Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 67 (2005) 75–83
Theoretical Population Biology
www.elsevier.com/locate/ytpbi

Backward simulation of ancestors of sampled individuals

Dario Gasbarra, Mikko J. Sillanpää*, Elja Arjas

Department of Mathematics and Statistics, Rolf Nevanlinna Institute, University of Helsinki, P.O. Box 68, FIN-00014 Helsinki, Finland

Received 27 June 2003
Available online 7 January 2005

Abstract

If the population is large and the sampling mechanism is random, the coalescent is commonly used to model the haplotypes in the sample. Ordered genotypes can then be formed by random matching of the derived haplotypes. However, this approach is not realistic when (1) there is departure from random mating (e.g., dominant individuals in breeding populations or monogamy in humans), or (2) the population is small and/or the individuals in the sample are ascertained by applying some particular non-random sampling scheme, as is usually the case when considering the statistical modeling and analysis of pedigree data. For such situations, we present here a data generation method where an ancestral graph with non-overlapping generations is first generated backwards in time, using ideas from coalescent theory. Alleles are randomly assigned to the founders, and subsequently the gene flow over the entire genome is simulated forwards in time by dropping alleles down the graph according to recombination model without interference. The parameters controlling the mating behavior of generated individuals in the graph (degree of monogamy) can be tuned in order to match a particular demographic situation, without restriction to simple random mating.

The performance of the approach is illustrated with a simulation example. The software (written in C-language) is freely available for research purposes at http://www.rni.helsinki.fi/~dag/.

Keywords: Coalescent; Chromosome simulation; Complex pedigrees; Dirichlet distribution; Pólya urn

1. Introduction

Simulation methods are generally used to answer questions which are not analytically tractable. Traditional population simulation programs create pedigrees (generations) forwards in time, starting from the founders of the population (Terwilliger et al., 1993; Hambe et al., 1998; Ollikainen, 2002). In contrast, simulation programs based on coalescent theory (Kingman, 1982; Hudson, 1993; Griffiths and Marjoram, 1996; Fu and Li, 1999; Nordborg and Tavaré, 2002; Möhle and Sagitov, 2003) start from a sample belonging to the youngest generation and work backwards in time.

Also the application areas of these methods are different. Forward algorithms (with recombinations) are used in population genetics typically over tens, or perhaps hundreds, of generations. The following types of tasks can then be considered: (1) computation of *P*-values for different statistics (e.g., length of shared segments) (Churchill and Doerge, 1994; Davies et al., 1995; van der Meulen and te Meerman, 1997; Zhao et al., 1999; Lange and Lange, 2004), (2) utilization in power calculations (Long and Langley, 1999) and design of experiments (including controlled mating schemes, see Luo et al., 2002), (3) study of population genetic forces or events (selection, migration, genetic drift, linkage disequilibrium, or segregation distortion), (4) study of population substructures (Pfaff et al., 2001), (5) study of population genetic parameters (e.g., effective population size) or population genetic summary statistics (e.g. F_{ST}) under various demographic scenarios (Strand, 2002), (6) study of influences of erroneous assumptions and model misspecification on genetic analysis (e.g., by omitting cross-over interference or sex-specific recombination fractions), (7) estimation of

*Corresponding author. Fax: +358 9 191 51400.
E-mail address: mjs@rolf.helsinki.fi (M.J. Sillanpää).

0040-5809/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2004.08.003

identity-by-descent distribution for individuals with a known pedigree, known as "gene dropping" (Maccluer et al., 1986; Meuwissen and Goddard, 2000), (8) utilization of simulations in teaching of genetics (Soderberg and Price, 2003), and (9) creation of simulated data sets for method comparison and testing.

Coalescent algorithms, with mutations as the driving mechanism, are mainly used in phylogenetics and in the study of population genetic parameters, such as effective population size and mutation frequency, but they have been applied also for studying linkage disequilibrium (Zöllner and von Haeseler, 2000) and for gene mapping (Larribe et al., 2002; Morris et al., 2002). The typical object of study is then the evolution of a narrow DNA segment in a time-span of perhaps tens of thousands of generations. Coalescent algorithms usually assume random mating but generalized methods allowing departure from random mating have also been considered (Cannings, 1975; Möhle and Sagitov, 2001). A nice feature in backward simulation is that the constructed pedigree contains only the ancestors of the individuals in the study sample. Branches which have become extinct or which have offspring only outside the study sample are not generated at all. This automatic "cut-off" property allows fast simulations.

Since coalescent methods like ancestral recombination graph (Hudson, 1993; Griffiths and Marjoram, 1996) are usually concerned with a narrow DNA segment followed during a large number of generations, it is reasonable to expect only a few recombination/mutation events. As we are interested in gene mapping in humans, we operate at the opposite end of the scale, considering either whole chromosomes with possibly multiple genetic events in one generation, or smaller regions over a time span of one hundred generations at most. Here we present a two-stage simulation method where an ancestral graph with non-overlapping generations is first generated backwards in time, using ideas from coalescent theory. Alleles are assigned randomly to the founders, and subsequently the gene flow over the entire genome or given segments is simulated forwards in time by dropping alleles down the graph according to a recombination model without interference. This simulator can be used for similar purposes as forward simulators, but it is especially useful in simulating ancestors of sampled individuals in the following situations: (1) Simulating an ancestral graph, or a set of extended families, in a large population. (2) When there is non-random mating in the population. (If the population size is moderate to large, the savings compared to simulating the entire population can be large in case reproduction in the population is mainly through a few dominant individuals, making the effective population size small.). (3) Simulating a sample from admixed populations or under an island model. (4) When the considered chromosomal segment is small (in this case the algorithm is similar to the ancestral recombination graph but on a different time/population scale).

