossover rates only. The work of Gay *et al.* marks important progress towards developing practical tools for studying gene conversion.

Our goal is to improve on the work of Gay *et al.* (2007) by introducing modifications to the model which we show are crucial to make the joint estimation of all three parameters feasible. Briefly, Gay *et al.* disallowed overlapping gene conversions in their model, for computational simplicity. We show that this simplification frequently leads to gross errors in the estimation of the gene conversion rate and the mean conversion tract length, when all three parameters are being estimated. In their paper, Gay *et al.* did not try to estimate the mean conversion tract length, but always fixed it to some reasonable value (actually, the true value in the case of simulation study). Therefore, they did not encounter this problem when testing their method. In this chapter, we devise algorithms to incorporate overlapping gene conversions into the PAC model and show that this modification dramatically improves the estimation of the gene conversion rate and the mean conversion tract length.

To test the performance of our method, we carry out a simulation study. We then apply our method to analyze real biological data from the telomere of the X chromosome of *Drosophila melanogaster*, and show that the ratio of the gene conversion rate to the crossover rate for the region may not be nearly as high as it was claimed to be by Gay *et al.* (2007).

2.2 Previous Methods

Throughout this chapter, the population-scaled crossover and gene conversion rates are denoted by $\rho = 4N_e c$ and $\gamma = 4N_e g$, respectively, where N_e is the effective population size, c is the per-generation probability of crossover per unit distance (kb in this chapter), and g is the per-generation probability of initiating a gene conversion per unit distance. The conversion tract length is assumed to follow a geometric distribution, and λ denotes the mean of that distribution.

2.2.1 An Overview of Previous Work

The coalescent is a retrospective process that can be used to describe the distribution of the underlying genealogy for a sample of chromosomes from unrelated individuals in an idealized population (Kingman, 1982) (Figure 2.4(a)). It has been proven extremely useful in a variety of applications in population genetics study (Hein *et al.*, 2004; Wakeley, 2008). As different sites on the same chromosome may have different genealogies in the presence of recombination, the genealogy of the recombined chromosomes is indeed a so-called ancestral recombination graph (ARG)[2] (Griffiths and Marjoram, 1996) instead of a coalescent tree, which includes a series of coalescent and recombination events until a most recent common ancestor (MRCA) is found (Figure 2.4(b)). There is a vast literature on estimating crossover rates ρ only based on the coalescent model with crossover recombination. Most of these statistical methods fall into one of three categories: moment-based estimators, full-likelihood approaches and approximate-likelihood approaches. See Stumpf and Mcvean (2003), Stephens (2008) and references therein.

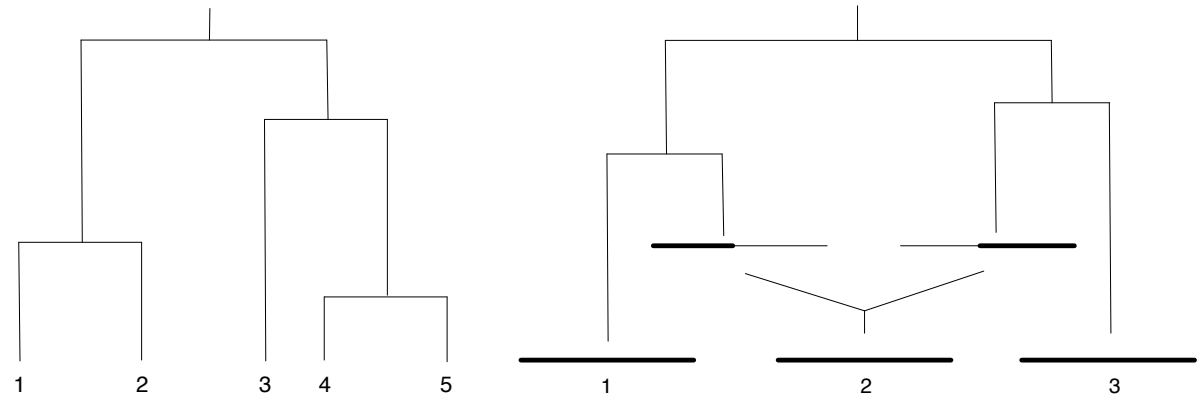


Figure 2.4: (a) A coalescent tree that relates 5 unrelated individuals. (b) An ancestral recombination graph (ARG) for a sample of three sequences that has one recombination event followed by two coalescent events. Each horizontal line represents a chromosome/sequence instead of a strand as in Figure 2.1 and Figure 2.2. Note that left and right parts of the second sequence have different marginal genealogies. Thick horizontal lines: genetic materials ancestral to the present-day sample; thin horizontal lines: non-ancestral materials. See Figure 2.7 for an illustration of the coalescent with gene conversions.

[2]Despite the name, Griffiths and Marjoram only considered incorporating crossover into the coalescent model.

There is a smaller literature on estimating gene conversion rates from population genetic data. Padhukasahasram *et al.* (2006) suggested using multiple summary statistics from SNP data to estimate crossover and gene conversion rates jointly. Despite the computational efficiency, this approach makes only partial use of the information in the data and can be strongly influenced by deviations from model assumptions.

The methods proposed by Frisse *et al.* (2001), Ptak *et al.* (2004) and Wall (2004b) generalize the composite-likelihood approach of Hudson (2001). Briefly, these methods break up the data set into smaller subsets (pairs or triplets of segregating sites), compute the likelihoods (as functions of ρ and γ, but with λ *fixed*) for the subsets, and then multiply those likelihoods together to form a composite likelihood. The point estimates of ρ and γ are then obtained by maximizing the composite likelihood over a suitably chosen finite grid. These methods don't take into account the dependency among the smaller subsets.

Assuming that each gene conversion tract contains a single SNP, Hellenthal (2006) incorporated gene conversion into the PAC framework, originally proposed by Li and Stephens (2003) to estimate crossover rates only. Gay *et al.* (2007) later generalized this PAC-likelihood based approach to allow for an arbitrary conversion tract length, and their method can be used to estimate ρ, γ and λ jointly from SNP data. The main advantage of these approaches is that they improve the statistical efficiency of the estimates by utilizing as much of the information in the data as possible. The work of Gay *et al.*, further detailed below, is most relevant to our own work.

2.2.2 The PAC Model with Gene Conversion

In principle, given a set $H = \{h_1, \ldots, h_n\}$ of haplotypes sampled from a population, the estimation of ρ, γ and λ can be obtained from maximizing the likelihood function $L(\rho, \gamma, \lambda \mid H) = \mathbb{P}(H \mid \rho, \gamma, \lambda)$. However, unless we could exam the true genealogical history of sampled sequences in the population, which is rarely available in a population genetics study, we cannot compute the likelihood function exactly in most cases of interest. To be precise,

$$L(\rho, \gamma, \lambda \mid H) = \mathbb{P}(H \mid \rho, \gamma, \lambda) = \int \mathbb{P}(H \mid G)\, \mathbb{P}(G \mid \rho, \gamma, \lambda)\, dG, \tag{2.1}$$

which involves an integral over all possible genealogies G. Hence, G could be viewed as a hidden variable or missing data, and $\mathbb{P}(G \mid \rho, \gamma, \lambda)$ is modeled by the coalescent with crossovers and gene conversions (Figure 2.4(b) and Figure 2.7).

Computing the integral in (2.1) is notoriously hard as the number of genealogies consistent with the sampled haplotypes H increases extremely fast as the length of sampled haplotypes grows (Song *et al.*, 2006). Therefore, several approximate-likelihood methods have been proposed. The PAC model (Li and Stephens, 2003), which is short for "Product of Approximate

Conditionals", begins with decomposing the likelihood function into a product of conditional probabilities:

$$\begin{aligned}\mathbb{P}(h_1,\ldots,h_n \mid \rho,\gamma,\lambda) &= \mathbb{P}(h_1 \mid \rho,\gamma,\lambda) \times \mathbb{P}(h_2 \mid h_1,\rho,\gamma,\lambda) \\ &\quad \times \cdots \times \mathbb{P}(h_n \mid h_1,\ldots,h_{n-1},\rho,\gamma,\lambda). \end{aligned} \tag{2.2}$$

Unfortunately, the exact conditional probabilities on the right hand side are unknown for the coalescent models with recombination. Li and Stephens (2003) proposed using efficiently computable approximations $\hat{\pi}$ to substitute for the exact probability distribution $\mathbb{P}$, thus obtaining the following approximation for the joint probability:

$$\begin{aligned}\mathbb{P}(h_1,\ldots,h_n \mid \rho,\gamma,\lambda) &\approx \hat{\pi}(h_1 \mid \rho,\gamma,\lambda) \times \hat{\pi}(h_2 \mid h_1,\rho,\gamma,\lambda) \\ &\quad \times \cdots \times \hat{\pi}(h_n \mid h_1,\ldots,h_{n-1},\rho,\gamma,\lambda). \end{aligned} \tag{2.3}$$

We denote the right hand side of (2.3) by $L_{\mathrm{PAC}}(\rho,\gamma,\lambda \mid H)$. The goal is to estimate ρ,γ and λ under the framework of maximum likelihood estimation (MLE), using L_{PAC} as a surrogate function for the original intractable likelihood function (2.2).

By exchangeability, the value of the right hand side of (2.2) is invariant under a permutation of the haplotype indices $1,\ldots,n$. However, because the $\hat{\pi}$ in (2.3) are not exact, the PAC likelihood L_{PAC} *does* depend on the order of haplotypes being considered. To account for this lack of exchangeability, Li and Stephens (2003) suggested averaging the PAC likelihood over several (say, between 10 and 20) random permutations of the input haplotypes.

The approximate conditional $\hat{\pi}(h_{k+1} \mid h_1,\ldots,h_k,\rho,\gamma,\lambda)$ is constructed by assuming that haplotype h_{k+1} is an imperfect mosaic of the first k haplotypes. That is, h_{k+1} is obtained by copying segments from $h_1,\ldots,h_k$; a crossover or a gene conversion can change the haplotype from which copying is performed. Furthermore, copying can be imperfect, corresponding to mutation. See Figure 2.5 for an illustration (adapted from Figure 2 of Li and Stephens (2003)). The copying process proceeds along the sequence from one end to the other, and it is assumed to be Markovian. This process can easily be modeled as a hidden Markov model (HMM) (Rabiner, 1989).

To compute $\hat{\pi}(h_{k+1} \mid h_1,\ldots,h_k,\rho,\gamma,\lambda)$, Gay *et al.* (2007) set up two hidden Markov chains along the sequence. This is illustrated in Figure 2.6(a), in which the "X chain" is for crossovers and the "G chain" is for gene conversions. The two chains evolve along the sequence independently of each other and, therefore, the model is a factorial HMM (Ghahramani and Jordan, 1997), satisfying the following identity:

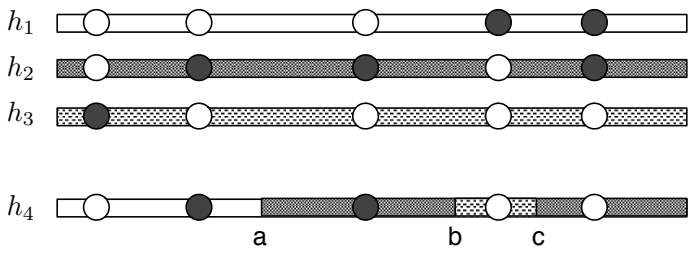


Figure 2.5: Illustration of the imperfect copying process with crossovers and gene conversions. Haplotype h_4 is created as a mosaic of fragments copied from haplotypes h_1, h_2, h_3. The shading shows from which haplotype each fragment is copied. The copying process is assumed to be Markovian along the sequence. Moving from left to right, there is a crossover event between h_1 and h_2 with a breakpoint at position "a". Then, there is a gene conversion event between h_2 and h_3, with a conversion tract between positions "b" and "c". Filled and unfilled circles represent different alleles. The second and the last circles in h_4 result from imperfect copying.

$$\mathbb{P}(X_{j+1}, G_{j+1} \mid X_j, G_j) = \mathbb{P}(X_{j+1} \mid X_j)\, \mathbb{P}(G_{j+1} \mid G_j), \tag{2.4}$$

where the index j denotes the position along the sequence, and $X_j \in \{1, \ldots, k\}$ and $G_j \in \{\emptyset, 1, \ldots, k\}$ are hidden states. The states X_j and G_j jointly determine the index c_j of the haplotype from which $h_{k+1,j}$ (allele at the jth site of h_{k+1}) is copied: If $G_j = \emptyset$ (the null state which indicates that the jth site is not in a gene conversion tract), then $c_j = X_j$; otherwise, $c_j = G_j$. To capture the imperfect nature of the copying process resulting from mutation, the emission probability of the HMM is set up as follows:

$$\mathbb{P}(h_{k+1,j} \mid X_j, G_j) = \begin{cases} \dfrac{\theta}{2(kL+\theta)}, & \text{if } h_{k+1,j} \neq h_{c_j,j}, \\ \dfrac{2kL+\theta}{2(kL+\theta)}, & \text{if } h_{k+1,j} = h_{c_j,j}, \end{cases} \tag{2.5}$$

where L is the number of polymorphic sites in the input data (i.e., the length of each haplotype) and θ/L is the rate of mutation per site. If θ is not specified, it is estimated by using Watterson's unbiased estimator (Watterson, 1975):

$$\hat{\theta} = L \left(\sum_{m=1}^{n-1} \frac{1}{m} \right)^{-1}. \tag{2.6}$$

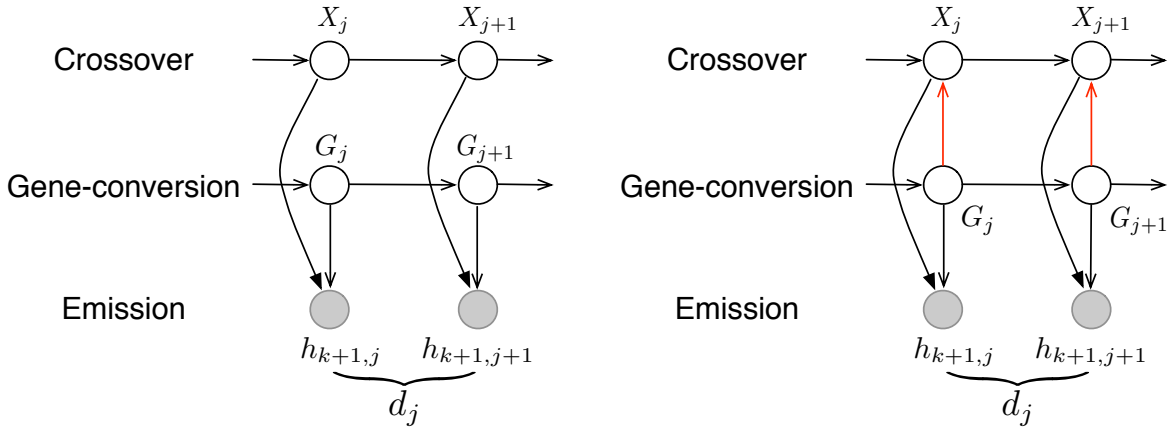


Figure 2.6: Two different versions of HMM for computing the conditional probability $\hat{\pi}(h_{k+1} \mid h_1, \ldots, h_k, \rho, \gamma, \lambda)$. Unshaded circles represent hidden variables, whereas shaded ones correspond to observed variables. The symbols d_j denotes the physical distance between sites j and $j+1$. In addition to a coupling of the two hidden chains, we allow pairwise overlaps of gene conversions. (a) A factorial HMM in which the two hidden chains are independent of each other. Gay *et al.* (2007) used this model. (b) An interleaved HMM with coupled hidden chains.

As in the original PAC model of Li and Stephens (2003), crossover is modeled as a Poisson process with rate ρ across the sequence. The transition probability of the X chain has only two distinct cases, depending on whether the hidden states of adjacent sites are the same or not:

$$\mathbb{P}(X_{j+1} \mid X_j) = \begin{cases} e^{-\frac{\rho d_j}{k}} + \frac{1}{k}\left(1 - e^{-\frac{\rho d_j}{k}}\right), & \text{if } X_j = X_{j+1}, \\ \frac{1}{k}\left(1 - e^{-\frac{\rho d_j}{k}}\right), & \text{if } X_j \neq X_{j+1}, \end{cases} \tag{2.7}$$

where d_j is the physical distance between sites $j-1$ and j.

