Statistica Sinica **21** (2011), 43-69

ON COMPOSITE LIKELIHOODS IN STATISTICAL GENETICS

F. Larribe and P. Fearnhead

UQAM and Lancaster University

Abstract: Due to the dimension and the dependency structure of genetic data, composite likelihood methods have found their natural place in the statistical methodology involving such data. After a brief description of the type of data one encounters in population genetic studies, we introduce the questions of interest concerning the main genetic parameters in population genetics, and present an up-to-date review on how composite likelihoods have been used to estimate these parameters.

Key words and phrases: Association, coalescent process, composite likelihood, fine genetic mapping, linkage disequilibrium.

1. Introduction

Composite likelihoods have been extremely influential in population genetics. A prime example of this is that such methods have led to the first ever fine-scale recombination map of the human genome (McVean et al. (2004), Myers et al. (2005)), which is useful for understanding the biology and evolution of recombination, and also for interpreting the results of genome-wide association studies. As ever-increasingly large and complicated data sets are collected, the use and importance of composite likelihood methods for analysing such data will also increase.

Composite likelihoods are based upon calculating likelihoods for a subset of the data, and then combining these likelihoods as if each subset of the data were independent. Parameter estimates are constructed by maximising the resulting composite likelihoods. Examples include taking the product of the marginal probability of each data point, or basing inference from likelihood for all pairs of data points (Cox and Reid (2004)). See Lindsay (1988), Varin and Vidoni (2005), Varin (2008), Varin, Reid, and Firth (2011) and references therein for further details.

In general there are two main motivations for using composite likelihood approaches. The first is computational, as calculating likelihoods for subsets of data is often substantially easier than calculating the full-likelihood for the complete

data set. The second is that it avoids the need to model higher order dependencies in the data, and thus gives inferences that are based only on modelling of appropriate marginal or low-dimensional aspects of the data.

Within genetics it is the first of these motives that has led to the widespread use of composite likelihoods. This article reviews some of the important applications of composite likelihoods within population genetics. We aim to give an overview of the extent to which composite likelihoods are used, as well as the different approaches used for constructing composite likelihoods. We also review the few existing theoretical results there are specifically for the application of these methods in genetics.

The article is organized as follows. In Section 2 we introduce the basic notions in genetics that are important for understanding the problems presented in this paper. The reader familiar with genetics can skip this section. Further, in Section 3, we describe typical genetic data, explain the types of inference problems that are of interest, and give a non-technical introduction to the first statistical methods that were used to analyze this type of data. In the next sections we explore how composite likelihoods are used in estimating the recombination rate (Section 4) and in genetic mapping (Section 5). Other uses of composite likelihoods for answering various problems in genetics are considered in Section 6.

2. Genetics: Basic Terminology

Although most of the methods presented here can be applied to non-human genetic data, we will, for simplicity, focus on human genetics. Each human has 46 chromosomes consisting of 23 pairs of chromosomes. Chromosomes are made up of sequences of nucleotides, the DNA bases. Chromosomes range in size from 250 million bases (250 Mb) for chromosome 1, to 50 Mb for chromosome 21. The total length of the genome is approximately 3,000 Mb. For each pair of chromosomes, one was inherited from the mother, and the other one from the father. Only part of this DNA is functional: the genes. Each gene, therefore, exists then in double: humans have one "variant" of a certain gene on one chromosome, and another "variant" (possibly the same) on the other chromosome. These "variants" are called *alleles*. Because humans have two versions of each gene, they are called *diploïd*.

Several biological events change the chromosomes over time when they are transmitted from one generation to another. The events that we consider here, which will have considerable importance in the rest of the paper, are *mutation* and *recombination*. A mutation changes the genetic code at one *locus* (a locus is a precise position on a chromosome), this can be a single nucleotide change or a more complex event. Further, when chromosomes are transmitted from parent

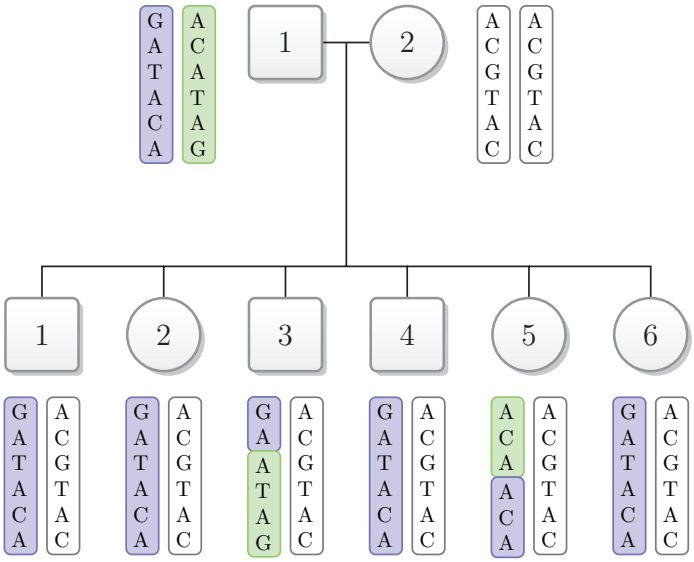

Figure 1. Illustration of the transmission of genetic material in a family unit. Square indicates male, and circle female. One can see that child 3 and 5 are recombinants.

to offspring, the offspring does not receive a copy of just one of that parent's pair of chromosomes, but a mosaic of that parent's pair of chromosomes. The process that creates this mosaic is recombination. For example, consider the child numbered 3 in Figure 1; his "left" chromosome is a mix of the two chromosomes his father has inherited. We say that a *recombination* has occurred between sites 2 and 3 (in child 3), because, on the chromosome inherited from his father, the DNA at site 2 does not come from the same parental chromosome as the DNA at site 3.

The *recombination fraction* between two loci is defined as the probability that a recombination occurs between the two loci. The expected number of recombination events between two loci defines the *genetic distance* between these two loci, expressed in Morgans. This is in contrast with the physical distance between the two loci, which is the number of base pairs (bp) separating them. There are no direct ways to translate genetic distances into physical distances because recombination rates vary over the genome (McVean et al. (2004)). However, we expect the recombination fraction between two loci to be very small if these two loci are close, and the recombination fraction to be 1/2 if these two loci are far apart (if the two loci are completely unlinked, due to independent assortment, the probability of a recombination is 1/2).

3. The Nature of Genetic Data and Statistical Tools

First, we describe how the data are collected. Here we focus on genetic data from one of two sources: individuals in pedigrees, and in population samples. Each of these sources induces some form of dependency in the data: in a pedigree, all individuals are directly linked; in a population, the sampled individuals are also related, albeit more distantly, but the nature of their relatedness is unknown. For both sources it is possible to collect phenotypic and genetic information for each individual under study. The phenotypic information usually consists of a trait under study (the disease status, or another measure of health, for example); genetic information is extracted from biological material (blood sample, for example). The available genetic information has changed over time, particularly with respect to the amount of data that is available. Today, the information can consist of a complete DNA sequence on a short scale (i.e. a chromosomal segment), data from thousands of sites across a larger region of a chromosome, or even data from hundreds of thousands of sites across the whole genome.

A *genetic marker* is a DNA sequence with a known location on the genome; the genetic marker is *segregating* if there exist variations in the sample at that site. One can imagine a set of genetic markers for one individual to be a sample of the genetic code along a chromosome at specific locus (see Figure 2). The markers most commonly used in genetics exhibit binary variation, and are called single-nucleotide polymorphisms (SNPs).

If we have genetic information from an individual at a set of SNPs, we know the two alleles that person has for each SNP – that is we know the DNA bases at each of the SNP locations that are present on that person's two chromosomes. However we commonly do not know on which chromosome each base is. For example, if we take individual 1 in Figure 2, we would not know that the two sequences are `TCTC` and `TTAG`, only that the alleles at the first marker are `T/T`, the alleles at the second marker are `T/C`, etc; this is called *genotype* data. Data that consisted of the sequences on each of the two chromosomes is called *haplotype* data, with each sequence being a haplotype. It is possible, albeit costly, to get *haplotype* data experimentally. Alternatively we can estimate the haplotypes from genotype data (see e.g., Stephens, Smith, and Donnelly (2001)).

