Available online at www.sciencedirect.com
ScienceDirect
Theoretical Population Biology 72 (2007) 305–322
Theoretical Population Biology
www.elsevier.com/locate/tpb

Estimating genealogies from unlinked marker data: A Bayesian approach

Dario Gasbarra[a,*], Matti Pirinen[a], Mikko J. Sillanpää[a], Elina Salmela[b,c], Elja Arjas[a,d]

[a]Department of Mathematics and Statistics, University of Helsinki, P.O.Box 68, FIN-00014, Finland
[b]Finnish Genome Center, Institute for Molecular Medicine Finland, University of Helsinki, Finland
[c]Department of Medical Genetics, University of Helsinki, Finland
[d]National Public Health Institute (KTL), Finland

Received 22 March 2006
Available online 22 June 2007

Abstract

An issue often encountered in statistical genetics is whether, or to what extent, it is possible to estimate the degree to which individuals sampled from a background population are related to each other, on the basis of the available genotype data and some information on the demography of the population. In this article, we consider this question using explicit modelling of the pedigrees and gene flows at unlinked marker loci, but then restricting ourselves to a relatively recent history of the population, that is, considering the genealogy at most some tens of generations backwards in time. As a computational tool we use a Markov chain Monte Carlo numerical integration on the state space of genealogies of the sampled individuals. As illustrations of the method, we consider the question of relatedness at the level of genes/genomes (IBD estimation), using both simulated and real data.

Keywords: Identity-by-descent estimation; Relatedness estimation; Pedigree reconstruction; Markov chain Monte Carlo

1. Introduction

Unobserved or unknown genetic links between the study individuals hamper many genetic analyses. Often this aspect is simply ignored, which may lead to incorrect statistical inferences. On the other hand, various approaches to estimating genetic relatedness between the study subjects, usually based on probabilities that their alleles are identical-by-descent (IBD), have been developed for carrying out a more elaborate analysis of such cases (Blouin, 2003; Weir et al., 2006). In this article we introduce a novel methodology for studying the relatedness of the sampled individuals when restricting to the relatively recent history of the population.

Our starting point is a genotyped sample of individuals from some background population with unknown relationships between them. The marker loci are assumed to be unlinked whereas no restrictions are posed on the number of segregating alleles. The recent history of the population is characterized by the following assumptions: (1) non-random mating with known values of mating parameters and known size of the population (Gasbarra et al., 2005) in each generation, (2) the time (expressed as the number of generations) from the founding of the population, and (3) the marker genotype frequencies in the founder population. Using such data, we consider the problem of estimating the pedigree structure and the corresponding IBD patterns over marker loci, performing Markov chain Monte Carlo (MCMC) sampling among the possible genealogies.

Independence of the marker loci is commonly assumed in the methods designed for relatedness estimation, and this is the case also in the present study. The assumption of unlinked loci is well justified if the study has been specifically designed so that only unlinked or nearly unlinked markers are collected, as is often the case in forensic applications, or if there is only limited knowledge of the marker distances and/or of the ordering of the markers. The main practical motivation for an assumption of unlinked markers, if it is not literally true, is that it will

*Corresponding author. Fax: +358 9 191 51400.
E-mail address: Dario.Gasbarra@rni.helsinki.fi (D. Gasbarra).

0040-5809/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2007.06.004

simplify the calculations and thereby lead to faster algorithms. Despite the assumption of no linkage, we model the gene flow jointly at all loci, taking into account the dependencies that diploid reproduction poses on the transmission of (unlinked) markers (i.e. each individual has two parents who contribute equally to his/her genetic composition). This is different from some approaches that first estimate IBD probabilities independently at different loci and then use some weighting scheme to derive genome level estimates (Lynch, 1988; Li et al., 1993; Lynch and Ritland, 1999; Wang, 2002). Note that there are also some approaches for estimating IBD probabilities at different genomic locations, by using linked multilocus markers/haplotypes and knowledge on recent population history (Meuwissen and Goddard, 2001; Hernández-Sánchez et al., 2006). We shall extend our model and the computational algorithm to the case of linked markers in the near future.

Another question concerning relatedness of individuals is the estimation of the genetic structure of a larger population. There the main goal is an assignment of the sampled individuals to internally more homogenous subpopulations. Several methods have been developed for such a purpose (for a review see Excoffier and Heckel, 2006), while in some cases the relevance of the concept of populations/subpopulations can be questioned (e.g. Barbujani and Belle, 2006; Waples and Gaggiotti, 2006). One of the difficulties lies in deciding when groups of individuals are genetically different enough to be considered as subpopulations. In the approach presented here we can largely avoid this semantic problem because we shall view the whole question from the perspective of estimating a large unknown pedigree, and a corresponding gene flow, which are then used to characterize the different levels of genetic similarity between the study individuals. This view has been earlier expressed by Aranzana et al. (2005) and Yu et al. (2006).

The main contribution of this work has been in designing and implementing an MCMC algorithm that works well on the vast state space of possible genealogies, and can cope with highly dependent variables like alleles of closely related individuals. We shall describe the underlying prior and posterior distributions in Sections 2 and 3, respectively. In Section 4 we shall give example analyses with both real and simulated data and then conclude the article with a discussion in Section 5. Algorithmic details will be given in the Appendix.

2. Prior distribution on the configuration space: a hierarchical construction

Consider a sample of $n(0)$ individuals belonging to the "current" generation of a population. In this section we define a probability (prior) model on the space of possible ancestral histories of this sample without considering any known genotype data.

The space Ω of possible ancestral histories for the sampled individuals consists of three components: the ancestral graph (pedigree) that specifies the relationships between individuals, the paths of alleles of these individuals at marker loci, and the types of the founder alleles introduced into the ancestral graph via the founder individuals.

2.1. Prior distribution for ancestral graphs

We use the random ancestral graph model introduced by Gasbarra et al. (2005). Here is a brief summary of the model.

We consider an isolated population with nonoverlapping generations, indexed backwards in time by $t = 0, 1, \ldots, T$ with $t = 0$ referring to the "present" and $t = T$ to the founder generation. We assume that we know (approximately) the number of males, N'_t, and the number of females, N''_t, in generations $t = 1, \ldots, T$. Each individual k in generation $t < T$ has a father $f(k) \in \{1, \ldots, N'_{t+1}\}$ and a mother $m(k) \in \{1, \ldots, N''_{t+1}\}$ within the population and we denote by $X_k = (f(k), m(k))$ the pair of parents of k. We fix two parameters α and β governing the mating behavior in the population. In short, α controls the distribution of offspring among males of the population and β is used to tune the degree of monogamy; see Gasbarra et al. (2005) for details. We assume that we have a sample of $n(0)$ individuals belonging to the present generation and then come up with probabilities for different possible pedigrees connecting these individuals to the founder generation T. An example of a pedigree with $T = 9$ and $n(0) = 39$ can be found in Fig. 10.

It seems most natural to describe the distribution on ancestral graphs by explaining how one can sample from it. The construction is sequential, proceeding one generation at a time. We start with $n(0)$ individuals belonging to the present generation, who will then choose sequentially their parents from amongst N'_1 potential fathers and N''_1 potential mothers belonging to generation 1 of the population. By default the first individual chooses the parental pair $X_1 = (1, 1)$. Suppose then that the first k children have been assigned altogether $F(k)$ fathers and $M(k)$ mothers. Then the $(k+1)$th child can choose either one or both of his/her parents from those already assigned to the previous k children, or from the potential parents who were not yet assigned to any children. Denote by $C_{fm}(k)$ the current number of children of parental pair (f, m), and let $C_f(k) = \sum_{m=1}^{M(k)} C_{fm}(k)$ for each $1 \leqslant f \leqslant F(k)$. Then the conditional distribution of the parental choice X_{k+1} of the $(k+1)$th child, given the choices of the previous k children, is specified by

$$
\begin{aligned}
&P(X_{k+1} = (f, m) | X_1, \ldots, X_k) \\
&\quad = P(X_{k+1} = (f, m) | C_{ij}(k),\ i = 1, \ldots, F(k), j = 1, \ldots, M(k)) \\
&\quad = \begin{cases}
\left(\frac{\alpha + C_f(k)}{N'_1 \alpha + k}\right)\left(\frac{\beta + C_{fm}(k)}{N''_1 \beta + C_f(k)}\right) & \text{if } f \leqslant F(k),\ m \leqslant M(k), \\
\left(\frac{\alpha + C_f(k)}{N'_1 \alpha + k}\right)\left(\frac{\beta (N''_1 - M(k))}{N''_1 \beta + C_f(k)}\right) & \text{if } f \leqslant F(k),\ m = M(k) + 1, \\
\left(\frac{\alpha (N'_1 - F(k))}{N'_1 \alpha + k}\right)\frac{1}{N''_1} & \text{if } f = F(k) + 1,\ m \leqslant M(k), \\
\left(\frac{\alpha (N'_1 - F(k))}{N'_1 \alpha + k}\right)\left(\frac{N''_1 - M(k)}{N''_1}\right) & \text{if } f = F(k) + 1,\ m = M(k) + 1, \\
0 & \text{otherwise.}
\end{cases}
\end{aligned}
$$

It follows that the parents' offspring counts $(C_i : i \in \text{males})$ and $(C_j : j \in \text{females})$ are exchangeable random vectors, whereas the couples' offspring counts (C_{ij}) are exchangeable only when $\alpha = \beta N_1''$, which corresponds to a $N_1' \cdot N_1''$-dimensional Dirichlet $(\beta, \ldots, \beta)$ distribution for the random probability of choosing each pair of parents. After all $n(0)$ individuals in generation 0 have chosen their parents, generation 1 is formed by including $n(1) = F(n(0)) + M(n(0))$ individuals in the pedigree.

We continue recursively by assigning parents to the $n(1)$ individuals we just included. The process is stopped after T generations with a random number $n(T)$ of founders. This process defines the prior distribution $P_{\mathscr{G}}(\cdot)$ on the set $\mathscr{G}$ of all sex-consistent ancestral graphs over $(T+1)$ generations, with $n(0)$ individuals in generation 0. Sex-consistent means that only a pair of opposite sexes can have a common child. Note that for fixed $n(0)$ and T, $\mathscr{G}$ is a finite set.

2.2. Flow of alleles on the ancestral graph

We continue our hierarchical specification of the prior distribution by considering the flow of alleles through the pedigree at L unlinked loci. By definition, the genome of each individual in the pedigree consists of a pair of haplotypes of which one is inherited from the father and the other from the mother. We assign labels to the haplotypes in such a way that $2k-1$ and $2k$ correspond to the paternal and maternal haplotypes of individual k, where the individuals are labelled progressively starting from the present generation. The flow of alleles is defined by haplotype-specific grandparental origin vectors indicating, for each locus, whether the allele is grandmaternal or grandpaternal.

In this article we consider only unlinked markers whence recombination fractions do not affect the probabilities of allelic paths. As a consequence each set of grandparental origins has the probability $2^{-2N_{(-T)}L}$, where $N_{(-T)}$ is the number of individuals in the pedigree belonging to generations $t<T$ (i.e. non-founders).

