European Journal of Human Genetics (2012) 20, 1283–1289

www.nature.com/ejhg

ARTICLE

Inferring separate parental admixture components in unknown DNA samples using autosomal SNPs

Daniel JM Crouch[1] and Michael E Weale*,[1]

The identification of ancestral admixture proportions for human DNA samples has recently had success in forensic cases. Current methods infer admixture proportions for the target sample, but not for their parents, which provides an additional layer of information that may aid certain forensic investigations. We describe new maximum likelihood methods (LEAPFrOG and LEAPFrOG Expectation Maximisation), for inferring both an individual's admixture proportions and the admixture proportions possessed by the unobserved parents, with respect to two or more source populations, using single-nucleotide polymorphism data typed only in the target individual. This is achieved by examining the increase in heterozygosity in the offspring of parents who are from different populations or who represent different mixtures from a number of source populations. We validated the methods via simulation; combining chromosomes from different Hapmap Phase III population samples to emulate first-generation admixture. Performance was strong for individuals with mixed African/European (YRI/CEU) ancestry, but poor for mixed Japanese/Chinese (JPT/CHB) ancestry, reflecting the difficulty in distinguishing closely related source populations. A total of 11 African-American trios were used to compare the parental admixture inferred from their own genotypes against that inferred purely from their offspring genotypes. We examined the performance of 34 ancestry informative markers from a multiplex kit for ancestry inference. Simulations showed that estimates were unreliable when parents had similar admixture, suggesting more markers are needed. Our results demonstrate that ancestral backgrounds of case samples and their parents are obtainable to aid in forensic investigations, provided that high-throughput methods are adopted by the forensic community.
European Journal of Human Genetics (2012) **20,** 1283–1289; doi:10.1038/ejhg.2012.134; published online 27 June 2012

Keywords: population genetics; SNPs; admixture; statistical genetics; genomics; forensic genetics

INTRODUCTION

Recent examples of forensic investigations inferring the genetic ancestry of DNA samples include the 11-M Madrid Bombing case[1,2] and Operation Minstead in the United Kingdom.[3] Both used small panels of ancestry informative markers (AIMs) made up of single-nucleotide polymorphisms (SNPs) or short tandem repeats (STRs) that are highly differentiated in allele frequency between human populations. These are suitable for forensic work where the amount of sample DNA is typically restricted. Currently, information about maternal and paternal ancestry is derived from the mitochondrial and Y haplotypes. These can be assigned to populations of origin, but represent a small fraction of the full genome, and so cannot be assumed to be representative of all the disparate ancestries that make up a single individual.[4] Whole-genome amplification, where PCR is performed using randomly generated primers, may facilitate the collection of genome-wide data from forensic samples,[5] and the nascent single-molecule sequencing technologies offer similar opportunities.

We describe a statistical method, LEAPFrOG (Likelihood Estimation of Admixture in Parents From Offspring Genotypes), for quantifying the admixture proportions of a sample's unobserved parents in addition to their personal admixture; designed to be computationally tractable over large numbers of autosomal SNPs. This parental admixture information could cast additional light on the identity of case samples, which is not possible using existing methods. We also present an alternative approach, LEAPFrOG Expectation Maximisation (EM), which is more appropriate when reliable phased data are available, for example from single-molecule sequencing. We demonstrate that the methods give useful predictions of parental admixture using genome-wide SNP platforms for divergent source populations such as those originating from different continents.

METHODS

When an offspring descends from parents with different ancestries, a departure from Hardy–Weinberg equilibrium (HWE) is induced in all autosomal loci that one might genotype in the offspring. To take an extreme case, if the parents are 100% divergent, then parent 1 will be homozygous AA, parent 2 will be homozygous BB and the offspring will be heterozygous AB at all autosomal SNP loci (arbitrary allele coding), representing a severe departure from HWE. Here we leverage this type of Wahlund effect[6] to infer the genetic make-up of the unobserved parents. We allow each parent to be a genetic mixture of a number of different source populations. We then estimate both the parental admixture proportions and the degree of genetic divergence between them, based only on SNP data obtained from a single offspring, together with data from a reference set of source populations.

We take a maximum likelihood estimation approach, based on the probability of seeing the observed genotypes conditional on the

[1]Department of Medical and Molecular Genetics, King's College London, London, UK
*Correspondence: Dr ME Weale, Department of Medical and Molecular Genetics, King's College London, 8th floor Tower Wing, Guy's Hospital, London SE1 9RT, UK.
Tel: +44 (0)20 7188 2601; E-mail: michael.weale@kcl.ac.uk
Received 27 October 2011; revised 24 April 2012; accepted 16 May 2012; published online 27 June 2012

unknown admixture proportions in the target (offspring) individual and parental divergences in admixture. These are governed by the parameters $m_{1\ldots J\text{-}1}$ and $D_{1\ldots J\text{-}1}$, respectively, where J is the number of source populations and m_J and D_J are constrained by the other parameters. m_j is the proportional contribution (admixture proportion) of source population j in the genome of the target individual, and by definition is made up of equal contributions from the genomes of the two parents. A_{1j} and A_{2j} are the proportional contributions of source population j in parents 1 and 2, respectively, and are determined by the underlying parental divergence parameter D_j as described by Equations 13 and 14.

For LEAPFrOG, maximum likelihood estimates are found using the first derivatives of the likelihood equations presented below. If phased data are available, we use an iterative EM algorithm,[7] LEAPFrOG EM, which assigns chromosomes probabilistically to the most likely parent of origin in the expectation step, and then optimises parental admixture proportions conditional on these probabilities in the maximisation step. Details of both LEAPFrOG and LEAPFrOG EM are as follows.

