

*Constructing Parental Linkage Phase And
Genetic Map Over Distances $< 1cM$
Using Pooled Haploid DNA*

Dario Gasbarra and Mikko J. Sillanpää

University of Helsinki, Finland

October 21, 2005

Short running title: Gametic Pools and Estimation

Corresponding author: Dario Gasbarra, Ph.D.,
Department of Mathematics and Statistics,
P.O. Box 68,
FIN-00014 University of Helsinki, Finland

E-mail: dag@rolf.helsinki.fi

Fax number: (358) 9-191-51400

Keywords: DNA pooling, genetic map, gene order, linkage phase

Abstract

A new statistical approach for construction of the genetic linkage map and estimation of the parental linkage phase based on allele frequency data from pooled gametic (sperm or egg) samples is introduced. The method can be applied for estimation of recombination fractions (over distances < 1 cM) and ordering of large number (even hundreds) of closely linked markers. This method should be extremely useful in species with a long generation interval and a large genome size like in dairy cattle or forest trees; the conifer species have haploid tissues available in megagametophytes. According to Mendelian expectation, two parental alleles should occur in gametes in 1:1 proportions, if segregation distortion does not occur. However, due to mere sampling variation, the observed proportions may deviate from their expected value in practice. These deviations and their dependence along the chromosome can provide information on the parental linkage phase and on the genetic linkage map. Usefulness of the method is illustrated with simulations. The role of segregation distortion as a source of these deviations is also discussed. The software implementing the method is freely available for research purposes from the authors.

Introduction

Estimation of recombination fraction over distances < 1 cM is important because current genetic maps are very inaccurate in such distances (Arnheim et al. 2003; Kong et al. 2002; Weber 2002) and because of current interest towards characterization and utilization of haplotype-block structure of the human genome (Arnheim et al. 2003; The International HapMap Consortium 2003). For excellent general review of characterization of recombination in small distances, see Arnheim et al. (2003).

In order to estimate parental linkage phase and short map distances in range < 1 cM one would ideally require data from several gametes (haploid tissues). In species with a short generation cycle it may be possible to produce "map expansion" (an excess of recombinants) by applying

a number of consecutive intercrosses (Darvasi 1998). Several haplotyping methods for pedigree data (Wijsman 1987; Sobel and Lange 1996; Sobel et al. 1996; Kruglyak et al. 1995,1996; Lin and Speed 1997; Tapadar et al. 2000; Qian and Beckmann 2002; Gao et al. 2004) and for general population samples (Clark 1990; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995; Kitamura et al. 2002; Stephens et al. 2001; Stephens and Scheet 2005) have been introduced during the last two decades. Similarly, there is a rich literature of methods for construction of the genetic linkage map (Lander and Green 1987; Stephens and Smith 1993; George et al. 1999; Jansen et al. 2001; Butcher et al. 2002; Wu et al. 2002; Rosa et al. 2002; Lu et al. 2004; George 2005). In species such as forest trees or dairy cattle, where the generation cycle is long and the genome size is large, all map distances and ordering of markers are generally difficult to estimate. However, certain types of haploid tissues like sperm, eggs, or megagametophytes, may be easily available in such species. Use of individual sperm typing (Navidi and Arnheim 1994, 1999; Lazzeroni et al. 1994; Cullen et al. 2002; Arnheim et al. 2003) and utilization of DNA in megagametophytes (Tulsieram et al. 1992; Yazdani et al. 1995) for the estimation of genetic map (order and distances) has been proposed in the literature earlier. However, extensive individual typing of haploid tissues is not cost-effective. Note that application of radiation-hybrid mapping has also been proposed for a similar purpose (Boehnke et al. 1991; Slonim et al. 1997).

Use of the pooled DNA data reduces the cost and time spent on typing by allowing for direct determination of allele frequencies at each locus (Shaw et al. 1998; Collins et al. 2000; Sham et al. 2002; Ritland 2002; Norton et al. 2004; Butcher et al. 2004). Pooled DNA data on diploid tissues can also be used to estimate haplotype frequencies (Ito et al. 2003; Yang et al. 2003). We assume here that allele frequency measurements from pooled DNA can be obtained with a high accuracy in the lab (Norton et al. 2002; Sham et al. 2002; Butcher et al. 2004; Yang et al. 2005). If haploid tissues from single parent are used in the pool and the possibility of segregation distortion is excluded, the average proportions of the two parental alleles is even at each locus. However, the observed proportions often deviate from expected proportions and are correlated between the loci (due to close linkage of the loci) which can provide information of the linkage

phase (haplotypes) of the parent. The information in these fluctuations is stochastic in nature and can vary somewhat from sample to sample. Note that these fluctuations, when measured from large samples, can be used to detect segregation distortion (McPeck 1999; and Discussion). When there is no significant evidence of segregation distortion, the asymptotic properties of these fluctuations (and their dependence) as a source of information in estimating linkage phase of the parent at short map distances are described in this article.

Hidden Markov models (HMMs) provide a flexible modeling framework for several types of problems (Rabiner 1989), including the construction of the linkage map (Lander and Green 1987). The Viterbi algorithm is a maximization algorithm which can be used in HMMs to determine the optimal path (having the highest likelihood value) through the hidden state structure. The same algorithm can also be used to sample directly from the posterior distribution. In the following, we present our haplotyping method which is based on the Viterbi algorithm. Note that also well known Genehunter program uses Viterbi algorithm for haplotyping (Kruglyak et al. 1995, 1996).

Model

We first describe the model for individual gametic observations and then introduce required modifications for that under pooled gametic observations.

Individual gametic observations: Assume that the sample consists of n gametes h^1, \dots, h^n , that have been collected from a diploid individual (the parent). The parent is assumed to be a fully heterozygote in a given set of L polymorphic marker loci so that at each locus there are two segregating alleles (denoted as 0 and 1) that can be distinguished in the sample. In practice, we omit the loci in our sample, where only a single allele can be identified (i.e. the parent is homozygote). Each gamete i can then be represented as a vector $h^i = (h_1^i, \dots, h_L^i) \in \{0, 1\}^L$.

Denote by X_l the grandparental origin of the 0-allele at locus l , with the convention that when $X_l = 0$ ($X_l = 1$) the origin of the 0-allele is grandpaternal (grandmaternal), respectively. Note that the assignment of the vector $X = (X_1, \dots, X_L) \in \{0, 1\}^L$ uniquely determines

the parental linkage phase. *A priori*, $(X_l : l = 1, \dots, L)$ are independent random variables, $p(X) = \prod_{l=1}^L p(X_l)$, with respective Bernoulli(π_l) distributions. We assume that both grandparental origins are *a priori* equally likely at each locus, i.e., $\pi_l = \frac{1}{2}$ corresponding to linkage equilibrium assumption for each locus l (for potential problems, see Schaid et al. 2002; Huang et al. 2004.) However, in case we have some prior information about parental linkage phase, e.g., genotype information from the grandparents, or knowledge about linkage disequilibrium among the loci, we could set $\pi_l \neq \frac{1}{2}$, or in the latter case, assume *a priori* a Markov model for $(X_l : l = 1, \dots, L)$ (see McPeck and Strahs 1999).

Denote the vector of recombination fractions by $\theta = (\theta_l : l = 2, \dots, L)$, where an element θ_l is the recombination fraction between two consecutive loci $(l-1)$ and l . The likelihood function is the following:

$$p(h^1, \dots, h^n | X, \theta) = \prod_{i=1}^n \left[p(h_1^i) \prod_{l=2}^L p(h_l^i | h_{l-1}^i, X_{l-1}, X_l, \theta_l) \right].$$

At the first locus above, gametes h_1^i , $i = 1, \dots, n$ are independent Bernoulli(1/2) variables. Then, given X_{l-1}, X_l and $(h_{l-1}^i : i = 1, \dots, n)$, gametes $(h_l^i : i = 1, \dots, n)$ are conditionally independent with the following distribution:

if $X_l = X_{l-1}$ (i.e. the grandparental origin of the 0-allele remains the same) then

$$p(h_l^i = h_{l-1}^i | h_{l-1}^i, X_{l-1}, X_l, \theta_l) = (1 - \theta_l) = 1 - p(h_l^i = 1 - h_{l-1}^i | h_{l-1}^i, X_{l-1}, X_l, \theta_l).$$

and if $X_l \neq X_{l-1}$ (i.e., the grandparental origins of the 0-allele differ) then

$$p(h_l^i = h_{l-1}^i | h_{l-1}^i, X_{l-1}, X_l, \theta_l) = \theta_l = 1 - p(h_l^i = 1 - h_{l-1}^i | h_{l-1}^i, X_{l-1}, X_l, \theta_l).$$

This formulation corresponds to the standard transmission model, e.g., Sillanpää and Arjas (1999); see their equations (4) and (5). Here, although the phases (X_l) are *a priori* independent, the vectors $(X_l, h_l^i), l = 1, \dots, L$, form a Markov process, which gives an Hidden Markov Model (HMM) with hidden layer (X_l) . Although the parametrization is different, this is the same model assumed in Lander and Green (1987) and in Kruglyak et al. (1995, 1996), where a bit in their inheritance vector includes information from both the grandparental origin (X_l) and

allele in the gamete (h_l^i). Given the individual gametic observations and the recombination fractions between the loci, it is possible to apply variant of the Viterbi algorithm to compute the posterior distribution of the parental linkage phase $p(X | h^1, \dots, h^n, \theta)$, which is proportional to $p(h^1, \dots, h^n | X, \theta) \times p(X)$. It is also possible to implement the EM-algorithm (Dempster et al. 1977) to compute the maximum likelihood estimator of the recombination fractions (θ).