The grandparental origins of alleles in generation t determine which alleles in the parents' generation $t+1$ have descendants in generation t. If an allele in generation $t>0$ is transmitted to some individual in the present generation $(t=0)$, we say that the allele is *ancestral* and otherwise that it is *censored*.

2.3. Types of alleles

Let E_l be the set of possible types of the alleles at locus l. If the allele at locus l of haplotype i is ancestral, we denote its type by $h_i(l) \in E_l$, otherwise we set $h_i(l) = \emptyset$. For each individual k in the ancestral graph we define a vector of genotypes $g_k = (g_k(l) : l = 1, \ldots, L)$, where $g_k(l) = \{h_{2k-1}(l), h_{2k}(l)\}$ is an unordered pair of alleles and the indexes $(2k-1)$ and $2k$ correspond to the haplotypes of individual k. Note that the genotypes can be censored or partially censored, that is, a genotype $g_k(l)$ takes one of the three possible forms $\{\emptyset,\emptyset\}, \{a,\emptyset\}$ and $\{a,b\}$, with $a, b \in E_l$.

Assuming linkage equilibrium for the founder generation, the types of the founders' ancestral alleles contribute to the prior probability of a configuration through the given population genotype frequencies $fr(\{a,b\}; l)$. (If Hardy–Weinberg equilibrium is assumed, we use the population allele frequencies instead.) These frequencies will be extended to partially censored and censored genotypes in the obvious way

$$fr(\{a,\emptyset\}; l) := \frac{1}{2} fr(\{a,a\}; l) + \frac{1}{2} \sum_{b \in E_l} fr(\{a,b\}; l),$$

which is simply the frequency of allele a in the population, and $fr(\{\emptyset,\emptyset\}; l) := 1$. If it is clear from the context which locus is under consideration, we shall omit the label l from the genotype frequencies. Note that the types of the ancestral alleles in the founder generation together with the allelic paths determine the types of all ancestral alleles in the pedigree. In other words, we are not modelling mutations.

To summarize, the prior distribution of a configuration $\omega \in \Omega$ consisting of ancestral graph G, grandparental origin vectors and the types of ancestral founder alleles is given by

$$\pi(\omega) = P_{\mathscr{G}}(G) \times 2^{-2N_{(-T)}L} \times \prod_{k \in \mathscr{F}} \prod_{l=1}^{L} fr(g_k(l); l),$$

where $\mathscr{F}$ is the set of founders in the ancestral graph and $N_{(-T)}$ is the number of non-founders.

3. Data and posterior distribution

Suppose now that we have been able to (partially) observe the genotypes $(g_k(l) : l = 1, \ldots, L,\ k = 1, \ldots, n(0))$ of the individuals in the current generation, and then want to study the conditional distribution on the configuration space given these data. The posterior distribution is simply

$$\pi(\omega | g_k(l) : l = 1, \ldots, L,\ k = 1, \ldots, n(0)) = \frac{\mathbf{1}(\omega \in \mathscr{C})\pi(\omega)}{\pi(\mathscr{C})}, \tag{3.1}$$

where $\mathscr{C} \subseteq \Omega$ is the set of configurations that are compatible with the genotype data. Note that here the indicator $\mathbf{1}(\omega \in \mathscr{C})$ has the role of a likelihood function.

We are able to sample efficiently from the prior distribution π. However, the proportion of configurations that are consistent with the genotype data is far too small for sampling from the prior to be a practical means for exploring the posterior. It is also clear that the space of consistent histories is too large for exact calculations in cases of practical interest. Thus we shall use an MCMC algorithm to estimate the quantities of interest that, generally speaking, are integrals with respect to the

posterior distribution. An overview of the algorithm is given in the Appendix.

4. Numerical examples

Complicated MCMC algorithms may suffer from poor mixing and their implementation may easily gather some programming errors. In order to identify possible problems, one can compare the results of the sampler to the underlying exact posterior distributions in some small test cases. Due to the complexity of our posterior distributions, we monitored the behavior of the sampler without any genetic data (i.e., all genotypes at the present generation were marked missing). In that case the posterior depends only on the pedigree model, for which we were able to calculate some summary statistics, like the distribution of family sizes in the youngest generation given the mating parameters and the population size. We tested the sampler with various values of α and β, the population sizes and the numbers of generations, by visually comparing the distribution of family sizes in the sampled configurations to the known ones. The two distributions seemed to match accurately (results not shown), suggesting that the corresponding parts of the sampler were working correctly. For more complicated data, testing was done by running the algorithm on simulated data for which the "correct results" were known and comparing the obtained values to the correct ones, as explained below.

4.1. The concept of relatedness

The concept of relatedness between two individuals can be defined accurately at the level of their genomes. Ultimately, relatedness between two alleles at the same locus can be measured by the time since their most recent common ancestral allele existed.

Two alleles are said to be identical-by-state (IBS) if they have the same allelic form (e.g. similar molecular composition). We adopt the traditional way in pedigree analyses and say that two alleles are identical-by-descent (IBD) if they descend from the same allele introduced into the pedigree via some founder individual. Since we ignore the possibility of mutations, it follows that IBD alleles are also IBS. However, two alleles may be IBS without being IBD, if two or more founder alleles happen to be of the same type. In our examples, the concept of IBD tells us whether two alleles have a common ancestor within some fixed number of generations of the population, but it does not estimate their possible coalescent times more accurately. However, if needed, our algorithm could also keep track of the actual coalescing events within the pedigree.

To quantify the concept of relatedness we denote by $r_{ij}(l)$ the probability that a randomly chosen allele from locus l of individual i has an IBD copy in individual j. For individuals i and j we define the locus-specific relatedness coefficient $R_{ij}(l) = \frac{1}{2}(r_{ij}(l) + r_{ji}(l))$ and the genome-level relatedness coefficient $R_{ij} = 1/L \sum_{l=1}^{L} R_{ij}(l)$, which is an estimate of the proportion of the genetic material that i and j share IBD. Note that in the presence of inbreeding we can have $r_{ij}(l) \neq r_{ji}(l)$. However, always $R_{ij}(l) = R_{ji}(l)$ and $R_{ij} = R_{ji}$.

4.2. Estimating population structure

We now consider three examples where our aim was to establish, at least qualitatively, a relatedness structure among the sampled individuals. The results are visualized using dendrograms and classical multidimensional scaling. The method was first tested on simulated data and then applied to two real data sets.

4.2.1. Example I: families from three populations

Data simulation: In this example we simulated data in such a way that there would be three different degrees of relatedness: between and within subpopulations, and between siblings. The data contained 90 individuals collected from three different population isolates all of which had their origin in the same founder population 20 generations ago. The simulation was done with the pedigree simulator of Gasbarra et al. (2005). This simulator actually samples a pedigree and a gene flow from a distribution that is similar to our prior distribution for the ancestral pedigree (see Section 2). Each of the three subpopulations had a constant size of 200 individuals (100 males and 100 females) for the first 10 generations and then grew exponentially by a factor of 1.1 for the last 10 generations. Mating parameters were given values $\alpha = 1.0$ and $\beta = 0.001$. In addition, in order to see whether our method could also identify close relatives, sampling from the current generation was done in groups of three siblings. Thus our final sample contained 30 individuals from each of the three subpopulations, and within each subpopulation the individuals were divided into 10 sibships each containing three children (see Fig. 1). The genotype data at 15 unlinked microsatellite markers were simulated by sampling all founder alleles using the same allele frequencies. The allele frequencies were taken from real genotype data on one hundred Finnish females, originally collected for paternity testing (M. Lukka, personal communication) and the number of alleles varied between 4 and 10 at different loci.

Reconstruction: Five independent runs of the MCMC algorithm were executed on the genotype data on the 90 sampled individuals. The length of each run was 500 000 iterations of which 100 000 were considered as a burn-in part and discarded from the results. The relatedness coefficients were saved from every tenth iteration. The reconstruction was done for nine generations and the mating parameters used were $\alpha = 1.0$ and $\beta = 0.01$. The population size was chosen to be small (100 males and 100 females in each generation) as we wanted to avoid large pedigrees and to squeeze the underlying relatedness structure into the framework of nine generations. Because of this, our results should be considered only as providing a

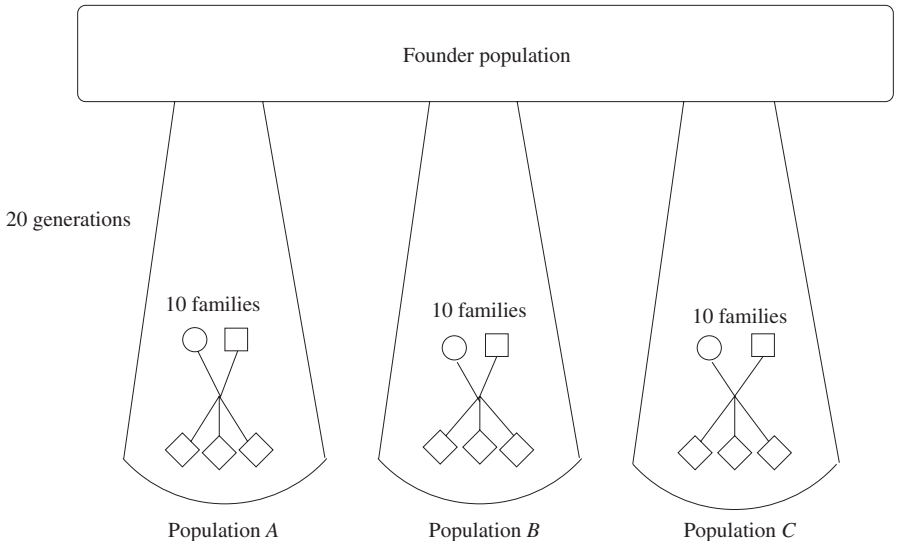

Fig. 1. Schematic representation of the structure of simulated data in Example I. Three populations (A, B, C) originated from the same founder population 20 generations ago. The sampling was done by collecting 10 sibships of 3 children from each of the three populations.

qualitative description of the relative levels of relatedness between pairs of sampled individuals. The genotype frequencies were estimated from the allele frequencies by assuming Hardy–Weinberg equilibrium. Each run took about 3 days on a Pentium-4 2.8 GHz processor.

Results: As a result of the MCMC runs we had estimates for the posterior expectations of the relatedness coefficients (averages over all five runs). In Fig. 2 we have summarized these values in the form of a dendrogram based on a simple hierachical clustering method. The clustering starts from each individual forming its own cluster and proceeds always by merging two existing clusters into one until all individuals belong to the same cluster. In our example the dissimilarity between two clusters C and D is defined as

$$\frac{1}{|C|\,|D|}\sum_{i\in C}\sum_{j\in D}(1-\hat{R}_{ij}),$$

where $\hat{R}_{ij}$ is the (estimated) relatedness coefficient between i and j. At each step, the clustering method merges the pair of clusters with the lowest dissimilarity measure.