LEAPFrOG

The Wahlund principle describes the departure of genotype frequencies from HWE in the first generation after divergent populations are fused. It follows that the probabilities of observing the three possible genotypes of a biallelic polymorphism are

$$p(AA) = \bar{P}^2 - w \tag{1}$$

$$p(AB) = 2\bar{P}\bar{Q} + 2w \tag{2}$$

$$p(BB) = \bar{Q}^2 - w \tag{3}$$

where $\bar{P}$ and $\bar{Q}$ are the average major and minor allele frequencies, respectively, and w is the magnitude of the departure, which is dependant on the divergence of the populations. A single offspring can be modelled as the product of an admixture event between two equal-sized 'populations', representing the populations from which parent 1 and 2, respectively, are drawn. If P_1 and P_2 are the major allele frequencies in these two populations, it can be shown that

$$w = \frac{(P_1 - P_2)^2}{4} \tag{4}$$

We now expand the model to allow both parental 'populations' to be admixtures of J distinct 'source' populations. We let p_j and q_j represent, respectively, the major and minor allele frequency in population j, m_j the average contribution of population j across the two parental 'populations' (and hence also the contribution of population j to the offspring's genome), and D_j the degree of divergence between the two parental 'populations', such that $D_j m_j$ and $(1-D_j)m_j$ are the weights for population j in parents 1 and 2, respectively. Expanding Equation 4:

$$w = \frac{\left(2\sum_{j=1}^{J} D_j m_j p_j - 2\sum_{j=1}^{J}(1-D_j)m_j p_j\right)^2}{4} \tag{5}$$

All values of D_j and m_j are constrained to be between 0 and 1. Each parent transmits half of its genetic material to the offspring, so $\sum_{j=1}^{J} D_j m_j$ and $\sum_{j=1}^{J}(1-D_j)m_j$ must be equal to 0.5. These terms are therefore multiplied by 2 in Equation 5 to obtain the allele frequencies in the parental 'populations'.

As all offspring and parental admixture proportions must sum to 1, we only estimate $J-1$ m and $J-1$ D parameters directly. D_J and m_J are replaced in the model as follows:

$$m_J = 1 - \left(\sum_{j=1}^{J-1} m_j\right) \tag{6}$$

$$D_J = \frac{0.5 - \sum_{j=1}^{J-1} D_j m_j}{m_J} \tag{7}$$

Furthermore, in Equation 5, there are limits on the values that D_j can take depending on m_j. For example, if m_j exceeds 0.5, D_j cannot be 1 as this implies that one parent is transmitting more genetic material to the offspring than the other. Therefore, we rearrange so that D_j governs the amount of possible admixture, for population j, originating from one parent. After some algebra, we define

$$W(D_{1\ldots J-1}, m_{1\ldots J-1}) = \left(\sum_{j=1}^{J-1}\left[p_j(2D_j-1) + p_J(1-2D_j)\right]\left[I(m_j \le 0.5)m_j + I(m_j > 0.5)(1-m_j)\right]\right)^2 \tag{8}$$

where $W(D_{1\ldots J-1}, m_{1\ldots J-1})$ is equal to $(P_1-P_2)^2/4$ as described in Equation 4 and I is an indicator variable, which takes the value of 1 if the condition in the parentheses are met and 0 otherwise. We use Equation 8 to measure the magnitude of the Wahlund effect, and rewrite Equations 1–3 as

$$p(AA) = \left(p_J + \sum_{j=1}^{J-1} m_j(p_j - p_J)\right)^2 - W(D_{1\ldots J-1}, m_{1\ldots J-1}) \tag{9}$$

$$p(AB) = 2\left(p_J + \sum_{j=1}^{J-1} m_j(p_j - p_J)\right)\left(q_J + \sum_{j=1}^{J-1} m_j(q_j - q_J)\right) + 2W(D_{1\ldots J-1}, m_{1\ldots J-1}) \tag{10}$$

$$p(BB) = \left(q_J + \sum_{j=1}^{J-1} m_j(q_j - q_J)\right)^2 - W(D_{1\ldots J-1}, m_{1\ldots J-1}) \tag{11}$$

Estimates for all m and D parameters are determined by maximising the following likelihood function using gradient optimisation (partial derivatives not shown):

$$l(D_{1\ldots J-1}, m_{1\ldots J-1}) = \sum_{i=1}^{N} \ln \begin{bmatrix} X_{i1}\left(\left(p_J + \sum_{j=1}^{J-1} m_j(p_j - p_J)\right)^2 - W(D_{1\ldots J-1}, m_{1\ldots J-1})\right) \\ + X_{i2}\left(2\left(p_J + \sum_{j=1}^{J-1} m_j(p_j - p_J)\right)\left(q_J + \sum_{j=1}^{J-1} m_j(q_j - q_J)\right) + 2W(D_{1\ldots J-1}, m_{1\ldots J-1})\right) \\ + X_{i3}\left(\left(q_J + \sum_{j=1}^{J-1} m_j(q_j - q_J)\right)^2 - W(D_{1\ldots J-1}, m_{1\ldots J-1})\right) \end{bmatrix} \tag{12}$$

where N denotes the number of SNPs, i is the SNP identifier and X_i is an indicator vector with three indices corresponding to the three genotypes, which takes the value 1 if that genotype is observed or 0 otherwise. SE for the parameter estimates are obtained from the square root of the inverted Hessian matrix of second-order partial derivatives, which are calculated numerically during optimisation.