Pooled gametic observations: Next we consider the case, where instead of observing the individual gametes (h_l^i) we observe their (allele) frequencies from a gametic pool, or equivalently

$N_l = \sum_{i=1}^n h_l^i$, the number of copies of the 1-allele at locus l . The likelihood function for pooled observations is the following:

$$p(N | X, \theta) = \left[p(N_1) \prod_{l=2}^L p(N_l | N_{l-1}, X_{l-1}, X_l, \theta_l) \right].$$

At the first locus above, the frequency N_1 is independent from phase vector X and it has the Binomial($n, 1/2$) distribution. For loci $l > 1$, given the phase information X_{l-1} , X_l and the frequency N_{l-1} at previous locus, N_l has the following conditional distribution (given by convolution of two binomials):

if $X_l = X_{l-1}$ (i.e. the grandparental origin of the 0-allele remains the same) then

$$p(N_l = s | N_{l-1}, X_{l-1}, X_l, \theta_l) = \sum_{i=0}^s \binom{N_{l-1}}{i} \binom{n - N_{l-1}}{s - i} \theta_l^{(N_{l-1} + s - 2i)} (1 - \theta_l)^{(n - N_{l-1} - s + 2i)}$$

and if $X_l \neq X_{l-1}$ (i.e., the grandparental origins of the 0-allele differ) then

$$p(N_l = s | N_{l-1}, X_{l-1}, X_l, \theta_l) = \sum_{i=0}^s \binom{N_{l-1}}{i} \binom{n - N_{l-1}}{s - i} (1 - \theta_l)^{(N_{l-1} + s - 2i)} \theta_l^{(n - N_{l-1} - s + 2i)}$$

Alternatively we can use the following notation:

$$N_l \sim \begin{cases} \text{Binomial}(N_{l-1}, 1 - \theta_l) * \text{Binomial}(n - N_{l-1}, \theta_l) & \text{if } X_l = X_{l-1} \\ \text{Binomial}(N_{l-1}, \theta_l) * \text{Binomial}(n - N_{l-1}, 1 - \theta_l) & \text{if } X_l \neq X_{l-1} \end{cases}$$

where $*$ denotes convolution. These convolution distributions can be computed efficiently using Discrete Fourier Transform (see e.g., Bremaud 2002).

Multiple pools: We can also consider the case where we observe different frequencies from n_{pools} distinct pools, each containing $n^{(k)}$ gametes, $k = 1, \dots, n_{pools}$. Accordingly, the pool-specific frequency vectors $N^{(k)} = (N_1^{(k)}, \dots, N_L^{(k)}) \in \{0, 1, \dots, n^{(k)}\}^L$, $k = 1, \dots, n_{pools}$, are conditionally independent given the phase vector X , starting with initial distribution $N_1^{(k)} \sim \text{Binomial}(n^{(k)}, 1/2)$ and then following independently the Markovian dynamics described above. To maintain simplicity of the notation in the rest of the paper, the theory is presented mainly for a single pool but it generalizes to multiple pools analogously.

Estimation of Linkage Phase: It may be somewhat surprising, but becomes clear from the formula above, that knowledge of the frequencies N_{l-1} and N_l gives some information about the change $|X_l - X_{l-1}|$ in the grandparental origin. The amount of such information is quantified by the total variation distance between the two distributions above; clearly it depends on N_{l-1} , n , and θ_l , and is random in this sense. This information is zero when $\theta_l = 1/2$, and it increases as the recombination fraction proceeds to 0. Also, it is zero when $N_{l-1} = n/2$ and increases as N_{l-1} goes to 0 or to n .

Again, the pairs $(X_l, N_l : l = 1, \dots, L)$ form an hidden Markov chain (Figure 1), and it is possible to apply HMM techniques and EM-algorithm to estimate parental linkage phase (X) and recombination fractions (θ), respectively. This practice is similar than in the full data case. The Viterbi algorithm is described in detail in the Appendix A. We show that in case the marker spacing is dense and several distinct pools are used, the information content of the pooled data about the parental linkage phase X is not much lower than the information carried by the full data.

(Figure 1 will be placed here)

Normal Approximation and Information about Linkage Phase

In following, we present a normal approximation for transitions between allele frequency measurements at different loci which may be used in order to speed up the numerical computations,

when the size of the gametic pools are large enough so that the binomial distribution can be approximated with the normal distribution. We also compute the *information in the data* and explain how to calculate the *mutual information* between linkage phase (X_1, \dots, X_L) and the observed allele frequency data (N_1, \dots, N_L) under the approximate model. Let

$$\tilde{N}_l \sim \begin{cases} N_l & \text{if } X_l = 0 \\ n - N_l & \text{if } X_l = 1 \end{cases}$$

be the frequency of the allele inherited from the grandfather at locus l , where n is the size of the gametic pool. The conditional distribution of \tilde{N}_l given \tilde{N}_{l-1} is a convolution of two binomials with mean $m(\theta) = \tilde{N}_{l-1} + (n - 2\tilde{N}_{l-1})\theta_l$ and variance $n(1 - \theta_l)\theta_l$. When n is large enough, we approximate the scaled process $(n^{-1}\tilde{N}_l)$ by a stationary Gaussian process $(\tilde{\xi}_l)$ with transition density

$$\tilde{\xi}_l \mid \tilde{\xi}_{l-1} \sim \mathcal{N}(\tilde{\xi}_{l-1}(1 - 2\theta_l) + \theta_l, \theta_l(1 - \theta_l)/n)$$

and invariant distribution $\mathcal{N}(1/2, (4n)^{-1})$.

Analogously the conditional distribution of the frequency N_l given N_{l-1} and $|\Delta X_l| = |X_l - X_{l-1}|$, is approximated by a normal distribution with mean $m(\theta) = N_{l-1} + (n - 2N_{l-1})r(|\Delta X_l|, \theta_l)$ and variance $n(1 - \theta_l)\theta_l$, where we denote $r(0, \theta) = \theta$ and $r(1, \theta) = 1 - \theta$.

After rescaling by $n^{-1/2}$, asymptotically we are in a conditionally Gaussian shift experiment (see van der Vaart 1998) with shift $(1 - 2\theta_l)(\sqrt{n} - 2N_{l-1}/\sqrt{n})$ and variance $\theta_l(1 - \theta_l)$. Note that under the marginal distribution the random variable $(2N_{l-1}/\sqrt{n})$ has mean \sqrt{n} and variance 1, therefore the shift is bounded in probability as n grows and the information about $|\Delta X_l|$ remains bounded too. On the other hand the experiment becomes more and more informative as the recombination fraction θ_l decreases, since the shift is a standardized random variable while the variance of the experiment goes to 0.

The observed information (information in the data) is the relative entropy (Kullback and Leibler 1951) of the prior distribution with respect to the posterior distribution, which is decomposable as follows

$$K(\text{Distr}_\theta(X_1, \dots, X_L | N_1, \dots, N_L); \text{Distr}(X_1, \dots, X_L)) = \sum_{l=1}^L \left\{ \frac{1}{p_\theta(N_l | N_{l-1})} \sum_{s=0,1} p(|\Delta X_l| = s) p_\theta(N_l | N_{l-1}, |\Delta X_l| = s) \log(p_\theta(N_l | N_{l-1}, |\Delta X_l| = s)) - \log(p_\theta(N_l | N_{l-1})) \right\}.$$

See Figure 2 for illustration. Note that the parameter vector θ is known here.

(Figure 2 will be placed here)

The mutual information is obtained by integrating out in each summand the pair (N_{l-1}, N_l) with respect to its joint distribution, that is N_{l-1} is $\mathcal{N}(n/2, n/4)$ and then N_l given N_{l-1} is mixed normal with mixing parameter $|\Delta X_l|$.

Remark. Here the size of the pool is related to the precision of the measurement, since it is implicitly assumed that one has perfect measurements with unit resolution. It would be more realistic to include measurement errors in the model (see Discussion).

Estimation of Recombination Fractions

While in the previous sections the recombination fractions were assumed to be known, we suppose them to be unknown here. By using a Monte Carlo EM-algorithm (Penttinen 1984; Geyer and Thompson 1992), we shall compute the maximum likelihood estimator of the recombination fractions $\theta = (\theta_l : l = 2, \dots, L) \in [0, 1/2]^{L-1}$ based on the pooled data.

Recall that conditionally on N_{l-1} and $|\Delta X_l|$,

$$N_l \sim \text{Binomial}(N_{l-1}, 1 - r(|\Delta X_l|, \theta_l)) * \text{Binomial}(n - N_{l-1}, r(|\Delta X_l|, \theta_l)).$$

Therefore we can write $N_l = y_l + (N_l - y_l)$ where, for $l > 1$, $y_l \sim \text{Binomial}(N_{l-1}, 1 - r(|\Delta X_l|, \theta_l))$ and $(N_l - y_l) \sim \text{Binomial}(n - N_{l-1}, r(|\Delta X_l|, \theta_l))$ are conditionally independent random variables. We set $y_1 = 0$. Note that $(X_l, y_l, N_l : l = 1, \dots, L)$ is a sufficient statistics for the recombination fractions $(\theta_l : l = 2, \dots, L)$.