The labelling of individuals in this example is such that each triple $\{3k-2, 3k-1, 3k\}$, $k = 1, \ldots, 30$, forms a sibship and each of the three sets of individuals $\{1, \ldots, 30\}$, $\{31, \ldots, 60\}$ and $\{61, \ldots, 90\}$ represents one population isolate (denoted by A, B and C, respectively). In Fig. 2 one can find the correct fine structure of the 30 sibships at the tips of the dendrogram, and by cutting the dendrogram at the level where only three clusters remain, the three population isolates can be separated quite accurately. (There is one exception as triple $\{52, 53, 54\}$ is not in the same cluster as the other sibships sampled from population B.) We have also plotted the estimated relatedness matrix in Fig. 3.

Another way of visualizing the pairwise relatedness estimates is provided by multidimensional scaling. There the aim is to represent the samples on some low-dimensional space (e.g. on a plane) in such a way that the Euclidean distances between the points approximate the dissimilarity measures (here $1-\hat{R}_{ij}$). Fig. 4 has been created by classical multidimensional scaling (Cox and Cox, 2001) as implemented in *cmdscale* function in R software package. The structure of the simulated data seems to be well reconstructed also in Fig. 4 as the three populations form their own clusters and the siblings are mainly represented close to each other.

For comparison we applied the population structure estimation program STRUCTURE (v.2.0.) (Pritchard et al., 2000) without admixture model on these data. STRUCTURE requires as an input parameter the number of populations, denoted by K, to which it attempts to assign the individuals. The results (Fig. 5) given by STRUCTURE are individual-specific estimates of the probabilities of belonging to each of the K postulated populations. One must keep in mind that STRUCTURE operates by estimating allele frequencies for each population and if the data contain only a few individuals from each (postulated) population (like in the lower picture of Fig. 5) the use of STRUCTURE may not be well justified.

4.2.2. *Example II: real data on six human populations*

Rosenberg et al. (2001) analyzed genotype data from eight human populations in order to explore the genetic relationships among them. Currently these data are publicly available on the Internet (see Acknowledgments). In this example we consider a subset of the data consisting of the following six groups: Druze, Ethiopian Jews, Iraqi Jews, Libyan Jews, Moroccan Jews and Yemenite Jews. Exclusion of the remaining two groups of the original data (Ashkenazi Jews and Palestinian Arabs) was made for computational reasons.

The Jewish samples were collected from the second-generation immigrants to Israel from the various source

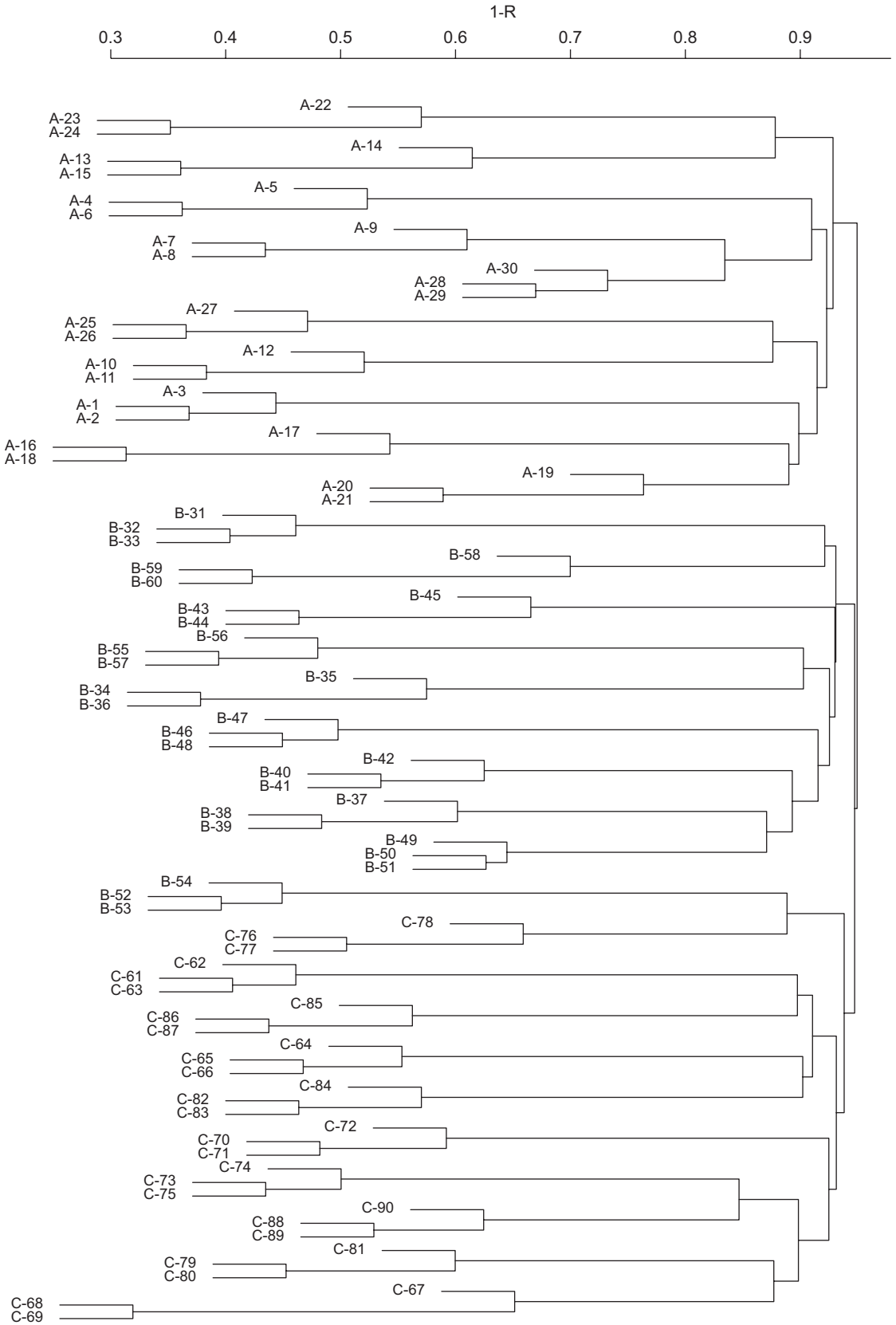

Fig. 2. Results of Example I visualized by a dendrogram. Each of the 90 labelled tips represents one individual. Labels refer to the sampling population (A, B, C) and to the individual's identifier. Two clusters coalesce when $1 - R$ equals the value on the vertical axis, where R refers to the average relatedness coefficient between the individuals belonging to different clusters. Thus, the lower the clusters coalesce, the more related the corresponding individuals are estimated to be.

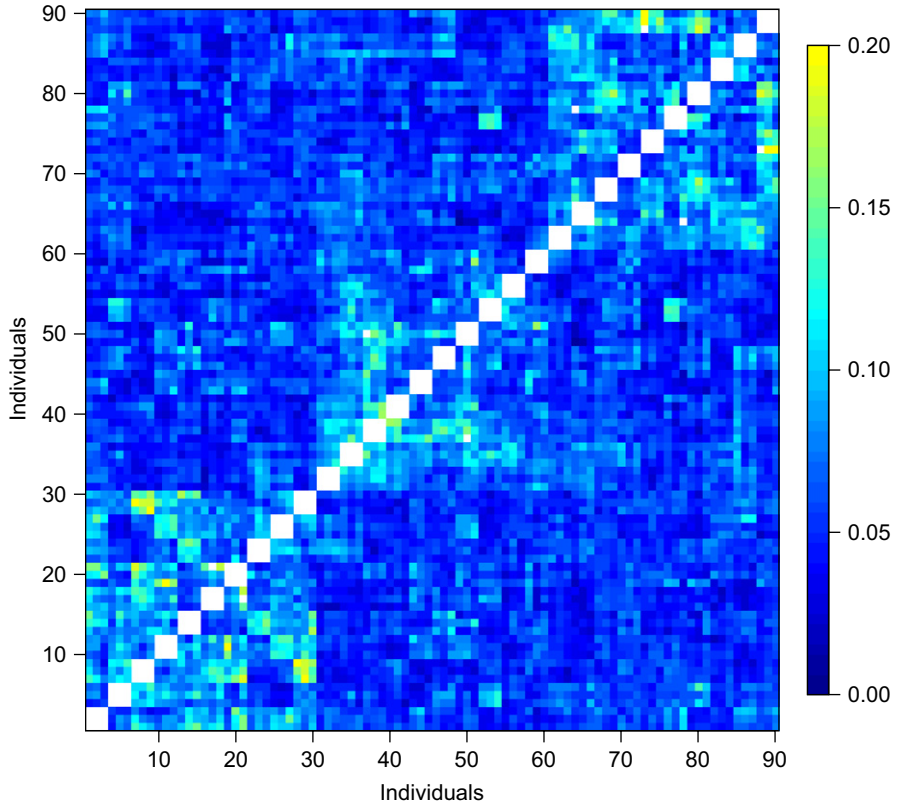

Fig. 3. Results of Example I visualized by a 90×90 matrix, whose entry (i, j) represents the estimated relatedness coefficient between individuals i and j. All values higher than 0.20 are colored white. Thirty white squares on the diagonal indicate the trios of siblings and three larger boxes with light colors result from the population structure.

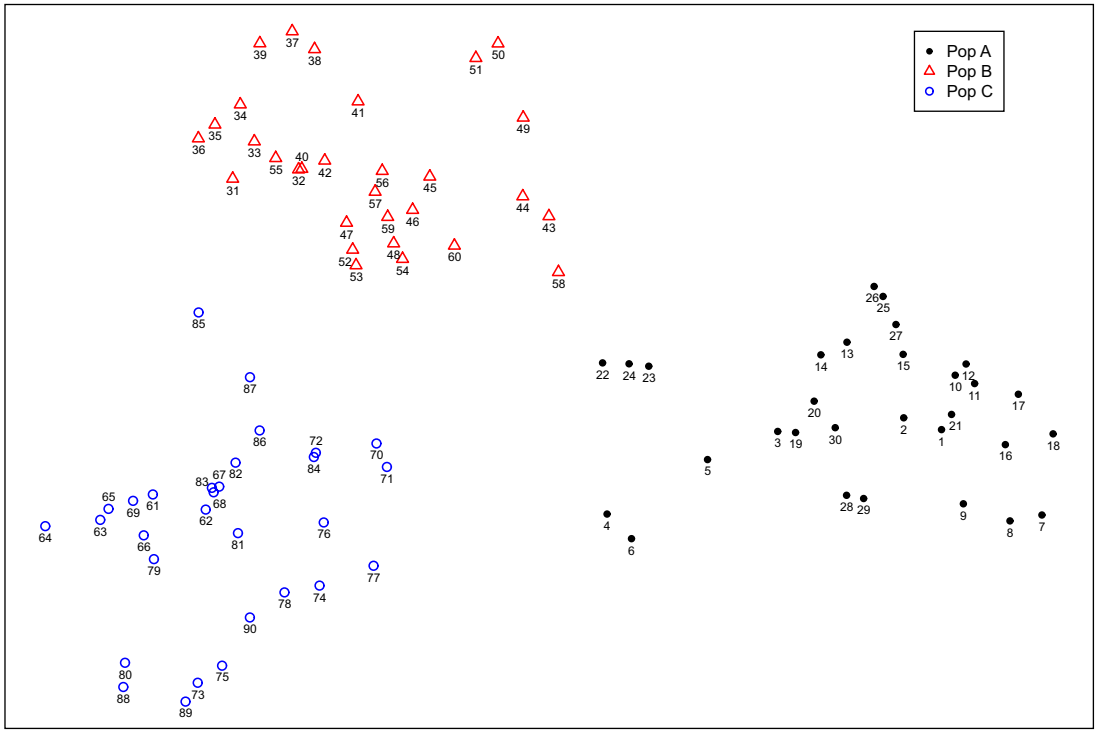

Fig. 4. Results of Example I visualized by classical multidimensional scaling. Each individual is represented as a point on the plane and the distance between two points approximates the dissimilarity $1 - \hat{R}$, where $\hat{R}$ is the estimated relatedness of the corresponding individuals. Thus the closer the two points are to each other, the more related the corresponding individuals are estimated to be.