The basic example of a *pedigree* is a collection of members from the same family. Conditional upon on the pedigree (and other factors, like allele frequencies in the population), a likelihood can be defined and one can estimate recombination rates. The form of this likelihood can be complex. Markov Chain Monte Carlo (MCMC) methods are often used in this context (Thompson (2000)). When we estimate the recombination fraction between some genetic markers and a disease

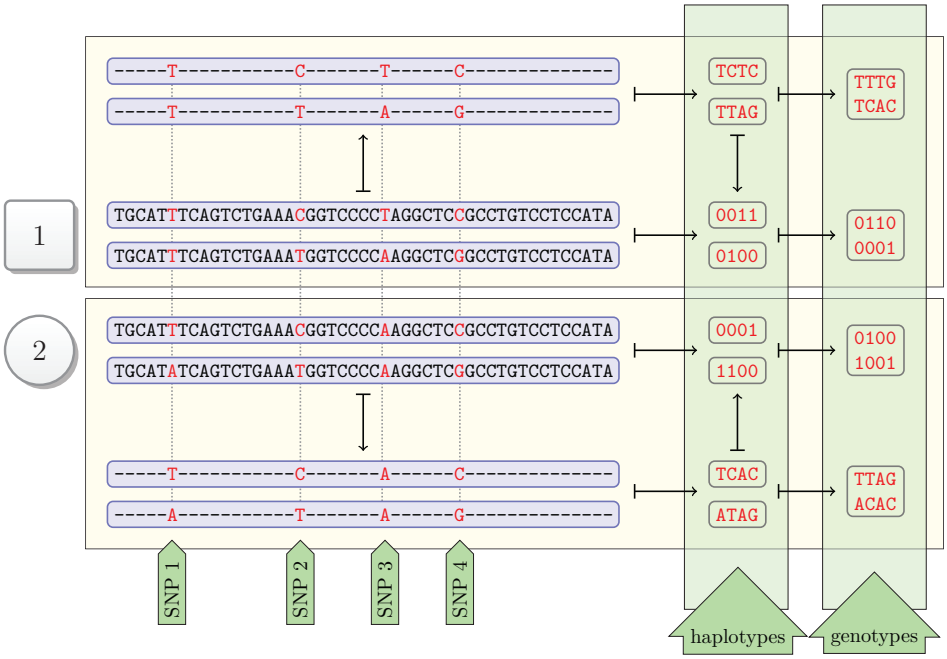

Figure 2. Illustration of a (very short) piece of chromosome (47 base pairs) for individuals 1 and 2; one can see four genetic markers (SNP1 to SNP4). Data would consist of either the haplotype or genotype for each person. At each SNP we observe just two of the four bases present. For most analyses the specific base (A, C, G or T) does not matter, and the data are summarised in binary for each SNP on each chromosome.

gene in a pedigree, we estimate the location of the disease gene on the chromosome. Such analyses, called "linkage analyses", have been successful in mapping disease genes but, because of the relatively small number of recombinations in a pedigree, the precision of the estimate is around 1cM, which corresponds to a million base pairs, and is much too coarse for physical mapping.

3.1. Population data

Population data consist of data from a sample of unrelated individuals. The size of such a sample ranges usually from a hundred to a thousand individuals. Whilst unrelated, the genetic data from these individuals will still be dependent. If we consider a single locus and trace the ancestry of the chromosomes, as we go back in time different pairs or sets of chromosome will share a common ancestor. The information about these common ancestors can be compactly described through a *genealogy*. This genealogy determines the dependence in the data at that locus (see Figure 3 (a) for an example of such a genealogy).

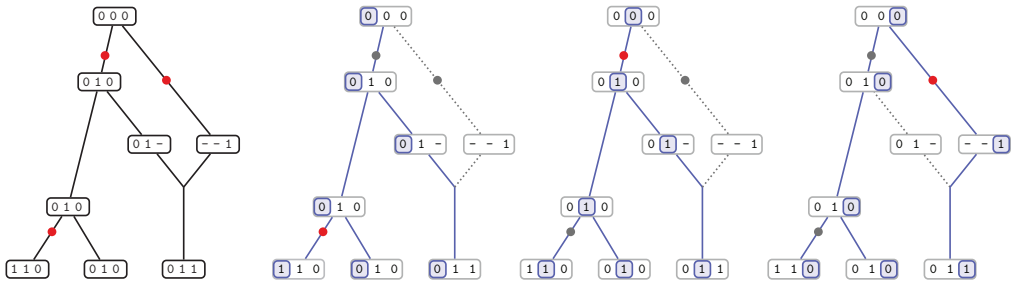

Figure 3. Example of a genealogy of three sequences of three genetic markers. Dots along the edges correspond to mutation events; moving back in time (from bottom to top), we have a coalescent event when two branches merge into one (two sequences find a common ancestor), and a recombination event when a branch split in two: we have here only one recombination event, between sites two and three. Solid lines (b, c, d) denote a genealogy at a particular site. (a) genealogy for the three sequences, (b) genealogy at site 1, (c) genealogy of site 2, (d) genealogy of site 3.

If there were no recombination, then we would have the same genealogy at all loci in our sample (like genealogies (b) and (c), for site 1 and 2, in Figure 3). If our data consisted of unlinked loci then we would be able to treat the data (and genealogies) as independent across loci. However for most data we have a small recombination fraction between loci. This means the genealogies at different loci are different, but dependent, and there are strong and complex dependencies in the data both across chromosomes and across loci.

This dependence, both across loci and across individuals, makes inference from population data challenging. The most common approach to inference is to introduce an appropriate stochastic model for the genealogy (or genealogies) of the sample. Given the genealogical information, calculating likelihoods is normally straightforward. Thus we have a classical missing-data framework, where the genealogical information is the missing data, and calculation of, say, the likelihood function requires averaging over all possible genealogies for the data.

The most widely used stochastic model for the genealogy is the *coalescent process* (Kingman (1982a,b)). This process can be derived from the Wright-Fisher model for the forward evolution of the genetic diversity within a population. In its simplest form it provides a model for the genealogy at a single locus under certain simplifying assumptions for the evolution of the underlying population such as no selection, random-mating and a constant population size. For more details see, for example, Nordborg (2007), Hein, Schierup, and Wiuf (2005),

Nordborg and Tavaré (2002), Tavaré, S. and Zeitouni, O. (2004), and Wakeley (2008).

The power and popularity of the coalescent derives first from a robustness result, showing that the coalescent process is the correct stochastic process for the genealogy of a sample for a wide-range of models for the forward-evolution of the population. Second, the coalescent can be easily extended to various *demographic models*, for example, allowing varying population sizes and certain forms of non-random mating. It can also be extended to model the joint distribution of genealogies at different loci in the presence of recombination. Lastly, in most cases it is easy and computationally efficient to simulate the coalescent, and hence simulate population genetic data under a range of different modelling assumptions.

The parameters of the coalescent are population-scaled rates. For example, rather than depending on the expected number of recombination events per meiosis between two loci, r say, the coalescent will depend on a rate $\rho = 4Nr$, where N is called the effective sample size. Similarly, the mutation model will be parameterised by a mutation rate $\theta = 4Nu$, where u is the mutation probability per meiosis.

3.2. Inference for population data

The estimation of the "full likelihood" (using all sequences and all sites in the sample) has stimulated much research over the past two decades; see Stephens (2007) for a comprehensive review. The key to these inference methods is how to efficiently average over all genealogies consistent with the data. In practice Monte Carlo methods are used to do this (at least approximately), but even here implementation is difficult due to the large set of possible genealogies at each locus.