In order to easily solve the maximization step in the EM-algorithm, we extend the state space by including the variables y_l to the hidden states X_l . The hidden Markov model $\{(X_l, y_l, N_l)\}$ has the following form: conditionally on $(X_{l-1}, y_{l-1}, N_{l-1})$,

$$X_l | X_{l-1}, y_{l-1}, N_{l-1} \sim \text{Bernoulli}(\pi_l)$$

$$y_l | X_{l-1}, X_l, y_{l-1}, N_{l-1} \sim \text{Binomial}(N_{l-1}, 1 - r(|X_l - X_{l-1}|, \theta_l)) \text{ and}$$

$$N_l | X_{l-1}, X_l, y_{l-1}, y_l, N_{l-1} \sim y_l + \text{Binomial}(n - N_{l-1}, r(|X_l - X_{l-1}|, \theta_l))$$

To sample $(X_l, Y_l : l = 1, \dots, L)$ conditionally on the pooled data (N_l) , we first sample (X_l) conditionally on (N_l) using the Viterbi algorithm described in the Appendix A, and then for $l = 1, \dots, L$ we sample value for Y_l conditionally on $X_{l-1}, X_l, N_{l-1}, N_l$ from a distribution

$$p(Y_l = y_l | X_{l-1}, X_l, N_{l-1}, N_l) \propto$$

$$\binom{N_{l-1}}{y_l} \binom{n - N_{l-1}}{N_l - y_l} r(|\Delta X_l|, \theta_l)^{N_{l-1} + N_l - 2y_l} (1 - r(|\Delta X_l|, \theta_l))^{n + 2y_l - N_{l-1} - N_l}.$$

Starting with some initial guess $\hat{\theta}_0$, at stage t , in the E-step we sample K independent identically distributed realizations $(X^{(j)}, Y^{(j)})$ conditionally on the pooled data (N) , and compute a Monte Carlo approximation of the integrated log-likelihood

$$\mathbb{E}_{\hat{\theta}_{t-1}} [\log(p_\theta(N, Y, X)) | N] \simeq \frac{1}{K} \sum_{j=1}^K \log(p_\theta(N, X^{(j)}, Y^{(j)})).$$

In the M-step we obtain the next estimate $\hat{\theta}_t$ by maximizing this expression over θ , taking $\hat{\theta} = (\hat{\theta}_t)$ with

$$\hat{\theta}_t := \frac{1}{nK} \left(\sum_{1 \leq j \leq K: \Delta X_t^{(j)} = 0} (N_{l-1} + N_l - 2Y_l^{(j)}) + \sum_{1 \leq j \leq K: |\Delta X_t^{(j)}| = 1} (n - N_{l-1} - N_l + 2Y_l^{(j)}) \right).$$

When we iterate the procedure, $\hat{\theta}_t \rightarrow \hat{\theta}_{ML} = \arg \max_{\theta} p_\theta(N)$.

This extends immediately to the situation where we observe distinct pooled gametic samples produced by unrelated individuals belonging to the same species.

Remark. For a given sample size, one can improve the estimation of the recombination fraction between two loci by measuring the frequencies at a finer map resolution between the two loci.

Simulation Experiment I: Estimation of Recombination Fractions and Parental Linkage Phase Under Different Pooling Strategies

Simulation were used to test performance of the algorithm under different pooling strategies. We simulated a data set containing 1000 haploid gametes across 200 evenly spaced loci where the recombination fraction between consecutive markers was determined to be 0.01. To illustrate the effect of different pooling strategies we computed the maximum likelihood estimator of recombination fraction in several situations, namely: (1) full data case, where we had observations from individual haplotypes (gametes), (2) pooled data cases, where we splitted the same data (1000 gametes) either into 1, 2, 4, or 8 different parts (pools) and had observed only the allele frequency of each pool. The Monte Carlo-EM algorithm starts from some arbitrary values for the recombination fractions and it is continued until convergence. In each Monte Carlo step, 5000 independent identically distributed realizations were drawn. The ML-estimates of the recombination fraction corresponding to these different situations are shown in Figure 3. One can see in the figure that the use of several smaller pools simultaneously improves the accuracy. In particular, the estimated recombination fraction, corresponding to a single pool (curve a), at locus 80 suddenly jumps towards infinity. This does not correspond to a change of chromosome but it is an artifact since at that locus the frequency is exactly 0.5 and there is no information about the parental linkage phase and the recombination fraction cannot be estimated. We see that when multiple pooling is used, we recover some information about the parental phase at that locus so that the “chromosome jump” disappears. By increasing the number of pools (up to 8) the parental phase was recovered without errors.

(Figure 3 will be placed here)

Simulation Experiment I: Robustness with respect to errors in data

From simulation experiment I, data of 8 pools (with $n = 125$ gametes in each) were further selected to study influence of errors to the estimation procedure. Two different kinds of errors were added to the observed allele frequencies in the pooled data.

Measurement errors: We first considered a perturbation in the data where the measured allelic frequencies are masked (at each locus and independently across loci) by adding a Gaussian error with mean 0 and standard deviation $\sqrt{n}\sigma$ with $\sigma = 0.01$. This was hoped to mimic an error attributable to the measuring instrument. In the numerical example, with this level of noise, the parental linkage phase was recovered without errors. However, the recombination fraction estimator is clearly not robust against this type of error and provided the values that were highly overestimated (see Figure 4). This is because after phase assignment the algorithm counts the remaining fluctuations in the data as recombinations.

DNA amplification errors: We also considered a second type of perturbation, which may occur during the DNA amplification process. In our model, pooled DNA sample is assumed to contain a single copy of each gametic observation (a haploid DNA) collected from the parent. To simulate DNA amplification errors, we add a simple perturbation to the pooled data sample by randomizing the number of copies of each gametic observation. Namely a collected gamete with index i is copied C^i times, where we assume that C^i are independently and identically distributed for some parameter $\varepsilon \in [0, 1]$,

$$P(C^i = k) = \exp(-\varepsilon) \left(\frac{\varepsilon^{k+1}}{k!} + \frac{(1-\varepsilon)\varepsilon^{k-1}}{(k-1)!} \right),$$

which is the convolution of a Poisson(ε) and a Bernoulli($1-\varepsilon$) distributions. This gives $E(C^i) = 1$ and $\text{Var}(C^i) = (2-\varepsilon)\varepsilon$. For more sophisticated modeling of DNA amplification, see Lalam et al. (2004). After DNA amplification, we obtain a pool containing $\check{n} = \sum_{i=1}^n C^i$ gametes, and we observe at each locus the perturbed counts $\check{N}_i = \sum_{i=1}^n C^i h_i^i$, with $E(\check{N}_i | N_i) = N_i$, $\text{Var}(\check{N}_i | N_i) = (2-\varepsilon)\varepsilon N_i$. Obviously these perturbations are correlated across loci, and this is the reason why the recombination fraction estimator is expected to be robust against DNA amplification errors.

In a numerical experiment that has been summarized in Figure 4, we tested several experiments with having different levels ($\varepsilon = 0.1, 0.2, 0.5, 1.0$) of DNA-amplification errors in each. In all these experiments, the recombination fraction estimator seemed to perform reasonably well and the linkage phase was recovered without errors.

(Figure 4 will be placed here)

Joint Estimation of Recombination Fractions, Parental Linkage Phase, and the Ordering of Markers

Next we consider the case where the ordering of markers is unknown and it is estimated from the data together with the recombination fractions and the parental linkage phase. We observe the allele frequencies from n_{pools} distinct gametic pools obtained from the same parental individual.

The state space is given by a permutation π of the marker loci indexes $\{1, \dots, L\}$, together with the recombination fractions θ_l between the consecutive markers indexed by $\pi(l-1)$ and $\pi(l)$, for $l = 2, \dots, L$, and the parental phase vector (X_1, \dots, X_L) .

In order to speed up the numerical computations we assume that the gametic pools have sizes large enough so that we can use the normal approximation.

Marginal likelihood and Pseudolikelihood

Consider a pair of markers, j and l , together with the corresponding parental linkage phases $X_j, X_l \in \{0, 1\}$. The loglikelihood contribution when marker l follows marker j with given phases is given by

$$L_\theta((j, X_j), (l, X_l)) = -n_{pools} \frac{1}{2} \log(\theta(1-\theta)) + \frac{-\bar{n}_{ll} - (2\theta - 1)^2 \bar{n}_{jj} - 2(2\theta - 1) \bar{n}_{jl} - \theta(\bar{n}\theta - 2(\bar{n}_l + (2\theta - 1)\bar{n}_j))}{2\theta(1-\theta)},$$

where θ is the recombination fraction between the markers j and l , $\bar{n} = \sum_{k=1}^{n_{pools}} n^{(k)}$ is the total number of gametes in the pools and we set

$$\begin{aligned}\bar{n}_l &= \sum_{k=1}^{n_{pools}} \left(N_l^{(k)} + (n^{(k)} - 2N_l^{(k)})X_l \right) \\ \bar{n}_{jl} &= \sum_{k=1}^{n_{pools}} \left(N_l^{(k)} + (n^{(k)} - 2N_l^{(k)})X_l \right) \left(N_j^{(k)} + (n^{(k)} - 2N_j^{(k)})X_j \right) / n^{(k)}.\end{aligned}$$

Note that (\bar{n}_l) and the matrix (\bar{n}_{jl}) are sufficient statistics, which need to be computed only once for every pair of markers. Their dimension does not depend on the number of the pools.