Once a pedigree has been generated it is possible to include phenotypes by sampling their values from a given genetic model in which some markers are treated as genes. Here genetic effects can be attached to alleles or genotypes. An alternative way of modeling or simulating phenotypes is to attach genetic effects to alleles identical by descent (with respect to the founding generation) given the pedigree and the allelic path configuration. This corresponds to a situation in which all founder alleles in a considered locus are different and have their individual effects. Note that this differs from the approaches of Podlich and Cooper (1998) and Chung et al. (2003), who used genetic-model-based simulation; see also Gauderman (1995). In order to illustrate the present simulation method, we provide an example where the expected average haplotype sharing around each locus is estimated for a given set of loci. This is done for an ascertained subsample from the youngest generation, consisting of individuals whose continuous phenotype value exceeds a given threshold. The expected average haplotype sharing will show a signal at the trait locus.

2. Methods

2.1. A model for the graph and the backward simulation procedure

This is a discrete time model in which reverse time t runs from 0 (the youngest generation) to T (the founder generation). Our starting point is a demographic process $(N'_t, N''_t : t = 1, \ldots, T)$, where N'_t and N''_t are the numbers of males and females in the population t generations ago. Depending on the assumed knowledge, we can either treat the demographic process as given or consider it as random with a given dynamics.

We start from a sample of n individuals belonging to the youngest generation (descendants) of the pedigree ($t = 0$), with the sample size being given by $n = n_0$. Our goal is to construct a random ancestral graph containing all the ancestors of the individuals in the sample up to the Tth generation. Note that maximally there are altogether $N'_1 \cdot N''_1$ possible pairs of parents. We shall assign to each of the n individuals in the sample a mother and a father from the females and males belonging to generation 1, by using a Pólya urn scheme (Blackwell and MacQueen, 1973). To make such a graph more realistic, parameters α and β can be tuned to control the mating behavior and to enforce a desired degree of monogamy.

2.1.1. Paternal assignment

We start by assigning inductively fathers to the children. The first father is the father of the first child. Suppose that after $k<n$ stages, the first k children have been assigned $F(k)$ fathers, with $F(k)\leqslant k$. At this stage the $F(k)$ fathers have been assigned, respectively, $C_1(k),\ldots,C_{F(k)}(k)$ children, where $C_i(k)\geqslant 1$ and $\sum_{i=1}^{F(k)} C_i(k)=k$.

Then the $(k+1)$th child can either be assigned to one of the $F(k)$ fathers already selected, or to a "new" father who was not yet assigned to any child. Under our model, these two possibilities have respective probabilities

$$\frac{\alpha+C_i(k)}{N_1'\alpha+k},\ i=1,\ldots,F(k),\text{and}\frac{\alpha(N_1'-k)}{N_1'\alpha+k}. \tag{1}$$

After n stages of the algorithm have been completed each child in the sample has been assigned to one the of $F(n)\leqslant n$ fathers, and the ith father has $C_i(n)$ children. This is equivalent to first sampling a random N_1'-dimensional probability distribution π from the Dirichlet$(\alpha,\ldots,\alpha)$ distribution and then, conditionally on π, sampling the parental assignments from the Multinomial(n,π) distribution. Here the parameter $\alpha\geqslant 0$ tunes the dependence of the parental assignments. In particular, the value $\alpha=\infty$ corresponds to the multinomial distribution, where each child chooses his or her father randomly among the N_1' candidates. Otherwise the parental choices become positively correlated, a small value of α corresponding to a mating type where only a few dominant fathers will reproduce and then have large numbers of offspring.

2.1.2. Maternal assignment

Next we assign inductively mothers to the children. The first mother is the mother of the first child. Suppose that, after $k<n$ steps, the first k children have been assigned to $M(k)\leqslant k$ mothers. From the earlier development we know already that the $(k+1)$th child has been assigned some father indexed by $I=I(k+1)\in\{1,\ldots,F(n)\}$. Denote

$C_{ij}(k)=$ number of children of the ith father who after k steps have been assigned to mother $j\leqslant M(k)$.

Then the $(k+1)$th child can either be assigned to one of the $M(k)$ mothers already selected, or he/she can find a "new" mother who was not yet assigned to any child. Under our model, this happens with respective probabilities

$$\frac{\beta+C_{Ij}(k)}{N_1''\beta+\sum_{j=1}^{M(k)}C_{Ij}(k)},\ j=1,\ldots,M(k),\quad\text{and}$$
$$\frac{\beta(N_1''-M(k))}{N_1''\beta+\sum_{j=1}^{M(k)}C_{Ij}(k)}. \tag{2}$$

The parameter $\beta\geqslant 0$ regulates the mating process, so that $\beta=\infty$ corresponds to random mating, and for small values of β the model is more likely to produce nuclear families where pair formation is permanent (individuals tend to have only one spouse).

In particular, the choice $\alpha=\beta N_1''$ corresponds to a model in which one first samples an $N_1'\cdot N_1''$-dimensional random probability vector $\pi=(\pi_{ij}:i=1,\ldots N_1', j=1,\ldots,N_1'')$ according to the Dirichlet$(\beta,\ldots,\beta)$ distribution, and then, conditionally on π, children are assigned independently to their parents by drawing pairs (i,j) from the distribution π. Under this special choice, the probability of sampling a particular parental configuration with $(C_{ij}(n);i=1,\ldots,N_1', j=1,\ldots,N_1'')$ is given by

$$\frac{\Gamma(\beta N_1'N_1'')}{\Gamma(\beta N_1'N_1''+n)}\prod_{i=1}^{N_1'}\prod_{j=1}^{N_1''}\frac{\Gamma(\beta+C_{ij}(n))}{\Gamma(\beta)C_{ij}(n)!},$$

where $\Gamma(.)$ is the gamma function. For $\beta=\infty$ this corresponds to multinomial sampling from the distribution in which all π_{ij}'s are equal.

When all children in generation $t=0$ have been assigned to parents, we set $t=1$, $n=n(t)=M_{n(t-1)}+F_{n(t-1)}$ and move to the next generation. The algorithm is continued until T generations has been completed.

Remark 1. It follows that the probability for two randomly sampled paternally transmitted haplotypes to coalesce (to be assigned to the same father) is $\frac{1}{2}(\alpha+1)/(N'\alpha+1)$. When we fix the number of fathers N' and let α go to ∞, this quantity goes to $1/(2N')$, which corresponds to random mating, while it goes to 1 as $\alpha\downarrow 0$. The corresponding *effective paternal population size* is given by $N'_{eff}=(N'\alpha+1)/(\alpha+1)$.

For two randomly sampled maternal haplotypes, we find the probability that they are assigned to the same mother by conditioning on the choice of the father, and it is given by

$$\frac{(\beta+1)}{2(N''\beta+1)}\times\frac{(\alpha+1)}{(N'\alpha+1)}+\frac{1}{2N''}\times\frac{(N'-1)\alpha}{(N'\alpha+1)}.$$

The corresponding *effective maternal population size* is given by

$$N''_{eff}=\left(\frac{(\beta+1)}{(N''\beta+1)}\times\frac{(\alpha+1)}{(N'\alpha+1)}+\frac{1}{N''}\times\frac{(N'-1)\alpha}{(N'\alpha+1)}\right)^{-1}$$

and the corresponding *total effective population size* by $N_{eff}=4N'_{eff}N''_{eff}/(N'_{eff}+N''_{eff})$.

For the special choice $\alpha=\beta N''$ we have

$$N'_{eff}=\frac{(\beta N'N''+1)}{(\beta N''+1)},\quad N''_{eff}=\frac{(\beta N'N''+1)}{(\beta N'+1)},$$
$$N_{eff}=\frac{4(\beta N'N''+1)}{(\beta(N'+N'')+2)}.$$

Remark 2. Within this framework it is possible to impose additional constraints, like avoiding mating between siblings or half-siblings. Then, as time runs backwards, a couple who have a common child cannot have a common parent. The parameters α, β can be estimated from real population data by using e.g., maximum likelihood (see Appendix).

The construction outlined here defines the distribution of the ancestral graph $\mathcal{G}$ with n_0 roots. Fig. 1 shows a pedigree consisting of 131 individuals in $T = 8$ generations with $n_0 = 8$ individuals at the root, and constrained to avoid mating between half and full siblings. The population was assumed to grow linearly according to $N'_t = N''_t = 7 \times (T - t + 1)$, $t = 1, \ldots, 8$, and the mating parameter values were chosen to be $\beta = 0.01$, and (in generation t) $\alpha = \beta N''_t$.

2.2. Sampling the meioses

All founders are assumed to exist at the very start of the population. The founder haplotypes are sampled independently from a known distribution by assuming that the founder population is in linkage equilibrium, or that founders come from a mixture of populations where there is linkage equilibrium within each. Conditionally on the ancestral graph $\mathcal{G}$, offspring haplotypes are sampled conditionally on the haplotypes of the parents, following the usual recombination model

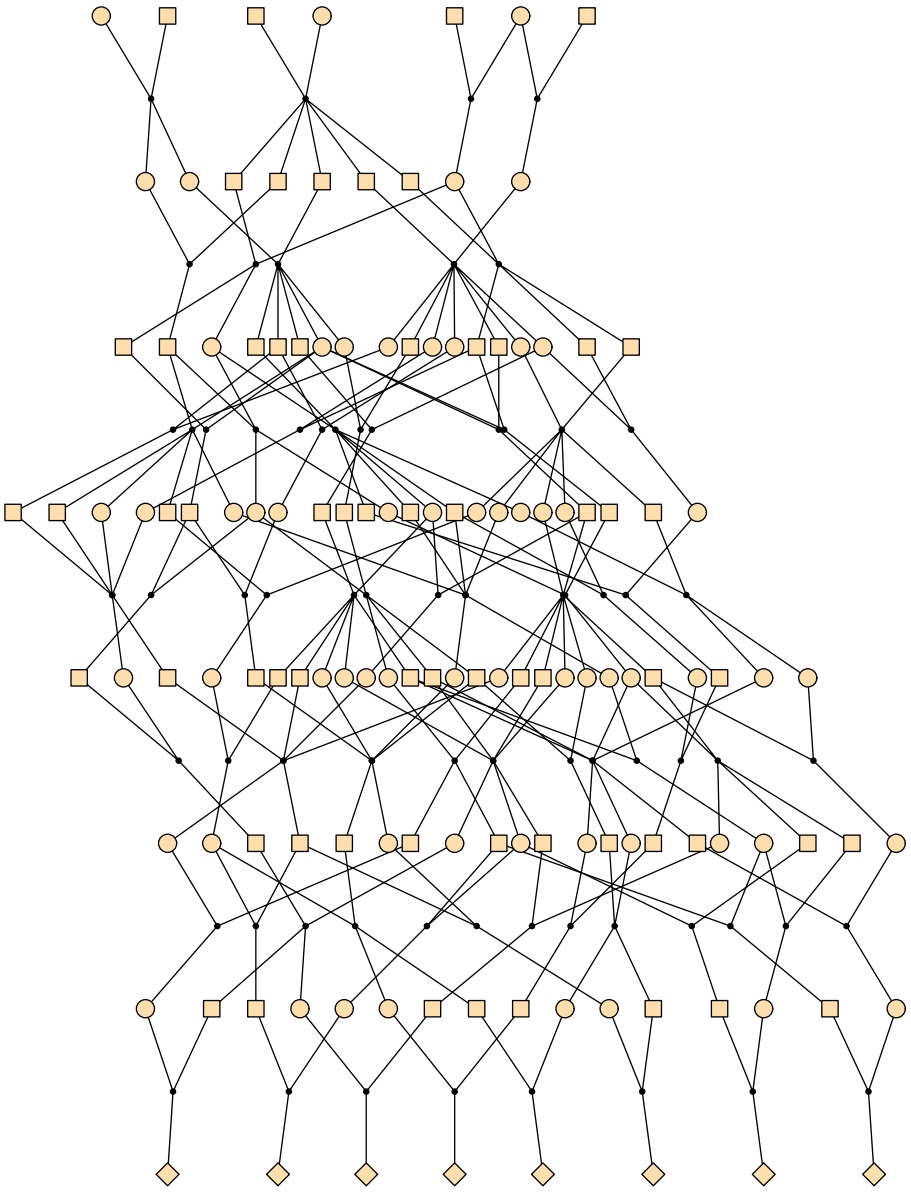

Fig. 1. Simulated pedigree. A realization of the constrained (no sib-mating) simulation with eight generations. The picture was drawn by using the software Pedfiddler freely available at http://stat.washington.edu/thompson/Genepi/pangaea.shtml.

without interference for a given genetic map (cf. Uimari and Sillanpää, 2001, Eq. (2)). This is done in all generations, down to the youngest one. To deal with population admixtures in the founding generation, we can consider a finite number of subpopulations, with different allele or genotype frequencies. Then each founder is assigned randomly to a subpopulation and its haplotypes are formed by sampling from the corresponding allele (genotype) frequencies under linkage equilibrium.

2.3. Generating only the ancestors who contributed genetic material

In a large population, the complete ancestry t generations back will often grow to an impractically large size with increasing t. On a long time span, the number of ancestors will be comparable with the population size. However, when considering a fixed set of marker loci, only a fraction of the ancestors transmit their alleles in these loci to the individuals in the sample at the present generation. This effect is particularly strong in a set of closely linked markers (within a short map distance), because recombinations are rare compared to coalescence events. In that case, the number of ancestors transmitting alleles at the markers of interest to some individual in the sample will be at most of the same order as the sample size (of present generation). In such a situation, we can speed up the simulation procedure in the following manner.

Start by assigning parents to the individuals in the study sample as described in the previous section. Then sample the meioses between marker loci for all haplotypes in the children. At each locus l, the recombination pattern on the haplotypes of the children determines whether an allele on a parental haplotype h has been transmitted to some of its children in the study sample. Accordingly, we set $A(t,h,l)=1$ or 0 to indicate whether the allele at locus l in haplotype h belonging to the tth ancestral generation was actually transmitted to an individual in the study sample or not. In case we find a haplotype h such that $A(t,h,l)=0$ for all loci l, we say that it is completely censored and we can stop tracking its genealogy. When continuing the recursive algorithm of assigning a next generation of parents, there is no need to look for the parental origin of the haplotypes which were completely censored. Once we reach the founder generation, the ancestral graph contains only the ancestors who contributed genetic material (at the marker loci) to the study sample.

The haplotypes of the founders are sampled independently, assuming linkage equilibrium, from the known allele (genotype) frequencies. This, together with the meiosis pattern which was already sampled, determines the haplotypes of all individuals in the ancestral graph, including those in the present generation. This variant of the algorithm should always be used when the population is large and the genetic distances are small.

2.4. Including a population substructure to the ancestral graph

Here we propose a simple extension of the ancestral graph construction, which is compatible with an island model.

Suppose there are J islands, each with a known demographic history, and denote by $N'_t(j), N''_t(j)$ the numbers of males and females in the jth island population t generations back. Children are allowed to choose their parents from any island, and moreover the parental choices of children belonging to different islands are assumed to be stochastically independent. Let $n(j)$ be the number of individuals in the sample who belong to the youngest generation and live in the jth island, with $n=\sum_{j=1}^{J} n(j)$.

Suppose the first child in the sample belongs to the jth island. With a probability that is proportional to $\alpha_{jh}N'_1(h)$, the first child from the jth island is assigned a generic father from the hth island. Here $\alpha_{jl}\geqslant 0$, $j,l\in\{1,\ldots,J\}$, and it is natural to assume that $\alpha_{jj}>\alpha_{jh}$ for $h\neq j$. For a general step of the induction, suppose that the first k children from the jth island have already been assigned to $F(j,k)$ fathers. For $i=1,\ldots,F(j,k)$, let $H(i)$ be the label of the island where the ith father lives, and let $C_i(j,k)$ be the current number of children of the ith father and belonging to the jth island. Then the $(k+1)$th child belonging jth island chooses the ith father with probability

$$\frac{\alpha_{jH(i)}+C_i(j,k)}{\sum_{h=1}^{J}\alpha_{jh}N'_1(h)+k}.$$

We postulate an analogous mechanism for the choice of mothers. The $(k+1)$th child from the jth island with father labeled by i chooses the lth mother belonging to the $H(l)$ island with probability

$$\frac{\beta_{jH(l)}+C_{il}(j,k)}{\sum_{h=1}^{J}\beta_{jh}N''_1(h)+C_i(j,k)},$$

where $C_{il}(j,k)$ is the current number of common children of the ith father and the lth mother and belonging to the jth island. Of the matrix (β_{jh}) we assume that $\beta_{jh}\geqslant 0$ and $\beta_{jj}>\beta_{jh}$ for $h\neq j$.

Once we have generated the ancestral graph, we can assume different allelic frequencies at the founder level for each island population.

Remark. It is natural to relate the parameters of the island model which runs backwards in time with the migration rates between the islands in a (continuous-time) forward population model. For $j\neq l$, let $\lambda_{ij}\geqslant 0$ be the rate at which one individual in the jth island emigrates to the l-island, and let $\lambda_{jj}=-\sum_{l\neq j}\lambda_{jl}\leqslant 0$.

Under this migration model, the population sizes follow the forward dynamics

$$E[N_{t+\Delta t}(j)|N_t(l), l=1,\ldots,J] \approx N_t(j)(1+\lambda_{jj}\Delta t)+\sum_{l\neq j} N_t(l)\lambda_{lj}\Delta t.$$

By taking $\Delta t = 1$, the time-unit of one generation, this means that a child from the j-island chooses a father from the same island with probability proportional to $N_t(j)(1+\lambda_{jj})$ and chooses a father from another island with index l with probability proportional to $N_t(l)\lambda_{lj}$. Therefore once a forward immigration model has been specified, it is natural to define the parameters in the island model as

$$\alpha_{jj} = (1+\lambda_{jj})\alpha, \quad \alpha_{jl} = \lambda_{lj}\ (l\neq j),$$

where $\alpha>0$ is a parameter determining the dependence of paternal assignments. We define analogously the matrix (β_{jl}) for the maternal choices.

2.5. Example: computation of expected average haplotype sharing

Since we have here excluded the possibility of mutations, at each locus an allele sampled from the youngest generation must be identical by descent (IBD) with some allele in the founder generation. Moreover, around each locus there is an entire IBD-segment which is inherited together with the considered allele and is an identical copy of the corresponding segment of the ancestor's haplotype. As the number of generations increases, the transmitted IBD-segments become shorter due to recombination which we call *erosion*.

On the other hand, if the population has been growing and one considers increasing values of T, the number of ancestors belonging to the founder generation tends to become smaller, and thereby the probability that two segments immediately linked to a given locus are IBD increases. We call this *bottleneck effect*. These effects can be studied using our simulation procedure. Namely, when alleles are transmitted from one generation to the next, for each locus and for each haplotype we only need to keep track of the boundaries of the IBD-segment immediately linked to the locus.

Considering two distinct haplotypes (i and j) and a given locus l, we say that their sharing $s_{ij}(l)$ is zero if the corresponding alleles at locus l are inherited from different founder haplotypes, otherwise $s_{ij}(l)$ is the length (in genetic distance) of the corresponding overlapping IBD-segments. The haplotype sharing statistic (HSS) at locus l computed from the sample of size n is then defined as the average sharing between all distinct haplotype pairs in the sample:

$$\mathrm{HSS}(l) = \frac{1}{n(2n-1)}\sum_{i=1}^{2n}\sum_{j<i} s_{ij}(l)$$

(cf. the statistics considered in van der Meulen and te Meerman (1997), and te Meerman and van der Meulen (1997)).

Consider now the following simple "dose effect" model for a monogenic trait: individual i will have phenotype ϕ_i according to

$$\phi_i = \beta_g 1_{\{x_{il}=Aa\}} + 2\beta_g 1_{\{x_{il}=AA\}} + \varepsilon_i,$$

where l is the gene position (trait locus) on the marker map, x_{il} is the corresponding genotype, β_g is the genetic dose effect of allele A, a refers to all other alleles except A, and ε_i is a standard normal random variable. At generation 0 we sample the haplotypes and the phenotypes of n_0 individuals, and then consider a subsample of "affected" individuals by selecting those with phenotype value ϕ larger than a given threshold τ.

We considered two examples, one with genetic effect $\beta_g = 10$ and threshold $\tau = 5$, (Fig. 2a) and one with genetic effect $\beta_g = 1$ and threshold $\tau = 1$, (Fig. 2b). In both cases we monitored 50 multiallelic marker loci, which were equally spaced at intervals of 5 centiMorgans. The trait locus was set between the 20th and the 21st marker loci.

We computed the expectation of HSS(l) as a function of l by averaging over 1000 simulation replicates. For the individuals in the selected subsample, we can see that the expected HSS curve has a peak exactly at the trait locus (Fig. 2), while for the complete sample the curve stays virtually constant. To illustrate the erosion process, expected HSS curves were calculated for different numbers of generations ($T = 10, 20, 70$). In this example, the expected HSS(l) values increase with the number of generations, suggesting that the bottleneck effect is stronger than erosion.

3. Discussion

In forward simulation it is common to determine the distribution, or compute the expectation, of some summary statistic by considering a large number of simulated data sets. As was illustrated above, the same goal can be achieved by our two-stage procedure, which however is generally computationally more efficient since only the ancestry of the study sample needs to be generated. This ancestry can be restricted to include only individuals that contribute genetic material to the sample.

The method proposed here is flexible and allows one to consider simulations of natural populations, as well as controlled line crosses like advanced intercrossed lines (Darvasi and Soller, 1995). The linkage disequilibrium patterns in simulated chromosomes of such populations can be studied in a detailed fashion using preferred marker maps and by following the route of transmission for each allele in the generated sample.

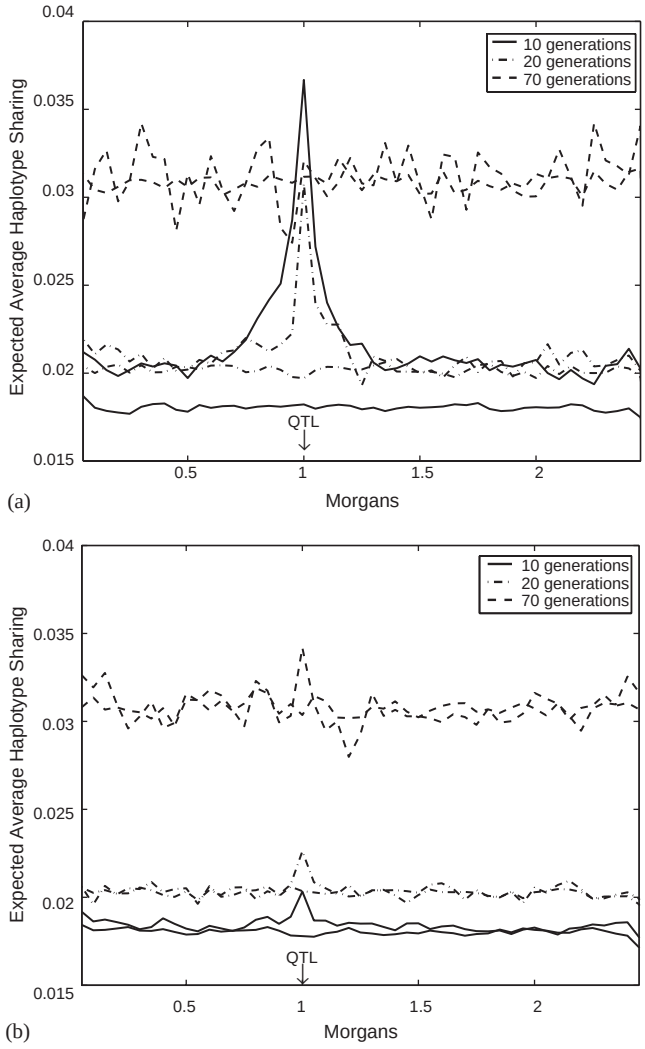

Fig. 2. (a, b). Haplotype sharing. Different curves of expected haplotype sharing statistic HSS(l) averaged over 1000 replicates and presented as a function of locus l calculated at marker positions. In (a) the genetic effect in the phenotype is 10 standard deviations and in (b) the genetic effect is 1 standard deviation. The selection threshold was set respectively to 5 and 1 standard deviations. Three different "population ages" ($T = 10$, 20, or 70) are considered. Two curves are drawn for each T, one corresponding to complete study sample and the other to the selected subsample (right tail of the phenotype distribution). The unknown true QTL position is marked with an arrow on the x-axis. When the population ages were 10 or 20 generations, the HSS-curves corresponding to the selected sample show clearly a peak around the QTL position, whereas after 70 generations there is no signal.

In order to arrive at realistic ancestral graphs, the mating parameters α and β should be estimated from real data. Such a method, based on parent/child data from two generations, is described in the Appendix. In case the sample is ascertained based on a phenotype, these mating parameters do not generally coincide with the mating parameters of the ancestral graph that would arise from randomly sampled individuals. To rule out matings between close relatives one can impose corresponding constraints for the ancestral graph. Extended families of a few generations (that are unrelated to each other) can be simulated by introducing such a taboo constraint, but will result in practice also if the population size is large. The offspring from controlled matings of inbred lines can be obtained with the present approach by assuming that the founder haplotypes come from a mixture of populations with fixed allele frequencies (a single segregating allele) in each.

In human genetics, mostly families rather than individuals are sampled. To mimic this kind of ascertainment strategy, it is possible to specify the mating parameters α and β for individuals in the present generation differently than for the general population, in such a way that coalescing is faster close to the present generation.

When the sample is ascertained from a larger population on the basis of a disease phenotype, a larger random sample is generated first and a subsample of "affected" individuals is then collected. However, the present approach will become computationally untractable if the disease is rare, because the first sample has to be so large. A more realistic approach for reconstructing ancestral graphs and a geneflow in the considered chromosomal segments containing the trait loci of a rare disease can be based on the simultaneous consideration of geneflow and ancestral graph, conditionally on the genetic data from the present generation.

Kuhner and Felsenstein (2000) combined an ancestral recombination graph simulation technique with haplotyping, conditionally on observed genotype data (see also Stephens et al., 2001). We are currently extending the present methodology, using explicit modeling of the whole recombination process and a Markov chain Monte Carlo algorithm to sample the ancestral graph and the allelic paths conditionally on the data. Possible application fields for such an approach include genotype elimination (Du and Hoeschele, 2000; Heath, 1998; Luo and Lin, 2003), relationship estimation (Blouin, 2003), haplotype estimation from population data (Lin et al., 2002), and gene mapping by combining pedigree and linkage disequilibrium information (Perez-Enciso, 2003; Meuwissen and Goddard, 2004).

The simulator described here was used to generate test data for the association mapping method in Kilpikari and Sillanpää (2003). The program source code (written in C language) can be downloaded freely for research purposes from Rolf Nevanlinna Institute's web page (http://www.rni.helsinki.fi/~dag/).

4. Acknowledgments

This work was supported by research Grants (nos. 72530, 52457, 202324) from the Academy of Finland.

Appendix. Maximum likelihood estimation of mating parameters

Here we comment briefly on the possibility of obtaining the values of the simulation parameters α and β from the data. These parameters control the mating behaviour of individuals within the graph (i.e., degree of monogamy).

Assume the following model: we have a sample consisting of two generations, n children and d fathers. For a parameter value $\alpha>0$, we consider a d-dimensional random probability vector $\pi=(\pi_1,\dots,\pi_d)\simeq$ Dirichlet$(\alpha,\dots,\alpha)$. Then, conditionally on π, children choose independently their father by sampling from the distribution π.

Denote $N_i := \#\{\text{children of } i\text{th father}\}$, $i=1,\dots,d$, and let $C(k) := \#\{i\leqslant n : N_i = k\}$ be the number of fathers with k children.

Since the Dirichlet and multinomial distributions are conjugate, we can integrate out π and compute the likelihood for the parameter α:

$$\begin{aligned} L(\alpha;n,d) &= \frac{n!}{N_1!\dots N_d!}\frac{\Gamma(\alpha d)}{\Gamma(\alpha)^d} \\ &\quad\times\int_{S^{d-1}} \pi_1^{N_1+\alpha-1}\dots\pi_d^{N_d+\alpha-1}d\pi_1\dots d\pi_d \\ &= \frac{n!}{N_1!\dots N_d!}\times\frac{\Gamma(\alpha d)}{\Gamma(\alpha)^d} \\ &\quad\times\frac{\Gamma(\alpha+N_1)\dots\Gamma(\alpha+N_d)}{\Gamma(\alpha d+n)}, \end{aligned}$$

where the integral is over the $(d-1)$-dimensional simplex $S_{d-1}=\{x\in[0,1]^d : x_1+\dots+x_d=1\}$.

Noting that

$$\frac{\Gamma(\alpha+N_1)\dots\Gamma(\alpha+N_d)}{\Gamma(\alpha)^d} = \prod_{k=1}^{n}\left(\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\right)^{C(k)}$$

and using the identity $\Gamma(\alpha)=(\alpha-1)\Gamma(\alpha-1)$ repeatedly, the log-likelihood is

$$\begin{aligned} l(\alpha) := \log L(\alpha) &= \text{const.} + \sum_{k=1}^{n}\{C(k)[\log(\alpha) \\ &\quad+\log(\alpha+1)+\dots+\log(\alpha+(k-1))] \\ &\quad-\log(\alpha d+k-1)\} \end{aligned}$$

so that the MLE $\hat\alpha$ is found by solving numerically the equation

$$\begin{aligned} 0 &= \frac{dl(\alpha)}{d\alpha} \\ &= \sum_{k=1}^{n}\{C(k)[\alpha^{-1}+(\alpha+1)^{-1}+\dots+(\alpha+(k-1))^{-1}] \\ &\quad-(\alpha d+k-1)^{-1}\}. \end{aligned}$$

The Fisher information at a parameter value α has the expression

$$\begin{aligned} I(\alpha;n,d) &= -\mathbb{E}_{\alpha,n,d}\frac{d^2l(\alpha)}{d\alpha^2} \\ &= \sum_{k=1}^{n}\{\mathbb{E}_{\alpha,n,d}(C(k))[\alpha^{-2}+(\alpha+1)^{-2}+\dots \\ &\quad+(\alpha+(k-1))^{-2}]-d(\alpha d+k-1)^{-2}\}, \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}_{\alpha,n,d}(C(k)) &= \mathbb{P}_{\alpha,n,d}(N_1=k)\times d \\ &= \binom{n}{k}\frac{\text{Beta}(\alpha+k,\alpha(d-1)+n-k)}{\text{Beta}(\alpha,\alpha(d-1))}\times d \\ &= \frac{n!}{k!(n-k)!}\times\frac{\Gamma(\alpha d)}{\Gamma(\alpha)\Gamma(\alpha(d-1))} \\ &\quad\times\frac{\Gamma(\alpha+k)\Gamma(\alpha(d-1)+n-k)}{\Gamma(\alpha d+n)}\times d. \end{aligned}$$

It follows that α can be estimated consistently from such a sample consisting of two generations as $I(\alpha;n,d)\to\infty$. This requires that d, $n\to\infty$ simultaneously. Heuristically, if the vector $\pi=(\pi_1,\dots,\pi_d)$ was observed directly, the information about α would grow linearly in d. By numerical evaluation (results not shown), it seems that for $n=\text{Const.}\times d$, the Fisher information grows linearly as $d\to\infty$.

For the parameter β we proceed analogously. Assuming that there are h mothers, the mating mechanism is the following: for $i=1,\dots,d$ (where d is the number of fathers) we sample i.i.d. h-dimensional random probability vectors $\rho_i\simeq$ Dirichlet$(\beta,\dots,\beta)$. Conditionally on ρ_i, the N_i children of the ith father choose their mother independently by sampling from the distribution ρ_i. If we denote by N_{ij} the number of common children of the ith father and the jth mother, the likelihood for β is given by the product

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{d}\frac{N_i!}{N_{i1}!\dots N_{ih}!}\times\frac{\Gamma(\beta h)}{\Gamma(\beta)^h} \\ &\quad\times\frac{\Gamma(\beta+N_{i1})\dots\Gamma(\beta+N_{ih})}{\Gamma(\beta h+N_i)}. \end{aligned}$$

References

Blackwell, D., MacQueen, J.B., 1973. Ferguson distributions via Polya urn schemes. Ann. Stat. 1, 353–355.

Blouin, M.S., 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol. Evol. 18, 503–511.

Cannings, C., 1975. Latent roots of certain Markov chains arising in genetics—new approach part 2: Further haploid models. Adv. Appl. Probab. 7, 264–282.

Chung, M.-H., Kim, C.K., Nahm, K., 2003. Fractional populations in multiple gene inheritance. Bioinformatics 19, 256–260.

Churchill, G.A., Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics 138, 963–971.

Darvasi, A., Soller, M., 1995. Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141, 1199–1207.

Davies, S., Schroeder, M., Goldin, L.R., Weeks, D.E., 1995. Nonparametric simulation-based statistics for detecting linkage in general pedigrees. Am. J. Hum. Genet. 58, 867–880.

Du, F.-X., Hoeschele, I., 2000. A note on algorithms for genotype and allele elimination in complex pedigrees with incomplete genotype data. Genetics 156, 2051–2062.

Fu, Y.-X., Li, W.-H., 1999. MINIREVIEW Coalescing into the 21st century: an overview and prospects of coalescent theory. Theor. Popul. Biol. 56, 1–10.

Gauderman, W.J., 1995. A method for simulating familial disease data with variable age at onset and genetic and environmental effects. Statist. Comput. 5, 237–243.

Griffiths, R.C., Marjoram, P., 1996. Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3, 479–502.

Hambe, J., Wienker, T., Schreiber, S., Nurberg, P., 1998. POPSIM: a general population simulation program. Bioinformatics 14, 458–464.

Heath, S.C., 1998. Generating consistent genotypic configuration for multi-allelic loci and large complex pedigrees. Hum. Hered. 48, 1–11.

Hudson, R.R., 1993. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23, 183–201.

Kilpikari, R., Sillanpää, M.J., 2003. Bayesian analysis of multilocus association in quantitative and qualitative traits. Genet. Epidemiol. 25, 122–135.

Kingman, J.F.C., 1982. The coalescent. Stochastic Proc. Appl. 13, 235–248.

Kuhner, M., Felsenstein, J., 2000. Sampling among haplotype resolutions in a coalescent-based genealogy sampler. Genet. Epidemiol. 19 (Suppl. 1), 515–521.

Lange, E.M., Lange, K., 2004. Powerful allele sharing statistics for nonparametric linkage analysis. Hum. Hered. 57, 49–58.

Larribe, F., Lessard, S., Schork, N.J., 2002. Gene mapping via the ancestral recombination graph. Theor. Popul. Biol. 62, 215–229.

Lin, S., Cutler, D.J., Zwick, M.E., Chakravarti, A., 2002. Haplotype inference in random population samples. Am. J. Hum. Genet. 71, 1129–1137.

Long, A.D., Langley, C.H., 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res. 9, 720–731.

Luo, Y., Lin, S., 2003. Finding starting points for Markov chain Monte Carlo analysis of genetic data from large and complex pedigrees. Genet. Epidemiol. 25, 14–24.

Luo, Z.W., Wu, C.-I., Kearsey, M.J., 2002. Precision and high-resolution mapping of quantitative trait loci by use of recurrent selection, backcross or intercross schemes. Genetics 161, 915–929.

Maccluer, J.W., Vandeberg, J.L., Read, B., Ryder, O.A., 1986. Pedigree analysis by computer simulation. Zoo. Biol. 5, 147–160.

Meuwissen, T.H.E., Goddard, M.E., 2000. Fine mapping of quantitative trait loci using linkage disequilibiria with closely linked marker loci. Genetics 155, 421–430.

Meuwissen, T.H.E., Goddard, M.E., 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. Genet. Sel. Evol. 36, 261–279.

Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29, 1547–1562.

Möhle, M., Sagitov, S., 2003. Coalescent patterns in diploid exchangeable population models. J. Math. Biol. 47, 337–352.

Morris, A.P., Whittaker, J.C., Balding, D.J., 2002. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am. J. Hum. Genet. 70, 686–707.

Nordborg, M., Tavaré, S., 2002. Linkage disequilibrium: what history has to tell us. Trends Genet. 18, 83–90.

Ollikainen, V., 2002. Simulation techniques for disease gene localization in isolated populations. Ph.D. Thesis, Department of Computer Science, University of Helsinki.

Perez-Enciso, M., 2003. Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. Genetics 163, 1497–1510.

Pfaff, C.L., Parra, E.J., Bonilla, C., Hiestrer, K., McKeigue, P.M., Kamboh, M.I., Hutchinson, R.G., Ferrell, R.E., Boerwinkle, E., Shriver, M.D., 2001. Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. Am. J. Hum. Genet. 68, 198–207.

Podlich, D.W., Cooper, M., 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. Bioinformatics 14, 632–653.

Soderberg, P., Price, F., 2003. An examination of problem-based teaching and learning in population genetics and evolution using EVOLVE, a computer simulation. Int. J. Sci. Educ. 25, 35–55.

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

Strand, A.E., 2002. METASIM 1.0: an individual-based environment for simulating population genetics of complex population dynamics. Mol. Ecol. Notes 2, 373–376.

te Meerman, G.J., van der Meulen, M.A., 1997. Genomic sharing surrounding alleles identical by descent: effects of genetic drift and population growth. Genet. Epidemiol. 14, 1125–1130.

Terwilliger, J.D., Speer, M., Ott, J., 1993. Chromosome-based method for rapid computer simulation in human genetic linkage analysis. Genet. Epidemiol. 10, 217–224.

Uimari, P., Sillanpää, M.J., 2001. Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genet. Epidemiol. 21, 224–242.

van der Meulen, M.A., te Meerman, G.J., 1997. Association and haplotype sharing due to identity by descent, with an application to genetic mapping. In: Pawlowitzki, I.H., Edwards, J.H., Thompson, E.A. (Eds.), Genetic Mapping of Disease Genes. Academic Press, San Diego, pp. 115–136.

Zhao, H., Merikangas, K.R., Kidd, K.K., 1999. On a randomization procedure in linkage analysis. Am. J. Hum. Genet. 65, 1449–1456.

Zöllner, S., von Haeseler, A., 2000. A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. Am. J. Hum. Genet. 66, 615–628.