The transition probability of the G chain is more complicated. By assuming that the conversion tract length follows a geometric distribution, both initiation and termination of a conversion tract are modeled as Poisson processes along the sequence, with rates γ and $1/\lambda$, respectively. Gay *et al.* used λ (not $1/\lambda$) to denote the termination rate and assumed that the termination process goes on all the time, even when the copying process is not in a

gene conversion state. Further, they make an additional assumption that conversion tracts from different gene conversion events cannot overlap. For example, consider the following probability of moving from state $g \in \{1, \ldots, k\}$ to state $g' \in \{1, \ldots, k\}$, where $g \neq g'$:

$$\mathbb{P}(G_{j+1} = g' \mid G_j = g) = \int_0^{d_j} \frac{e^{-x/\lambda}}{\lambda} \frac{(1 - e^{-\gamma x/k})}{k} \, \mathrm{d}x. \tag{2.8}$$

This formulation requires terminating the gene conversion tract from g before initiating a new one from g'. The integrand corresponds to the probability of there being *at least* one gene conversion event *after* the last termination event at distance x to the left of site $j + 1$. In general, Gay *et al.* (2007)'s formulation implicitly allows for an infinite number of gene conversion initiation events to occur before the last termination event.

Lastly, the initial probability of the G chain depends on how the rate of starting a gene conversion tract compares to the rate of ending one, i.e.,

$$\mathbb{P}(G_1 = g) = \begin{cases} \dfrac{1/\lambda}{1/\lambda + \gamma/k}, & \text{if } g = \emptyset, \\ \dfrac{\gamma/k}{k(1/\lambda + \gamma/k)}, & \text{if } g \neq \emptyset. \end{cases} \tag{2.9}$$

In the above HMM formulation, it is straightforward to compute the conditional probability $\hat{\pi}(h_{k+1} \mid h_1, \ldots, h_k, \rho, \gamma, \lambda)$ by using the standard forward-backward algorithm.

2.3 Our Model

As described above, the work of Gay *et al.* (2007) assumes that crossovers and gene conversions are independent, and that gene conversion tracts cannot overlap. In this section, we construct a new model, called OVERPAINT, which couples the crossover and gene conversion processes. We then describe how overlapping gene conversions can be incorporated into the model.

2.3.1 Interleaved HMM

By assuming independence of the two hidden chains, the factorial HMM formulation of Gay *et al.* (2007) cannot model the typical alternating pattern of gene conversion; i.e., a prefix of haplotype h followed by an internal fragment of a haplotype h', which is then followed by a suffix of the first haplotype h. To remedy this, we couple the two hidden chains by using an interleaved HMM, illustrated in Figure 2.6(b). Direct edges from the G chain to the X chain constrain the X chain to stay in its previous state whenever the G chain is "active."

More precisely,

$$\mathbb{P}(X_{j+1} \mid X_j, G_{j+1}) = \begin{cases} \mathbf{1}_{X_{j+1}=X_j}, & \text{if } G_{j+1} \neq \emptyset, \\ \mathbb{P}(X_{j+1} \mid X_j), & \text{if } G_{j+1} = \emptyset, \end{cases} \tag{2.10}$$

where $\mathbb{P}(X_{j+1} \mid X_j)$ in the second line is the same as in (2.7). If site $j+1$ is in a conversion tract (i.e., $G_{j+1} \neq \emptyset$), the G chain is "active" and the copying process keeps track of the previous state of the X chain (i.e., $X_{j+1} = X_j$). If $G_{j+1} = \emptyset$, the X chain evolves according to the usual transition probability $\mathbb{P}(X_{j+1} \mid X_j)$.

We point out that coupling the two hidden chains *alone* does not increase the complexity of the forward-backward computation. Even in the factorial HMM, the two hidden chains become dependent upon conditioning on the observed variables. It is modeling of overlapping gene conversions that increases the computational complexity, as the state space of hidden chains are explicitly augmented.

2.3.2 Modeling Overlapping Gene Conversions

The key new feature of our model is that it allows for overlapping gene conversion events in the copying process. This means that the copying process does not need to terminate a gene conversion event before initiating another gene conversion event.

Figure 2.7 shows two examples of genealogies that can generate overlapping gene conversion tracts in the coalescent model with gene conversion (Wiuf and Hein, 2000; Wiuf, 2000). In Figure 2.7(a), two gene conversion events have conversion tracts that overlap partially, while in Figure 2.7(b), one conversion tract is entirely nested inside the other conversion tract.

Motivated by the common belief that the conversion tract length is typically short, between 50 and 2000 bp (Jeffreys and May, 2004; Hilliker *et al.*, 1994), we restrict each overlap to involve only a pair of gene conversion events, although a generalization to more than two gene conversion events can easily be achieved at the expense of more computation time.

In terms of the underlying HMM, we augment the state space of the G chain as follows. When computing $\hat{\pi}(h_{k+1} \mid h_1, \ldots, h_k, \rho, \gamma, \lambda)$, we include ordered pairs $\{(g, g') \mid g, g' = 1, \ldots, k\}$ in the state space of the G chain, in addition to the singlet states $\{g \mid g = \emptyset, 1, \ldots, k\}$ considered in Gay *et al.*'s model. If $G_j = (g, g')$, then site j of haplotype h_{k+1} is within a region of overlapping gene conversion events involving two haplotypes h_g and $h_{g'}$. The second entry g' in a doublet state (g, g') is said to be "active" and it indicates that the conversion tract from $h_{g'}$ overwrites the conversion tract from h_g at marker j of h_{k+1}. In Figure 2.7(a), g is active in the region of overlapping gene conversions, while in Figure 2.7(b) g' is active in the region of overlap. As in Gay *et al.*'s model, the hidden states $X_j \in \{1, \ldots, k\}$ and G_j jointly

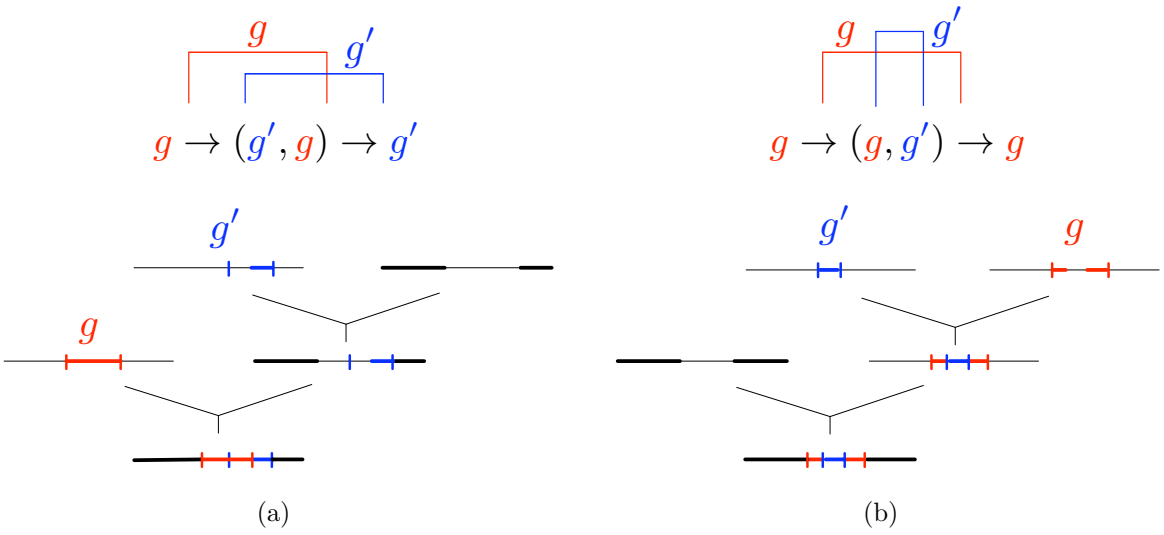


Figure 2.7: Genealogical interpretations of overlapping gene conversions. Each genealogy contains two gene conversion events. Thin horizontal lines represent genetic material non-ancestral to the present-day sample, whereas thick horizontal lines correspond to ancestral material. Short vertical lines mark the boundaries of gene conversion tracts. (a) Two gene conversion tracts partially overlap. The left part of the blue conversion tract is non-ancestral because it is overwritten by the red conversion tract from a more recent gene conversion event. The "active" haplotype in the region of overlapping gene conversion is g. (b) One conversion tract is completely nested inside the other conversion tract. The blue conversion tract overwrites the middle part of the red conversion tract. The "active" haplotype in the region of overlap is g'.

determine the index c_j of the haplotype from which $h_{k+1,j}$ is copied. In our model,

$$c_j = \begin{cases} X_j, & \text{if } G_j = \emptyset, \\ g, & \text{if } G_j = g \neq \emptyset, \\ g', & \text{if } G_j = (g, g'). \end{cases} \tag{2.11}$$

We use the same emission probability as that shown in (2.5).

2.3.3 Transition Probabilities for the Augmented G Chain

We now describe the transition probabilities $\mathbb{P}(G_{j+1} = s' \mid G_j = s)$ for the augmented G chain in the computation of $\hat{\pi}(h_{k+1} \mid h_1, \ldots, h_k, \rho, \gamma, \lambda)$. Instead of using the form of integral

Algorithm 1: EnumeratePaths

Input: a, b; `InitialState` $\in \{\emptyset, (1), (1,1), (1,2)\}$.
Output: All the valid paths starting with `InitialState` with **at most** a initiations and b terminations.

```
Initialize a hash table AllPaths with keys represented by pair of integers
(i, j), i ∈ {0, 1, ··· , a} and j ∈ {0, 1, ··· , b}.
for i = 0, 1, ··· , a do
    for j = 0, 1, ··· , b do
        if i == 0 and j == 0 then
            n ← number of distinct symbols in InitialState.
            AllPaths[(i, j)] ← [(n, [InitialState])].
        else
            ExtendPaths(i, j, AllPaths).
        end
    end
end
P ← ∅
for i = 0, 1, ··· , a do
    for j = 0, 1, ··· , b do
        foreach AugPath ∈ AllPaths[(i, j)] do
            P ← P ∪ AugPath[1].
        end
    end
end
return P.
```

as in (2.8), which implicitly allows for infinitely many gene conversion events between two adjacent sites, we explicitly enumerate all possible "valid" paths of events defined to satisfy the following two properties: 1) Each "valid" path starts in state s and ends in state s', and 2) contains at most a initiations and b terminations of gene conversions. In our implementation, we use $a = b = 1$ for simplicity, but it is also possible to consider larger values of a and b without increasing the asymptotic complexity of the forward-backward algorithm in our HMM (Algorithm 1 and Algorithm 2).

For $a = b = 1$, the path $(g, g') \to g' \to (g', g'')$ is valid since it contains exactly one initiation event and one termination event. In contrast, the path $g \to \emptyset \to g' \to (g, g')$ is not valid since it contains two initiation events.

For a given pair of states s, s' of the G chain (and for given values of a and b), all valid paths

starting in s and ending in s' can be enumerated using dynamic programming (Algorithm 1 and Algorithm 2). We use $\mathcal{P}_{s,s'}$ to denote the set of all such valid paths. To compute the probability $\mathbb{P}(\Gamma)$ for a given path $\Gamma \in \mathcal{P}_{s,s'}$, we make the following assumptions:

- If the current state in Γ is the $\emptyset$ state, then the initiation process has rate γ/k and an initiation event uniformly chooses the next state from $\{1, \ldots, k\}$; no termination event can occur, which contrasts with Gay *et al.* (2007) that allows the termination process to run all the time.

- If the current state in Γ is a singlet g, then the initiation process has rate γ/k and an initiation event uniformly chooses $g' \in \{1, \ldots, k\}$ to create either (g, g') or (g', g) with equal probability; the termination process has rate $1/\lambda$.

- If the current state in Γ is a doublet (g, g'), then no initiation can occur, since we assume only pairwise overlaps of gene conversions. The termination process has rate $2/\lambda$, and when a termination event occurs, one makes a transition from (g, g') to either g or g' with equal probability.

With the above assumptions, $\mathbb{P}(\Gamma)$ can be computed by integrating over all possible positions along the sequence where the events in Γ can happen. In contrast, recall that Gay *et al.* (2007) only integrate over the position of the last termination event. The main computation involves a symbolic convolution of exponential functions, which can be easily evaluated. The transition probability $\mathbb{P}(G_{j+1} = s' \mid G_j = s)$ can then be obtained by adding up the probability of all valid paths in $\mathcal{P}_{s,s'}$ and then normalizing to make sure that the outgoing probabilities sum to 1, that is,

$$\mathbb{P}(G_{j+1} = s' \mid G_j = s) = \frac{\sum_{\Gamma \in \mathcal{P}_{s,s'}} \mathbb{P}(\Gamma)}{\sum_{s'} \sum_{\Gamma \in \mathcal{P}_{s,s'}} \mathbb{P}(\Gamma)}. \tag{2.12}$$

As a concrete example, consider the transition probability $\mathbb{P}(G_{j+1} = g' \mid G_j = g)$, where $g, g' \in \{1, \ldots, k\}$ and $g \neq g'$. For $a = b = 1$, $\mathcal{P}_{g,g'}$ contains three valid paths, namely $\Gamma_1 = g \to \emptyset \to g'$, $\Gamma_2 = g \to (g, g') \to g'$, and $\Gamma_3 = g \to (g', g) \to g'$. The probability of Γ_1 is given by

$$\begin{aligned}
\mathbb{P}(\Gamma_1) &= \int_0^{d_j} \int_0^{d_j - x} \left[\frac{1}{\lambda} e^{-x/\lambda} \cdot e^{-(d_j - x - y)/\lambda} \right] \\
&\quad \times \left[e^{-\gamma x/k} \cdot \frac{\gamma}{k} e^{-\gamma y/k} \frac{1}{k} \cdot e^{-\gamma(d_j - x - y)/k} \right] \mathrm{d}y \, \mathrm{d}x \\
&= \frac{\lambda\gamma \, e^{-\gamma d_j/k - d_j/\lambda}}{k^2} \int_0^{d_j} \frac{1}{\lambda} \int_0^{d_j - x} \frac{1}{\lambda} e^{y/\lambda} \mathrm{d}y \, \mathrm{d}x \\
&= \frac{\lambda\gamma \, e^{-\gamma d_j/k - d_j/\lambda}}{k^2} \left(-1 + e^{d_j/\lambda} - \frac{d_j}{\lambda} \right). \qquad (2.13)
\end{aligned}$$

The integrand corresponds to the probability of there being exactly one termination event and exactly one initiation event, with the termination (respectively, initiation) event occurring at distance x (respectively, $x + y$) to the right of site j. Integrating over all possible values of x and y yields the probability of Γ_1. In a similar vein, one can show that the probabilities $\mathbb{P}(\Gamma_2)$ and $\mathbb{P}(\Gamma_3)$ are given by

$$\mathbb{P}(\Gamma_2) = \mathbb{P}(\Gamma_3) = \frac{1}{2} \frac{\lambda\gamma \, e^{-\gamma d_j/k - d_j/\lambda}}{k^2} \left(-1 + e^{-d_j/\lambda} + \frac{d_j}{\lambda} \right). \qquad (2.14)$$

Hence the transition probability $\mathbb{P}(G_{j+1} = g' \mid G_j = g)$ is proportional to $\mathbb{P}(\Gamma_1) + \mathbb{P}(\Gamma_2) + \mathbb{P}(\Gamma_3)$.

Table 2.1 lists the transition probabilities in the G chain of our implementation with $a = b = 1$. In the table, g, g', g'' denote distinct elements of $\{1, \ldots, k\}$.

2.3.4 Initial Probabilities of the G Chain

We wish to use the stationary distribution of the transition matrix of the G chain as the initial probability at site 1. However, in the computation of $\hat{\pi}(h_{k+1} \mid h_1, \ldots, h_k, \rho, \gamma, \lambda)$, the size of the transition matrix is $(k^2+k+1) \times (k^2+k+1)$, since there are 1 null state $\emptyset$, k singlet states (g), k degenerate doublet states (g, g), and $k^2 - k$ non-degenerate doublet states (g, g'), where $g \neq g'$. Finding an eigenvector of that transition matrix could be computationally expensive for moderate values of k. Therefore, we make the following approximation: we collapse the transition matrix to a 4×4 matrix, whose rows and columns are indexed by "null", "singlet", "degenerate doublet," and "non-degenerate doublet." Each entry in the collapsed matrix is obtained by summing over the corresponding entries in the original transition matrix. We first find the left eigenvector $v = (v_0, v_1, v_2, v_3)$ of the collapsed matrix with eigenvalue 1. Then, for distinct $g, g' \in \{1, \ldots, k\}$, the initial probabilities of the G chain are specified as

Algorithm 2: ExtendPaths

Input: i, j; `AllPaths`.