We apply an additional constraint to prevent D_1 taking values <0.5. The problem then becomes asymmetrical and a parameter nonidentifiability issue is circumvented. The admixture proportions for parents 1 and 2 (A_{1j} and A_{2j}, respectively) can be easily calculated

from the estimates of D_j and m_j:

$$A_{1j} = 2I(m_j \leq 0.5)D_j m_j + 2I(m_j > 0.5)(D_j(1-m_j) + (m_j - 0.5)) \tag{13}$$

$$A_{2j} = 2I(m_j \leq 0.5)(1-D_j)m_j + 2I(m_j > 0.5)((1-D_j)(1-m_j) + (m_j - 0.5)) \tag{14}$$

LEAPFrOG EM

We begin with phased parental chromosomes, where we know that alleles on the same chromosome must come from the same parent, but alleles on separate chromosomes can come from different parents. The objective is to separate the chromosomes into two groups, each representing a parent, and estimate admixture proportions for both. We assume that the admixture pattern for each transmitted chromosome is the same as the admixture pattern for the full parental genome.

Expectation step

We consider two alternatives $H=1$ and $H=2$ for each chromosome, which are the two ways of allocating the homologous chromosomes to parent 1 and 2 respectively. Denote $p(H=1)$ and $p(H=2)$ as the likelihoods under these respective hypotheses. We calculate the responsibility $\hat{\gamma}_c$ for the cth chromosome as the ratio $p(H=1)/(p(H=1)+p(H=2))$. This can also be written as

$$\ln \hat{\gamma}_c = \ell_{1c} - \ell_{0c} \tag{15}$$

where c is the chromosome pair identifier and

$$\ell_{1c} = \ln p(H=1 \mid \mu_1^t, \mu_2^t, x_{c1}, x_{c2}) \tag{16}$$

$$\ell_{0c} = \ln\left(p(H=1 \mid \mu_1^t, \mu_2^t, x_{c1}, x_{c2}) + p(H=2 \mid \mu_1^t, \mu_2^t, x_{c1}, x_{c2})\right) \tag{17}$$

where t denotes the iteration (initially 1), x_{c1} and x_{c2} are the data for chromosomes 1 and 2 in the homologous pair and $\mu_1$and μ_2 are vectors of population proportions in parents 1 and 2 respectively. It follows that

$$\hat{\gamma}_c = e^{\ell_{1c} - \ell_{0c}} \tag{18}$$

$$\ell_{1c} = \sum_{i=1}^{N_c} \ln \left[\begin{array}{l} \left(X_{i1}\left(\sum_{j=1}^{J} \mu_{1j}^t p_{ij}\right) + (1-X_{i1})\left(\sum_{j=1}^{J} \mu_{1j}^t q_{ij}\right)\right) \\ \times \left(X_{i2}\left(\sum_{j=1}^{J} \mu_{2j}^t p_{ij}\right) + (1-X_{i2})\left(\sum_{j=1}^{J} \mu_{2j}^t q_{ij}\right)\right) \end{array} \right] \tag{19}$$

$$= \sum_{i=1}^{N_c} \ln \left[\begin{array}{l} \left(X_{i1}\left(p_{ij} + \left(\sum_{j=1}^{J-1} \mu_{1j}^t (p_{ij} - p_{ij})\right)\right) + (1-X_{i1})\left(q_{ij} + \left(\sum_{j=1}^{J-1} \mu_{1j}^t (q_{ij} - q_{ij})\right)\right)\right) \\ \times \left(X_{i2}\left(p_{ij} + \left(\sum_{j=1}^{J-1} \mu_{2j}^t (p_{ij} - p_{ij})\right)\right) + (1-X_{i2})\left(q_{ij} + \left(\sum_{j=1}^{J-1} \mu_{2j}^t (q_{ij} - q_{ij})\right)\right)\right) \end{array} \right] \tag{20}$$

where X_{i1} and X_{i2} are indicator values for each chromosome in the homologous pair, which are 1 if the major allele for SNP i is present on that chromosome or 0 otherwise, and N_c is the number of SNPs on the chromosome. l_{0c} is calculated as

$$\ell_{0c} = \ln \left(\begin{array}{l} \prod_{i=1}^{N_c} \left[\begin{array}{l} \left(X_{i1}\left(p_{ij} + \left(\sum_{j=1}^{J-1} \mu_{1j}^t (p_{ij} - p_{ij})\right)\right) + (1-X_{i1})\left(q_{ij} + \left(\sum_{j=1}^{J-1} \mu_{1j}^t (q_{ij} - q_{ij})\right)\right)\right) \\ \times \left(X_{i2}\left(p_{ij} + \left(\sum_{j=1}^{J-1} \mu_{2j}^t (p_{ij} - p_{ij})\right)\right) + (1-X_{i2})\left(q_{ij} + \left(\sum_{j=1}^{J-1} \mu_{2j}^t (q_{ij} - q_{ij})\right)\right)\right) \end{array} \right] \\ + \prod_{i=1}^{N_c} \left[\begin{array}{l} \left(X_{i1}\left(p_{ij} + \left(\sum_{j=1}^{J-1} \mu_{2j}^t (p_{ij} - p_{ij})\right)\right) + (1-X_{i1})\left(q_{ij} + \left(\sum_{j=1}^{J-1} \mu_{2j}^t (q_{ij} - q_{ij})\right)\right)\right) \\ \times \left(X_{i2}\left(p_{ij} + \left(\sum_{j=1}^{J-1} \mu_{1j}^t (p_{ij} - p_{ij})\right)\right) + (1-X_{i2})\left(q_{ij} + \left(\sum_{j=1}^{J-1} \mu_{1j}^t (q_{ij} - q_{ij})\right)\right)\right) \end{array} \right] \end{array} \right) \tag{21}$$