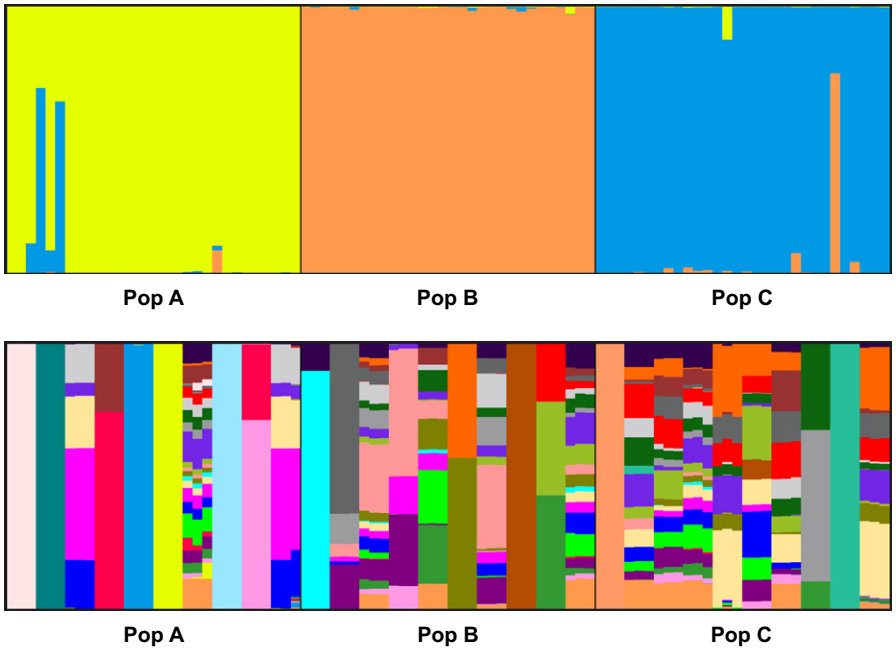

Fig. 5. Results of Example I given by STRUCTURE (no admixture model, burn-in part: 10^5, MCMC iterations: 10^6). In the upper picture the postulated number of populations was $K = 3$ (the number of original populations in the data) and in the lower picture $K = 30$ (the number of trios in the data). In the pictures each postulated population is labelled by its own color, and each of the 90 individuals is represented by one column. The individuals are in ascending order (from left to right). For each individual the proportion of each color describes the (estimated) probability of assigning the individual to the corresponding population. The upper picture identifies the population structure but not the sibling trios, whence in the lower picture the sibling trios are found to have similar genetic composition, but the original population structure is more blurred.

populations and the Druze were collected from the large Druze settlement in Northern Israel. All sampled individuals were males and altogether there were 119 individuals in our analysis (19 representatives of the Ethiopian Jews and 20 from each of the other five groups). Individuals were genotyped for 20 unlinked microsatellites that were spread across 14 autosomes. More details of the data as well as a historical account of these populations can be found in Rosenberg et al. (2001).

Reconstruction: The reconstruction was carried out for nine generations. The parameter values were $N'_t = N''_t = 100 + 10(9 - t)$, $\beta_t = 0.001$ and $\alpha_t = N''_t \beta_t$ and the allele frequencies were estimated from the data. The algorithm was run for 1 000 000 iterations of which 500 000 were considered as a burn-in part. The run took about 7 days on a Pentium-4 2.8 GHz processor.

Results: Rosenberg et al. (2001) analyzed these data by the program STRUCTURE which clustered the data into three genetically distinct groups. The first cluster almost coincided with the group of Libyan Jews, the second cluster contained only Ethiopian (11 individuals) and Yemenite Jews (four individuals) whereas most sampled individuals fell into the third cluster. It was also reported by Rosenberg et al. that the third cluster could not be further divided by STRUCTURE even when analyzed as a separate data set.

When interpreting our results (Figs. 6 and 7) one may first consider how similar the individuals from the same group are to each other (when the frame of reference is defined by the relatedness structure of the whole sample). The results shown in Fig. 7 (top) are consistent with the origins of the samples in that the representatives of each group are mainly concentrated on a relatively small area (excepting maybe the Iraqi Jews for whom the points are spread more widely). It is then another issue to try to determine whether these internally homogenous groups can be separated from each other. At that level of population structure estimation our results are in good agreement with those reported by Rosenberg et al. Most of the Libyan Jews clearly form a cluster of their own and a large proportion of the Ethiopian and Yemenite Jews are reasonably well separated from the other samples. In addition, Fig. 7 (top) suggests that the Libyan Jews can actually be further divided into two distinct groups. This phenomenon was not reported in the analysis performed with STRUCTURE by Rosenberg et al. (2001), but there seems to be an indication of it in the neighbor-joining tree of individual genotypes also reported by Rosenberg et al. (2001).

Rosenberg et al. considered it puzzling that the Druze could not be separated from the other groups. The Druze are an endogameous sect of an Arab origin with apparently a few thousand founder members in the 11th century. In our results the majority of the Druze indeed seem to form a relatively homogeneous group as they are represented close to each other in Figs. 6 and 7. However, our analyses could not separate them from the Iraqi and Moroccan Jews either. Possible explanations for this difficulty have been discussed by Rosenberg et al. (2001).

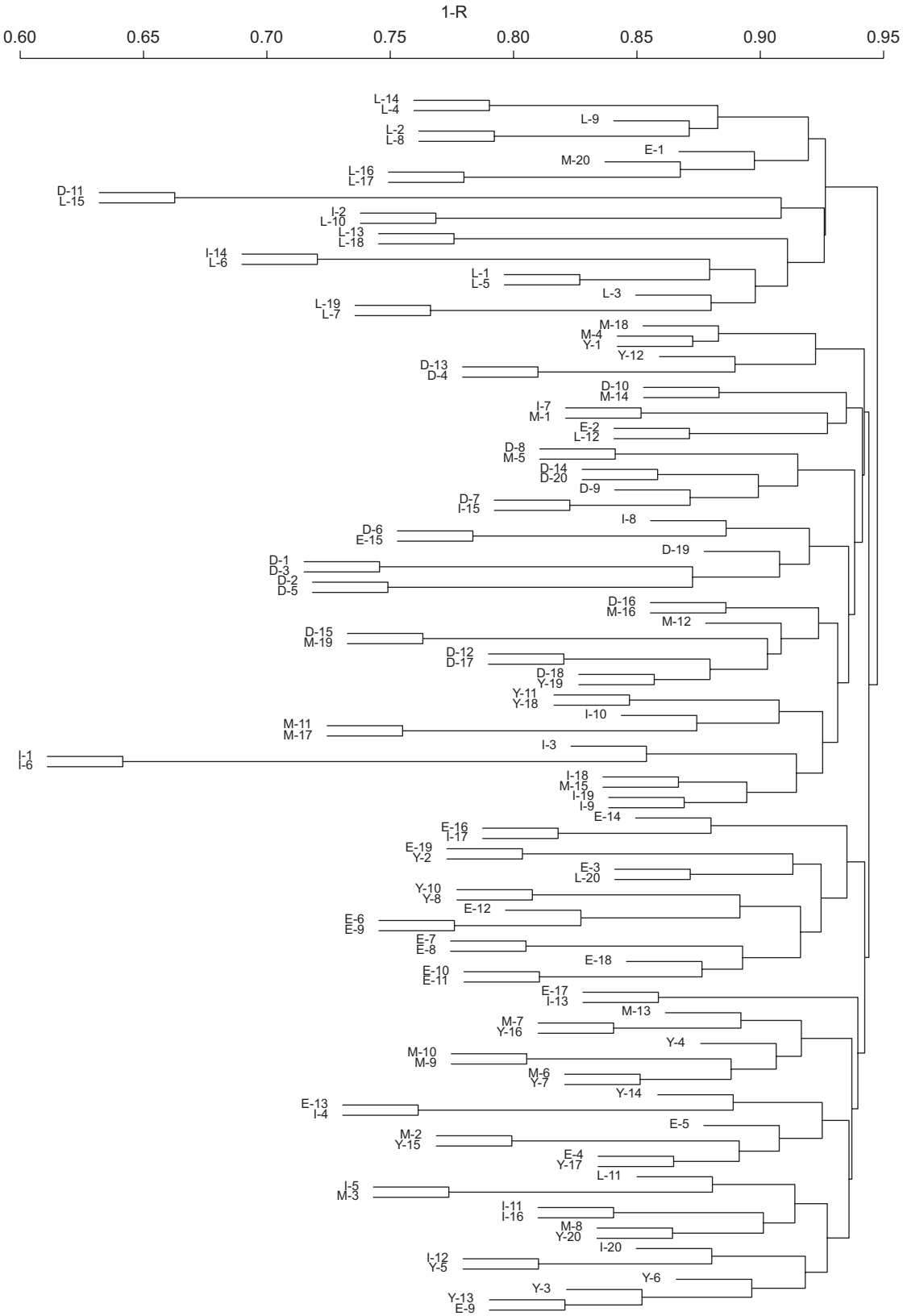

Fig. 6. Results of Example II visualized by a dendrogram. Each of the 119 labelled tips represents one individual. Letters refer to the sampling population: Druze (D), Ethiopian Jews (E), Iraqi Jews (I), Libyan Jews (L), Moroccan Jews (M) and Yemenite Jews (Y).