The first full-likelihood method was an importance sampling method due to Griffiths and Tavaré (1994a,b,c). Stephens and Donnelly (2000) proposed improvements to the method by suggesting an approximation to an optimal proposal density for importance sampling. The method of Griffiths and Tavaré (1994a,b,c) has been extended to the coalescent with recombination (Griffiths and Marjoram (1996, 1997), Fearnhead and Donnelly (2001)). An alternative approach is based upon Markov chain Monte Carlo (Kuhner, Yamato, and Felsenstein (1995, 1998, 2000); Wilson and Balding (1998)).

These methods can be computationally efficient for analysing data at a single locus, and can analyse small sample sizes (up to 100 chromosomes) at a small number of bi-allelic loci (of the order of 5 to 10). However, they do not scale to analysing the size of data currently being generated: data from 100s of chromosomes at maybe 1000s of loci. It is this that has motivated the use of composite

likelihood methods for analysing population genetic data. Indeed, composite likelihood approaches enable us to split a large data set into smaller pieces, for each of which a likelihood function can be calculated.

3.3. Notation and results on dependence

In our discussion of composite likelihood methods for population genetic data we focus on methods for analysing haplotype data. This is both for simplicity, and because many composite likelihood methods require such data. In practice using such methods is not problematic, as there are efficient procedures that can accurately infer haplotypes from genotype data (Stephens, Smith, and Donnelly (2001)).

We will focus on data from N chromosomes, corresponding to $N/2$ individuals. Each chromosome will be typed at L loci. For concreteness we will assume each locus is a SNP, and thus the genetic type of a chromosome at each locus will be one of two possibilities. Mathematically we can represent this data by a binary matrix D of dimension $L \times N$, with D_{in} denoting the genetic type (allele) of the nth chromosome at the ith SNP. For a given SNP, i say, we know that for any n and m that $D_{in} = D_{im}$ means that the nth and mth chromosome share the same genetic type at this SNP, while $D_{in} \neq D_{im}$ means that they have different genetic types at that SNP. The actual value of D_{in} is arbitrary (that is if we switched all the 0s and 1s within any row of D then we have an identical data set). Most methods we describe allow for missing data – though for simplicity we will ignore this possibility.

We will use D_i to denote that data just at the ith SNP, and $D_{(i,j)}$ to denote the data just at the ith and jth SNP and so on. For consecutive SNPs we will use $D_{(i:j)}$ to denote data from SNPs $i, i+1, \ldots, j$. To help distinguish $D_{(i,j)}$ from D_{in} we use the convention that chromosomes are denoted by n or m, and SNPs by i, j or k.

We will denote the physical location of SNP i by x_i, and assume that $x_1 < x_2 < \cdots < x_L$. We introduce a recombination rate function $\rho(x)$ defined so that the recombination rate between SNPs i and j, with $i < j$ is $\rho(x_j) - \rho(x_i)$. Note that $\rho(x)$ is an increasing function. If the recombination rate per bp is assumed constant then $\rho(x_j) - \rho(x_i) = \bar{\rho}[x_j - x_i]$, for an appropriate constant $\bar{\rho}$.

A key result which affects the implementation of composite likelihood methods for population genetic data is the following, which demonstrates long-range dependence across loci.

Theorem 1. (Fearnhead (2003)) *Consider the coalescent with recombination for a sample of N chromosomes. Let $\pi_1(\cdot)$ denote the marginal probability mass function of data at a single SNP, and $\pi_2(\cdot, \cdot; \lambda)$ the joint probability of data at*

two SNPs separated by a recombination rate λ. Then there exists a K such that for any D_i and D_j,

$$|\pi_1(D_i)\pi_1(D_j) - \pi_2(D_{(i,j)};\lambda)| < \frac{K}{\lambda}.$$

Thus if we have a constant recombination rate, the dependence between two SNPs decays like the inverse of the distance between the SNPs. The result in Fearnhead (2003) is for the simplest coalescent with recombination, but this result holds generally (Wiuf (2006)).

4. Estimation of the Recombination Rate

We first focus on composite likelihood methods for estimating recombination rates. There are two basic approaches that have been used: one based on analysing data from pairs of SNPs, and the other based on analysing data from contiguous regions of 5 to 10 SNPs. First we will introduce both these methods for estimating the recombination rate under a model where the rate between two SNPs is proportional to the physical distance between the SNPs. This is termed a constant recombination rate model, as the rate per bp is constant. We will then discuss empirical and theoretical results for these methods. Finally we will briefly discuss extensions of these methods, particularly to estimating more general recombination rate models.

4.1. Composite likelihood methods for constant recombination rate models

The first composite likelihood method for estimating recombination rates was introduced by Hudson (2001), see also McVean, Awadalla, and Fearnhead (2002). This is based on combining the likelihood for all pairs of SNPs in the data. We will describe the idea under a constant recombination rate model, which is parameterised by a recombination rate per bp, $\bar{\rho}$, and a set of nuisance parameters, ϕ. The latter can include parameters governing the mutation model, such as mutation rate, and the demographic model, such as details of how the population size has varied over time, or parameters for models which allow non-random mating. We describe inference conditional on an estimated (or assumed) value of ϕ. For simplicity we drop this conditioning from our notation.

As above let $\pi_2(D_{(i,j)};\bar{\rho}(x_j - x_i))$ be the probability of the data at SNPs i and j, given the recombination rate between them is $\bar{\rho}(x_j - x_i)$. Then the composite likelihood of Hudson (2001), which we will call a *pairwise likelihood* (Cox and Reid (2004)) is

$$\mathrm{PwL}(\bar{\rho}) = \prod_{i=1}^{L}\prod_{j=i+1}^{L} \pi_2(D_{(i,j)};\bar{\rho}(x_j - x_i)).$$

The key to the computational efficiency of this model is that for data from N chromosomes there are of the order of just N^3 possible values for the data $D_{(i,j)}$. Thus the individual likelihood functions $\pi_2(D_{(i,j)};\lambda)$ can be tabulated for all possible values of $D_{(i,j)}$ and for a suitable grid of λ, for sample sizes up to $N \approx 100$. Once stored, evaluating the pairwise likelihood is computationally cheap unless L is extremely large.

The likelihoods for each pair of SNPs can be calculated in a number of ways. The simplest is by simulating from the appropriate coalescent model and storing the observed frequencies of each value for $D_{(i,j)}$. Hudson (2001) uses an improved version of this idea, where genealogies at each SNP are simulated, the probability of $D_{(i,j)}$ given each pair of genealogies is calculated and then averaged over a large sample of pairs. This approach can be used for any demographic model (see also Carvajal-Rodriguez, Crandall, and Posada (2006)), but is currently restricted to mutation models that do not allow more than one mutation on the genealogy at each SNP. An alternative approach was taken by McVean, Awadalla, and Fearnhead (2002), who use importance sampling to calculate the likelihood function under a more general mutation model, but only allow the simplest demographic model for the coalescent.

The second composite likelihood method is based on splitting the data in subregions of consecutive SNPs (Fearnhead and Donnelly (2002)). Define the size of each subregion as containing k SNPs. Further define the joint probability mass function of data at k SNPs as $\pi_k(D_{(i:i+k-1)}|\rho_{i:i+k-1})$, where $\rho_{i:i+k-1} = \bar{\rho}(x_{i+1}-x_i,\ldots,x_{i+k-1}-x_{i+k-2})$ is the vector of the $k-1$ recombination rates between each pair of consecutive SNPs. Then the composite likelihood is defined as

$$L_C(\bar{\rho}) = \prod_{i=1}^{L-k+1} \pi_k(D_{(i:i+k-1)}|\rho_{i:i+k-1}).$$

Calculating this composite likelihood is computationally more expensive than calculating the pairwise likelihood, as it is no longer possible to tabulate the likelihoods for each possible value of $D_{(i:i+k-1)}$. Instead the terms $\pi_k(D_{(i:i+k-1)}|\rho_{i:i+k-1})$ are calculated using the importance sampling method of Fearnhead and Donnelly (2002). Obviously by setting $k = L$ we recover the full-likelihood for the data, and in general the larger the value of k the more accurate will the resulting inference be. However, the larger k the harder it is to accurately estimate $\pi_k(D_{(i:i+k-1)}|\rho_{i:i+k-1})$. In practice, Fearnhead and Donnelly (2002) suggest taking k between 5 and 10 as giving a reasonable compromise between statistical and computational efficiency.