Maximisation step

Here we choose the parameter vectors μ_1^{t+1} and μ_2^{t+1} using gradient optimisation as follows:

$$\mu_1^{t+1}, \mu_2^{t+1} = \underset{\mu_{1,1\ldots J-1}^{t+1}, \mu_{2,1\ldots J}^{t+1}}{\arg\max} \sum_{c=1}^{C}\sum_{i=1}^{N_c} \ln \left[\begin{array}{l} \hat{\gamma}_c \left(\begin{array}{l} \left(X_{i1}\left(\sum_{j=1}^{J} \mu_{1j}^{t+1} p_{ij}\right) + (1-X_{i1})\left(\sum_{j=1}^{J} \mu_{1j}^{t+1} q_{ij}\right)\right) \\ \times \left(X_{i2}\left(\sum_{j=1}^{J} \mu_{2j}^{t+1} p_{ij}\right) + (1-X_{i2})\left(\sum_{j=1}^{J} \mu_{2j}^{t+1} q_{ij}\right)\right) \end{array} \right) \\ + (1-\hat{\gamma}_c) \left(\begin{array}{l} \left(X_{i1}\left(\sum_{j=1}^{J} \mu_{2j}^{t+1} p_{ij}\right) + (1-X_{i1})\left(\sum_{j=1}^{J} \mu_{2j}^{t+1} q_{ij}\right)\right) \\ \times \left(X_{i2}\left(\sum_{j=1}^{J} \mu_{1j}^{t+1} p_{ij}\right) + (1-X_{i2})\left(\sum_{j=1}^{J} \mu_{1j}^{t+1} q_{ij}\right)\right) \end{array} \right) \end{array} \right] \tag{22}$$

Where C is the number of chromosome pairs. Again, we only estimate μ for the first $J-1$ populations:

$$\mu_1^{t+1}, \mu_2^{t+1} = \underset{\mu_{1,1\ldots J-1}^{t+1}, \mu_{2,1\ldots J-1}^{t+1}}{\arg\max} \sum_{c=1}^{C}\sum_{i=1}^{N_c} \ln \left[\begin{array}{l} \hat{\gamma}_c \left(\begin{array}{l} \left(X_{i1}\left(p_{ij} + \sum_{j=1}^{J-1} \mu_{1j}^{t+1}(p_{ij} - p_{ij})\right) + (1-X_{i1})\left(q_{ij} + \sum_{j=1}^{J-1} \mu_{1j}^{t+1}(q_{ij} - q_{ij})\right)\right) \\ \times \left(X_{i2}\left(p_{ij} + \sum_{j=1}^{J-1} \mu_{2j}^{t+1}(p_{ij} - p_{ij})\right) + (1-X_{i2})\left(q_{ij} + \sum_{j=1}^{J-1} \mu_{2j}^{t+1}(q_{ij} - q_{ij})\right)\right) \end{array} \right) \\ + (1-\hat{\gamma}_c) \left(\begin{array}{l} \left(X_{i1}\left(p_{ij} + \sum_{j=1}^{J-1} \mu_{2j}^{t+1}(p_{ij} - p_{ij})\right) + (1-X_{i1})\left(q_{ij} + \sum_{j=1}^{J-1} \mu_{2j}^{t+1}(q_{ij} - q_{ij})\right)\right) \\ \times \left(X_{i2}\left(p_{ij} + \sum_{j=1}^{J-1} \mu_{1j}^{t+1}(p_{ij} - p_{ij})\right) + (1-X_{i2})\left(q_{ij} + \sum_{j=1}^{J-1} \mu_{1j}^{t+1}(q_{ij} - q_{ij})\right)\right) \end{array} \right) \end{array} \right] \tag{23}$$

The full process is iterated over t until the change in parameters becomes negligible. We skip the first expectation step, assigning responsibilities ($\hat{\gamma}$) of 0.5 to all homologous chromosomal pairs save the first, which we set to 1. This is helpful because μ_1 and μ_2 are designated arbitrarily to the parents, and thus it avoids a nonidentifiability issue. SEs are calculated using the Hessian matrix produced by gradient optimisation during the maximisation step.

Equation 21 involves calculating the product of many probabilities, which results in extremely small numbers. To achieve this, we use a python script which can operate with arbitrary floating point precision.

The computation time is largely dependant on the ancestral similarity of the parents. Convergence for well-diverged parents typically occurs at three iterations, but can take longer when divergence is low.

Application to real and simulated data

To validate the methods, we first took 100 phased haploid genomes from unrelated individuals in Hapmap Phase III CEU (www.hapmap.org) and combined them with 100 unrelated YRI haploid genomes to create synthetic first-generation admixed individuals, which were analysed using both LEAPFrOG and LEAPFrOG EM. The source populations were East Asia, Europe and West Africa, represented by 170 JPT + CHB, 112 CEU and 113 YRI individuals. SNPs with genotyping rates of <98% across the populations were removed, leaving 994 200 SNPs at the analysis stage. As the parents of the synthetic individuals come from the same source data sets, both of these were excluded before calculating source-population allele frequencies for use in maximum likelihood estimation. We also analysed 100 first-generation admixed individuals with parents from Japan and China (86 JPT and 84 CHB individuals), simulated using the same SNPs.

We examined African, East Asian and European parental admixture in 83 African-American (ASW) individuals from the Hapmap phase III data set. The analysis was performed on 1 078 914 autosomal SNPs (1 011 119 when using phased data for LEAPFrOG EM), which exceeded a 98% genotyping rate. Real genome-wide SNP data contain dependencies between SNPs due to linkage disequilibrium (LD),

which violates an assumption made by the LEAPFrOG methods. Although it is possible to 'prune' markers from the data in order to reach some nominal low value of LD, here we are more interested in maximising the accuracy of the point estimates, at the expense of some underestimation of the SEs. We analysed the effect of using independent SNPs by comparing LEAPFrOG results for the ASW individuals with those for an 'LD pruned' version of the data described above, where 53 783 SNPs with pairwise $r^2 < 0.1$ were retained. The LD pruned data was also used to compare LEAPFrOG predictions of offspring admixture proportions for the ASW individuals with those from ADMIXTURE,[8] an alternative method, which estimates admixture proportions in the target individual only.