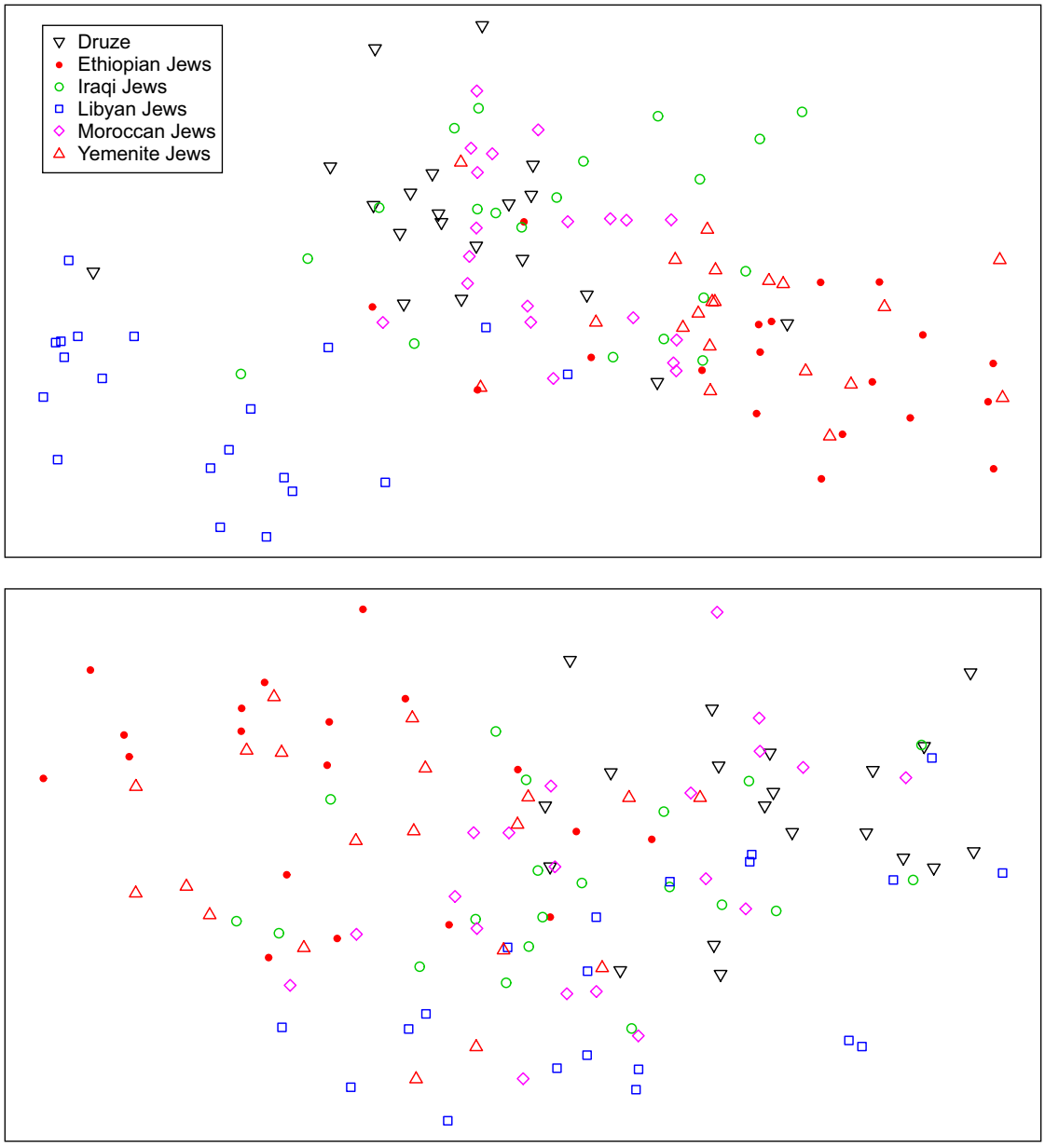

Fig. 7. Classical multidimensional scaling applied to results of Example II (top) and IBS data of Example II (bottom).

To investigate the differences between our results and simple IBS sharing statistics we also calculated the relatedness estimates by defining the similarity of the alleles directly from their IBS status (Fig. 7, bottom). It seems that, when using only the IBS information, the populations appear less homogeneous than in the results which were based on the estimated IBD status. Also the distinctive features of the Libyan Jews are considerably less noticeable in the lower panel than in the upper one in Fig. 7.

4.2.3. *Example III: Finnish data*

In this example we tested our method on real genotype data of 27 microsatellite loci from 34 Finnish males. This is a subset of the data described in Salmela et al. (2006).

Data set: The subjects were Finnish blood donor males who had given an informed consent. They were considered eligible for the sampling if the birthplaces of their all four grandparents were located within the same geographical area of Finland, and if, to their knowledge, none of their relatives had participated in the study earlier. The samples were collected in 1998–1999 when the subjects were between 40 and 55 years of age, and they can therefore be considered to represent a single generation.

For this study, we chose 16 samples with grandparental origin in North Karelia, and 18 samples whose grandparents originated from Ostrobothnia and represented the Swedish-speaking minority. North Karelia is an eastern Finnish province, currently located partly on the Russian side of the border. It is a part of the late-settlement area of

northern and eastern Finland, populated from the 16th century onwards by emigrants from the eastern province of Southern Savo, and their descendants (Norio, 2003a). The province of Ostrobothnia is located on the western coast of Finland. Its Swedish-speaking minority descends from immigrants who arrived from Sweden in the 13th and early 14th century (Virtaranta-Knowles et al., 1991; Pitkänen, 1994). Undoubtedly, both areas have been subject to later immigration and population admixture, but probably to a relatively low extent (Norio, 2003a).

East-west differences are known to exist within Finland in several non-genetic phenomena (Norio, 2003b), as well as Y-chromosomal haplotype frequencies (Kittles et al., 1998; Lappalainen et al., 2006), but it is unclear to what extent the same applies to autosomal DNA markers.

The 27 markers considered here were non-coding autosomal dinucleotides from 14 chromosomes. Even when multiple markers resided on the same chromosome, they were considered to be distributed sparsely enough to justify their treatment as unlinked. The markers were originally chosen based on their rare alleles seen in previous studies on Finns (Finnish Genome Center, unpublished data). Details of the genotyping are given by Salmela et al. (2006). The number of alleles varied between 5 and 16 per locus, and the overall genotyping success was 99.2%.

Reconstruction: We analyzed the data by running five independent chains, each for 500 000 iterations, and took the average of the resulting relatedness coefficients after burn-in parts of 100 000 iterations. The parameters used in the reconstruction were $T = 9, N'_t = N''_t = 100 + 10(T - t), \beta_t = 0.001$ and $\alpha_t = \beta_t N''_t$. The population allele frequencies were estimated based on 465 males from nine Finnish provinces, sampled according to the procedure described above. The genotype frequency estimates for the founder generation were obtained by assuming Hardy–Weinberg equilibrium. The qualitative results were quite similar across different runs, suggesting that the algorithm performed consistently with different starting values. Each run took about 60 h on a Pentium-4 2.8 GHz processor.

Results: The estimated dendrogram is shown in Fig. 8. In the dendrogram the individuals originating from North Karelia are labelled with E (East) and those from coastal Ostrobothnia with W (West). The tree structure does not indicate a clear clustering of the samples according to their geographical origin, even though such a clustering can be observed in several distinct branches. Visualization of the results by classical multidimensional scaling (Fig. 9) suggests that the Eastern samples are more homogeneous than the Western ones, but it does not show a clear separation of the two groups.

The lack of geographical structure could result from insufficient information content of the data, as the number of samples and markers is relatively low. It could also be that the geographical scale of analysis was too broad to reveal significant differences between the regions: at least in blood group markers, the allele frequency differences within Finland are most pronounced on a narrow geographical scale (between communities and villages) and become more subtle when wider regions are compared (Nevanlinna, 1972). Unfortunately, a narrower scale of analysis was not feasible on these data due to the sampling design. Furthermore, the high mutation rate of microsatellites and the slight violation of the assumption of no linkage between the markers could have blurred the genetic structure.

We also analyzed these data with the program STRUCTURE v.2.0 (Pritchard et al., 2000) using the admixture model. For each number of (postulated) populations (K) between 2 and 6, we executed a run of 1 000 000 iterations preceded by a burn-in part of 100 000 iterations. For each value of K, STRUCTURE estimated that each individual originates from each of the K populations with equal proportions (with the precision of 1%), suggesting that the program did not find any indication of a substructure in the data.

4.3. IBD estimation

To compare our method with some existing relatedness estimators we simulated an example for which we were able to calculate the "true" values of relatedness coefficients.

4.3.1. Example IV: 13 nuclear families

Data simulation: We considered an isolated population that was founded nine generations ago by two hundred founders. The population was supposed to have grown by a factor of 1.2 at each generation, yielding approximately one thousand individuals in the current generation. We decided to consider a study sample consisting of the children of 13 nuclear families (four families with two children, five families with three children and four families with four children). We then used the pedigree simulator of Gasbarra et al. (2005) to create an ancestral pedigree that connected these families to the founders of the population. The mating parameters needed in the simulation were set to $\alpha = 0.1$ and $\beta = 0.001$ in order to model a human pedigree with a considerable degree of monogamy. We also imposed the constraint that neither full nor half siblings could have common children. The simulation resulted in a 439-member pedigree in which 86% of the matings occurred in nuclear families (Fig. 10). It also turned out after the simulation that all $\binom{39}{2} = 741$ pairs of individuals from the current generation shared at least one common ancestor within this pedigree. The gene flow on the pedigree was simulated at 20 unlinked marker loci. All markers were polymorphic with 10 equally frequent alleles at the population level and Hardy–Weinberg equilibrium was assumed in the genotype simulation.

Reconstruction: The mating parameters α and β can be conveniently estimated by a maximum-likelihood estimator (Gasbarra et al., 2005) from observed family structures if the size of the base population is known. We used this approach to estimate β using the family structures between generations $t = 1$ and 2 in the simulated pedigree, resulting

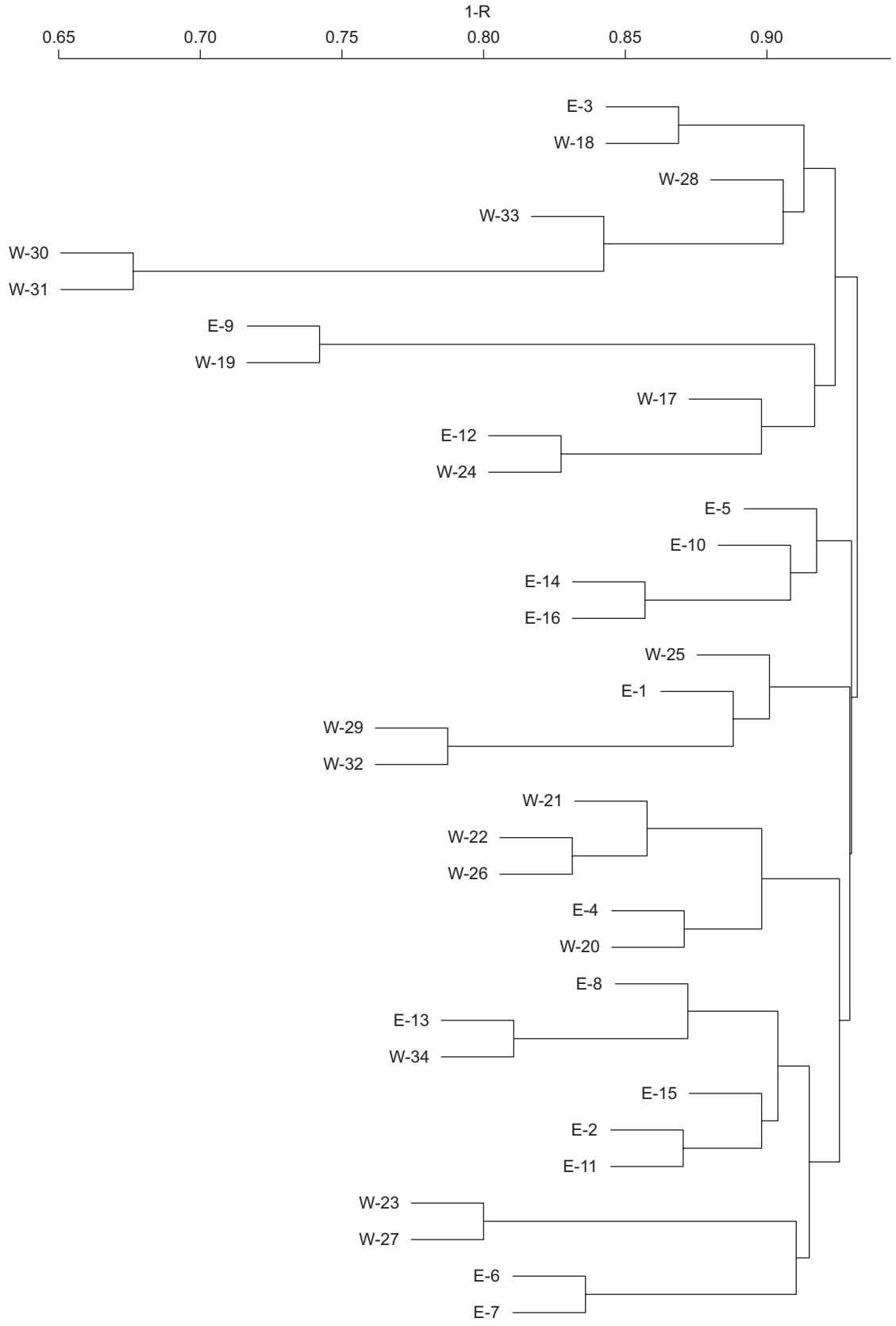

Fig. 8. Results of Example III with 34 Finnish males visualized by a dendrogram. Individuals $1,\ldots,16$ are collected from Eastern Finland and $17,\ldots,34$ from Western Finland.

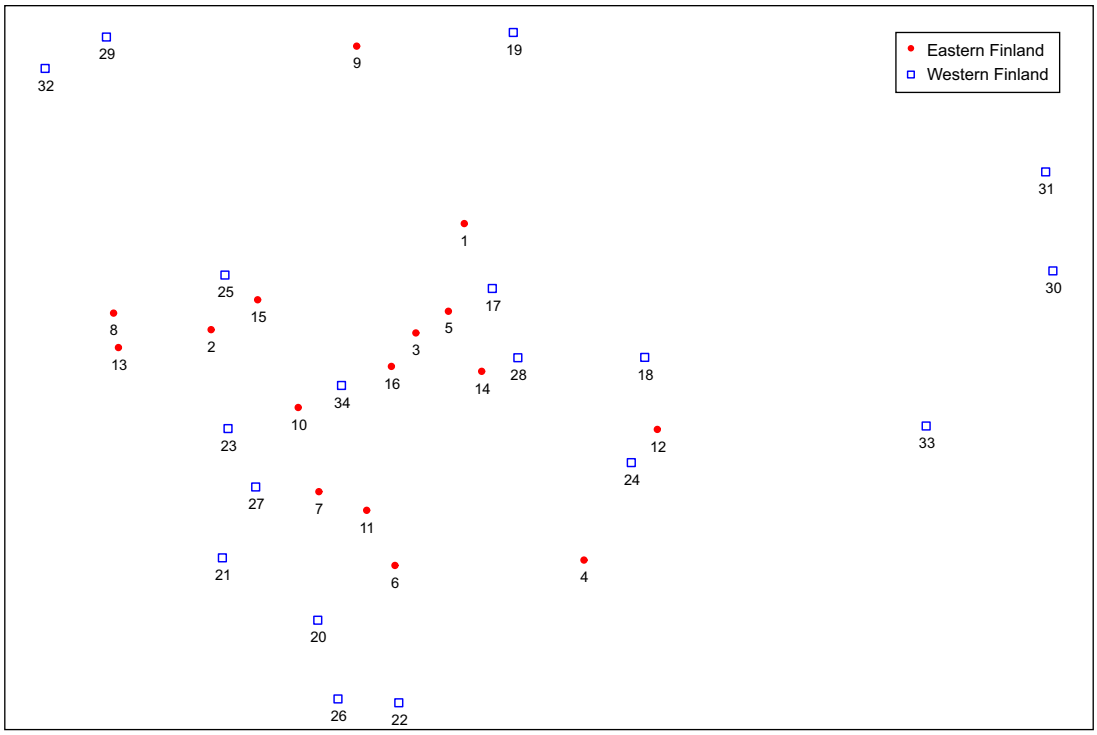

Fig. 9. Results of Example III visualized by classical multidimensional scaling.

in the value $\beta = 4 \times 10^{-4}$. For α we used generation dependent values $\alpha_t = \beta N_t''$, where N_t'' was the number of females belonging to generation t, $t = 1, \ldots, 9$. The population allele frequencies and the size of the population were assumed known in the reconstruction.

The lengths of all five independent sample chains were 100 000 iterations and each run took about 6 h on a Pentium-4 2.8 GHz processor. The monitored statistics behaved very similarly across different runs, suggesting that with these data the method performs consistently regardless of the initial state.

Results: As our input data contain no pedigree information, it seems that other methods available for IBD estimation from such data are based on different formulas that combine in some fashion the IBS status of the markers and the known population allele frequencies to an estimate of the IBD probability (usually R_{ij}). We compared the estimates given by our algorithm to three such moment estimators.

The simplest of the moment estimators (referred here as LL) was developed by Lynch (1988) and later improved by Li et al. (1993). The other two estimators are by Lynch and Ritland (1999) (LR) and Wang (2002) (W). All these methods assume unlinked loci and then combine the locus-specific results according to some weighting schemes in order to obtain estimates of the genome-level relatedness coefficients. In addition, the derivations of both LR and W are based on the assumption that there is no inbreeding. This is violated in our data, which may lead to some additional error in the results reported here for LR and W. Moreover, it is unclear to what extent these moment estimators actually answer the exact question of IBD sharing when restricted to the latest 10 generations (see also Rousset, 2002). The estimates which they give are relative to the base population defined by the allele frequencies and there is no exact reference point of IBD sharing (like the founder generation in our example). On the other hand, polymorphic data sets like the one used here are advantageous for the moment estimators since in these cases the degree of IBS sharing gives already a fairly good approximation of the actual IBD sharing.

We implemented the LL and LR estimators according to the formulas given by Lynch and Ritland (1999) and used the publicly available program MER (v.3.0) to calculate the estimates of W. For our method we estimated the relatedness coefficients by taking averages of the corresponding values over all five MCMC-samples.

The accuracy of the relatedness estimates was measured by squared error. Namely, we computed the true values of the coefficients R_{ij} for each pair of individuals from the original genealogy and compared then the distribution of the squared differences $(R_{ij} - \hat{R}_{ij})^2$ that were obtained by our method and by the three above-mentioned moment estimators. These distributions are shown with boxplots in Fig. 11, where the letter G refers to our method. The sums of squared errors over all 741 pairs of individuals were 1.58, 3.21, 3.35 and 3.58 for our method, LL, LR and W, respectively. For a comparison, the corresponding sum of squared errors is 30.31 if one considers all pairs of individuals as unrelated.

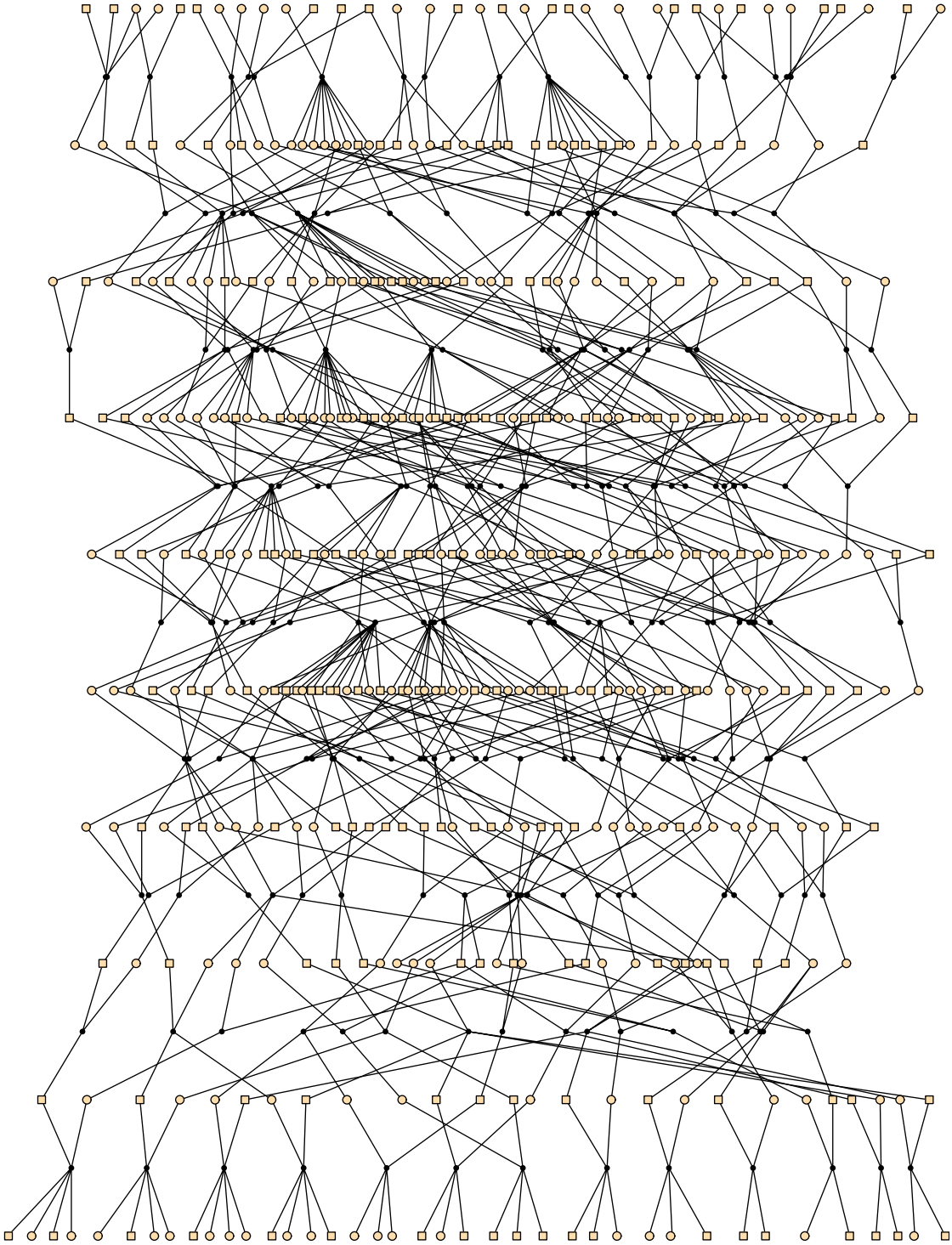

Fig. 10. Simulated pedigree from Example IV. Four hundred and thirty nine individuals and 10 generations of which the youngest one consists of the children of 13 nuclear families. Squares denote males, circles denote females. The pedigree is drawn by program Pedfiddler (J.C. Loredo-Osti and K. Morgan).