Output: None; Fill `AllPaths`$[(i,j)]$ with all the augmented paths with **exactly** i initiations and j terminations.

```
AllPaths[(i,j)] ← ∅.
if i >= 1 then
    foreach AugPath ∈ AllPaths[(i−1,j)] do
        Path ← AugPath[1]
        LastState ← last state in Path
        n ← AugPath[0]  //the number of distinct symbols in Path
        if LastState is an empty state then
            for k = 1,2,··· ,n do
                Append(AllPaths[(i,j)], (n, Path + (k)))
            end
            Append(AllPaths[(i,j)], (n+1, Path + (n+1)))
        else if LastState is a singlet then
            for k = 1,2,··· ,n do
                Append(AllPaths[(i,j)], (n, Path + (LastState, k)))
                Append(AllPaths[(i,j)], (n, Path + (k, LastState)))
            end
            Append(AllPaths[(i,j)], (n+1, Path + (LastState, n+1)))
            Append(AllPaths[(i,j)], (n+1, Path + (n+1, LastState)))
        end
    end
end
if j >= 1 then
    foreach AugPath ∈ AllPaths[(i,j−1)] do
        Path ← AugPath[1]
        LastState ← last state in Path
        n ← AugPath[0]  //the number of distinct symbols in Path
        if LastState is a singlet then
            Append(AllPaths[(i,j)], (n, Path + ∅))
        else if LastState is a doublet then
            Append(AllPaths[(i,j)], (n, Path + (LastState[0])))
            Append(AllPaths[(i,j)], (n, Path + (LastState[1])))
        end
    end
end
```

state G_j	state G_{j+1}	$\mathbb{P}(G_{j+1} = s' \mid G_j = s)$ up to normalization
$\emptyset$	$\emptyset$	$e^{-\gamma d_j/k} + \frac{e^{-\gamma d_j/k}\gamma\lambda}{k}(-1 + e^{-d_j/\lambda} + d_j/\lambda)$
$\emptyset$	g	$\frac{e^{-\gamma d_j/k - d_j/\lambda}\gamma\lambda}{k^2}(-1 + e^{d_j/\lambda})$
g	(g, g)	$\frac{e^{-\gamma d_j/k - 2d_j/\lambda}\gamma\lambda}{k^2}(-1 + e^{d_j/\lambda})$
g	$(g, g'), (g', g)$	$\frac{e^{-\gamma d_j/k - 2d_j/\lambda}\gamma\lambda}{2k^2}(-1 + e^{d_j/\lambda})$
g	g	$e^{-\gamma d_j/k - d_j/\lambda} + \frac{e^{-\gamma d_j/k - d_j/\lambda}\lambda\gamma}{k^2}\left[(k+1)(-1 + e^{-d_j/\lambda} + d_j/\lambda) + (-1 + e^{d_j/\lambda} - d_j/\lambda)\right]$
g	g'	$\frac{e^{-\gamma d_j/k - d_j/\lambda}\lambda\gamma}{k^2}\left[(-1 + e^{-d_j/\lambda} + d_j/\lambda) + (-1 + e^{d_j/\lambda} - d_j/\lambda)\right]$
g	$\emptyset$	$e^{-\gamma d_j/k}(1 - e^{-d_j/\lambda})$
(g, g)	(g, g)	$e^{-\gamma d_j/k - 2d_j/\lambda} + \frac{2e^{-\gamma d_j/k - 2d_j/\lambda}\gamma\lambda}{k^2}(-1 + e^{d_j/\lambda} - d_j/\lambda)$
(g, g)	$(g, g'), (g', g)$	$\frac{e^{-\gamma d_j/k - 2d_j/\lambda}\gamma\lambda}{k^2}(-1 + e^{d_j/\lambda} - d_j/\lambda)$
(g, g)	g	$2e^{-\gamma d_j/k - d_j/\lambda}(1 - e^{-d_j/\lambda})$
(g, g')	$(g, g), (g', g'), (g', g)$	$\frac{e^{-\gamma d_j/k - 2d_j/\lambda}\gamma\lambda}{k^2}(-1 + e^{d_j/\lambda} - d_j/\lambda)$
(g, g')	(g, g')	$e^{-\gamma d_j/k - 2d_j/\lambda} + \frac{e^{-\gamma d_j/k - 2d_j/\lambda}\gamma\lambda}{k^2}(-1 + e^{d_j/\lambda} - d_j/\lambda)$
(g, g')	$(g, g''), (g', g'')$ $(g'', g), (g'', g')$	$\frac{e^{-\gamma d_j/k - 2d_j/\lambda}\gamma\lambda}{2k^2}(-1 + e^{d_j/\lambda} - d_j/\lambda)$
(g, g')	g, g'	$e^{-\gamma d_j/k - d_j/\lambda}(1 - e^{-d_j/\lambda})$

Table 2.1: $\mathbb{P}(G_{j+1} = s' \mid G_j = s)$ for the gene conversion chain in the computation of $\hat{\pi}(h_{k+1} \mid h_1, \ldots, h_k, \rho, \gamma, \lambda)$, assuming at most one initiation and at most one termination of gene conversions between adjacent sites.

$$\begin{aligned}
\mathbb{P}(G_1 = \emptyset) &= v_0, \\
\mathbb{P}(G_1 = g) &= \frac{v_1}{k}, \\
\mathbb{P}(G_1 = (g, g)) &= \frac{v_2}{k}, \\
\mathbb{P}(G_1 = (g, g')) &= \frac{v_3}{k^2 - k}.
\end{aligned} \tag{2.15}$$

2.3.5 Prior for Mean Tract Length λ

To take into account the prior information that the tract length typically ranges between 0.05 and 2kb (Hilliker *et al.*, 1994; Jeffreys and May, 2004), we assign an uninformative prior (Figure 2.8) for λ with

$$\log_{10}(\lambda) \sim \mathcal{N}(-0.5, 0.4^2), \tag{2.16}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a standard normal distribution with mean μ and variance σ^2. This prior is carefully chosen so that $\mathbb{P}(\lambda \in [0.05, 2]) = 95\%$.

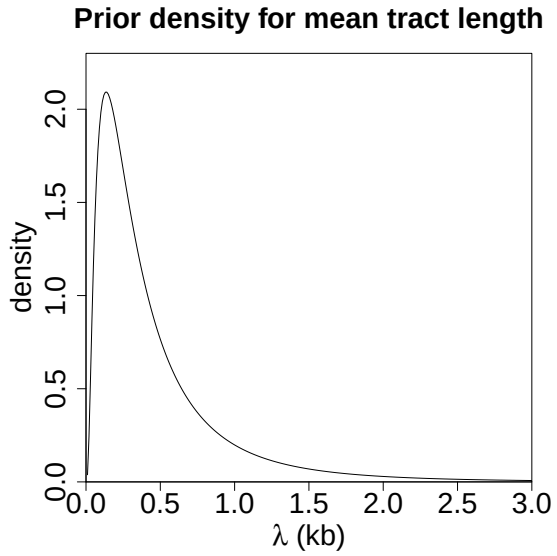


Figure 2.8: Prior density for mean tract length λ.

Then we use a standard derivative-free optimization procedure (the Nelder-Mead simplex-reflection method (Nocedal and Wright, 2000)) to find estimates of ρ, γ, λ based on the posterior

$$\tilde{L}_{\text{OVERPAINT}}(\rho, \gamma, \lambda \mid H) \propto f(\lambda) \times L_{\text{OVERPAINT}}(\rho, \gamma, \lambda \mid H), \tag{2.17}$$

where $L_{\text{OVERPAINT}}(\rho, \gamma, \lambda \mid H)$ denotes the likelihood function of our model OVERPAINT and $f(\lambda)$ is the density of λ that corresponds to (2.16). The prior can also be interpreted as a regularizer, penalizing very small or large values of λ.

2.3.6 Hypothesis Testing for the Boundary Cases

As observed by Gay *et al.* (2007) and ourselves, the gene conversion rates tend to be overestimated when there is actually no gene conversion present, that is, $\gamma = 0$. This is inevitable as the true value lies at the boundary of possible range. Here, we devise a hypothesis testing procedure based on a likelihood ratio test. In our case, the null hypothesis is $H_0 : \gamma = 0$ and the test statistic is the likelihood ratio statistic:

$$\Lambda(H) = -2 \log \left(\frac{\sup_\rho L_{\text{OVERPAINT}}(\rho, 0, 0 \mid H)}{\sup_{\rho,\gamma,\lambda} L_{\text{OVERPAINT}}(\rho, \gamma, \lambda \mid H)} \right), \tag{2.18}$$

where $L_{\text{OVERPAINT}}(\rho, 0, 0 \mid H)$ simply denotes the likelihood function that is computed with crossovers only. Large value of observed statistic $\Lambda(H)$ intuitively should lead to the rejection of the null hypothesis H_0, but the key question is: what is the threshold value for $\Lambda(H)$ to reject H_0?

Shown in Figure 2.9 and Figure 2.10 are the empirical distributions of test statistic $\Lambda(H)$ for sample size $n = 20$ and $n = 35$, respectively. Interestingly, the 95% quantile of empirical distribution is not a monotone function of the nuisance parameter ρ: as ρ goes beyond some value, it begins to increase. It could be possibly explained by the fact that high crossover rate tends to yield two crossover events that are nearby along the chromosome, resulting in a descendant sequence that has an alternating pattern between two parental genomes, which is the main characteristic of gene conversion recombinants. Thus, for haplotype data H generated by high crossover rate ρ *only*, the maximum probability of H under the alternative (the denominator in (2.18)), which allows gene conversions, is more likely to increase by a large amount compared to H generated by medium crossover rate ρ *only*; the overall effect is to shift the empirical distribution of $\Lambda(H)$ to the right and increase its 95% quantile. However,

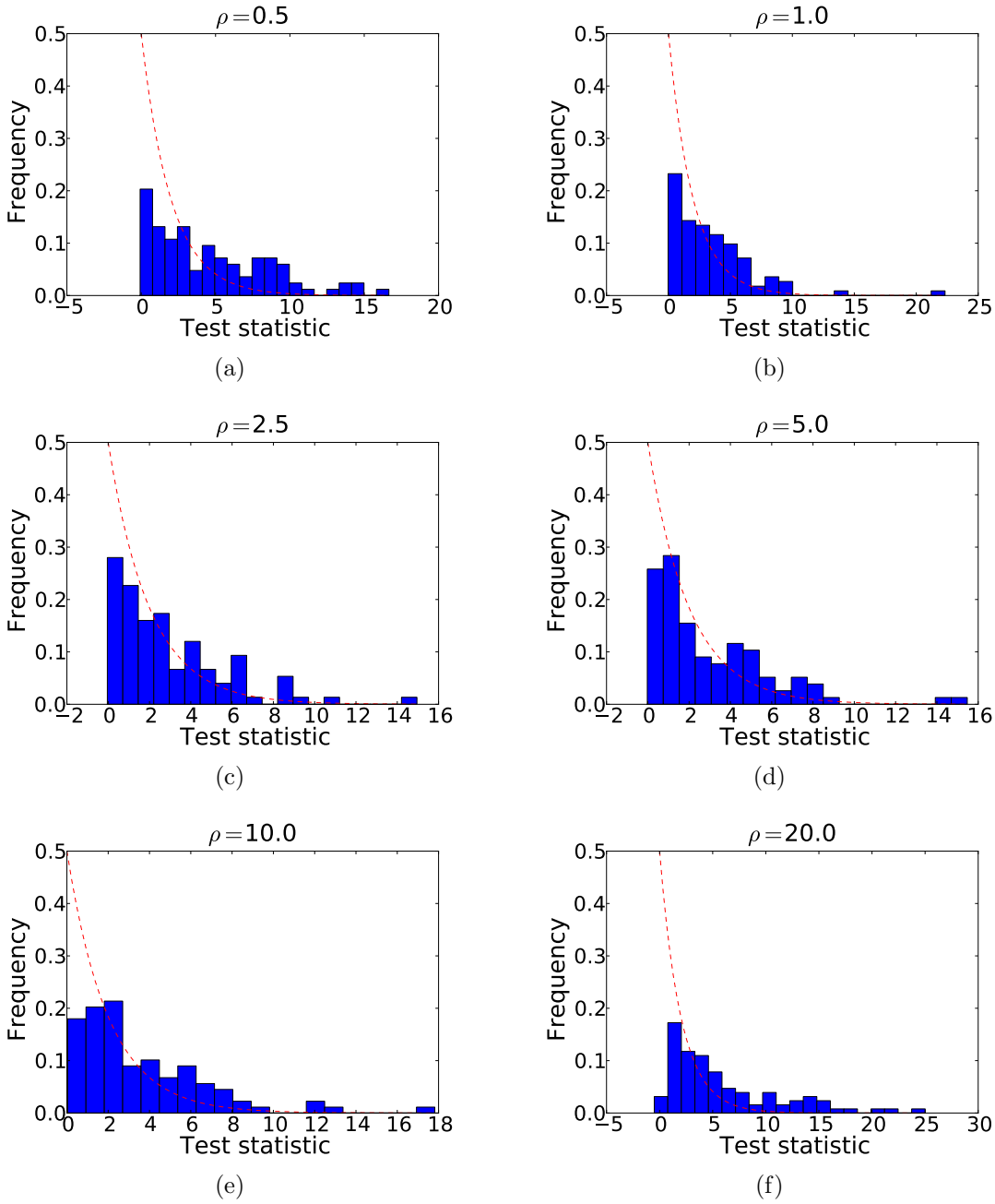


Figure 2.9: Empirical distributions of likelihood ratio statistic $\Lambda(H)$ (2.18) under the null hypothesis $H_0 : \gamma = 0$ for different values of nuisance parameter ρ. Based on 100 simulations with sample size $n = 20$. The red dashed lines correspond to the density of χ^2_2 distribution. The 95% empirical quantiles are: (a) 13.3; (b) 8.95; (c) 8.53; (d) 8.17; (e) 9.06; (f) 16.47.

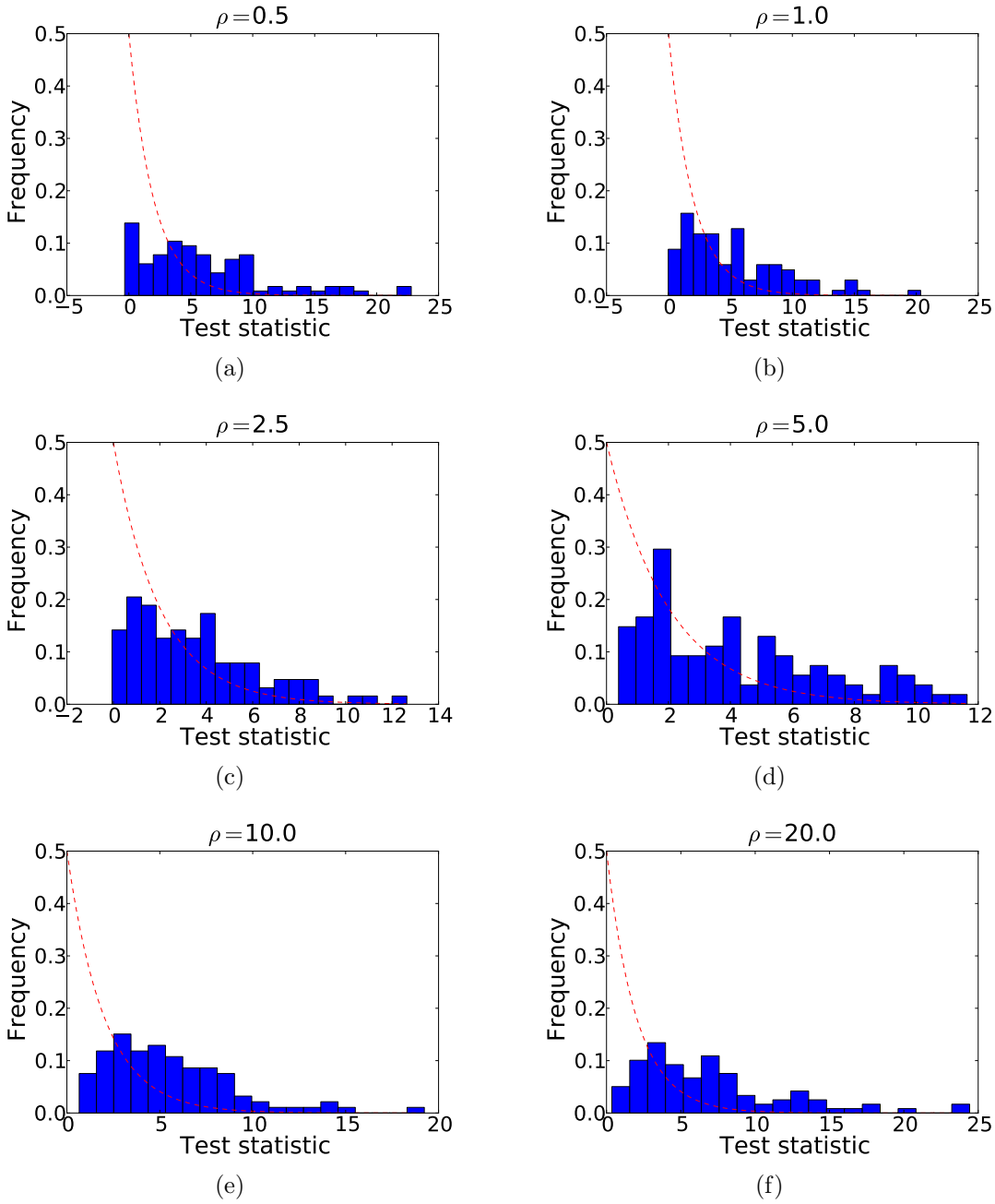


Figure 2.10: Empirical distributions of likelihood ratio statistic $\Lambda(H)$ (2.18) under the null hypothesis $H_0 : \gamma = 0$ for different values of nuisance parameter ρ. Based on 100 simulations with sample size $n = 35$. The red dashed lines correspond to the density of χ^2_2 distribution. The 95% empirical quantiles are: (a) 16.99; (b) 14.34; (c) 8.32; (d) 9.76; (e) 13.04; (f) 17.24.

Algorithm 3: ParametricBootstrap

Input: A haplotype data H.

Output: The bootstrap estimation of p-value.

1 Compute $\hat{\rho} = \text{argmax}_\rho L_{\text{OVERPAINT}}(\rho, 0, 0)$, which is the MLE of ρ under H_0.

2 Draw B bootstrap samples $H_1^*, \cdots, H_B^*$ of size n using MS (Hudson, 2002) with $\hat{\rho}$.

3 For each bootstrap sample H_b^*, compute $\Lambda(H_b^*)$ (2.18), $b = 1, \cdots, B$.

4 Return the estimated p-value as

$$\frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\Lambda(H_b^*)>\Lambda(H)} \tag{2.19}$$

these empirical distributions of $\Lambda(H)$ under H_0 cannot be obtained in practice since they depend on the known nuisance parameter ρ. If we don't know the distribution of the test statistic under H_0, we are unable to compute the significance of the test, such as p-value. Hence, we use a parametric bootstrap procedure (Efron and Tibshirani, 1994) as outlined in Algorithm 3 to obtain an approximate p-value.

2.3.7 Complexity of the Algorithm

Since the augmented HMM has $O(k^3)$ states when computing $\hat{\pi}(h_{k+1} \mid h_1, \ldots, h_k, \rho, \gamma, \lambda)$, a naive implementation of the forward-backward algorithm takes $O(k^6 L)$ time, where L is the number of polymorphic sites in the input data (i.e., the length of each haplotype). Hence the computational complexity of the (posterior) likelihood $\tilde{L}_{\text{OVERPAINT}}$ (for fixed parameters ρ, γ, λ) is $O(n^7 L)$, where n is the total number of input haplotypes. However, the computational complexity can be reduced to $O(n^4 L)$ by exploiting the sparsity and regularity of transition probabilities.

As a concrete example, consider the computation of intermediate term $\alpha_j(x, s) = \mathbb{P}(h_{k+1,1:j}, X_j = x, G_j = s)$ in the forward computation, where $h_{k+1,1:j}$ denotes the alleles of the first j sites in h_{k+1}. By recursion, for $s = \emptyset$,

$$\alpha_j(x, \emptyset) = \left(\sum_{x', s'} \alpha_{j-1}(x', s')\, \mathbb{P}(X_j = x \mid X_{j-1} = x')\, \mathbb{P}(G_j = \emptyset \mid G_{j-1} = s')\right) \gamma_j(x), \tag{2.20}$$

where $\gamma_j(x) = \mathbb{P}(h_{k+1,j} \mid X_j = x, G_j = \emptyset)$ is given in (2.5) and (2.11).

Denote $g_{00} = \mathbb{P}(G_j = \emptyset \mid G_{j-1} = \emptyset), g_{10} = \mathbb{P}(G_j = \emptyset \mid G_{j-1} = g)$[3], and $x_0 = \mathbb{P}(X_j \mid X_{j-1})$ if $X_j = X_{j-1}$ and $x_1 = \mathbb{P}(X_j \mid X_{j-1})$ if $X_j \neq X_{j-1}$ (2.7). We can break the terms in the summation above into four distinct classes:

- $s' = \emptyset$ and $x' = x$: the single term in the summation is

$$\alpha_{j-1}(x, \emptyset)\, x_0\, g_{00}. \tag{2.21}$$

- $s' = g \in \{1, \cdots, k\}$ and $x' = x$: the partial sum in the summation is

$$\left(\sum_g \alpha_{j-1}(x, g) \right) x_0\, g_{10}. \tag{2.22}$$

- $s' = \emptyset$ and $x' \neq x$: the partial sum in the summation is

$$\left(\sum_{x' \neq x} \alpha_{j-1}(x', \emptyset) \right) x_1\, g_{00} = \left(\sum_{x'} \alpha_{j-1}(x', \emptyset) - \alpha_{j-1}(x, \emptyset) \right) x_1\, g_{00}. \tag{2.23}$$

- $s' = g \in \{1, \cdots, k\}$ and $x' \neq x$: the partial sum in the summation is

$$\left(\sum_{x' \neq x, g} \alpha_{j-1}(x', g) \right) x_1\, g_{10} = \left(\sum_{x', g} \alpha_{j-1}(x', g) - \sum_g \alpha_{j-1}(x, g) \right) x_1\, g_{10}. \tag{2.24}$$

A brute force computation of $\alpha_j(x, \emptyset)$ in (2.20) for $x = 1, \cdots, k$ takes $O(k^3)$ time. However, note that the first terms in the parentheses of (2.23) and (2.24) are independent of x hence need to be computed only once. Moreover, the second term in the parentheses of (2.24) is exactly the term in the parentheses of (2.22). These algorithmic shortcuts help us reduce the complexity of computing $\alpha_j(x, \emptyset)$ for $x = 1, \cdots, k$ to $O(k^2)$.

2.4 Results

In this section, we summarize the performance of our method on simulated data and then consider a real biological application. For non-boundary cases of the simulation ($\rho \neq 0, \gamma \neq 0$) and the real dataset, we compare our method with GenCo, the method developed by Gay *et al.* (2007).

[3]This is unambiguous as the transition probability $\mathbb{P}(G_j = \emptyset \mid G_{j-1} = g)$ does not depend on g (Table 2.1).

2.4.1 Simulation Study (Non-boundary Cases)

To test the performance of our method, we used Hudson's (2002) coalescent simulation program MS to generate simulated data sets. In general, it is possible that the evolutionary history of a particular region R in a genome involves gene conversions with one end of the conversion tract falling outside R and the other end falling within R. To account for such events, we simulated a 30 kb region and then discarded 5 kb from each end. In all simulations, we used $\theta = 1.0$/kb for mutation rate, which is relevant to humans (see Ptak *et al.* 2004 and Frisse *et al.* 2001, respectively). For each data set, both GenCo and our method were each run 10 times, taking 20 random permutations of haplotype order in each iteration. The same permutations were used in the two methods. In the first iteration, both GenCo and our method started the optimization procedure at the true values of ρ, γ and λ, while in the subsequent iterations, the maximum likelihood estimates from the previous iteration were used as initial values.

The mean tract length λ was set to $0.3, 0.5$ or 1.5 kb in the simulation. For the crossover rate, we used $\rho = 0.5$ or 1.0 per kb, while for the gene conversion rate, we used $\gamma = 0.5, 1.0, 2.5, 5.0$ or 10.0 per kb. Different combinations of ρ and γ result in the ratio $f = \gamma/\rho$ ranging from 0.5 to 10. For each parameter setting, we generated 100 simulated data sets each with 20 haplotypes. For each simulated data set, we estimated all three parameters ρ, γ, and λ, while θ was set to Watterson's unbiased estimator (2.6).

Results are summarized in Table 2.2, Table 2.3 and Table 2.4. The columns labeled $\hat{\rho}, \hat{\gamma}$, and $\hat{\lambda}$ display the mean and standard deviation (shown in parentheses) of the corresponding estimates, whereas the column labeled $\hat{f}$ corresponds to the median of the estimates $\hat{\gamma}/\hat{\rho}$. The column labeled $\#\hat{\rho}_k$ shows the number of data sets with crossover estimates $\hat{\rho}$ within a factor of k from the true ρ; the columns labeled $\#\hat{\gamma}_k$ and $\#\hat{\lambda}_k$ are similarly defined for gene conversion rate γ and the mean tract length λ, respectively.

Estimation of ρ: Both our method and GenCo produced reasonable estimates of ρ. The two estimates had similar means for most parameter settings, but our estimate generally had a smaller variance than that of GenCo.

Estimation of γ: Our improvement over GenCo is clearly illustrated in the estimation of γ. For small values of γ, GenCo's estimates were substantially biased upward, with means above the true γ by factors of tens to thousands. In most cases, this significant bias was not a result of only a few outliers; as the column labeled $\#\hat{\gamma}_{10}$ in the tables and the histogram in Figure 2.11(a) show, GenCo produced very large estimates of γ for a significant fraction of simulated data sets. In contrast, as the tables and the histogram in Figure 2.11(b) indicate, our estimates of γ were much more well-behaved for all parameter settings, though it was slightly biased upward for small γ and biased downward for large γ. Interestingly,

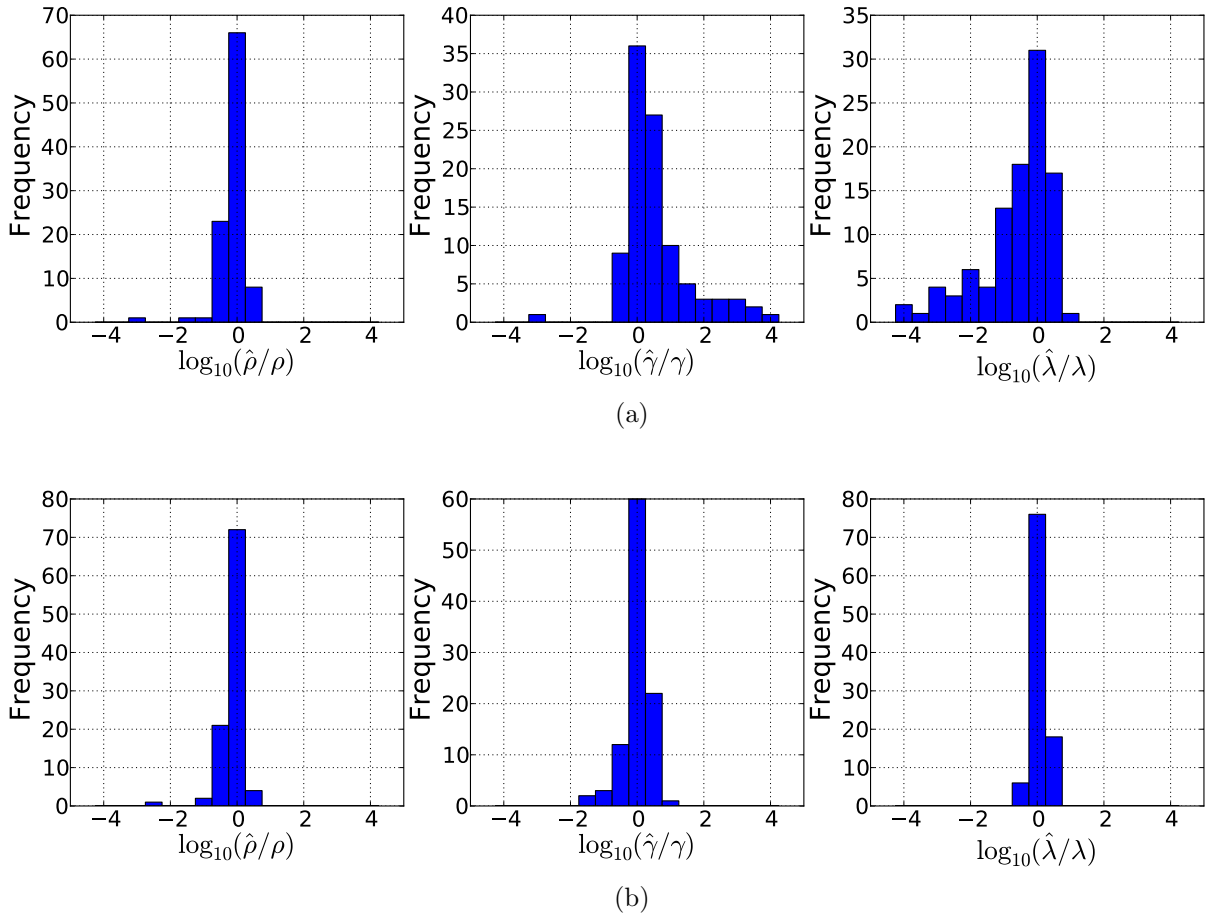
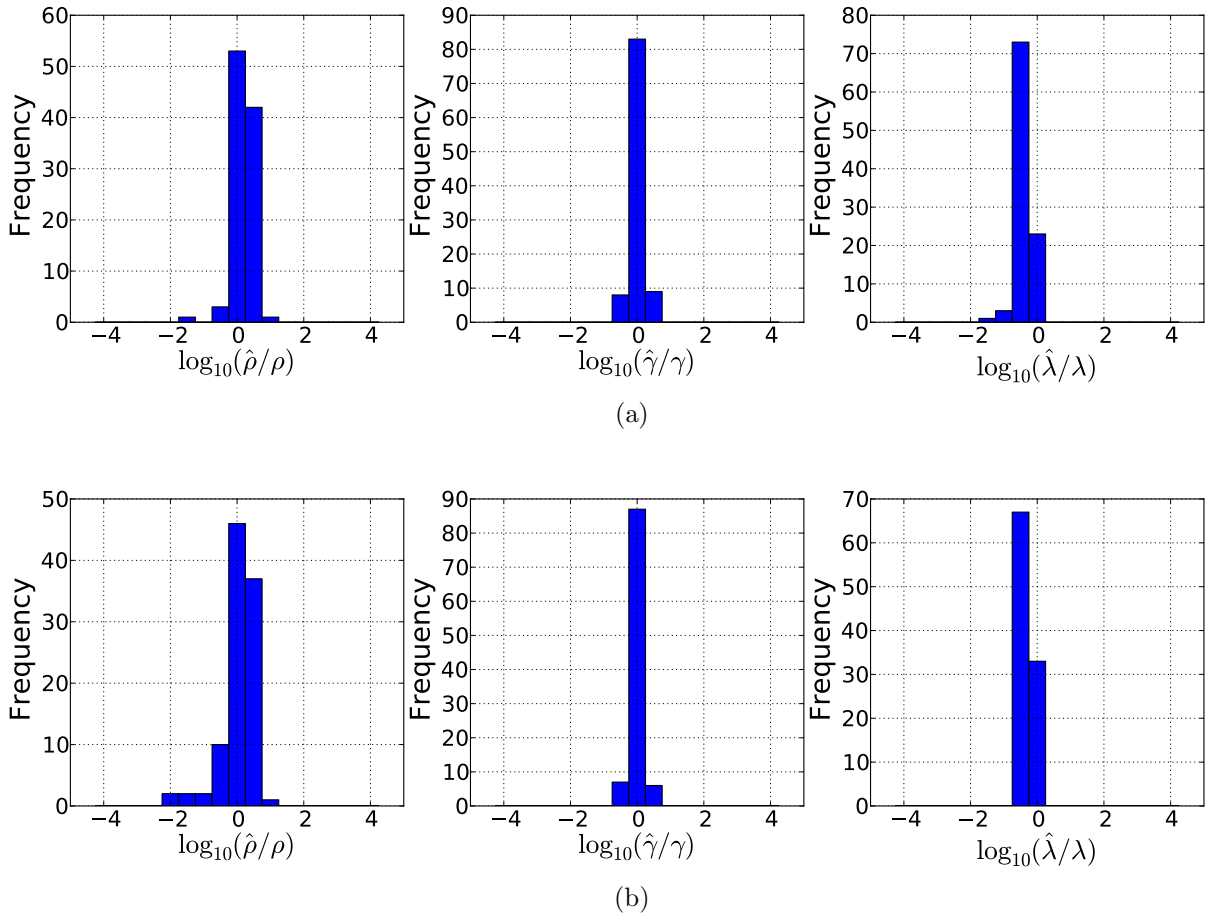


Figure 2.11: Histograms of crossover rate estimates $\hat{\rho}$, gene conversion rate estimates $\hat{\gamma}$ and mean conversion tract length estimates $\hat{\lambda}$ relative to their true values. Based on 100 simulations with $n = 20, \rho = 0.5/\text{kb}, \gamma = 1.0/\text{kb}$ and $\lambda = 0.5$ kb. (a) GenCo; (b) OVERPAINT.

the performance of GenCo was better for large values of true γ, as shown in the tables and Figure 2.12(a).

