

Backward Simulation of Ancestors of Sampled Individuals

Dario Gasbarra, Mikko J. Sillanpää, Elja Arjas

Rolf Nevanlinna Institute,
FIN-00014 University of Helsinki, Finland

January 29, 2004

Right running head: Simulation of pedigree and gene flow

Left running head: Gasbarra, Sillanpää, and Arjas

Corresponding author: Mikko J. Sillanpää
Rolf Nevanlinna Institute,
Research Institute of Mathematics, Statistics,
and Computer Science,
P.O. Box 4,
FIN-00014 University of Helsinki, Finland

Telephone number: (358) 9-191-22764

Fax number: (358) 9-191-22779

E-mail: mjs@rolf.helsinki.fi

Abstract

If the population is large and the sampling mechanism is random, the haploid 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 some systematics in the matings (e.g., in animal breeding populations), 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 diploid 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 Haldane's recombination model. 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, and is not restricted 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: diploid coalescent, chromosome simulation, complex pedigrees, Dirichlet distribution, Pólya urn

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; Hampe et al., 1998; Ollikainen, 2002). In contrast, simulation programs based on haploid coalescent theory (Kingman, 1982; Hudson, 1993; Griffith and Marjoram, 1996; Fu and Li, 1999; Nordborg and Tavaré, 2002) start from a sample belonging to the youngest generation and work backwards in time. The mathematical properties of diploid coalescent models have recently been studied by Möhle and Sagitov (2003).

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), (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), and (5) 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 tens of thousands of generations. 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.

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 diploid coalescent theory. Alleles are assigned randomly 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 Haldane’s recombination model. 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.

Once a pedigree has been generated it is possible to include phenotypes by sampling values from a given genetic model, where some markers are treated as genes. Note that this differs from the approaches of Podlich and Cooper (1998) and Chung et al. (2003), which 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.

Methods

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 $(N'_t, N''_t : t = 1, \dots, T)$, 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 T^{th} 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.

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 $F(k) \leq k$. At this stage the $F(k)$ fathers have been assigned, respectively, $C_1(k), \dots, C_{F(k)}(k)$ children, where $C_i(k) \geq 1$ and $\sum_{i=1}^{F(k)} C_i(k) = k$.

Then the $(k+1)^{\text{st}}$ 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

$$\frac{\alpha + C_i(k)}{N'_1\alpha + k}, \quad i = 1, \dots, F(k), \quad \text{and} \quad \frac{\alpha(N'_1 - k)}{N'_1\alpha + k}. \quad (1)$$

After n stages of the algorithm have been completed each child in the sample has been assigned to one the of $F(n) \leq n$ fathers, and the i^{th} father has $C_i(n)$ children. This is equivalent to first sampling a random N'_1 -dimensional probability distribution π from the Dirichlet(α, \dots, α) distribution and then, conditionally on π , sampling the parental assignments from the Multinomial(n, π)

distribution. Here the parameter $\alpha \geq 0$ 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.

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 $M(k) \leq k$ mothers. From the earlier development we know already that the $(k + 1)^{\text{st}}$ child has been assigned some father indexed by $I = I(k + 1) \in \{1, \dots, F(n)\}$. Denote

$C_{ij}(k)$ = number of children of the i^{st} father who after k steps have been assigned to

mother $j \leq M(k)$.

Then the $(k + 1)^{\text{st}}$ 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

$$\frac{\beta + C_{Ij}(k)}{N''_1 \beta + \sum_{j=1}^{M(k)} C_{Ij}(k)}, j = 1, \dots, M(k), \text{ and } \frac{\beta(N''_1 - M(k))}{N''_1 \beta + \sum_{j=1}^{M(k)} C_{Ij}(k)}. \quad (2)$$

The parameter $\beta \geq 0$ 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 $\pi = (\pi_{ij} : i = 1, \dots, N'_1, j = 1, \dots, N''_1)$ according to the Dirichlet (β, \dots, β) 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 $(C_{ij}(n); i = 1, \dots, N'_1, j = 1, \dots, N''_1)$ 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 $\Gamma(\cdot)$ 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.

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 maximum likelihood (see the Appendix).

The construction outlined here defines the distribution of the ancestral graph \mathcal{G} with n_0 roots. Figure 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 $N'_t = N''_t = 7 \times (T - t + 1)$, $t = 1, \dots, 8$, and the mating parameter values were chosen to be $\beta = 0.01$, and (in generation t) $\alpha = \beta N''_t$.

(Figure 1 will be placed here)

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 Haldane's recombination model 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.

Extension: 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 j^{th} 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 j^{th} island, with $n = \sum_{j=1}^J n(j)$.

Suppose the first child in the sample belongs to the j^{th} island. With a probability that is proportional to $\alpha_{jh}N_1'(h)$, the first child from the j^{th} island is assigned a generic father from the h^{th} island. Here $\alpha_{jl} \geq 0$, $j, l \in \{1, \dots, J\}$, 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 j^{th} island have already been assigned to $F(j, k)$ fathers. For $i = 1, \dots, F(j, k)$, let $H(i)$ be the label of the island where the i^{th} father lives, and let $C_i(j, k)$ be the current number of children of the i^{th} father and belonging to the j^{th} island. Then the $(k + 1)^{\text{st}}$ child belonging j^{th} island chooses the i^{th} 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 $(k + 1)^{\text{st}}$ child from the j^{th} island with father labeled by i chooses the l^{th} mother belonging to 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 i^{th} father and the l^{th} mother and belonging to the j^{th} island. Of the matrix (β_{jh}) we assume that $\beta_{jh} \geq 0$ 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.

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 (*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 around a given locus are IBD increases (increasing *founder effect*).

These effects can be studied using our simulation procedure. Namely, when alleles are transmitted from one generation to the next, for each haplotype we only need to record the extremes of the IBD-segment around every 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:

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

(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 $\tau = 5$, (Figure 2a) and one with genetic effect $\beta_g = 1$ and threshold $\tau = 1$, (Figure 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 (Figure 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 founder effect is stronger than erosion.

(Figure 2 will be placed here)

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 computationally more efficient since only the ancestry of the study sample needs to be generated.

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.

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 a 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 and without marriages between can be simulated introducing such a 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 present approach by assuming that the founder haplotypes come from a mixture of populations with fixed allele frequencies (a single segregating allele) in each.

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 gene flow in the chromosomes containing the trait loci of a rare disease can be based on the simultaneous consideration of gene flow and ancestral graph, conditionally on the genetic data from the present generation.

A haplotype sampling approach using observed marker data as the starting point and based on explicit modeling of the whole recombination process was recently described by Kuhner and Felsenstein (2000), see also Stephens et al. (2001). We are currently extending the present methodology in a comparable manner, developing 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), relationship estimation, haplotype estimation from population data, and gene mapping by combining pedigree and linkage disequilibrium information (Perez-Enciso, 2003).

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/>).

Acknowledgments

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

References

- Blackwell, D., and MacQueen, J. B., 1973. Ferguson distributions via Polya urn schemes. *Ann. Stat.* 1, 353-355.
- Chung, M-H., Kim, C. K., and Nahm, K., 2003. Fractional populations in multiple gene inheritance. *Bioinformatics* 19, 256-260.
- Churchill, G.A., and Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963-971.
- Darvasi, A., and Soller, M., 1995. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141, 1199-1207.

- Davies, S., Schroeder, M., Goldin, L.R., and Weeks, D.E., 1995. Nonparametric simulation-based statistics for detecting linkage in general pedigrees. *Am. J. Hum. Genet.* 58, 867-880.
- Du, F.-X., and 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., and Li, W.-H., 1999. MINIREVIEW Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Pop. 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. *Statistics and Computing* 5, 237-243.
- Griffiths, R.C, and Marjoram, P., 1996. Ancestral inference from samples of DNA sequences with recombination. *J. Comp. Biol.* 3, 479-502.
- Hambe, J., Wienker, T., Schreiber, S., and 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. Pop. Biol.* 23, 183-201.
- Kuhner, M., and Felsenstein, J., 2000. Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genet. Epidemiol.* 19 (Suppl. 1), 515-521.
- Kilpikari, R., and 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.
- Larribe, F., Lessard, S., and Schork, N.J., 2002. Gene mapping via the ancestral recombination graph. *Theor. Pop. Biol.* 62, 215-229.

- Long, A.D., and 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, Z.W., Wu, C.-I., and 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.
- Morris, A.P., Whittaker, J.C., and Balding, D.J., 2002. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am. J. Hum. Genet.* 70, 686-707.
- Möhle, M. and Sagitov, S., 2003. Coalescent patterns in diploid exchangeable population models. *J. Math. Biol.* 47, 337-352.
- Nordborg, M., and Tavaré, S., 2002. Linkage disequilibrium: what history has to tell us. *Trends Genet.* 18, 83-90.
- 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., and 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., and Cooper, M., 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. *Bioinformatics* 14, 632-653.
- Ollikainen, V., 2002. Simulation techniques for disease gene localization in isolated populations. Ph. D. thesis, Dept. of Computer Science, University of Helsinki.
- Stephens, M., Smith, N.J., and Donnelly, P., 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68, 978-989.
- Terwilliger, J.D., Speer, M., and Ott, J., 1993. Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet. Epidemiol.* 10, 217-224.

- Uimari, P., and 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., and 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.
- te Meerman, G.J., and 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.
- Zhao, H., Merikangas, K.R., and Kidd, K.K., 1999. On a randomization procedure in linkage analysis. *Am. J. Hum. Genet.* 65, 1449-1456.
- Zöllner, S., and von Haeseler, A., 2000. A coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 66, 615-628.

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 $\pi = (\pi_1, \dots, \pi_d) \simeq \text{Dirichlet}(\alpha, \dots, \alpha)$. Then, conditionally on π , children choose independently their father by sampling from the distribution π .

Denote $N_i := \#\{\text{children of } i^{\text{th}} \text{ father}\}$, $i = 1, \dots, d$, and let $C(k) := \#\{i \leq n : N_i = k\}$ 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 :

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

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

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

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

$$0 = \frac{dl(\alpha)}{d\alpha} = \sum_{k=1}^n \left\{ C(k)[\alpha^{-1} + (\alpha + 1)^{-1} + \dots + (\alpha + (k - 1))^{-1}] - (\alpha d + k - 1)^{-1} \right\}$$

The Fisher information at a parameter value α has the expression

$$I(\alpha; n, d) = -\mathbb{E}_{\alpha, n, d} \frac{d^2 l(\alpha)}{d\alpha^2}$$

$$\begin{aligned}
&= \sum_{k=1}^n \left\{ \mathbb{E}_{\alpha,n,d}(C(k))[\alpha^{-2} + (\alpha+1)^{-2} + \dots + (\alpha+(k-1))^{-2}] - d(\alpha d + k - 1)^{-2} \right\}, \text{ where} \\
\mathbb{E}_{\alpha,n,d}(C(k)) &= \mathbb{P}_{\alpha,n,d}(N_1 = k) \times d = \binom{n}{k} \frac{\text{Beta}(\alpha+k, \alpha(d-1)+n-k)}{\text{Beta}(\alpha, \alpha(d-1))} \times d = \\
&\frac{n!}{k!(n-k)!} \times \frac{\Gamma(\alpha d)}{\Gamma(\alpha)\Gamma(\alpha(d-1))} \times \frac{\Gamma(\alpha+k)\Gamma(\alpha(d-1)+n-k)}{\Gamma(\alpha d + n)} \times d
\end{aligned}$$

It follows that α can be estimated consistently from such a sample consisting of two generations as $I(\alpha; n, d) \rightarrow \infty$. This requires that $d, n \rightarrow \infty$ 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 $d \rightarrow \infty$.

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 $\rho_i \simeq \text{Dirichlet}(\beta, \dots, \beta)$. Conditionally on ρ_i , the N_i children of the i^{th} 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 i^{th} father and the j^{th} mother, the likelihood for β is given by the product

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

Figure legends:

Figure 1. Simulated pedigree.

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

Figures 2a, 2b. Haplotype sharing.

Different curves of expected haplotype sharing statistic $HSS(l)$ averaged over 1000 replicates and presented as a function of locus l .

In Figure 2a the genetic effect in the phenotype is 10 standard deviations and in Figure 2b the genetic effect is 1 standard deviation. The selection threshold was set respectively to 5 and 1 standard deviations.

Three different “population ages” ($T=10, 20, \text{ or } 70$) 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 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.