5. Discussion

In this paper we have introduced a method for estimating the recent genealogy of a sample of individuals, given that there are available multilocus unlinked genotype data on these individuals and some information on the underlying population. Several current methods analyse relatedness between individuals from different perspectives. For instance, different methods have been introduced for IBD estimation, relationship estimation and for population structure estimation. A central motivation behind our study was to build a unifying framework within which we

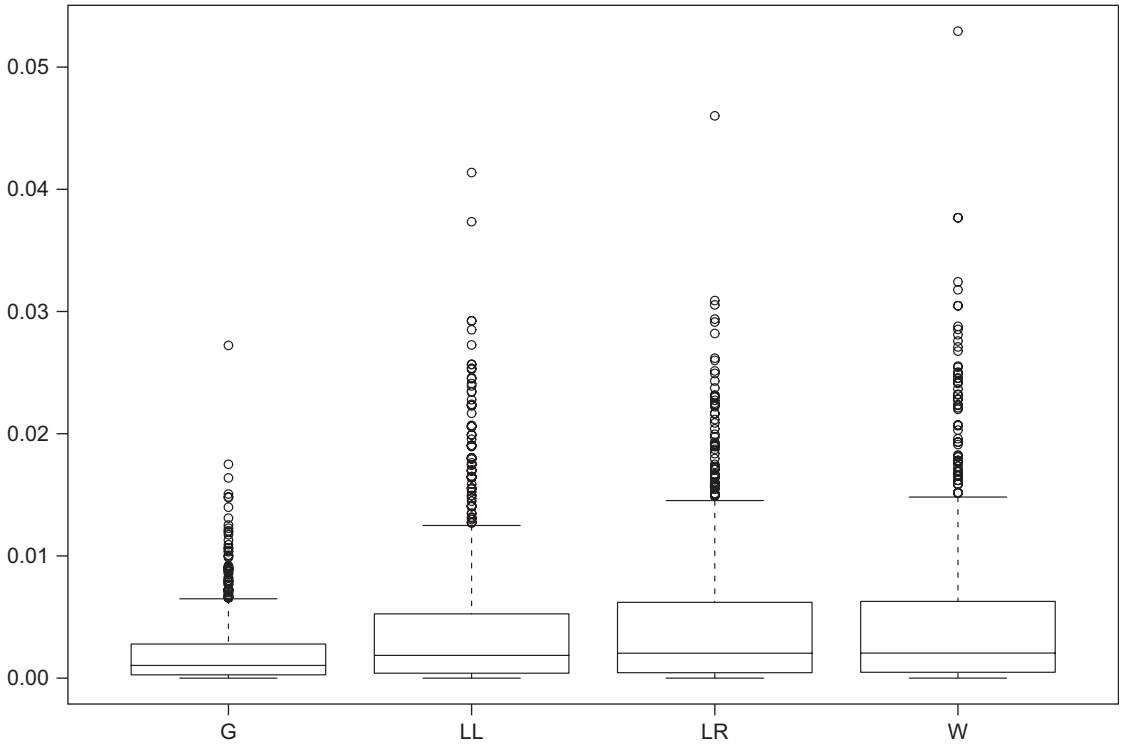

Fig. 11. Squared errors of IBD probability estimates from Example IV. Boxplots show squared errors of all 741 pairwise relatedness coefficients R_{ij}, where i and j are different individuals from generation 0. The boxes indicate the quartiles (1st, 2nd and 3rd) and the 'whiskers' cover the errors whose distance from the box is less than 1.5 times the box size. The outliers are indicated with single points. Methods used: ours (G), Lynch and Li's (LL), Lynch and Ritland's (LR) and Wang's (W).

could simultaneously answer several different questions about relatedness. Our approach towards this goal is a detailed imitation of the true diploid reproduction process, by an explicit modelling of the unknown pedigree and the gene flow therein. Although this effort is technically and computationally demanding, in view of the above examples it seems to form a promising stepping stone to the further development of the method. Naturally our approach has also some limitations, the most obvious one being the number of generations in the reconstruction which currently, for computational reasons, needs to be restricted to about 10. Other restrictions are, for instance, our assumptions of non-overlapping generations and of unrelated founders all belonging to the same population. An obvious limitation in practice, the postulate that the markers are unlinked, will be removed in a forthcoming extension of the MCMC algorithm which applies also to linked marker data.

A central concept in several existing methods for estimating population structure is the number of (sub)-populations represented in the data. This number may either be fixed (Pritchard et al., 2000) or the methods may try to infer it from the data (Corander et al., 2003). These approaches may be successful if the data indeed have a clear population structure. However, in reality the levels of relatedness will vary both within and between any fixed groups of individuals. In order to be able to identify structures that may represent different scales of magnitude (e.g. family structures versus demes) the methods may need to be run several times with different input data and parameters (see Fig. 5). Here we have taken an opposite approach to the problem, by proceeding from the level of individuals to the level of populations. The fundamental question in our framework is, simultaneously for each possible set of individuals, how the individuals are related to each other. This information can then be used, for instance, for assigning the individuals to different groups (or subpopulations) depending on their genetic relatedness. Even in a case where we are not able to clearly detect subpopulations from the data (as in our Finnish data example), the estimated relatedness structure nevertheless gives some information on the differing levels of relatedness between the considered individuals.

Reconstructing plausible pedigrees of the sampled individuals conditionally on the observed marker data always requires some knowledge about the recent history of the population. Here such information is provided in the form of postulated parameter values for controlling population growth and mating, as well as assuming that the members of the founder generation are in linkage equilibrium. As explained by Gasbarra et al. (2005), the mating parameters can be estimated from pedigree data by maximum likelihood, and this method was adapted in the last of our numerical examples. If there were no pedigree data available, one could use formulas linking the parameters to the effective population size (Gasbarra

et al., 2005). One can also use the population size and mating parameters to control the width of the pedigrees and the rate of coalescing events of the alleles. When the study sample is large it may be for computational reasons necessary to "squeeze" the pedigree tighter, by setting the size of the population to be relatively small. As a result, the generations simulated in the reconstruction may differ significantly from the generations in the real population, and thus the numerical values of relatedness estimates with respect to a certain number of generations must be interpreted with care. However, as the first example indicated, the relative relatedness between pairs of individuals (i.e. the population structure) may be found also when the population parameters are used to restrict the size of the pedigrees.

We have assumed known population genotype frequencies in the founder generation. These are important pieces of information in the case of rare genotypes. If there is no reasonable prior information on the genotype frequencies, we could either use the uniform distribution or consider random genotype frequencies with a non-informative Dirichlet prior distribution. In the latter case the frequencies are included in the MCMC updates, and proposed conditionally on the data and on the allelic paths.

Our numerical examples illustrated how one can usefully summarize the relevant posterior information contained in an MCMC sample of ancestral genealogies by considering certain statistics of interest, such as those describing the relatedness between a pair of individuals. In view of the enormous size of the sample space, it is the relative robustness of these summary statistics to the exact pedigree and gene flow information which makes our approach based on MCMC sampling at all feasible.

In addition to the case of linked markers, there are several obvious directions towards which we could extend the model. In a near future our plan is to modify the present reconstruction method in a way which allows us to fix the known parts of the pedigree and the corresponding marker information to the extent in which it is known, and then apply the reconstruction algorithm for building "bridges between these islands". This will be only a technical modification of the algorithm, but there is also a need for extensions of the probability model. For example, we could condition on data which include also some phenotype observations and which might contain genotyping errors. In such a case, up to a normalizing constant, the posterior distribution is given by

$$\pi(\omega|\text{data}) \propto \pi(\omega) \prod_{k=1}^{n(0)} \left\{ P(\phi_k|\omega) \prod_{l=1}^{L} P(g_k(l)|\omega) \right\},$$

where ϕ_k and g_k are the observed phenotypes and genotypes, respectively, and we would need to specify models for phenotypes and genotyping errors.

Acknowledgments

We would like to thank Matti Taskinen for his technical assistance in the software development; Pertti Sistonen, Marja-Liisa Savontaus, Päivi Lahermo and Juha Kere for kindly providing the Finnish data set for analysis, and Matti Lukka for providing the allele frequencies used in Example I. The microsatellite genotyping of the Finnish data set was performed in Finnish Genome Center. The data set for Example II was obtained from Noah Rosenberg's website.[1] We are also grateful to the reviewers for their useful comments on the earlier versions of the manuscript. This research was supported by Grant nos. 50178, 202324, and 53297 (Centre of Population Genetic Analyses) from the Academy of Finland and by the ComBi Graduate School (ES and MP).

Appendix A. An overview of the MCMC sampling method

Conditionally on the observed genetic data from the present generation, the ancestral graph and the paths of the observed alleles are strongly dependent. In order to construct a Metropolis algorithm (see Robert and Casella, 1999) that would be sufficiently well mixing in practice, it is necessary to use large block-updates. These block-updates should also be computable in a reasonable time and have reasonable acceptance probabilities. The craft of MCMC design consists of finding a balance between these requirements. Our general approach is to proceed having in mind the update which is in some sense ideal, viz., the Gibbs update, where the proposal distribution incorporates all the information using Bayes' rule, and then to make compromises when necessary, which means the possibility to include only partial information in the proposal distribution and to use approximations to Bayes' rule.

In our proposal moves a set of children in the pedigree are allowed to change their parents. We use several different ways to choose the set of children involved in the block-update. In one update, we sample a random number of children from the pedigree, who try to change (both of) their parents. In another update we sample a random set of fathers (or mothers) from the pedigree and let all their children change mothers (fathers). In all updates allelic paths from the children to their new ancestors are proposed.

It turns out that when the local move involves only a single child (and free recombination is assumed), we are able to use Gibbs sampling, and we shall next give the details of this case. After that we shall briefly describe a generalization to several children, and explain how an initial configuration for the chain is created.

A.1. Proposal move for a single child

First we shall fix some notation. Suppose that the chain is currently in state $\omega \in \Omega$. We denote by $fr(a|k,\omega)$ the

[1] http://www.rosenberglab.bioinformatics.med.umich.edu/

conditional probability that individual k transmits an allele of type a to his/her hypothetical new child (not necessarily included in the ancestral graph), given the ancestral graph, the ancestral alleles carried by k and his/her ancestors and the genotype frequencies in the founder population. Analogously, we denote by $fr(a,a'|k,k',\omega)$ the joint conditional probability that individuals k and k' transmit the alleles a and a', respectively, to a hypothetical common child. Finally

$$P(\{a,b\}|f,m,\omega) := fr(a,b|f,m,\omega) + \mathbf{1}(a \neq b) fr(b,a|f,m,\omega)$$

is the conditional probability that a hypothetical child with parents f and m inherits genotype $\{a,b\}$. These quantities can be computed recursively across the pedigree starting from the founders.