Estimation of f: The ratio of γ and ρ was estimated by $\hat{f} = \hat{\gamma}/\hat{\rho}$. As it is a ratio of two estimates, we report the median instead of the mean since the median is more robust to large estimates of γ and small estimates of ρ. In general, our median estimates of f were closer to true f than those of GenCo.

Estimation of λ: In GenCo, a very large $\hat{\gamma}$ was usually accompanied by a very small $\hat{\lambda}$. Although the means of GenCo's estimate of λ were closer to true values than ours in some



(a)

(b)

Figure 2.12: Histograms of crossover rate estimates $\hat{\rho}$, gene conversion rate estimates $\hat{\gamma}$ and mean conversion tract length estimates $\hat{\lambda}$ relative to their true values. Based on 100 simulations with $n = 20, \rho = 0.5/\text{kb}, \gamma = 5.0/\text{kb}$ and $\lambda = 1.5$ kb. (a) GenCo; (b) OVERPAINT.

settings, our estimates have smaller variances. The means of estimates are not particularly useful here since a large proportions of small estimates can be readily remedied by a few large estimates. For example, the mean of $\hat{\lambda}$ in Figure 2.11(a) is 0.47, closer to 0.5 than that of ours, which is 0.64. However, our estimates of γ are much more concentrated around the true value. As Table 2.2, Table 2.4 and Figure 2.12(b) suggest, our method tends to overestimate the mean tract length λ when it is small ($\lambda = 0.3$ kb) and underestimate it when it is large ($\lambda = 1.5$ kb).

ρ	γ	f	Method	$\hat{\rho}$	$\hat{\gamma}$	$\hat{f}$	$\hat{\lambda}$	$\#\hat{\rho}_2$	$\#\hat{\gamma}_2$	$\#\hat{\lambda}_2$	$\#\hat{\rho}_{10}$	$\#\hat{\gamma}_{10}$	$\#\hat{\lambda}_{10}$
0.5	0.5	1.0	OVERPAINT	0.47(0.23)	1.05(0.93)	1.96	0.65(0.29)	82	44	51	100	99	100
			GenCo	0.58(0.62)	682.96(3555.37)	8.45	0.28(0.50)	80	20	28	97	54	56
0.5	1.0	2.0	OVERPAINT	0.44(0.21)	1.24(1.30)	2.40	0.63(0.27)	81	60	49	98	98	100
			GenCo	0.54(0.75)	277.54(635.55)	6.37	0.41(0.74)	78	38	26	98	66	64
0.5	2.5	5.0	OVERPAINT	0.41(0.25)	2.03(1.12)	4.64	0.53(0.29)	68	69	74	96	100	100
			GenCo	0.43(0.27)	152.02(766.02)	9.08	0.50(1.22)	69	65	69	97	86	84
0.5	5.0	10.0	OVERPAINT	0.46(0.21)	4.02(2.54)	7.93	0.45(0.18)	82	77	82	100	99	100
			GenCo	0.81(3.36)	231.94(1602.24)	12.56	0.29(0.19)	81	79	71	99	94	93
1.0	0.5	0.5	OVERPAINT	0.77(0.32)	0.81(0.54)	0.89	0.71(0.26)	81	63	39	99	98	100
			GenCo	0.82(0.40)	157.25(586.59)	2.83	0.53(1.25)	80	26	32	99	71	70
1.0	1.0	1.0	OVERPAINT	0.78(0.33)	1.08(0.71)	1.43	0.66(0.24)	83	65	48	100	95	100
			GenCo	0.79(0.34)	296.10(1037.20)	3.30	0.32(0.46)	84	40	35	100	72	71
1.0	2.5	2.5	OVERPAINT	0.79(0.34)	1.97(1.16)	2.56	0.58(0.21)	76	69	63	100	98	100
			GenCo	0.82(0.38)	157.18(585.82)	4.78	0.34(0.34)	80	59	51	100	89	87
1.0	10.0	10.0	OVERPAINT	0.85(0.44)	7.55(3.08)	9.21	0.39(0.15)	76	81	94	97	100	100
			GenCo	0.90(1.00)	297.49(1930.14)	12.44	0.29(0.30)	75	77	75	99	95	92

Table 2.2: Comparison of different methods on simulated data ($\lambda = 0.3$ kb). The estimates of ρ and γ are per kb. For each triplet (ρ, γ, λ), we generated 100 simulated data sets using MS (Hudson, 2002) for $\theta = 1.0$/kb and 20 haplotypes. Shown in the columns labeled $\hat{\rho}$, $\hat{\gamma}$ and $\hat{\lambda}$ are the mean and standard deviation (shown in parenthesis) of the corresponding parameter estimates. The column labeled $\hat{f}$ corresponds to the median of the estimates $\hat{\gamma}/\hat{\rho}$. The symbol $\#\hat{\rho}_k$ denotes the number of data sets with an estimate $\hat{\rho}$ within a factor of k from the true ρ. The symbols $\#\hat{\gamma}_k$ and $\#\hat{\lambda}_k$ are similarly defined for γ and λ, respectively.

ρ	γ	f	Method	$\hat{\rho}$	$\hat{\gamma}$	$\hat{f}$	$\hat{\lambda}$	$\#\hat{\rho}_2$	$\#\hat{\gamma}_2$	$\#\hat{\lambda}_2$	$\#\hat{\rho}_{10}$	$\#\hat{\gamma}_{10}$	$\#\hat{\lambda}_{10}$
0.5	0.5	1.0	OVERPAINT	0.48(0.28)	0.96(0.67)	1.78	0.68(0.27)	71	55	88	99	98	100
			GenCo	0.67(1.86)	248.45(980.26)	5.35	0.49(0.85)	71	26	36	99	67	68
0.5	1.0	2.0	OVERPAINT	0.45(0.24)	1.40(1.09)	3.10	0.64(0.28)	79	70	87	98	97	100
			GenCo	0.47(0.29)	156.39(770.66)	6.15	0.47(0.59)	78	48	35	97	78	76
0.5	2.5	5.0	OVERPAINT	0.51(0.24)	2.12(1.20)	4.02	0.63(0.19)	84	77	98	100	100	100
			GenCo	0.52(0.30)	198.24(1205.65)	5.50	0.52(0.55)	81	75	72	100	94	91
0.5	5.0	10.0	OVERPAINT	0.46(0.35)	4.32(1.83)	11.19	0.56(0.21)	61	90	95	93	100	100
			GenCo	0.46(0.32)	50.55(447.27)	12.18	0.46(0.25)	61	82	79	97	99	99
1.0	0.5	0.5	OVERPAINT	0.79(0.29)	0.93(0.75)	1.11	0.73(0.27)	84	49	84	99	98	100
			GenCo	0.87(0.57)	404.04(1358.41)	3.89	0.38(0.61)	83	18	27	99	59	58
1.0	1.0	1.0	OVERPAINT	0.76(0.33)	1.22(0.75)	1.59	0.74(0.32)	81	75	83	99	100	100
			GenCo	0.78(0.36)	447.16(3288.46)	2.67	0.56(0.56)	75	60	49	98	84	76
1.0	2.5	2.5	OVERPAINT	0.71(0.34)	2.34(1.23)	3.33	0.67(0.25)	72	83	88	99	100	100
			GenCo	0.71(0.33)	209.49(844.96)	4.52	0.49(0.36)	73	73	66	99	88	86
1.0	10.0	10.0	OVERPAINT	0.87(0.49)	8.88(3.99)	10.20	0.46(0.14)	74	85	96	98	100	100
			GenCo	0.81(0.41)	20.07(101.95)	11.74	0.39(0.16)	74	85	81	99	99	99

Table 2.3: Comparison of different methods on simulated data ($\lambda = 0.5$ kb). The estimates of ρ and γ are per kb. For each triplet (ρ, γ, λ), we generated 100 simulated data sets using MS (Hudson, 2002) for $\theta = 1.0$/kb and 20 haplotypes. Shown in the columns labeled $\hat{\rho}$, $\hat{\gamma}$ and $\hat{\lambda}$ are the mean and standard deviation (shown in parenthesis) of the corresponding parameter estimates. The column labeled $\hat{f}$ corresponds to the median of the estimates $\hat{\gamma}/\hat{\rho}$. The symbol $\#\hat{\rho}_k$ denotes the number of data sets with an estimate $\hat{\rho}$ within a factor of k from the true ρ. The symbols $\#\hat{\gamma}_k$ and $\#\hat{\lambda}_k$ are similarly defined for γ and λ, respectively.

ρ	γ	f	Method	$\hat{\rho}$	$\hat{\gamma}$	$\hat{f}$	$\hat{\lambda}$	$\#\hat{\rho}_2$	$\#\hat{\gamma}_2$	$\#\hat{\lambda}_2$	$\#\hat{\rho}_{10}$	$\#\hat{\gamma}_{10}$	$\#\hat{\lambda}_{10}$
0.5	0.5	1.0	OVERPAINT	0.46(0.29)	1.02(0.60)	2.57	0.98(0.46)	68	43	59	99	98	100
			GenCo	0.56(0.58)	354.61(2111.31)	4.62	1.54(5.91)	61	31	44	93	82	70
0.5	1.0	2.0	OVERPAINT	0.54(0.28)	1.45(0.72)	2.79	0.90(0.38)	72	80	62	98	100	100
			GenCo	0.67(0.71)	148.36(806.17)	3.96	1.25(3.72)	72	58	47	96	86	81
0.5	2.5	5.0	OVERPAINT	0.74(1.05)	2.74(1.12)	4.42	0.94(0.38)	62	89	67	92	100	100
			GenCo	0.76(0.79)	3.88(8.00)	4.49	0.91(0.79)	61	89	57	95	99	97
0.5	5.0	10.0	OVERPAINT	0.85(0.67)	5.18(1.97)	6.19	0.79(0.24)	55	92	55	95	100	100
			GenCo	0.93(0.59)	5.54(2.43)	5.89	0.70(0.30)	64	90	40	99	100	99
1.0	0.5	0.5	OVERPAINT	0.80(0.35)	1.07(0.60)	1.33	0.83(0.33)	82	41	48	97	97	100
			GenCo	0.89(0.59)	144.44(453.63)	2.42	0.70(0.97)	80	24	30	99	75	62
1.0	1.0	1.0	OVERPAINT	0.76(0.40)	1.47(0.85)	2.01	0.96(0.36)	73	73	69	95	99	100
			GenCo	5.10(42.01)	67.96(320.83)	2.93	1.29(3.28)	66	59	48	95	89	81
1.0	2.5	2.5	OVERPAINT	0.94(0.46)	2.78(1.29)	3.00	0.84(0.28)	78	86	59	99	100	100
			GenCo	1.01(0.70)	98.85(854.27)	4.07	0.78(0.46)	77	80	49	100	97	89
1.0	10.0	10.0	OVERPAINT	1.60(2.04)	11.08(4.91)	9.54	0.64(0.21)	55	88	21	92	99	100
			GenCo	1.93(1.98)	11.68(5.33)	7.37	0.51(0.30)	56	87	13	96	100	95

Table 2.4: Comparison of different methods on simulated data ($\lambda = 1.5$ kb). The estimates of ρ and γ are per kb. For each triplet (ρ, γ, λ), we generated 100 simulated data sets using MS (Hudson, 2002) for $\theta = 1.0$/kb and 20 haplotypes. Shown in the columns labeled $\hat{\rho}$, $\hat{\gamma}$ and $\hat{\lambda}$ are the mean and standard deviation (shown in parenthesis) of the corresponding parameter estimates. The column labeled $\hat{f}$ corresponds to the median of the estimates $\hat{\gamma}/\hat{\rho}$. The symbol $\#\hat{\rho}_k$ denotes the number of data sets with an estimate $\hat{\rho}$ within a factor of k from the true ρ. The symbols $\#\hat{\gamma}_k$ and $\#\hat{\lambda}_k$ are similarly defined for γ and λ, respectively.

2.4.2 Simulation Study (Boundary Cases)

There are two separate cases with parameters at the boundary: either $\gamma = 0$ or $\rho = 0$. Table 2.5 shows that, for those data sets generated with gene conversions only ($\rho = 0$), the estimates of parameters by OVERPAINT were well behaved: most of $\hat{\rho}$ were close to the true value 0. On the other hand, for the setting of $\gamma = 0$, the estimates of γ were significantly far away from 0 (Table 2.6), which was also observed in Gay *et al.* (2007) by using their method GenCo.

γ	$\hat{\rho}$	$\hat{\gamma}$	$\hat{\lambda}$	$\#(\hat{\rho}; 0.05)$	$\#(\hat{\rho}; 0.1)$
0.5	0.03(0.05)	1.50(1.21)	0.56(0.23)	60	74
1.0	0.03(0.05)	1.81(2.01)	0.59(0.22)	77	90
2.5	0.05(0.06)	3.08(1.77)	0.54(0.19)	90	99
5.0	0.05(0.07)	4.55(1.69)	0.52(0.14)	96	99
10.0	0.12(0.15)	9.31(4.18)	0.48(0.15)	97	100

Table 2.5: Summary of results on simulated data sets with $\rho = 0.0$/kb, $\theta = 1.0$/kb and $\lambda = 0.5$ kb. For each possible value of γ, 100 data sets were independently generated by MS program (Hudson, 2002), each of size 20 haplotypes. Shown in the columns labeled $\hat{\rho}$, $\hat{\gamma}$ and $\hat{\lambda}$ are the mean and standard deviation (shown in parenthesis) of the corresponding parameter estimates. The symbol $\#(\hat{\rho}; k)$ denotes the number of data sets with an estimate $\hat{\rho}$ in the range between 0 and $k\gamma$.

ρ	$\hat{\rho}$	$\hat{\gamma}$	$\hat{\lambda}$	$\#(\hat{\gamma}; 0.05)$	$\#(\hat{\gamma}; 0.1)$
0.5	0.45(0.22)	0.71(0.62)	0.66(0.25)	6	11
1.0	0.75(0.29)	0.71(0.60)	0.73(0.28)	4	10
2.5	1.54(0.68)	0.78(0.61)	0.81(0.25)	14	19
5.0	2.59(0.96)	1.21(0.79)	0.79(0.22)	7	20
10.0	5.24(8.94)	2.89(2.81)	0.75(0.29)	4	13

Table 2.6: Summary of results on simulated data sets with $\gamma = 0.0$/kb and $\theta = 1.0$/kb. For each possible value of ρ, 100 data sets were independently generated by using MS program (Hudson, 2002), each of size 20 haplotypes. Shown in the columns labeled $\hat{\rho}$, $\hat{\gamma}$ and $\hat{\lambda}$ are the mean and standard deviation (shown in parenthesis) of the corresponding parameter estimates. The symbol $\#(\hat{\gamma}; k)$ denotes the number of data sets with an estimate $\hat{\gamma}$ in the range between 0 and $k\rho$.

Therefore, it is the case of $\gamma = 0$ that we mainly focus on in this section, but it is straightforward to develop a similar hypothesis testing procedure (as outlined in Algorithm 3) for the setting of $\rho = 0$.

We first tested on the datasets that were generated under the null hypothesis $\gamma = 0$. The nuisance parameter ρ was set to $0.5, 2.5$ and 10/kb. For each value of ρ, we generated 100 simulated datasets each with 20 haplotypes. For each simulated dataset, we applied a parametric bootstrap procedure (Algorithm 3) with $B = 50$ to obtain an estimate of p-value (2.19). The resulting p-values are plotted in Figure 2.13. Note that these null p-values were approximately uniformly distributed.

Next we considered the datasets simulated under the alternative hypothesis $\gamma \neq 0$. Different combinations of ρ and γ were chosen in the simulation so that the resulting ratio $f = \gamma/\rho$ is $0.5, 1.0, 2.5, 5.0$ and 10.0. The mean tract length λ was set to 0.5 kb and the mutation rate θ was set to 1.0/kb. For each parameter setting, we generated 100 datasets with sample size $n = 20$ and $n = 35$, respectively. Figure 2.14 shows the bootstrap estimates of alternative p-values. As expected, the distributions of alternative p-values tend to concentrate closer to 0. Furthermore, the larger the ratio f or the more samples we have, the more likely that the null hypothesis will be rejected by our test procedure.

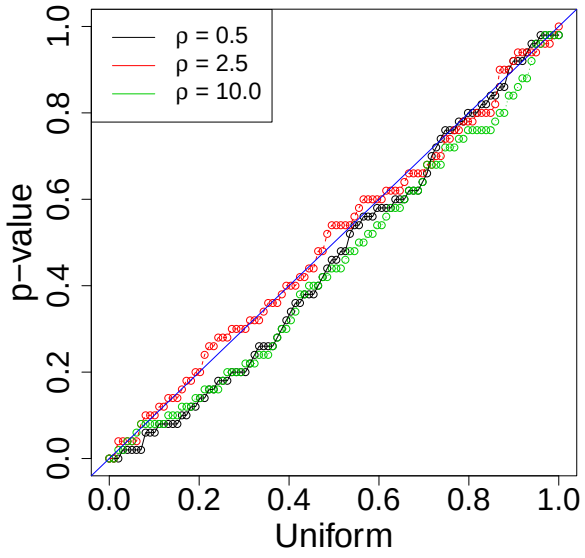


Figure 2.13: Bootstrap estimates of p-value under the null hypothesis $H_0 : \gamma = 0$. For each value of ρ, 100 data sets were independently generated by using MS program (Hudson, 2002), with sample size $n = 20$ and $\theta = 1.0$/kb. Shown in the figure is a Q-Q plot of p-value estimates by parametric bootstrap with $B = 50$ versus a uniform distribution.

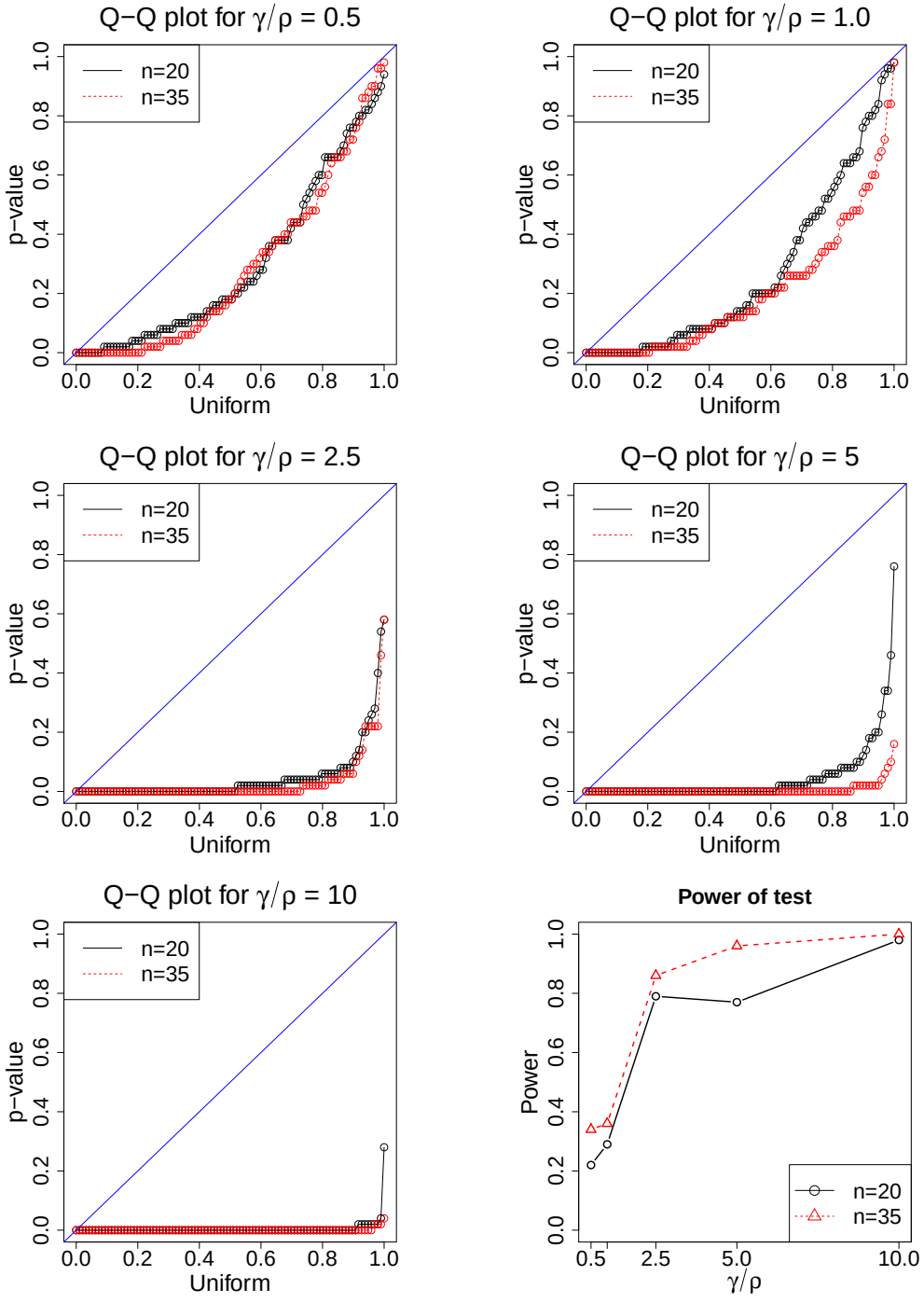


Figure 2.14: Bootstrap estimates of p-value under the alternative hypothesis $H_1 : \gamma \neq 0$. The first five sub-figures show the Q-Q plots of 100 p-value estimates versus a uniform distribution for different settings of f. The last sub-figure plots the power of the test when setting the p-value threshold at 0.05.

2.4.3 A Real Biological Application

Gay *et al.* (2007) used their method to study recombination patterns in two genes—namely, $su(s)$ and $su(w^a)$ surveyed by Langley *et al.* (2000)—located near the telomere of the X chromosome of *Drosophila melanogaster*. The $su(s)$ and $su(w^a)$ loci are about 4.1 kb and 2.5 kb long, respectively, and are about 400 kb apart. Langley *et al.* (2000) surveyed samples from both an African and a European population, but only the African sample was considered by Gay *et al.*, and we do the same here. The $su(s)$ data set contains 50 haplotypes and 41 SNPs, while the $su(w^a)$ data set contains 50 haplotypes and 46 SNPs.

Gay *et al.* reported that, upon fixing the mean tract length to 0.352 kb (Hilliker *et al.*, 1994), they obtained $\hat{\rho} = 0.067$/kb and $\hat{\gamma} = 26.9$/kb, thus concluding $\hat{\gamma}/\hat{\rho} = 432$. In their paper, Gay *et al.* did not specify whether the above estimates were for the $su(s)$ locus or the $su(w^a)$ locus. To compare their method GenCo with our method, we redid the analysis, following the same procedure as in Section 2.4.1, i.e., taking 20 random permutations of haplotype order and iterating the computation 10 times. We used $\rho = 1.0$/kb and $\gamma = 1.0$/kb as the starting values of the optimization procedure in the first iteration. The results are summarized in Table 2.7. Assuming $\lambda = 0.352$ kb, GenCo and OVERPAINT produced similar estimates for the $su(s)$ locus, though the ratio of the gene conversion rate to the crossover rate estimated by OVERPAINT was not as high as that of GenCo. The estimates for the $su(w^a)$ are more different: OVERPAINT yielded a much smaller estimate of ρ than that of GenCo, resulting in a much larger estimate of the ratio f.

We also performed analysis with λ as a free parameter; Gay *et al.* (2007) did not consider this analysis in their study. In this case, we used $\rho = 5.0$/kb, $\gamma = 5.0$/kb, and $\lambda = 0.352$ kb as the starting values of the optimization procedure in the first iteration. The corresponding maximum likelihood estimates of ρ, γ, and λ are shown in Table 2.8. For the $su(s)$ locus, GenCo and our method produced similar estimates of λ, but OVERPAINT still produced a smaller estimate of f compared to GenCo. For the $su(w^a)$ locus, though the estimate of

Gene	Method	$\hat{\rho}$	$\hat{\gamma}$	$\hat{\gamma}/\hat{\rho}$
$su(s)$	GenCo	1.63	12.82	7.87
	OVERPAINT	2.24	11.51	5.14
$su(w^a)$	GenCo	0.48	27.85	58.02
	OVERPAINT	0.033	27.04	819.40

Table 2.7: Estimates of ρ and γ for the $su(s)$ and $su(w^a)$ loci in *Drosophila melanogaster*, with λ held fixed at 0.352 kb. The estimates of ρ and γ are per kb.

Gene	Method	$\hat{\rho}$	$\hat{\gamma}$	$\hat{\gamma}/\hat{\rho}$	$\hat{\lambda}$
$su(s)$	GenCo	0.92	11.60	12.61	0.48
	OVERPAINT	1.29	9.86	7.64	0.56
$su(w^a)$	GenCo	8.52	251.13	29.48	0.005
	OVERPAINT	1.45	41.09	28.34	0.162

Table 2.8: Estimates of ρ, γ, and λ for the $su(s)$ and $su(w^a)$ loci in *Drosophila melanogaster*. The estimates of ρ and γ are per kb, while the estimate of λ is in kb.

ratio by OVERPAINT was slightly smaller than that of GenCo, GenCo produced much larger estimates of ρ and γ, while the opposite is true for the mean tract length λ. The estimate of λ by GenCo was extremely small, outside the typical range $50 \sim 2000$ bp, whereas our estimate seemed more biologically reasonable. This could be an artifact of the model GenCo, which tends to produce small estimates of λ when estimates of γ are large.

The fact that both methods detected strong signals of gene conversion suggests that gene conversion is likely to have played an important role in shaping the observed pattern of genetic variation in the two genes, which agrees with Langley *et al.*'s conclusion. However, when treating the mean tract length as a free parameter, our analysis implies that crossover may not have been greatly suppressed in the $su(s)$ and $su(w^a)$ loci, as Gay *et al.* (2007) concluded.

2.5 Summary

In this chapter, we extended a PAC-based model by explicitly allowing overlapping gene conversions. This extension significantly improves the previous work for the task of jointly estimating three parameters essential to recombination: the crossover rate, the gene conversion rate and the mean conversion tract length; although the previous model can produce reasonable estimates of parameters in some regimes, our estimates are more robust and well-behaved for almost all parameter settings. We believe that this aspect of our model is crucial in making the joint estimation of the gene conversion rate and the mean conversion tract length feasible. Along the way, we also demonstrated a parametric bootstrap procedure for testing the parameters at the boundary of the range of possible values, which can provide a statistical significance score on whether the region of interest is subject to the effects of gene conversion.

Although the joint estimation of the three parameters ρ, γ, and λ is indeed a very difficult

problem, and the model proposed here is unlikely to be optimal, we believe that we have taken an important step towards devising a more realistic and reliable model.

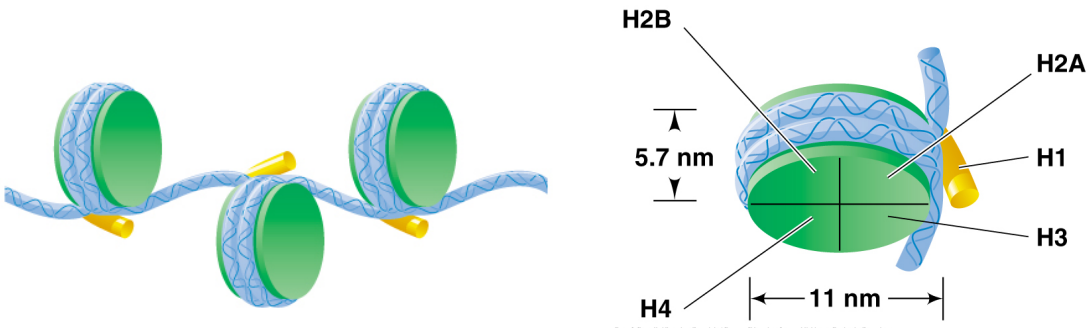
Chapter 3

Recombination and Nucleosome Positioning

3.1 Introduction

The backbone of the DNA molecule is negatively charged due to the presence of phosphate ions, hence long strands of eukaryotic DNA by itself do not have the ability to fold up to fit into the tiny cell nucleus. Several levels of compaction via DNA-protein interactions enable the DNA to fit inside the cell, resulting in a DNA-protein complex known as *chromatin*. In the first level of compaction, the DNA of eukaryotic genome wraps around many octamers of histone proteins to form a beads-on-a-string structure (Figure 3.1(a)). The histones are highly positively charged so that they can bind to and neutralize the negatively charged DNA. Each basic DNA packaging unit, called *nucleosome*, consists of a 147-bp long stretch of DNA wrapped twice around a core composed of eight histones, two copies each of H2A, H2B, H3 and H4 (Figure 3.1(b)). Consecutive nucleosome core particles are connected by a stretch of intervening DNA termed "linker DNA", which is typically about 20-50 bp long. This beads-on-a-string structure can be further compacted by the linker histone H1 into higher-order 30 nm fibres.

Characterizing the organization of nucleosome *in vivo* is crucial for understanding gene regulation as the positions can strongly influence the DNA-binding ability of transcription factors (Jiang and Pugh, 2009). Recent advanced in tiling microarrays and massively parallel DNA sequencing technologies have provided genome-wide mapping of nucleosome locations for many different species (Yuan *et al.*, 2005; Mavrich *et al.*, 2008; Schones *et al.*, 2008; Zhang *et al.*, 2008; Segal *et al.*, 2006; Kaplan *et al.*, 2008). According to the distribution of their positions in a population, nucleosomes can be classified either as localized (phased or well-positioned) or as delocalized (fuzzy) (Figure 3.2). Localized nucleosomes reside within a small range of a genomic coordinate, whereas the positions of delocalized nucleosomes



(a) (b)

Figure 3.1: (a) Beads-on-a-string form of chromatin. Adjacent nucleosomes are connected by a stretch of linker DNA. (b) A schematic of nucleosome structure. Each nucleosome core particle consists of 147 bp of DNA wrapped around a histone octamer, two each of H2A, H2B, H3 and H4. Histone tails (not shown) are subject to different types of covalent modifications, such as acetylation and methylation. Image source: Russell (2009).

Figure 3.2: Localized (left) and delocalized (right) nucleosomes. At a population level, the positions of localized (phased) nucleosomes are tightly concentrated around a genomic coordinate, while the positions of delocalized (fuzzy) nucleosomes are more randomly distributed. Image source: (Jiang and Pugh, 2009).

are more randomly distributed. The distribution for the positions of a localized nucleosome is commonly summarized by a consensus nucleosome position in a nucleosome position map. Evidence has accumulated recently that various factors, including the intrinsic binding affinity of DNA sequence, DNA methylation, histone variants, post-translational tail modifications, and DNA transcription factors, can affect nucleosome positioning (Segal and Widom, 2009). It is indicated that, among all factors, the intrinsic DNA sequence preference has a major role in determining nucleosome positions *in vivo* (Kaplan *et al.*, 2008). Several studies showed that nucleosome occupancy could be predicted from the underlying DNA sequence alone (Gupta *et al.*, 2008; Peckham *et al.*, 2007; Yuan and Liu, 2008). A more

sophisticated model, assuming thermodynamic equilibrium, integrates all these factors into a unified framework to yield a probabilistic distribution over all possible configurations of nucleosomes and transcription factors on the DNA (Segal and Widom, 2009).

In this chapter, we focus on the correlation of nucleosome positions and recombination rates instead of the mechanism of how nucleosomes regulate gene transcription, which has been the subject of an extensive literature (see, for example, Jiang and Pugh (2009) and references therein). Recombination rates can vary widely along chromosomal DNA at several scales. Advances in high-resolution recombination map construction techniques showed that mammalian meiotic recombination events are not randomly distributed, but instead cluster into short (1-2 kb) regions known as recombination hotspots (Kauppi *et al.*, 2004; Paigen and Petkov, 2010). In humans, more than $30,000$ hotspots have been identified and these hotspots typically occur every 50 to 100 kb. These highly localized regions capture a majority of recombination events: 80% of recombination occurs in only 10 to 20% of the sequence (Myers *et al.*, 2005). More recently, much interest has arisen in how hotspot activity is regulated. Several *cis* and *trans*-acting regulatory factors for the activity of hotspots in humans and mice have been identified (Paigen *et al.*, 2008; Baudat and de Massy, 2007; Myers *et al.*, 2008). Remarkably, three independent groups discovered that the *trans*-acting factor, PRDM9, is a major determinant that controls recombination hotspot activation in mice and humans (Myers *et al.*, 2010; Baudat *et al.*, 2010; Parvanov *et al.*, 2010). The precise mechanism of PRDM9 is still not known, but it was suggested that the binding of highly polymorphic PRDM9 zinc-finger protein to the distinct DNA sequences (such as the degenerate 13-mer motif CCNCCNTNNCCNC for human recombination hotspots (Myers *et al.*, 2008)) can facilitate the initiation of DNA double-strand break (DSB) near the binding sites. However, it has been reported that the site selection of DNA DSBs by SPO11 depends on the position of nucleosomes, which behave as a physical barrier for the cleavage activity of SPO11 *in vivo* (Di Felicei *et al.*, 2008) (see top of Figure 2.1). This suggests that nucleosome positions could also affect, at least partly, the activity of recombination, which is supported by recent studies of recombination hotspots in mice (Getun *et al.*, 2010; Wu *et al.*, 2010). In particular, the work showed that the recombination hotspot cores in mice are generally open, with nucleosomes located at genomic positions where crossover activity is relatively suppressed.

However, to the best of our knowledge, such an observation has not been reported in species other than yeast (Wu and Lichten, 1994). Due to previously observed rapidly decaying patterns of LD and the absence of long haplotype blocks, it is generally assumed that *Drosophila melanogaster* does not have similar pattern of recombinational landscape as humans and mice at a global scale, where a majority of recombination events occur within highly localized hotspots. However, recent studies have shown that there still exists significant variation in recombination rate at a fine scale in some genomic regions of *Drosophila melanogaster* (Singh *et al.*, 2009; Kulathinal *et al.*, 2008). Here, we superpose a genome-wide high-resolution ref-

erence map of H2A.Z nucleosomes in *Drosophila melanogaster* (Mavrich *et al.*, 2008) with fine-scale recombination rates estimated by the Drosophila Population Genomics Project (DPGP), and find that the majority of regions that exhibit highly elevated rates of recombination are depleted of localized nucleosomes. Interestingly, we find many regions with low recombination rates are also nucleosome-free.

3.2 Background

The coalescent-based program, *LDhat* (McVean *et al.*, 2004), was used by the Drosophila Population Genomics Project (DPGP) to infer the fine-scale recombination rate variation in 37 RAL lines of 50 *Drosophila melanogaster* Genomes Project (`http://www.dpgp.org/1K_50genomes.html`). It is worth pointing out that *LDhat* can only estimate the rate of recombination that combines effects of crossover and gene conversion. In what follows in this chapter, we will use ρ for this combined rate. This is different from the ρ defined in Chapter 2, where it denotes the crossover rate only. Detailed below is their analysis method.

Dealing with missing data: The computational complexity of handling missing data in *LDhat* is exponential in the number of missing entries, as all the unknown variables need to be marginalized out to compute the likelihood. To circumvent this problem, missing data was removed by using the following procedure. First, for each chromosome, the end points of missing intervals (contiguous missing entries) were found and then used to partition the chromosome into a set of non-overlapping blocks. Within each block, completely missing haplotypes were removed. Finally, the loci with at least one missing entry were also removed. Table 3.1 summarizes the resulting data:

Chromosome	# Blocks	# non-missing haplotypes			Average distance between adjacent SNPs (bps)
		Min	Max	Mean	
2L	20	32	35	33.6	130
2R	14	32	36	34.2	145
3L	18	31	35	33.2	146
3R	21	32	34	33.3	151
X	2	34	35	34.5	406

Table 3.1: Summary of 37 RAL lines of 50 *Drosophila melanogaster* Genomes after data reduction.

Inference of fine-scale recombination rate: Two-locus likelihood lookup tables were generated by *LDhat*'s subprogram *complete*, with ρ ranging from 0 to 500 and step size 0.5. The mutation rate θ was set to 0.006 for autosomes and 0.004 for the X chromosome. To estimate genome-wide fine-scale recombination rates, a sliding-window scheme was applied, with 1000 SNPs in each window and 250 SNPs in adjacent overlapping windows. The reversible-jump MCMC algorithm in *LDhat*'s subprogram *interval* was used to estimate variable recombination rate. The number of iterations for the burn-in period and the total number of iterations to run the chain was set to 200,000 and 5 million, respectively. Successive samples were taken every 2000 iterations. The estimates of recombination rate for 125 SNPs from the ends of each window were discarded in order to stitch together the estimates in adjacent windows. To avoid over-fitting, the prior for the number of change-points in the map of recombination rates is set to a Poisson distribution with mean $(S-2)\exp(-\xi)$, where S is the number of SNPs in each window and ξ is a penalty parameter. Large values of penalty can detect change-points where there is strong signal for changes in recombination rate, but the estimates could be too smooth and lose detail. Small values of penalty could display detail, but might introduce noise.

3.3 Results and Discussion

ChIP-Seq mapping technology was used by Mavrich *et al.* (2008) to obtain a genome-wide high-resolution reference map of H2A.Z nucleosomes in *Drosophila melanogaster*. H2A.Z is a variant of core histone H2 that has been reported as a hallmark of active genes and is widely distributed in *Drosophila* (Leach *et al.*, 2000). Shown in Table 3.2 is a summary of nucleosome positioning dataset.

Chromosome	Length (Mb)	# Localized H2A.Z nucleosomes
2L	23	38550
2R	21	41001
3L	24.5	42128
3R	28	49305
X	22.5	34306

Table 3.2: Summary of H2A.Z nucleosomes in *Drosophila melanogaster*.

The recombination rate estimates discussed in Section 3.2 are piecewise constant along the chromosome. For each chromosome, we found the set of change-points in the estimated recombination rates and used the resulting change-points to partition the chromosome into

non-overlapping blocks. Hence, recombination rates are constant within each block but vary across consecutive blocks. For each block, the number of localized H2A.Z nucleosomes (including the ones lying on the boundaries) is recorded. We regard the blocks where the estimated recombination rate ρ satisfies $\log_{10}(\rho \text{ per kb}) \geq 2$ (about 10-fold higher than the genome-wide average of recombination rates) as being the regions that exhibit highly elevated activity of recombination. The number of such blocks for each chromosome and the summary statistics for the number of localized nucleosomes in these blocks, are reported in Table 3.3. Note that deceasing ξ from 45 to 15 increases the number of such blocks by a factor of about 6-7.

Chr	ξ	#Blocks	#nuc	%(#nuc = 0)	%(#nuc $\leq$ 1)	%(#nuc $\leq$ 2)	%(#nuc $\leq$ 5)
2L	15	277	0.67 (1.3)	59.6	89.5	96	98.2
	45	41	1.22 (1.9)	34.1	80.5	90.2	95.1
2R	15	174	0.53 (1.2)	63.8	92	97.1	98.9
	45	20	0.45 (0.8)	70	90	95	100
3L	15	377	0.64 (1.6)	62.1	88.1	96	98.9
	45	63	1.2 (2.0)	35	81	90.5	95.2
3R	15	242	0.54 (0.9)	60.7	93	95.9	99.6
	45	41	0.95 (2.0)	58.5	80.5	90.2	97.6
X	15	290	1.37 (3.0)	48.6	74.5	85.2	95.9
	45	53	1.74 (3.3)	43.4	66	79.2	94.3

Table 3.3: Summary for the number of localized H2A.Z nucleosomes within the blocks that exhibit highly elevated rates of recombination: $\log_{10}(\rho \text{ per kb}) \geq 2$. The column labeled #Blocks shows the total number of such blocks for each chromosome. The column labeled #nuc displays the mean and SD (shown in parentheses) for the number of localized nucleosomes in these blocks. The column labeled %(#nuc $\leq k$) denotes the percentage of these blocks with no more than k nucleosomes.

As shown in the tables, the majority of blocks that exhibit highly elevated rates of recombination are nucleosome-depleted, which coincides with the previous observation in four hotspots of mice (Getun *et al.*, 2010). Figure 3.3, Figure 3.5 and Figure 3.4 depict the nucleosome landscapes around several typical blocks where both estimates (with $\xi = 15$ or 45) show strong evidence for elevated recombination rates. Interestingly, it seems that localized nucleosomes tend to reside close to the boundaries of blocks, which suggests that the presence of

localized nucleosomes could influence the local activity of recombination process, possibly by constraining strand invasion and D-loop extension (Figure 2.1). In addition, we also found some localized nucleosomes located at the center of a long block (top right in Figure 3.3). This could be due to the limited resolution of recombination intensity that the statistical approach can provide.

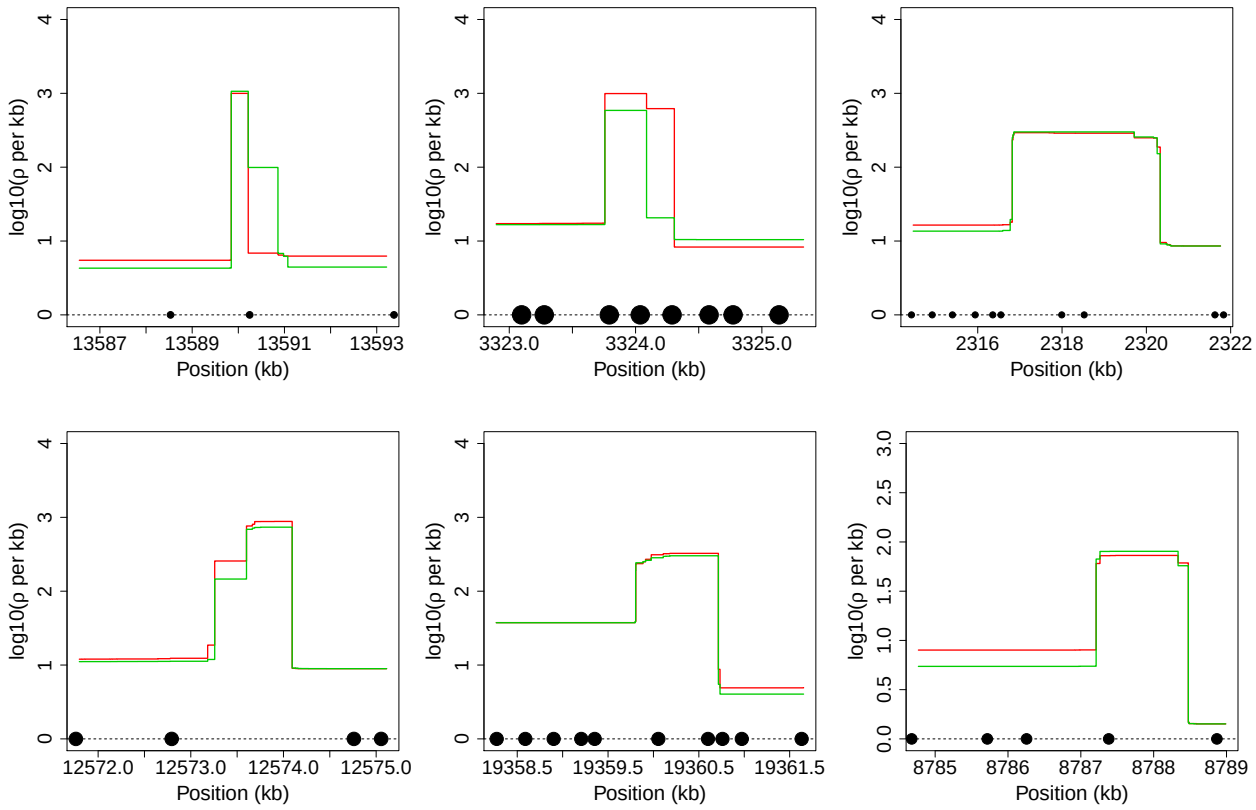


Figure 3.3: Nucleosome occupancy near the regions that exhibit highly elevated rates of recombination in chromosome 2 of *Drosophila melanogaster*. Nucleosomes are drawn to scale. Red: estimates with $\xi = 15$; green: estimates with $\xi = 45$. Top: chromosome 2L; bottom: chromosome 2R.

On the other hand, among the blocks where recombination is suppressed (say, $\log_{10}(\rho \text{ per kb}) \leq -1$), the blocks with few nucleosomes are also observed frequently (Table 3.4), although less frequently than among the blocks where the rates are elevated (Table 3.3). This is also consistent with what has been observed in mice hotspots: "recombinationally code regions can also have large nucleosome-free domains" (Getun *et al.*, 2010).

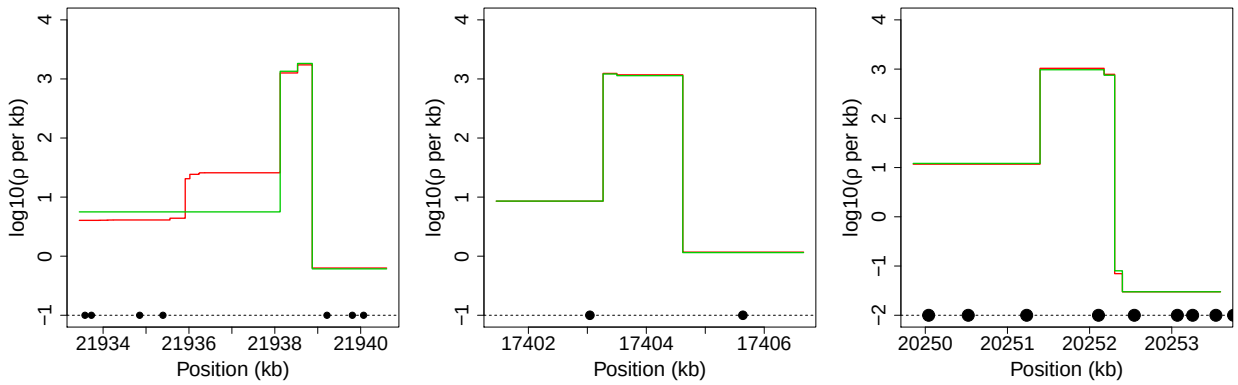


Figure 3.4: Nucleosome occupancy near the regions that exhibit highly elevated rates of recombination in chromosome X of *Drosophila melanogaster*. Nucleosomes are drawn to scale. Red: estimates with $\xi = 15$; green: estimates with $\xi = 45$.

Chr	ξ	#Blocks	#nuc	%(#nuc = 0)	%(#nuc $\leq$ 1)	%(#nuc $\leq$ 2)	%(#nuc $\leq$ 5)
2L	15	56	35.38 (109.6)	37.5	58.9	62.5	75
	45	30	40.73 (85.3)	36.7	60	63.3	66.7
2R	15	83	38.08 (110.7)	43.4	55.4	57.8	66.3
	45	42	63.8 (159.3)	45.2	52.4	54.8	57.1
3L	15	82	40.89 (121.6)	31.7	47.5	53.7	61
	45	26	118 (248.7)	38.5	46.2	46.2	46.2
3R	15	181	12.17 (56.7)	41.4	59.7	65.2	71.8
	45	68	18.5 (78.3)	38.2	58.8	60.3	63.3
X	15	113	18.45 (77.6)	25.7	39.8	51.3	62.8
	45	29	27.07 (56.6)	13.8	20.7	27.6	44.8

Table 3.4: Summary for the number of localized H2A.Z nucleosomes within the blocks where recombination is suppressed: $\log_{10}(\rho \text{ per kb}) \leq -1$. The column labeled #Blocks shows the total number of such blocks for each chromosome. The column labeled #nuc displays the mean and SD (shown in parentheses) for the number of localized nucleosomes in these blocks. The column labeled %(#nuc $\leq k$) denotes the percentage of these blocks with no more than k nucleosomes.

Chr	ξ	$\log_{10}(\rho$ per kb)	#Blocks	Q_1	Q_2	Q_3	p-value$_>$	p-value$_<$
2L	15	≥ 2	277	0	0	5.3	0.94	0.06
		≤ -1	56	0	1.5	2.9		
2L	45	≥ 2	41	0	3.3	6.0	0.16	0.84
		≤ -1	30	0	1.7	2.8		
2R	15	≥ 2	174	0	0	4.3	0.90	0.10
		≤ -1	83	0	1.2	2.9		
2R	45	≥ 2	20	0	0	2.2	0.82	0.18
		≤ -1	42	0	1.1	2.9		
3L	15	≥ 2	377	0	0	4.0	0.99	0.01
		≤ -1	82	0	2.1	3.2		
3L	45	≥ 2	63	0	2.8	6.1	0.08	0.92
		≤ -1	26	0	2.5	2.8		
3R	15	≥ 2	242	0	0	4.1	0.97	0.03
		≤ -1	181	0	1.0	2.8		
3R	45	≥ 2	41	0	0	3.1	0.85	0.15
		≤ -1	68	0	1.1	2.9		
X	15	≥ 2	290	0	0.6	2.4	0.99	0.01
		≤ -1	113	0	1.2	2.3		
X	45	≥ 2	53	0	0.9	2.4	0.92	0.08
		≤ -1	29	0.8	1.3	2.5		

Table 3.5: Summary for the density of localized H2A.Z nucleosomes. Nucleosome density of a block is described by the number of nucleosomes per kb. The columns labeled Q_1, Q_2 and Q_3 are the first, the second and the third quartile of nucleosome densities, respectively. The column labeled p-value$_>$ (p-value$_<$) shows the 1-sided p-value of Wilcoxon rank sum test where the alternative is the nucleosome density in the blocks with $\log_{10}(\rho$ per kb$) \geq 2$ is higher (lower) than that in the blocks with $\log_{10}(\rho$ per kb$) \leq -1$.

We also compare the density of nucleosomes within those two different types of blocks. Table 3.5 indicates that, in most settings (8 out of 10), the 1-sided p-value suggests that the nucleosome density in the blocks with elevated rates of recombination is generally lower

than that in the blocks where recombination is suppressed.

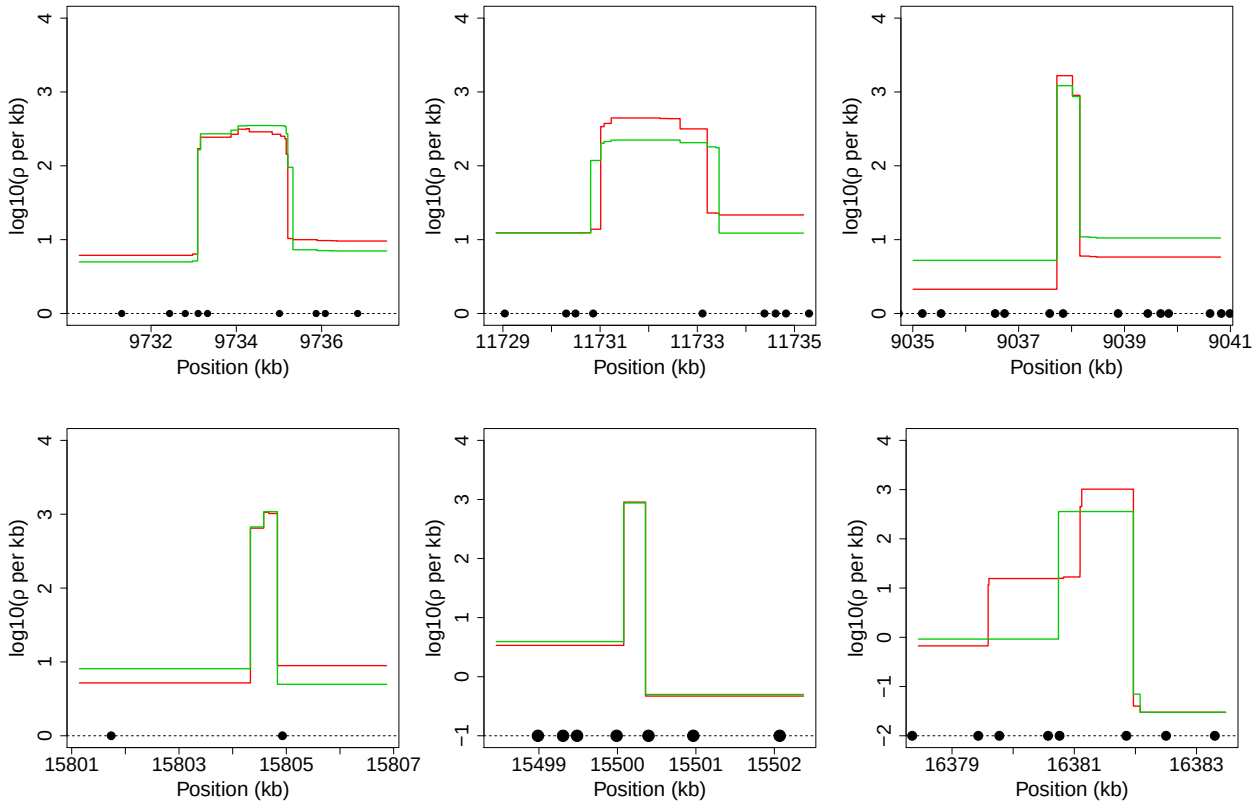


Figure 3.5: Nucleosome occupancy near the regions that exhibit highly elevated rates of recombination in chromosome 3 of *Drosophila melanogaster*. Nucleosomes are drawn to scale. Red: estimates with $\xi = 15$; green: estimates with $\xi = 45$. Top: chromosome 3L; bottom: chromosome 3R.

Together, this suggests that nucleosome-depleted regions might be more favored by recombination process. However, they are not sufficient for recombination to occur. Similarly, it has been reported that DNA sequence motif for hotspots cannot incur recombination by itself since it is also found in genomic regions devoid of recombination (Myers *et al.*, 2008). Hence it is tempting to speculate that both conditions, specific DNA sequence and nucleosome depletion, are required for a site to initiate recombination: the binding of *trans*-acting factors to specific target DNA sequence could modify local chromatin structure so that the nucleosome-free site could be exposed and recognized by SPO11 for DSB formation. In humans and mice, hotspot motifs have been identified. There might also exist such a consensus sequence in *Drosophila melanogaster*, which is enriched in the blocks with high recombination rates but is lacking in the blocks where recombination is suppressed. This could partially

explain why the blocks with low density of nucleosomes are also observed frequently among those blocks with low recombination rates.

Understanding the precise regulatory mechanism of recombination is yet far from complete, our finding suggests that in addition to *cis* and *trans*-acting elements, nucleosome positioning could be another important layer of control for recombination activity, at least in *Drosophila melanogaster*.

Chapter 4

Conclusions

4.1 Summary

In this thesis, we investigate two computational problems that arise in studying meiotic recombination. In Chapter 2, we have developed a likelihood-based model for jointly estimating three fundamental parameters to recombination: the crossover rate, the gene conversion rate and the mean conversion tract length. In particular, we show that modeling overlapping gene conversions is essential for accurate and robust estimation of these parameters. We then apply the method to two genes located near the telomere of the X chromosome of *Drosophila melanogaster*, and the results imply that the ratio of the gene conversion rate to the crossover rate for these two genomic regions may not be nearly as high as previously claimed. In Chapter 3, we study the correlation of nucleosome positions and meiotic recombination rates. We superpose a high-resolution reference map of H2A.Z nucleosomes in *Drosophila melanogaster* with fine-scale estimated recombination rates. We find that the majority of genomic regions with highly elevated rates of recombination are depleted of well-positioned nucleosomes, which suggests that nucleosome occupancy can influence, at least in part, the activity of meiotic recombination.

4.2 Future Directions

High-throughput sequencing and high-resolution nucleosome mapping technology have advanced remarkably in the past few years. It will soon become routine to obtain whole-genome sequence and nucleosome position information. Such fine-scale data will allow us to gain more insights in the mechanisms and properties of meiotic recombination. There are several directions for future investigation under the same theme of this thesis:

- *Coarse-to-fine implementation of likelihood computation for* OVERPAINT. Despite the use of algorithmic shortcuts to reduce the burden of computation for likelihood, the

computational complexity is still $O(n^4L)$, where n is the number of sampled haplotypes and L is the length of each haplotype. Coarse-to-fine scheme to find the best path in HMM (Raphael, 2001) might be generalized to forward-backward algorithm for computing the likelihood.

- *Estimation of variable recombination rates.* The model OVERPAINT has limits in that it can only estimate constant crossover and gene conversion rate. It would be useful to extend the model to allow variable rates by adapting the approach of (Li and Stephens, 2003). Together with the development of coarse-to-fine approximation for likelihood computation, it would be desirable to apply the method to provide a genome-wide map of crossover and gene conversion rates across the human genome, as well as characterize the distribution of conversion tract lengths.

- *Consensus sequence motif for regions with elevated recombination rates in Drosophila melanogaster.* We have depicted the nucleosome landscapes in these regions, but it is unknown whether there exists a DNA sequence motif that intrinsically encodes these regions. If there is one such speculated motif, it could be interesting to examine its richness within the regions where recombination process is suppressed and localized nucleosomes are depleted.

- *Incorporate nucleosome position information into estimation of recombination rates or vice versa.* Change-points detection in estimating recombination rates is a hard problem and depends heavily on the choice of penalty value. The positions of nucleosomes, could be taken into account to facilitate better detection of change-points. For example, those change-points within a localized nucleosome core could be penalized less. Conversely, recombination rates might be incorporated into the framework of (Wasson and Hartemink, 2009) to predict nucleosome positions.

Bibliography

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

Baudat, F., Buard, J., Grey, C., Fledel-Alon, A., Ober, C., Przeworski, M., Coop, G., and de Massy, B. (2010). PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. *Science*, **327**(5967), 836–840.

Carlson, C. S., Eberle, M. A., Kruglyak, L., and Nickerson, D. A. (2004). Mapping complex disease loci in whole-genome association studies. *Nature*, **429**(6990), 446–452.

Chen, J.-M., Cooper, D. N., Chuzhanova, N., Ferec, C., and Patrinos, G. P. (2007). Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics*, **8**(10), 762–775.

Crawford, D. C., Bhangale, T., Li, N., Hellenthal, G., Rieder, M. J., Nickerson, D. A., and Stephens, M. (2004). Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature Genetics*, **36**(7), 700–706.

Di Felicei, F., Chiani, F., and Camilloni, G. (2008). Nucleosomes represent a physical barrier for cleavage activity of DNA topoisomerase I in vivo. *Biochemical Journal*, **409**(3), 651–656.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, First edition.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Rienzo, A. D. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *American Journal of Human Genetics*, **69**(4), 831–843.

Gay, J. C., Myers, S., and McVean, G. (2007). Estimating meiotic gene conversion rates from population genetic data. *Genetics*, **177**(2), 881–894.

Getun, I. V., Wu, Z. K., Khalil, A. M., and Bois, P. R. J. (2010). Nucleosome occupancy landscape and dynamics at mouse recombination hotspots. *EMBO Reports*, **11**(7), 555–560.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden markov models. *Machine Learning*, **29**, 245–273.

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, **3**(4), 479–502.

Gupta, S., Dennis, J., Thurman, R. E., Kingston, R., Stamatoyannopoulos, J. A., and Noble, W. S. (2008). Predicting human nucleosome occupancy from primary sequence. *PLoS Computational Biology*, **4**(8), e1000134.

Hartwell, L., Hood, L., Goldberg, M., and Reynolds, A. (2006). *Genetics: From Genes to Genomes*. McGraw-Hill Science, Third edition.

Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, UK.

Hellenthal, G. (2006). *Exploring rates and patterns of variability in gene conversion and crossover in the human genome*. Ph.D. thesis, University of Washington, Seattle.

Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M., Clark, S. H., and Chovnick, A. (1994). Meiotic gene conversion tract length distribution within the rosy locus of drosophila melanogaster. *Genetics*, **137**(4), 1019–1026.

Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159**(4), 1805–1817.

Hudson, R. R. (2002). Generating samples under the Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**(2), 337–338.

Hwang, D. G. and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences*, **101**(39), 13994–14001.

International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, **437**(7063), 1299–1320.

Jeffreys, A. J. and May, C. A. (2004). Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genetics*, **36**(2), 151–156.

Jiang, C. and Pugh, B. F. (2009). Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics*, **10**(3), 161–172.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., and Segal, E. (2008). The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, **458**(7236), 362–366.

Kauppi, L., Jeffreys, A. J., and Keeney, S. (2004). Where the crossovers are: recombination distributions in mammals. *Nature Reviews Genetics*, **5**(6), 413–424.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, **13**(3), 235–248.

Kulathinal, R. J., Bennett, S. M., Fitzpatrick, C. L., and Noor, M. A. F. (2008). Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence . *Proceedings of the National Academy of Sciences*, **105**(29), 1005110056.

Langley, C. H., Lazzaro, B. P., Phillips, W., Heikkinen, E., and Braverman, J. M. (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.

Leach, T. J., Mazzeo, M., Chotkowski, H. L., Madigan, J. P., Wotring, M. G., and Glaser, R. L. (2000). Histone H2A.Z is widely but nonrandomly distributed in chromosomes of *Drosophila melanogaster*. *Journal of Biological Chemistry*, **275**(30), 23267–23272.

Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**(4), 2213–2233.

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*, **454**(7203), 479–485.

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., Zanton, S. J., Tomsho, L. P., Qi, J., Glaser, R. L., Schuster, S. C., Gilmour, D. S., Albert, I., and Pugh, B. F. (2008). Nucleosome organization in the Drosophila genome. *Nature*, **453**(7193), 358–362.

McMahill, M. S., Sham, C. W., and Bishop, D. K. (2007). Synthesis-dependent strand annealing in meiosis. *PLoS Biology*, **5**(11), e299.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., and Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, **304**(5670), 581–584.

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

Myers, S., Freeman, C., Auton, A., Donnelly, P., and McVean, G. (2008). A common sequence motif associated with recombination hot spots and genome instability in humans. *Nature Genetics*, **40**(9), 1124–1129.

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, **327**(5967), 876–879.

Nocedal, J. and Wright, S. J. (2000). *Numerical Optimization*. Springer, New York, USA, Second edition.

Padhukasahasram, B., Wall, J. D., Marjoram, P., and Nordborg, M. (2006). Estimating recombination rates from single-nucleotide polymorphisms using summary statistics. *Genetics*, **174**(3), 1517–1528.

Paigen, K. and Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. *Nature Reviews Genetics*, **11**(3), 221–233.

Paigen, K., Szatkiewicz, J. P., Sawyer, K., Leahy, N., Parvanov, E. D., Ng, S. H., Graber, J. H., Broman, K. W., and Petkov, P. M. (2008). The recombinational anatomy of a mouse chromosome. *PLoS Genetics*, **4**(7), e1000119.

Parvanov, E. D., Petkov, P. M., and Paigen, K. (2010). Prdm9 controls activation of mammalian recombination hotspots. *Science*, **327**(5967), 835.

Peckham, H. E., Thurman, R. E., Fu, Y., Stamatoyannopoulos, J. A., Noble, W. S., Struhl, K., and Weng, Z. (2007). Nucleosome positioning signals in genomic DNA. *Genome Research*, **17**(8), 1170–1177.

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, **69**(1), 1–14.

Ptak, S. E., Voelpel, K., and Przeworski, M. (2004). Insights into recombination from patterns of linkage disequilibrium in humans. *Genetics*, **167**(1), 387–397.

Rabiner, L. (1989). A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.

Raphael, C. (2001). Coarse-to-fine dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **23**(12), 1379–1390.

Russell, P. J. (2009). *iGenetics: A Molecular Approach*. Benjamin Cummings, Third edition.

Schones, D., Cui, K., Cuddapah, S., Roh, T., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**(5), 887–898.

Segal, E. and Widom, J. (2009). What controls nucleosome positions? *Trends in Genetics*, **25**(8), 335–343.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, **442**(7104), 772–778.

Singh, N., Aquadro, C., and Clark, A. (2009). Estimation of fine-scale recombination intensity variation in the *white–echinus* interval of *D. melanogaster*. *Journal of Molecular Evolution*, **69**(1), 42–53.

Song, Y. S., Lyngsø R., and Hein, J. (2006). Counting all possible ancestral configurations of sample sequences in population genetics. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **3**(3), 239–251.

Song, Y. S., Ding, Z., Gusfield, D., Langley, C. H., and Wu, Y. (2007). Algorithms to distinguish the role of gene-conversion from single-crossover recombination in the derivation of SNP sequences in populations. *Journal of Computational Biology*, **14**(10), 1273–86.

Stahl, F. W. (1994). The Holliday junction on its thirtieth anniversary. *Genetics*, **138**(2), 241–246.

Stephens, M. (2008). Inference under the coalescent. In D. J. Balding, M. Bishop, and C. Cannings, editors, *Handbook of Statistical Genetics*, volume 1. John Wiley & Sons, Ltd, Chichester, UK, Third edition.

Stumpf, M. P. and Mcvean, G. A. (2003). Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, **4**(12), 959–968.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, **4**(3), e72.

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts & Company Publishers, First edition.

Wall, J. D. (2004a). Close look at gene conversion hot spots. *Nature Genetics*, **36**(2), 114–115.

Wall, J. D. (2004b). Estimating recombination rates using three-site likelihoods. *Genetics*, **167**(3), 1461–1473.

Wall, J. D. and Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, **4**(8), 587–597.

Wasson, T. and Hartemink, A. J. (2009). An ensemble model of competitive multi-factor binding of the genome. *Genome Research*, **19**(11), 2101–2112.

Watterson, G. (1975). On the number of segregation sites. *Theoretical Population Biology*, **7**(2), 256–276.

Wiuf, C. (2000). A coalescence approach to gene conversion. *Theoretical Population Biology*, **57**(4), 357–367.

Wiuf, C. and Hein, J. (2000). The coalescent with gene conversion. *Genetics*, **155**(1), 451–462.

Wu, T. C. and Lichten, M. (1994). Meiosis-induced double-strand break sites determined by yeast chromatin structure. *Science*, **263**(5146), 515–518.

Wu, Z. K., Getun, I. V., and Bois, P. R. J. (2010). Anatomy of mouse recombination hot spots. *Nucleic Acids Research*, **38**(7), 2346–2354.

Yuan, G.-C. and Liu, J. S. (2008). Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Computational Biology*, **4**(1), e13.

Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., and Rando, O. J. (2005). Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, **309**(5734), 626–630.

Zhang, Y., Shin, H., Song, J. S., Lei, Y., and Liu, X. S. (2008). Identifying positioned nucleosomes with epigenetic marks in human from ChIP-Seq. *BMC genomics*, **9**(1), 537.