Next, we assume a uniform prior on $[0, 1/2]$ on the recombination fractions, and by integrating the recombination fraction parameter θ with respect to the prior we obtain the marginal likelihood contribution for the joint choice of the ordering and the linkage phase, and we define the cost function by

$$C((j, X_j), (l, X_l)) = -\log \left(\int_0^{1/2} \exp[L_\theta((j, X_j), (l, X_l))] d\theta \right).$$

Maximizing the likelihood correspond to solving a traveling salesman problem, i.e., finding an ordering π of loci $\{1, \dots, L\}$ and a phase vector (X_1, \dots, X_L) which minimizes the total cost

$$C_0(\pi(1)) + \sum_{l=2}^L C((\pi(l-1), X_{l-1}), (\pi(l), X_l)),$$

where the cost for the first locus $C_0(l) := 2\bar{n}_{ll} + \bar{n}/2 - 2\bar{n}_l$.

For a small number of markers it is possible to find the optimal ordering in the maximum likelihood sense by using dynamic programming. However dynamic programming becomes unfeasible as the number of markers grows. Instead we develop a Markov chain Monte Carlo (MCMC) method (Hastings 1970) to sample from the posterior distribution of the configuration $((\pi(l), X_l : l = 1, \dots, L), (\theta_l : l = 2, \dots, L))$.

Another problem is that the integral in the cost function does not have an analytic expression. For a large number of markers, it becomes numerically expensive to evaluate accurately the cost matrix $\{C((j, X_j), (l, X_l)) : j, l = 1, \dots, L\}$. We consider instead the pseudo-loglikelihood

$$(1) \quad \Psi(\pi) := -\sum_{l=2}^L (\bar{n}_{\pi(l), \pi(l)} + \bar{n}_{\pi(l-1), \pi(l-1)} - 2\bar{n}_{\pi(l-1), \pi(l)}).$$

Intuitively, the optimal ordering should be also close to optimal in the least-squares sense, and this gives a guideline to set the proposal distribution for the ordering π in the McMC algorithm. For updating steps and the proposal distributions, see Appendix B.

Simulation Experiment II: Joint Estimation of Recombination Fractions, Parental Linkage Phase, and the Ordering of Markers

We computed a numerical experiment with 400 closely linked marker loci. The markers were not equally spaced, instead the distance between consecutive markers was either 0.001 or 0.01 Morgans, systematically in the proportion of 10 to 1. We simulated data from 500 pools, with 200 gametes in each pool, and then applied a random permutation to the marker indices and to the parental phases before collecting the pooled data.

The McMC estimation algorithm found quickly (say, in a few thousands iterations) some good permutations, quite close to the true ordering of the markers. We stopped the experiment after two weeks and several millions of McMC cycles, because the algorithm was clearly stucked in a local maxima. Since the mixing of the McMC seems to be very slow, instead of using the empirical distribution as an approximation to the posterior, we look only at the maximum a posteriori (MAP), estimated by the sampled configuration with the highest posterior density.

The pseudo-loglikelihood values for initial, true, and estimated MAP configurations were -19747072 , -75385 , and -121663 , respectively. In Figure 4, we can see that the estimated MAP configuration is not far away from the true ordering, since all the markers are placed extremely close to their true positions. In fact, many markers are placed correctly while some relatively short segments are placed in their right positions but in the reverse direction.

The parental phase was recovered without errors, and in Figure 5 we plot the genetic distances estimated without any prior knowledge on the ordering of the markers, compared with the true map and the map obtained by using the full data and knowing the true ordering of the markers. These estimates are comparable with the corresponding estimates obtained in the simulation

experiment I where the pooled data was used and ordering of the markers was known *a priori*. It seems that when the number of pools is large enough the marker order can be recovered so well that the two pooling estimators of the recombination fraction (with and without prior knowledge of the ordering) behave very similarly.

(Figures 5 and 6 will be placed here)

Discussion

We have presented the method to estimate parental linkage phase, sex-specific recombination fractions and ordering of markers using pooled haploid DNA available from sperm, egg, or megagametophyte samples. The presented method has lot of promises because potential accuracy provided by the method has not been available without individual genotyping before. Use of single pool seems to result some information gaps along the chromosome where the linkage phase (see Figure 2) and recombination fractions (see Figure 3) cannot be determined. However, these gaps can be effectively avoided by using multiple (2 or more) pools simultaneously because random gap positions are likely to differ between pools.

The parental linkage phase and recombination frequency estimation in this method is based on random fluctuations in transmission ratios and their correlation between the loci under Mendelian inheritance. The information content in the pooled data seems to depend on how much the observed frequency deviates from its expectation. To maximize the information content we propose the following strategy which requires individual genotyping at the first locus. One could classify individuals into two sets of divergent gametic pools according to the typing of allele at the first locus. These two pools would then be used to measure pooled frequencies from other locus. This would enhance (locally) the information of the pooled frequencies, without need of any corrections.

We comment here shortly about segregation distortion. Segregation distortion may follow as result of a selection process on the gametes produced in the meiosis. Since we are considering a

short chromosomal segment, we may assume that the selection probability depends only on the alleles at a single locus. Having assumed that the recombination process occurs independently from the haplotypes of parent, it follows that the distribution of the recombination pattern is not affected by the selection process. In that case we can apply our method without changes. In fact, this kind of segregation distortion would produce data which on average is more informative around the selective locus than in the neutral case. Note however that in case the selection probability depends on several loci, the distribution of the recombination pattern on the selected gametes may be perturbed in a way depending on the alleles, violating our assumptions. Because this method is based on typing gametic samples (of sperm or egg) rather than living progeny, it can be used to estimate offspring ratios and sex-specific recombination frequencies in meiotic state under transmission equilibrium. These estimates can then be compared to similar estimates obtained from living organisms (i.e., under postmeiotic state). Statistically significant departure in estimates between these two states can then be taken as evidence in favor of postmeiotic selection as a source of transmission distortion. For other methods to determine origin of transmission ratio distortion, see de Villena et al. (2000).

We hope that the proposal distributions in the McMC algorithm could be improved in the future to achieve a faster relaxation. One possibility to be explored would be to use particle filters (say genetic algorithm) techniques, which can be used successfully for the traveling salesman problem (see Del Moral 2004). The linkage phase estimator is robust to errors and estimator for genetic distance is robust to DNA amplification errors although the sources of errors in the allele frequency measurements from gametic pools are not taken into account in the current model. We are currently investigating the possible further extension of the method using errors-in-variable modeling (Fuller 1987; Biemer et al. 1991; Carroll et al. 1995). In any case, the measurement errors can be controlled in the design of the pooling experiment - by taking several pools with pool size small enough. Of course, we are looking forward to test our method with real data when available. The software implementing the method is freely available for research purposes from the authors.

Acknowledgements We are grateful to two anonymous referees for their constructive comments on the manuscript. This work was supported by research grant (202324) from the Academy of Finland and by the Centre of Population Genetic Analyses, University of Oulu, Finland.

References

- Arnheim, N., P. Calabrese, and M. Nordborg, 2003 Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *Am. J. Hum. Genet.* **73**: 5-16.
- Biemer, P. P., R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman, 1991 *Measurement Errors in Surveys*. New York: John Wiley & Sons.
- Boehnke, M., K. Lange, and D. Cox, 1991 Statistical methods for multipoint radiation hybrid mapping. *Am. J. Hum. Genet.* **49**: 1174-1188.
- Bremaud, P., 2002 *Mathematical Principles of Signal Processing: Fourier and Wavelet Analysis*. New York: Springer-Verlag.
- Butcher, P. A., E. R. Williams, D. Whitaker, S. Ling, T. P. Speed, and G. F. Moran, 2002 Improving linkage analysis in outcrossed forest trees - an example from *Acacia mangium*. *Theor. Appl. Genet.* **104**: 1185-1191.
- Butcher, L. M., E. Meaburn, L. Liu, C. Fernandez, L. Hill, A. AL-Chalabi, R. Plomin, L. Schalkwyk, and I. W. Craig, 2004 Genotyping pooled DNA on microarrays: a systematic genome screen of thousands of SNPs in large samples to detect QTLs for complex traits. *Behav. Genet.* **34**: 549-555.
- Carroll, R. J., D. Ruppert, and L. A. Stefanski, 1995 *Non-Linear Measurement Error Models*. London: Chapman & Hall.
- Clark, A. G., 1990 Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7**: 111-122.

Collins, H. E., H. Li, S. E. Inda, J. Anderson, K. Laiho, J. Toumilehto, and M. F. Seldin, 2000 A simple and accurate method for determination of microsatellite total allele content differences between DNA pools. *Hum. Genet.* **106**: 218-226.

Cullen, M., S.P. Perfetto, W. Klitz, G. Nelson, and M. Garrington, 2002 High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am. J. Hum. Genet.* **71**: 759-776.

Darvasi, A., 1998 Experimental strategies for the genetic dissection of complex traits in animal models. *Nat. Genet.* **18**: 19-24.

Del Moral, P., 2004 *Feynman-Kac formulae: Genealogical and Interacting Particle Systems with Applications*. Springer.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**: 1-38.

de Villena, F. P-M., E. de la Casa-Esperon, T. L. Briscoe, and C. Sapienza, 2000 A general test to determine the origin of maternal transmission ratio distortion: meiotic drive at the mouse Om Locus. *Genetics* **154**: 333-342.

