# Scalable Statistical Methods for Ancestral Inference from Genomic Variation Data

*Andrew Chan*

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-193
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-193.html

December 1, 2014

# Scalable Statistical Methods for Ancestral Inference from Genomic Variation Data

by

Andrew Hans Chan

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

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Professor Satish Rao
Professor Haiyan Huang

Fall 2013

**Scalable Statistical Methods for Ancestral Inference from Genomic Variation Data**

## Abstract

Scalable Statistical Methods for Ancestral Inference from Genomic Variation Data

by

Andrew Hans Chan

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Yun S. Song, Chair

Developments in DNA sequencing technology over the last few years have yielded unprecedented volumes of genetic data. The resulting datasets are indispensable for a variety of purposes, from understanding cancer to answering questions about evolution. Despite the ease with which one can obtain these large quantities of data, the task of extracting meaning from the data remains an open and challenging problem. In this work, we develop statistical methods to infer population genetic parameters from high-throughput sequencing data through the use of coalescent theory, which stochastically models the evolution of DNA from generation to generation. Because closed analytic formulas are unknown for many parameters of interest, computational methods such as Markov Chain Monte Carlo and Sequential Importance Sampling become particularly relevant.

We develop a method using reversible jump MCMC to infer genome-wide variable recombination rates and apply it to data from two *Drosophila melanogaster* populations. Our analysis of the results reveals several interesting findings. A systematic search for hotspot regions reveals only a few occurrences along the genome, far less than that observed in human. We apply a wavelet analysis to quantify the differences between the recombination maps of the two populations, and find that although there is high variability at the fine scales, the recombination maps demonstrate general agreement at the broad scales. The correlation between various genomic features is also assessed using the wavelet analysis, and we find, in contrast to humans, a correlation between recombination and diversity.

In addition, we describe a particle filtering method to sample genealogies from the posterior distribution. Particle filtering is a model estimation technique in the family of sequential importance sampling methods. It provides the ability to perform inference on a continuous state space where the distributions under consideration are complex enough such that exact inference is intractable. The sequentially Markov coalescent, an approximation to the coalescent model where the Markov property is imposed along the sequence, is used to decompose the likelihood of the data into the product of conditional densities and allows inference on otherwise intractably long sequences of genomic data.

# Contents

# List of Figures

# List of Tables

# Acknowledgements

I would like to express my appreciation for those who have given me support during my years at Berkeley. Without their support, I would not be where I am today. Berkeley presented me with an abundance of challenging opportunities, many of which I could not have overcome without the help and support of others.

I would especially like to thank my adviser Yun Song, who has provided me invaluable guidance and support throughout the years. I was very new to the field of population genetics when I started at Berkeley, and he provided many challenging and exciting projects to work on. It was a pleasure to work with him, and I learned a great deal from his advice and mentorship throughout the years. I particularly appreciate the free reign he gave me to explore research topics I found of interest, and his unwavering and tireless support in their pursuit.

I would further like to express my gratitude to the colleagues with whom I have had the pleasure of working during my time at Berkeley, including Anand Bhaskar, Fumei Lam, Jun-ming Yin, Ma'ayan Bresler, Joshua Paul, Sara Sheehan, Jack Kamm, Matthias Steinrücken, Wei-Chun Kao, Atif Rahman, Adam Roberts, Jonathan Terhorst, Jasmine Nirody, and Paul Jenkins. I want to thank them for many stimulating and thoughtful discussions on a variety of topics relating to population genetics and computational biology.

In particular, I would like to acknowledge Wei-Chun Kao and Paul Jenkins. Wei-Chun and I worked on high-throughput error correction methods, and spent many long nights developing and coding our methods. We had many productive discussions bouncing ideas back and forth and learning from each other, and I am grateful that we had the opportunity to work together.

My primary research in population genetics was in collaboration with Paul Jenkins, who was not only a colleague but also a mentor of sorts and taught me virtually everything I know about population genetics. He had an almost magical way of explaining difficult and complex concepts intuitively, and, as importantly, was extremely patient with any questions I had, no matter how mundane or poorly thought through they were. We had many interesting discussions together that would usually conclude in my realizing a subtlety I never noticed before or a lack of understanding of a concept I thought I had mastered. Without his tutorship, I would not have the sense for population genetics I have today.

Finally, I would like to thank the members of my qualifying and dissertation committees: Satish Rao, Haiyan Huang, and Lior Pachter. The Center for Theoretical and Evolutionary Genomics and their many seminars were an effective venue for hearing about others' research and gave me an opportunity to interact with researchers outside my group. The Designated Emphasis in Computational and Genomic Biology, and their well-run and thoroughly organized retreats, was also an invaluable resource for collaboration and productive discussions with researchers in the field of computational biology at large.

# Chapter 1

# Introduction

The past several years have experienced a tremendous growth in the availability of genomic data. High-throughput sequencing technologies developed by companies such as Illumina, Life Technologies, Roche, and many others have provided data on a greater scale than ever seen before. But collecting genomic data is not sufficient; the data must also be analyzed to answer biological questions of interest. Such questions include those about mutation rate, recombination rate, ancient population structure, natural selection, and many more. Typically, the experiment involves collecting genetic data from a relatively sample of individuals from a much larger population and examining the patterns of variation to infer population-genetic parameters of interest.

In order to answer these questions rigorously and quantitatively, we must first construct an appropriate stochastic process to serve as a lens through which we can analyze the data. One such model is called the Wright-Fisher diffusion, which naturally describes the evolution of genetic information from generation to generation. The Wright-Fisher diffusion is capable of incorporating many aspects of interest to population geneticists, such as mutation, recombination, demography and so on. In its most basic form, it describes a randomly mating (i.e., *panmictic*) population that allows for a rigorous analysis of the probabilities of sampling a given observation from the population.

Kingman's coalescent [46] is the dual process to the Wright-Fisher diffusion, and in many settings is more convenient or even the only feasible approach to understanding or computing quantities of interest. Whereas the Wright-Fisher diffusion describes the evolution of the population *forward* in time, the coalescent explains the history of the observed sample from the population *backward* in time. Because many genomic studies are concerned primarily with the relationships among a sample of individuals, rather than the population as a whole, the coalescent is often a more convenient framework under which to understand the data.

Despite the theoretical advances made over the last several decades in population genetics, efficient statistical inference under the coalescent with recombination remains a challenging open problem. Closed-form analytic formulas for sampling probabilities are known only for the simplest of cases under the coalescent. For more complex models, heuristics and approximations have been employed to gain a handle on quantifying the data. Many such methods are computationally expensive methods, involving some form of sampling such as Markov Chain Monte Carlo or importance sampling. However, despite the computational expense, often paired with potentially extreme simplifications to the underlying model, such methods have proved highly useful and effective at answering many population genetic questions.

In this thesis, we describe two methods for statistical inference under the coalescent. The first employs the method of composite likelihood to estimate variable recombination maps. The second uses particle filtering to approximate the posterior distribution on genealogies along the genome. We applied the composite likelihood method to estimate variable recombination rates in two populations of *Drosophila melanogaster* and analyzed the resulting recombination maps toward answering several biologically relevant questions relating to recombination, such as occurrence of recombination hotspots and the relationship of recombination to a variety of genomic features.

The structure of the thesis is as follows. The remainder of the chapter will provide an introduction to population genetics and the coalescent, Chapter 2 describes work on estimating fine-scale recombination rates, Chapter 3 describes work on particle filtering techniques for population-genetic inference, and Chapter 4 concludes with a discussion.

## 1.1 Wright-Fisher Model and the Coalescent

The Wright-Fisher diffusion is derived from the Wright-Fisher model, a discrete time process on a finite population of $2N$ individuals. The population evolves by generation, and in every generation, a new population is constructed from the previous generation. Because the population is assumed to be random-mating, every individual in the previous generation is equally like to be the parent of a given individual in the current generation. An individual inherits the genetic properties of his parent, with a small probability $\mu$, called the *mutation rate*, which introduces a change to the genetic material. This process continues in non-overlapping generations until equilibrium is achieved, at which point a relatively small sample (compared to the population size) of $n$ individuals is taken from the population and sequenced. This then serves as the genetic data on which inference is performed. Note that there are many variations and extensions that allow for more biologically realistic and sophisticated models of population evolution, some of which will be described later. When the number of individuals is taken to infinity and the time of each generation is scaled to 0, we obtain the Wright-Fisher diffusion.

The relationship between the coalescent and the Wright-Fisher model is in the ancestry of the sample of $n$ individuals. Tracing the ancestry backward in time of the individuals in the sample, we find a *genealogy* relating the individuals. Given any two individuals, they will eventually find a most recent common ancestor (MRCA), and this information can be encoded in a tree with branch lengths in units of coalescent time. This tree also contains information on mutation events, where the genetic material of lineage differs from its ancestor.

It can be shown that the rate of *coalescent* events between any pair of lineages in the sample is $\binom{k}{2}$, where $k$ is the number of lineages in the ancestry. In other words, every pair of lineages coalesces at a rate of 1. Mutation events occur according to a Poisson process with rate $\theta/2$, where $\theta = 2N\mu$, the population-scaled mutation rate. Here, $N$ is the *effective* population size rather than the *census* population size. (Note that the limit as $N \to \infty$ of the population size in the Wright-Fisher model, stated above as $2N$, is taken to obtain the continuous time processes of the Wright-Fisher diffusion and the coalescent.) In practice, the effective population size must be inferred through other means besides coalescent-based methods, but with the effective population size in hand, a coalescent time unit can be interpreted as $2N$ generations.

Coalescent and mutation events together create the following process that proceeds from the present to the past: initially there are $n$ lineages that extend backward in time. Uniformly chosen pairs of lineages *coalesce* at rate $\binom{k}{2}$, where $k$ is the number of remaining lineages at any given time, and every lineage *mutates* at rate $\theta/2$. Once the coalescent process reaches the state of one lineage, an ancestral type is chosen for that lineage and its genetic material is propagated down the tree, incorporating the effects of coalescence and mutation.

## 1.2 Recombination

Although the model just described is tractable and convenient to work with, it lacks several aspects of biological realism, of which recombination may be the greatest. Recombination occurs during the process of meiosis in diploid species and results in gametes that are mosaics of their parental homologous chromosomes. A recombination *breakpoint* occurs along the forming gamete, where the genetic material prior to the breakpoint is inherited from one chromosome and that after the breakpoint inherits from the homologous chromosome.

Where mutation events occur with probability $\mu$ in the Wright-Fisher model, recombination events occur with probability $r$, with the individual in the new generation taking on two parents from the previous generation. The breakpoint occurs uniformly along the new individual's genome, either continuously or along discrete points depending on the model, and the genetic information before the breakpoint comes from one parent and after it the other parent.

Recombination changes the coalescent process from a pure death process to a birth-death process. Whereas without recombination, the coalescent process involves only coalescent events that reduce the number of remaining lineages, recombination events provide a way for process to gain lineages. The rates of coalescent and mutation events remain the same as before, but every lineage *recombines* at a rate of $\rho/2$, where $\rho = 2Nr$. $\rho$ is the population-scaled recombination rate and determines the rate in coalescent time units at which recombination events happen. These two lineages each carry only a portion of ancestral material depending on which sides of the breakpoint they represent. One lineage represents the ancestral material before (or to the left) of the breakpoint and the other represents the ancestral material after (or to the right) of the breakpoint.

Besides the significance of recombination in the context of evolution, recombination has important implications for inference procedures based on patterns of genetic variation. This stems from the fact that individuals are not merely related to one another in a tree-based genealogy but in a much more complex genealogy represented by a directed acyclic graph known as an ancestral recombination graph. This poses both challenges and opportunities in inference procedures. On the one hand, the underlying model becomes increasingly complicated. On the other, the richness of the relationships among individuals produces more complex and informative data for inference methods.

A key result of the thesis is a method to infer *variable* recombination rates. Rather than assume the recombination breakpoint occurs uniformly along the genome, we allow for the possibility that the breakpoint may occur with higher probability in some regions of the genome than others. We model this heterogeneity using an inhomogeneous Poisson process, and the inference is performed in a Bayesian framework using a composite likelihood approximation.

## 1.3 Composite Likelihood

Composite likelihood methods are motivated by the computational infeasibility of standard likelihood methods in high-dimensional inference. The basic concept of composite likelihood methods is to project high-dimensional likelihood functions to more computationally tractable low-dimensional likelihood functions. If the projection is performed appropriately, the composite likelihood approximation can provide a more easily implementable approach to finding parameters of interest.

A common technique for simplifying the full likelihood is to assume subsets of the components of the likelihood are independent when they are in fact not. The likelihood for each of the subsets is computed, and the product over all such likelihoods is taken to serve as the pseudo-likelihood. In many cases, the marginal likelihoods for these subsets are substantially easier to compute than the full likelihood, and this reduction in complexity lends composite likelihood methods their effectiveness.

In our setting, we use pairwise composite likelihood to decompose the full likelihood into more manageable components. By computing the likelihoods for pairs of sites (or pairs of loci) in the genome and taking the product, we obtain a pseudo-likelihood for use in a Bayesian framework. Computing the likelihood for two-locus data is much more tractable than for data with more than two loci because only a limited dependency structure needs to be accounted for, and the computational performance becomes all the more significant when the computation of such a likelihood is a subroutine of a time-intensive inference scheme such as MCMC.

## 1.4 Reversible Jump Markov Chain Monte Carlo

Standard Markov Chain Monte Carlo (MCMC) schemes are restricted to problems where the joint distribution of the parameters have a density with respect to a fixed standard underlying measure. In inference settings where the dimensionality of the parameter vector is not fixed, a more sophisticated form of MCMC called *reversible jump* MCMC (rjMCMC) [29] must be used. rjMCMC methods have been applied successfully to multiple change point analysis, where the dimensionality is not fixed due to the varying number of change points representing the parameter vector. In particular, rjMCMC is well suited for inference on Poisson processes where the rate is assumed to be piecewise constant but changes an unknown number of times. The state space is then the set of step functions, where the number of change points is unbounded. Because the dimensionality of the parameter vector can be unbounded, rjMCMC is often described as a *non-parametric* inference method.

Reversible jump MCMC is as an extension to the standard Metropolis-Hastings method. The Metropolis-Hastings method generates samples from the target distribution by proposing locally modified samples sequentially. Given the current state of the Markov Chain, local modifications are proposed, and an acceptance ratio is computed for the proposed state. On acceptance, the Markov Chain moves to the new state, otherwise it remains at the current state. Many iterations of this procedure produce a large set of (dependent) samples from the target distribution.

The implicit assumption in Metropolis-Hastings is that the proposed state has the same dimensionality as the current state, such that the ratio of densities is a valid quantity to consider. If the proposed state is composed of a different number of dimensions, then this assumption no longer holds and the method breaks down. Reversible jump MCMC seeks to address this issue by creating bijections between subspaces of the state space having different dimensionality. From a lower dimensional subspace to a higher dimensional subspace, auxiliary random variables are sampled to *match* the dimensionality between the lower dimensional subspace and the higher dimensional subspace, and a bijection is used to map the lower dimensional state to the higher dimensional state. Likewise when mapping higher dimensional states to lower dimensional states: auxiliary random variables are used to match the dimensionality, and a bijection bridges the higher dimensional states from one subspace to the lower dimensional states in the other subspace. These bijections are then used to make the Markov chain reversible across these subspaces. In practice, for most situations this results in an addition to the Metropolis-Hasting acceptance ratio: the Jacobian of the bijection from the two subspaces is multiplied into the acceptance ratio. In the setting of estimating variable recombination rates, the prior on the recombination map is a step function with an unknown number of change points, and hence rjMCMC is a natural choice for the inference method.

## 1.5 Particle Filtering

Particle filtering is a statistical inference approach for general state space hidden Markov models. Traditionally, it has been used for time series data, but in the setting of genomics, position along the genome rather than time is used. Particle filtering techniques for sampling from the posterior on genealogies is particularly effective when used in conjunction with the sequentially Markov coalescent, described in Section 1.6.

### 1.5.1 Importance Sampling

Particle filtering techniques are special instances of Sequential Importance Sampling, a form of importance sampling where samples are constructed by sampling from a sequence of proposal distributions. Importance sampling is a technique for approximating otherwise intractable distributions. The process involves biasing samples toward regions of high density and weighting them according to the ratio of their likelihood to their proposal density.

Suppose we wish to compute $\mathbb{E}(f(X))$ and a closed analytic form is not available. To approximate the expectation by direct Monte Carlo, we would sample $X^{(i)}$ from $X$ $M$ times and the following would be the approximation to $\mathbb{E}(f(X))$:

$$\mathbb{E}(f(X)) \approx \frac{1}{M}\sum_{i=1}^{M} f(X^{(i)}).$$

However, in cases where $f(x)$ is close to 0 for many values of $x \in \mathcal{X}$, where $\mathcal{X}$ is the support of $X$, we might instead desire to sample primarily those values for which $f(x)$ makes a significant contribution to the expectation. Hence a proposal distribution focusing on high density regions of the space could be more effective than sampling from $X$ directly. Sampling from a distribution different from the one naturally associated with $X$ results in biased samples that must then be corrected by the *importance weight*. The importance weight can be derived by considering the following set of equations:

$$\begin{aligned}\mathbb{E}(f(X)) &= \int_{\mathcal{X}} f(x)p(x)dx \\ &= \int_{\mathcal{X}} f(x)\frac{p(x)}{q(x)}q(x)dx \\ &\approx \sum_{i=1}^{M} f(X^{(i)})\frac{p(X^{(i)})}{q(X^{(i)})},\end{aligned}$$

where $X^{(i)}$ is sampled from $q(X)$, the proposal distribution, rather than directly from $p(x)$, the density of $X$. Note that for the above equations to hold, we must have $q(x) > 0$ whenever $p(x) > 0$.

### 1.5.2 Filtering

A similar argument applies when approximating distributions using importance sampling. Suppose we wish to approximate a density $p(X)$. The direct Monte Carlos approach would sample from $p(X)$, and each sample would be a discrete atom in the approximation with probability $1/M$, where $M$ is the number of samples. A better approximation would be to bias the samples toward high density regions of the distribution. An importance sampling approach samples from $q(X)$, and assigns a weight of $p(X)/q(X)$ to each atom. The associated probability of each atom is its normalized weight, and this often provides a better approximation to the target distribution. In particle filtering, these atoms are called particles, and they provide a discrete approximation to a continuous distribution.

*Filtering* refers to computing a given quantity in an online manner, that is, sequentially updating the estimate on the parameters conditioned on the available data seen so far. If the data is available all at once, then a technique called *smoothing* can be used to incorporate all the data for the estimation at every genomic site (or every time point). These methods implicitly rely on a Markov property in the underlying sequence of distributions, and this assumption distinguishes particle filtering from the more general sequential importance sampling techniques. The sequentially Markov coalescent allows particle filtering approaches to be applied in the context of the coalescent.

## 1.6 Sequentially Markov Coalescent

The sequentially Markov coalescent (SMC) [58] is an approximation to the full coalescent model that imposes a Markov constraint on the dependency of genealogies along the sequence. Every site along the genome has an associated marginal genealogy, represented by a tree. The collection of marginal genealogies and recombination events is represented by an ancestral recombination graph (ARG), which fully captures the relationships among the individuals in the sample along their genomes.

The SMC seeks to simplify the dependency structure of the marginal trees. In the SMC, the joint distribution on marginal trees can be decomposed into a product of conditional densities such that the conditional density for each marginal tree depends only on the marginal tree immediately before it. Let $p(T_1)$ be the marginal probability for the tree at the first site, and $p(T_i \mid T_{i-1})$ be the conditional probability for the tree at site $i$ given the tree at site $i-1$ (where $T_i$ represents the tree at site $i$). The joint distribution on trees along a genome of $L$ sites is

$$p(T_1, \ldots, T_L) = p(T_1) \prod_{i=2}^{L} p(T_i \mid T_{i-1}).$$

The Markov nature of the sequentially Markov coalescent is generally more amenable to inference procedures than the full coalescent. The SMC has been shown to preserve many of the salient features of the full coalescent, and hence serves as a good model for inference despite losing some of the dependency structure. It is difficult to perform inference under the full coalescent for a number of reasons, including the enormous state space of ARGs and the weakly informative nature of data to infer the ARG. This is due to the fact that many different ARGs result in the same observable data, and hence the extent to which one can infer the ARG is limited by the degeneracy of the mapping from ARGs to data. The pattern observed in the data is only dependent on the marginal genealogies along the genome and not directly on the associated ARG tying the marginal genealogies together. In other words, conditioned on the marginal genealogies, the likelihood of the data is independent of the additional information contained in the ARG. The SMC seeks to mitigate these problems by providing a model that retains many of the features necessary for the inference of desired parameters while at the same time discarding features with possibly less relevant effects. Furthermore, the Markov nature of the SMC is particularly convenient because inference in the hidden Markov setting is well studied in the literature.

# Chapter 2

# Fine-scale Recombination Rate Variation

In this chapter, we describe the method we developed, called `LDhelmet`, for estimating fine-scale recombination maps of *Drosophila melanogaster* from population genomic data. The task of estimating recombination maps in *Drosophila* is challenging, in part because of the high background recombination rate. We first provide a description of the method and then provide an extensive simulation study to demonstrate that it allows more accurate inference and exhibits greater robustness to the effects of natural selection and noise compared to a previous method for studying fine-scale recombination rate variation in the human genome called `LDhat` [59, 62].

As an application of our method, a genome-wide analysis of genetic variation data is performed for two *Drosophila melanogaster* populations, one from North America (Raleigh, USA) and the other from Africa (Gikongoro, Rwanda). It is shown that fine-scale recombination rate variation is widespread throughout the *D. melanogaster* genome, across all chromosomes and in both populations. At the fine-scale, a conservative, systematic search for evidence of recombination hotspots suggests the existence of a handful of putative hotspots each with at least a tenfold increase in intensity over the background rate.

We perform a wavelet analysis, described in Section 2.21 to compare the estimated recombination maps in the two populations and to quantify the extent to which recombination rates are conserved. In general, similarity is observed at very broad scales, but substantial differences are seen at fine scales. The average recombination rate of the X chromosome appears to be higher than that of the autosomes in both populations, and this pattern is much more pronounced in the African population than the North American population. The correlation between various genomic features—including recombination rates, diversity, divergence, GC content, gene content, and sequence quality—is examined using the wavelet analysis, and it is shown that the most notable difference between *D. melanogaster* and humans is in the correlation between recombination and diversity.

## 2.1 Motivation

Recombination is a biological process of fundamental importance in population genetic inference. The crossing-over of homologous chromosomes during meiosis results in the exchange of genetic material and the formation of new haplotypes. Accurate estimates of the recombination rate in different regions of the genome help us to understand the molecular and evolutionary mechanisms of recombination, as well as a host of other important phenomena. For example, recombination rate estimates are needed in assessing the impacts of natural selection [34, 69], admixture [67], and disease associations [77].

Recombination rates have been observed to exhibit a number of interesting heterogeneities: they are known to vary in magnitude and distribution between species (e.g., [64, 63, 5]), between populations within species [76, 47], and between individuals within populations [7, 10, 19, 47]. There is also substantial variation in different regions of the genome at different scales. At the broad-scale, for example, recombination rates in humans are known to be correlated negatively with the distance from telomeres [62], while at the fine-scale, recombination events cluster in narrow *hotspots* of $\sim 2$ kb width [59, 62, 77]. In humans, hotspots are typically defined as those with statistical support in favor of at least a five-fold increase of the recombination rate [62] over the background or surrounding region, and many hotspots suggest a ten- or even hundred-fold increase. Such hotspots exhibit a powerful influence on the recombination landscape; 70–80% of recombination events in humans occur in 10% of the total sequence [76]. Extensive fine-scale variation and recombination hotspots have also been found in other species, including chimpanzees [5], *Arabidopsis thaliana* [22] and yeast [80].

### 2.1.1 *Drosophila melanogaster*

The picture in *Drosophila* is however less clear. Broad-scale maps of recombination have been constructed for *D. melanogaster* by fitting a third-order polynomial to each chromosome arm [27, 54]. These give an overview of the distribution of recombination along each arm, quantifying for example earlier observations of declining recombination rates with proximity to the telomeres and centromeres. Variation on finer scales has been inferred by studies of linkage disequilibrium (LD) and by breeding experiments. Rapid and consistent decay in LD [50] leads to an absence of long haplotype blocks. There is scant evidence for hotspots either at the intensity or prevalence of those found in humans. Experimental studies of variation have produced local, fine-scale maps in *D. melanogaster* [71], *D. persimilis* [74], and *D. pseudoobscura* [18, 48], providing a resolution typically on the order of 100 kb in the regions analyzed. These experimental results suggest that regions of fine-scale variation—including some mild "hotspots" [18]—do exist in several *Drosophila* species. For example, Singh *et al.*[71] study a 1.2 Mb region of the X chromosome in *D. melanogaster*, and find 3.5-fold variation in this region, though no hotspots by the criterion mentioned above. These experimental approaches are cumbersome to recapitulate, however.

A number of crucial questions concerning *Drosophila* therefore remain unanswered. It is not known to what extent this variation is further localized to finer scales, or how common such variation is across the genome. Further, intra-specific differences in recombination rate have not been characterized. However, the advent of ambitious projects (e.g., see the Drosophila Genetic Reference Panel [54] and the Drosophila Population Genomics Project [49]) sequencing tens of *D. melanogaster* genomes each from different global populations raises the exciting prospect of addressing these and other questions. The patterns of LD in a random sample of contemporary genome sequences taken from a population contain a great deal of information regarding historical recombination events, and from these we can infer recombination rates across the genome. A number of sophisticated and computationally-intensive statistical approaches have been developed for inferring recombination rates from such data [4, 52, 59, 82] and for testing for the presence of recombination hotspots [26, 25], and are ostensibly suitable for this task. In particular, `LDhat` [60, 59, 4] is a useful software package which scales well to large datasets, and it has therefore been applied to estimating recombination rates in humans [59, 62, 77, 76], chimpanzees [5], dogs [6], yeast [80], and microbes [44], among others.

### 2.1.2 Challenges

Estimating fine-scale recombination rates from newly published *D. melanogaster* genomes is, however, challenging for several reasons: First, these data exhibit a much higher density of single nucleotide polymorphisms (SNPs) than those of other species and of earlier technologies. For example, the African data exhibits a mean SNP rate of about 1 SNP per 38 bp for a sample of size 22, far higher than those of other recent sequencing projects (e.g., [76]). This promises an unprecedented opportunity to localize recombination rate variation to very fine scales, but making full use of these data raises further challenges in computational and statistical efficiency. Second, data generated from short-read sequencing technologies give rise to numerous missing alleles. It would be highly advantageous to be able to make use of sites in which some alleles are missing without the exponential increase in `LDhat`'s running time that this entails. Third, the background recombination parameter in *D. melanogaster* is known to be an order of magnitude higher than in humans (the species for which `LDhat`'s prior distributions and parameters are typically calibrated) and it is not clear how this will affect the accuracy of subsequent rate estimates. Fourth, there is a growing consensus that a considerable fraction of the genome of some *Drosophila* species is influenced by adaptive substitutions [70, 69]. Recurrent selective sweeps combined with genetic hitchhiking affect patterns of variation across many kilobases of sequence and have the potential to invalidate inferences of recombination, even leading to the possibility of spurious signals of recombination hotspots [68, 73]. By contrast, the footprints of positive selection in recent human evolution are less widespread [34]. The model underlying `LDhat` assumes a neutrally evolving population of constant size. While `LDhat` is known to be robust to mis-specification of the demographic model [59], its susceptibility to the effects of selection is less clear cut.

### 2.1.3 Approach

We develop a new method, called `LDhelmet`, which addresses the above critical issues. While it employs a reversible-jump Markov Chain Monte Carlo (rjMCMC) mechanism similar to that of `LDhat`, our method has a number of modifications that render key advantages. Briefly, by utilizing recent theoretical advances in asymptotic sampling distributions [40, 39, 41, 12, 42, 11], we introduce several analytic improvements to the computation of likelihoods in the underlying population genetic model, which reduce Monte Carlo errors and simultaneously provide likelihoods for all relevant samples with an arbitrary number of missing alleles. Our refinements further improve accuracy by allowing us to make full use of a tetra-allelic mutation model in which realistic mutation patterns between the four nucleotides A, C, G, T can be taken into account. Additionally, we utilize information from the available genomes of outgroup species by using them to infer a distribution on the ancestral allele at each polymorphic site in *D. melanogaster*. Taken together, our method enables us to compute fine-scale, genome-wide recombination rates with considerably improved accuracy and efficiency. `LDhelmet` generally produces recombination maps that are less noisy than that of `LDhat`'s. In particular, while `LDhat` can infer spurious hotspots under certain types of selection, we demonstrate that our approach is much more robust.

We apply our method to data taken from two *D. melanogaster* populations, one from North America and the other from Africa, and estimate fine-scale recombination maps for each population. Then, through a wavelet analysis, we capture levels of variability and correlation of the two recombination maps, and provide a quantitative view of genome-wide inter-population comparison of recombination rates in *D. melanogaster*. We also employ the wavelet analysis to examine the correlation between various genomic features, including recombination rates, diversity, divergence, GC content, gene content, and sequence quality. At the fine-scale, we perform a conservative, systematic search for evidence of the existence of recombination hotspots and find a handful of putative hotspots each with at least a tenfold increase in intensity over the background rate. Also, we compare our recombination rate estimates with existing experimental genetic maps.

## 2.2 Outline of Method

Given a sample of chromosomes from a population, `LDhat` estimates the recombination map $\rho$ within a Bayesian setting, placing a prior on the map. To avoid overfitting, $\rho$ is assumed to be a step function (i.e., a piecewise constant function). The prior is a distribution on the number of times $\rho$ changes value, the locations of such changes, and the value of each piecewise constant segment. We employ reversible-jump MCMC (rjMCMC) [29] to sample from a posterior distribution over a sample space of step functions where different parts of the space have different numbers of parameters.

Denote the likelihood of $\rho$ and $\theta$ by $\mathbb{P}(D \mid \rho, \theta)$, where $D$ represents a set of phased haplotypes. Rather than compute the full likelihood, which is in general intractable except for a very small sample, we compute an approximation known as the *pairwise composite likelihood* [38, 60]. For every pair of SNPs in a short region, the pairwise likelihood is computed under the coalescent with recombination, and the product over all such pairwise likelihoods serves as an approximation to the full likelihood. This approach scales well to large datasets, and has been demonstrated through simulation studies to provide a reasonable approximation to the full likelihood [60]. The two-locus likelihoods are precomputed and stored in a lookup table for computational efficiency. There is one likelihood table for every choice of mutation parameter $\theta$, and likelihoods are precomputed over a grid of the recombination parameter $\rho$.

## 2.3 Two-locus Recursion Relation

We generate two-locus likelihood lookup tables by solving solving recursion relations [28] (see also [23, 40, 39, 41, 12]). These recursion relations necessitate the solution of large systems of equations in the possible observed sample configurations. However, the one-mutation-per-site assumption leads to gains in efficiency that make such systems soluble.

### 2.3.1 One-locus Model

To illustrate, consider first a random sample drawn from a single locus. We use the notation $q(\boldsymbol{m};\theta)$ to denote the probability that a sample of $m$ alleles taken at random from the population in some fixed order leads to the one-locus configuration $\boldsymbol{m} = (m_j)_{j=1,\dots,K}$, where $m_j$ is the number of samples with allele $j$; if we are modeling, say, the evolution of DNA nucleotides, then $K = 4$ and $j \in \{A, C, G, T\}$. (It is implicit that this probability is also a function of the mutation transition matrix $\boldsymbol{P}$ at this locus.) It is well known (e.g., [30]) that $q(\boldsymbol{m};\theta)$ satisfies

$$m(m-1+\theta)q(\boldsymbol{m};\theta) = \sum_{i=1}^{K} m_i(m_i-1)q(\boldsymbol{m}-\boldsymbol{e}_i;\theta) + \theta \sum_{i,j=1}^{K} m_j P_{ij} q(\boldsymbol{m}-\boldsymbol{e}_j+\boldsymbol{e}_i;\theta), \quad (2.1)$$

for which a closed-form solution is not known in general. Here, $\boldsymbol{e}_i$ denotes a unit vector with $i$th entry 1 and the rest zero. In a later section, we describe a method for using outgroup data to infer which of the alleles in our samples is ancestral. When the identity of the ancestral allele (i.e., the allele of the most recent common ancestor of the sample) is presumed known, say type $a$, the appropriate boundary condition for use with (2.1) is

$$q(\boldsymbol{e}_j;\theta) = \begin{cases} 1, & \text{if } j = a, \\ 0, & \text{otherwise.} \end{cases}$$

As an alternative to working with (2.1), we can seek a solution for the joint probability of obtaining the configuration $\boldsymbol{m}$ with the event that it arose as the result of precisely $s$

mutation events in the history of the sample, a probability we denote by $q(\boldsymbol{m}, s; \theta)$. Then we have [30]:

$$m(m-1+\theta)q(\boldsymbol{m}, s; \theta) = \sum_{i=1}^{K} m_i(m_i-1)q(\boldsymbol{m}-\boldsymbol{e}_i, s; \theta) + \theta \sum_{i,j=1}^{K} m_j P_{ij} q(\boldsymbol{m}-\boldsymbol{e}_j+\boldsymbol{e}_i, s-1; \theta), \tag{2.2}$$

with

$$q(\boldsymbol{e}_j, s; \theta) = \begin{cases} 1, & \text{if } j = a \text{ and } s = 0, \\ 0, & \text{otherwise.} \end{cases}$$

The advantage of the one-mutation-per-site assumption is then apparent: $q(\boldsymbol{m}, 1; \theta)$ is known in closed-form [42, 11]:

$$q(\boldsymbol{m}, 1; \theta) = P_{ad} \frac{\theta m_a! m_d!}{m(\theta+1)(\theta+2)\cdots(\theta+m-1)} \sum_{l=1}^{m_a} \binom{m_a-1}{l-1} \binom{m-1}{l}^{-1} \frac{1}{\theta+l}, \tag{2.3}$$

where the only nonzero entries of $\boldsymbol{m}$ are $m_a$ and $m_d$, corresponding to a sample comprising $m_a$ copies of the ancestral allele type $a$ and $m_d$ copies of a derived allele type $d$. Hence, in this case we entirely circumvent the need for a numerical solution to a large system of linear equations. Provided the mutation rate per site is sufficiently small, the error $|q(\boldsymbol{m}; \theta) - q(\boldsymbol{m}, 1; \theta)|$ should be negligible.

We can make similar gains in a two-locus model by reducing a large system of equations to a much smaller system, albeit one that still requires a numerical solution. The idea is similar to that described above, though notation is more complicated: the precise form of the system is provided in Section 2.3.2. The largest sample size we work with is $n = 37$. This leads to a very large system of equations that must be solved: Accounting for symmetries, the total number of complete configurations of size $n = 37$ is approximately 1,300. When we count all configurations encountered in the RAL data—including those with missing alleles—this number rises to $27 \times 10^6$. In the two locus case, the quantity of interest is $q(\boldsymbol{n}, 1, 1; \theta, \rho)$, the probability of obtaining the two-locus configuration $\boldsymbol{n}$ together with the events that there was precisely one mutation event at each of the two loci. Here, $\theta$ denotes the mutation rate and $\rho$ denotes the recombination rate between the two loci. Provided we work with the reduced system of equations for $q(\boldsymbol{n}, 1, 1; \theta, \rho)$ as outlined above, it becomes feasible to solve the system for every sample of size $n = 37$, and thus to generate *exactly* solved lookup tables for later use. Table 2.1 shows the running time of this recursion-based likelihood computation as a function of sample size $n$.

### 2.3.2 Two-locus Model

Suppose we sample $n$ haplotypes, observing their alleles at each of *two* loci and obtaining configuration $\boldsymbol{n} = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. Here $\boldsymbol{c} = (c_{ij})$ is a matrix of the counts of haplotypes for which both alleles were observed; $c_{ij}$ is the number of haplotypes with allele $i$ at the first locus and allele $j$ at the second locus. We also allow for the possibility that a haplotype had data missing at one locus: $\boldsymbol{a} = (a_i)_{i=1\ldots,K}$ is the vector of counts of haplotypes with allele $i$ observed at the first locus and missing data at the second locus, and $\boldsymbol{b} = (b_j)_{j=1,\ldots,L}$ is the vector of counts of haplotypes with allele $j$ observed at the second locus and missing data at the first locus.

Further, let:

$$a = \sum_{i=1}^{K} a_i, \quad c_{i\cdot} = \sum_{j=1}^{L} c_{ij}, \quad c = \sum_{i=1}^{K}\sum_{j=1}^{L} c_{ij},$$
$$b = \sum_{j=1}^{L} b_j, \quad c_{\cdot j} = \sum_{i=1}^{K} c_{ij}, \quad n = a + b + c.$$

The probability that, when we sample $n$ haplotypes in some fixed order, we obtain a set consistent with configuration $\boldsymbol{n}$, is denoted by $q(\boldsymbol{n};\theta_A,\theta_B,\rho)$. This probability is a function of $\theta_A$, $\theta_B$, and $\rho$: the mutation rates at the two loci, and the recombination rate between them. The respective mutation transition matrices at the two loci, which we denote $\boldsymbol{P}^A$ and $\boldsymbol{P}^B$, are fixed. A system of equations for $q(\boldsymbol{n};\theta_A,\theta_B,\rho)$ is given in [40]. We denote by $q(\boldsymbol{n}, s_1, s_2;\theta_A,\theta_B,\rho)$ the joint probability of obtaining $\boldsymbol{n}$ with the events that there were precisely $s_1$ mutations in the history of the sample at the first locus and $s_2$ mutations in the history of the sample at the second locus. The corresponding system of equations for $q(\boldsymbol{n}, s_1, s_2;\theta_A,\theta_B,\rho)$ is:

$$\begin{aligned}
&[n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c]q((\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}), s_1, s_2;\theta_A,\theta_B,\rho) = \\
&\sum_{i=1}^{K} a_i(a_i - 1 + 2c_{i\cdot})q((\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b},\boldsymbol{c}), s_1, s_2;\theta_A,\theta_B,\rho) \\
&+ \sum_{j=1}^{L} b_j(b_j - 1 + 2c_{\cdot j})q((\boldsymbol{a},\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{c}), s_1, s_2;\theta_A,\theta_B,\rho) \\
&+ \sum_{i=1}^{K}\sum_{j=1}^{L}\Bigg[c_{ij}(c_{ij}-1)q((\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}-\boldsymbol{e}_{ij}), s_1, s_2;\theta_A,\theta_B,\rho) \\
&\qquad\qquad + 2a_ib_jq((\boldsymbol{a}-\boldsymbol{e}_i,\boldsymbol{b}-\boldsymbol{e}_j,\boldsymbol{c}+\boldsymbol{e}_{ij}), s_1, s_2;\theta_A,\theta_B,\rho)\Bigg] \\
&+ \theta_A\sum_{i=1}^{K}\Bigg[\sum_{j=1}^{L} c_{ij}\sum_{t=1}^{K} P^A_{ti}q((\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}-\boldsymbol{e}_{ij}+\boldsymbol{e}_{tj}), s_1-1, s_2;\theta_A,\theta_B,\rho) \\
&\qquad\qquad + a_i\sum_{t=1}^{K} P^A_{ti}q((\boldsymbol{a}-\boldsymbol{e}_i+\boldsymbol{e}_t,\boldsymbol{b},\boldsymbol{c}), s_1-1, s_2;\theta_A,\theta_B,\rho)\Bigg] \\
&+ \theta_B\sum_{j=1}^{L}\Bigg[\sum_{i=1}^{K} c_{ij}\sum_{t=1}^{L} P^B_{tj}q((\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}-\boldsymbol{e}_{ij}+\boldsymbol{e}_{it}), s_1, s_2-1;\theta_A,\theta_B,\rho)
\end{aligned}$$

$$+ b_j \sum_{t=1}^{L} P^B_{tj} q((\boldsymbol{a}, \boldsymbol{b} - \boldsymbol{e}_j + \boldsymbol{e}_t, \boldsymbol{c}), s_1, s_2 - 1; \theta_A, \theta_B, \rho)\Bigg]$$

$$+ \rho \sum_{i=1}^{K} \sum_{j=1}^{L} c_{ij} q((\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{c} - \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho), \tag{2.4}$$

where $\boldsymbol{e}_{ij}$ is a unit matrix whose $(i, j)$th entry is one and the rest are zero. As before, we suppose that we know the identity of the ancestral haplotype, say $(\lambda_A, \lambda_B)$. Then we replace the relevant instances of (2.4) with the following:

$$q((\boldsymbol{0}, \boldsymbol{b}, \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} q((\boldsymbol{0}, \boldsymbol{b} + \boldsymbol{e}_j, \boldsymbol{0}), 0, s_2; \theta_A, \theta_B, \rho) & \text{if } i = \lambda_A \text{ and } s_1 = 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$q((\boldsymbol{a}, \boldsymbol{0}, \boldsymbol{e}_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} q((\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}), s_1, 0; \theta_A, \theta_B, \rho) & \text{if } j = \lambda_B \text{ and } s_2 = 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$q((\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} 1 & \text{if } i = \lambda_A \text{ and } s_1 = s_2 = 0, \\ 0 & \text{otherwise,} \end{cases}$$

$$q((\boldsymbol{0}, \boldsymbol{e}_j, \boldsymbol{0}), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} 1 & \text{if } j = \lambda_B \text{ and } s_1 = s_2 = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

Table 2.1: **Running times (in seconds) for solving recursions and computing Padé coefficients.** The second column is the time to solve the two-locus recursion to compute the likelihood of a *single* value of $\rho$ for all sample configurations of size $n$. The third column is the time to compute 11 Padé coefficients for all sample configurations of size $n$. Recall that the recursion must be solved afresh for every value of $\rho$ in the lookup table. On the other hand, the Padé coefficients are used to construct a rational function of $\rho$ that approximates the likelihood; once the Padé coefficients are determined, evaluating the likelihood is instantaneous. A single 2.5 Ghz core was used in this benchmarking to provide representative estimates of the running time. However, note that both the recursion and Padé coefficient computations are highly parallelizable, which we exploit in the implementation of `LDhelmet`. Also note that the presence of missing data does not increase the running time for either computation.

| Sample size $n$ | Two-locus recursion (seconds) | Padé coefficients (seconds) |
|:-:|--:|--:|
| 10 | 0.1 | 5 |
| 20 | 11 | 429 |
| 30 | 189 | 5271 |
| 40 | 1523 | 26405 |
| 50 | 7755 | 75704 |

## 2.4 Population-scaled Recombination Parameter

Our method seeks to infer the fine-scale map of the population-scaled recombination rate in *D. melanogaster*, in which recombination occurs only in females. The population-scaled recombination rate between a pair of sites in the $X$ chromosome is defined as $\rho_X = \frac{8}{3}N_e^X r_f^X$, where $N_e^X$ is the effective population size for X and $r_f^X$ is the probability of recombination between the sites per generation per X chromosome in females. The population-scaled recombination rate between a pair of sites in an autosome is defined as $\rho_A = 2N_e^A r_f^A$, where $N_e^A$ is the effective population size for the autosome and $r_f^A$ is the recombination rate between the sites per generation per autosome in females. Furthermore, $N_e^X$ and $N_e^A$ are defined as $N_e^X = 9N_f N_m/(4N_m + 2N_f)$ and $N_e^A = 4N_f N_m/(N_f + N_m)$, where $N_f$ and $N_m$ denote the effective number of female and male individuals in the population. If we assume $N_f = N_m = N_e/2$, we obtain $\rho_X = 2N_e r_f^X$ and $\rho_A = 2N_e r_f^A$.

In contrast to recombination, mutation occurs in both males and females. We denote the X chromosome mutation rates in females and males as $\mu_f^X$ and $\mu_m^X$, respectively, and the autosomal mutation rates in females and males as $\mu_f^A$ and $\mu_m^A$, respectively. Then, the population-scaled mutation rates for X and the autosomes are given by $\theta_X = \frac{4}{3}N_e^X(2\mu_f^X + \mu_m^X)$ and $\theta_A = 2N_e^A(\mu_f^A + \mu_m^A)$, respectively. Further, if $N_f = N_m = N_e/2$, then the expressions simplify to $\theta_X = N_e(2\mu_f^X + \mu_m^X)$ and $\theta_A = 2N_e(\mu_f^A + \mu_m^A)$.

In our statistical model, we allow the recombination rate to vary across the genome. We use $\rho$ to denote generically the population-scaled recombination map, which is a function of genomic position. For ease of notation, we do not add a subscript to $\rho$ to distinguish between X and autosome; it should be clear from the context which is intended. Similarly, we use $\theta$ to denote generically the population-scaled mutation rate.

Our objective is to estimate the recombination map $\rho$ from population genomic DNA sequence data. Our approach introduces several key improvements to the method `LDhat` [60, 59] (v2.1 used throughout), which was first developed for estimating fine-scale recombination maps in humans.

## 2.5 Features of the Method

To accommodate the higher recombination rate observed in *D. melanogaster*, we introduce several key modifications and additions to `LDhat` to improve the accuracy and robustness of recombination map estimation. Instead of using importance sampling to compute the two-locus likelihoods, we compute them by solving a systems of recursion relations, thereby producing more accurate lookup tables. An additional benefit of this approach is that we can handle large amounts of missing data at no additional computational cost, since the likelihoods of configurations with missing data naturally appear in the system of recursions. Our method incorporates a general tetra-allelic mutation model, whereas `LDhat` assumes a diallelic model. As a consequence, we can handle complex mutation patterns between the A, C, G, T nucleotides. Furthermore, our method can use different mutation transition matrices for different sites at no extra computational cost. We make use of the recent work [40, 39, 41, 12] on asymptotic sampling distributions to incorporate a larger range of $\rho$ values in the lookup table in a computationally tractable manner. The lookup table exhibits a finer grid resolution for values of $\rho$ in regions of higher likelihood curvature, for improved accuracy. We infer a distribution on the ancestral allele at each site and use this information to compute more refined likelihoods. The prior for the recombination map is more flexible and can be tailored to the particular species under analysis. For example, when analyzing a species that is believed to have significantly higher recombination rates than that of humans, as is the case for *D. melanogaster*, one should not use the same prior as for humans.

To improve computational tractability, we assume that each site underwent at most one mutation in the entire genealogical history of the sample. This assumption is reasonable for small values of $\theta$, as is the case for *D. melanogaster*, and it provides several computational advantages, described in the following sections.

## 2.6 Missing Data

Because the two-locus recursion relation is solved jointly for every configuration, this also gives us exact solutions for every subconfiguration at no extra computational cost. In particular, we emphasize that we also obtain likelihoods for all relevant configurations with any *missing* data, at no extra computational cost. By contrast, when `LDhat` encounters a configuration in which some alleles are missing, its approach is to marginalize over missing alleles by summing over the relevant entries in its lookup table for fully-specified haplotypes, but the time required for this computation scales poorly with the number of missing alleles. The extent of missing data in the *D. melanogaster* genomes is such that this approach is impracticable. On the data we analyzed, we masked all alleles with a quality less than 30. For the RAL lines, about 20% of the data was missing, and for the RG lines, about 8% of the data was missing. The more missing data there is, the more expensive marginalization becomes, and the greater the number of distinct configurations present in the data.

## 2.7 Incorporating a Tetra-allelic Mutation Model

One key advantage of our approach is that it can make use of all four alleles (A, C, G, T) in sequence data, together with the ancestral alleles inferred from outgroup sequences. This is achieved with modifications to the boundary conditions of the appropriate two-locus recursion described above. In combination with the one-mutation-per-site assumption, this allows us to use a full 4 × 4 transition matrix $\boldsymbol{P} = (P_{ij})_{i,j\in\{A,C,G,T\}}$ to model realistic mutation patterns between nucleotides, with no significant amount of extra computation: Suppose the ancestral allele at each of a given pair of segregating sites is known to be $A$ and $C$, respectively. At the first site some chromosomes exhibit a derived $G$ allele, and at the second site some chromosomes exhibit a derived $T$ allele. Because of the one-mutation-per-site assumption and the decoupling of the genealogical and mutational processes under neutrality, it is easy to see that the likelihood of this two-locus configuration has a dependence on $\boldsymbol{P}$ only through a single multiplicative factor $P_{AG}P_{CT}$. Hence, this expression can be factorized completely out of the two-locus likelihoods and hence from our lookup tables. The remaining quantity, which represents the probability of observing a particular configuration up to the identities of the alleles involved, can be multiplied by the relevant pair of entries in $\boldsymbol{P}$ for *any* observed combination of nucleotides. To be precise, if $q(\boldsymbol{n}, 1, 1; \theta, \rho)$ denotes our solution to the system of equations described above, this argument shows we can write

$$q(\boldsymbol{n}, 1, 1; \theta, \rho) = P_{AG}P_{CT}F(\boldsymbol{n}; \theta, \rho), \tag{2.6}$$

for some function $F$ independent of $\boldsymbol{P}$. [The single-locus analogue of this result is evident in equation (2.3).] We then need to store only $F(\boldsymbol{n}; \theta, \rho)$. If later we see the same combination of haplotype counts but for a different combination of nucleotides, we can reuse this quantity and multiply it by different relevant entries in $\boldsymbol{P}$. For simplicity, in our analysis we used the same $\boldsymbol{P}$ for each site in the genome, but note that, because of the factorization in (2.6), it is possible to use different mutation transition matrices for different sites at no extra computational cost.

This approach easily generalizes to the case where the ancestral allele is not known or where we only have a distribution on the ancestral allele at each site. We can simply take the weighted average over each of the four possible combinations of ancestral alleles, weighted with respect to their distributions. In the case where no information is known about the ancestral alleles, this reduces to using the stationary distribution of $\boldsymbol{P}$ as the distribution over ancestral alleles at each site.

## 2.8 Estimation of Mutation Transition Matrices

Because we are now able to make full use of a tetra-allelic mutation model, we developed a method to estimate the $4 \times 4$ mutation transition matrix $\boldsymbol{P}$ from empirical data, for subsequent use in our recombination rate inference. We use the following parsimony-based method to estimate $\boldsymbol{P}$ by inferring the ancestral allele at each site in *D. melanogaster* by comparison with aligned outgroup reference genomes of *D. simulans*, *D. erecta*, and *D. yakuba*. We designate the ancestral allele at each dimorphic site in *D. melanogaster* using the following rule. If the alleles of the three outgroups are not all missing at this site and together exhibit precisely one of the four possible nucleotides, and if this allele agrees with one of the two observed in *D. melanogaster*, then this is designated as the ancestral allele. Otherwise, it is considered unknown and discarded from the analysis. (We also discarded triallelic and tetra-allelic sites.) A related approach is used in the Drosophila Population Genomics Project in the estimation of divergence. We tried both more and less restrictive parsimony rules, as well as excluding CpG sites from our analysis; neither variation substantially altered our results.

Given a large collection of SNPs in our dataset for which the ancestral allele is known, we can infer the identities of the alleles involved in the mutation event at each polymorphic site. For example, an A/G polymorphism with A ancestral implies a historical A$\mapsto$G transition. The relative frequencies of each type of event, normalized to account for varying genomic content of the four nucleotides, determines our empirical estimate of $\boldsymbol{P}$. To be precise, let $f_{\mathrm{A}}$ denote the total number of A nucleotides in the *D. melanogaster* genome, of which $f_{\mathrm{AC}}$, $f_{\mathrm{AG}}$, and $f_{\mathrm{AT}}$ have been inferred to be A$\mapsto$C, A$\mapsto$G, and A$\mapsto$T polymorphisms, respectively. (For consistency we restrict all these definitions only to those monomorphic or dimorphic sites for which sufficient, consistent outgroup information is also available, as required above.) We make analogous definitions for $f_i$ and $f_{ij}$, for each $i, j \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Finally, let $M = \max_{i \in \{A,C,G,T\}} \{(\sum_{j \neq i} f_{ij})/f_i\}$, the largest empirical frequency of mutation away from any particular nucleotide. The appropriate choice for $\boldsymbol{P}$ is given by

$$P_{ij} = \begin{cases} \dfrac{f_{ij}}{f_i M}, & i \neq j, \\ 1 - \displaystyle\sum_{j \neq i} \dfrac{f_{ij}}{f_i M}, & i = j. \end{cases}$$

Division by $M$ ensures that, without loss of generality, one entry in the diagonal of $\boldsymbol{P}$ is zero. By allowing the diagonal entries of $\boldsymbol{P}$ to be nonzero, different nucleotides can have different overall mutation rates. The total “effective” mutation rate—that is, mutations not involving the diagonal entries of $\boldsymbol{P}$—is calibrated against classical infinite-sites-based estimators: for RAL this is $\theta_{\text{effective}} = 0.006$ per bp (autosomes) and $\theta_{\text{effective}} = 0.004$ per bp (X chromosome). For RG we used $\theta_{\text{effective}} = 0.006$ per bp for all chromosomes. Since we are to use a general tetra-allelic model in which both effective and ineffective mutations are permitted to occur, the appropriate choice of $\theta$ for use with $\boldsymbol{P}$ is such that it exhibits the same overall rate of effective mutations:

$$\theta_{\text{effective}} = \theta \sum_i \left( \frac{f_i}{\sum_k f_k} \sum_{j \neq i} P_{ij} \right).$$

## 2.9 Ancestral Allele Distribution

When it is not known which of the two alleles at a polymorphic site is ancestral, one can use the stationary distribution of $\boldsymbol{P}$ as a prior distribution over the ancestral allele. However, when additional information is available, such as sequence data from an outgroup, we can use the information to update our prior beliefs about the identity of the ancestral allele, thus allowing a more accurate estimate of the recombination map. In our application, we used the *D. simulans* outgroup information to update our prior distributions on the ancestral alleles of the *D. melanogaster* samples. Specifically, for each *D. melanogaster* genome, we used the software `psmc` [51] to estimate, at each site, a distribution on the time to the most recent ancestor (TMRCA) of the *D. melanogaster* and *D. simulans* genomes. Given the TMRCA, we integrate over possible mutations occurring according to $\boldsymbol{P}$ along the two branches, to obtain a distribution on the ancestral allele. Finally, for each site, we aggregate each of these pairwise distributions into a single distribution on the ancestral allele, and use this distribution in the computation of our likelihoods.

Suppose we have one genomic sequence of *Drosophila simulans* and $n$ sequences of *Drosophila melanogaster*. Let $S$ represent the sequence of *D. simulans* and $M^{(k)}$ represent the sequence of the $k$th *D. melanogaster*, where $S_l$ denotes the $l$th base of the sequence, and $S_{\hat{l}}$ represents the sequence with the exclusion of the $l$th base. Given $(S, M^{(k)})$, let $T_l^{(k)}$ be the time to the most recent common ancestor (TMRCA) at locus $l$; $f_l^{(k)}(t \mid M_{\hat{l}}, S_{\hat{l}})$ be the density of the TMRCA conditioned on both their sequences but *excluding* the $l$th locus; and $A_l^{(k)}$ be the ancestral allele at the $l$th locus, i.e., the allele of the most recent common ancestor (MRCA).

To compute the distribution on the ancestral allele at the $l$th locus conditioned on $M^{(k)}$ and $S$, we use Bayes' theorem to obtain

$$\begin{aligned}
&\mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S) \\
&= \frac{\int_0^\infty p(A_l^{(k)} = i, M^{(k)}, S, T_l^{(k)} = t)dt}{\mathbb{P}(M^{(k)}, S)} \\
&= \frac{\int_0^\infty \mathbb{P}(M_l^{(k)}, S_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)})p(A_l^{(k)} = i, T_l^{(k)} = t)dt}{\mathbb{P}(M^{(k)}, S)} \\
&= \frac{\int_0^\infty \mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)} = t)\mathbb{P}(S_l \mid A_l^{(k)} = i, T_l^{(k)} = t)\mathbb{P}(A_l^{(k)} = i) f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})dt}{\sum_j \int_0^\infty \mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = j, T_l^{(k)} = t)\mathbb{P}(S_l \mid A_l^{(k)} = j, T_l^{(k)} = t)\mathbb{P}(A_l^{(k)} = j) f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})dt}.
\end{aligned} \tag{2.7}$$

In equation (2.7), the prior on the ancestral allele at locus $l$, $\mathbb{P}(A_l^{(k)} = i)$, is given by the stationary distribution of the allele frequencies from the mutation matrix $\boldsymbol{P}$. (In the above, $p$ denotes a joint probability of discrete events together with the density for $T_l^{(k)}$.) The density on the TMRCA, $f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})$, is estimated using Li and Durbin's `psmc` [51]. In practice, we use `psmc` to compute $f_l^{(k)}(t \mid M^{(k)}, S)$ and assume $f_l^{(k)}(t \mid M^{(k)}, S) \approx f_l^{(k)}(t \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})$.

The remaining two probabilities, $\mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)} = t)$ and $\mathbb{P}(S_l \mid A_l^{(k)} = i, T_l^{(k)} = t)$, are computed as follows. For the computation of $\mathbb{P}(M_l^{(k)} \mid A_l^{(k)} = i, T_l^{(k)} = t)$, let $\boldsymbol{P} = (P_{ij})$ denote the mutation matrix, and let $r_l^{(k)}$ specify the number of mutations that have occurred at the $l$th locus of the $k$th *D. melanogaster* sequence during time $T_l^{(k)}$.

Then we have

$$\begin{aligned}
\mathbb{P}(M_l^{(k)} = j \mid A_l^{(k)} = i, T_l^{(k)} = t) &= \sum_{s=0}^{\infty} \mathbb{P}(r_l^{(k)} = s \mid T_l^{(k)} = t)(\boldsymbol{P}^s)_{ij} \\
&= \sum_{s=0}^{\infty} \left(\frac{\theta t}{2}\right)^s \frac{e^{-\theta t/2}}{s!} (\boldsymbol{P}^s)_{ij} \\
&= \sum_{s=0}^{\infty} \left[\left(\frac{\theta t}{2}\boldsymbol{P}\right)^s\right]_{ij} \frac{e^{-\theta t/2}}{s!} \\
&= \left[e^{\frac{\theta t}{2}(\boldsymbol{P}-\mathbf{I})}\right]_{ij},
\end{aligned}$$

where $\mathbf{I}$ is the identity matrix with the same dimensions as $\boldsymbol{P}$. The computation for $\boldsymbol{P}(S_l \mid A_l^{(k)} = j, T_l^{(k)} = t)$ is analogous.

After computing $\mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S)$ for every $k$ and given $l$, we heuristically aggregate these pairwise probabilities to estimate $\mathbb{P}(A_l^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S)$ as follows. Let $\bar{t}_l^{(k)}$ be the posterior mean of $f_l^{(k)}(t \mid M^{(k)}, S)$, i.e.:

$$\bar{t}_l^{(k)} = \int_0^{\infty} t f_l^{(k)}(t \mid M^{(k)}, S) dt,$$

and define $\tau_l = \max_k \bar{t}_l^{(k)}$. We approximate $\mathbb{P}(A_l^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S)$ as

$$\mathbb{P}(A_l^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S) \approx \frac{\sum_{k=1}^{n} \mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S) f_l^{(k)}(\tau_l \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})}{\sum_j \sum_{k=1}^{n} \mathbb{P}(A_l^{(k)} = j \mid M^{(k)}, S) f_l^{(k)}(\tau_l \mid M_{\hat{l}}^{(k)}, S_{\hat{l}})},$$

which is a weighted average of $\mathbb{P}(A_l^{(k)} = i \mid M^{(k)}, S)$ over $k$, weighted by the density of the TMRCA evaluated at $\tau_l$ for each $k$. This averaging should mitigate effects such as genotyping errors and incomplete lineage sorting in individual *D. melanogaster* genomes.

## 2.10 Padé Approximants

Recall that `LDhat`'s lookup tables are precomputed over a grid: $\rho = 0, 1, \ldots, 100$. For a pair of sites with a recombination rate greater than 100, the likelihood at $\rho = 100$ is used as an approximation. This can create systematic errors in the likelihood [40]. Instead, for $\rho > 100$ we compute accurate approximations to the two-locus likelihood using the method of Padé summation described in Jenkins & Song [41]. Briefly, one Taylor expands $q(\boldsymbol{n}, s_1, s_2; \theta, \rho)$ about $\rho = \infty$ and uses the method of Jenkins & Song to compute the first few terms in the expansion. In practice this Taylor series rapidly diverges for values of $\rho$ of interest, but it can be made into an accurate, convergent approximation of $q(\boldsymbol{n}, s_1, s_2; \theta, \rho)$ by replacing this truncated series with a rational function approximation whose own Taylor series agrees as far as possible, a technique known as Padé summation. We modified the analysis of Jenkins & Song to account for our new system of equations. We precompute 11 Padé coefficients (up to $1/\rho^{10}$ in the Taylor series expansion of the likelihood about $\rho = \infty$) for every sample configuration of size $n$, which gives an extremely accurate approximation for every $\rho > 100$ (not just integral values). Usually, the "join" between the Padé approximant for $\rho > 100$ and the true likelihood for $\rho \leq 100$ is indistinguishable. We also employ a "defect heuristic" [41] with threshold parameter $\epsilon = 40$ to correct for potential effects from singularities in the Padé approximants. As in the direct computation of the likelihoods from the system of equations, obtaining the Padé coefficients for a given configuration also yields the coefficients for all its subconfigurations. This approach is therefore well-suited to data with a large proportion of missing data. Table 2.1 shows the running time for the Padé coefficient computation as a function of sample size $n$.

Modifications to the approach described in [41] are made, following from the boundary conditions given above. These can be converted into modifications of entries of the dynamic programming tables given in [41]. For example, using (2.5) we have that

$$\begin{aligned} q((\boldsymbol{a}, \boldsymbol{0}, \boldsymbol{e}_{i\lambda_B}), 1, 0; \theta_A, \theta_B, \rho) &= q((\boldsymbol{a} + \boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{0}), 1, 0; \theta_A, \theta_B, \rho) \\ &= q(\boldsymbol{a} + \boldsymbol{e}_i, 1; \theta_A) + \frac{0}{\rho} + \frac{0}{\rho^2} + \ldots, \end{aligned}$$

where $q(\boldsymbol{a} + \boldsymbol{e}_i, 1; \theta_A)$ is the one-locus solution given by equation (2.3). Notice that this expansion is in fact independent of $\rho$, from which it follows (by comparison with eq. (3.7) of [41]) that a number of entries in the dynamic programming tables are modified. For example, the second row in the dynamic programming table for the configuration $(\boldsymbol{a}, \boldsymbol{0}, \boldsymbol{e}_{i\lambda_B})$ is set to zero. Other boundary conditions may be interpreted in a similar fashion.

## 2.11 Lookup Table Grid Resolution

One can imagine that it would be useful to have a more refined lookup table in regions of higher curvature of the likelihood. In such regions simply using integral values of $\rho$ might be too coarse. Since the lookup tables will be used for every conceivable pairwise dataset, we should be interested in the expected curvature of the likelihood curve at $\rho$, across datasets drawn under a model with the same $\rho$. (That is, the curvature at some $\rho_0$ is most important for datasets that we are likely to see when the recombination rate really is $\rho_0$.) This is reflected by Fisher's information:

$$I(\rho_0) = -\mathbb{E}_{\rho_0}\left[\frac{\partial^2}{\partial \rho^2} \ln L(\rho; \mathcal{D})\right],$$

which can be estimated from an existing lookup table using the second-order central difference operator. As is evident from Figure 2.1, curvatures are generally higher in the range $0 \leq \rho \leq 10$, and so we changed the increment between $\rho$ values in the lookup table from 1 to 0.1 in this range.

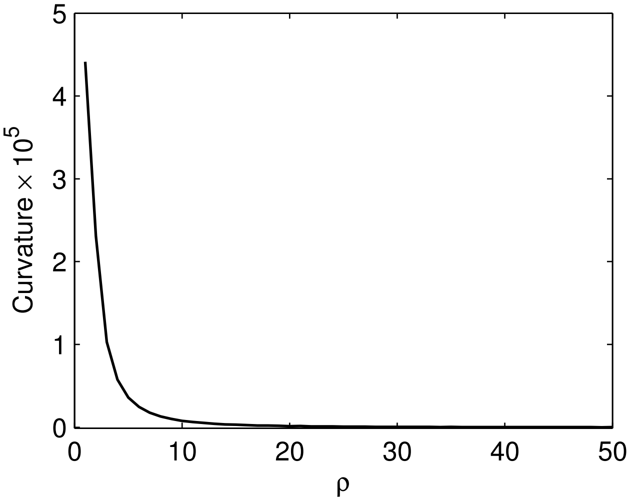

Figure 2.1: **Fisher's information for two=-locus samples.** Fisher's information for two-locus samples of size $n = 37$ using lookup tables for $\theta = 0.006$ and under the infinite-sites assumption. The ancestral haplotype is assumed to be known.

## 2.12 Prior and Block Penalty

`LDhelmet` places a prior distribution on the number of change points, the positions of the change points, and the heights of the change points in the recombination map. The prior on the number of change points is, as in `LDhat`, a Poisson distribution with mean equal to $(S-2)\exp(-\xi)$, where $S$ is the number of SNPs in the data and $\xi$ is a user-defined parameter called the *block penalty*. The positions of the change points are distributed uniformly, and the distribution on the heights of the change points is user-settable as exponential, gamma or log-normal.

One should be mindful that `LDhat` was designed for background recombination rates an order of magnitude less than that used in the simulations. In particular, `LDhat` implements the exponential prior but the mean is hard-coded for human data. Adjusting the mean of the prior according to the expected background recombination rate is necessary to obtain meaningful results. For example, using a prior suitable for humans on *Drosophila*-type data produces poor estimates with little to none of the true variation in the underlying recombination map (simulations not shown). To facilitate a comparison, we modified the source code of `LDhat` such that its prior was similar to the one used by `LDhelmet`. Without such modifications, the estimates from `LDhat` were not comparable to `LDhelmet`'s estimates. In the simulations and analysis, we used an exponential prior with the mean adjusted for the expected background rate of *D. melanogaster*.

## 2.13 Proposal Distribution and Metropolis-Hastings Ratios

The rjMCMC procedure proposes one of four moves per iteration:

1. **Change** the rate of a change point.

2. **Reposition** a change point.

3. **Split** a change point into two change points.

4. **Merge** two change points into one change point.

The Change move is selected with probability 0.4 and the others are each selected with probability 0.2. The proposed state is then made according to the chosen move.

The Change move selects a change point uniformly at random (excluding the initial change point and the final change point) and proposes changing its rate $r$ to

$$r' = e^U r,$$

where $U \sim \text{Uniform}([-1/2, 1/2])$.

The Metropolis-Hastings ratio for the Change move is

$$\frac{l'r'}{lr}(r' - r)^{\frac{1}{\overline{r}}},$$

where $l$ is the likelihood of the current state, $l'$ is the likelihood of the proposed state, $r$ is the rate of the selected change point, and $\overline{r}$ is the mean of the prior on rates.

The Reposition move selects a change point uniformly at random (excluding the initial change point and the final change point) and proposes moving it to another SNP. The new SNP is chosen uniformly at random from among those strictly between the change points to the left and right of the selected change point.

The Metropolis-Hastings ratio for the Reposition move is

$$\frac{l'}{l},$$

where $l$ is the likelihood of the current state and $l'$ is the likelihood of the proposed state.

The Split move selects a change point uniformly at random (excluding the initial change point and the final change point) and adds a new change point at a uniformly random SNP between the selected change point and the change point to its right. The rate of the selected change point is proposed to be

$$r' = r\left(\frac{u}{1-u}\right)^{\frac{p_r - p_n}{p_r - p}},$$

where $r$ is the current rate of the selected change point, $u$ is a sample from Uniform($[0,1]$), $p_r$ is the position of the right change point, $p_n$ is the position of the new change point, and $p$ is the position of the selected change point.

The rate of the new change point is then set to

$$r_n = \frac{1-u}{u}.$$

The Metropolis-Hastings ratio for the Split move is

$$\frac{l'(S-k)(k-1)}{l(S-2)^2} e^{-\xi - \frac{r'+r_n-r}{\overline{r}}} \frac{\mathbb{P}_{\text{merge}}}{\mathbb{P}_{\text{split}}} \frac{(r'+r_n)^2}{r\overline{r}},$$

where $l$ is the likelihood of the current state, $l'$ is the likelihood of the proposed state, $S$ is the number of SNPs, $k$ is the number of change points before proposing to add the new change point, $\xi$ is a user-specified parameter called the block penalty, $\overline{r}$ is the mean of the prior on rates, $r$ is the current rate of the selected change point, $\mathbb{P}_{\text{merge}}$ is the probability of proposing a Merge move, and $\mathbb{P}_{\text{split}}$ is the probability of proposing a Split move.

The Merge move selects a change point uniformly at random (excluding the initial change point and the final change point) and removes it. The rate of the change point to the left of the removed change point is set to

$$r' = r_l^{\frac{p-p_l}{p_r-p_l}} \cdot r^{\frac{p_r-p}{p_r-p_l}},$$

where $r$ is the rate of the selected change point, $r_l$ is the rate of the change point to the left of the selected change point, $p$ is the position of the selected change point, $p_l$ is the position of the change point to the left, and $p_r$ is the position of the change point to the right.

The Metropolis-Hastings ratio for the Merge move is

$$\frac{l'(S-2)^2}{l(S-k)(k-1)} e^{\xi - \frac{r_l+r-r'}{\overline{r}}} \frac{\mathbb{P}_{\text{split}}}{\mathbb{P}_{\text{merge}}} \frac{r'\overline{r}}{(r_l+r)^2},$$

where $l$ is the likelihood of the current state, $l'$ is the likelihood of the proposed state, $S$ is the number of SNPs, $k$ is the number of change points before proposing to remove the selected change point, $\xi$ is a specified parameter called the block penalty, $r_l$ is the rate of the change point to the left of the selected change point, $\overline{r}$ is the mean of the prior on rates, $\mathbb{P}_{\text{split}}$ is the probability of proposing a Split move, and $\mathbb{P}_{\text{merge}}$ is the probability of proposing a Merge move.

## 2.14 Selecting Parameters

The block penalty controls the extent of variation in the estimated recombination map. In general, the higher the block penalty, the smoother the estimated map. We carried out a simulation study to choose a conservative penalty value to reduce false positive inference of hotspots, at the expense of tolerating more false negatives.

We considered the following three scenarios: no recombination variation (constant rate), moderate variation (with a hotspot of width 2 kb and intensity 10× the background rate), and high variation (with a hotspot of width 2 kb and intensity of 50× the background rate, such as that seen in humans). We simulated 100 datasets of each kind, with a fixed background rate of $\rho = 10$ per kb in all cases. After considering a variety of evaluation metrics for measuring the accuracy of an estimated map, we found the $\ell_1$-distance between the true map and the estimated map to be the simplest to interpret and assess, where the $\ell_1$-distance is the sum of the point-wise differences between the true and estimated maps. For the three scenarios described above, Figure 2.2 shows the average $\ell_1$-distances between the true recombination maps and the estimated maps for various block penalty values and recombination landscapes. For each dataset, we ran `LDhelmet` for 250,000 iterations after a 50,000 iteration burn-in. We observed that noise from overfitting is reduced for higher block penalties. Based on our simulation study, we chose a conservative block penalty of 50 in our analysis of the real data.

In our simulation study for evaluating the choice of block penalty on realistic data (Figure 2.5), we used the program `MaCS` [17] to simulate a 1 Mb region with a highly variable recombination map. (We used $n = 22$ and $\theta = 0.008$; output was postprocessed to incorporate an empirical tetra-allelic mutation model.) The map's variability was taken from a 1 Mb excerpt of the estimated recombination map of the X chromosome for the RAL sample. The total recombination rate for the region was then rescaled to match the mean (per Mb) rate of the RAL X chromosome (to create a "RAL-like" map) or the RG X chromosome (to create a "RG-like" map; see Figure 2.5).

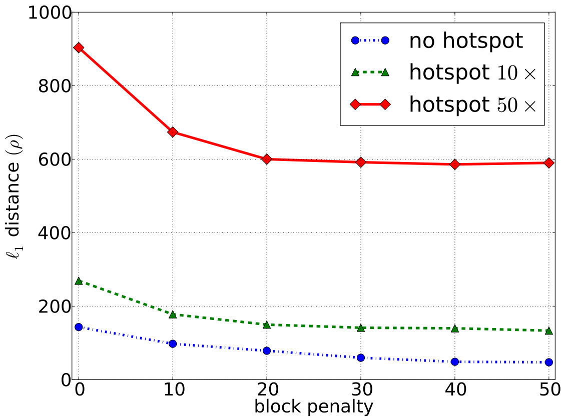

Figure 2.2: **Plot of the average $\ell_1$-distance between the true and estimated recombination maps.** Each plot shows the results averaged over 100 simulated datasets per block penalty for a given recombination landscape. In each simulation, we considered a 25 kb region with the background recombination rate of $\rho = 10/\text{kb}$. "no hotspot": The true recombination map is constant. "hotspot 10×": In the middle of the 25 kb region, the true recombination map has a hotspot of width 2 kb and intensity 10× the background rate. "hotspot 50×": In the middle of the 25 kb region, the true recombination map has a hotspot of width 2 kb and intensity 50× the background rate.

## 2.15 Data

The mean coverage of the RAL data was $\geq 10\times$. Regions of residual heterozygosity and regions of identity-by-descent between genomes were masked in the RAL data, in addition to a quality filter of Q30 applied to both populations. Preliminary analysis by the DPGP2 group found evidence of admixture among 5 of the 22 RG lines we considered, in addition to evidence for minor levels of identity-by-descent between genomes. To maintain a reasonable sample size, these regions were not masked in the results. We did repeat several of our analyses with these regions excluded and generally found little difference. Despite the extensive filtering, which increases the amount of missing data, the runtime complexity of our method does not increase from a lack of data, as it does for `LDhat`.

The data were divided into overlapping blocks of 4,400 SNPs each, with 200 SNPs of overlap on either end of a block. For every block, `LDhelmet` was run for 3,000,000 iterations after 300,000 iterations of burn-in. The map for each chromosome or chromosome arm was constructed by removing 200 SNPs from the ends of the blocks and concatenating the blocks together.

## 2.16 Simulation Study on the Impact of Natural Selection

In order to simulate datasets that had been affected by natural selection, we focused on modeling the effects of sites experiencing positive, genic selection, i.e. selective sweeps. We investigated two modeling scenarios: First, the effect of a single, strong sweep with its strength, fixation time, and location treated as fixed parameters. Under some parameter combinations, we expect such sweeps to substantially reduce observed polymorphism levels. Second, we considered data generated under the influence of a recurrent sweep model, in which the ages and genomic locations of sweeps occur randomly. In this scenario, we chose the parameters of the model (selection coefficient and rate of fixation of beneficial mutations) such that expected polymorphism levels were concordant with observations in *D. melanogaster*. While the second scenario is likely to be a more realistic model for the forces affecting variation in *D. melanogaster* genomes, its inherent randomness introduces additional noise. The first scenario allows us to study the effects of a sweep with particular characteristics under a controlled environment.

Under both scenarios, we again simulated data under three possible recombination landscapes: a flat recombination rate of $\rho = 10$ per kb except for a 2 kb-wide hotspot at the center of the sequence, of relative strength 1 (no hotspot), 10, or 50; we also post-processed all outputs to allow for a full tetra-allelic mutation model, using the mutation transition matrix $\boldsymbol{P}_{\text{RAL}}$. To reconstruct the recombination maps of simulated data, we used the following parameters for `LDhelmet` and `LDhat`: 250,000 iterations after 50,000 iterations of burn-in for `LDhelmet`, and 1,000,000 iterations after 100,000 iterations of burn-in for `LDhat`. We chose the number of iterations such that the two methods would require about the same computational time.

### 2.16.1 Single Sweep Model

In order to simulate datasets that had experienced a single hard sweep, we used the software `mbs` [75]. Using `mbs`, we simulated the trajectory of a selected allele backwards from its fixation time at the present back to the random time of its birth, then post-processed the software's output to translate the trajectory such that its fixation time was instead at time $T_{\text{fix}}$ in the past. (This lets us condition on $T_{\text{fix}}$, which is otherwise not possible using the software.) Subject to this trajectory we then used `mbs` to simulate $n = 37$ samples of 25 kb of sequence in the vicinity of the selected site. We simulated 100 trajectories for each possible combination of the following choices of parameter: selection strength $4N_es = 0$, 10, $10^2$, $10^3$, and $10^4$ (where $s$ is the relative fitness); $T_{\text{fix}} = 0.01$, 0.1, 0.5, 1, and 5, in units of $4N_e$ generations; and three possible recombination landscapes (see above). For each trajectory we simulated, independently, a 25 kb sample with the selected site at coordinate $-100$, $-50$, $-10$, 0, 5, or 12.5 kb with respect to the start coordinate of the sequences. In total, this procedure generated $100 \times 6 \times 5 \times 5 \times 3 = 45,000$ independent datasets for input into `LDhelmet` and `LDhat`.

### 2.16.2 Recurrent Sweep Model

In order to simulate datasets experiencing hard sweeps at random times and locations, we modified the software `rsweep` [43] to allow for a recombination hotspot rather than a constant recombination landscape. As above, we simulated datasets of $n = 37$ samples of 25 kb of sequence, this time under three realistic recurrent sweep models:

(RS1) $s = 10^{-5}$, $2N_e\lambda = 2 \times 10^{-3}$,

(RS2) $s = 10^{-4}$, $2N_e\lambda = 2 \times 10^{-3}$,

(RS3) $s = 10^{-2}$, $2N_e\lambda = 2 \times 10^{-5}$,

where $s$ is the selection coefficient of new beneficial mutations and $2N_e\lambda$ is the rate of fixation of beneficial mutations. In each case we took $N_e = 2.5\times10^6$. The first parameter combination is one of frequent, weak sweeps, and similar to the parameters estimated in [2]. The third combination is one of infrequent but stronger sweeps and similar to the parameters estimated in [43]. The second combination is intermediate between the two. Under a recurrent sweep model, selective sweeps occur at random times at a rate governed by $2N_e\lambda$ and at a location in the genome chosen uniformly at random. Sweeps both within the sequenced 25 kb and in flanking sequence can affect the observed data and are accounted for in the simulation software [43].

We considered $\rho = 10$ per kb for comparison with the single sweep model. As in the single sweep model, we simulated 100 datasets under each parameter combination, generating $100 \times 3 \times 3 = 900$ independent datasets for input into `LDhelmet`. As for the single sweep simulations, we ran `LDhelmet` for 250,000 iterations after 50,000 iterations of burn-in.

## 2.17 Simulation Study on the Impact of Demographic History

In order to simulate datasets that had been affected by a nonstandard demographic history, we used the software `msHOT` [33]. We investigated four realistic demographic histories:

(G1) Exponential growth at rate 100 initiated $0.023N_e$ generations ago (a tenfold increase by the present time),

(G2) Exponential growth at rate 10 initiated $0.161N_e$ generations ago (a fivefold increase by the present time),

(B1) A bottleneck initiated $0.5N_e$ generations ago, with a transient reduction to size $0.00001N_e$ lasting $0.00002N_e$ generations,

(B2) A bottleneck initiated $0.0055N_e$ generations ago, with a transient reduction to size $0.029N_e$ lasting $0.00375N_e$ generations.

The first three models were proposed by Haddrill *et al.* [32] as reasonable fits to their (African) data, while the fourth is taken from [78] for a European population. We note that the precise demographic history of *D. melanogaster* populations remains poorly understood, and that these models simply serve as reasonable examples for investigating the robustness of our method. It is probable that there exist better fitting demographic models; indeed, Haddrill *et al.* ultimately favor their bottleneck model over any growth model.

We simulated 100 datasets under each model and under each of three recombination landscapes: a flat recombination rate of $\rho = 10$ per kb except for a 2 kb-wide hotspot at the center of the sequence, of relative strength 1 (no hotspot), 10, or 50. This provided $100 \times 4 \times 3 = 1,200$ independent datasets in total. We also post-processed all outputs from the infinite-sites-based software to allow for a full tetra-allelic mutation model, using the mutation transition matrix $\boldsymbol{P}_{\mathrm{RAL}}$ and the mutation rate $\theta = 0.008$ per bp. We ran `LDhelmet` for 250,000 iterations after 50,000 iterations of burn-in.

## 2.18 Search for Recombination Hotspots

We used a conservative approach to identify candidate recombination hotspots. From the recombination maps for RAL and RG we first identified putative hotspots—regions in which the recombination rate exceeded ten times the mean for that chromosome arm, and which were greater than 500 bp in length. We discarded regions of length less than 500 bp on the grounds that such narrow peaks can be produced occasionally as spurious artifacts of the rjMCMC procedure.

To further filter the remaining candidate hotspots, we applied an independent method, `sequenceLDhot` [25], to the same data, in order to test for the presence of hotspots in these regions. The software uses a computationally-intensive importance sampling framework to construct likelihood ratios in sliding windows to evaluate the evidence for the presence of a hotspot in that window. To reduce computation time we focused on 50 kb regions centered on the autosomal putative hotspots. We modified `sequenceLDhot`'s default parameters, which are tuned for interrogating human data, as follows. We used $\theta = 0.008$ per site, and for the background recombination rate we used the estimated mean across the local 50 kb containing the hotspot of interest. We specified the software's grid for hotspot likelihoods to be in the range 10–100 times the background rate, and tested windows of 500 bp sliding in steps of 250 bp, using a composite likelihood comprising ten SNPs. Other parameters were unchanged. We reduced SNP density to be comparable to the data on which the software had been calibrated [25], by discarding sites with any missing alleles and singleton SNPs, though we obtained similar results without such a reduction (not shown). In constructing our final list of candidate hotspots, we retained only those which overlapped one of `sequenceLDhot`'s 'extended hotspot regions', constructed conservatively from windows with a likelihood ratio greater than 10. To improve power in the search for hotspots, we included five additional lower coverage RG genomes in this analysis.

## 2.19 Wavelet Analysis

To put the recombination maps into a suitable time-series format, we used the (log-transformed) cumulative recombination rate across each $\delta t = 250$ bp window. We found that this provided good resolution at high frequencies, with little further improvement using smaller bins. To facilitate a comparison between RAL and RG, we used the maps estimated from sample size $n = 22$ in both populations.

### 2.19.1 Continuous Wavelet Transform

Continuous wavelet transforms are useful for visualization purposes and for feature extraction, and the methods of *wavelet coherence* [79, 31] are based on them. All our plots of wavelet power are therefore based on continuous transforms, using software provided by [31] which convolves the data with the Morlet wavelet (parametrized by a frequency parameter $\omega_0$; we take $\omega_0 = 6$). This choice of wavelet is reasonable because it is simple, widely used, and provides a sensible balance between time and frequency localization.

At large scales, the wavelet transform is influenced by data distant from a given position—possibly even outside the range of the data. The region of the time-frequency domain distorted by the consequent introduction of unwanted edge effects is said to be inside the *cone of influence*, which we define following [79] as the region in which wavelet power for a discontinuity at the edge drops by a factor of $e^{-2}$. Results using the transform inside the cone of influence should be treated with caution.

To assess the significance of regions of the local wavelet power spectrum of high power, we assumed as a background power spectrum that of an autoregressive process of order 1 (AR(1)), whose underlying power spectrum is red noise. This serves as a simple, parametric way of positing an expected power spectrum for a dataset varying about some mean value and allowing for some autocorrelation. The distribution of the observed wavelet power taken with respect to the Morlet wavelet is, for each position and scale, then proportional to a $\chi^2_2$ distribution under this model [79]. The autoregression parameter of the null model was estimated as that which best fit our observed data.

In order to identify regions of correlation of the wavelet transforms for the RAL and RG data, we performed a *wavelet coherence* analysis. Wavelet coherence is a (smoothed) measure of correlation which is computed as a function of both position and scale; we used the formulation given in [31]. To assess the significance of regions of high coherence, we again assume AR(1) models underlying the two datasets, and obtain critical coherence values using Monte Carlo simulation as described in [31] (with 1,000 Monte Carlo samples and 10 scales per octave).

### 2.19.2 Discrete Wavelet Transform

Because the scale index of a continuous wavelet transform varies continuously, coefficients at nearby scales encode similar information and a great deal of the transformed data is superfluous. On the other hand, the discrete wavelet transform provides a decomposition of the data into a minimal number of independent coefficients. It is therefore suitable for modeling purposes, since the transform is constructed so that variation in a signal at one scale is orthogonal to that at a different scale. Within the discrete set of scales chosen, those with important or significant variation can be identified unambiguously. In our linear model analyses we take the discrete wavelet transform based on the Haar wavelet, using methods and `R` scripts provided by [72]. Indeed, the paper by Spencer *et al.* [72] provides an excellent overview of the use of the discrete wavelet transforms in analyzing genomic data, and we refer the interested reader there for further details. Our analysis differs from theirs in several respects: (i) We analyzed five chromosome arms from two populations, giving ten datasets in total compared to their two, (ii) Since our data has much improved SNP density, we binned our data into 250 bp windows rather than 1 kb, giving a fourfold improved resolution, (iii) To control for the influence of local sequence quality, we used quality score information directly rather than read depth.

In addition to wavelet transforming the 250 bp-binned recombination map, we also binned and transformed a number of other genomic features: Diversity was computed as the mean, across pairs of samples within the population, of the fraction of sites that differed between the pair, out of a total of the number of sites for which both samples had data available. Divergence was computed as the diversity between the *D. melanogaster* and *D. simulans* reference sequences, which were available as part of a multiple sequence alignment along with the data from [49]. GC content was computed as the fraction of the total number of sequenced positions in the window (across all samples within the population) that were called as G or C. Gene content was computed for each window as the fraction of the window annotated as exonic; genome annotations were obtained from FlyBase (release 5.45, `http://www.flybase.org` [56]). Sequence quality scores were taken directly from the FASTQ files of the original data. Note that divergence and gene content data are the same for RAL and RG, explaining their identical power spectra in Figure 2.24.

## 2.20 Results

We applied our method to samples from two populations of *D. melanogaster*: Raleigh, USA (RAL) and Gikongoro, Rwanda (RG). The RAL dataset consisted of the genomes (Release 1.0) of 37 inbred lines sequenced at a coverage of $\geq 10\times$ by the Drosophila Population Genomics Project [49] (DPGP, `http://www.dpgp.org/`). The RG dataset comprised 22 genomes (Release 2.0) from haploid embryos sequenced at a coverage of $\geq 25\times$ by the Drosophila Population Genomics Project 2 (DPGP2, `http://www.dpgp.org/dpgp2/DPGP2.html`).

### 2.20.1 Mutation Transition Matrices

We were able to designate the ancestral allele in 1,755,040 of 2,475,674 high quality (quality score $Q \geq 30$) SNPs in the RAL sample (70.9%), and 2,213,312 out of 3,134,295 high quality SNPs in the RG sample (70.6%). These collections of polarized SNPs yielded the following estimates for the mutation transition matrix $\boldsymbol{P}$, with rows and columns ordered as A, C, G, T:

$$\boldsymbol{P}_{\text{RAL}} = \begin{bmatrix} 0.47 & 0.10 & 0.23 & 0.19 \\ 0.27 & 0.00 & 0.14 & 0.59 \\ 0.59 & 0.14 & 0.00 & 0.27 \\ 0.20 & 0.23 & 0.10 & 0.47 \end{bmatrix} \quad \text{and} \quad \boldsymbol{P}_{\text{RG}} = \begin{bmatrix} 0.48 & 0.09 & 0.24 & 0.20 \\ 0.24 & 0.00 & 0.14 & 0.62 \\ 0.62 & 0.14 & 0.00 & 0.24 \\ 0.20 & 0.24 & 0.09 & 0.47 \end{bmatrix}.$$

These results imply that simple diallelic models are inadequate for the *Drosophila* populations. As expected, we see a transition:transversion bias. We also observe a higher overall mutation rate away from C and G nucleotides—this pattern persists even after excluding CpG sites from our analysis (not shown). Indeed, each of the four nucleotides exhibits its own characteristic mutation distribution. There appears to be no significant difference between the transition matrices for the two populations. This is partly explained by the shared history of the two populations: There were 2,990,025 sites for which: (i) data were available in both populations, (ii) two alleles were observed in the combined sample, and (iii) one of the two alleles was assignable as ancestral. Of these, 925,569 (31.0%) were polymorphic in both populations, 800,118 (26.8%) were private to RAL, 1,262,109 (42.2%) were private to RG, and 2,229 (0.1%) were fixed differences.

For simplicity, we used the same mutation transition matrix for all sites in the genome. However, we note that our method can easily handle site-specific mutation transition matrices at no extra computational cost.

### 2.20.2 Accuracy of the Method in the Neutral Case

To assess the accuracy of estimated recombination maps, we carried out an extensive simulation study with various simple recombination patterns, first assuming selective neutrality (the case with natural selection is discussed in the subsequent section).

The simulations assumed a finite-sites, tetra-allelic mutation model, with the mutation transition matrix $\boldsymbol{P}_{\text{RAL}}$ shown above and the population-scaled mutation rate $\theta = 0.008$ per bp. We used these transition matrix and mutation rate in LDhelmet's inference. For LDhat, we used the corresponding effective mutation rate $\theta_{\text{effective}} = 0.006$ per bp (see Section 2.8). Incidentally, we note that 0.006 per bp is the estimated effective mutation rate for the autosomes of RAL lines [49].

Figure 2.3 shows representative examples of LDhelmet's and LDhat's results. As the figure illustrates, our method LDhelmet generally produces recombination maps that are less noisy than that of LDhat's; in particular, LDhelmet produces spurious "spikes" less frequently than does LDhat. To illustrate the impact of the spikes on the total genetic distance, the corresponding cumulative recombination maps comparing LDhelmet and LDhat are shown in Figure 2.4. Additional comparisons between LDhelmet and LDhat can be found in Table 2.2, and SNP statistics of the datasets are listed in Table 2.3.

In general, we observed that LDhelmet is able to identify the location of hotspots reliably. Furthermore, in the scenario considered in the second row of Figure 2.3, the width and height of the hotspot could be estimated very accurately; on average the total rate in the hotspot region could be estimated within 2.5% of the true value.

To test the performance of LDhelmet in a more realistic scenario, we simulated 1 Mb regions each with a substantial amount variation in recombination rate and with a high average rate representative of the interior of the *D. melanogaster* X chromosome. To specify realistic levels of recombination rate variability in these regions, we took as the true recombination map a 1 Mb excerpt from our estimated map for the RAL sample. To specify realistic absolute levels of recombination, we rescaled this map to match the mean (per megabase) recombination rates inferred for the X chromosomes of RAL and of RG. In Figure 2.5, LDhelmet's estimated recombination maps for these two scenarios are illustrated in blue, while the true maps are shown in red. These results demonstrate that, even when the average recombination rate is high, LDhelmet with our chosen block penalty in the rjMCMC is able to capture the pattern of fine-scale variation rather well. However, we note that in the top plot of Figure 2.5, in which case the true average rate is $\rho = 21$ per kb, the estimated map tends to be slightly more variable than the true map. In contrast, if the true average recombination rate is substantially higher, as in the bottom plot of Figure 2.5 with the true average rate of 170 per kb but otherwise the same pattern of variation, the estimated map tends to be somewhat smoother than the true map. Clearly, there is no single block penalty value that is universally optimal in all cases, but the value we have chosen seems to yield reasonable results for *D. melanogaster*.

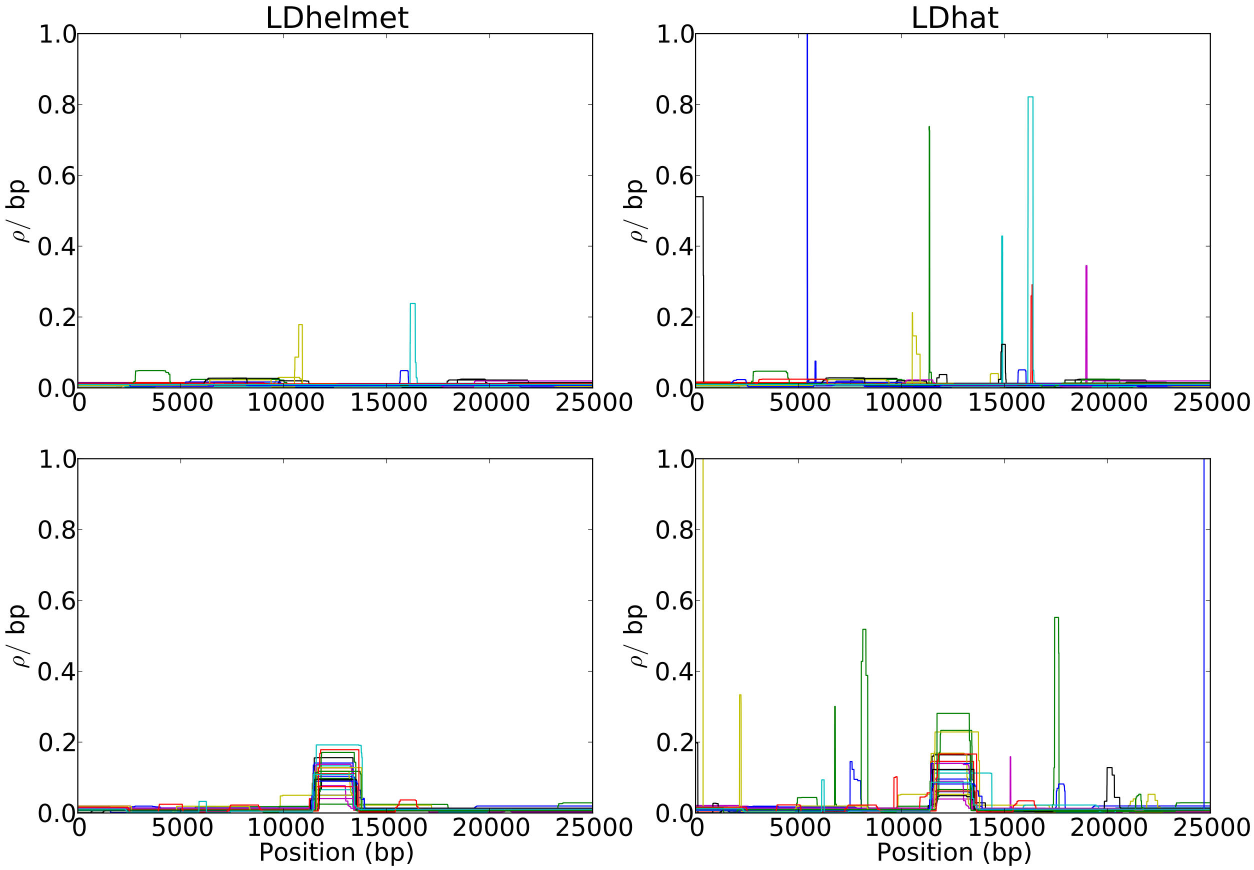

Figure 2.3: **Comparison of the results of** `LDhelmet` **and** `LDhat` **for 25 datasets simulated under neutrality.** In each plot, different colors represent the results for different datasets. The left and right columns show the estimated recombination maps of `LDhelmet` and `LDhat`, respectively, using the same block penalty of 50. Our method `LDhelmet` generally produces less noisy estimates than that produced by `LDhat`. (First Row) Each dataset was simulated with a constant recombination rate of 0.01 per bp. (Second Row) Each dataset was simulated with a hotspot of width 2 kb starting at location 11 kb. The background recombination rate was 0.01 per bp, while the hotspot intensity was 10× the background rate, i.e., 0.1 per bp. The maps are shown in their entirety, including potential edge effects.

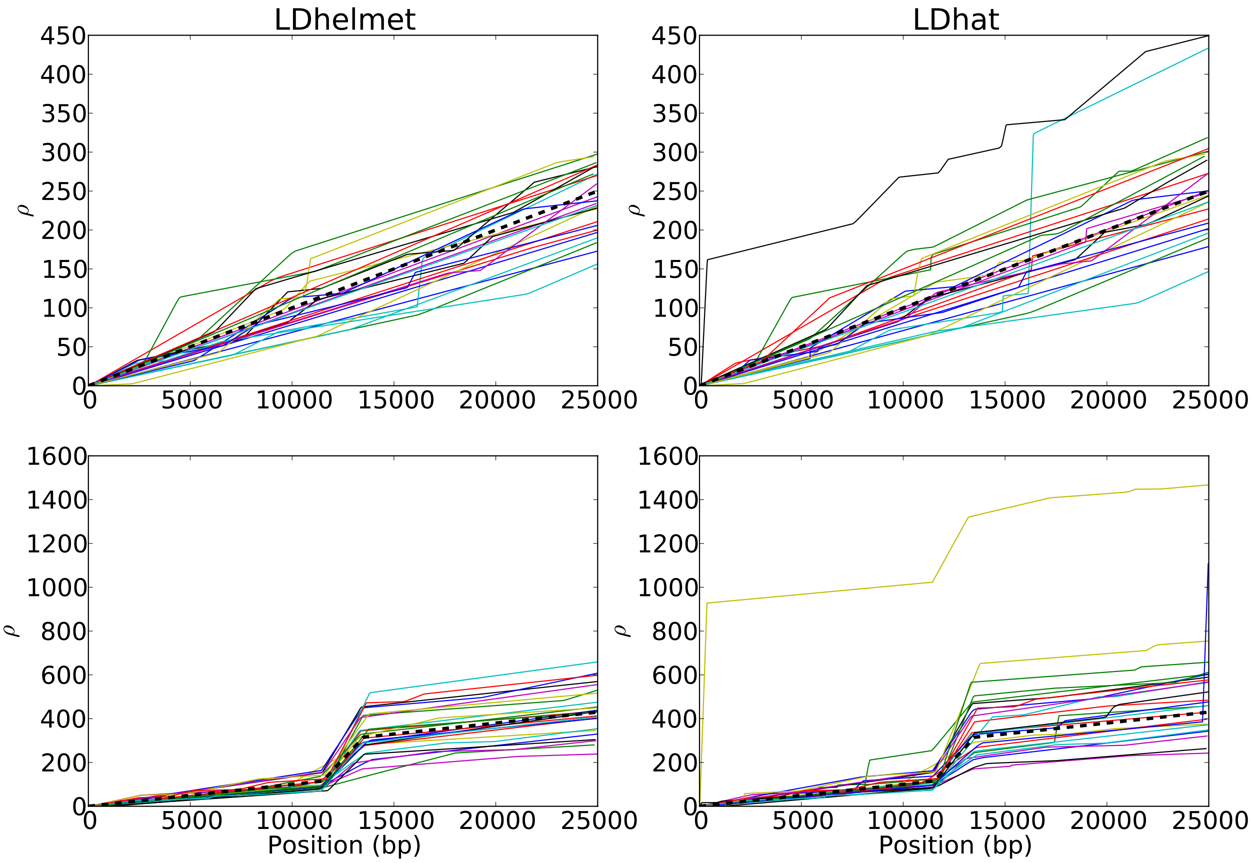

Figure 2.4: **Comparison of the cumulative recombination maps of** **`LDhelmet`** **and** **`LDhat`** **for 25 datasets simulated under neutrality** In each plot, different colors represent the cumulative recombination maps for different datasets. The datasets in these plots correspond to the same datasets used in Figure 2.3. The thick dashed line indicates the true cumulative recombination map for the given recombination landscape. The left and right columns show the estimated recombination maps of `LDhelmet` and `LDhat`, respectively, using the same block penalty of 50. (First Row) Each dataset was simulated with a constant recombination rate of 0.01 per bp. (Second Row) Each dataset was simulated with a hotspot of width 2 kb starting at location 11 kb. The background recombination rate was 0.01 per bp, while the hotspot intensity was 10× the background rate, i.e., 0.1 per bp. The cumulative maps are shown in their entirety, including potential edge effects.

Table 2.2: **Summary of comparison between `LDhelmet` and `LDhat` in the neutral case.** Based on 100 simulated datasets for a 25 kb region. "No Hotspot" corresponds to the case of a constant recombination map, whereas "Hotspot 10×" corresponds to the case with a 2 kb wide hotspot situated at the center of the region. The first row shows the regional average of $\rho$ obtained by `LDhelmet` and `LDhat`, averaged over the 100 datasets. The second row shows the total rate in the hotspot region, averaged over the datasets. The third row shows the percentage of datasets for which the estimate had at least one false peak with height $\geq 5$ times the background rate. The fourth row shows the percentage of datasets for which the estimate had at least one false peak with height $\geq 10$ times the background rate. The fifth row shows the percentage absolute error of the estimated $\rho$ average outside the hotspot region from the true $\rho$ average outside the hotspot region. The true $\rho$ average outside the hotspot region is $\rho = 0.01$/bp. To account for edge effects, 2.5 kb from each end of the map were removed prior to computing the statistics.

| | No Hotspot | | | Hotspot 10× | | |
|---|---|---|---|---|---|---|
| Measure of Accuracy | True Value | LDhelmet | LDhat | True Value | LDhelmet | LDhat |
| $\rho$ average (per bp) | 0.01 | 0.0097 | 0.0109 | 0.0172 | 0.0184 | 0.0203 |
| Total hotspot area | 20.0 | 19.0 | 20.3 | 200.0 | 195.2 | 210.0 |
| % with false peak $\geq 5\times$ | | 5% | 30% | | 4% | 30% |
| % with false peak $\geq 10\times$ | | 2% | 21% | | 4% | 21% |
| % abs. error outside hotspot | | 14% | 23% | | 15% | 20% |

Table 2.3: **SNP densities (per kb) of neutral and single-sweep simulations.** The mean, minimum, maximum and standard deviation of the SNP density for the datasets used in Tables 2.2 and 2.4. The simulations assumed a finite-sites, tetra-allelic mutation model, with mutation matrix $\boldsymbol{P}_{\text{RAL}}$ and $\theta = 0.008$, which is the effective population-scaled mutation rate adjusted for $\boldsymbol{P}_{\text{RAL}}$ (see Section 2.8).

| | Neutral | | Single-Sweep Model | |
|---|---|---|---|---|
| | No Hotspot | Hotspot 10× | No Hotspot | Hotspot 10× |
| Mean | 21.82 | 21.68 | 18.15 | 18.38 |
| Min | 18.32 | 17.40 | 14.84 | 14.68 |
| Max | 26.12 | 25.72 | 24.08 | 22.76 |
| Std dev | 1.71 | 1.38 | 1.64 | 0.61 |

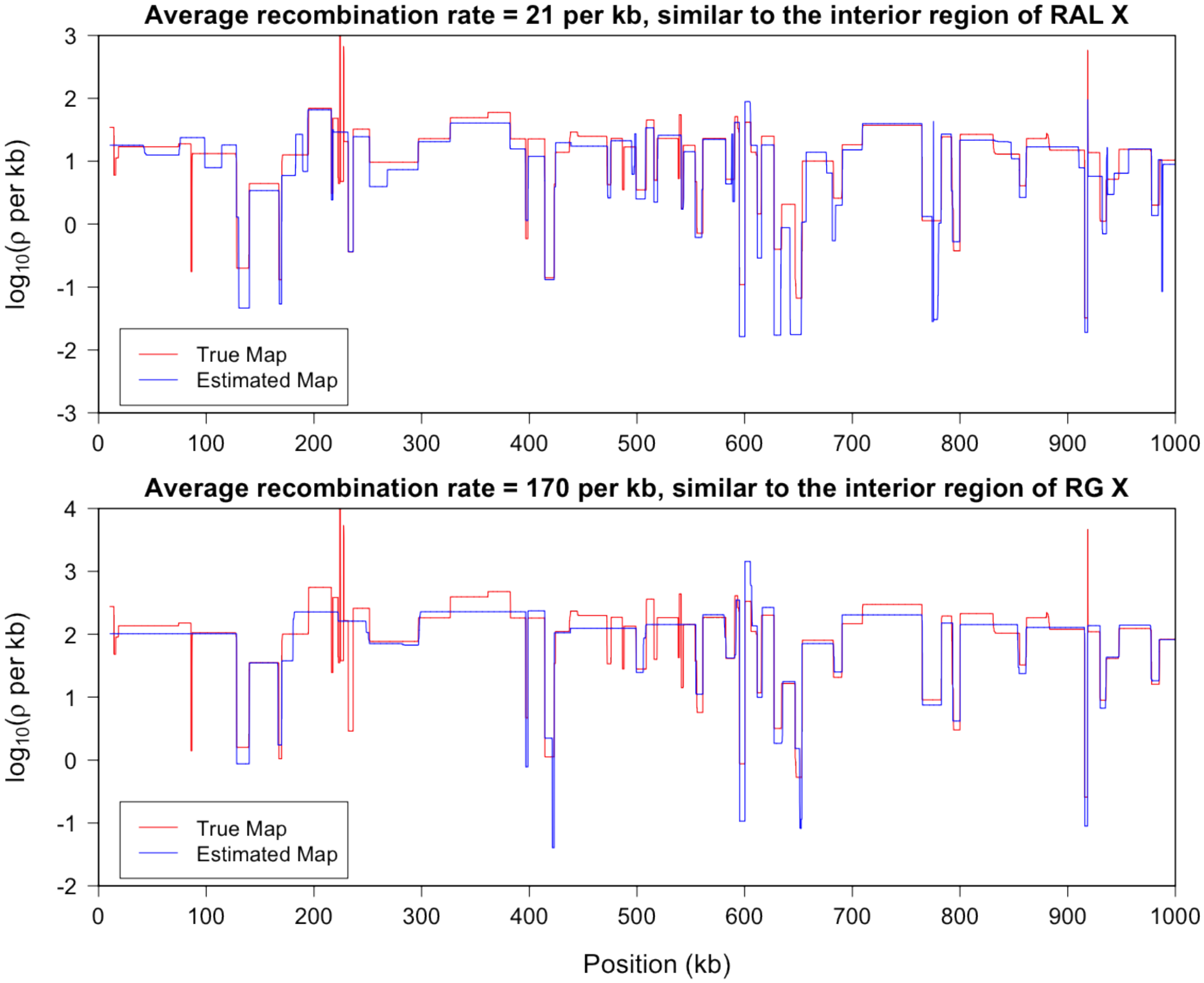

Figure 2.5: `LDhelmet` **results on simulations with realistic variable recombination rates.** In each study, the program `MaCS` [17] was used to simulate data, with sample size 22, for a 1 Mb region with the variable recombination map shown in red. (We used $\theta = 0.008$; output was postprocessed to incorporate an empirical tetra-allelic mutation model.) Estimated recombination maps are shown in blue. The same block penalty of 50 was used in both cases. (Top) The average recombination rate for the region is about 21 per kb, representative of the interior of the North American X. (Bottom) The average recombination rate for the region is 8× higher than the above case, representative of the interior of the African X.

### 2.20.3 Impact of Positive Selection on Inference

It has been previously shown [45, 68, 73] that hitchhiking can induce seemingly similar patterns of linkage disequilibrium as that created by recombination hotspots, while McVean [57] has argued that the precise signatures of selective sweeps and hotspots actually differ. To test the robustness of our method to natural selection, we simulated data under various scenarios with positive selection and recombination rate variation, and assessed the impact on our estimates of recombination rates. We generated data using a range of values for the selection strength and fixation time. See Section 2.16.2 for details of the simulation setup.

The results of `LDhelmet` and `LDhat` for a few cases are shown in Figure 2.6; each plot shows the results for 25 simulated datasets illustrated in 25 different colors. The corresponding cumulative recombination maps are shown in Figure 2.7. For both methods, the estimated recombination maps are in general noisier than that for the neutral case (c.f., Figure 2.3), though `LDhelmet` is still more robust than `LDhat`. As can be seen in Figure 2.6, `LDhat` tends to produce false inference of elevated recombination rates near the selected site more frequently than does `LDhelmet`. A more detailed comparison is provided in Table 2.4 and SNP statistics of the datasets are listed in Table 2.3. Overall, although strong positive selection causes more noise and fluctuations in our estimates, it does not seem to produce a strong bias to the extent that would consistently lead to false inference of recombination hotspots.

The noise in our estimates of the recombination rate in the presence of selection depends on several factors. Specifically, we observed that the accuracy of our estimates decreases as the selection strength increases, whereas the accuracy improves as the distance between the selected site and the region of estimation increases. Furthermore, the more recent the time of fixation, the noisier are the estimates.

In addition to the case of a single, recent selective sweep, we also assessed the impact of recurrent selective sweeps [2, 43] on the estimation of recombination rates. Assuming that beneficial mutations fixate randomly at a given rate, we simulated three different sets of datasets with a background recombination rate of 10 per kb, as detailed in Section 2.16.2. The degree to which recurrent sweeps reduce diversity in each model is summarized in Table 2.5. In model RS3, which has infrequent but strong sweeps, the mean number of SNPs reduces by more than a factor of 8 relative to the neutral model. Such a drastic drop in diversity significantly reduces the ability to perform accurate statistical inference of recombination. To infer the location of a recombination hotspot, for example, at least a few SNPs must be present in the hotspot and near its edges.

The results of recombination rate estimation under recurrent sweep models are summarized in Tables 2.6 and 2.7. Compared to a single sweep, recurrent selective sweeps tend to decrease the accuracy of recombination rate estimates more noticeably. Furthermore, infrequent but strong selective sweeps (model RS3) have more severe impact on the accuracy than do frequent but weaker selective sweeps (model RS1). As discussed above and can be seen in Table 2.7, detecting recombination hotspots in model RS3 would pose a great challenge. Overall, `LDhelmet` generally underestimates the recombination rate in the presence of selection, suggesting that it is unlikely to produce spurious hotspots because of selection.

Table 2.4: **Summary of comparison between `LDhelmet` and `LDhat` in the case of single selective sweep.** Based on 100 simulated datasets for a 25 kb region. For each dataset, a selected site was placed at position 5 kb and the population-scaled selection coefficient was set to 1000. The fixation time of the selected site was 0.01 coalescent units in the past. The column and the row labels are the same as in Table 2.2. As for Table 2.2, 2.5 kb from each end of the map were removed prior to computing the statistics to account for edge effects.

| | No Hotspot | | | Hotspot 10× | | |
|---|---|---|---|---|---|---|
| Measure of Accuracy | True Value | LDhelmet | LDhat | True Value | LDhelmet | LDhat |
| $\rho$ average (per bp) | 0.01 | 0.0079 | 0.0108 | 0.0172 | 0.0162 | 0.0220 |
| Total hotspot area | 20.0 | 14.7 | 15.4 | 200.0 | 169.8 | 224.6 |
| % with false peak $\geq 5\times$ | | 10% | 42% | | 8% | 34% |
| % with false peak $\geq 10\times$ | | 6% | 39% | | 5% | 24% |
| % abs. error outside hotspot | | 39% | 58% | | 30% | 56% |

Table 2.5: **SNP densities (per kb) of recurrent-sweep and demography simulations.** The statistics for each selection or demography scenario are merged over the three recombination landscapes (i.e., no hotspot, hotspot 10× and hotspot 50×). The simulations use $\theta_{\text{RAL}}$ and $\boldsymbol{P}_{\text{RAL}}$ as parameters. The third column shows the SNP density per kb across the hundred datasets, and the fourth column shows the standard deviation. For the definitions of the scenario names, refer to Section 2.16.2 and Section 2.17 of the main text. "Control" refers to a control dataset with constant population size and no selection.

| Simulation Type | Model | Mean | Std dev |
|---|---|---|---|
| Recurrent Sweeps | RS1 | 18.22 | 1.66 |
| | RS2 | 4.10 | 1.05 |
| | RS3 | 2.71 | 1.24 |
| Demography | G1 | 12.86 | 1.07 |
| | G2 | 15.85 | 1.24 |
| | B1 | 13.84 | 2.78 |
| | B2 | 5.53 | 2.14 |
| Neutral | Control | 22.51 | 1.49 |

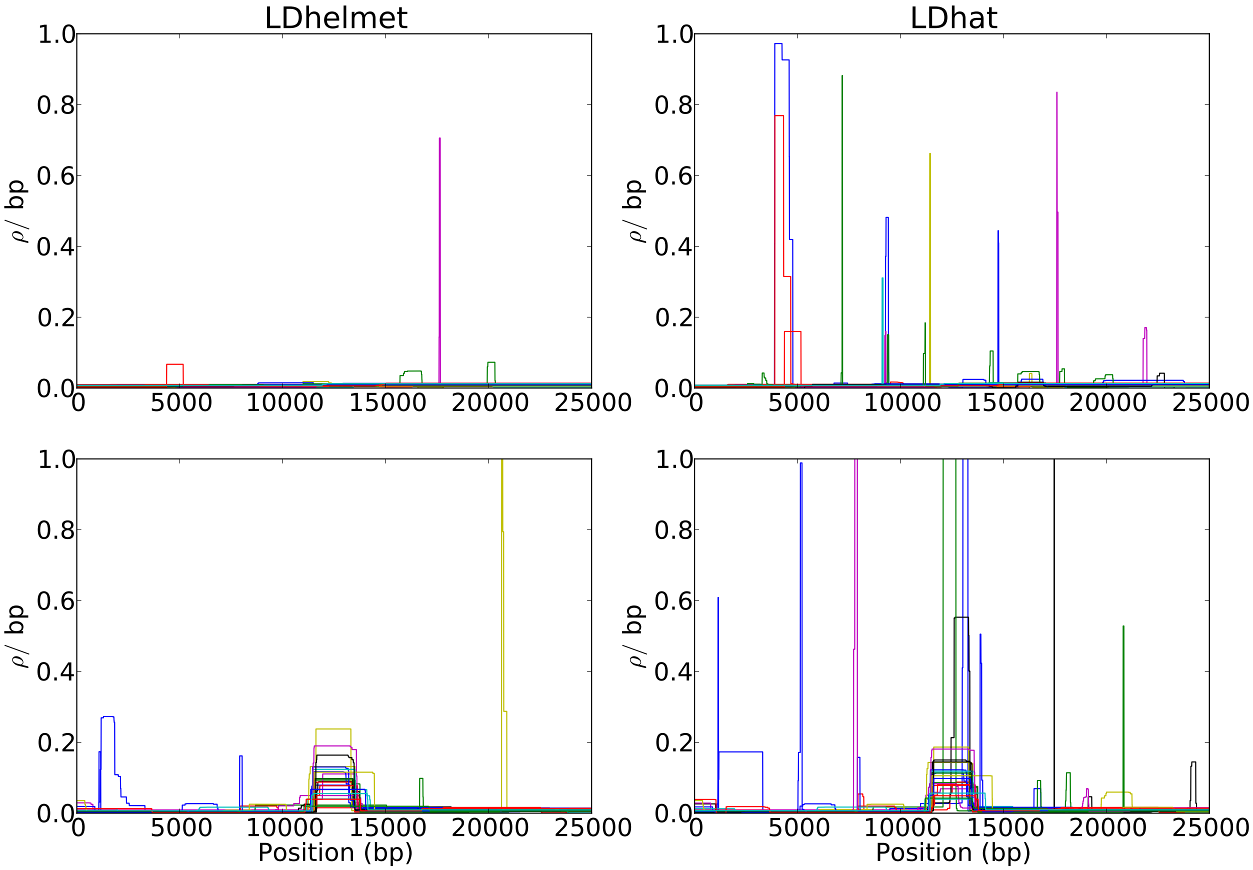

Figure 2.6: **Comparison of the results of** **`LDhelmet`** **and** **`LDhat`** **for 25 datasets simulated under strong positive selection.** In each plot, different colors represent the results for different datasets. The left and right columns show the estimated recombination maps of `LDhelmet` and `LDhat`, respectively, using the same block penalty of 50. In each simulation, the selected site was placed at position 5 kb and the population-scaled selection coefficient was set to 1000. The fixation time of the selected site was 0.01 coalescent unit in the past. Although the estimated recombination maps are in general noisier than that for the neutral case (c.f., Figure 2.3), `LDhelmet` is more robust than `LDhat`. As illustrated in the plots, `LDhat` produces false inference of elevated recombination rates near the selected site more frequently than does `LDhelmet`. The same scenarios of recombination patterns as in Figure 2.3 were considered: (First Row) with a constant recombination rate of 0.01 per bp, and (Second Row) with a hotspot of width 2 kb starting at location 11.5 kb. The background recombination rate was 0.01 per bp, while the hotspot intensity was 10× the background rate, i.e., 0.1 per bp. The maps are shown in their entirety, including potential edge effects.

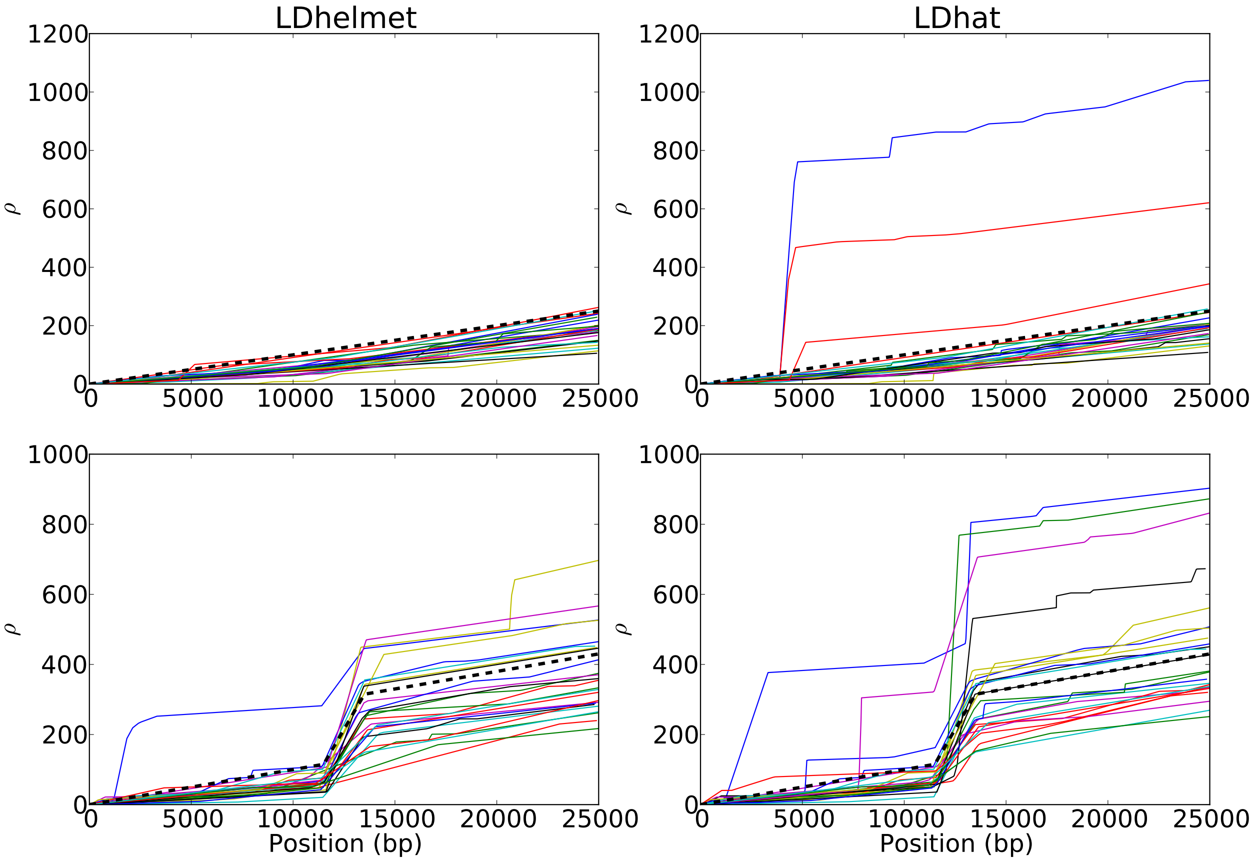

Figure 2.7: **Comparison of the cumulative recombination maps of** `LDhelmet` **and** `LDhat` **for 25 datasets simulated under strong positive selection.** In each plot, different colors represent the results for different datasets. The datasets in these plots correspond to the same datasets used in Figure 2.6. The thick dashed line indicates the true cumulative recombination map for the given recombination landscape. The left and right columns show the estimated recombination maps of `LDhelmet` and `LDhat`, respectively, using the same block penalty of 50. In each simulation, the selected site was placed at position 5 kb and the population-scaled selection coefficient was set to 1000. The fixation time of the selected site was 0.01 coalescent units in the past. The same scenarios of recombination patterns as in Figure 2.3 were considered: (First Row) with a constant recombination rate of 0.01 per bp, and (Second Row) with a hotspot of width 2 kb starting at location 11.5 kb. The background recombination rate was 0.01 per bp, while the hotspot intensity was 10× the background rate, i.e., 0.1 per bp. The cumulative maps are shown in their entirety, including potential edge effects.

Table 2.6: **Average recombination rates for recurrent sweeps simulations.** The accuracy of the recombination rate estimate for model RS3, containing infrequent but strong selective sweeps, was considerably worse than that for model RS1, containing frequent but weaker selective sweeps. The mean number of SNPs in model RS3 was a factor of 8 less than that in the selectively neutral "Control" model, thus reducing the ability to perform accurate statistical inference of recombination. See Section 2.16.2 for the details of the models. For each recombination landscape, the median estimated average recombination rate is shown in the left column ("est.") and the percent error is shown in the right ("% err."). The true average recombination rate for each recombination landscape is shown in parenthesis.

| | No Hotspot | | Hotspot 10× | | Hotspot 50× | |
|---|---|---|---|---|---|---|
| | (10 per kb) | | (17.2 per kb) | | (49.2 per kb) | |
| Model | est. | % err. | est. | % err. | est. | % err. |
| RS1 | 8.5 | −15.0 | 15.6 | −9.3 | 44.9 | −8.7 |
| RS2 | 4.4 | −56.0 | 8.6 | −50.0 | 45.0 | −8.5 |
| RS3 | 0.9 | −91.0 | 1.3 | −92.4 | 2.3 | −95.3 |
| Control | 9.3 | −7.0 | 16.4 | −4.7 | 53.9 | 9.6 |

Table 2.7: **Hotspot areas for recurrent sweeps simulations.** For each recombination landscape, the median estimated hotspot area is shown in the left column ("est.") and the percent error is shown in the right ("% err."). The true hotspot area for each recombination landscape is shown in parenthesis. "Control" refers to a neutral model. See Section 2.16.2 for the details of the models and Table 2.6 for related results.

| | No Hotspot | | Hotspot 10× | | Hotspot 50× | |
|---|---|---|---|---|---|---|
| | (20) | | (200) | | (1000) | |
| Model | est. | % err. | est. | % err. | est. | % err. |
| RS1 | 16.6 | −16.8 | 179.8 | −10.1 | 889.6 | −11.0 |
| RS2 | 8.9 | −55.5 | 38.2 | −80.9 | 773.0 | −22.7 |
| RS3 | 1.7 | −91.4 | 2.6 | −98.7 | 4.5 | −99.5 |
| Control | 18.2 | −9.0 | 183.5 | −8.3 | 1100.0 | 10.0 |

### 2.20.4 Impact of Demography on Inference

We also tested our method on datasets simulated under a variety of demographic scenarios. Specifically, the demographic models we considered are those proposed by Haddrill *et al.* [32], and by Thornton & Andolfatto [78], comprising two exponential growth models and two bottleneck models. As in the neutral simulations, we assumed a finite-sites, tetra-allelic mutation model, with the mutation transition matrix $\boldsymbol{P}_{\mathrm{RAL}}$ and the mutation rate $\theta = 0.008$ per bp. See Section 2.17 for details on the other parameters used in the simulations.

Tables 2.8 and 2.9 show the results of recombination rate estimation in this simulation study. Although the estimates are clearly less accurate compared to the case with constant population size, they are reasonably accurate in most cases. Note that the overall trend is to underestimate the true rates, in some cases only slightly.

As in the case of recurrent selective sweeps, demography may decrease diversity, thus hindering statistical inference of recombination. Table 2.5 includes the SNP statistics for the demography models we considered. In model B2, which involves a very recent bottleneck, a reduction in diversity by about a factor of 4 was observed, partly explaining the particularly poor estimates of the recombination rate. Table 2.10 shows that the average SNP density of the *D. melanogaster* data; note that the average SNP density of each chromosome is substantially greater than the SNP density observed in simulation model B2.

Table 2.8: **Average recombination rates for demography simulations.** Here, "Control" refers to a neutral model with constant population size. Model B2 involved a very recent bottleneck, and we observed a reduction in diversity by about a factor of 4 relative to the Control model. This reduction in diversity partly explains the particularly poor estimates of the recombination rate for model B2. The estimates for the other models are reasonably accurate, although they are clearly nosier compared to that for the Control model. See Section 2.17 for the details of the models. For each recombination landscape, the median estimated average recombination rate is shown in the left column ("est.") and the percent error is shown in the right ("% err."). The true average recombination rate for each recombination landscape is shown in parenthesis.

| | No Hotspot (10 per kb) | | Hotspot 10× (17.2 per kb) | | Hotspot 50× (49.2 per kb) | |
|---|---|---|---|---|---|---|
| Model | est. | % err. | est. | % err. | est. | % err. |
| G1 | 5.8 | −42.0 | 10.1 | −41.3 | 38.6 | −21.5 |
| G2 | 7.7 | −23.0 | 12.8 | −25.6 | 52.2 | 6.1 |
| B1 | 7.2 | −28.0 | 10.2 | −40.7 | 28.8 | −41.5 |
| B2 | 1.2 | −88.0 | 3.9 | −77.3 | 20.0 | −59.3 |
| Control | 9.3 | −7.0 | 16.4 | −4.7 | 53.9 | 9.6 |

Table 2.9: **Hotspot areas for demography simulations.** For each recombination landscape, the median estimated hotspot area is shown in the left column ("est.") and the percent error is shown in the right ("% err."). The true hotspot area for each recombination landscape is shown in parenthesis. "Control" refers to a neutral model with constant population size. See Section 2.17 for the details of the models and Table 2.8 for related results.

| | No Hotspot (20) | | Hotspot 10× (200) | | Hotspot 50× (1000) | |
|---|---|---|---|---|---|---|
| Model | est. | % err. | est. | % err. | est. | % err. |
| G1 | 11.6 | −41.9 | 116.6 | −41.7 | 752.0 | −24.8 |
| G2 | 15.2 | −23.8 | 131.9 | −34.1 | 1032.6 | 3.3 |
| B1 | 14.2 | −29.1 | 25.6 | −87.2 | 471.0 | −52.9 |
| B2 | 1.6 | −92.2 | 31.0 | −84.5 | 205.2 | −79.5 |
| Control | 18.2 | −9.0 | 183.5 | −8.3 | 1100.0 | 10.0 |

Table 2.10: **SNP densities (per kb) of the real *Drosophila* data.**

| Chromosome Arm | RAL | RG |
|---|---|---|
| 2L | 24.54 | 25.49 |
| 2R | 22.56 | 24.21 |
| 3L | 22.29 | 25.20 |
| 3R | 19.77 | 20.79 |
| X | 14.92 | 28.15 |

### 2.20.5 Population-specific Average Recombination Rates

The population-specific average recombination rate for each major chromosome arm is summarized in Table 2.11, which shows that the average rate for the African (RG) population is higher than that for the North American (RAL) population. This difference could be explained partially, but not entirely, by a difference in population size. Note that the average recombination rate in the X chromosome appears to be higher than that in the autosomes, much more so in RG than in RAL. Table 2.11 shows the ratio of the average recombination rate of RG to that of RAL for each chromosome arm. Although the ratio is more or less consistent for the autosomes, the ratio for the X chromosome is significantly higher. Hence, a difference in population size could explain the higher recombination rate estimates in RG for the autosomes, but it does not explain the significant increase in the recombination rate for the X chromosome of RG over RAL. Furthermore, for RAL, that the observed average recombination rate of the X chromosome is higher than that of autosomes is unexpected given that an excess of LD is observed on the X chromosome of this population [49, 54]. In both populations, arm 3R has a notably reduced recombination rate compared to the other arms. This reduction is more pronounced in RG than in RAL, which could be partly explained by the fact that, in African populations, arm 3R has the largest number of common inversions [3].

To study the effect of sample size on the estimation of recombination rates, we subsampled a 2 Mb excerpt of chromosome arm 2L from each population over several repeated trials. We performed the subsampling on an excerpt rather than the entire genome for computational reasons. The averages of the estimates are shown in Table 2.12. Despite a slight increase in the estimate as sample size increases, the effect is small and appears to diminish with increasing sample size. We also analyzed the whole-genome RAL dataset down-sampled to match the sample size (i.e., 22) of RG. As Table 2.11 shows, the genome-wide average estimates produced using 22 genomes of RAL were only slightly lower than those produced using all 37 genomes. Encouragingly, our estimate (107.3 per kb) of the recombination rate for the X chromosome of RG is similar to the previous estimates for other African populations obtained using a different method: Haddrill *et al.* [32] estimated $84, 89$, and 47 per kb for the X recombination rate in three African populations.

To assess the effect of SNP density, we thinned the SNPs on chromosome arm 2L and chromosome X of the RG dataset to the corresponding SNP densities of RAL, and performed inference on the resulting data. The results summarized in Table 2.13 show that although SNP density indeed influences the ability to estimate recombination rates, the impact is not nearly large enough to account for the difference between the observed recombination rates of RAL and RG on the X chromosome.

Finally, as there exist several inversions in *D. melanogaster*, we analyzed regions of inversion excluding individuals known to carry the inversion [20]. The comparison of excluding individuals with inversions and the original analysis is shown in Table 2.14. Note that for each inversion, only a small number of individuals carry it. We found that excluding the individuals with inversions did not significantly affect the recombination rate estimates.

Table 2.11: **The average recombination rate for each major chromosome arm.** Note that RG has higher recombination rates than that of RAL. This difference could be explained partially, but not entirely, by a difference in population size. In RG, the average recombination rate of X is substantially higher than that of the autosomes. In both populations, arm 3R has a notably lower recombination rate than do the other arms. We also analyzed a smaller RAL dataset down-sampled to match the sample size of RG. The numbers in parentheses denote sample sizes.

| | $\rho$ per kb | | | Ratio | |
|---|---|---|---|---|---|
| Chromosome arm | RAL(37) | RAL(22) | RG(22) | RG(22):RAL(37) | RG(22):RAL(22) |
| 2L | 13.3 | 12.4 | 33.2 | 2.5 | 2.7 |
| 2R | 13.4 | 12.4 | 34.5 | 2.6 | 2.8 |
| 3L | 13.4 | 12.1 | 44.9 | 3.4 | 3.7 |
| 3R | 9.6 | 8.1 | 17.8 | 1.9 | 2.2 |
| X | 14.8 | 13.4 | 107.3 | 7.3 | 8.0 |

Table 2.12: **Subsampling of real data.** To assess the effect of subsampling individuals, we subsampled a 2 Mb excerpt from chromosome arm 2L for both the RAL and RG datasets. We performed subsampling four times, and each row is the average of the four subsampled datasets. The column labeled $n$ is the number of individuals in each subsample. The percentiles are given in the three rightmost columns. The results show that sample size has a slight positive bias, but does not impact estimates greatly.

| | | Percentile ($\rho$ per kb) | | |
|---|---|---|---|---|
| | $n$ | 2.5% | 50% | 97.5% |
| RAL | 17 | 6.1 | 6.2 | 6.5 |
| | 27 | 7.2 | 7.3 | 7.4 |
| | 37 | 7.8 | 7.8 | 7.9 |
| RG | 12 | 8.1 | 8.4 | 9.2 |
| | 17 | 9.0 | 9.0 | 9.2 |
| | 22 | 9.2 | 9.3 | 9.4 |

Table 2.13: **Thinned SNPs on RG dataset.** To assess the effect of SNP density on the recombination rate inference, we thinned the SNPs on chromosome arm 2L and chromosome X of RG to the SNP density of RAL. The 2.5%, 50% and 97.5% percentiles are shown for estimates. The number of SNPs in the original dataset and in the thinned dataset are shown in the fourth column. For chromosome arm 2L, the change in SNP density is negligible. For chromosome X, the difference in SNP density is significant. The results show that SNP density impacts the estimate, but not to the extent of the difference observed between RAL and RG on chromosome X.

| | | | | Percentile ($\rho$ per kb) | | |
|---|---|---|---|---|---|---|
| Dataset | Arm | Type | # SNPs | 2.5% | 50% | 97.5% |
| RG | 2L | Original | 586476 | 33.0 | 35.9 | 39.4 |
| | | Thinned | 564673 | 32.5 | 35.5 | 38.9 |
| | X | Original | 631205 | 110.0 | 121.4 | 134.1 |
| | | Thinned | 334647 | 97.5 | 106.8 | 117.4 |

Table 2.14: **Exclusion of individuals with inversions.** To assess the effect of inversions on the recombination rate estimate, we excluded individuals known to carry the given inversion, and performed inference on the remaining sample. 0.5 Mb was added to both ends of the region to eliminate possible edge effects.The $\rho$ average is over the inversion region only. The column labeled **Original** gives the estimate using the entire sample. The column labeled **Excluded** gives the estimate excluding the individuals with the given inversion. The inversion region length and the number of individuals with the inversion are provided in the rightmost two columns.

| | | | **Original** | **Excluded** | Inversion | # with |
|---|---|---|---|---|---|---|
| Dataset | Arm | Inversion | $\rho$ per kb | $\rho$ per kb | length (Mb) | inversion |
| RAL | 2L | 2Lt | 16.97 | 16.45 | 10.9 | 3 |
| | 2R | 2RNS | 17.34 | 16.66 | 4.9 | 2 |
| | 3R | 3RK | 11.80 | 11.39 | 14.4 | 1 |
| | 3R | 3RMO | 12.51 | 14.56 | 14.6 | 7 |
| | 3R | 3RP | 12.49 | 11.35 | 8.3 | 1 |
| RG | 2L | 2Lt | 54.44 | 50.80 | 10.9 | 2 |
| | 2R | 2RNS | 53.93 | 50.81 | 4.9 | 1 |
| | 3R | 3RP | 22.44 | 17.24 | 8.3 | 4 |
| | X | 1Be | 106.26 | 103.21 | 1.8 | 3 |

### 2.20.6 Comparison with Experimental Genetic Maps

`LDhelmet`'s fine-scale recombination maps for RAL and RG are illustrated in Figure 2.8; files containing the corresponding numerical values are publicly available. To assess the accuracy of our recombination estimates obtained via statistical analysis of population genetic variation data, we compared them to genetic maps obtained experimentally.

Singh *et al.*[71] examined the fine-scale recombination rate variation over a 1.2 Mb region of the *D. melanogaster* X chromosome using a genetic mapping approach, by crossing an African line with a line presumably of North American origin (a cross between two lines from Bloomington Drosophila Stock Center). For their experiment, Singh *et al.* genotyped 8 SNPs and identified two flanking genes, *white* and *echinus*, with visible phenotypes. They found statistically significant heterogeneity in this region, with differences in rate up to 3.5-fold. In Figure 2.9, estimates from `LDhelmet` for both the RAL and RG samples are shown, along with the genetic map from [71]. Both estimates from `LDhelmet` mostly fall within the 95% confidence intervals of the empirical estimate, with the exception of the outermost intervals. The three maps share the same overall shape, including the location of the highest peak. We find 4.5-fold variation in the RG estimate, which is comparable to the 3.5-fold variation obtained by Singh *et al.* The high correlation among the three maps give us confidence that our estimates from the statistical analysis of population genetic data accurately represent the true underlying recombination map.

In a second study, we compared our chromosome-wide recombination estimates with a consensus genetic map for each chromosome arm based on data hosted at the FlyBase website (`http://www.flybase.org` [56]). To facilitate a comparison with this map, resolution of which is roughly 200 kb, we binned our data into the same cytogenetic subdivisions [49] and LOESS-smoothed the results, with a span of 15%; a correspondingly LOESS-smoothed version of the FlyBase data was kindly provided to us by C.H. Langley. A comparison of the maps is shown in Figure 2.10; evidently, the three estimates show broad agreement, each capturing key features such as the spike in recombination near position 10 Mb on arm 2L, as well as a series of dramatic changes in recombination rate across chromosome X. When the recombination map for RAL is regressed on the FlyBase maps, the coefficient of determination, or proportion of variability explained by the simple linear regression model, is $R^2 = 0.54, 0.57, 0.37, 0.53$ and 0.50 for chromosome arms 2L, 2R, 3L, 3R, and X, respectively; the corresponding values for RG are $R^2 = 0.55, 0.63, 0.45, 0.42,$ and 0.41. These correlations are lower than those seen in a comparison of statistically- versus experimentally-derived maps in humans (e.g. $R^2 = 0.97$ [62]), though in that case the experimental data from pedigrees were of higher quality. As noted by Langley *et al.* [49], data on which the FlyBase map is based is highly edited and based on heterogeneous experimental conditions with sometimes conflicting results.

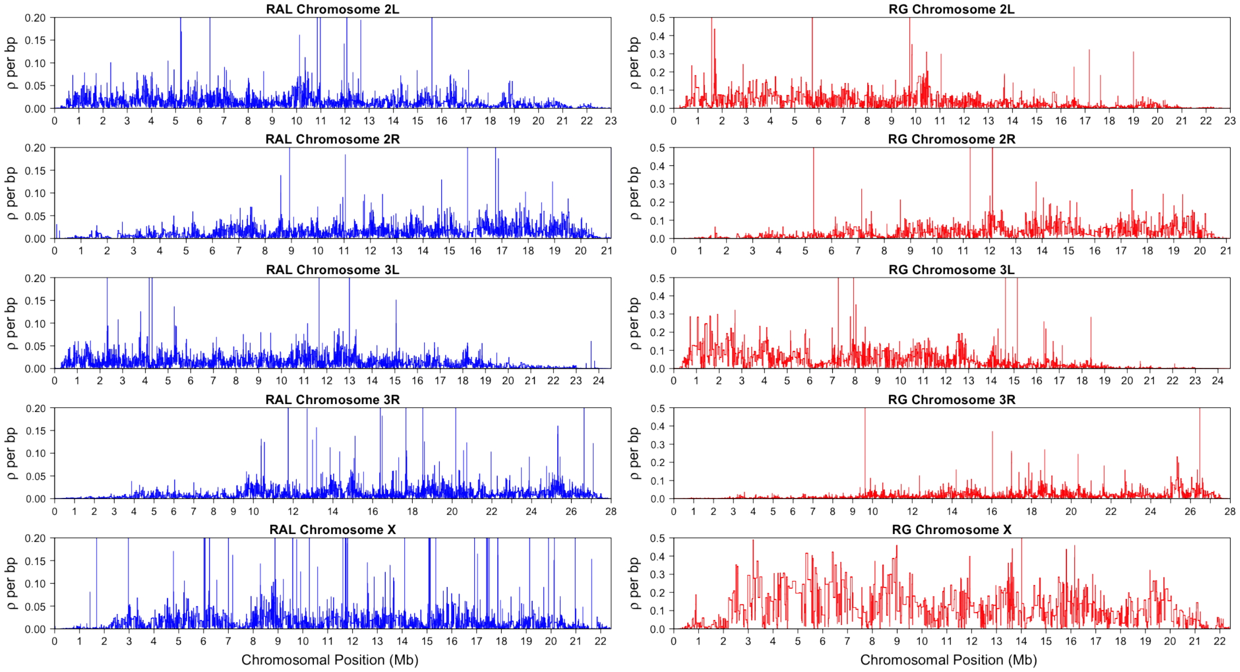

Figure 2.8: `LDhelmet`**'s estimated fine-scale recombination maps for RAL and RG populations of *D. melanogaster*.** The North American sample (RAL) comprised 37 genomes, while the African sample (RG) comprised 22 genomes.

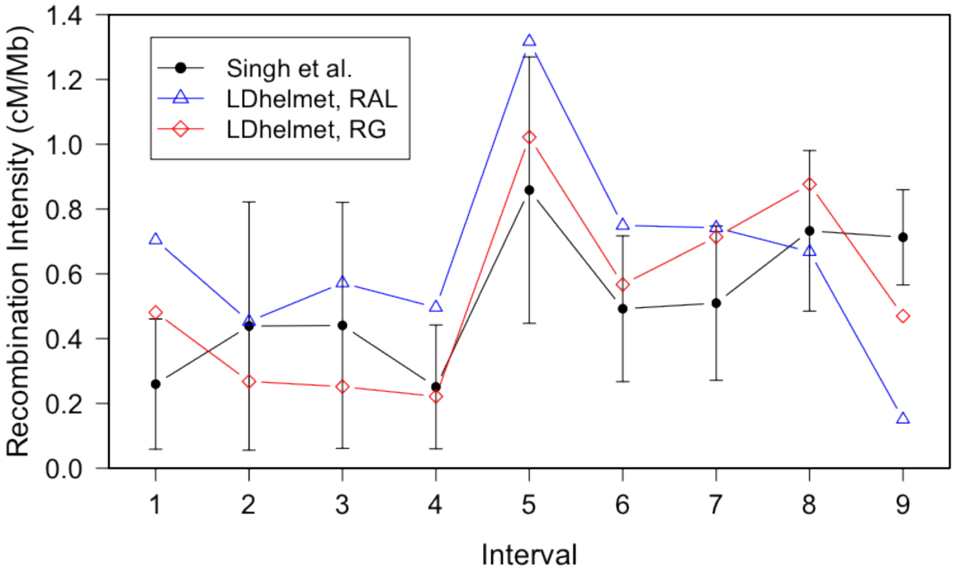

Figure 2.9: **Comparison of `LDhelmet` estimates to the empirical genetic map of Singh *et al.*[71].** The experimental genetic map of Singh *et al.*[71] is shown in black with 95% confidence intervals. The `LDhelmet` estimate for the RAL sample is shown in blue, while the estimate for the RG sample is shown in red. The `LDhelmet` estimates were converted into units of cM/Mb by normalizing them to have the same total genetic distance as the empirical map for the region. The three maps demonstrate high correlation, especially near the center of the region, where they share the highest peak in the same interval.

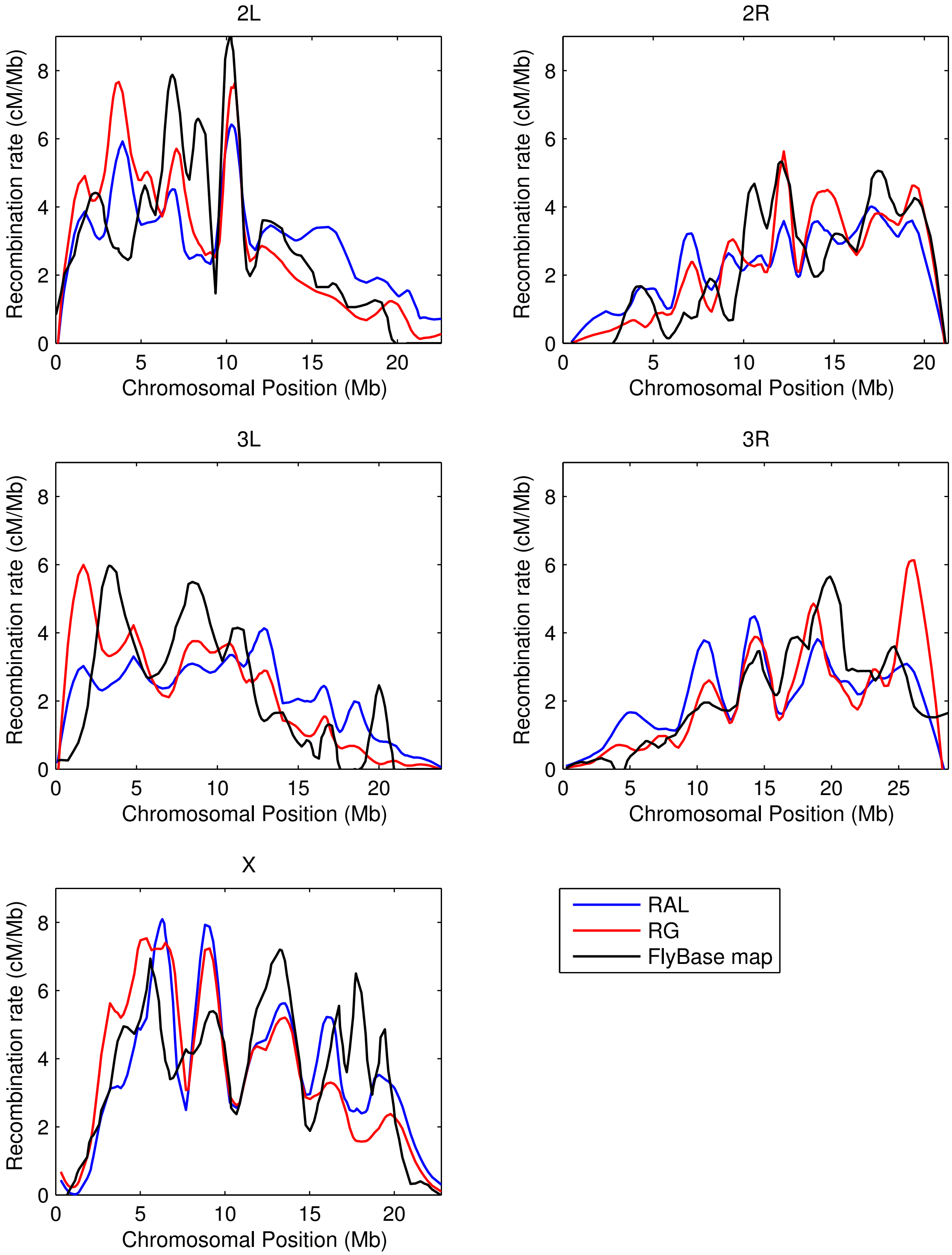

Figure 2.10: **Comparison with FlyBase genetic map.** Plotted for each chromosome arm are the estimated recombination maps using our method and the consensus experimental map hosted at FlyBase [56]. To ease comparison each map is LOESS-smoothed using a span of 15%. `LDhelmet` estimates were converted into units of cM/Mb by normalizing them to have the same total genetic distance as the empirical map.

### 2.20.7 Recombination Hotspots

As discussed in Section 2.1, it is well known that in humans and many other eukaryotes recombination tends to cluster in recombination hotspots, regions of approximately 2 kb wide in which the rate of recombination may be one or more orders of magnitude higher than the background rate [59, 62, 77, 19]. However, it is an open question whether hotspots exist in the *D. melanogaster* genome, or to what extent recombination rates vary on a fine scale.

We first searched for the most extreme forms of recombination rate variation—namely, recombination hotspots. Using a highly conservative approach, we initially identified nineteen and five putative autosomal recombination hotspots from the RAL and RG data, respectively. Of these, respectively six and four were also detected by the hotspot detection software `sequenceLDhot` [25]. These ten hotspots, the width of which ranges between 0.5 kb and 6.8 kb, are listed in Table 2.15. All were found in genic regions, with all except one overlapping exons and one contained within an intron. An example of a RAL hotspot is shown in Figure 2.11, where we also show the RG recombination map. The fine-scale recombination maps in this region for the two populations are clearly highly correlated, with both RAL and RG exhibiting a tenfold increase in recombination rate within almost identical 4 kb intervals, though only the hotspot of RAL was also found by `sequenceLDhot`. We note that the power of `sequenceLDhot` may be further reduced by the apparent preference (not shown) for putative hotspots to reside in regions in which the "local" background rate is higher than that of the chromosome arm as a whole. In light of these factors, it is likely that several more hotspots would have been found in one or both populations under a more relaxed definition, though it is clear that they are far scarcer, and less hot, than in humans.

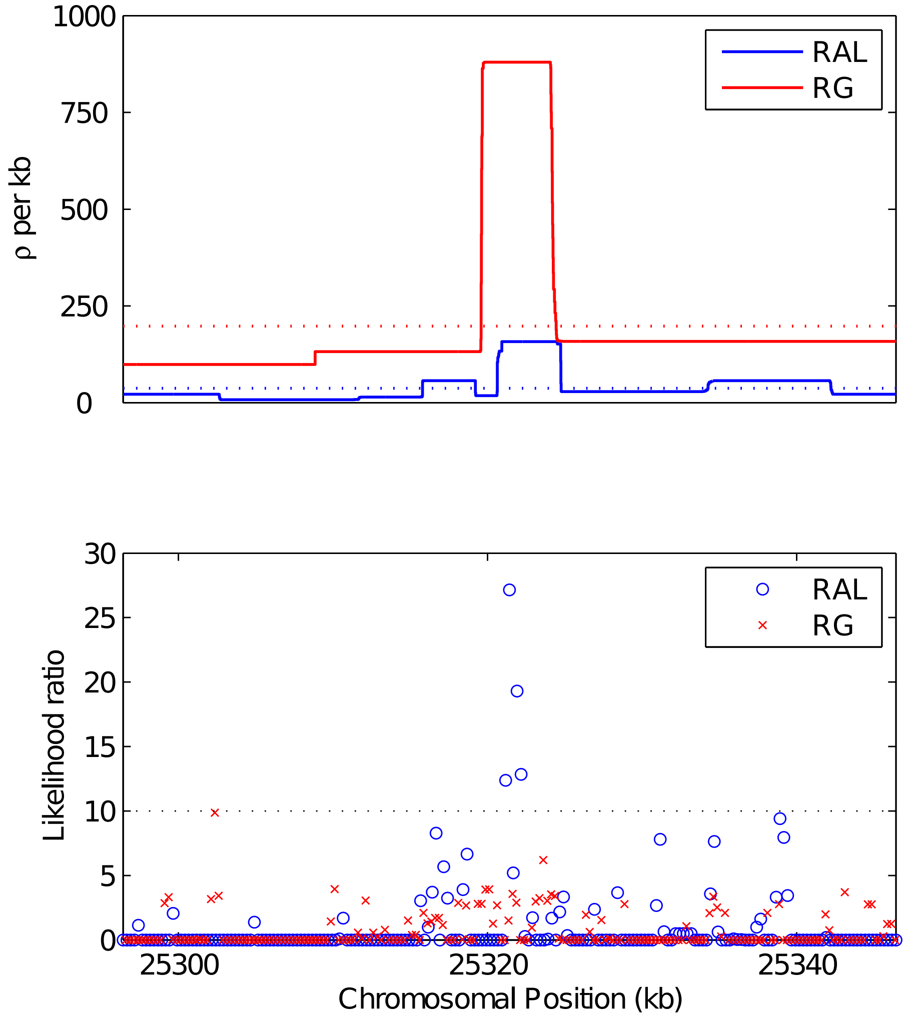

Figure 2.11: **A putative hotspot found by `LDhelmet` and confirmed by `sequenceLDhot`.** (Top): Estimated recombination rate for RAL (blue) and RG (red) in a 50 kb region of chromosome 3R, and their respective mean recombination rates in this region (dotted). (Bottom): Evidence of recombination hotspots in the same region, evaluated according to `sequenceLDhot`. The dotted line shows the likelihood ratio cutoff we used.

### 2.20.8 Genome-wide Fine-scale Recombination Rate Variability

It is apparent from both RAL and RG maps shown in Figure 2.8 that recombination rates vary on multiple scales, from the very fine to the very broad, as has been observed in a number of other species [59, 62, 22, 80, 5]. It is clear, for example, that recombination rates tail off towards each end of the arm, with the reduction towards the telomere much more precipitous than the pericentromeric reduction. The latter reduction is evident from as far as the start of heterochromatic sequence a few megabases from the centromere, in agreement with other broad-scale estimates of recombination [27, 54], although we do not find a complete absence of recombination here.

Figure 2.12 shows that the recombination rate in the X chromosome between positions 10 kb and 20 kb is noticeably higher than the rate in the subtelomeric region to the right. This trend is much more pronounced in the North American X than in the African X, consistent with a previous study by Anderson *et al.*[1]. The telomere-associated sequence (TAS), located to the left of position 10 kb, was not available in our data, but Anderson *et al.* provided evidence that the TAS region in the North American X exhibits even higher recombination rate than that in the subtelomeric region between positions 10 kb and 20 kb.

As shown in Figure 2.8, the largest difference between the estimated recombination maps of the two populations is observed in the X chromosome. First, the recombination map in the African X is generally much higher than that in the North American X. Second, there is noticeably less variation in the estimated African X recombination map. As mentioned earlier in the discussion of our simulation study, when the average recombination rate is as high as that of the African X, the amount of variation in our estimated map tends to be somewhat lower than the true variation. Hence, the observed reduction in variation could be partially attributed to our method being not sensitive enough in that range of very high rates. More generally, it is also true that Fisher's information for data on sequence variation is lower in regions of high recombination (Figure 2.1), which could create an inherent limitation in our ability to infer recombination rate changes here.

Table 2.15: **Putative recombination hotspots in *D. melanogaster* found by our method.** These putative hotspots were confirmed by the hotspot detection software `sequenceLDhot` [25].

| Dataset | Arm | Gene | Start | End | Width (kb) | #SNPs | $\rho$ per kb | Ratio to arm mean |
|---|---|---|---|---|---|---|---|---|
| RAL | 2L | CR43314 | 11966311 | 11966880 | 0.6 | 20 | 140.8 | 11 |
| | 3L | CG9384, CG17173 | 14759823 | 14761142 | 1.3 | 30 | 177.9 | 13 |
| | 3R | Cys | 10394533 | 10395940 | 1.4 | 42 | 100.8 | 10 |
| | 3R | CG7530 | 10552022 | 10553677 | 1.7 | 65 | 110.6 | 11 |
| | 3R | Ccap | 18526587 | 18527115 | 0.5 | 23 | 122.1 | 13 |
| | 3R | CG2010, Trc8 | 25320629 | 25324745 | 4.1 | 169 | 154.9 | 16 |
| RG | 2R | DJ-1$\alpha$, AGO1 | 9830014 | 9830946 | 0.9 | 53 | 547.3 | 14 |
| | 2R | CG15706, Tsf3 | 12109706 | 12116536 | 6.8 | 344 | 545.2 | 14 |
| | 2R | CG4927, CG8317 | 12460329 | 12466422 | 6.1 | 255 | 431.4 | 11 |
| | 3R | nAcR$\beta$-96A | 20339494 | 20340164 | 0.7 | 33 | 219.7 | 12 |

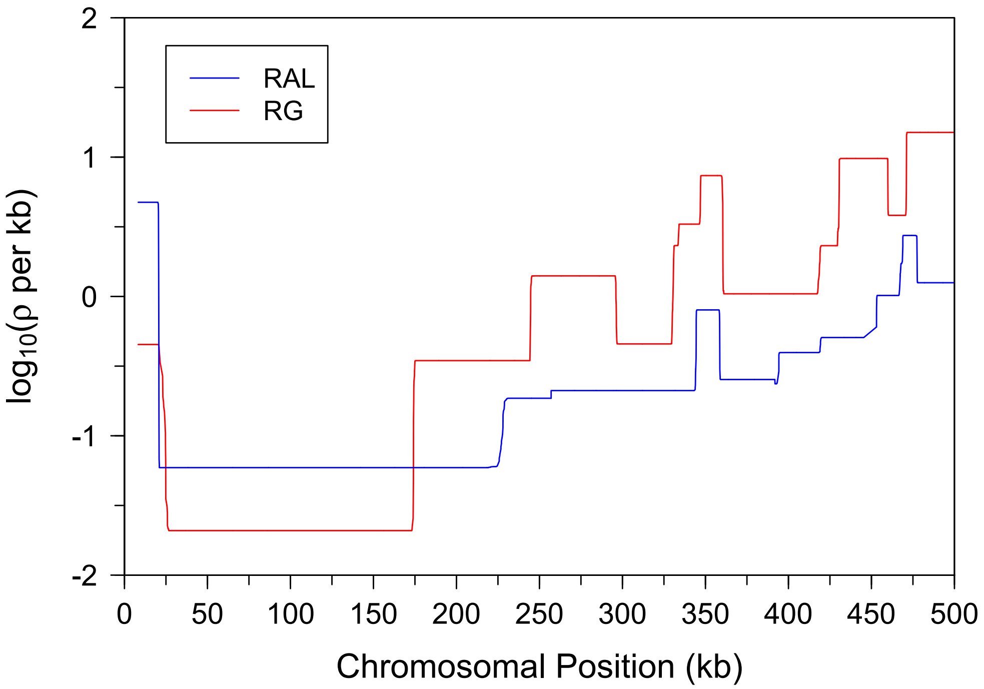

Figure 2.12: **Fine-scale recombination maps for the X chromosome subtelomeric region.** The telomere is at the left end of the region. The recombination rate between positions 10 kb and 20 kb is considerably higher than the rate in the subtelomeric region immediately to the right. This trend is much more pronounced in the North American X than in the African X, consistent with a previous study [1].

### 2.20.9 Recombination around Transcription Start Sites

To assess the pattern of recombination around genes, we plotted the average recombination rate as a function of distance from the transcription start sites (TSS). As shown in Figure 2.13, the plots for RAL and RG show high similarity in shape, despite differences between their fine-scale recombination maps. Also, note that the plots follow a similar pattern as in human [62, 77, 19], chimpanzee [5], and mouse [13], although the gene density of *D. melanogaster* is much higher than that of the other species.

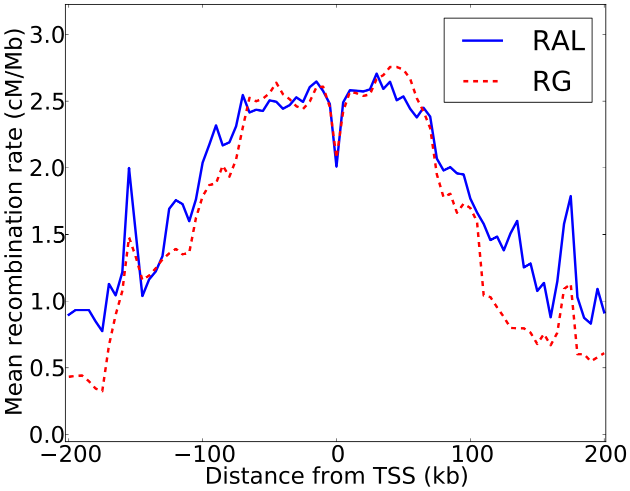

Figure 2.13: **Distribution of recombination rates relative to transcription start sites.** Plots for RAL (solid) and RG (dashed) of the average estimated recombination rate as a function of distance from the midpoint of the nearest transcription start site (TSS) to the left (negative x-axis) and to the right (positive x-axis) of every base. A 5-kb averaging window was used to smooth the estimates.

## 2.21 A Wavelet Analysis

To carry out a more methodical analysis of recombination rate variation within and between the two populations, and its correlation with other genomic features, we performed a wavelet analysis. Wavelet analyses are suitable for detecting localized, intermittent periodicities embedded in the data, across a range of possible scales. Our inputs are two sets of discrete "time"-series data representing the recombination maps of RAL and RG, binned into a recombination rate in each 250 bp window. Each is transformed into a collection of coefficients indexed by position ("time") and scale, and describe the variation in the input signal at each position and scale. The scale index may be discrete or continuous, and we make use of both types of transform as appropriate. Although the wavelet transform may be complex-valued, it can be summarized by a plot of its *(local) power*: the square of the norm of the wavelet coefficients at each position and scale. Taking the mean power across all positions yields the *(global) wavelet power spectrum*, which summarizes how the total variability in the signal is explained by heterogeneity at different scales. Further, a correlation between the wavelet coefficients from two different "time"-series datasets can identify how a change in one signal predicts a change in the other, at a given scale. One advantage of the wavelet approach is that one does not have to choose the appropriate window size in advance, which is important since analyses of genomic data on different pre-chosen scales can give conflicting results (e.g., [72, 48, 69]).

### 2.21.1 Interpretation

To illustrate, continuous wavelet transforms of the recombination maps of chromosome arm 2L are shown in Figure 2.14; wavelet transforms for the rest of the genome are shown in Figures 2.15–2.18. For brevity we focus on chromosome arm 2L throughout. We can interpret these transforms with reference to the wavelet transform of a constant recombination map, which would yield essentially zero power (dark blue) everywhere. Clearly the transform is highly inconsistent with a constant map. Regions of high power, shown at the red end of the spectrum and corresponding to wavelet coefficients of large magnitude, are consistent with variation in recombination rate at the given location ($x$-axis) and at the given scale ($y$-axis). Intuitively, a location of high local power in the wavelet transform suggests that a useful proportion of the variability in our dataset is well-explained if we track it by placing a wavelet function at this position and with the appropriate width corresponding to this scale. One way to evaluate the most significant regions of the time-frequency domain is to compare the transformed data with the transform of a null first-order autoregressive process with the same variance; thus, we allow for some variability as we scan along the data from left to right, and identify those regions (black contours in the figures) with wavelet power significantly above the null expectation.

Observe that highest power (red color) is seen in Figure 2.14 at the broadest scales (long periods) and at very fine scales. The former reflects the centromeric and telomeric decline in recombination rate, and we see that the centromeric decline has a more pronounced effect on the largest periods (though we caution that these signals are below the *cone of influence*, a region whose wavelet transform may be unduly distorted by edge effects [79]). Analogous patterns are evident in the other chromosome arms (Figures 2.15–2.18). Notice also that very fine-scale variation is manifested in high power regions at small periods (e.g., Figure 2.14, right-hand plots). While there exists some previous evidence for localized fine-scale variation in recombination rate in *D. melanogaster* [71], our finding that it is widespread across the genome is novel.

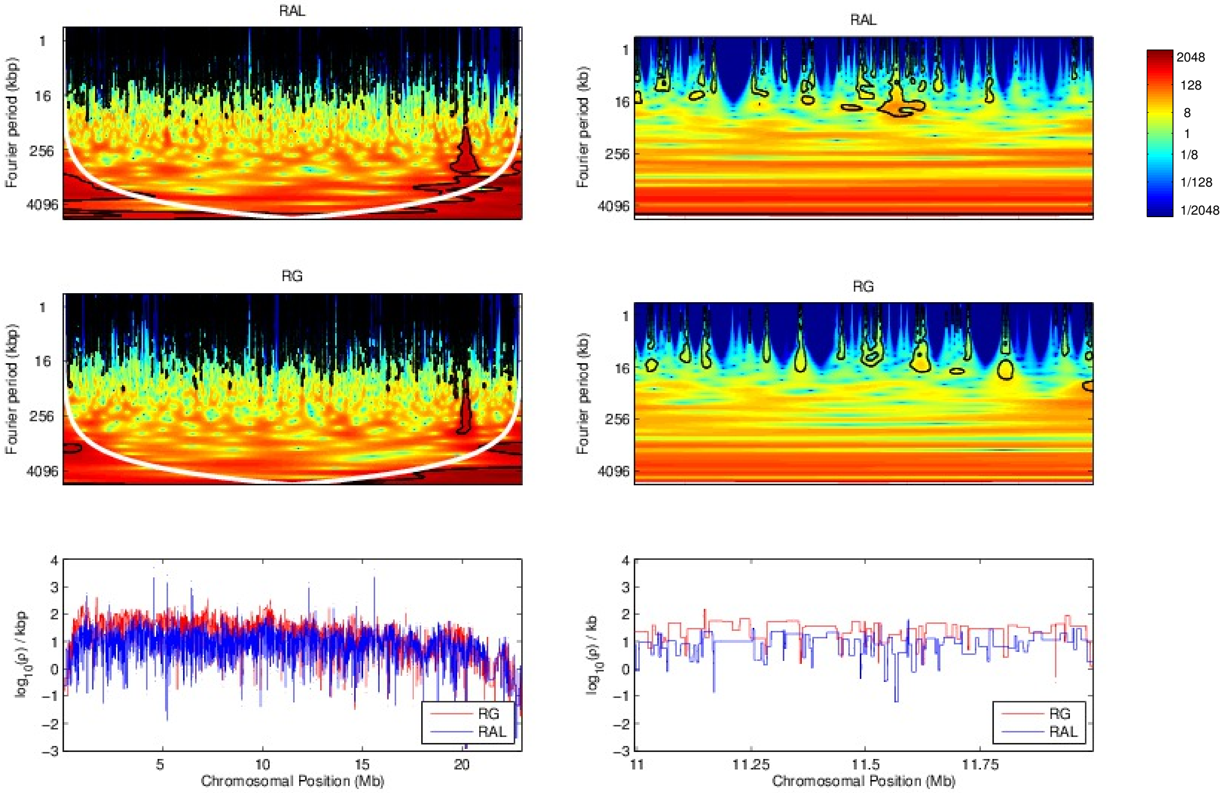

Figure 2.14: **Local wavelet power spectrum of recombination rate variation across chromosome arm 2L.** The whole arm is shown on the left, and a detailed (central) 1 Mb is shown on the right, for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white noise process with the same variance. Lower panels show estimates of the corresponding recombination maps.

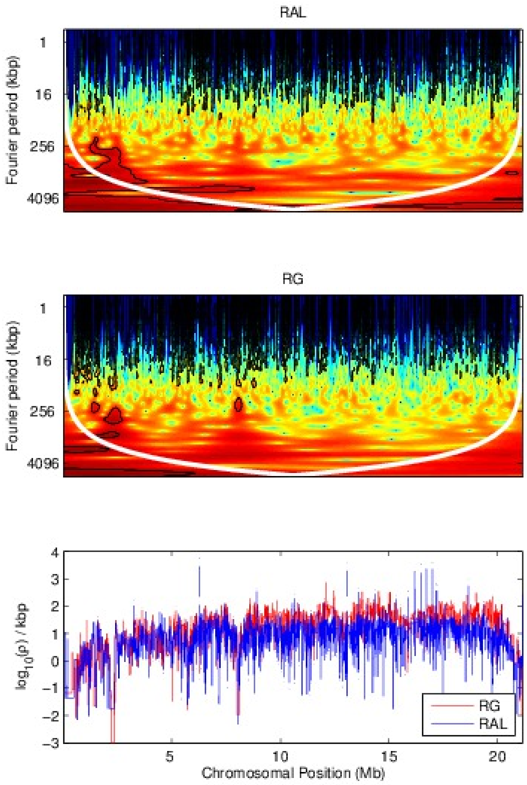

Figure 2.15: **Local wavelet power spectrum of recombination rate variation in chromosome arm 2R.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

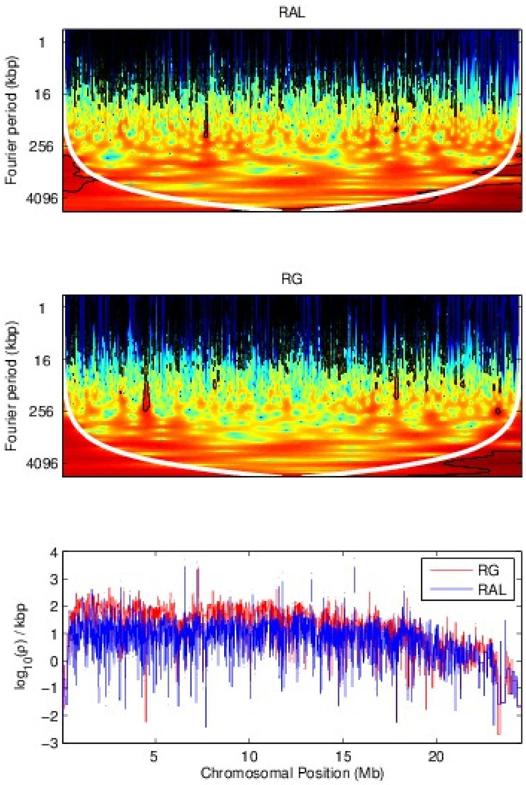

Figure 2.16: **Local wavelet power spectrum of recombination rate variation in chromosome arm 3L.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

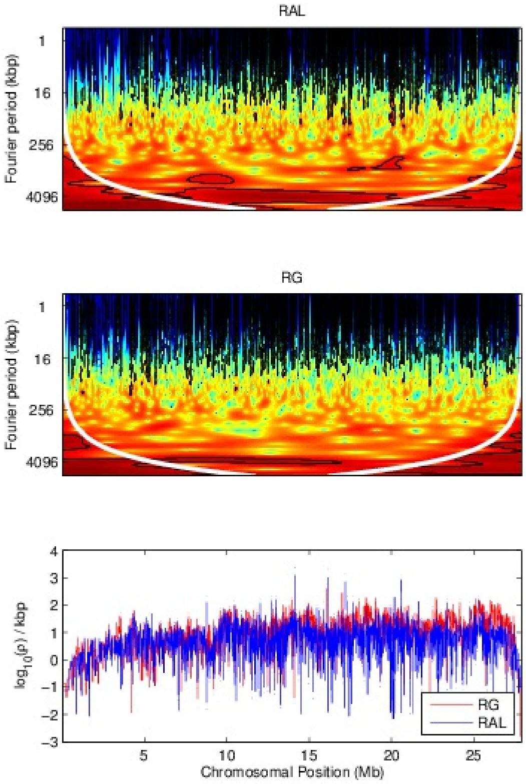

Figure 2.17: **Local wavelet power spectrum of recombination rate variation in chromosome arm 3R.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

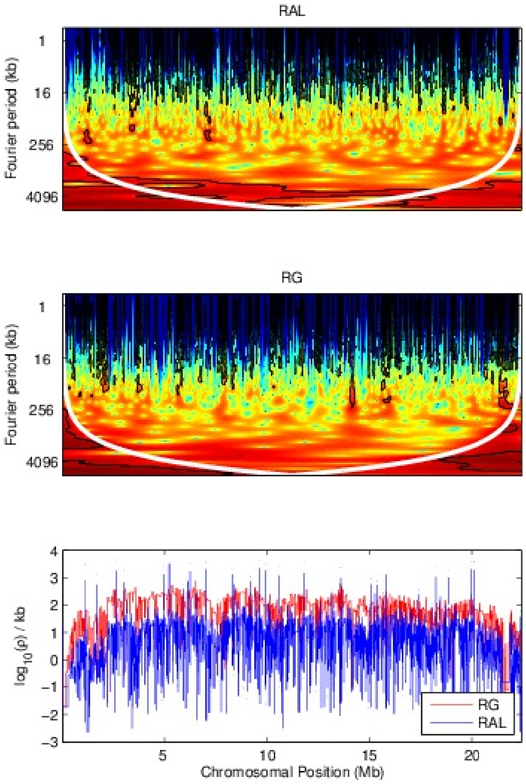

Figure 2.18: **Local wavelet power spectrum of recombination rate variation in chromosome X.** A power spectrum is shown for RAL and RG. Black contours denote regions of significant power at the 5% level, and the white contour denotes the cone of influence. Color scale is relative to a white-noise process with the same variance. The lower panels shows estimates of the corresponding genetic maps.

### 2.21.2 Correlation at Various Scales

Although there is some correlation in fine-scale variation between the two populations (for example, its lower volatility in region 11.2–11.25 Mb of arm 2L; see the right column of Figure 2.14), it is far from strong. To explore how well correlated the two maps are at each scale, we computed the pairwise correlations between wavelet coefficients of the two maps, after applying a discrete (Haar) wavelet transform following [72] (Figures 2.19, 2.20). This choice of transform decomposes a dataset into a series of wavelet coefficients for each of a discrete set of scales. The decomposition provides a series of *detail* coefficients measuring changes between neighboring observations, and a series of *smooth* coefficients which provides a smooth approximation of the original signal [24]. The correlation, at a given scale, between the detail coefficients of the wavelet transform of two maps can then be computed, and those with significantly high correlation identify the scales at which the two maps do covary. Across all arms and across all except the broadest scales there is a highly significant correlation in the variability of the two maps (Kendall's rank correlation, two-tailed test at 1% significance). The lack of correlation at broader scales is probably due to lack of power: for example, at the 1% level there are too few data points for this test to have any power at any scale broader than 4 Mb.

Given the similarities between the two populations, it is perhaps not surprising that their recombination rates are highly correlated when assessed globally. To further elucidate how this correlation varies in different regions of the genome, we performed a *wavelet coherence* analysis, which can be regarded informally as calculating a squared correlation coefficient between the variation of the two maps at each position as well as at each scale. Wavelet coherence analysis thus evaluates correlations in local, rather than global, power. Results are shown in Figures 2.21 and 2.22. It is clear that the correlation between the two maps is found nonuniformly along the chromosome. While there is high correlation at all positions at the broadest (megabase) scales, at smaller scales there exist regions of very low correlation, even when the overall correlation between the two maps at this scale is high. For example, the average coherence between the two maps at the 256 kb scale is 0.59 over the whole of 2L, compared to only 0.19 in the region 5–6 Mb. (Note that the persistently high correlation seen near position 20 Mb across many scales, reflects a particular region of missing data in both populations, and hence flat recombination.) Although the existence of regions of low coherence is partly explained by statistical error (Figure 2.23), it does not explain the drop fully. Thus, at least some isolated regions of low correlation are consistent with the idea that biological differences between the two populations create local differences in the recombination rate.

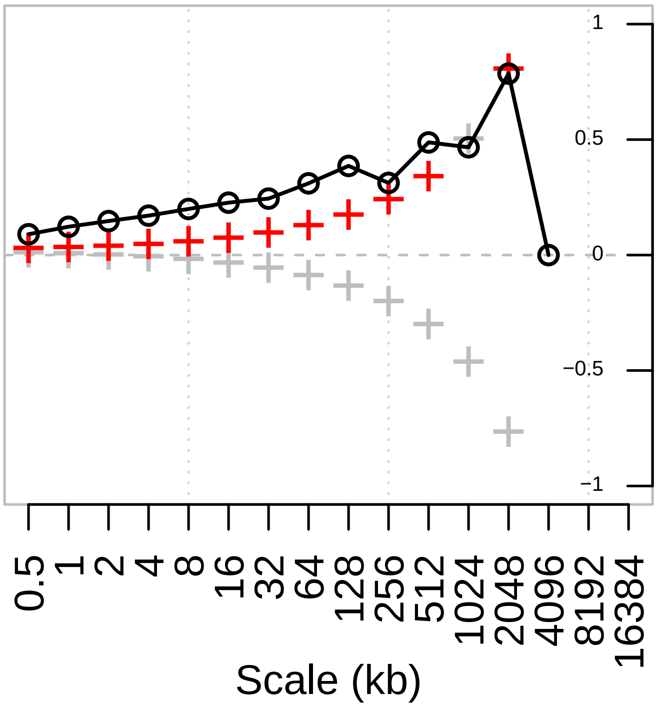

Figure 2.19: **Pairwise correlation of detail wavelet coefficients of RAL and RG recombination maps for chromosome arm 2L.** Black circles denote Kendall's rank correlation between pairs of detail coefficients at each scale. Crosses denote the correlation that would be required for significance at the 1% level in a two-tailed test; red crosses are those scales at which the correlation is in fact significant.

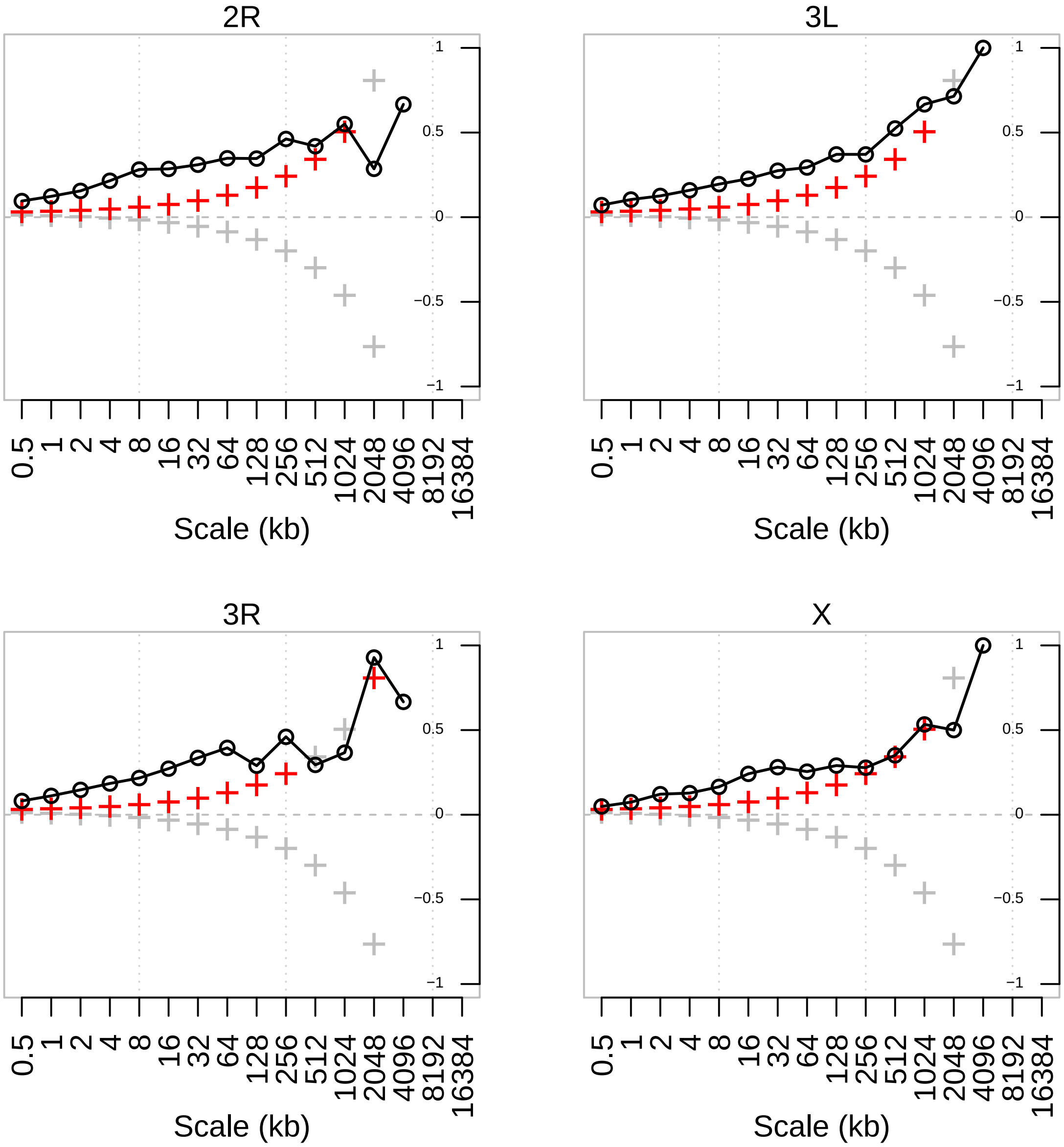

Figure 2.20: **Pairwise correlation of detail wavelet coefficients of RAL and RG recombination maps for chromosome arms 2R, 3L, 3R, and X.** Black circles denote Kendall's rank correlation between pairs of detail coefficients at each scale. Crosses denote the correlation that would be required for significance at the 1% level in a two-tailed test; red crosses are those scales at which the correlation is in fact significant.

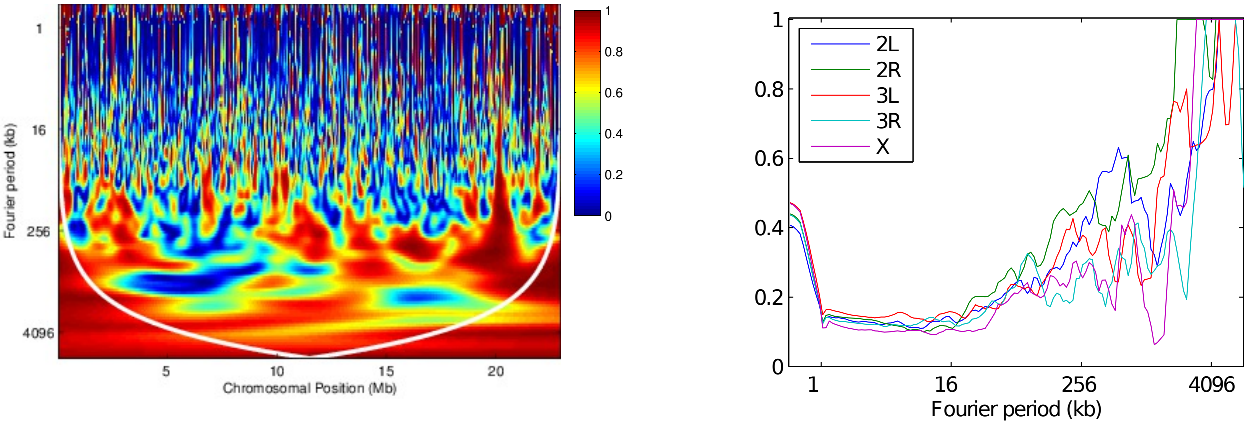

Figure 2.21: **Wavelet coherence analysis comparing RAL against RG.** (Left): Wavelet coherence of the two maps for chromosome arm 2L. The cone of influence is shown in white. (Right): For each arm, the plot shows the fraction of the genome with significantly high coherence at the 5% level, at each scale.

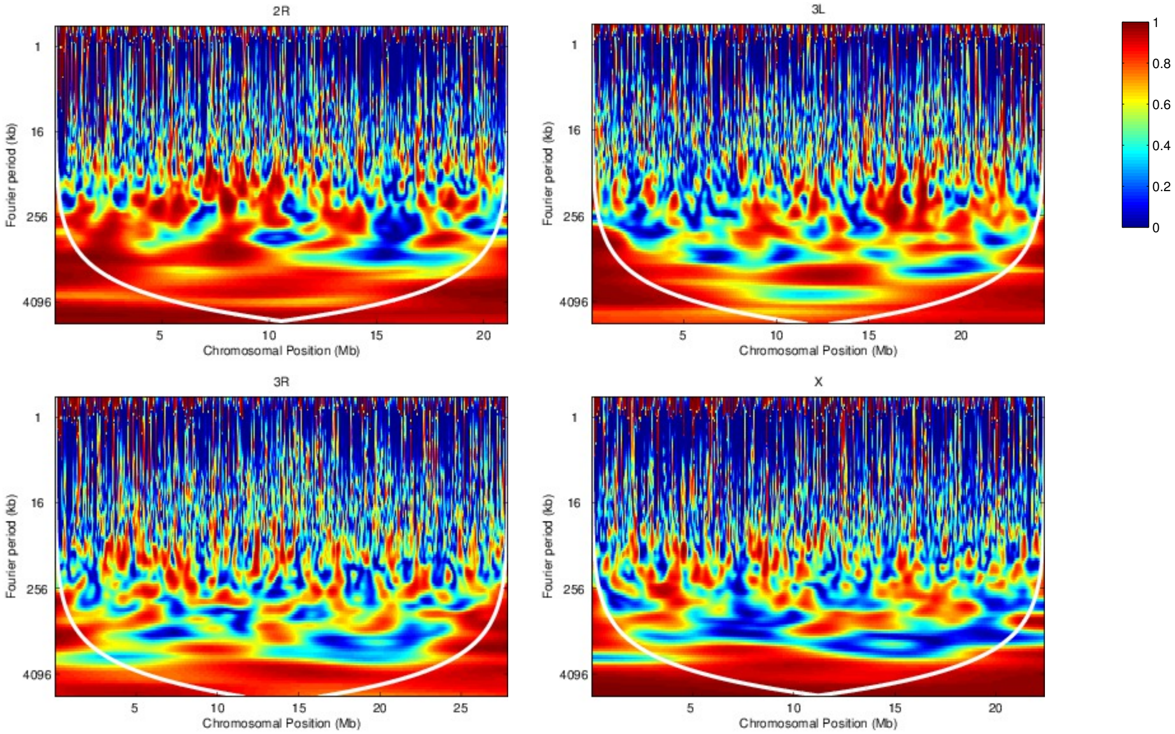

Figure 2.22: **Wavelet coherence analysis comparing RAL against RG for chromosome arms 2R, 3L, 3R, X.** The cone of influence is shown in white.

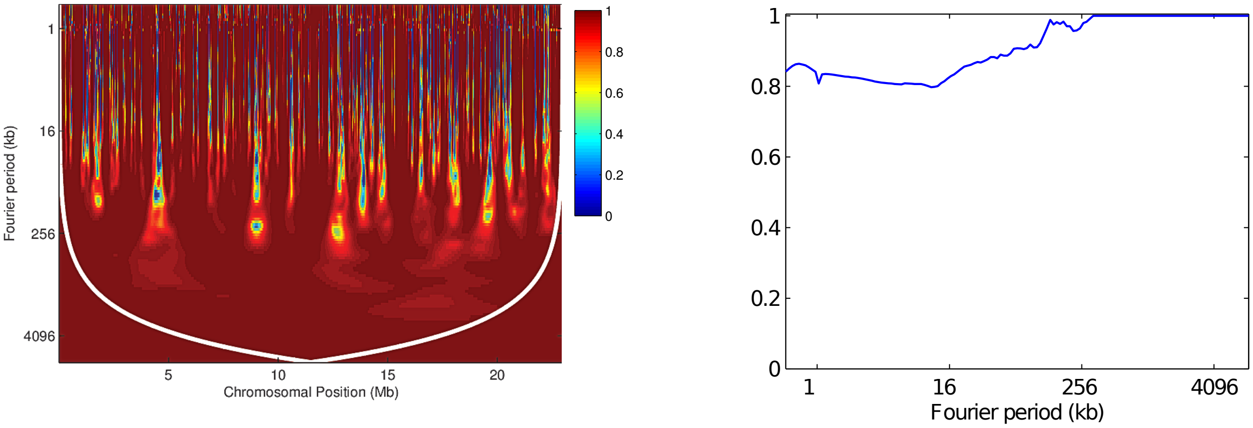

Figure 2.23: **Positive control for wavelet coherence analysis.** (Left): Coherence plot for two independent estimates of the recombination map across chromosome arm 2L using the same (RG) dataset. (Right): The fraction of chromosome arm 2L with significantly high coherence at the 5% level, at each scale.

## 2.22 Correlation of Recombination Rates with Other Genomic Features

The use of wavelets enables us to compare how changes in the rate of recombination along the genome correlate with other genomic features. For each population we computed pairwise correlations between the detail coefficients of the following features: diversity (mean fraction of pairwise differences between each individual in the population, within sequenced nucleotides), divergence (fraction of differences between the reference sequences of *melanogaster* and *simulans*), GC content, gene content (fraction of sites annotated as exonic), and sequence quality (Phred score), as well as the recombination rate, with each feature measured in 250 bp windows. Results are shown in Figures 2.24 and 2.25, and follow a similar analysis performed by Spencer *et al.* [72] on human data. From these results we can make a number of observations detailed below.

### 2.22.1 The Power Spectra of Each Genomic Feature

As in humans, we find the greatest heterogeneities in divergence and GC content at the finest scales, and in gene content at intermediate scales. Heterogeneity in diversity and recombination are strikingly different when we compare RAL and RG: recombination shows the greatest heterogeneity at fine scales in RAL and at intermediate scales in RG (as in humans); the reverse is true of diversity. These patterns are broadly repeated for each arm (Figure 2.25), although it should be noted that the lack of heterogeneity in recombination at fine scales in the RG data may partly be a consequence of its high background recombination rate leading to lower resolution (as discussed above; see Figure 2.1). Limitations such as these notwithstanding, the broad agreement between chromosome arms gives ground for optimism that the signals are not swamped by noise.

### 2.22.2 Pairwise Covariation of Genomic Features

The off-diagonal plots in Figure 2.24 provide a great deal of information about the covariation of several pairs of genomic features. Some are predictable and also found in humans [72]. For example, there is a strong positive correlation between diversity and divergence at fine and intermediate scales, consistent with variation in mutation rates at different positions in the genome. As a second example, both the negative correlation between gene content and diversity and the negative correlation between gene content and divergence are predicted by the observation that exons tend to be under greater selective constraint.

Perhaps the most notable difference between *D. melanogaster* and humans is seen when we examine the correlation between recombination and diversity. In humans this correlation is weak and extends only up to approximately the 4 kb scale. Spencer *et al.* [72] therefore infer that the influence of recombination on changes in diversity is primarily local in nature and driven by recombination hotspots. In *D. melanogaster*—for both the RAL and RG

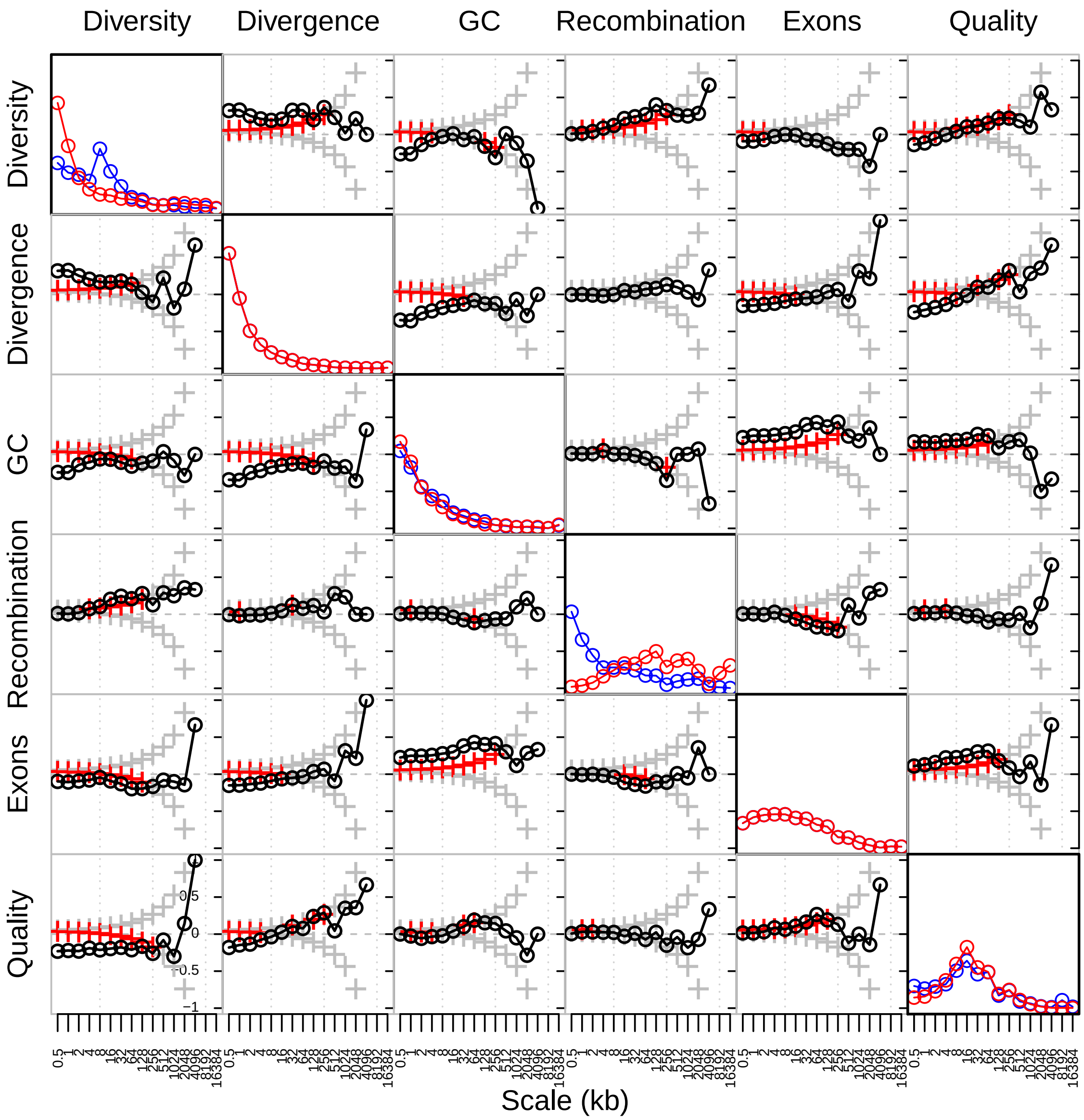

Figure 2.24: **Global wavelet power spectrum and pairwise correlations of detail wavelet coefficients of RAL and RG data for chromosome arm 2L.** Diagonal plots show the global wavelet power spectrum of each feature of the RAL (blue) and RG (red) data. Off-diagonal plots show Kendall's rank correlation between pairs of detail coefficients at each scale, with respect to the wavelet decomposition of the two indicated features. Crosses denote the correlation that would be required for significance at the 1% level in a two-tailed test; red crosses are those scales at which the correlation is in fact significant. The bottom left and top right plots correspond to RAL and RG, respectively.

data—the positive correlation between recombination and diversity is stronger and acts up to intermediate scales, approximately 2–256 kb. Interestingly, the correlation at very fine scales, $< 2$ kb, is weaker and for some chromosome arms nonsignificant (see Figure 2.25). These findings suggest both that a local influence of recombination hotspots on diversity is weaker or absent in *D. melanogaster*, consistent with the paucity of hotspots found in our search described above, and that some other phenomenon exerts an effect on diversity, but not divergence, over much larger scales. Clearly, one candidate is the action of selection, whose impact on the correlation between recombination and diversity is well appreciated [8, 9, 48, 71, 74, 49]. The scale up to which we have been able to detect this correlation, around 256 kb (with some differences according to the population and chromosome arm examined), is surprisingly large given that the footprints of selective sweeps are typically in the region of up to $\sim 20$ kb [69, 49].

Finally, it is notable that there is a significant negative correlation between the recombination rate and gene content at intermediate scales, in both RAL and RG and across all chromosome arms (though the signal is weaker on the X chromosome). This is consistent with the apparent preference for crossovers to occur outside exonic sequence [61], although we note that the effect does not appear to act at the finest scales—recall also that all but one of the putative hotspots identified in Table 2.15 do in fact overlap with exonic sequence.

### 2.22.3 A Linear Model Analysis

Given the strong but imperfect correlation between the recombination maps of RAL and RG, can we use the same genomic features to predict the regions in which the two maps might differ? To extend the analysis above and to address this question, we used a linear model analysis of the wavelet coefficients of each recombination map, using wavelet coefficients of other features as predictors. This analysis is similar to that described in [72], though their interest was in the prediction of changes in diversity rather than recombination. For each population and at each scale, we fit a linear model for the detail coefficients of the recombination map using as predictors the detail coefficients of wavelet transforms of sequence quality, gene content, GC content, divergence, and diversity (Tables 2.16A, 2.17A–2.20A). We find changes in diversity to be a strong predictor of changes in recombination across all chromosome arms and across many scales, though the effect is on some arms somewhat weaker (and nonsignificant) at the finest scales. Again, this is in contrast to the primarily local relationship between changes in diversity and recombination in humans. In addition to diversity, there are additional positive influences of GC content and sequence quality at fine scales; a weak negative influence of gene content at intermediate scales; and, in RG only, a negative influence of sequence quality at broad scales. Each of these signals is much weaker on the X chromosome (Table 2.20A), except the influence of diversity as a predictor of recombination, which still extends up to the megabase scale despite much higher absolute rates of recombination on this chromosome. The positive association between GC content and recombination is consistent with biased gene conversion [72, 74] and/or codon bias [18, 74], though we note an apparent negative correlation between GC content and recombination at broader scales (Figures 2.24, 2.25).

When the recombination map from the other population is added as an additional covariate, it is the strongest predictor of recombination rate at all but the broadest scales (Tables 2.16B, 2.17B–2.20B). Of the remaining covariates, those which were previously highly significant predictors now generally have reduced impact. However, their $p$-values at several scales are still highly significant, indicating that they offer explanatory power of the recombination rate over and above that provided by the recombination map of the other population. In particular, diversity remains a strong positive predictor of levels of recombination over most scales.

Table 2.16: **Linear model for wavelet transform of recombination map of chromosome arm 2L.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm 2L, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the –$\log_{10} p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted $r^2$. (B) As above, but with the recombination map of the other population as an additional covariate.

## A

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 9.12 | 24.88 | 28.77 | 15.26 | 8.04 | 1.54 | 1.09 | 1.31 | 0.10 | 1.78 | 0.65 | 0.05 |
| Exons | 0.41 | 1.26 | 0.97 | 2.08 | 1.48 | 4.92 | 8.73 | 6.24 | 0.73 | 0.92 | 0.11 | 0.04 |
| GC | 2.69 | 7.53 | 5.38 | 4.56 | 0.07 | 0.39 | 0.37 | 1.61 | 0.68 | 0.16 | 0.46 | 0.89 |
| Divergence | 0.54 | 0.97 | 0.09 | 0.35 | 0.13 | 0.76 | 0.64 | 1.65 | 0.36 | 1.16 | 0.11 | 0.04 |
| Diversity | 5.58 | 4.96 | 10.00 | 14.84 | 17.00 | 13.08 | 17.72 | 6.76 | 12.34 | 1.29 | 4.92 | 3.04 |
| Adjusted $r^2$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.05 | 0.10 | 0.25 | 0.24 | 0.39 | 0.20 | 0.50 | 0.67 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 16.21 | 18.92 | 7.47 | 1.13 | 0.25 | 6.56 | 0.37 | 4.26 | 4.36 | 2.66 | 2.34 | 0.69 |
| Exons | 1.03 | 0.65 | 0.70 | 0.33 | 0.77 | 1.30 | 3.92 | 1.52 | 0.84 | 0.68 | 0.31 | 0.33 |
| GC | 2.64 | 4.08 | 3.11 | 3.70 | 1.00 | 0.07 | 0.05 | 0.12 | 0.18 | 0.32 | 1.90 | 0.52 |
| Divergence | 0.04 | 0.28 | 0.15 | 1.02 | 0.33 | 0.40 | 0.63 | 0.37 | 0.01 | 0.27 | 0.06 | 0.14 |
| Diversity | 6.25 | 3.67 | 6.53 | 17.21 | 17.18 | 33.27 | 17.58 | 19.02 | 16.15 | 5.11 | 5.93 | 4.41 |
| Adjusted $r^2$ | 0.00 | 0.01 | 0.01 | 0.03 | 0.05 | 0.15 | 0.23 | 0.38 | 0.48 | 0.38 | 0.61 | 0.81 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

## B

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RG map | 23.24 | 45.06 | 57.62 | 47.49 | 43.80 | 35.86 | 15.95 | 11.87 | 5.46 | 2.92 | 2.91 | 1.38 |
| Quality | 8.01 | 20.57 | 23.00 | 11.61 | 4.78 | 1.01 | 0.47 | 0.78 | 0.27 | 1.10 | 0.61 | 0.03 |
| Exons | 0.33 | 1.21 | 0.92 | 2.17 | 1.28 | 3.71 | 4.73 | 3.58 | 0.29 | 0.64 | 0.25 | 0.04 |
| GC | 2.49 | 6.28 | 4.26 | 3.34 | 0.06 | 0.30 | 0.16 | 1.32 | 0.60 | 0.25 | 0.14 | 0.35 |
| Divergence | 0.52 | 0.80 | 0.00 | 0.25 | 0.09 | 0.89 | 0.59 | 0.94 | 0.65 | 0.90 | 0.19 | 0.14 |
| Diversity | 5.13 | 4.20 | 7.64 | 10.73 | 11.31 | 7.82 | 11.14 | 1.96 | 5.93 | 0.31 | 2.22 | 0.52 |
| Adjusted $r^2$ | 0.00 | 0.02 | 0.05 | 0.08 | 0.14 | 0.23 | 0.34 | 0.38 | 0.48 | 0.32 | 0.66 | 0.76 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RAL map | 23.67 | 45.22 | 58.36 | 46.84 | 41.39 | 31.87 | 14.06 | 9.17 | 4.05 | 3.02 | 1.97 | 1.46 |
| Quality | 15.72 | 15.23 | 3.52 | 0.25 | 0.14 | 7.04 | 0.48 | 2.75 | 2.98 | 1.41 | 1.11 | 0.11 |
| Exons | 1.00 | 0.45 | 0.37 | 0.07 | 0.09 | 0.03 | 1.52 | 0.36 | 0.53 | 0.25 | 0.50 | 0.26 |
| GC | 2.47 | 3.30 | 2.27 | 2.18 | 0.34 | 0.26 | 0.02 | 0.30 | 0.15 | 0.22 | 1.20 | 0.10 |
| Divergence | 0.07 | 0.23 | 0.28 | 0.70 | 0.10 | 0.46 | 0.51 | 0.39 | 0.12 | 0.68 | 0.06 | 0.73 |
| Diversity | 5.94 | 3.25 | 5.34 | 13.01 | 11.13 | 23.81 | 10.44 | 11.85 | 8.38 | 4.31 | 2.84 | 1.26 |
| Adjusted $r^2$ | 0.01 | 0.02 | 0.04 | 0.08 | 0.14 | 0.26 | 0.32 | 0.47 | 0.53 | 0.48 | 0.68 | 0.87 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

Table 2.17: **Linear model for wavelet transform of recombination map of chromosome arm 2R.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm 2R, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the –$\log_{10}$ $p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted $r^2$. (B) As above, but with the recombination map of the other population as an additional covariate.

**A**

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 6.58 | 15.49 | 16.81 | 8.92 | 14.46 | 8.80 | 7.43 | 3.41 | 4.87 | 8.08 | 1.87 | 1.83 |
| Exons | 0.85 | 0.81 | 1.86 | 2.04 | 6.98 | 8.46 | 6.06 | 4.21 | 4.16 | 3.86 | 1.17 | 0.38 |
| GC | 4.04 | 1.76 | 1.80 | 1.45 | 3.00 | 2.28 | 0.22 | 0.60 | 0.35 | 0.71 | 0.08 | 0.21 |
| Divergence | 0.17 | 0.04 | 0.07 | 1.91 | 1.11 | 0.53 | 0.54 | 0.69 | 0.16 | 0.31 | 0.14 | 0.03 |
| Diversity | 4.60 | 2.95 | 2.73 | 5.35 | 8.12 | 6.47 | 12.05 | 6.43 | 2.13 | 1.87 | 1.48 | 2.09 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.01 | 0.02 | 0.07 | 0.11 | 0.24 | 0.27 | 0.31 | 0.58 | 0.37 | 0.68 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 1.27 | 4.36 | 1.33 | 0.32 | 0.11 | 0.24 | 1.53 | 0.32 | 0.23 | 2.46 | 0.01 | 0.20 |
| Exons | 0.14 | 2.28 | 2.66 | 1.39 | 0.93 | 4.25 | 0.94 | 0.89 | 1.03 | 1.64 | 0.07 | 0.19 |
| GC | 2.00 | 3.36 | 3.43 | 0.31 | 0.60 | 0.40 | 0.01 | 1.77 | 0.07 | 0.01 | 0.94 | 0.00 |
| Divergence | 0.42 | 1.29 | 0.20 | 1.14 | 0.01 | 0.13 | 0.15 | 3.18 | 0.02 | 0.32 | 0.17 | 0.30 |
| Diversity | 4.66 | 7.62 | 12.99 | 16.49 | 18.25 | 16.75 | 25.88 | 16.68 | 11.90 | 2.83 | 6.13 | 1.44 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.01 | 0.02 | 0.05 | 0.12 | 0.28 | 0.38 | 0.44 | 0.56 | 0.67 | 0.51 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**B**

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RG map | 26.17 | 35.32 | 47.13 | 56.43 | 56.31 | 43.22 | 24.41 | 13.19 | 19.89 | 5.76 | 5.53 | 3.08 |
| Quality | 6.41 | 13.74 | 15.17 | 6.56 | 8.49 | 3.74 | 4.17 | 2.18 | 0.66 | 3.51 | 0.28 | 1.30 |
| Exons | 0.87 | 0.61 | 1.34 | 1.43 | 5.53 | 4.38 | 3.56 | 1.79 | 0.77 | 2.25 | 1.12 | 0.10 |
| GC | 3.79 | 1.38 | 1.17 | 1.35 | 2.09 | 2.58 | 0.70 | 0.75 | 0.47 | 0.64 | 0.32 | 0.53 |
| Divergence | 0.14 | 0.01 | 0.05 | 2.45 | 1.08 | 0.77 | 0.48 | 1.46 | 0.75 | 0.10 | 0.12 | 0.22 |
| Diversity | 4.36 | 2.45 | 1.60 | 3.12 | 4.43 | 2.63 | 3.43 | 2.35 | 1.44 | 1.37 | 0.23 | 1.41 |
| Adjusted $r^2$ | 0.00 | 0.01 | 0.04 | 0.08 | 0.17 | 0.26 | 0.39 | 0.41 | 0.66 | 0.71 | 0.72 | 0.89 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RAL map | 25.95 | 35.44 | 46.67 | 55.63 | 56.75 | 40.87 | 21.47 | 7.92 | 16.56 | 4.34 | 3.31 | 3.12 |
| Quality | 0.96 | 3.25 | 0.46 | 0.29 | 0.38 | 0.72 | 1.77 | 0.11 | 0.26 | 0.42 | 0.27 | 0.71 |
| Exons | 0.19 | 2.08 | 2.19 | 0.75 | 0.00 | 1.56 | 0.14 | 0.26 | 0.39 | 0.07 | 0.43 | 0.04 |
| GC | 1.78 | 3.14 | 3.22 | 0.07 | 0.24 | 0.57 | 0.48 | 1.69 | 0.33 | 0.14 | 0.60 | 0.11 |
| Divergence | 0.41 | 1.26 | 0.20 | 1.62 | 0.25 | 0.28 | 0.04 | 4.04 | 1.11 | 0.47 | 0.13 | 0.83 |
| Diversity | 4.32 | 7.03 | 11.83 | 13.79 | 12.66 | 9.62 | 13.96 | 7.99 | 4.18 | 1.48 | 2.74 | 0.14 |
| Adjusted $r^2$ | 0.00 | 0.01 | 0.03 | 0.08 | 0.16 | 0.26 | 0.40 | 0.45 | 0.69 | 0.66 | 0.79 | 0.84 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

Table 2.18: **Linear model for wavelet transform of recombination map of chromosome arm 3L.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm 3L, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the $-\log_{10}$ $p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted $r^2$. (B) As above, but with the recombination map of the other population as an additional covariate.

**A**

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 5.37 | 12.20 | 14.07 | 20.63 | 2.92 | 12.72 | 6.05 | 7.51 | 0.83 | 1.99 | 1.25 | 0.38 |
| Exons | 0.43 | 0.78 | 2.09 | 1.53 | 2.01 | 7.33 | 6.95 | 2.57 | 3.70 | 2.20 | 0.66 | 0.99 |
| GC | 5.97 | 6.33 | 5.22 | 1.35 | 0.08 | 0.68 | 0.19 | 0.07 | 0.11 | 1.21 | 0.78 | 1.19 |
| Divergence | 0.54 | 2.58 | 0.35 | 0.32 | 2.31 | 0.17 | 0.29 | 1.00 | 1.15 | 0.44 | 0.16 | 0.72 |
| Diversity | 4.43 | 4.38 | 4.75 | 7.53 | 9.94 | 11.01 | 10.18 | 9.91 | 5.50 | 4.15 | 4.29 | 5.65 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.01 | 0.03 | 0.04 | 0.15 | 0.20 | 0.25 | 0.43 | 0.51 | 0.67 | 0.88 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 5.10 | 14.22 | 4.30 | 2.43 | 0.04 | 1.09 | 6.09 | 2.31 | 3.42 | 0.64 | 0.61 | 1.39 |
| Exons | 0.68 | 1.03 | 0.66 | 1.09 | 1.96 | 5.03 | 4.20 | 1.71 | 0.95 | 0.21 | 0.08 | 1.44 |
| GC | 2.55 | 6.96 | 1.92 | 2.04 | 0.35 | 0.17 | 0.51 | 0.06 | 0.63 | 0.48 | 1.17 | 2.18 |
| Divergence | 0.02 | 1.23 | 0.51 | 1.68 | 0.38 | 1.21 | 0.02 | 1.04 | 0.17 | 0.08 | 1.38 | 0.22 |
| Diversity | 1.32 | 2.15 | 0.98 | 7.62 | 11.60 | 16.57 | 21.79 | 17.20 | 15.68 | 6.32 | 9.02 | 7.14 |
| Adjusted $r^2$ | 0.00 | 0.01 | 0.00 | 0.01 | 0.03 | 0.11 | 0.23 | 0.32 | 0.51 | 0.41 | 0.79 | 0.92 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**B**

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RG map | 18.62 | 48.20 | 38.39 | 50.24 | 44.70 | 18.86 | 10.64 | 11.14 | 2.40 | 1.23 | 3.20 | 0.96 |
| Quality | 4.78 | 9.81 | 12.15 | 15.93 | 1.19 | 9.82 | 5.68 | 5.86 | 0.63 | 1.57 | 0.83 | 0.22 |
| Exons | 0.49 | 0.69 | 1.96 | 1.19 | 1.26 | 5.20 | 4.60 | 1.39 | 3.29 | 1.78 | 0.67 | 0.31 |
| GC | 5.57 | 4.87 | 4.41 | 0.65 | 0.03 | 0.37 | 0.08 | 0.35 | 0.07 | 1.12 | 0.71 | 0.58 |
| Divergence | 0.54 | 2.53 | 0.31 | 0.09 | 3.10 | 0.03 | 0.15 | 0.68 | 0.98 | 0.49 | 0.59 | 0.29 |
| Diversity | 4.04 | 4.04 | 4.19 | 5.13 | 5.46 | 5.89 | 6.51 | 4.43 | 3.62 | 3.16 | 1.01 | 1.69 |
| Adjusted $r^2$ | 0.00 | 0.02 | 0.03 | 0.08 | 0.13 | 0.21 | 0.26 | 0.37 | 0.46 | 0.53 | 0.78 | 0.90 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RAL map | 18.81 | 47.84 | 38.36 | 52.65 | 45.56 | 22.02 | 10.01 | 11.77 | 1.12 | 0.91 | 2.29 | 0.68 |
| Quality | 4.64 | 11.65 | 2.28 | 0.51 | 0.12 | 3.09 | 7.53 | 3.66 | 3.53 | 0.72 | 1.06 | 0.68 |
| Exons | 0.74 | 0.86 | 0.36 | 0.66 | 1.16 | 2.21 | 1.98 | 0.44 | 0.41 | 0.01 | 0.13 | 1.24 |
| GC | 2.28 | 6.23 | 1.53 | 2.27 | 0.43 | 0.55 | 0.57 | 0.34 | 0.65 | 0.73 | 0.70 | 1.33 |
| Divergence | 0.04 | 0.84 | 0.34 | 1.40 | 0.95 | 0.97 | 0.03 | 0.65 | 0.12 | 0.01 | 1.58 | 0.15 |
| Diversity | 1.22 | 1.84 | 0.79 | 6.78 | 8.08 | 14.22 | 17.11 | 12.63 | 12.85 | 4.94 | 4.50 | 2.29 |
| Adjusted $r^2$ | 0.00 | 0.02 | 0.02 | 0.07 | 0.12 | 0.19 | 0.29 | 0.44 | 0.52 | 0.42 | 0.84 | 0.93 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

Table 2.19: **Linear model for wavelet transform of recombination map of chromosome arm 3R.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm 3R, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the –$\log_{10}$ $p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted $r^2$. (B) As above, but with the recombination map of the other population as an additional covariate.

## A

### RAL

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 9.15 | 14.12 | 11.51 | 12.13 | 10.37 | 6.99 | 8.05 | 4.26 | 2.38 | 0.94 | 0.05 | 1.62 |
| Exons | 0.43 | 0.27 | 1.90 | 2.09 | 1.55 | 5.62 | 2.18 | 7.94 | 3.02 | 2.05 | 0.77 | 1.07 |
| GC | 2.98 | 2.98 | 1.60 | 0.70 | 0.03 | 0.96 | 0.77 | 0.15 | 0.25 | 0.27 | 0.22 | 1.69 |
| Divergence | 0.20 | 1.30 | 0.19 | 0.12 | 0.10 | 0.21 | 0.31 | 0.50 | 1.58 | 0.05 | 0.03 | 0.90 |
| Diversity | 0.80 | 8.81 | 10.73 | 18.82 | 27.24 | 22.71 | 14.83 | 5.81 | 5.20 | 2.82 | 0.97 | 1.43 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.01 | 0.03 | 0.09 | 0.16 | 0.21 | 0.30 | 0.41 | 0.40 | 0.28 | 0.30 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

### RG

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 9.14 | 13.46 | 15.50 | 4.82 | 2.54 | 1.17 | 0.31 | 0.69 | 0.63 | 0.97 | 0.24 | 1.03 |
| Exons | 0.91 | 1.49 | 2.89 | 1.74 | 2.21 | 4.51 | 4.51 | 6.63 | 1.60 | 2.60 | 0.33 | 0.01 |
| GC | 0.83 | 1.20 | 0.97 | 0.78 | 0.40 | 0.45 | 0.29 | 0.56 | 0.88 | 0.76 | 0.04 | 1.95 |
| Divergence | 0.23 | 0.10 | 0.36 | 0.38 | 0.25 | 0.37 | 0.41 | 0.22 | 0.49 | 0.03 | 0.33 | 0.45 |
| Diversity | 8.78 | 9.10 | 13.83 | 17.23 | 28.09 | 20.90 | 18.50 | 13.66 | 8.64 | 7.76 | 2.90 | 3.86 |
| Adjusted $r^2$ | 0.00 | 0.01 | 0.02 | 0.04 | 0.10 | 0.17 | 0.26 | 0.43 | 0.37 | 0.58 | 0.33 | 0.74 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

## B

### RAL

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RG map | 29.79 | 58.35 | 45.48 | 34.45 | 46.73 | 19.13 | 20.15 | 7.80 | 3.20 | 3.26 | 1.88 | 2.45 |
| Quality | 8.38 | 11.65 | 7.59 | 9.29 | 6.03 | 3.49 | 5.18 | 3.95 | 1.89 | 0.57 | 0.19 | 1.89 |
| Exons | 0.39 | 0.17 | 1.31 | 1.59 | 0.70 | 2.93 | 0.10 | 2.93 | 2.16 | 1.34 | 0.94 | 0.82 |
| GC | 2.82 | 2.56 | 1.21 | 0.51 | 0.07 | 0.93 | 1.49 | 0.35 | 0.21 | 0.01 | 0.32 | 1.32 |
| Divergence | 0.20 | 1.50 | 0.16 | 0.33 | 0.24 | 0.18 | 0.27 | 1.04 | 1.42 | 0.18 | 0.16 | 1.84 |
| Diversity | 0.62 | 7.46 | 9.44 | 14.87 | 17.74 | 17.88 | 11.07 | 2.81 | 3.70 | 1.04 | 0.30 | 0.07 |
| Adjusted $r^2$ | 0.01 | 0.02 | 0.04 | 0.07 | 0.18 | 0.23 | 0.33 | 0.39 | 0.46 | 0.50 | 0.41 | 0.68 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

### RG

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RAL map | 29.31 | 57.37 | 44.51 | 33.51 | 43.57 | 17.01 | 14.23 | 6.82 | 1.69 | 1.97 | 1.82 | 0.56 |
| Quality | 8.19 | 10.13 | 12.58 | 3.36 | 1.65 | 0.79 | 0.51 | 1.11 | 0.60 | 0.61 | 0.36 | 0.38 |
| Exons | 0.85 | 1.37 | 2.28 | 1.27 | 1.36 | 2.58 | 3.07 | 3.26 | 0.49 | 1.17 | 0.91 | 0.03 |
| GC | 0.70 | 1.14 | 0.82 | 0.85 | 0.43 | 0.58 | 0.26 | 0.36 | 0.39 | 0.45 | 0.17 | 0.61 |
| Divergence | 0.20 | 0.25 | 0.39 | 0.45 | 0.18 | 0.34 | 0.27 | 0.40 | 0.29 | 0.06 | 0.27 | 0.62 |
| Diversity | 8.47 | 8.34 | 12.01 | 13.28 | 16.81 | 14.23 | 9.86 | 10.33 | 5.87 | 4.80 | 2.44 | 2.29 |
| Adjusted $r^2$ | 0.01 | 0.02 | 0.04 | 0.07 | 0.19 | 0.23 | 0.34 | 0.49 | 0.39 | 0.62 | 0.45 | 0.75 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

Table 2.20: **Linear model for wavelet transform of recombination map of chromosome X.** (A) In a linear model for the detail coefficients of the wavelet transform of the recombination map of chromosome arm X, covariates are the detail coefficients of wavelet transforms of data quality, gene content, GC content, divergence, and diversity. Shown is the –$\log_{10} p$-value of the regression coefficient at the given scale, as determined by a t-test. Colored boxes indicate significant relationships, with red positive and blue negative. Also shown in the adjusted $r^2$. (B) As above, but with the recombination map of the other population as an additional covariate.

**A**

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 2.46 | 4.71 | 3.64 | 2.48 | 0.82 | 0.62 | 0.36 | 0.11 | 2.11 | 0.54 | 0.35 | 1.57 |
| Exons | 1.43 | 2.11 | 0.03 | 1.18 | 0.31 | 0.10 | 2.18 | 1.95 | 0.79 | 2.11 | 0.53 | 1.51 |
| GC | 1.54 | 2.02 | 2.11 | 0.91 | 0.71 | 0.01 | 1.61 | 0.68 | 0.01 | 0.01 | 0.47 | 1.19 |
| Divergence | 0.11 | 0.03 | 0.05 | 0.13 | 0.02 | 0.01 | 0.65 | 0.02 | 0.26 | 0.45 | 0.08 | 1.01 |
| Diversity | 0.55 | 0.59 | 2.26 | 5.30 | 11.04 | 13.94 | 19.30 | 11.85 | 6.16 | 1.93 | 0.24 | 3.21 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.05 | 0.18 | 0.23 | 0.25 | 0.28 | 0.03 | 0.57 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Quality | 1.68 | 0.81 | 0.36 | 1.41 | 4.81 | 8.64 | 10.49 | 3.12 | 1.48 | 0.12 | 0.15 | 1.15 |
| Exons | 0.48 | 0.90 | 0.06 | 0.53 | 0.16 | 0.11 | 0.11 | 2.35 | 0.62 | 2.73 | 0.67 | 0.01 |
| GC | 0.32 | 2.13 | 1.30 | 1.18 | 1.67 | 0.65 | 0.12 | 0.36 | 0.11 | 0.63 | 0.95 | 1.70 |
| Divergence | 0.04 | 0.14 | 0.61 | 0.97 | 1.24 | 0.08 | 0.02 | 0.31 | 0.94 | 0.45 | 0.13 | 0.23 |
| Diversity | 3.33 | 5.88 | 3.24 | 5.29 | 15.37 | 26.97 | 24.56 | 17.46 | 12.69 | 4.25 | 2.75 | 3.89 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.11 | 0.20 | 0.37 | 0.43 | 0.43 | 0.51 | 0.75 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**B**

**RAL**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RG map | 1.75 | 3.74 | 11.81 | 13.72 | 25.39 | 17.48 | 13.42 | 8.96 | 6.38 | 2.70 | 2.45 | 0.82 |
| Quality | 2.39 | 4.36 | 3.13 | 2.23 | 0.81 | 0.36 | 0.20 | 0.25 | 1.02 | 0.22 | 0.24 | 0.40 |
| Exons | 1.44 | 2.17 | 0.02 | 1.32 | 0.24 | 0.07 | 1.49 | 0.53 | 0.38 | 1.28 | 0.01 | 1.26 |
| GC | 1.53 | 1.93 | 1.90 | 0.77 | 0.34 | 0.26 | 1.58 | 0.70 | 0.31 | 0.34 | 0.21 | 0.84 |
| Divergence | 0.11 | 0.03 | 0.02 | 0.06 | 0.16 | 0.02 | 0.81 | 0.00 | 0.41 | 0.73 | 0.64 | 0.56 |
| Diversity | 0.52 | 0.53 | 1.98 | 4.71 | 8.50 | 8.98 | 12.52 | 4.94 | 3.23 | 0.91 | 0.07 | 1.44 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.01 | 0.02 | 0.07 | 0.12 | 0.27 | 0.33 | 0.39 | 0.38 | 0.24 | 0.62 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

**RG**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RAL map | 1.79 | 4.04 | 12.42 | 14.08 | 25.45 | 17.84 | 14.48 | 10.32 | 5.67 | 3.11 | 3.02 | 0.49 |
| Quality | 1.64 | 0.74 | 0.22 | 1.71 | 5.11 | 8.45 | 10.35 | 2.62 | 1.60 | 0.03 | 0.01 | 0.66 |
| Exons | 0.50 | 0.96 | 0.04 | 0.74 | 0.02 | 0.17 | 0.24 | 1.58 | 0.37 | 1.24 | 0.36 | 0.18 |
| GC | 0.30 | 2.05 | 1.12 | 1.03 | 1.33 | 0.58 | 0.31 | 0.13 | 0.07 | 0.38 | 0.78 | 1.44 |
| Divergence | 0.04 | 0.14 | 0.61 | 0.97 | 1.22 | 0.12 | 0.11 | 0.30 | 1.10 | 0.80 | 0.36 | 0.24 |
| Diversity | 3.33 | 5.91 | 3.39 | 4.85 | 13.05 | 22.64 | 18.24 | 12.28 | 9.12 | 3.38 | 3.33 | 1.89 |
| Adjusted $r^2$ | 0.00 | 0.00 | 0.01 | 0.02 | 0.09 | 0.18 | 0.29 | 0.47 | 0.52 | 0.53 | 0.67 | 0.75 |
| Scale (kb) | 0.5 | 1 | 2 | 4 | 8 | 16 | 32 | 64 | 128 | 256 | 512 | 1024 |

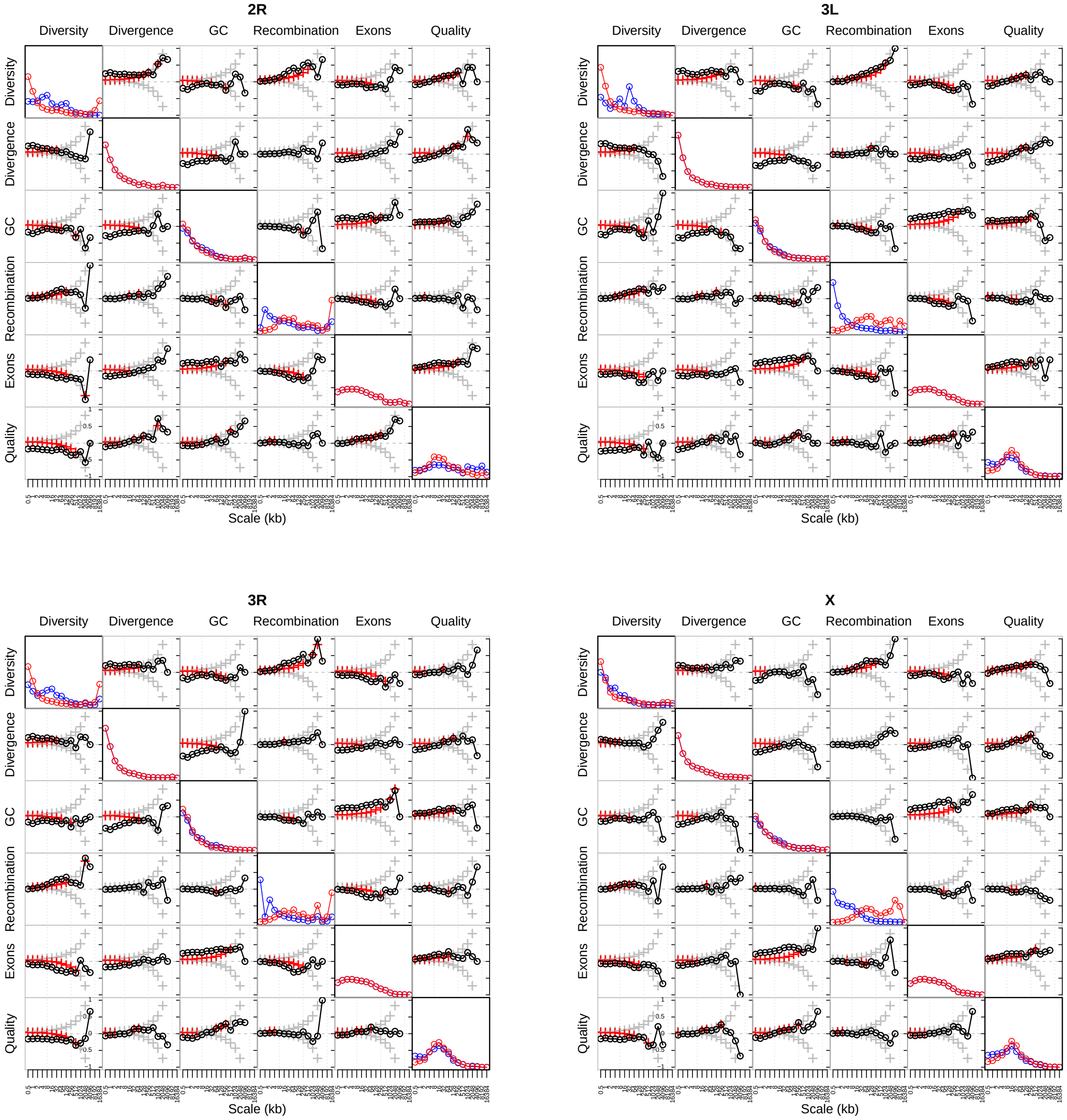

Figure 2.25: **Global wavelet power spectrum and pairwise correlations of detail wavelet coefficients of RAL and RG data for chromosome arms 2R, 3L, 3R, and X.** Diagonal plots show the global wavelet power spectrum of each feature of the RAL (blue) and RG (red) data. Off-diagonal plots show Kendall's rank correlation between pairs of detail coefficients at each scale, with respect to the wavelet decomposition of the two indicated features. Crosses denote the correlation that would be required for significance at the 1% level in a two-tailed test; red crosses are those scales at which the correlation is in fact significant. The lower left triangle and upper right triangle of plots correspond to RAL and RG, respectively.

# Chapter 3

# Particle Filtering in the SMC

## 3.1 Introduction

We use a special case of sequential importance sampling (SIS) called particle filtering [21] to perform Bayesian inference on haplotype data under the sequentially Markov coalescent (SMC) [58, 14]. We propose genealogies with at most one mutation to reduce the state space of the inference procedure, and consider only the segregating sites and the number of bases between them. We further assume that the ancestral allele is known, though it is straightforward to place a prior on the ancestral allele and adjust the inference accordingly.

Consider a discrete-time Markov process $\{X_j\}_{j=1}^{L}$, where $X_j$ is the two-locus ARG for SNPs $j$ and $j-1$, and

$$\begin{gathered} X_1 \sim \mu(x_1) \\ X_j \mid (X_{j-1} = x_{j-1}) \sim f(x_j \mid x_{j-1}). \end{gathered}$$

The variable $j$ indexes the SNPs in the haplotype data, which consists of $L+1$ SNPs, starting from SNP 0 and ending at SNP $L$.

We wish to estimate $\{X_j\}$ but only observe $\{Y_j\}_{j=0}^{j=L}$. Each haplotype is in $\{0,1\}^{L+1}$, where 0 denotes an ancestral allele and 1 denotes a derived allele. We assume that given $\{X_j\}$, the observations $\{Y_j\}$ are statistically independent and their marginal densities are given by

$$Y_j \mid (X_j = x_j) \sim g(y_j \mid x_j).$$

We assume a constant-sized population and that the population-scaled mutation rate $\theta$ and the population-scaled recombination rate $\rho$ are known.

## 3.2 Sampling Procedure

Particle filtering requires sampling independent particles from a proposal distribution and weighting them according to an importance weight. The weighted particles, with normalized weights, serve as an approximation to the target distribution. The sampling algorithm operates as follows.

For $n = 1$:

1. Sample trees $T_0^i$ conditioned on $y_0$ according to Wiuf and Donnelly [83].
2. Apply the transition function to $T_0^i$ to obtain $X_1^i$.
3. Compute the weights
$$w_1(X_1^i) = \frac{g(y_1 \mid X_1^i)\mu(X_1^i)}{q(X_1^i)}.$$
and let $W_1^i$ be the normalized weights.

For $j = 2 \dots L$:

1. Sample $X_j^i \sim q(X_j \mid y_j, X_{j-1}^i)$.
2. Compute the incremental weights
$$\alpha_j(X_{j-1:j}^i) = \frac{g(y_j \mid X_j^i) f(X_j^i \mid X_{j-1}^i)}{q(X_j^i \mid y_j, X_{j-1}^i)}.$$
and let $W_j^i$ be the normalized weights.

Denote by $t_{j,l}$ the first (or left) marginal tree of $x_j$ and by $t_{j,r}$ the second (or right) marginal tree of $x_j$. Then $g(y_j \mid x_j)$ is

$$\begin{aligned} g(y_j \mid x_j) &= p(y_j \mid t_{j,r}) \\ &= \frac{\gamma_j}{\bar{t}_{j,r}} \left[ \frac{\theta}{2} \bar{t}_{j,r} e^{-\frac{\theta}{2}\bar{t}_{j,r}} \right] \\ &\approx \frac{\gamma_j}{\bar{t}_{j,r}} \left[ \frac{\frac{\theta}{2}\bar{t}_{j,r}}{1 + \frac{\theta}{2}\bar{t}_{j,r}}, \right] \end{aligned}$$

where $\gamma_j$ is the length of the branch subtending all and only the derived alleles in $t_{j,r}$, and $\bar{t}_{j,r}$ is the total branch length of the tree $t_{j,r}$. The approximation is made assuming $\theta$ is small. One way to view the approximation is as the probability of one mutation occurring conditioned on the event that at most one mutation occurs. The transition probability $f(x_j \mid x_{j-1})$ is

$$f(x_j \mid x_{j-1}) = p(x_j \mid t_{j-1,r}) = \frac{p(x_j)}{p(t_{j-1,r})}.$$

The incremental weight $\alpha_j(x_{j-1:j})$ is given by

$$\begin{aligned}\alpha_j(x_{j-1:j}) &= \frac{g(y_j \mid x_j) f(x_j \mid x_{j-1})}{q(x_j \mid y_j, x_{j-1})}\\ &= \frac{p(y_j \mid t_{j,r}) p(x_j \mid t_{j-1,r})}{q(x_j \mid y_j, x_{j-1})}\\ &= \frac{\gamma_j}{\bar{t}_{j,r}}\left[\frac{\theta}{2}\bar{t}_{j,r} e^{-\frac{\theta}{2}\bar{t}_{j,r}}\right]\frac{p(x_j)}{p(t_{j-1,r}) q(x_j \mid y_j, x_{j-1})}\\ &\approx \frac{\gamma_j}{\bar{t}_{j,r}}\left[\frac{\frac{\theta}{2}\bar{t}_{j,r}}{1+\frac{\theta}{2}\bar{t}_{j,r}}\right]\frac{p(x_j)}{p(t_{j-1,r}) q(x_j \mid y_j, x_{j-1})}\end{aligned}$$

The prior on the initial state, $\mu(x_1)$, is the likelihood of $x_1$,

$$\mu(x_1) = p(x_1),$$

and the weight $w_1$ is given by

$$\begin{aligned}w_1(X_1^i) &= \frac{g(y_1 \mid X_1^i)\mu(X_1^i)}{q(X_1^i)}\\ &= \frac{\gamma_j}{\bar{t}_{j,r}}\left[\frac{\frac{\theta}{2}\bar{t}_{j,r}}{1+\frac{\theta}{2}\bar{t}_{j,r}}\right]\frac{p(x_1)}{q(x_1)}.\end{aligned}$$

The following sections describe the proposal distribution we use in the particle filtering.

## 3.3 Continuous-time Markov Chain

The process generating ARGs can be viewed as a continuous-time Markov chain since the process backward in time depends only on the current state. We begin with a description of the state space, describe the construction of the infinitesimal generator, and demonstrate applications of the generator. In the following, an ARG specifically refers to a two-locus ARG unless otherwise noted.

### 3.3.1 State Space

Let $[n]$, where $n$ is a positive integer, be the set of integers from 1 to $n$. Let $\mathcal{B}$ be the power set of $[n]$ and define $\mathcal{Y}$ to be subsets $Y \subseteq \mathcal{B}$ where

$$\begin{aligned}&\forall y, y' \in Y, y \neq y' : y \cap y' = \emptyset\\ &\bigcup_{y \in Y} y = [n]\\ &\forall y \in Y : y \neq \emptyset\end{aligned}$$

In other words, $Y$ is a partition of $[n]$ and $\mathcal{Y}$ is the set of possible partitions. The elements of $\mathcal{Y}$ are called tree states and represent the state of a tree at any given point in time.

To incorporate the data into the Markov chain, define a *color* to be an element in $\mathcal{C} = \{\emptyset, 0, 1\}$, where 0 indicates an ancestral lineage or allele, 1 indicates a derived lineage or allele, and $\emptyset$ indicates an *uncolored* lineage or allele. Then the data for a given site can be defined as $D \in \mathcal{C}^n$, i.e. the alleles of the individuals at a given site. In this context, we condition only on the data for the right tree of the ARG.

For notational purposes, we will need to define the following projection functions: $\mathcal{P}_\varsigma((l, r, c)) = l$ and $\mathcal{P}_\Omega((l, r, c)) = r$, where $(l, r, c) \in \mathcal{V} = \mathcal{B} \times \mathcal{B} \times \mathcal{C}$. Informally, these projection functions extract the left tree state and the right tree state, respectively, from an ARG state, which is defined formally in the following. Furthermore, define $C((l, r, c)) = c$ to be a function that extracts the color from $(l, r, c) \in \mathcal{V}$.

Define $\mathcal{S}$ to be subsets $S \subseteq \mathcal{V}$ where

$$\begin{aligned}
&\forall s, s' \in S, s \neq s' : \mathcal{P}_\varsigma(s) \cap \mathcal{P}_\varsigma(s') = \emptyset \\
&\forall s, s' \in S, s \neq s' : \mathcal{P}_\Omega(s) \cap \mathcal{P}_\Omega(s') = \emptyset \\
&\bigcup_{s \in S} \mathcal{P}_\varsigma(s) = [n] \\
&\bigcup_{s \in S} \mathcal{P}_\Omega(s) = [n] \\
&\forall s \in S : \mathcal{P}_\varsigma(s) \neq \emptyset \vee \mathcal{P}_\Omega(s) \neq \emptyset \\
&\forall s \in S : \mathcal{P}_\Omega(s) = \emptyset \to C(s) = \emptyset \\
&\forall s \in S : \mathcal{P}_\Omega(s) \neq \emptyset \to C(s) \in \{0, 1\}
\end{aligned}$$

The elements of $\mathcal{S}$ are called ARG states and serve as the states in the continuous-time Markov chain for generating ARGs. Another way to look at the ARG state is as a bipartite matching between blocks of two partitions, where every block is labeled with a color (possibly the null color). The blocks represent lineages, where the elements of each block indicate the lineages that have coalesced to produce the combined lineage. It is possible that a block on one side of the ARG will not be matched with a block from the other side of the block. In that case, it is matched with the null set $\emptyset$ to indicate that it is not a joint lineage.

### 3.3.2 Infinitesimal Generator

Define the infinitesimal generator $Q$ as follows. Let the components of $Q$ be indexed by $(S, S') \in \mathcal{S} \times \mathcal{S}$.

Define $s \cup s'$ for $s, s' \in \mathcal{V}$ as

$$s \cup s' = (l, r, c) \cup (l', r', c') = \begin{cases} (l \cup l', r \cup r', c') & \text{if } c = c' \\ (l \cup l', r \cup r', c') & \text{if } c = \emptyset \\ (l \cup l', r \cup r', c) & \text{if } c' = \emptyset \\ \emptyset & \text{otherwise} \end{cases}$$

Intuitively, this is corresponds to the coalescence of two lineages. The color must match for two lineages to coalesce, and once coalesced, a new lineage is formed and is represented by the set of leaves contained by the lineage.

The components of the infinitesimal generator are then as follows.

$$Q_{S,S'} = \begin{cases} 1, & \text{if } \exists s \in S, u \in S, s' \in S' : s \cup u = s' \wedge S \backslash \{s, u\} = S' \backslash \{s'\}, \\ \frac{\rho}{2}, & \text{if } \exists s \in S, s' \in S', u' \in S' : s = s' \cup u' \wedge S \backslash \{s\} = S' \backslash \{s', u'\}, \\ \frac{\theta}{2}, & \exists s \in S, s' \in S' : \widehat{C}(s) = 1 \wedge \widehat{C}(s') = 0 \wedge \mathcal{P}_\zeta(s) = \mathcal{P}_\zeta(s'), \\ & \qquad\qquad \wedge \mathcal{P}_\Omega(s) = \mathcal{P}_\Omega(s') \wedge S \backslash \{s\} = S' \backslash \{s'\}, \\ -\sum_{S''} Q_{S,S''}, & \text{if } S = S', \\ 0, & \text{otherwise.} \end{cases}$$

The above formula defines the infinitesimal generator of the continuous-time Markov chain for constructing an ARG jointly with the data at its right locus. The first case with rate 1 represents a coalescence event, the second case with rate $\rho/2$ represents a recombination event, and the third case with rate $\theta/2$ represents a mutation event. These events are transitions in the continuous-time Markov chain.

An ARG can be constructed from a continuous-time Markov chain of ARG states. The initial state is $w_1^a = \{(\{i\}, \{i\}, D_i)\}_{i=1}^n$ and the transitions follow $Q$. Recall that $D \in \mathcal{C}^n$ and represents the data at the right locus of the ARG. A realization of the continuous-time Markov chain with infinitesimal generator $Q$ generates an ARG from the joint distribution of the ARG and the data at its right locus. By running the Markov chain from the initial state to a state containing only one lineage on each side of the tree, one can generate an ARG from distribution on ARGs conditioned on the data. By using this infinitesimal generator in the following sections, we compute several densities and likelihoods that are useful for the proposal distribution.

## 3.4 ARG Densities

With $Q$ in hand, we can now compute several quantities of interest. However, we first need to define some additional notation. For joint event $i$, define the transition of the event to be $w_i = (w_i^t, w_i^a, w_i^b) \in \mathbb{R} \times \mathcal{S} \times \mathcal{S}$, where $w_i^t$ is the waiting time until the event occurs, $w_i^a$ is the source state, and $w_i^b$ is the destination state. Define a sequence of joint events as $\boldsymbol{w} = \{w_i\}_{i=1}^m$.

Define the following additional projection functions. For $X \in \mathcal{S}$, let $\widehat{\mathcal{P}}_\zeta(X) = \{\mathcal{P}_\zeta(x) \mid x \in X\} \in \mathcal{V}$ and let $\widehat{\mathcal{P}}_\zeta(w_i) = (w_i^t, \widehat{\mathcal{P}}_\zeta(w_i^a), \widehat{\mathcal{P}}_\zeta(w_i^b)) \in \mathbb{R} \times \mathcal{Y} \times \mathcal{Y}$. These projection functions extend the previous projection functions to operate on transitions (pairs of states). Finally, given a matrix $A$, let $A_{i,j}$ denote component $(i, j)$ of the matrix. For an ARG described by $\boldsymbol{w}$,

$$p(\boldsymbol{w}, D) = \prod_{i=1}^m Q_{w_i^a, w_i^b} \exp(-Q_{w_i^a, w_i^b} w_i^t).$$

$m$ is the total number of transitions in $\boldsymbol{w}$. As an aside, $m$ can be arbitrarily large because the transitions in $\boldsymbol{w}$ might be recombination events, and the number of recombination events in an ARG is unbounded. Again, recall that $D \in \mathcal{C}^n$ represents the alleles at the right locus of the ARG. The above equation follows from the product of the densities of exponential waiting times.

We now wish to compute the likelihood for a set of ARGs instead of a single ARG. We will specify the set of ARGs with $\boldsymbol{h}$, a subset of events describing an ARG, and integrate over all the ARGs consistent with $\boldsymbol{h}$. Namely, for a given $\boldsymbol{h}$,

$$p(\boldsymbol{h}, D) = \int_{f(\boldsymbol{h})} p(\boldsymbol{w}, D) d\boldsymbol{w}$$

where

$$f(\boldsymbol{h}) = \{\boldsymbol{w} \mid \forall h \in \boldsymbol{h}, \exists w \in \boldsymbol{w} : h = w\}.$$

$f(\boldsymbol{h})$ is the set of ARGs consistent with $\boldsymbol{h}$.

To compute the above integral, we will use matrix exponentiation as in [35, 55]. Let $M^{(t)} = \exp(Qt)$. Then we have

$$p(\boldsymbol{h}, D) = \prod_{i=1}^{m} \sum_{\substack{r \in \mathcal{S} \\ r \neq h_i^b}} M_{h_i^a, r}^{h_i^t} Q_{r, h_i^b},$$

where $m$ is the number of events in the ARG.

We now need to integrate over all ARGs consistent with a given left tree. For marginal event $i$, define the transition to be $v_i = (v_i^t, v_i^a, v_i^b) \in \mathbb{R} \times \mathcal{Y} \times \mathcal{Y}$. Define a sequence of marginal events as $\boldsymbol{v} = \{v_i\}_{i=1}^m$. For $\boldsymbol{v}$ describing the tree at the left locus of an ARG,

$$p(\boldsymbol{v}, D) = \int_{g(\boldsymbol{v})} p(\boldsymbol{w}, D) d\boldsymbol{w}$$

where

$$g(\boldsymbol{v}) = \{\boldsymbol{w} \mid \forall v \in \boldsymbol{v}, \exists w \in \boldsymbol{w} : v = \widehat{\mathcal{P}}_\zeta(w)\}.$$

$g(\boldsymbol{v})$ is the set of ARGs consistent with a given left tree described by $\boldsymbol{v}$. To compute the above integral using matrix exponentiation, define

$$\xi(\boldsymbol{v}) = \{\boldsymbol{h} \mid \forall v \in \boldsymbol{v}, \exists h \in \boldsymbol{h} : v = \widehat{\mathcal{P}}_\zeta(h) \wedge v^t = h^t\}.$$

Then we have

$$p(\boldsymbol{v}, D) = \sum_{\boldsymbol{h} \in \xi(\boldsymbol{v})} p(\boldsymbol{h}, D). \tag{3.1}$$

This can be computed efficiently using dynamic programming by recording the probabilities for $h \in \mathcal{S}$ for every $v_i^t$ and using the Markov property.

## 3.5 Conditioned Paths

The first step in sampling from the proposal distribution is to sample joint transitions conditioned on the left tree. For every transition $v$ in $\boldsymbol{v}$ describing the left tree, there is a set of ARG transitions $w$ consistent with $v$,

$$\eta(v) = \{w \in \mathcal{V} \mid \widehat{\mathcal{P}}_\zeta(w) = v\}$$

The computation of (3.1) requires recording the probabilities for $h \in \mathcal{S}$ for every $v_i^t$. These can be used to compute the probability of being in a given state at some $v_i^t$. The density of starting in $S$ and making a transition from any state to $S'$ at time $t$ is

$$p(S \xrightarrow{t} S', D) = \sum_{\substack{r \in \mathcal{S} \\ r \neq S'}} M_{S,r}^t Q_{r,S'}.$$

In the computation of (3.1), we constructed a dynamic programming table that records the probabilities for every ARG state indexed by the times in $\boldsymbol{v}$, meaning that we have a separate set of probabilities for every $v_i^t$. Recall that $\boldsymbol{v}$ is the sequence of coalescence transitions that describe the left of the ARG, $m$ is the number of elements in the sequence. In the dynamic programming table, at each time $v_i^t$, which are the times of the coalescence transitions in the left tree of the ARG, we record the probability of being any state in $\mathcal{S}$, which is the space of all ARG states. (In practice, we only record states relevant to the transitions in $\boldsymbol{v}$, since conditioned on $\boldsymbol{v}$, many states are will have zero probability and do not need to be recorded.) To sample $w_m$, we sample a final state from this table of probabilities indexed at time $v_m^t$ (the last time of $\boldsymbol{v}$). Conditioned on the final state ($S_m$), we sample the previous state $S_{m-1}$ at time $m-1$ from the following distribution,

$$p(S_{m-1}, D) \propto \prod_{i=1}^{m-1} \sum_{\substack{r \in \mathcal{S} \\ r \neq S_m}} M_{S_{m-1},r}^{v_{m-1}^t} Q_{r,S_m}.$$

Once $S_{m-1}$ is sampled, we recurse to sample the rest of $S_i$ for $i = 1, \ldots, m-2$. These $S_i$ then lead directly to $\boldsymbol{w}$ because any consecutive pair of $S_i, S_{i+1}$ defines a transition in the Markov chain.

Note that this is only a subset of the entire number of transitions needed to define an ARG uniquely, and this is where we use matrix exponentiation to integrate over all the other possible events.

The normalization constant is computed by summing over the joint densities, which is straightforward to execute. Therefore, this recursively samples conditioned transitions from the final state back to the initial state. Although we now have a subset of the transitions sampled, we still need to sample the remaining of the transitions in the ARG to produce a particle. To do this, we will use a technique called uniformization.

## 3.6 Uniformization

Uniformization [36] is a technique for transforming a continuous-time Markov chain into a discrete-time Markov chain. Once in discrete-time form, a Markov chain is often easier to analyze or use. In our context, we use uniformization to sample a path in a continuous-time Markov chain conditioned on the initial state and the final state, and the total time of the path $t$.

Define

$$P = I + \frac{1}{\lambda} Q,$$

where $I$ is the identity matrix and $\lambda$ is

$$\lambda = \max_{S} |Q_{S,S}|.$$

Let $Z$ be the discrete-time Markov process, and let the initial state be $S_0$ and the final state be $S_f$. Then the number of transitions $N$ occurring in time $t$ is

$$\mathbb{P}(N = n \mid Z_0 = S_0, Z_n = S_f) = e^{-\lambda t} \frac{(\lambda t)^n}{n!} \frac{P^n_{S_0,S_f}}{M^t_{S_0,S_f}}.$$

The times of the transitions $t_1, \ldots, t_n$ are uniformly distributed over $[0, t]$. The transition probabilities are

$$\mathbb{P}(Z_i = S_i \mid Z_{i-1} = S_{i-1}, Z_n = S_f) = \frac{P_{S_{i-1},S_i} P^{n-i}_{S_i,S_f}}{P^{n-i+1}_{S_{i-1},S_f}}$$

Given the initial and final states, and the time between the two states, one can sample a path satisfying the boundary conditions using the above transition probabilities.

## 3.7 Proposal Distribution

The procedure for sampling from the proposal distribution is summarized as follows.

1. Use matrix exponentiation to compute the quantities described above.
2. Sample a conditioned path given the left tree.
3. Conditioned on this path, use uniformization to sample the remaining joint transitions.
4. Project the resulting ARG onto the right tree.

This produces a particle from the optimal proposal distribution

$$p(T_n | D_n, T_{n-1})$$

for $n \geq 2$. For $n = 1$, we use the method from [83], described in Section 3.9, to sample directly from the posterior.

## 3.8 Importance Weights

Because we sample from the optimal proposal distribution, the importance weights for sites $n \geq 2$ are

$$\begin{aligned}
\frac{p(D_n, T_n \mid T_{n-1})}{q(T_n \mid D_n, T_{n-1})} &= \frac{p(D_n, T_n \mid T_{n-1})}{p(T_n \mid D_n, T_{n-1})} \\
&= \frac{p(D_n, T_n, T_{n-1}) p(D_n, T_{n-1})}{p(T_{n-1}) p(T_n, D_n, T_{n-1})} \\
&= \frac{p(D_n, T_{n-1})}{p(T_{n-1})}
\end{aligned}$$

It is straightforward to compute $p(T_{n-1})$. We can compute $p(D_n, T_{n-1})$ using (3.1) since $p(D_n, T_{n-1}) = p(\boldsymbol{v}, D_n)$, where $T_{n-1}$ is described by $\boldsymbol{v}$.

## 3.9 Initial Tree

Wiuf and Donnelly [83] provide a way to sample from the posterior distribution on trees conditioned on one mutation occurring on the tree. Let $j$ be the number of derived lineages remaining, and let $k$ be the total number of lineages remaining (both ancestral and derived). When $j \geq 2$, two ancestral lineages coalesce with probability

$$\frac{k - j - 1}{k - 1}$$

and two derived lineages coalesce with probability

$$\frac{j}{k - 1}.$$

When one derived lineage remains, the probability that two ancestral lineages coalesce is

$$\frac{k - 2}{k - 1}$$

and the probability that the remaining derived lineage mutates to an ancestral lineage is

$$\frac{1}{k - 1}.$$

The times between transitions are exponentially distributed with rate

$$\binom{k}{2}.$$

We marginalize over the mutation event as follows. First consider the proposal density without marginalization. The density is the product of the following components: the jump chain transitions, the waiting times, the ancestral lineage selection probabilities, and the derived lineage selection probabilities.

The jump chain transitions and the derived lineage selection probabilities are the same regardless of the tree (for a fixed number of derived alleles and ancestral alleles in the sample). Namely, the probability of the jump chain transitions is

$$\frac{j(j-1)\dots 1\cdot(l-1)(l-2)\dots 1}{(j+l-1)(j+l-2)\dots 1}, \tag{3.2}$$

where $j$ is the initial number of derived alleles in the sample and $l$ is the initial number of ancestral alleles (hence $j+l=n$, where $n$ is the sample size). The probability of selecting the pairs of derived lineages for coalescence is

$$\binom{j}{2}^{-1}\binom{j-1}{2}^{-1}\dots\binom{2}{2}^{-1}. \tag{3.3}$$

The waiting time for epoch $r$, where $r$ indexes the epochs from the present to the past, is the sum of two exponential random variables with rate $\binom{k_r}{2}$, where $k_r$ is the number of remaining lineages in the epoch.

The density of the waiting times is

$$\begin{aligned}&\binom{n}{2}\binom{n-1}{2}\dots\binom{k_{r^*}}{2}\binom{k_{r^*}}{2}\binom{k_{r^*}-1}{2}\dots\binom{2}{2}\\ &\times\exp\left(-\left[\binom{n}{2}u_1+\binom{n-1}{2}u_2+\dots+\binom{k_{r^*}}{2}u_{r_1^*}+\binom{k_{r^*}}{2}u_{r_2^*}\right.\right.\\ &\qquad\left.\left.+\binom{k_{r^*}-1}{2}u_{r^*+1}+\dots+\binom{2}{2}u_{n-1}\right]\right),\end{aligned} \tag{3.4}$$

where $r^*$ is the epoch in which the mutation event occurs, and $u_i$ is the waiting time for epoch $i$. $u_{r_1^*}$ is the waiting time from the coalescence event just before the mutation until the mutation occurs, and $u_{r_2^*}$ is the waiting time from the time of the mutation until the next coalescence event. (Note that the mutation event does not start another epoch.)

The probability of selecting the pairs of ancestral lineages for coalescence is

$$\binom{l}{2}^{-1}\binom{l-1}{2}^{-1}\dots\binom{k_{r^*}}{2}^{-1}\binom{k_{r^*}}{2}^{-1}\binom{k_{r^*}-1}{2}^{-1}\dots\binom{2}{2}^{-1}. \tag{3.5}$$

Consider the epochs in which the mutation can occur for a given tree. These epochs are the ones starting from the time when one derived lineage remains until the time this lineage coalesces with the rest of the tree (after it experiences a mutation event). Supposing that the mutation occurs in epoch $r'$, the density of the waiting time $u_{r'}$ is

$$\binom{k_{r'}}{2}^2 u_{r'}e^{-\binom{k_{r'}}{2}u_{r'}}.$$

The proposal density $q_{\text{initial}}$ for the initial tree where the mutation event occurs in epoch $r^*$ is the product of the probabilities for the jump chain (3.2), derived lineage pair selection (3.3), ancestral lineage pair selection (3.5), and the waiting times (3.4). The marginalized proposal density is

$$\frac{q_{\text{initial}} \cdot \binom{k_{r^*}}{2}}{\binom{k_{r^*}}{2}} \sum_{r'} \frac{\binom{k_{r'}}{2}^2 u_{r'} \exp\left(-\binom{k_{r'}}{2} u_{r'}\right)}{\binom{k_{r'}}{2} \cdot \binom{k_{r'}}{2} \exp\left(-\binom{k_{r'}}{2} u_{r'}\right)} = q_{\text{initial}} \sum_{r'} u_{r'} = q_{\text{initial}} \cdot \gamma,$$

where $r'$ is over all the epochs in which the mutation event can occur, and $\gamma$ is the length of the branch subtending all the derived alleles. Therefore, the importance weight, integrated over the mutation event, is simply the non-integrated importance weight multiplied by $\gamma$.

## 3.10 Systematic Resampling

Systematic resampling is a resampling method that attempts to minimize the variance of the importance weights. Rather than resampling according to the multinomial distribution, systematic resampling proceeds by sampling

$$U_1 \sim \text{Uniform}([0, 1/N])$$

and defining

$$U_i = U_1 + \frac{i-1}{N}$$

for $1 \leq i \leq N$, where $N$ is the number of particles. The number of times particle $j$ is resampled is then

$$N_j = \left| \left\{ U_k \middle| \sum_{i=1}^{j-1} W_j \leq U_k < \sum_{i=1}^{j} W_j \right\} \right|,$$

where $W_j$ are the normalized particle weights.

This resampling approach is straightforward to implement and provides good performance in a variety of situations. We perform resampling whenever the effective sample size (ESS) falls below $N/2$, where the ESS is defined to be

$$ESS = \frac{1}{\sum_j W_j^2}.$$

For the initial site, because we use the method in [83] to sample from the optimal proposal distribution, the ESS is always $n$, which is the heighest achievable ESS. For subsequent sites, the ESS depends on the variance of $p(D_n \mid T_{n-1})$, which could potentially be large, leading to a lower ESS. Intuitively, if the data at site $n$ cannot be explained by the tree at the previous site, then the particle filter will suffer from greater degeneracy. In other words, particle filtering works best when the sequential distributions across sites do not vary too quickly.

## 3.11 Results

To evaluate the particle filtering method, we generated datasets using Hudson's `ms` coalescent simulator [37]. We fixed the parameters to $\rho = 0.01$ and $\theta = 0.001$, with a constant population size. We also conditioned on the number of segregating sites by repeatedly generating datasets until a dataset of the desired number of segregating sites was obtained. Note that this is different from the option in `ms` to "condition" on the number of segregating sites, as `ms` will first generate a tree and place ea fixed number of mutations on it. This does not produce the trees from the distribution conditioned on a given number of segregating sites. We conditioned on 5 segregating sites and used a sample size of 4. We used a particle filter wtih 1000 samples and compared the the estimate from using data at only one site compared to conditioning on all the data. The summary statistics we used were tree length (the sum of the branch lengths), time to most recent common ancestor (TMRCA), and the expected age of of the mutation. Note that given a genealogy and assuming exactly one mutation, the expected age of the mutation is the average between the TMRCA of the derived alleles and the time when the MRCA of the derived alleles coalesces with an ancestral lineage. Figures 3.1, 3.2, and 3.3 show the posterior distributions for a single dataset.

The results show that the posterior distribution with this sample size is relatively dispersed, and that the actual variance of the summary statistics computed above is fairly high. However, as one would expect, the use of more data to infer the properties of the distribution on genealogies at any given site provides more accurate posterior inference. Conditioning on a single site does not incorporate linkage disequilibrium information present in the data set. When the recombination rate is low, the effects of linkage disequilibrium are stronger, and the patterns of variation from the linkage disequilibrium can provide supporting evidence for the posterior distribution.

Table 3.1 compares between the particle filtering and conditioning on a single site the absolute relative error of the posterior mean with respect to the tree length, the TMRCA, and the expected age of mutation, averaged over 100 datasets. The estimates from the particle filtering can still have high variance due to the nature of the model and data. That is, even the exact posterior distribution potentially has high variance and any estimate will deviate somewhat from the truth. Due to the computational expense of the particle filtering, conditioning on a dataset with a large sample size and many segregating sites would require additional algorithmic improvements to scale the method.

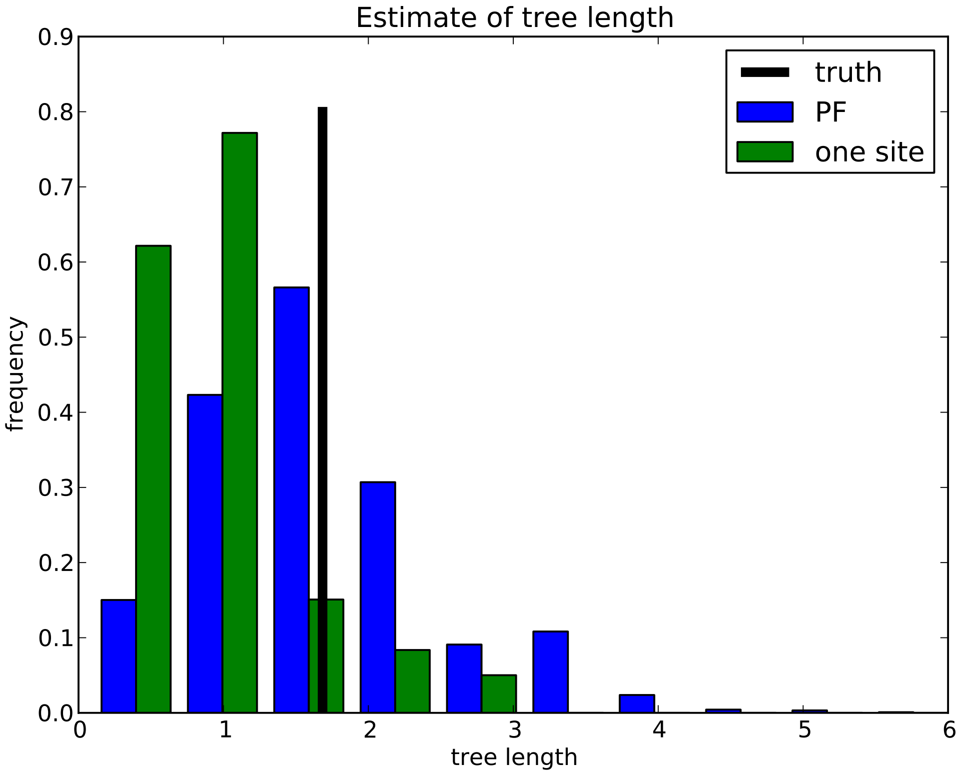

Figure 3.1: **Estimate of tree length using particle filtering.** The histogram shows the posterior distribution on the tree length of the last site conditioned on all the data, using particle filtering (PF), shown in blue; the posterior distribution on tree length conditioned only on the last site, shown in green; and the true tree length, shown in black. The parameters are $\rho = 0.01$, $\theta = 0.001$, 5 segregating sites, a sample size of 4, and 1000 particles.

Table 3.1: **Absolute relative error of estimates.** The absolute relative error of the estimates for tree length, TMRCA, and expected age of mutation are shown, comparing the particle filtering method, which conditions on multiple sites, and the posterior estimate conditioned on a single site. The relative error is averaged over 100 datasets, with $\rho = 0.01$, $\theta = 0.001$, 5 segregating sites, sample size of 4, and 1000 particles.

| | Tree length | TMRCA | Expected age of mutation |
|---|---|---|---|
| PF | 0.362 | 0.370 | 0.260 |
| one site | 0.750 | 0.715 | 0.492 |

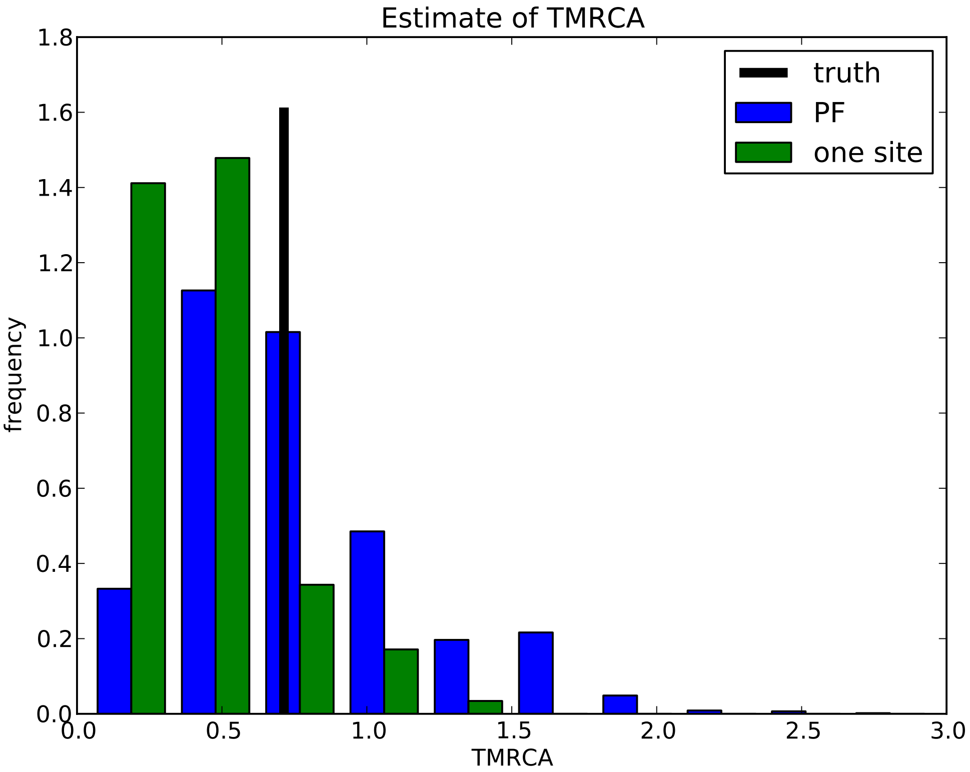

Figure 3.2: **Estimate of TMRCA using particle filtering.** The histogram shows the posterior distribution on the TMRCA of the last site conditioned on all the data, using particle filtering (PF), shown in blue; the posterior distribution on the TMRCA conditioned only on the last site, shown in green; and the true TMRCA, shown in black. The parameters are $\rho = 0.01$, $\theta = 0.001$, 5 segregating sites, a sample size of 4, and 1000 particles.

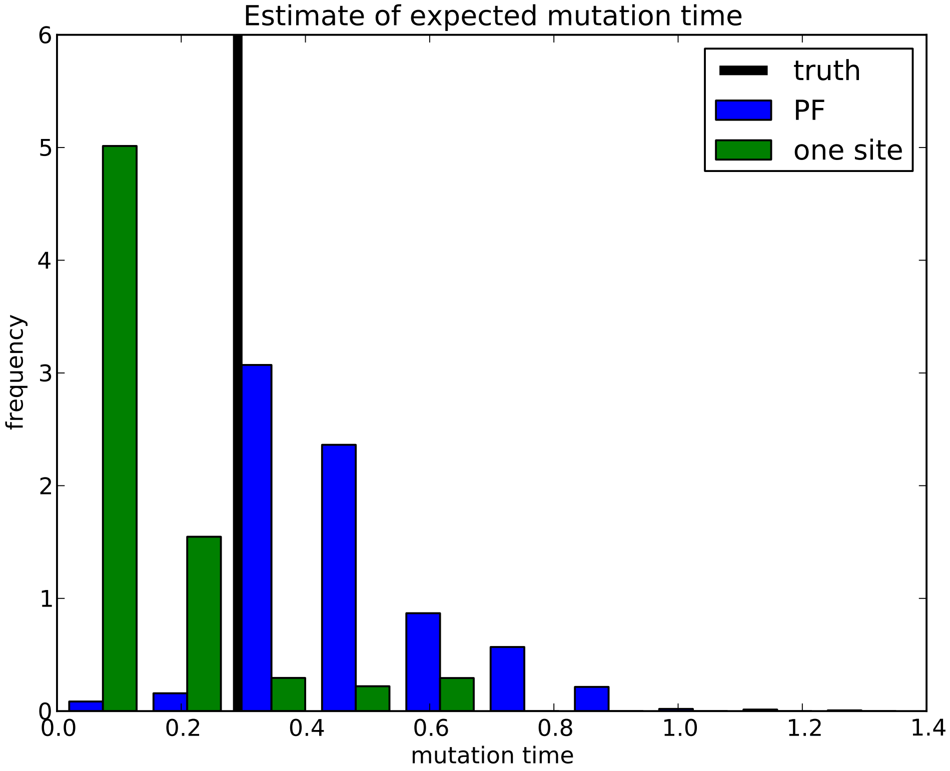

Figure 3.3: **Estimate of expected age of mutation using particle filtering.** The histogram shows the posterior distribution on the expected age of the mutation for the last site conditioned on all the data, using particle filtering (PF), shown in blue; the posterior distribution on the expected age of the mutation conditioned only on the last site, shown in green; and the true expected age of the mutation, shown in black. The parameters are $\rho = 0.01$, $\theta = 0.001$, 5 segregating sites, a sample size of 4, and 1000 particles.

# Chapter 4

# Discussion

We have developed a new method, LDhelmet, which is able to provide accurate estimates of recombination rates using genomic data from *D. melanogaster*. Although our focus has been on this species, the features of our method should offer improvements in the estimation of recombination in other species too. For example, the desire to efficiently incorporate sites in which some alleles are missing is a common issue when data are generated by next-generation sequencing technologies. We believe that our method will find many further applications in other datasets.

In addition, we have described a method using particle filtering for approximating posterior genealogies under the SMC. The particle filtering proposal distribution uses matrix exponentiation to integrate over the recombination events, and to compute the joint likelihood of a marginal tree and the data. These quantities are used to sample from the optimal proposal distribution, which is the posterior distribution on particles given the particle at the previous site and the current data. Although our method in its current form works only with small samples, we believe that with additional well-motivated approximations to the method, it can be scaled to larger sample sizes.

## 4.1 Population Comparison

Using our method based on the composite likelihood approximation, LDhelmet, we have performed a genome-wide comparison of fine-scale recombination rates between two populations of *D. melanogaster*, one from Raleigh, USA (labeled RAL) and the other from Gikongoro, Rwanda (labeled RG). While earlier studies have largely been confined to regions of moderate resolution, we find extensive fine-scale variation across all chromosomes and in both populations. A notable difference between the two recombination maps is the higher overall recombination rate in RG than in RAL. Our method estimates the composite parameter $2N_e r_f$, where $N_e$ is the effective population size and $r_f$ is the (female) rate of recombination per generation, so this difference is partly explained by a difference in effective population size. However, further differences between chromosomes—namely, the inflated recombination rates in the X chromosome relative to autosomes—lead us to invoke biological differences too, particularly the role of polymorphic inversions. There may also be other, unappreciated, biological factors causing an increase in $r_f$ on the X chromosome.

### 4.1.1 X Chromosome

In addition to the higher absolute rate of recombination in RG, a further difference between the populations merits discussion: the relative increase in recombination on the X chromosome compared to the autosomes is much more pronounced in RG than in RAL. In the African population, estimates of the ratio $\rho_X/\rho_A$ lie in the range $2.4 \sim 6.0$, whereas in the North American population they lie in the range $1.1 \sim 1.5$ (Table 2.11). There are several possible explanations for the difference between the two populations.

First, RAL may have experienced a historical population bottleneck. The effect of a population bottleneck on LD is stronger on the X chromosome than on the autosomes [81] (a similar effect on diversity is also seen [65]). Thus, a population bottleneck leads to an increase in LD on the X chromosome over and above the increase on the autosomes. A bottleneck in the non-African population is a sensible proposition since *D. melanogaster* is a human commensal of African origin which has colonized North America more recently. Bottlenecks in non-African populations of *D. melanogaster* have been inferred from genetic data by others [32, 78]. As shown in our simulation study, bottlenecks tend to cause our method to underestimate the true recombination rate, so the bottleneck explanation would be consistent with the fact that our recombination rate estimates for RAL are lower than that for RG. Second, the impact of polymorphic inversions may be greater in RG, since the African population has a high frequency of polymorphic inversions in the autosomes and in the centromere-proximal X. The observed increase in the recombination rate in the African X could be partially attributed to *interchromosomal effect* [53, 66]. A third possible explanation is the more efficient role of selection on the X chromosome when non-neutral mutations are recessive: such mutations can more easily be exposed to the action of selection in their hemizygous state in males. This effect will be more pronounced in RAL if it has undergone greater selective pressures, as seems likely in its adaptation to a new environment. Unraveling the relative importance of these possible explanations merits further investigation.

### 4.1.2 Fine-scale Differences

At fine-scales, we also find extensive differences between the recombination maps of the two populations, for which a simple difference in effective population size is not a sufficient explanation. Wavelet coherence analysis reveals high correlation at broad scales but regions of low correlation at fine scales, as has been documented among human populations, and in comparison between humans and chimpanzees [76, 47]. The advantage of a wavelet coherence approach is that it further identifies the locations of similarities and differences. However, the causes of these differences remain to be understood. One noteworthy result of our analysis is that changes in diversity are a strong positive predictor of changes in recombination in one population, even when the recombination map of the other population is included as a covariate. A possible explanation for this observation is that the two populations have undergone separate selective sweeps, with sufficient impact on the genome that the correlation between recombination and diversity can still be detected even when the recombination map of the other population is used as a covariate. We note that a partial overlap in the signature of selective sweeps was also found by Langley *et al.* [49]. Using a metric based on valleys of diversity, they found that 44% of diversity valleys in RAL overlapped with those found in an African sample. There are of course other possible explanations for the observed correlations between diversity and recombination; it is known that background selection—the loss of neutral diversity due to linked *deleterious* mutations—can also induce such a correlation (see Charlesworth [15, 16] and references therein). The relative importance of these types of selection in distinguishing the two populations is obviously deserving of further study.

## 4.2 Recombination Hotspots

Access to a fine-scale map lets us address a crucial question of the distribution of recombination in *Drosophila*: whether they localize into recombination hotspots. Using a conservative approach, we found a few regions with solid statistical support for a local elevation of at least 10 times the background recombination rate (see Table 2.15). With the caveat that we used a high block penalty in the rjMCMC and employed a stringent hotspot detection strategy, overall our findings support the belief that extreme localization of recombination into hotspots is not prevalent in *D. melanogaster*; in humans, on the other hand, the list of well-supported hotspots exceeds 30,000 [77], many of which exhibit much more than a tenfold increase and have a common mechanism for recruiting the recombination machinery [7, 10, 63]. Singh *et al.*[71] therefore reserve the term “recombination peaks” for the milder variability they find, and it could be the case that what we have found are the most extreme examples of these peaks. Having said that, we also note that, as discussed earlier in our simulation study, the ability to perform accurate statistical inference of recombination (in particular, detecting hotspots) gets significantly reduced when recurrent strong selective sweeps are in play. It is hence possible that there are actually more hotspots in the *D. melanogaster* genome than our study could find.

## 4.3 Motifs

We have focused on estimating and characterizing the recombination map itself and on its correlation with a set of important genomic annotations, but given such a map one can tackle many further problems. The question of primary sequence influences of recombination localization can now be addressed with much greater power. In humans, the 13 bp motif CCNCCNTNNCCNC has been found to be over-represented in hotspots, consistent with its recruitment of the protein PRDM9 which has been implicated in the hotspot usage [10, 63]. Searches for motifs in *Drosophila* that correlate with fine-scale recombination rate have been undertaken in *D. pseudoobscura* [18, 48], *D. persimilis* [74], and *D. melanogaster* [61]. Motifs that correlate with fine-scale recombination in humans are also significant in some of these species [48, 74], which is unexpected given the rapid turnover of motif usage in humans and chimpanzees [63]. In a recent pedigree study, Miller *et al.* [61] were able to localize with high precision fifteen crossover events on the X chromosome of *D. melanogaster*. From these they identified the 7 bp motif GTGGAAA as significantly enriched in the vicinity of these crossovers. Further study is required to validate this motif and to search for others, and our maps should prove useful in this regard.

## 4.4 Natural Selection

Finally, our work should be of interest since a fine-scale recombination map is a prerequisite of studies seeking to estimate the influence of natural selection on the genome [34]; those lacking such a map retain this caveat [69]. Although these inferences of recombination and selection rely on the same data and have the potential to distort each other, it is reassuring that our method is robust to the influence of positive selection, and that it shows good agreement with existing experimental estimates of recombination. In our simulation studies we focused on the effects of hard sweeps, since they are thought to be an important mode of adaptation in *Drosophila* [43, 70, 69] and are expected to have the strongest effect on patterns of variation. Aside from additional noise resulting from a reduction in diversity, there is little bias introduced by failing to include selection in the assumed model, at least under the parameters we considered. This is consistent with the observation that a recurrent sweep model does not have a striking effect on LD beyond that predicted by the reduction in diversity [81]. Nonetheless, further investigation is warranted on the effects of other types of selection, and on the development of methods that can account for recombination and selection jointly.

## 4.5 Particle Filtering in the SMC

The method we employed for variable recombination rate estimation relies on the composite likelihood approximation. As demonstrated in in this thesis, this approach works particularly well for the inference of recombination rates. However, for the inference of other parameters of interest, such as population size, demography and selection, a composite likelihood approximation may be less effective and might in fact lead to significant biases. Furthermore, a simple and sensible interpretation of the underlying model that a composite likelihood approach implies is often difficult to find.

The SMC is a well-motivated approximation to the full coalescent. It begins by approximating the full coalescent by imposing the Markov condition along the genome. This has many benefits to tractability in addition to being motivated by several important reasons in the context of inference, as described in [58, 14]. The model furthermore circumvents one of the most prohibitive difficulties in the inference of any parameters of interest on full genome data: the significantly long length of the genome. While it does not simplify the matter of inference on large sample sizes, the full coalescent model rapidly becomes intractable. Under the SMC, however, the length of the genome in many cases will only be a linear term in the runtime of an inference algorithm.

We have shown that the use of matrix exponentiation as used in [35, 55] provides a method to sample directly from the optimal proposal distribution in the context of particle filtering. Although computationally intensive, several approximations can be made to improve tractability. One such approximation is to assume at most one mutation per site, which dramatically decreases the size of the state space. This is a reasonable assumption when dealing with low mutation rates. The recombination events in a two-locus ARG can then be integrated out of the likelihood, resulting in a much better proposal distribution within the particle filtering framework. Combined with uniformization [36], it is possible to sample from the optimal proposal distribution, which minimizes the variance of the importance weights.

Compared to a composite likelihood method, particle filtering is orders of magnitude slower, and our method in its current form can handle only small sample sizes. Hence, it cannot be used easily in a more general inference framework such as rjMCMC, which requires that every sample's likelihood be computed very quickly. Nonetheless, the approximation that particle filtering and the SMC provide could be considered much more rigorous and easier to analyze because it rests on an arguably stronger mathematical foundation.

The posterior distribution on genealogies can prove useful in certain analyses and can also be used to guide parameter inference. Although several methods exist for particle filtering to estimate parameters (see [21] for a review), they are not straightforward to apply to our problem. However, the posterior genealogies, even without direct parameter inference, still provide insight on likely parameters and can be useful in downstream analysis.

# Bibliography

[1] J. A. Anderson, Y. S. Song, and C. H. Langley. “Molecular population genetics of *Drosophila* subtelomeric DNA”. In: *Genetics* 178.1 (2008), pp. 477–487.

[2] P. Andolfatto. “Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome”. In: *Genome Research* 17 (2007), pp. 1755–1762.

[3] S. Aulard, J.R. David, F. Lemeunier, et al. “Chromosomal inversion polymorphism in Afrotropical populations of *Drosophila melanogaster*”. In: *Genetical Research* 79.1 (2002), pp. 49–63.

[4] A. Auton and G. McVean. “Recombination rate estimation in the presence of hotspots”. In: *Genome Research* 17 (2007), pp. 1219–1227.

[5] A. Auton et al. “A Fine-Scale Chimpanzee Genetic Map from Population Sequencing”. In: *Science* 336.6078 (2012), pp. 193–198.

[6] E. Axelsson et al. “Death of PRDM9 coincides with stabilization of the recombination landscape in the dog genome”. In: *Genome Research* 22 (2012), pp. 51–63.

[7] F. Baudat et al. “PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice”. In: *Science* 327 (2010), pp. 836–840.

[8] D. J. Begun and C. F. Aquadro. “Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*”. In: *Nature* 356 (1992), pp. 519–520.

[9] D. J. Begun et al. “Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*”. In: *PLoS Biology* 5.11 (2007), e310.

[10] I. L. Berg et al. “PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans”. In: *Nature Genetics* 42.10 (2010), pp. 859–863.

[11] A. Bhaskar, J. A. Kamm, and Y. S. Song. “Approximate sampling formulas for general finite-alleles models of mutation”. In: *Advances in Applied Probability* 44 (2012), pp. 408–428.

[12] A. Bhaskar and Y. S. Song. “Closed-form asymptotic sampling distributions under the coalescent with recombination for an arbitrary number of loci”. In: *Advances in Applied Probability* 44 (2012), pp. 391–407.

[13] H. Brunschwig et al. “Fine-scale maps of recombination rates and hotspots in the mouse genome”. In: *Genetics* 191.3 (2012), pp. 757–764.

[14] N. Cardin. “Approximating the coalescent with recombination”. PhD thesis. University of Oxford, 2006.

[15] B. Charlesworth. “The effects of deleterious mutation on evolution at linked sites”. In: *Genetics* 190 (2012), pp. 5–22.

[16] B. Charlesworth. “The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila X* chromosome”. In: *Genetics* 191 (2012), pp. 233–246.

[17] G.K. Chen, P. Marjoram, and J.D. Wall. “Fast and flexible simulation of DNA sequence data”. In: *Genome Research* 19.1 (2009), pp. 136–142.

[18] E. T. Cirulli, R. M. Kliman, and M. A. F. Noor. “Fine-scale crossover rate heterogeneity in *Drosophila pseudoobscura*”. In: *Journal of Molecular Evolution* 64 (2007), pp. 129–135.

[19] G. Coop et al. “High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans”. In: *Science* 319 (2008), pp. 1395–1398.

[20] Russell B. Corbett-Detig, Charis Cardeno, and Charles H. Langley. “Sequence-based detection and breakpoint assembly of polymorphic inversions”. In: *Genetics, in press* (2012). in press.

[21] A. Doucet and A. M. Johansen. “A tutorial on particle filtering and smoothing: fifteen years later”. In: *The Oxford handbook of nonlinear filtering.* Ed. by D. Crisan and B. Rozovskii. Oxford University Press, 2011.

[22] J. Drouaud et al. “Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination “hot spots””. In: *Genome Research* 16 (2006), pp. 106–114.

[23] S. N. Ethier and R. C. Griffiths. “On the two-locus sampling distribution”. In: *Journal of Mathematical Biology* 29 (1990), pp. 131–159.

[24] M. Farge. “Wavelet transforms and their applications to turbulence”. In: *Annual Review of Fluid Mechanics* 24 (1992), pp. 395–457.

[25] P. Fearnhead. “SequenceLDhot: detecting recombination hotspots”. In: *Bioinformatics* 22.24 (2006), pp. 3061–3066.

[26] P. Fearnhead and N. G. C. Smith. “A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes”. In: *American Journal of Human Genetics* 77 (2005), pp. 781–794.

[27] A.-S. Fiston-Lavier et al. “*Drosophila melanogaster* recombination rate calculator”. In: *Gene* 463 (2010), pp. 18–20.

[28] G. B. Golding. "The sampling distribution of linkage disequilibrium". In: *Genetics* 108 (1984), pp. 257–274.

[29] P. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82 (1995), pp. 711–732.

[30] R. C. Griffiths and S. Tavaré. "Simulating probability distributions in the coalescent". In: *Theor. Popul. Biol.* 46 (1994), pp. 131–159.

[31] A. Grinsted, J. C. Moore, and S. Jevrejeva. "Application of the cross wavelet transform and wavelet coherence to geophysical time series". In: *Nonlinear Processes in Geophysics* 11 (2004), pp. 561–566.

[32] P. R. Haddrill et al. "Multi-locus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations". In: *Genome Research* 15 (2005), pp. 790–799.

[33] G. Hellenthal and M. Stephens. "msHOT: modifying Hudson's ms simulator to incorporate crossover and gene conversion hotspots". In: *Bioinformatics* 23.4 (2007), pp. 520–521.

[34] R. D. Hernandez et al. "Classic selective sweeps were rare in recent human evolution". In: *Science* 331 (2011), pp. 920–924.

[35] A. Hobolth, L. N. Andersen, and T. Mailund. "On computing the coalescence time density in an isolation-with-migration model with few samples". In: *Genetics* 187.4 (2011), pp. 1241–1243.

[36] A. Hobolth and E. A. Stone. "Simulation from endpoint-conditioned, continuous-time Markov chains on a finite state space, with applications to molecular evolution". In: *Annals of Applied Statistics* 3.3 (2009), pp. 1204–1231.

[37] R. R. Hudson. "Generating samples under a Wright-Fisher neutral model of genetic variation". In: *Bioinformatics* 18 (2002), pp. 337–338.

[38] R. R. Hudson. "Two-locus sampling distributions and their application". In: *Genetics* 159 (2001), pp. 1805–1817.

[39] P. A. Jenkins and Y. S. Song. "An asymptotic sampling formula for the coalescent with recombination". In: *Annals of Applied Probability* 20.3 (2010), pp. 1005–1028. ISSN: 1050-5164. DOI: 10.1214/09-AAP646.

[40] P. A. Jenkins and Y. S. Song. "Closed-form two-locus sampling distributions: accuracy and universality". In: *Genetics* 183 (2009), pp. 1087–1103.

[41] P. A. Jenkins and Y. S. Song. "Padé approximants and exact two-locus sampling distributions". In: *Annals of Applied Probability* 22.2 (2012), pp. 576–607.

[42] P. A. Jenkins and Y. S. Song. "The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele". In: *Theor. Popul. Biol.* 80.2 (2011), pp. 158–173.

[43] J. D. Jensen, K. R. Thornton, and P. Andolfatto. “An approximate Bayesian estimator suggests strong, recurrent selective sweeps in Drosophila”. In: *PLoS Genetics* 4.9 (2008), e10000198.

[44] P Johnson and Montgomery Slatkin. “Inference of microbial recombination rates from metagenomic data”. In: *PLoS Genetics* 5.10 (2009), e1000674.

[45] Y. Kim and R. Nielsen. “Linkage disequilibrium as a signature of selective sweeps”. In: *Genetics* 167 (2004), pp. 1513–1524.

[46] J. F. C. Kingman. “On the genealogy of large populations”. In: *Journal of Applied Probability* 19 (1982), pp. 27–43.

[47] A. Kong et al. “Fine-scale recombination rate differences between sexes, populations and individuals”. In: *Nature* 467 (2010), pp. 1099–1103.

[48] R. J. Kulathinal et al. “Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence”. In: *Proc. Nat. Acad. Sci.* 105.29 (2008), pp. 10051–10056.

[49] C. H. Langley et al. “Genomic variation in natural populations of *Drosophila melanogaster*”. In: *Genetics, in press* (2012).

[50] C. H. Langley et al. “Linkage disequilibria and the site frequency spectra in the $su(s)$ and $su(w^a)$ regions of the *Drosophila melanogaster* X chromosome”. In: *Genetics* 156 (2000), pp. 1837–1852.

[51] H. Li and R. Durbin. “Inference of human population history from individual whole-genome sequences”. In: *Nature* 475 (2011), pp. 493–496.

[52] N. Li and M. Stephens. “Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data”. In: *Genetics* 165 (2003), pp. 2213–2233.

[53] J.C. Lucchesi and D.T. Suzuki. “The interchromosomal control of recombination”. In: *Annual Review of Genetics* 2.1 (1968), pp. 53–86.

[54] T. F. C. Mackay et al. “The *Drosophila melanogaster* genetic reference panel”. In: *Nature* 482 (2012), pp. 173–178.

[55] T. Mailund et al. “Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a Coalescent hidden Markov model”. In: *PLoS Genetics* 7.3 (2011).

[56] P. McQuilton et al. “FlyBase 101—the basics of navigating FlyBase”. In: *Nucleic Acids Research* 40.D1 (2012), pp. D706–D714.

[57] G. McVean. “The structure of linkage disequilibrium around a selective sweep”. In: *Genetics* 175.3 (2007), pp. 1395–1406.

[58] G. A. T. McVean and N. J. Cardin. “Approximating the coalescent with recombination”. In: *Philosophical Transactions of the Royal Society B* 360 (2005), pp. 1387–1393.

[59] G. A. T. McVean et al. "The fine-scale structure of recombination rate variation in the human genome". In: *Science* 304 (2004), pp. 581–584.

[60] G. McVean, P. Awadalla, and P. Fearnhead. "A coalescent-based method for detecting and estimating recombination from gene sequences". In: *Genetics* 160 (2002), pp. 1231–1241.

[61] D. E. Miller et al. "A whole-chromosome analysis of meiotic recombination in *Drosophila melanogaster*". In: *G3: Genes—Genomes—Genetics* 2 (2012), pp. 249–260.

[62] S. Myers et al. "A fine-scale map of recombination rates and hotspots across the human genome". In: *Science* 310 (2005), pp. 321–324.

[63] S. Myers et al. "Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination". In: *Science* 327.5967 (2010), pp. 876–879.

[64] D. Ortiz-Barrientos, A. S. Chang, and M. A. F. Noor. "A recombinational portrait of the *Drosophila pseudoobscura* genome". In: *Genetics Research Cambridge* 87 (2006), pp. 23–31.

[65] J. E. Pool and R. Nielsen. "Population size changes reshape genomic patterns of diversity". In: *Evolution* 61.12 (2007), pp. 3001–3006.

[66] P. Portin and M. Rantanen. "Further studies on the interchromosomal effect on crossing over in *Drosophila melanogaster* affecting the preconditions of exchange". In: *Genetica* 82.3 (1990), pp. 203–207.

[67] A. L. Price et al. "Sensitive detection of chromosomal segments of distinct ancestry in admixed populations". In: *PLoS Genetics* 5.6 (2009), e1000519.

[68] F. A. Reed and S. A. Tishkoff. "Positive selection can create false hotspots of recombination". In: *Genetics* 172 (2006), pp. 2011–2014.

[69] S. Sattath et al. "Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*". In: *PLoS Genetics* 7.2 (2011), e1001302.

[70] G. Sella et al. "Pervasive natural selection in the *Drosophila* genome?" In: *PLoS Genetics* 5.6 (2009), e10000495.

[71] N. D. Singh, C. F. Aquadro, and A. G. Clark. "Estimation of fine-scale recombination intensity variation in the *white-echinus* interval of *D. melanogaster*". In: *Journal of Molecular Evolution* 69 (2009), pp. 42–53.

[72] C. C. A. Spencer et al. "The influence of recombination on human genetic diversity". In: *PLoS Genetics* 2.9 (2006), e148.

[73] W. Stephan, Y. S. Song, and C. H. Langley. "The hitchhiking effect on linkage disequilibrium between linked neutral loci". In: *Genetics* 172.4 (2006), pp. 2647–2663.

[74] L. S. Stevison and M. A. F. Noor. "Genetic and evolutionary correlates of fine-scale recombination rate variation in *Drosophila persimilis*". In: *Journal of Molecular Evolution* 71 (2010), pp. 332–345.

[75] K. M. Teshima and H. Innan. "mbs: modifying Hudson's ms software to generate samples of DNA sequences with a biallelic site under selection". In: *BMC Bioinformatics* 10 (2009), p. 166.

[76] The 1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing". In: *Nature* 467 (2010), pp. 1061–1073.

[77] The International HapMap Consortium. "A second generation human haplotype map of over 3.1 million SNPs". In: *Nature* 449.7164 (2007), pp. 851–861.

[78] K. Thornton and P. Andolfatto. "Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*". In: *Genetics* 172 (2006), pp. 1607–1619.

[79] C. Torrence and G. P. Compo. "A practical guide to wavelet analysis". In: *Bulletin of the American Meteorological Society* 79 (1998), pp. 61–78.

[80] I. J. Tsai, A. Burt, and V. Koufopanou. "Conservation of recombination hotspots in yeast". In: *Proc. Nat. Acad. Sci.* 107.17 (2010), pp. 7847–7852.

[81] J. D. Wall, P. Andolfatto, and M. Przeworski. "Testing models of selection and demography in *Drosophila simulans*". In: *Genetics* 162 (2002), pp. 203–216.

[82] Y. Wang and B. Rannala. "Bayesian inference of fine-scale recombination rates using population genomic data". In: *Philosophical Transactions of the Royal Society B* 363.1512 (2008), pp. 3921–3930.

[83] C. Wiuf and P. Donnelly. "Conditional genealogies and the age of a neutral mutant". In: *Theor. Popul. Biol.* 56 (1999), pp. 183–201.