Assume that the selected child has genotypes $g_c(l) = \{g_c^0(l), g_c^1(l)\}$, with $g_c^0(l), g_c^1(l) \in E_l \cup \{\emptyset\}$, at loci $l = 1,\ldots,L$. Then he/she chooses the parental pair (f,m) with a probability proportional to

$$P(X_c = (f,m)|X_{c'}, c' \neq c) \prod_{l=1}^{L} P(g_c(l)|f,m,\bar{\omega}), \tag{A.1}$$

where the first factor incorporates the prior probability for the ancestral graph, given the parental choices of the other children in the same generation. The configuration $\bar{\omega}$ is obtained from the current configuration ω by erasing from the pedigree the edges between the child and his/her current parents, and consequently truncating the paths of the ancestral alleles carried by the child. In this way the child's current ancestors lose all the ancestral alleles which were currently inherited only by this child.

Having chosen a new father f and a new mother m for child c, we sample new parental origins of the ancestral alleles of c as follows: when both $g_c^0(l)$ and $g_c^1(l)$ are ancestral, we have

$$P(g_c^0(l) \text{ paternal}, \ g_c^1(l) \text{ maternal } |g_c(l), f, m, \bar{\omega}) = \frac{fr(g_c^0(l), g_c^1(l)|f,m,\bar{\omega})}{fr(g_c^0(l), g_c^1(l)|f,m,\bar{\omega}) + fr(g_c^1(l), g_c^0(l)|f,m,\bar{\omega})}. \tag{A.2}$$

Otherwise, if $g_c^0(l)$ is ancestral and $g_c^1(l)$ is censored, we have

$$P(g_c^0(l) \text{ paternal}, \ g_c^1(l) \text{ maternal } |g_c(l), f, m, \bar{\omega}) = \frac{fr(g_c^0(l)|f,\bar{\omega})}{fr(g_c^0(l)|f,\bar{\omega}) + fr(g_c^0(l)|m,\bar{\omega})}, \tag{A.3}$$

and symmetrically if $g_c^1(l)$ is ancestral and $g_c^0(l)$ is censored. After this, in case any ancestral allele was transmitted, we must update the genotypes of the new parents f and m, and eventually also the genotypes of their ancestors.

Assume first that the parents belong to the founder generation. In this case we can update the genotypes of the parents separately. For example, consider the case where $h_{2c-1}(l) = a$, that is, allele a came from the father, and let $g_f(l) = \{x,y\}$ be the current genotype of the father. There are three cases to consider. (i) If x and y are both ancestral, nothing needs to be done. (ii) If $x = y = \emptyset$, we set $g_f(l) = \{a,\emptyset\}$. (iii) If $x \in E_l$ and $y = \emptyset$, and if $x \neq a$, then the genotype of the father must be $g_f(l) = \{x,a\}$, whereas if $x = a$, we set $g_f(l) = \{a,a\}$ with probability

$$\frac{fr(\{a,a\})}{fr(\{a,\emptyset\}) + fr(\{a,a\})}$$

and otherwise $g_f(l) = \{a,\emptyset\}$ remains unchanged.

Next we consider the case when the parents are not founders. Let the alleles in locus l of child c be $h_{2c-1}(l) = a$ (paternal) and $h_{2c}(l) = b$ (maternal), with $a,b \in E_l \cup \{\emptyset\}$, and suppose that at least one of them is ancestral. Consider first the case in which only one of a and b is ancestral and, for example, that it is paternal. If father f has already two ancestral alleles, nothing needs to be done. Otherwise, let $h_{2f-1}(l) = x$ and $h_{2f}(l) = y$, $x,y \in E_l \cup \{\emptyset\}$, where x or y or both are censored. We denote by f' and m' the father and the mother of f, respectively. With probability

$$\frac{\mathbf{1}(x=a) + \mathbf{1}(x=\emptyset) fr(a|f',\bar{\omega})}{\mathbf{1}(x=a) + \mathbf{1}(x=\emptyset) fr(a|f',\bar{\omega}) + \mathbf{1}(y=a) + \mathbf{1}(y=\emptyset) fr(a|m',\bar{\omega})}$$

the ancestral allele a was inherited from the grandfather f', and in case $h_{2f-1}(l)$ was censored, we update it to $h_{2f-1}(l) = a$, and leave $h_{2f}(l) = y$. Otherwise a was inherited from the grandmother m', and if $h_{2f}(l)$ was censored, we update it by setting $h_{2f}(l) = a$, and leave $h_{2f-1}(l) = x$.

Next we consider the case where the child has two ancestral alleles a and b. If only one parent has censored alleles, we are back to the previous case, since at most one parental allele will be updated. Otherwise we have to follow the origins of two censored alleles simultaneously. Assume therefore that both parents f and m have censored alleles, that is, $h_{2f-1}(l) = x, h_{2f}(l) = y$ and $h_{2m-1}(l) = x', h_{2m}(l) = y'$, with x or y or both censored, and x' or y' or both censored. Let f' and m' be the father and the mother of f, and let f'' and m'' be the father and the mother of m. Then with probability

$$\begin{aligned}\{&\mathbf{1}(x=a)\mathbf{1}(x'=b) + \mathbf{1}(x=a)\mathbf{1}(x'=\emptyset) fr(b|f'',\bar{\omega}) \\ &+ \mathbf{1}(x=\emptyset)\mathbf{1}(x'=b) fr(a|f',\bar{\omega}) \\ &+ \mathbf{1}(x=\emptyset)\mathbf{1}(x'=\emptyset) fr(a,b|f',f'',\bar{\omega})\}/C\end{aligned}$$

a was inherited from f' and b was inherited from f'', and the probabilities for the other three cases are formulated analogously. Here C is the joint normalizing constant for all four cases. In each of the cases, the corresponding alleles of the parents become ancestral if they were censored before and then we need to repeat the same procedure further backwards in time until both transmitted alleles have coalesced with an ancestral allele or reached the founder generation.

A.2. Proposal move for several children

Consider now a proposal move where n children are changing their parents simultaneously. In order to use a Gibbs update, it would be necessary to compute at each marker locus the joint conditional law for the transmission of up to $2n$ alleles carried by the children, given the

structure of the ancestral graph above the candidate parents, the ancestral alleles already carried by the candidate parents and their ancestors, and the population genotype frequencies in the founder generation. Instead of doing that we approximate the joint conditional transmission law of $2n$ alleles by the n-fold product of the conditional transmission probabilities for the pairs of ancestral alleles carried by each child. Such an approximation is used in a sequential scheme, where each child in turn chooses his/her parents and transmits the ancestral alleles to them, conditionally on the current state of the ancestral graph and the ancestral alleles already transmitted to the candidate parents. Finally the block-update is accepted or rejected in a single step according to the Metropolis rule.

A.3. Initial configuration

It is not completely trivial to construct an initial configuration for the Markov chain, since the allelic paths and the ancestral graph have to be compatible with the data. We do this sequentially by starting from the current generation. At each step a child chooses his/her parents from the population and transmits his/her ancestral alleles to them by using the prior distribution of the graph and conditioning on the ancestral alleles which the parents had already received from their earlier children, and the genotype frequencies in the population. This procedure differs from the MCMC move described above, since at each stage the parents have no ancestors, and a priori their genotypes follow the population frequencies. This sequential procedure does not produce a sample from the posterior distribution, since at any stage it does not take into account the alleles carried by the later children in the list. However, when the population of candidate parents is large enough, the method is guaranteed to produce a configuration that is compatible with the data.

References

Aranzana, M.J., Kim, S., Zhao, K., Bakker, E., Horton, M., et al., 2005. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 1, e60.

Barbujani, G., Belle, E.M.S., 2006. Genomic boundaries between human populations. Hum. Hered. 61, 15–21.

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

Corander, J., Waldmann, P., Sillanpää, M.J., 2003. Bayesian analysis of genetic differentiation between populations. Genetics 163, 367–374.

Cox, M.F., Cox, M.A.A., 2001. Multidimensional Scaling. Chapman and Hall, London.

Excoffier, L., Heckel, G., 2006. Computer programs for population genetics data analysis: a survival guide. Nat. Rev. Genet. 7, 745–758.

Gasbarra, D., Sillanpää, M.J., Arjas, E., 2005. Backward simulation of ancestors of sampled individuals. Theor. Popul. Biol. 67, 75–83.

Hernández-Sánchez, J., Haley, C.S., Woolliams, J.A., 2006. Prediction of IBD based on population history for fine gene mapping. Genet. Sel. Evol. 38, 231–252.

Kittles, R.A., Perola, M., Peltonen, L., Bergen, A.W., Aragon, R.A., et al., 1998. Dual origins of Finns revealed by Y chromosome haplotype variation. Am. J. Hum. Genet. 62, 1171–1179.

Lappalainen, T., Koivumäki, S., Salmela, E., Huoponen, K., Sistonen, P., et al., 2006. Regional differences among the Finns: a Y-chromosomal perspective. Gene 376, 207–215.

Li, C.C., Weeks, D.E., Chakravarti, A., 1993. Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered. 43, 45–52.

Lynch, M., 1988. Estimation of relatedness by DNA fingerprinting. Mol. Biol. Evol. 5, 584–599.

Lynch, M., Ritland, K., 1999. Estimation of pairwise relatedness with molecular markers. Genetics 152, 1753–1766.

Meuwissen, T.H.E., Goddard, M.E., 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol. 33, 605–634.

Nevanlinna, H.R., 1972. The Finnish population structure. A genetic and genealogical study. Hereditas 71, 195–236.

Norio, R., 2003a. Finnish Disease Heritage I: characteristics, causes, background. Hum. Genet. 112, 441–456.

Norio, R., 2003b. Finnish Disease Heritage II: population prehistory and genetic roots of Finns. Hum. Genet. 112, 457–469.

Pitkänen, K., 1994. Suomen väestön historialliset kehityslinjat. In: Koskinen, S., et al. (Eds.), Suomen väestö. Gaudeamus, Hämeenlinna, pp. 19–63.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods. Springer, New York.

Rosenberg, N.A., Woolf, E., Pritchard, J.K., Schaap, T., Gefel, D., et al., 2001. Distinctive genetic signatures in the Libyan Jews. Proc. Natl. Acad. Sci. USA 98, 858–863.

Rousset, F., 2002. Inbreeding and relatedness coefficients: what do they measure? Heredity 88, 371–380.

Salmela, E., Taskinen, O., Seppänen, J.K., Sistonen, P., Daly, M.J., et al., 2006. Subpopulation difference scanning: a strategy for exclusion mapping of susceptibility genes. J. Med. Genet. 43, 590–597.

Virtaranta-Knowles, K., Sistonen, P., Nevanlinna, H.R., 1991. A population genetic study in Finland: comparison of the Finnish- and Swedish-speaking populations. Hum. Hered. 41, 248–264.

Wang, J., 2002. An estimator for pairwise relatedness using molecular markers. Genetics 160, 1203–1215.

Waples, R.S., Gaggiotti, O., 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Mol. Ecol. 15, 1419–1439.

Weir, B.S., Anderson, A.D., Hepler, A.B., 2006. Genetic relatedness analysis: modern data and new challenges. Nat. Rev. Genet. 7, 771–780.

Yu, J., Pressoir, G., Briggs, H.W., Vroh Bi, I., Yamasaki, M., et al., 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38, 203–208.