Durbin, R., Eddy, S., Krogh, A., and G. Mitchison, 1998 *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge.

Excoffier, L., and M. Slatkin, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**: 921-927.

Fuller, W. A., 1987 *Measurement Error Models*. New York: John Wiley & Sons.

Gao, G., I. Hoeschele, P. Sorensen, and F. Du, 2004 Conditional probability methods for haplotyping in pedigrees. *Genetics* **167**: 2055-2065.

George, A. W., 2005 A novel MCMC approach for constructing accurate meiotic maps. *Genetics* (in press).

- George, A. W., K. L. Mengersen, and G. P. Davis, 1999 A Bayesian approach to ordering gene markers. *Biometrics* **55**: 419-429.
- Geyer, C. J., and E. A. Thompson, 1992 Constrained Monte-Carlo maximum-likelihood for dependent data. *J. R. Stat. Soc. B* **54**: 657-699.
- Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97-109.
- Hawley, M. E., and K. K. Kidd, 1995 HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. *J. Hered.* **86**: 409-411.
- Huang, Q., S. Shete, and C. I. Amos, 2004 Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis. *Am. J. Hum. Genet.* **75**: 1106-1112.
- Ito, T., S. Chiku, E. Inoue, M. Tomita, T. Morisaki, H. Morisaki, and N. Kamatani, 2003 Estimation of haplotype frequencies, linkage-disequilibrium measures, and combination of haplotype copies in each pool by use of pooled DNA data. *Am. J. Hum. Genet.* **72**: 384-398.
- Jansen, J., A. G. de Jong, and J. W. van Ooijen, 2001 Constructing dense genetic linkage maps. *Theor. Appl. Genet.* **102**: 1113-1122.
- Kitamura, Y., M. Moriguchi, H. Kaneko, H. Morisaki, T. Morisaki, K. Toyama, and N. Kamatani, 2002 Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm. *Ann. Hum. Genet.* **66**: 183-193.
- Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S. T. Palsson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, and K. Stefansson, 2002 A high-resolution recombination map of the human genome. *Nat. Genet.* **31**: 241-247.

- Kruglyak, L., M. J. Daly, and E. S. Lander, 1995 Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am. J. Hum. Genet.* **56**: 519-527.
- Kruglyak, L., M. J. Daly, M. P. Reeve-Daly, and E. S. Lander, 1996 Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**: 1347-1363.
- Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. *Ann. Math. Stat.* **22**: 79-86.
- Lalam, N., C. Jacob, and P. Jagers, 2004 Modelling the PCR amplification process by a size-dependent branching process and estimation of the efficiency. *Adv. Appl. Prob.* **36**: 602-615.
- Lander, E. S., and P. Green, 1987 Construction of multilocus genetic maps in humans. *Proc. Natl. Acad. Sci. USA* **84**: 22363-22367.
- Lazzeroni, L. C., N. Arnheim, K. Schmitt, and K. Lange, 1994 Multipoint mapping calculations for sperm-typing data. *Am. J. Hum. Genet.* **55**: 431-436.
- Lin, S., and T. P. Speed, 1997. An algorithm for haplotype analysis. *J. Comput. Biol.* **4**: 535-546.
- Long, J. C., R. C. Williams, and M. Urbanek, 1995 An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**: 799-810.
- Lu, Q., Y. Cui, and R. Wu, 2004 A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. *BMC Genetics* **5**: 20.
- McPeck, M. S., 1999 SPERMSEG: analysis of segregation distortion in single-sperm data. *Am. J. Hum. Genet.* **65**: 1195-1197.
- McPeck, M. S., and A. Strahs, 1999 Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine scale genetic mapping. *Am. J. Hum. Genet.* **65**: 858-875.
- Navidi, W., and N. Arnheim, 1994 Analysis of genetic data from the polymerase chain reaction. *Stat. Sci.* **9**: 320-333.

- Navidi, W., and N. Arnheim, 1999 Combining data from polymerase chain reaction DNA typing experiments: application to sperm typing data. *J. Am. Stat. Assoc.* **94**: 726-733.
- Norton, N., N. M. Williams, H. J. Williams, G. Spurlock, G. Kirov, D. W. Morris, B. Hoogenboom, M. J. Owen, M. C. O'Donovan, 2002 Universal, robust, highly quantitative SNP allele frequency measurement in DNA pools. *Hum. Genet.* **110**: 471-478.
- Norton, N., N. M. Williams, M. C. O'Donovan, and M. J. Owen, 2004 DNA pooling as a tool for large-scale association studies in complex traits. *Ann. Med.* **36**: 146-152.
- Penttinen, A., 1984 *Modelling Interactions in Spatial Point Patterns: Parameter Estimation by the Maximum Likelihood Method*. Jyväskylä studies in Computer Science, Economics and Statistics 7. University of Jyväskylä.
- Qian, D., and L. Beckmann, 2002 Minimum-recombinant haplotyping in pedigrees. *Am. J. Hum. Genet.* **70**: 1434-1445.
- Rabiner, L., 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**: 257-286.
- Ritland, K., 2002 Estimation of gene frequency and heterozygosity from pooled samples. *Mol. Ecol. Notes* **2**: 370-372.
- Rosa, G. J. M., B. S. Yandell, and D. Gianola, 2002 A Bayesian approach for constructing genetic maps when markers are miscoded. *Gen. Sel. Evol.* **34**: 353-369.
- Schaid, D. J., S. K. McDonnell, L. Wang, J. M. Cunningham, S. N. Thibodeau, 2002 Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am. J. Hum. Genet.* **71**: 992-995.
- Sham, P., J. S. Bader, I. Craig, M. O'Donovan, M. Owen, 2002 DNA pooling: a tool for large-scale association studies. *Nat. Rev. Genet.* **3**: 862-871.

Shaw, S. H., M. M. Carrasquillo, C. Kashuk, E. G. Puffenberger, and A. Chakravarti, 1998 Allele frequency distribution in pooled DNA samples: applications to mapping complex disease genes. *Genome Res.* **8**: 111-123.

Slonim, D., L. Kruglyak, L. Stein, and E. Lander, 1997 Building human genome maps with radiation hybrids. *J. Comput. Biol.* **4**: 487-504.

Sobel, E., K. Lange, J. R. O'Connell, and D. E. Weeks, 1996 Haplotyping algorithms. In *Genetic Mapping and DNA Sequencing*, IMA Volume 81 in Mathematics and its applications, TP Speed & MS Waterman, eds. Springer-Verlag, New York, pp. 89-110.

Sobel, E., and K. Lange, 1996 Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *Am. J. Hum. Genet.* **58**: 1323-1337.

Stephens, D. A. and A. F. M. Smith, 1993 Bayesian inference in multipoint gene mapping. *Ann. Hum. Genet.* **57**: 65-82.

Stephens, M., and P. Scheet, 2005 Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76**: 449-462.

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

Sillanpää, M. J., and E. Arjas, 1999 Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**: 1605-1619.

Tapadar, P., S. Ghosh, P. P. Majumder, 2000 Haplotyping in pedigrees via a genetic algorithm. *Hum. Hered.* **50**: 43-56.

The International HapMap Consortium, 2003 The international HapMap project. *Nature* **426**: 789-796.

Tulsieram, L. K., J. C. Glaubitz, G. Kiss, and J. E. Carlson, 1992 Single tree genetic linkage mapping in conifers using haploid DNA from megagametophytes. *Bio/Technology* **10**: 686-690.

- van der Vaart, A. W., 1998 *Asymptotic Statistics*. Cambridge: Cambridge University Press.
- Weber, J. L., 2002 The Iceland map. *Nat. Genet.* **31**: 225-226.
- Wijsman, E. M., 1987 A deductive method of haplotype analysis in pedigrees. *Am. J. Hum. Genet.* **41**: 356-373.
- Wu, R., C-X. Ma, S. S. Wu, and Z-B. Zeng, 2002 Linkage mapping of sex-specific differences. *Genet. Res.* **79**: 85-96.
- Yang, H-C., C-C. Pan, R. C. Y. Lu, and C. S. J. Fann, 2005 New adjustment factors and sample size calculation in DNA-pooling experiment with preferential amplification. *Genetics* **169**: 399-410.
- Yang, Y., H. Zhang, J. Hoh, F. Matsuda, P. Xu, M. Lantrop, and J. Ott, 2003 Efficiency of single-nucleotide polymorphism haplotype estimation from pooled DNA. *Proc. Natl. Acad. Sci. USA* **100**: 7225-7230.
- Yazdani, R. F., C. Yeh, and J. Rimsha, 1995 Genomic mapping of *Pinus sylvestris* (L.) using random amplified polymorphic DNA markers. *For. Genet.* **4**: 209-215.

APPENDIX A: VITERBI ALGORITHM

Here we describe the Viterbi algorithm for the hidden Markov model $\{(X_l, N_l) : l = 1, \dots, L\}$, where we assume that the recombination fractions $(\theta_l : l = 2, \dots, L)$ are known in advance. The problem is to sample the parental linkage phase (X_1, \dots, X_L) conditionally on the observed allele frequencies (N_1, \dots, N_L) which are obtained from the pool of gametes. Because we are on a discrete state space we can use standard setting; see Durbin et al. (1998) and Rabiner (1989).