Phillips and colleagues designed a panel of 34 ancestrally informative SNPs from Hapmap and the 1000 Genomes Project.[2,9] With these, we predicted African and European admixture proportions in simulated individuals, using 225 Africans from Nigeria, Senegal, The Democratic Republic of Congo, Kenya, Somalia, The Central African Republic, Mozambique, South Africa and Namibia, and 278 Europeans from Spain, France, Scotland, Italy, Denmark and Russia. These genotypes were collated by the SNPforID consortium (http://www.snpforid.org). Two triallelic SNPs were removed, being incompatible with the likelihood method, plus a biallelic SNP, which has poor genotyping rates in casework,[1] leaving 31 for analysis. A total of 100 simulations of both first-generation admixture and 0% parental divergence (each parent having 50/50 African/European ancestry) were performed.

To explore the relationship between population differentiation and admixture predictability, we simulated two ancestral populations under a Balding–Nichols model[10] and then generated 100 individuals with first-generation admixture. This simulation was then repeated with 0% parental divergence, where the parents of the individuals are both 50% admixed between the two Balding–Nichols populations. Allele frequencies from the two populations were used to predict admixture and parental admixture from the offspring genotypes, and F_{st} between the ancestral populations was varied from 0 to 0.1 in increments of 0.005. A further simulation was performed where the true allele frequency, derived from the Balding–Nichols model, was replaced with one calculated from 400 simulated individuals (200 from each population) to approximate the sampling error that we would expect for real data sets. Simulations were for 80 000 independent SNPs, roughly equivalent to the number of independent regions in the human autosomal genome.

Further Balding–Nichols simulations were performed to assess reliability across various parental divergences and number of SNPs. Two populations were simulated with an F_{st} of 0.15, and the target sample had 50% admixture from each. We recorded the mean squared error between the estimated and simulated parameters across 1000 simulated individuals, under various parental admixture proportions (mean across parents always 0.5/0.5) and numbers of SNPs. We also calculated the proportion of times that the 95% confidence intervals for the estimates covered the simulated parameter value (coverage probability).

RESULTS

With genome-wide SNP data, the LEAPFrOG method gave accurate prediction results for YRI/CEU synthetic first-generation admixture (Figure 1a, Supplementary Table 2), making it possible to reliably infer admixture proportions in both the unobserved parents and the genotyped target individual. For all the simulations, East Asian ancestry was correctly found to be low (mean 0.006), and African and European admixture proportions approximately equal (mean 0.495, 0.499), but highly divergent in terms parental origin. Results from the LEAPFrOG EM method were similar (results not shown).

JPT/CHB first-generation admixture was harder to predict (Figure 1b, Supplementary Table 3). Target individual admixture proportions were correctly distributed around 0.5 but variation in these estimates was greater than in the YRI/CEU results. Parental predictions were highly variable with LEAPFrOG, although LEAPFrOG EM returned more consistent estimates, with mean CHB admixture of 0.767 (0.766–0.768) in the first parent and 0.224 (0.223–0.225) in the second (full results not shown).

The African-American ASW individuals (Figure 1c, Supplementary Table 4) demonstrated admixture proportions approximately in line with prior expectations[11] (mean African admixture of 0.773 in target individuals), and a range of parental divergence. The mean difference in African admixture between parents was 0.118, and this increased to 0.172 when using LEAPFrOG EM (Supplementary Figure 1).

The ASW set contains 11 trio sets, allowing a comparison to be made between the parental admixture proportions predicted from the target and parent genotypes (Figure 2, Supplementary Figure 2). LEAPFrOG EM results had marginally lower variance around the expected (full concordance between estimates) than LEAPFrOG (0.004 *versus* 0.005).

Comparing LEAPFrOG results for the LD pruned and unpruned ASW data set showed that African admixture proportions for the offspring were on average 2.1% (1.9–2.1) higher for the latter (Supplementary Figure 4). We also estimated African admixture proportions using the ADMIXTURE program[8] and found that they were on average 1.7% (0.69–1.71) higher than when using LEAPFrOG (Supplementary Figure 5). These deviations are relatively small, and we consider the reasons why they occur in the Discussion.

We investigated LEAPFrOG prediction results from 31 AIM SNPs for simulated admixture between Africa and Europe (Figure 3, Supplementary Tables 5 and 6). Results were quite accurate for first-generation admixture (parents either fully 'African' or 'European') but highly variable when both parents were simulated with identical admixture proportions.

The Balding–Nichols simulation results across various levels of population divergence (Supplementary Figure 3) showed that first-generation admixture could be accurately inferred using LEAPFrOG for F_{st} greater than approximately 0.02. Non-divergent parental admixture is more difficult to predict, and even distantly related populations ($F_{st} = 0.1$) fail to provide mean admixture proportions of 0.50 in each parent. The simulations across various parental divergences and numbers of SNPs (Table 1) show that 60 000 SNPs provide excellent accuracy in terms of mean squared error for parental divergence, but good results can be obtained with fewer markers, particularly at higher levels of divergence. Confidence interval coverage was poor at the boundary values but good otherwise, with the true parameter value within the 95% confidence intervals in approximately 95% of simulations. Parameter estimates and coverage for offspring admixture were highly accurate (Supplementary Table 1), though confidence intervals appeared slightly conservative when parental divergence was high.

DISCUSSION

The identification of ethnic or geographic origins has proved useful in some forensic investigations.[1] Our method takes this one step further by inferring the origins of the unknown DNA sample's parents, something not attempted by any of the existing methods for admixture estimation (eg Frappe,[12] Structure[13] and ADMIXTURE[8]). Different admixture proportions from multiple source populations are allowed in either parent, thus allowing for complex admixture patterns rather than 'all-or-nothing' classification to a single-source population.

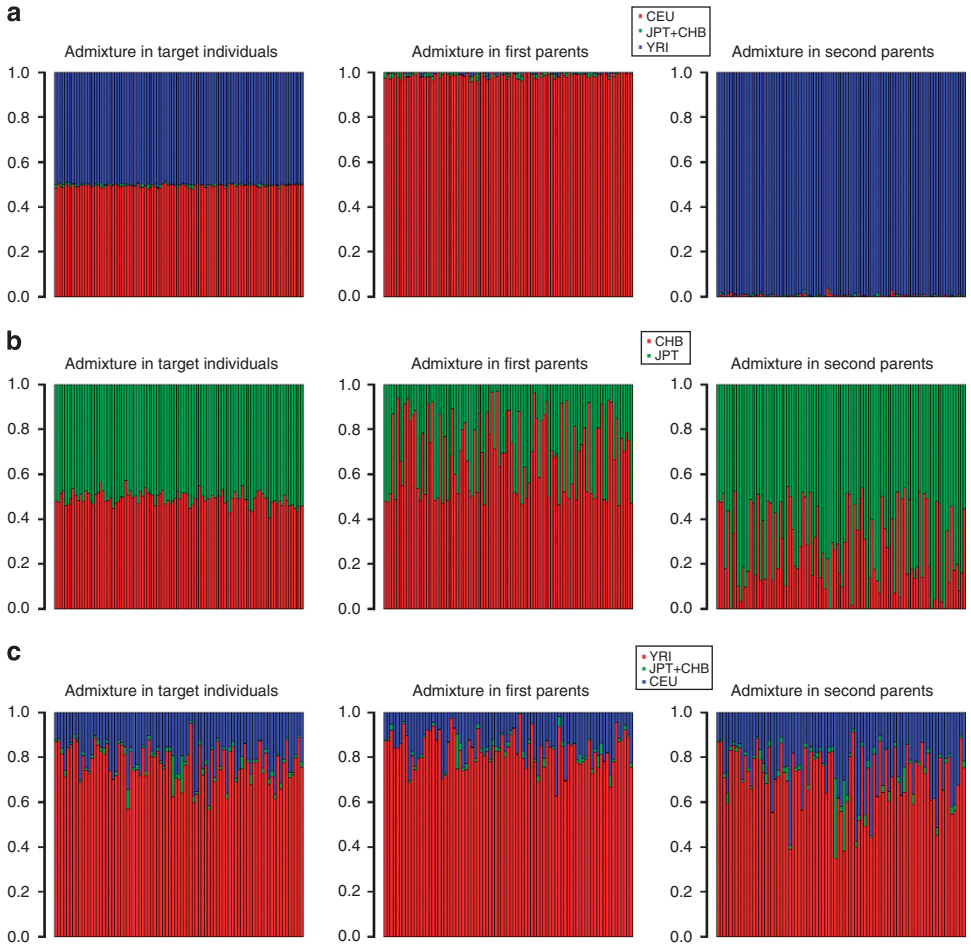

Figure 1 (**a**, **b**) LEAPFrOG-estimated admixture proportions in synthetic individuals with first-generation admixture. Each vertical bar shows the admixture proportions for one individual, and the order of bars in each panel is consistent with the parental/offspring relationships. (**a**) 100 individuals with European/West African (CEU/YRI) admixture using 994,200 SNPs from Hapmap phase III. (**b**) 100 individuals with Japanese/Chinese (JPT/CHB) admixture using the same SNPs. (**c**) LEAPFrOG parameter estimation for 83 African-American (ASW) individuals using 1,078,914 SNPs from Hapmap phase III. For full model estimates see Supplementary Tables 2–4.

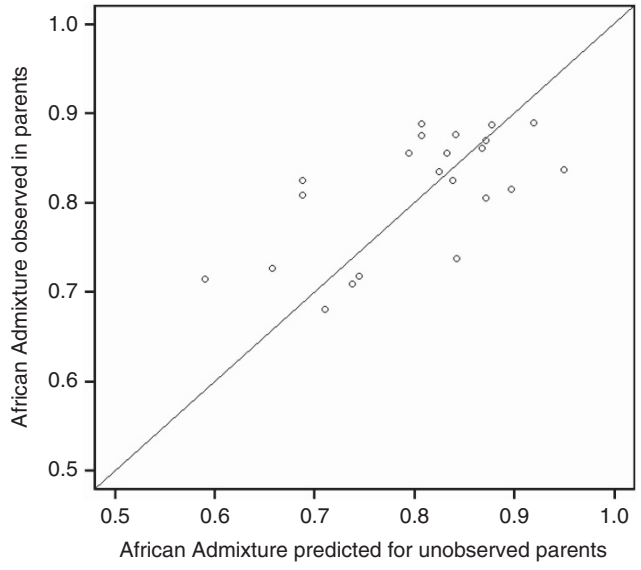

Figure 2 Comparison of LEAPFrOG parental admixture proportions estimated from offspring genotypes and directly from parental genotypes in 11 African-American trios genotyped at 1,078,914 SNPs.

Our results demonstrate the efficacy of LEAPFrOG and LEAPFrOG EM for well-differentiated populations, such as Africa and Europe, using genome-wide SNP data; particularly when parents are highly divergent for ancestry. LEAPFrOG EM gave more consistent parental admixture estimates than LEAPFrOG, underlining the potential utility of pre-phased data. However, neither method was able to accurately identify first-generation admixture events between China and Japan, showing that parental ancestry prediction may be hard to achieve when the source populations are closely related.

Comparison of parental admixture predictions from their observed genotypes with those from the genotypes of their offspring, using the 11 African-American trios, suggest a reasonable ability to estimate parental admixture proportions using offspring data alone (Figure 2, Supplementary Figure 2). LEAPFrOG estimates for all 83 African-Americans reveal extensive diversity in both offspring admixture and parental admixture divergence. Three individuals were predicted to have a considerable East Asian component (Figure 1c), the most likely explanation of which is Native American ancestry, as this population would be most similar to JPT/CHB among the source groups. LEAPFrOG EM detected greater African admixture divergence between parents than LEAPFrOG, demonstrating the advantages conferred by phase information when the admixture is complex.

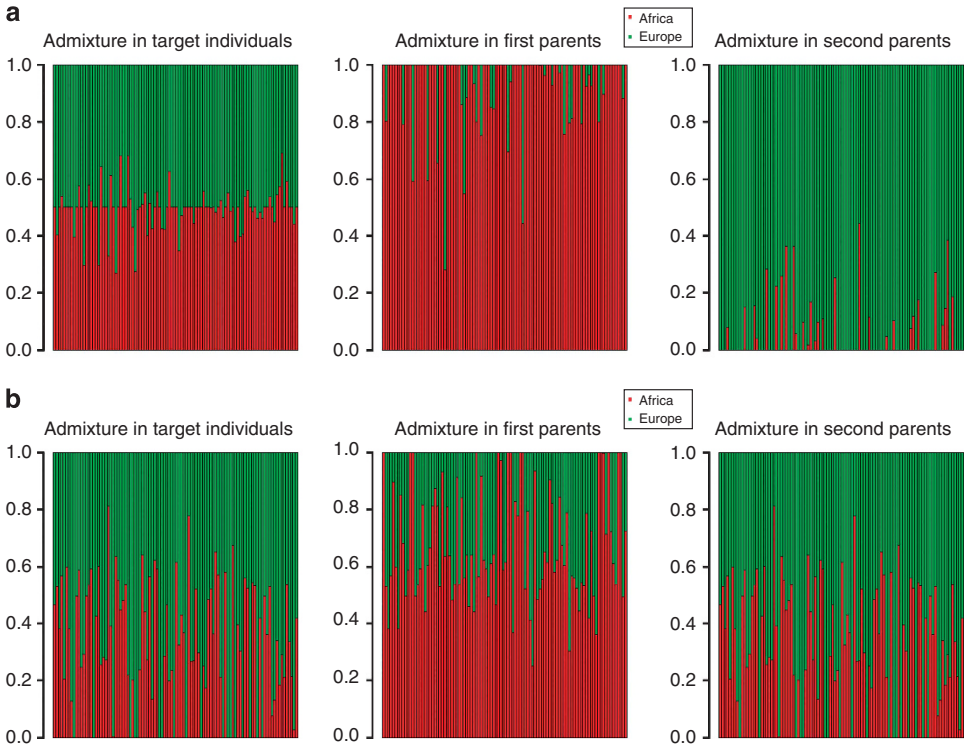

Figure 3 LEAPFrOG estimates of offspring and parental admixture proportions from 31 ancestrally informative markers, for (**a**) 100 simulated individuals with first-generation admixture between Africa and Europe; (**b**) 100 simulated individuals whose parents have 50% African and 50% European admixture. Each bar shows the admixture proportions for one individual. For full model estimates see Supplementary Tables 5 and 6.

Table 1 Mean squared error (MSE) and confidence interval coverage probability of *D* estimates for individuals with equal admixture from two simulated source populations (50% from each)

D/SNPs	*2000*	*5000*	*20 000*	*60 000*	*100 000*	F_{st}
MSE						
0.5	0.028	0.017	0.009	0.005	0.004	0
0.625	0.017	0.012	0.005	0.003	0.002	0.01
0.75	0.016	0.008	0.001	0	0	0.04
0.875	0.007	0.002	0.001	0	0	0.087
1	0	0	0	0	0	0.15
Coverage probability						
0.5	0.812	0.825	0.801	0.832	0.823	0
0.625	0.845	0.879	0.921	0.941	0.947	0.01
0.75	0.886	0.925	0.952	0.964	0.944	0.04
0.875	0.885	0.925	0.931	0.946	0.938	0.087
1	1	1	1	1	1	0.15

Parental divergence was varied between 0.5 (no parental divergence) and 1 (maximal parental divergence), denoted by *D*, and the number of SNPs used for estimation was varied between 2000 and 100 000. The F_{st} between the parents is provided – when divergence is maximal this is equal to the F_{st} between the source populations.

However, phased data in these applications should be treated with care, as existing statistical phasing methods assume a simple population history at odds with the complex admixture patterns being investigated here. This obstacle should be removed by the advent of experimentally determined phase from single-molecule sequencing technology.[14]

Balding–Nichols simulations indicated that accurate LEAPFrOG prediction of first-generation admixture with genome-wide microarray panels may be possible with F_{st} as low as 0.01, the same as between Latvia and Spain,[15] two widely spaced European countries. The sampling error ($n = 200$ for each source population) had a considerable effect on prediction when $F_{st} < 0.03$, suggesting that our results for JPT/CHB (combined $n = 170$) admixture might have been more accurate if larger samples were available. The F_{st} between these populations is 0.0069.[16]

Our simulations showed that the mean squared error for the *D* parameters depends largely on the number of SNPs and extent of parental divergence, with low divergence requiring more markers to accurately estimate. The confidence interval coverage (the proportion of times the confidence intervals capture the simulated parameter value) is poor at the parameter boundaries but otherwise good. D_1 cannot take values < 0.5 or > 1 (see Methods), so the distribution of estimates is not normal here, which explains why the confidence intervals do not behave as expected. The anti-conservative confidence intervals, which occur when parental divergence is 0%, are the result of a bimodal distribution for the parameter estimates, which disappears when an extremely high number of independent SNPs (500 000) is simulated. Both mean squared error and confidence interval coverage are generally excellent for *m*, the offspring admixture proportions (Supplementary Table 1), although confidence intervals appear somewhat conservative when parental divergence is maximal. We have also implemented a Bayesian model with user customisable Dirichlet priors for parental admixture, which can be used to generate credible intervals, but this method does not scale to full genome-wide data.

Our methods do not allow one to determine which of the two inferred parental admixture patterns belongs to the mother and which to the father. However, it may be possible to cross-compare with mtDNA or Y-chromosome data to shed light on this. Furthermore, if the DNA sample is male then the X-chromosome is guaranteed to come from the mother, and so separate analysis of this chromosome may allow one to distinguish the maternal and paternal admixture pattern.

Both LEAPFrOG and LEAPFrOG EM assume independence between markers, in contrast with real genome-wide data, which contain dependencies among markers due to LD. Using all available data, while violating the assumption of independence, provides the best point estimates at the expense of underreporting the width of the confidence intervals. Users should decide which of these quantities they consider to be more important and filter the data accordingly (for example by 'pruning' the data for LD). Using African-American individuals, we demonstrated that the average discrepancy between African admixture estimates from LD 'pruned' and 'unpruned' data was 2.1%. The systematic difference is because the uncertainty in the largest admixture proportion is predominantly one-sided, as estimates cannot be >1.

The innovative aspect of LEAPFrOG lies in its ability to predict admixture in the unobserved parents of a target DNA sample, but it is nevertheless of interest to compare the target individual's estimated admixture proportions with those obtained with an existing method such as ADMIXTURE.[8] Our comparison revealed small but consistent discrepancies, which we attribute to differences in the approach taken to modelling the underlying source populations. LEAPFrOG assumes that the allele frequencies in these are known, whereas ADMIXTURE takes a clustering approach in which they are re-estimated based on the admixture across all individuals. LEAPFrOG could be extended to treat the source populations in the same way as ADMIXTURE. Although this could be advantageous in some circumstances, we found only small differences in the admixture estimates when both methods were applied to African-American individuals.

Prediction of parental admixture proportions was not accurate using a 34-AIM (31 after quality control) PCR multiplex panel. The increase in heterozygosity one sees after admixture events is subtle and requires large numbers of markers to exploit. The ability to make inferences about parental ancestry therefore seems limited to situations where genome-wide data, for example from SNP microarrays or whole-genome sequencing, are available to investigators. It should be possible to extend the methods to facilitate the analysis of STRs, but at present they are limited to biallelic markers. Although the ability to collect genome-wide SNP data for the target individual depends largely on the quantity and quality of DNA, there must also be a sufficiently large intersection with SNPs genotyped at the desired source populations. Furthermore, many individuals are needed at each population to accurately estimate allele frequencies. Most applications will therefore be limited to cases where genome-wide data is publicly available for these populations. The best current resources are the International Hapmap and 1000 Genomes projects.

Our results, as examples of the detailed information attainable from DNA samples, advocate the integration of high-throughput technology into forensic casework. The methods described in this paper are implemented as an R package, LEAPFrOG, available from CRAN (http://cran.r-project.org/web/packages/LEAPFrOG). This includes a Bayesian method not presented in detail here.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank the Biotechnology and Biological Sciences Research Council (BBSRC) for funding. We also thank Chris Phillips, Barbara Daniel, Denise Syndercombe Court and members of the Statistical Genetics Unit for advice and discussion.

1 Phillips C, Prieto L, Fondevila M *et al*: Ancestry analysis in the 11-M Madrid bomb attack investigation. *PLoS One* 2009; **4**: e6583.
2 Phillips C, Salas A, Sanchez JJ *et al*: Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Sci Int Genet* 2007; **1**: 273–280.
3 Jacobson P: Investigation: Stalker in the suburbs. *The Sunday Times* 2005.
4 King TE, Parkin EJ, Swinfield G *et al*: Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy. *Eur J Hum Genet* 2007; **15**: 288–293.
5 Schneider PM, Balogh K, Naveran N *et al*: Whole genome amplification—the solution for a common problem in forensic casework? *Int Congr Ser* 2004; **1261**: 24–26.
6 Wahlund S: Composition of populations and correlation appearances viewed in relation to the studies of inheritance. *Hereditas* 1928; **11**: 65–106.
7 Dempster AP, Laird NM, Rubin DB: Maximum likelihood from incomplete data via em algorithm. *J Roy Stat Soc B Met* 1977; **39**: 1–38.
8 Alexander DH, Novembre J, Lange K: Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009; **19**: 1655–1664.
9 Durbin RM, Abecasis GR, Altshuler DL *et al*: A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
10 Balding DJ, Nichols RA: A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 1995; **96**: 3–12.
11 Price AL, Tandon A, Patterson N *et al*: Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet* 2009; **5**: e1000519.
12 Tang H, Peng J, Wang P, Risch NJ: Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* 2005; **28**: 289–301.
13 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945–959.
14 Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H: Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* 2009; **4**: 265–270.
15 Nelis M, Esko T, Magi R *et al*: Genetic structure of Europeans: a view from the North-East. *PLoS One* 2009; **4**: e5472.
16 Heath SC, Gut IG, Brennan P *et al*: Investigation of the fine structure of European populations with applications to disease association studies. *Eur J Hum Genet* 2008; **16**: 1413–1429.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)