Currently this composite likelihood method can only be calculated under the simplest demographic model for the coalescent, however it does allow for

mutation models with recurrent mutation, and allows for joint inference of the mutation and recombination parameters.

4.1.1. Example

To demonstrate the application of composite likelihood methods, we considered analysing SNP data from the HABP2 gene from the SeattleSNPs database (Crawford et al. (2004)). The original data consist of genotypes from 23 Europeans and 24 African Americans. We focus on analysing the African American data, and first use PHASE (Stephens, Smith, and Donnelly (2001)) to infer haplotypes.

The resulting data set consists of 188 SNPs across an approximately 40kb region. A graphical deciption of the data is shown in Figure 4 (left-hand plot), where we plot the correlation in genetic types for each pair of SNPs. High correlation between a pair of SNPs reflects the idea that the frequency of genetic type at one SNP depends strongly on the genetic type at the other SNP on that haplotype. Low correlation suggests that there is little dependence between the genetic types on a haplotype at the two SNPs. This is a common plot for data as the more the recombination between two SNPs then the smaller the amount of correlation that would be expected to be observed. The idea of the pairwise likelihood is that for each pair of SNPs a log-likelihood curve is calculated, and these are summed to provide a composite log-likelihood curve from which the recombination rate is estimated. This can be viewed as a theoretically justified way of combining the information in the amount of correlation between each pair of SNPs, which is shown in Figure 4, to provide an estimate of the recombination rate.

For simplicity we focus on analysing data from a 10kb region, starting at position 10kb (this region corresponds to the highlighted square in Figure 4.) This region itself has 79 SNPs, which is substantially more than could be analysed via full-likelihood methods. We first implemented the composite likelihood method of Fearnhead and Donnelly (2002), using 72 overlapping subregions each of which contained 8 consecutive SNPs. Results are shown in Figure 4 (right-hand plot), where we show a random sample of log-likelihood curves for the sub-regions, together with the final composite log-likelihood curve. The final estimate of the recombination rate is 4.0 per kb. However, there is substantial variation in the log-likelihood curves, and the position of their maximum, across the individual sub-regions. Possible explanations include it being a feature of the sampling variability in the log-likelihood curves, but also that the assumption of a constant recombination-rate across this 10kb region may not be appropriate. We return to the latter point in Section 4.3.

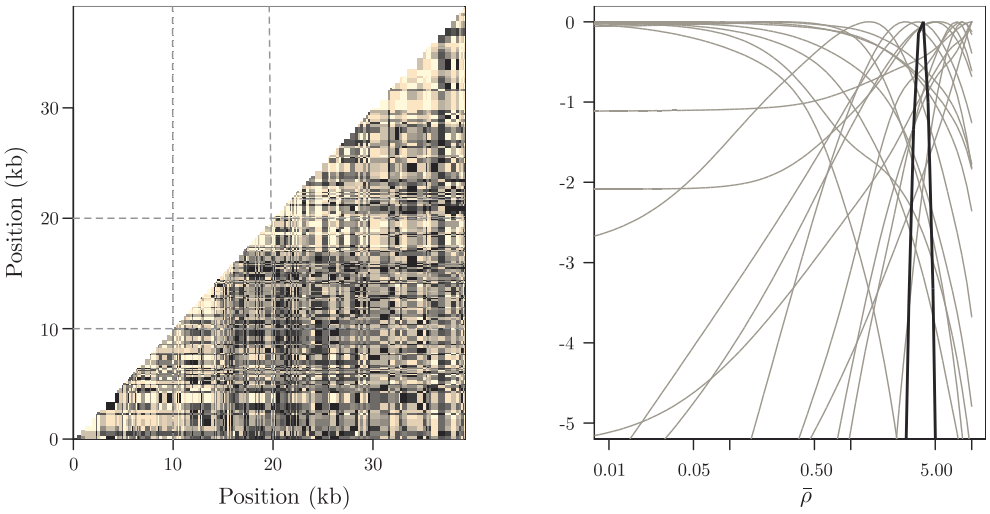

Figure 4. SNP Data from 24 African Americans for the HABP2 gene. Left-hand plot shows the pairwise correlation between each pair of SNPs: light shading shows high correlation, darker shading shows lower correlation. Right-hand plot shows results from analysing data from positions 10kb-20kb using the composite likelihood method of Fearnhead and Donnelly (2002). In grey are a sample of 20 log-likelihood curves from sub-regions; in bold is the composite log-likelihood curve obtained by summing log-likelihood curves from all 72 sub-regions. All log-likelihood curves have been shifted to have a maximum value of 0.

We also used the pairwise likelihood method to estimate the recombination rate, using the method of McVean, Awadalla, and Fearnhead (2002). This method gave an estimate of the recombination rate of 2.5 per kb, which is noticeably different from that of the first composite likelihood method. The reason for a difference is that the two methods use different information. The pairwise likelihood methods use information from quite distant SNPs. The method of Fearnhead and Donnelly (2002) uses information only from nearby SNPs, but uses the additional information that is contained in the haplotype patterns of multiple SNPs. Note that it could be possible to combine the information from the two composite likelihood approaches, again using the idea of composite likelihood to sum the composite likelihood curves from the two methods. This has been considered by Smith and Fearnhead (2005) who showed that such an approach can improve the accuracy of estimates, but that the improvement is quite small.

4.2. Empirical and theoretical results

One empirical study of the accuracy of composite likelihood methods for

estimating recombination rates is found in Smith and Fearnhead (2005). The main conclusion from that was that the two composite methods described here were roughly as accurate as each other: the pairwise likelihood method does better for regions were the recombination rate is small, and the method of Fearnhead and Donnelly (2002) does better when the recombination rate is larger. The explanation for this is that the information from pairs of SNPs decays as the recombination rate between them increases. As such the pairwise likelihood method benefits more from using information from SNPs that are a large physical distance apart when the recombination rate is lower. Overall, the main advantage of the pairwise method is its speed, which can be orders of magnitude quicker than the alternative composite likelihood approach.

The only theoretical results about the above two composite likelihood methods concern the consistency of estimates for $\bar{\rho}$. These results come from Fearnhead (2003). A necessary condition for consistent estimators is that the marginal likelihoods used within the composite likelihood model are correctly specified. Essentially this means that the expected log-likelihood curve, with expectation over repeated draws of the data, from a sub-region or pair of SNPs will have its maximum at the true parameter value. If we analyse sequence data using the pairwise method, then it is common just to use pairs of sites that segregate within the sample. Theorem 4 of Fearnhead (2003) shows that if we do not account for the conditioning on segregation in defining our log-likelihood then any resulting pairwise likelihood estimator will not be consistent.

For a similar reason, satisfying this condition is non-trivial for SNP data due to needing to account for the ascertainment procedure used in choosing the SNP locations (see e.g., Nielsen and Signorovitch (2003)). More importantly, if the demographic model that is used is incorrect, then again we will not get consistent estimators of the recombination rate. This is investigated empirically by Fearnhead and Smith (2005), who show that the resulting biases, for large samples, in the composite likelihood methods are generally small across a range of demographic scenarios. Furthermore, these are particularly small when trying to estimate how recombination rates vary between different genomic regions.

In the following we now ignore this issue and assume that these likelihoods are correctly specified, and for the pairwise likelihood that we know, or have consistent estimators of, the nuisance parameters (mutation rates). Under these assumptions we give a brief overview of existing theoretical results.

The asymptotic regime considered is as the number of SNP loci tends to infinity. The information in the data is bounded as the number of chromosomes $N \to \infty$, and we do not obtain consistent estimators as $N \to \infty$. Fearnhead (2003) shows that the composite likelihood estimator of the recombination rate

of Fearnhead and Donnelly (2002) is consistent as $L \to \infty$. This follows almost directly from the result of Theorem 1, which is sufficient to show that

$$\mathrm{Var}\left(\frac{1}{L-k+1}L_C(\bar{\rho})\right) \to 0,$$

as $L \to \infty$. As a result we have that $L_C(\bar{\rho})/(L-k+1)$ converges in probability to the expected log-likelihood for a sub-region, and hence (under further mild conditions) that the maximum value of $L_C(\bar{\rho})$ is attained for the value of $\bar{\rho}$ for which the expected log-likelihood has its maximum. Under the assumptions above this is the true parameter value. However, the rate of decay of this variance is $\log(L)/L$, which suggests that the variance of the composite likelihood is also of order $\log(L)/L$, as compared to the more normal $1/L$ rate of convergence. The extra $\log(L)$ term is due to the slow rate of decay of dependence in population genetic data.

Fearnhead (2003) was unable to show that the pairwise likelihood estimator of $\bar{\rho}$ was consistent. This is due to a combination of the slow rate of decay in dependence plus also that the information contained in $D_{(i,j)}$ about $\bar{\rho}$ also decays as $x_j - x_i$ gets large. However, estimators based on a weighted pairwise likelihood,

$$\mathrm{WPwL}(\bar{\rho}) = \prod_{i=1}^{L}\prod_{j=i+1}^{L} \pi_2(D_{(i,j)};\bar{\rho}(x_j-x_i))^{w(x_j-x_i)},$$

are consistent for appropriate weight functions $w(x)$. In particular we get consistent estimators if we choose $w(x)$ such that

$$w(x) = \begin{cases} 1 & \text{if } x < C, \\ 0 & \text{if } x \geq C, \end{cases}$$

for some C. Empirical results in Smith and Fearnhead (2005) show that this choice of weight function leads to more accurate estimates of the recombination rate for values of C of the order of 10,000bps. Note that this choice leads to pairwise likelihood methods that scale well to large L as their complexity is linear, rather than quadratic, in L; this is what is used in McVean et al. (2004).

Important open theoretical questions include results on the asymptotic distribution for estimators of $\bar{\rho}$, and also deriving consistent estimators for the variance of the composite likelihood based estimates of $\bar{\rho}$. The former is non-trivial due to the dependence structure across loci within population genetic models; the latter is difficult because of the dependence that exists both across loci and across chromosomes.

4.3. Estimating varying recombination rates

An important extension of these composite likelihood methods has been the inference about more general recombination models. Of particular importance, as motivated by the data in Section 4.1.1, is to infer whether, and how, recombination rates vary across a region of interest. Pioneering work in McVean et al. (2004) shows how the pairwise likelihood can be used to fit a model where the recombination rate function $\rho(x)$ is piecewise linear, thus allowing the recombination rate per bp to vary across the chromosome.

The pairwise likelihood can be extended to a pairwise likelihood for a general recombination rate function, $\rho(x)$, in an obvious way. The inference mechanism is to introduce a prior on $\rho(x)$, defined by a distribution for the number and position of changepoints where the recombination rate per bp is allowed to change, and a distribution for this rate per bp between successive changepoints. If we denote this prior by $\pi(\rho(x))$, then McVean et al. (2004) base inference on a (pseudo) posterior which is proportional to $\pi(\rho(x))\mathrm{PwL}(\rho(x))$. Samples from this posterior were generated by Markov chain Monte Carlo.

The use of the pairwise likelihood to define this posterior cannot really be justified in a formal sense. In particular, the resulting posterior substantially under-estimates the uncertainty in the $\rho(x)$. In practice this approach is closely related to basing inference on a penalised likelihood, with the prior being the mechanism that penalises how much $\rho(x)$ varies (through a probabilistic penalty on the number of changepoints allowed). MCMC then just gives an efficient procedure for finding values of $\rho(x)$ that have a high penalised likelihood. However, even if viewed as a penalised likelihood approach, it is unclear how to choose the penalty function (prior), and what the resulting theoretical properties of the estimate of $\rho(x)$ are.

To demonstrate this approach, we return to the HABP2 data set of Section 4.1.1. Here we analyse data from the whole gene, using the method of McVean et al. (2004) We implemented their procedure a number of times with different degrees of penalty in the definition of the prior. For each choice of prior we ran the MCMC algorithm, using the software LDhat, for 10 million iterations. Each run took a matter of minutes on a desktop PC. The results are displayed in Figure 5.

We notice the effect that the size of penalty within the prior has on the estimates, with the estimates of the recombination rate varying substantially more for the smallest penalty. The differences are practically important. For example, a region of high recombination rate (often called a recombination hotspot, see below), at around 25kb is only detected with two of the three choices of penalty. Also the different penalties give different inferences for the region of high recombination close to 15kb, with alternative explanations of either one or two distinct

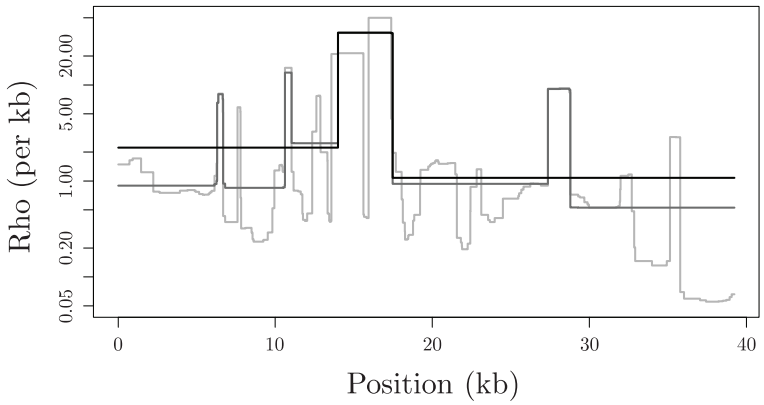

Figure 5. Results of analysis of variable recombination within the HABP2 gene using pairwise likelihood. We present three point estimates of how the recombination rates vary across the gene, each for different degrees of penalty. Numerical values were 2 (lightest grey, smallest penalty), 20, and 100 (black, largest penalty).

regions of high recombination. Currently, the choice of penalty is based purely on empirical evidence – with suggestions of values close to the middle of the three choices we used for Figure 5.

This approach to estimating how the recombination rate varies across the genome has proved immensely influential. McVean et al. (2004) was one of the first papers to quantify the substantial rate variation within the human genome, and Myers et al. (2005) produced the first fine-scale recombination map of the human genome. Related composite likelihood approaches have attempted to detect regions of the genome with substantially higher recombination rates than expected, so called *recombination hotspots* (Fearnhead and Smith (2005), Myers et al. (2005), Fearnhead (2006), Auton and McVean (2007)). These have led to the first human recombination hotspot map (Myers et al. (2005)), and understanding about the cause and evolution of recombination hotspots (see e.g., Winckler et al. (2005), Myers et al. (2008)).

4.4. Software

Software is freely available for implementing the composite likelihood methods. The pairwise likelihood approach can be implemented using code from Dick Hudson, available from `http://home.uchicago.edu/~rhudson1/`, or within the software `LDhat`, available from `http://www.stats.ox.ac.uk/~mcvean/LDhat/`. The latter is able to perform inference for variable recombination rates, and is arguably the most popular software for estimating recombination rates from population data.

Software implementing the composite likelihood method of Fearnhead and Donnelly, including methods for detecting hotspots, is available from `http://www.maths.lancs.ac.uk/~fearnhea/Software.html`.

We used the program `sequenceLDhat`, available from the latter website, and the `LDhat` software to produce the results for the HABP2 gene.

5. Association and Fine Genetic Mapping

In medical genetics, one of the main interests is to map a gene, i.e. to locate the position of a gene causing a disease, or any phenotype of interest, on a chromosome. When this is done, in subsequent research, molecular biologists can examine the code of the gene in the hope of preventing the disease. Linkage analysis in families has been the traditional way of mapping genes. However, because of the relatively limited number of recombination events, it is difficult to precisely locate the position of the gene that one is looking for; this is why population data is often used instead. Usually, the result of a linkage analysis is to identify one or several chromosomal regions of interest likely to contain the "disease gene".

Mapping of disease-genes works because of what is called *linkage disequilibrium* (LD). This concept refers to the non-random association of alleles at nearby loci. To understand the concept of LD, one has to imagine that, at some time in the past, a mutation causing the trait of interest appeared on a particular chromosome of the population. There is consequently dependence between this mutation site and nearby sites, and because of recombination, this dependence will decrease over time. Indeed, in proportion to the genetic distance between the mutation site and the considered site, recombinations can break the original mutant chromosome into pieces, and eventually the two sites are no longer transmitted together. The dependence between two alleles, due to their linkage or for other reasons, is called *allelic association* (see Elston (2000) for more details).

Association and mapping studies are usually conducted to study rare diseases. In order to get enough cases, the data usually come from a case/control study. The phenotype of interest can be quantitative or qualitative (usually diseased/non-diseased). For each individual in the sample, genetic markers are obtained for the region of interest. This region either corresponds to a region previously identified by linkage analysis in families, or it corresponds to the whole genome. The latter is called a *genome scan*.

We can classify methods using composite likelihoods for association or mapping in three categories. The first one assumes independence of marker loci, the second one assumes independence of individuals, and the last one takes into account the dependence between markers and individuals, but works only for small

regions in the sequence.

5.1. Independence of marker loci

The first approach is to build a marginal likelihood for each marker, and then form a composite likelihood by assuming independence of markers. In short, we can view this methodology as a pairwise likelihood approach, where the pair is formed by a genetic marker and the disease gene. Most of these methods use a special parameter to model the association between a marker and the disease gene, a parameter that is assumed to be a function of the distance between the genetic marker and the disease gene. Then, these approaches define a probability model, or use an existing one, to describe the data given the chosen parameter. Assuming independence of the markers, a composite likelihood is then built and used to obtain an estimate of the parameter of interest. Although all these methods have their own particularities, we present only a brief outline; the interested reader can refer to the included bibliography.

The first one of these methods was proposed by Terwilliger (1995): for a given genetic marker, and assuming that a particular allele is associated with the disease gene, the LD is measured by the increase in the frequency of this allele on the case chromosome; he defines a parameter λ_k, which depends on the recombination fraction between the disease gene and the marker k. Using simple probability rules and some assumptions about the genetic model, the probability of the data can be evaluated. A composite likelihood is then formed to combine the probability over all the markers loci, and a maximum composite likelihood estimate of the parameter is obtained. Xiong and Guo (1997) define a likelihood for one marker under a Wright-Fisher model, and approximate this likelihood using a first order Taylor series; then, a composite likelihood over all the markers loci is defined. They show that Terwilliger's method is actually a particular case of their method. Service et al. (1999) extend the Terwilliger approach by using triplets of markers, instead of pairs. Furthermore, a similar approach to Terwilliger's can be found in Devlin, Risch, and Roeder (1996), except that they take into account the stochastic aspect of the genealogy. Their LD parameter is the robust attributable risk (estimated by using joint frequencies of alleles between a marker and the disease), and this risk can be formulated in terms of a recombination fraction between the marker and the disease; they show (by simulation) that, under a Wright-Fisher model, a function of the robust attributable risk approximately follows a gamma distribution. As the probability of the data at one marker can be approximated given the distance between the marker and the disease gene, a composite likelihood is constructed by multiplying the likelihoods for each marker, and the location of the disease gene can then be estimated. Another important contribution to this type of method is the

work of Collins and Morton (1998); the general idea of their work is similar to the methods described above. Details of all these methods in the context of composite likelihood can be found in Lazzeroni (2001), Rannala and Slatkin (2000), and in Li (2008).

5.2. Independence of sequences

The second type of method using composite likelihoods consists of building a marginal likelihood for each sequence, and forming a composite likelihood by assuming that the sequences are independent, conditional on the type of the common ancestor of the sample. This approach is equivalent to assuming that the genealogy of the sampled chromosomes is star shaped (all the sequences are directly related to a common ancestor, giving a star shaped genealogy - $*$ - where the ancestor is the centre of the star, and the branches are the sequences), which in practice is likely not to be the case given that the sequences share the same history by the hypothesis of LD. In contrast, Slatkin and Rannala (2000) note that the star genealogy makes sense in a very rapidly growing population. The methods of Terwilliger (1995), described above, actually fall into this second category: the independence of markers within sequences, and the independence of sequences themselves is assumed (Rannala and Slatkin (2000)).

A different and more interesting example of this type of method is the one proposed by McPeek and Strahs (1999). They elegantly model the probability to observe a sequence, taking into account recombination and mutations, if the ancestral haplotype is known. Conditional upon a common ancestor and the location of the disease mutation, they develop a marginal likelihood for a sequence of markers using a coalescent model. Considering the ancestor and the location of the disease mutation as parameters, they propose a composite likelihood to combine the likelihood for all the sequences, evaluated by a hidden Markov algorithm. They obtain a point estimate and confidence interval; to take into account the shared ancestry of chromosomes (ignored by the composite likelihood), they proposed a correction factor that inflates standard error. Morris, Whittaker, and Balding (2000) and Liu et al. (2001) have transposed the McPeek and Strahs (1999) approach into a Bayesian framework.

5.3. Full likelihood on a subset of the data

Attempting to use all the information, while taking into account the dependence existing in both marker loci and sequences, comes to a relatively similar problem to the one described in Section 4, where the recombination rate was the parameter of interest. The main difference is that in the present case, the recombination rate is assumed to be known, and we wish to estimate the location of a disease mutation, while in the previous approach one estimated the unknown

recombination rate. The inference methods presented in Section 3.2 could be used here by changing the parameter of interest.

The first attempt to map a gene with a full likelihood approach was done in Larribe, Lessard, and Schork (2002). They describe the genealogy by using the coalescent with recombination, and adapt the importance sampling procedure proposed by Griffiths and Marjoram (1996) to estimate the position r_T of a disease mutation. Assuming a simple genetic model, the joint probability of the genetic data D and the phenotypic data T assuming the disease mutation is at position r_T, $\pi(D, T|r_T)$, can be estimated. The procedure involves estimating the likelihood independently for each of the $L-1$ inter marker intervals; note that the resulting likelihood is not a continuous function of r_T.

Although this approach is appealing because it uses all available information and takes the structure of the genealogy and the dependence between markers correctly into account, the method is so computer intensive that it is difficult to use it with sequences containing more than a few markers.

To reduce the computation size, while keeping the coalescent with recombination as a model for the genealogy, Larribe and Lessard (2008) propose to use windows of markers, similarly to Fearnhead and Donnelly (2002) (see Section 4.1). Let D_g be the genetic data within the gth window, then the likelihood contribution from this window is $\pi(D_g, T|r_T)$. In practice there are two problems in estimating these likelihoods, and thus the composite likelihood. First the likelihood is evaluated independently for each marker interval in each window, and these can be difficult to estimate accurately. This poses a real problem given that we have to compare likelihoods from different intervals and windows. Second, $\pi(D_g, T|r_T)$ does not use the same data from one window g to another. Therefore, a conditional marginal likelihood is defined for marker interval m in window g:

$$L_{m,g}(r_T) = \frac{\pi_g(D_g, T|r_T)}{\pi_g(D_g)},$$

where r_T lies within marker interval m in window g, and $\pi_g(D_g)$ is the observed sample configuration in windows g. This conditional marginal likelihood is evaluated in two steps for each interval m in window g: the numerator and the denominator have to be evaluated independently. Larribe and Lessard (2008) define a *composite-conditional-likelihood*:

$$CCL(r_T) = \prod_{m=1}^{L-1} \left[\prod_{g \in G(m)} L_{m,g}(r_T) \right]^{w_m},$$

where $G(m)$ is the set of windows that cover interval m. This composite-conditional-likelihood is a weighted product of conditional likelihood functions

associated with windows of k contiguous marker loci, where one conditions on the observed sample configuration at those genetic markers; the weight w_m is inversely proportional to the number of windows of that size encompassing the location of the mutation.

For example, if windows of size two are considered, then the likelihood is evaluated using pairs of contiguous markers, and the pairs are considered independent; this likelihood is very quick to compute, but does not give better estimates than simple pairwise statistics. If only one window of L contiguous markers is taken, then $CCL(r_T)$ is proportional to the full likelihood and would give the same estimate as before. Although windows can be disjoint, the current implementation uses overlapping windows. It is shown by simulation that the larger the window size, the closer $CCL(r_T)$ estimates the full likelihood $L(r_T)$ and the more difficult–computationally it is to evaluate the likelihood.

Simulation results show that the composite likelihood method gives similar results to the full likelihood as long as the windows are not too small (windows had to be approximately greater than 6 markers in size in the examples considered). Moreover, the improvement in calculation time is drastic, and the composite likelihood method allows inference from large data sets that would be impossible to analyse using full likelihood methods.

5.4. Comparison of methods

Few studies compare how different approaches behave. Garner and Slatkin (2002) estimate the position of a disease gene by using information on two markers, under a coalescent framework. They then compare this method to a single-marker likelihood and a composite likelihood by simulations. As expected, the method using two markers at the same time give a smaller confidence interval; their results show that the composite likelihood estimate can be biased.

Currently there are no theoretical results for any of these composite likelihood methods. One problem with deriving such results is that many of the composite likelihood methods are based on combining approximate likelihoods for subsets of the data – and it is not even clear what the properties of these approximate likelihoods are. There are also issues about appropriate asymptotic regimes. For example, increasing data by typing more markers that will be further and further away from the disease gene is unlikely to lead to consistent estimates of the position of the disease gene because the information in each marker will decay quickly with its distance from the disease gene. While asymptotics based on increasing sample size of individuals are likely to have substantially different properties depending on whether the disease gene is one of the marker loci or not. In the former case it should be possible to consistently estimate the position of the gene. By comparison, if the disease gene is not a marker locus,

consistent estimators are unlikely as, by chance, the markers in highest linkage disequilibrium with the disease gene may not be those closest to it.

6. Other Uses of Composite Likelihoods

Beside estimation of recombination rates, association and fine mapping methods, composite likelihoods have also been used in other contexts in statistical genetics. Here we give a brief overview of two areas, with indicative references.

6.1. Detecting genes under selection

Finding genes which are under selective pressure is of interest, as these genes are more likely to be linked to diseases. Also understanding the selective pressures on the human genome is informative about human evolution. Such genes can be detected by looking at the patterns of diversity at neutral markers near the gene. For example, a gene which has undergone a recent *selective sweep*, that is a favourable gene allele has rapidly increased in frequency in the population due to positive selection, will be surrounded by less genetic variation than other genes. Population genetic models can describe the likely patterns of genetic diversity at a single marker around a selective sweep, and composite likelihood methods can be used to combine such information across multiple marker loci. Kim and Stephan (2002) propose such a composite likelihood approach to test the hypothesis of selective sweep, the distribution under the null hypothesis being obtained by simulation (see also Zhu and Bustamante (2005)). Meiklejohn et al. (2004) use similar techniques for a related model of positive selection. In a similar context of selection, Kim and Nielsen (2004) propose a new composite likelihood by using pairs of sites (instead of a single site), where the distribution of various allelic configurations is obtained by simulation, in the way Hudson (2001) computed recombination (see Section 4.1).

Patterns of selection are often similar to those obtained under certain demographic models (e.g., how the population size has varied over time). Related work has been done to make the above inferences about selection robust to the underlying demographic model. The key idea is that demography affects diversity across the whole genome, while selection affects diversity locally. See for example Jensen et al. (2005). Composite likelihood methods have contributed to the production of genome-wide maps of genes under recent selection (Pickrell et al. (2009)).

6.2. Inference of demography

Composite likelihood methods have been used to infer the demographic history of a population, or a number of populations. Questions addressed include

how the population size has varied over time, to what extent there is random mating within a population, how much migration is there between populations, and how and when did current distinct populations evolve from ancestral populations. Different demographic models will potentially affect the observed genetic diversity in a sample in different ways, though there is difficulty in inferring between different demographic, biological, and selective models that produce similar patterns in data. There is great scope for composite likelihood methods and the effect of demography should be seen on markers across the whole genome, hence there is a need for practicable methods that can combine information across a, potentially large, number of marker loci.

Composite likelihood methods are most naturally defined based on likelihood data at a single marker, or for data at a small number of closely linked markers. For example, Gray et al. (2009) use the approach of Sawyer and Hartl (1992) to model the data at individual sites, and then derive a composite likelihood under the assumption of independence of data across sites. Plagnol and Wall (2006) infer ancestral human population structure based on data from 135 genes. They summarise the data within each gene by a set of summary statistics, calculate the likelihood of this set of summary statistics under different models, and then combine these likelihoods to produce a composite likelihood. Alternatively, Meligkotsidou and Fearnhead (2007) use composite likelihood to estimate models of spatial spread of a population, where they combine likelihoods from pairs of chromosomes.

Wiuf (2006) gives theoretical results supporting approaches that combine likelihoods at single markers, or regions of markers. Using a generalisation of Theorem 1, Wiuf shows that dependence across a genome decays as the inverse of genetic distance, and thus composite likelihood estimates of parameters in demographic models will be consistent as the number of markers increases.

7. Conclusion

Composite likelihoods are extensively used within genetics. The main motivation for this is computational: composite likelihood methods can scale so that they are computationally feasible for the large genetic data sets being generated. As data sets get bigger we would expect the importance and use of composite likelihood methods to also increase.

Currently, different composite likelihood methods are justified by empirical results (mainly through simulation studies) rather than theoretically. While the complex dependencies, across both individuals and markers, can make theoretical analysis of composite likelihood methods challenging, we feel there is an important role for developing practically relevant theory to underpin the application of these methods within genetics.

Important open questions relate to the asymptotic distribution of composite likelihood estimators: under what conditions will this be Gaussian, and how can we consistently estimate the variance of this distribution? For pairwise likelihood methods (e.g., Section 4.1) can we characterise what the optimal weights would be for the likelihood term for each pair of markers, and how these should decay with the distance between the markers? For applications in disease-gene mappings, results on the regimes under which we get consistent estimators, and how the rate of convergence differs for different approaches (such as assuming independence across markers or across individuals) are also important.

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (CRSNG) and the Fonds québécois de la recherche sur la nature et les technologies (FQRNT). We are thankful to N. Reid and S. Froda for their support.

References

Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots. *Genome Res.* **17**, 1219-1227.

Carvajal-Rodriguez, A., Crandall, K. A. and Posada, D. (2006). Recombination estimation under complex evolutionary models with the coalescent composite-likelihood. *Mol. Biol. Evol.* **23**, 817-827.

Collins, A. and Morton, N. E. (1998). Mapping a disease locus by allelic association. *Proc. Natl. Acad. Sci. USA* **95**, 1741-1745.

Cox, D. R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.

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

Devlin, B., Risch, N. and Roeder, K. (1996). Disequilibrium mapping: composite likelihood for pairwise disequilibrium. *Genomics* **36**, 1-16.

Elston, R. C. (2000). Introduction and overview. Statistical methods in genetic epidemiology. *Statist. Meth. Medical Res.* **9**, 527-541.

Fearnhead, P. (2003). Consistency of estimators of the population-scaled recombination rate. *Theor. Popul. Biol.* **64**, 67-79.

Fearnhead, P. (2006). SequenceLDhot: detecting recombination hotspots. *Bioinformatics* **22**, 3061-3066.

Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics* **159**, 1299-1318.

Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *J. Roy. Statist. Soc. Ser. A* **64**, 657-680.

Fearnhead, P. and Smith, N. G. C. (2005). A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. *Amer. J. Hum. Genet.* **77**, 781-794.

Garner, C. and Slatkin, M. (2002). Likelihood-based disequilibrium mapping for two-marker haplotype data. *Theor. Popul. Biol.* **61**, 153-161.

Gray, M. M., Granka, J. M., Bustamante, C. D., Sutter, N. B., Boyko, A. R., Zhu, L., Ostrander, E. A. and Wayne, R. K. (2009). Linkage disequilibrium and demographic history of wild and domestic canids. *Genetics* **181**, 1493-1505

Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479-502.

Griffiths, R. C. and Marjoram, P. (1997). An ancestral recombination graph. In *Progress in Population Genetics and Human Evolution,* (Edited by P. Donnelly and S. Tavaré), IMA Volumes in Mathematics and its Applications, 257-270. Springer Verlag, Berlin.

Griffiths, R. C. and Tavaré, S. (1994a). Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* **46**, 131-159.

Griffiths, R. C. and Tavaré, S. (1994b). Ancestral inference in population genetics. *Statist. Sci.* **9**, 307-319.

Griffiths, R. C. and Tavaré, S. (1994c). Sampling theory for neutral alleles in a varying environment. *Phil. Trans. R. Soc. Lond. B* **344**, 403-410.

Hein, J., Schierup, M. and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory.* Oxford University Press, Oxford.

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

Jensen, J. D., Kim, Y., DuMont, V. B., Aquadro, C. F. and Bustamante, C. D. (2005). Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**, 1401-1410.

Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513-1524.

Kim, Y. and Stephan, W. (2002). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* **155**, 1415-1427.

Kingman, J. F. C. (1982a). The coalescent. *Stochastic Process. Appl.* **13**, 235-248.

Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics.* (Edited by G. Koch and F. Spizzichino), 97-112. North-Holland, Amsterdam.

Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421-1430.

Kuhner, M. K., Yamato, J. and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**, 429-434.

Kuhner, M. K., Yamato, J. and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics* **156**, 1393-1401.

Larribe, F. and Lessard, S. (2008). A composite-conditional-likelihood approach for gene mapping based on linkage disequilibrium in windows of marker loci. *Statistical Applications in Genetics and Molecular Biology* **7**, 27.

Larribe, F., Lessard, S. and Schork, N. J. (2002). Gene mapping via the ancestral recombination graph. *Theor. Popul. Biol.* **62**, 215-229.

Lazzeroni, L. C. (2001). A chronology of fine-scale gene mapping by linkage disequilibrium. *Statist. Meth. Medical Res.* **10**, 57-76.

Li, N. (2008). The promise of composite likelihood methods for addressing computationally intensive challenges. *Adv. Genetics* **60**, 637-654.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Math.* **80**, 221-239.

Liu, J. S., Sabatti, C., Teng, J., Keats, B. J. and Risch, N. (2001). Bayesian analysis of haplotypes for linkage disequilibrium mapping. *Genome Res.* **11**, 1716-1724.

McPeek, M. S. and Strahs, A. (1999). Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale gene mapping. *Amer. J. Hum. Genet.* **65**, 858-875.

McVean, G., Awadalla, P. and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231-1241.

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

Meligkotsidou, L. and Fearnhead, P. (2007). Postprocessing of genealogical trees. *Genetics* **177**, 347-358.

Meiklejohn, C. D., Kim, Y., Hartl, D. L. and Parsh. J. (2004). Identification of a locus under complex positive selection in drosophila simulans by haplotype mapping and composite-likelihood estimation. *Genetics* **168**, 265-279.

Morris, A. P., Whittaker, J. C. and Balding, D. J. (2000). Bayesian fine-scale mapping of disease loci, by hidden Markov models. *Amer. J. Hum. Genet.* **67**, 155-169.

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

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

Nielsen, R. and Signorovitch, J. (2003). Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* **63**, 245-255.

Nordborg, M. (2007). Coalescent theory. In *Handbook of Statistical Genetics* (Edited by D. J. Balding, M. J. Bishop, and C. Cannings), 843-877. John Wiley & Sons, Inc., Chichester, U.K.

Nordborg, M. and Tavaré, S. (2002). Linkage disequilibrium: What history has to tell us. *Trends in Genetics* **18**, 83-90.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W. and Pritchard, J. K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19**, 826-837.

Plagnol, V. and Wall, J. D. (2006). Possible ancestral structure in human populations. *PLoS Genetics* **2**, e105.

Rannala, B. and Slatkin, M. (2000). Methods for multipoint disease mapping using linkage disequilibrium. *Genet. Epidemiol.* **19**, S71-S77.

Sawyer, S. A. and Hartl, D. L. (1992). Population-genetics of polymorphism and divergence. *Genetics* **132**, 1161-1176.

Service, S. K., Temple, D. W., Freimer, N. B. and Sandkuijl, L. A. (1999). Linkage disequilibrium mapping of disease genes by reconstruction of ancestral haplotypes in founder populations. *Amer. J. Hum. Genet.* **64**, 1728-1738.

Slatkin, M. and Rannala, B. (2000). Estimating allele age. *Annual Review of Genomics and Human Genetics* **1**, 225-249.

Smith, N. G. C. and Fearnhead, P. (2005). A comparison of three estimators of the population-scaled recombination rate: accuracy and robustness. *Genetics* **171**, 2051-2062.

Stephens, M. (2007). Inference under the coalescent. In *Handbook of Statistical Genetics* (Edited by D. J. Balding, M. Bishop and C. Cannings), 878-908. Wiley, Inc., Chichester, U.K.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. Roy. Statist. Soc. Ser. B* **62**, 605-655.

Stephens, M., Smith, N. J. and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Amer. J. Hum. Genet.* **68**, 978-989.

Tavaré, S. and Zeitouni, O. (2004). *Lectures on Probability Theory and Statistics: Ecole d'Eté de Probabilités de Saint-Flour XXXI* – 2001. (Edited by J. Picard), Lecture Notes in Mathematics, Vol. 1837. Springer-Verlag, Berlin and Heidelberg.

Terwilliger, J. D. (1995). A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *Amer. J. Hum. Genet.* **56**, 777-787.

Thompson, E. A. (2000). Statistical inferences from genetic data on pedigrees. *NSF-CBMS Regional Conference Series in Probability and Statistics*. Volume 6. IMS, Beachwood, OH.

Varin, C. (2008). On composite marginal likelihoods. *Adv. Statist. Anal.* **92**, 1-28.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods, *Statist. Sinica* **21**, 5-42.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika* **92**, 519-528.

Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado.

Wilson, I. J. and Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics* **150**, 499-510.

Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., McVean, G. A. T., Gabriel, S. B., Reich, D., Donnelly, P. and Altshuler, D. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107-111.

Wiuf, C. (2006). Consistency of estimators of population scaled parameters using composite likelihood. *J. Math. Biol.* **53**, 821-841.

Xiong, M. and Guo, S. W. (1997). Fine-scale genetic mapping based on linkage disequilibrium: Theory and applications. *Amer. J. Hum. Genet.* **60**, 1513-1531.

Zhu, L. and Bustamante, C.D. (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* **170**, 1411-1421.

Département de Mathématiques, Université du Québec à Montréal, Montréal, H3C 3P8, Canada.

E-mail: larribe.fabrice@uqam.ca

Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YW, UK.

E-mail: p.fearnhead@lancaster.ac.uk

(Received October 2009; accepted August 2010)