Nothing is computed for locus $l = 1$, because linkage phase X_1 and observed frequency N_1 are independent there. But for all subsequent loci $l = 2, \dots, L$, the (conditional) transition probabilities $p(X_l = x | X_{l-1} = y, N_{l-1}, N_l)$ are recursively computed for all possible values of (x, y) . The term X_{l-1} is then integrated out from this expression so that we obtain

$$p(X_l = x | N_1, N_2, \dots, N_{l-1}, N_l) = \sum_{y=0,1} p(X_l = x | X_{l-1} = y, N_{l-1}, N_l) p(X_{l-1} = y | N_1, N_2, \dots, N_{l-2}, N_{l-1}).$$

After these L forward steps, linkage phase X_L is sampled from $p(X_L | N_1, N_2, \dots, N_{L-1}, N_L)$ and the backward algorithm continues for $l = L - 1, \dots, 1$ as follows:

Linkage phase X_l is drawn given the data and X_{l+1} with probability proportional in x to

$$p(X_l = x | N_1, N_2, \dots, N_{l-1}, N_l) p(X_{l+1} | X_l = x, N_l, N_{l+1}).$$

After L backward steps we obtain the sample (X_1, \dots, X_L) from the posterior distribution.

By dynamic programming it is also possible to compute *a posteriori* the most probable path, and its posterior probability.

APPENDIX B: UPDATE STEPS AND PROPOSAL DISTRIBUTIONS FOR THE METROPOLIS ALGORITHM

A McMC cycle consists of several proposed moves.

i) A block update for the parental phase vector (X_1, \dots, X_L) which samples from the fully conditional posterior distribution given the data and the current values of π and $(\theta_1, \dots, \theta_L)$. If there is no prior information about the phase vector this is very simple since under the uniform prior for X , $\Delta X_{\pi(l)}$ will be conditionally independent given π . If the prior for X is not uniform, we use the Viterbi algorithm (see Appendix A).

ii) We update independently the recombination fraction parameters θ_l conditionally on the data, π and on the phases $X_{\pi(l-1)}, X_{\pi(l)}$. This is a Metropolis move where as proposal distribution for θ_l we take an histogram approximation over a finite grid $G \subset [0, 1/2]$ of the full conditional density $p(\theta_l | N_{\pi(l-1)}^{(k)}, N_{\pi(l)}^{(k)}, k = 1, \dots, n_{pools}, X_{\pi(l-1)}, X_{\pi(l)})$.

The two following moves update the ordering. For each of these, we resample simultaneously the parental phases and the recombination fraction parameters around the markers involved.

iii) As a proposal distribution, we apply a random permutation to the current ordering. For each loci involved, we resample jointly the phase and the recombination fractions between the new neighbouring markers (that is, by using the proposal distribution described in ii) given the new ordering and phase).

iv) As a proposal distribution we randomly select a segment, we cut it off from the chromosome, join the extremes, choose a random location in the chromosome and insert there the segment, eventually also inverting the direction of the segment. We also have a possibility to flip simultaneously all the parental phases of the loci belonging to the segment. Finally new recombination fraction parameters are sampled for the markers around the two cuts points by using the proposal distribution described in ii).

The random permutation in move iii) and the random segment in move iv) together with the new phases are not selected uniformly at random. Instead in the proposal distribution we take in account the pseudo-loglikelihood (Equation 1).

For each of these move, for a given starting configuration (π, X) we denote by $F(\pi, X)$ the set of configurations which are reachable in one step. By using the pseudo-loglikelihood we define

the proposal distribution as follows:

$$q_t((\pi, X) \rightarrow (\pi', X')) = \begin{cases} Z_t(\pi, X) \exp(-t\Psi(\pi', X')) & \text{if } (\pi', X') \in F(\pi, X) \\ 0 & \text{otherwise,} \end{cases}$$

where $t \geq 0$ is an inverse temperature parameter $t = 0$ corresponding to the uniform distribution over $F(\pi, X)$, and $Z_t(\pi, X)$ is the normalizing constant. For high values of t , the proposal distribution choose among the reachable configurations will choose with high probability those with the smallest energy. Since we want to have also a chance to propose any reachable configuration, we let the inverse temperature parameter t vary cyclically in a finite set.

The Hastings ratio for such moves is given by

$$\frac{\exp(-C_0(\pi'(1))) \pi(x'_{\pi(1)}) \prod_{l=2}^L \frac{\pi_l(x'_{\pi(l)})}{\pi_l(x_{\pi(l)})} \frac{L_{\theta'_l}(\pi'(l-1), x'_{\pi'(l-1)}, \pi'(l), x'_{\pi'(l)}; \text{data})}{L_{\theta_l}(\pi(l-1), x_{\pi(l-1)}, \pi(l), x_{\pi(l)}; \text{data})}}{\exp(-C_0(\pi(1))) \pi(x_{\pi(1)}) \prod_{l=2}^L \frac{\pi_l(x_{\pi(l)})}{\pi_l(x'_{\pi(l)})} \frac{L_{\theta_l}(\pi(l-1), x_{\pi(l-1)}, \pi(l), x_{\pi(l)}; \text{data})}{L_{\theta'_l}(\pi'(l-1), x'_{\pi'(l-1)}, \pi'(l), x'_{\pi'(l)}; \text{data})}} \frac{q_t((\pi', x') \rightarrow (\pi, X))}{q_t((\pi, x) \rightarrow (\pi', x'))} \\ \times \prod_{l=2}^L \frac{\tilde{q}(\theta_l | \pi, x, \text{data})}{\tilde{q}(\theta'_l | \pi', x', \text{data})}.$$

Remark. In case the data contain gametic pools sampled from J different parental individuals, it is clear that one should extend the MCMC algorithm by keeping one permutation vector (π_1, \dots, π_L) and introducing J independent parental phase vectors $(X_1^{(j)}, \dots, X_L^{(j)})$, $j = 1, \dots, J$.

Figure legends

FIGURE 1. Graphical representation of the hidden state structure. At locus L , hidden parental linkage phase is indicated as X_L and number of offspring having 1-allele as N_L .

FIGURE 2. Observed information about the parental linkage phase in pooled haplotype data. Pooled data consist of simulated haploid offspring group with 100 haplotypes (gametes). Allele frequency of the allele with lower frequency (top panel) and corresponding observed information content (bottom panel) are shown for 200 heterozygote loci. The limit of maximal information in $\log(2)$ is indicated with broken line in the bottom panel. A genetic map distances (in Morgans) are also shown on the x-axis.

FIGURE 3. Simulation Experiment I. The curve of estimated genetic distance over marker loci (measured cumulatively from the left) in case of different pooling strategies: (a) 1 pool of 1000 gametes, (b) 2 pools of 500 gametes, (c) 4 pools of 250 gametes, (d) 8 pools of 125 gametes, (e) full data of 1000 individual gametes, (f) true genetic distance. Recombination fractions were converted to genetic distance using Haldane's mapping function.

FIGURE 4. Simulation Experiment I - Erroneous Data. The curve of estimated genetic distance over marker loci (measured cumulatively from the left) in case of different degree of errors in the data. The curves (d),(e), and (f) coincide with those in Figure 3: (f) true genetic distances, (e) estimator based on 1000 individual gametes (correct data), (d) estimator based on 8 pools of 125 gametes in each (correct data). In curves (g)-(k), data set with 8 pools of 125 gametes are also used but with errors in data: (g) Gaussian random error with mean 0 and standard deviation $\sqrt{n}\sigma$ where $\sigma = 0.01$, (h) DNA amplification error with $\epsilon = 0.1$, (i) DNA amplification error with $\epsilon = 0.2$, (j) DNA amplification error with $\epsilon = 0.5$, and (k) DNA amplification error with $\epsilon = 1.0$. Recombination fractions were converted to genetic distance using Haldane's mapping

function.

FIGURE 5. Simulation Experiment II. Estimated order of 400 marker loci using pooled data from 500 pools, each containing 200 gametes.

FIGURE 6. Simulation Experiment II. Genetic distance curves over marker loci (measured cumulatively from the left): (a) true genetic distances, (b) estimated genetic distance using the full data with known marker ordering, (c) the genetic distance is estimated simultaneously with the ordering of the markers by using pooled data.

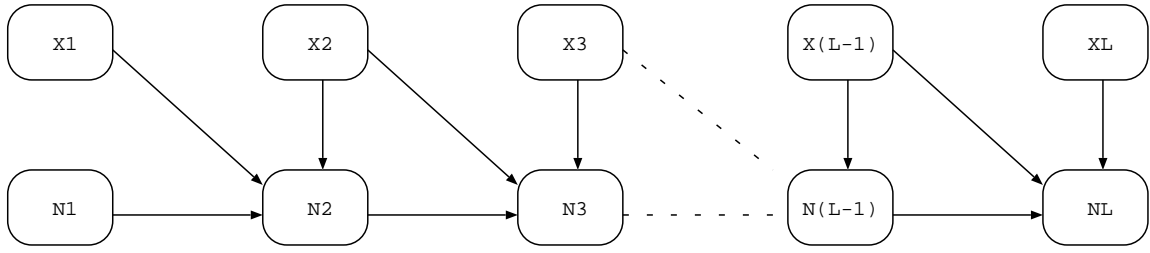


Figure 1:

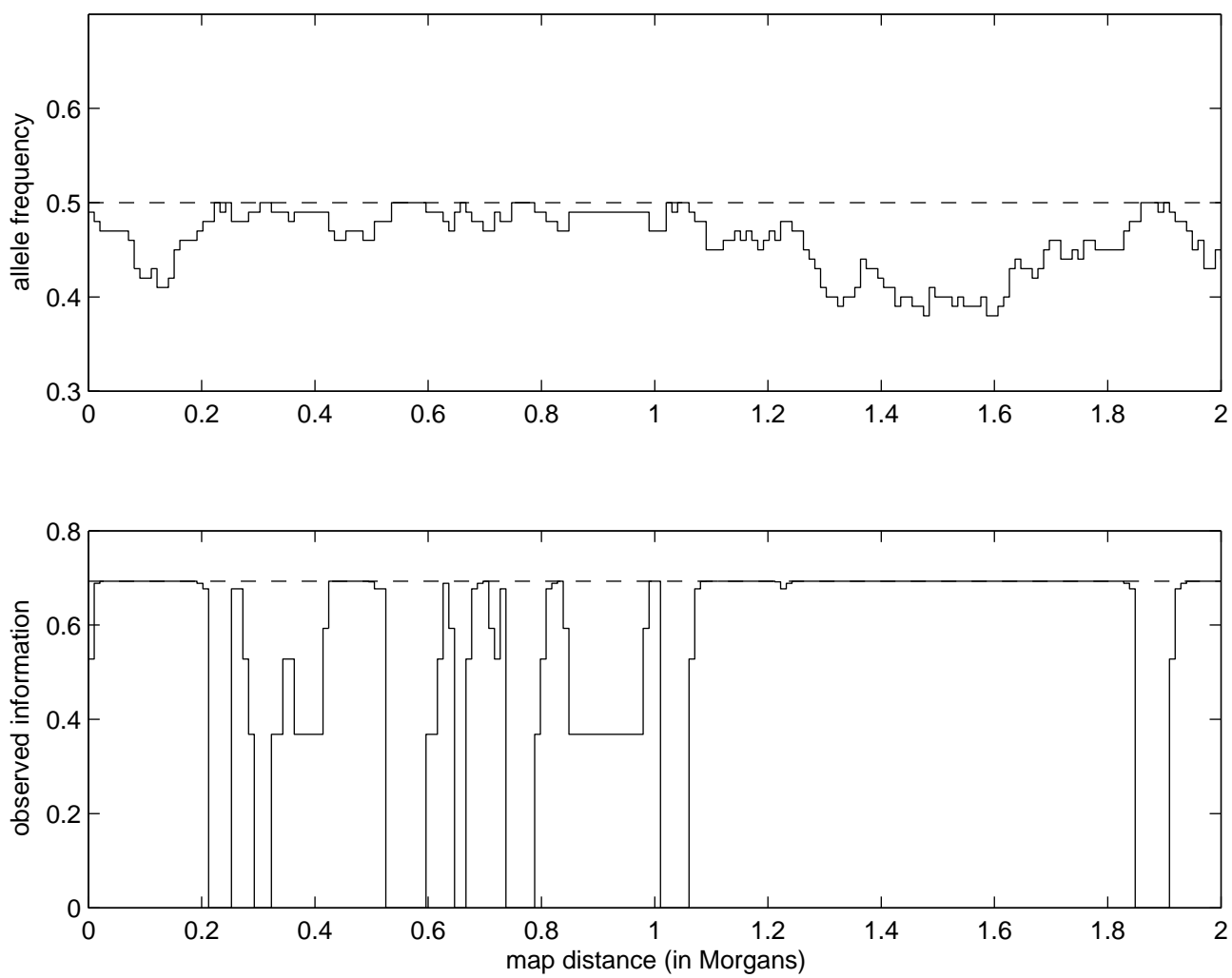


Figure 2:

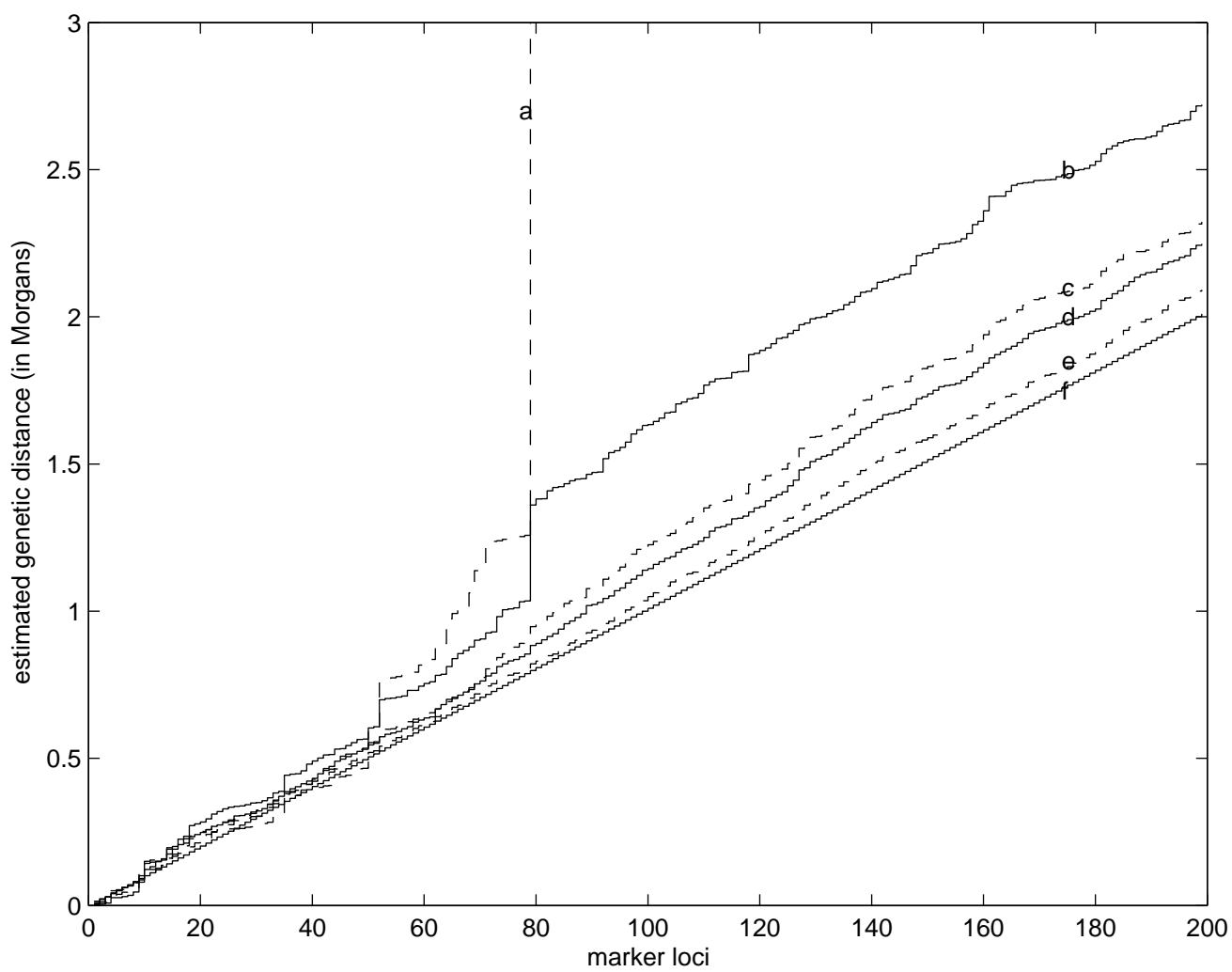


Figure 3:

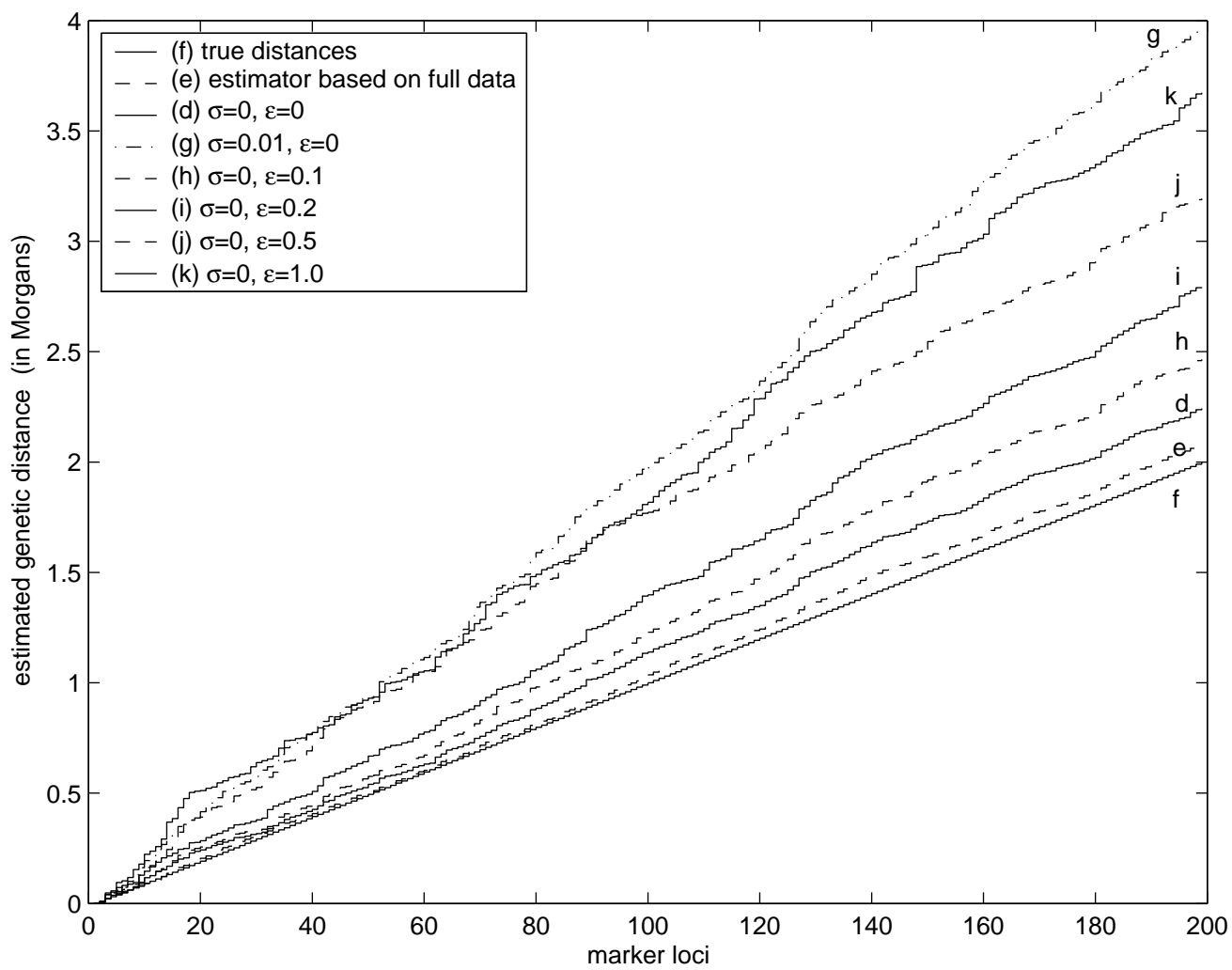


Figure 4:

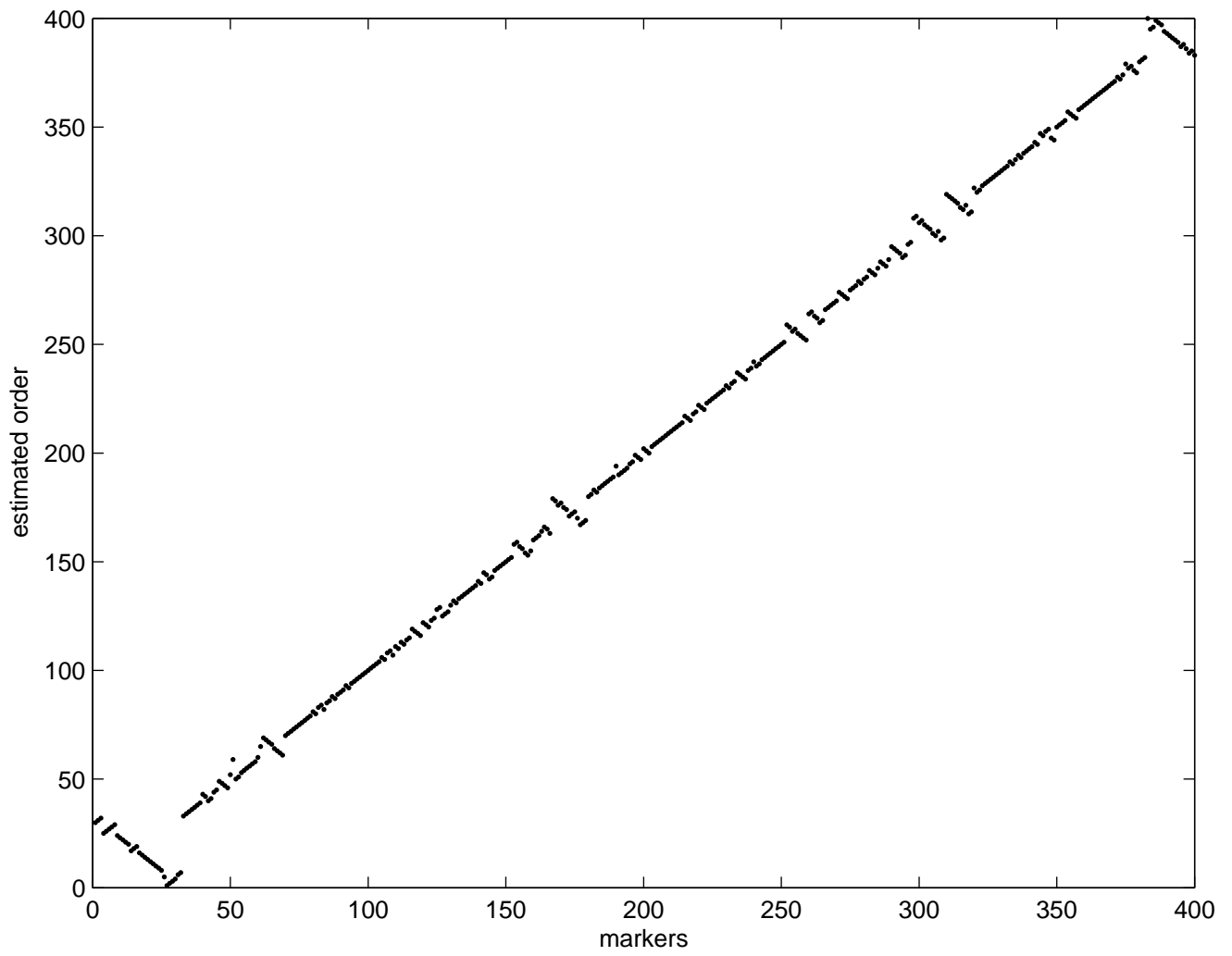


Figure 5:

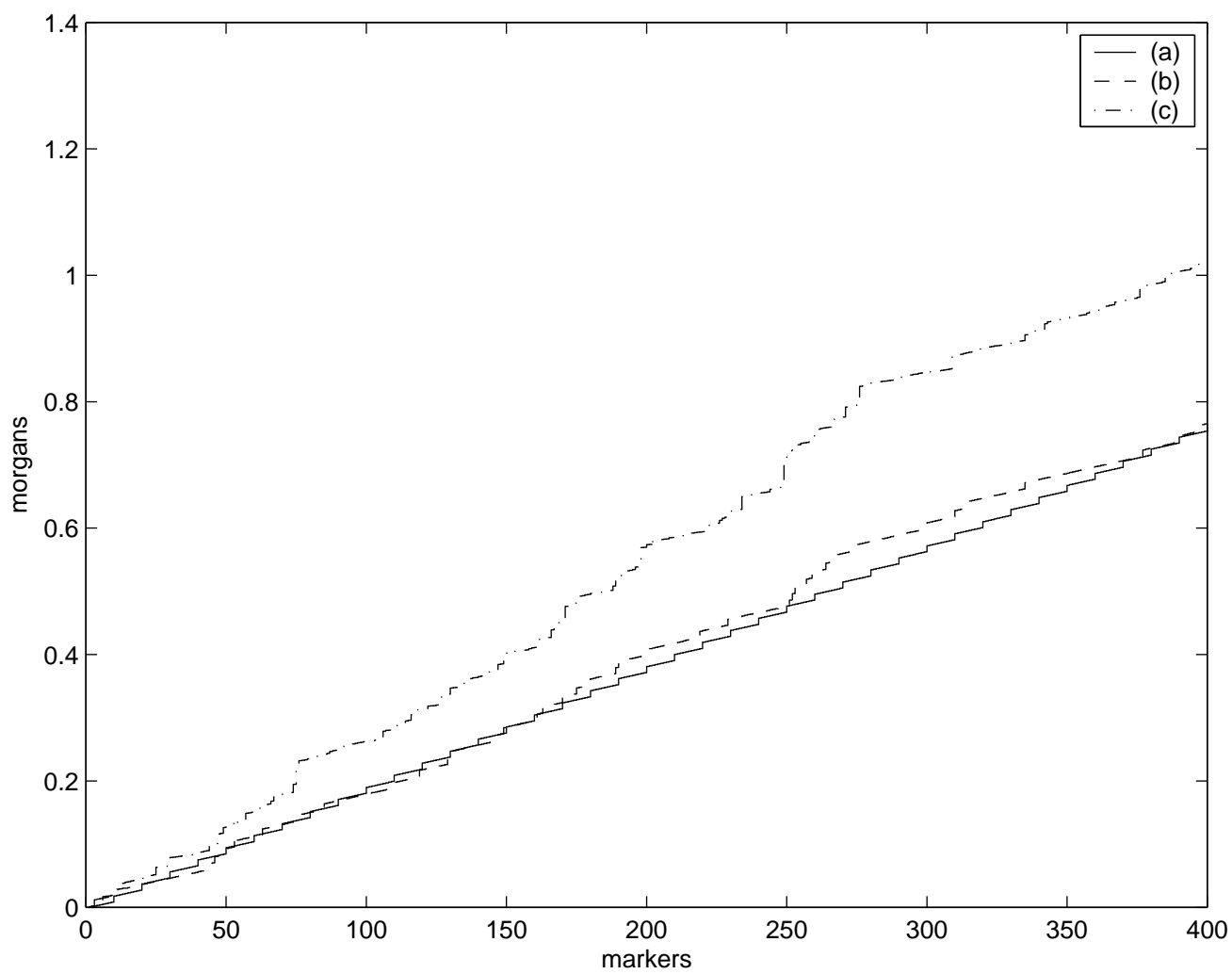


Figure 6: