Testing hypotheses about the genetic background of individuals and populations using ancestry-informative markers

Paul M. McKeigue, David O'Donnell

Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland

paul.mckeigue@ucd.ie

Clive J. Hoggart

Department of Epidemiology & Public Health, Imperial College, St Mary's Campus Norfolk Place, London W2 1PG, UK

Ray Fysh

Forensic Science Service, 109 Lambeth Road, London SE1 7LP, UK

Mark D. Shriver

Department of Anthropology, Penn State University, 409 Carpenter Building, State College, PA 16801

Gerome Breen, Camila Guindalini

Institute of Psychiatry, King's College London SE5 8AF, UK

Homero Vallada

Department of Psychiatry, Universidade de São Paulo, Brazil

Tony N. Frudakis

DNAPrint Inc, Sarasota, FL 34236

ABSTRACT

Testing hypotheses about the genetic background of individuals and populations is equivalent to comparing models of parental admixture. Formally, the strength of evidence favouring one model over another is given by the Bayes factor (ratio of marginal likelihoods). We have extended the ADMIXMAP program to calculate Bayes factors between different models of parental admixture, and demonstrate applications to inferring individual ancestry, inferring the demographic origin of an unknown individual, classifying individuals as unadmixed or admixed, and inferring the number of subpopulations ancestral to an admixed population. From comparisons based on Bayes factors it was possible to infer parental ancestry of each individual, except where the information

content of the marker panel was inadequate. Inference of individual genetic background from parameter estimates was often misleading: in a Brazilian population, estimated admixture proportions were as high as 25% in some individuals for whom Bayes factors favoured a model with no admixture. In a Hispanic-American population, a widely-used approximate method for inferring the number of ancestral subpopulations was shown to yield misleading results, in comparison with direct calculation of Bayes factors. It is thus both feasible and necessary to calculate Bayes factors when testing hypotheses about the genetic background of individuals and populations.

Subject headings: admixture, ancestry, forensic

1. Corresponding author

Paul M McKeigue
University College Dublin
Belfield, Dublin 4, Ireland
paul.mckeigue@ucd.ie
phone +353 1 716 6952
fax +353 1 716 6950

2. INTRODUCTION

Using genotype data to model genetic admixture or population stratification has several applications, including admixture mapping to localize genes underlying ethnic variation in disease risk (1), controlling for hidden population stratification in genetic association studies (2), and inferring the genetic background of a single individual. To allow for linkage and for uncertainty in allele frequencies, Bayesian computationally-intensive methods are required. Three such programs are available, all based on the same statistical model for variation of ancestry on admixed gametes: STRUCTURE (3), ADMIXMAP (4) and ANCESTRYMAP (5). Tests of hypotheses about the genetic background of an individual or a population are equivalent to comparing different models of admixture. However, as others have noted (6), the methods for model comparison implemented in these programs have limitations, relying either on model diagnostics (7) or on global tests of "fit" penalized by "complexity" (6). In a Bayesian framework, comparison between models is based on the marginal likelihood of each model given the data (8). However the marginal likelihood is rarely calculated when modeling the genetic structure of populations, mainly because of computational difficulties. In this paper we describe methods implemented in the program ADMIXMAP (4) to test hypotheses about the genetic background of an individual or population by calculating the marginal likelihood for any specified model of admixture, given genotype data on one or more admixed individuals. We demonstrate application of these methods to practical problems, including inferring the ancestry or demographic background of an unknown individual, classifying individuals as admixed or unadmixed, and inferring the number of subpopulations ancestral to an admixed population. These applications illustrate how hypothesis tests can be formulated in terms of model comparisons, how these Bayesian hypothesis tests differ from classical inference based on the "best-fit" model, and that a widely-used method for inferring the number of ancestral subpopulations by calculating an approximation to the marginal likelihood may yield misleading results.

3. MATERIAL AND METHODS

3.1. Individuals and markers

The datasets used in this study come from four sources. In the first dataset, the DNAPrint panel of 176 SNPs described below was typed on a family of five individuals: parent 1 (author TNF) who reports Greek ancestry, parent 2 (his wife) who reports Mexican ancestry, and their three offspring. In the second dataset, the DNAPrint panel was typed on a sample of 142 individuals resident in five Caribbean countries and 41 individuals of Caribbean origin working in the Metropolitan Police Service (London, UK), who donated DNA samples to provide a database for inference of genetic background from genotype data, in response to an appeal for assistance with a criminal investigation. Questionnaire data on origin of each grandparent were obtained, and the 142 participants were grouped into five categories based on country of origin of both parents: Jamaica (21), St Lucia (31), St Vincent (28), Trinidad (29), Guyana (31). The third dataset was a case-

control study of cocaine addiction among residents of São Paulo, Brazil, typed at 72 loci from the Penn State panel of ancestry-informative marker loci supplemented with 3 other loci in candidate genes (9). The fourth dataset has been described previously (7): 446 Hispanic American individuals resident in San Luis Valley, Colorado were typed at 22 marker loci chosen to be informative for three-way admixture between European, Native American and west African ancestry.

The DNAPrint marker panel (DNAPrint, Sarasota, FL) consists of 176 autosomal single-nucleotide polymorphisms (SNPs) selected from some 27000 SNPs for which allele frequency data in African-American, European-American, and East Asian populations were available on the SNP Consortium database in 2002. Markers were selected based on TSC allele frequency differentials and suitability for multiplex amplification panels. Allele frequencies were re-estimated in a sample of 70 west Africans, 59 Native Americans (Mayans from southern Mexico), 58 Europeans (European-American college students), and 59 East Asians (Coriell Institute panel augmented with Chinese-Americans). Allele frequencies in other populations were estimated from 56 South Asians, 14 Pacific islanders, and 20 individuals of Middle Eastern origin from the Coriell repository.

Linkage map positions of all markers were estimated by interpolation between markers in the Rutgers Combined Linkage-Physical Map (10). Before interpolation, the linkage map position estimates in the Rutgers map were smoothed to use prior information from physical map distances over short intervals where little information is available from the linkage map. An online resource for calculating smoothed estimates of map positions is available. Any subsequence of loci separated by less than 0.1 cM was modeled as a "compound locus" for which haplotype assignments and haplotype frequencies were modeled as described previously. There were 9 such pairs in the DNAPrint marker panel and one in the marker panels used for the San Paulo and San Luis Valley studies.

3.2. Statistical model for admixture

For each gamete transmitted from an admixed parent with ancestry from K subpopulations, the model specifies a vector $\boldsymbol{\mu}$ for the proportions of the parent's genome that have ancestry from these subpopulations. The stochastic variation of states of ancestry on chromosomes of this gamete is modeled as generated by K independent Poisson arrival processes with intensities $\rho\boldsymbol{\mu}$ per morgan, where ρ is the sum of the intensities of the arrival processes. The arrival rate parameter ρ can be interpreted as the effective number of generations back to unadmixed ancestors (3). To model data on a single individual, we specify for each gamete a uniform Dirichlet prior on parental admixture proportions $\boldsymbol{\mu}$, and a gamma prior on the arrival rate parameter ρ with mean 6 and variance 24. To model data on a a sample of individuals, we specify a hierarchical model for $\boldsymbol{\mu}$ and ρ. The prior distribution of gamete admixture proportions in the population is specified as Dirichlet, with Gamma(1,1) priors on the elements of the parameter vector (equivalent to specifying a flat prior on the mean admixture proportions). The distribution of the arrival rate parameter ρ over gametes in the population is specified as gamma with shape parameter 6 and a gamma (5, 4) prior on the rate parameter (equivalent to a gamma-gamma prior on ρ with mean 6 and

variance 14). These priors on the arrival rate parameter ρ are based on previous studies of admixed populations in the Americas which have typically yielded estimates of about 6 for the mean arrival rate, and on the demographic history of the Americas (from which we can infer that the maximum possible time back to unadmixed ancestors is about 20 generations). For the São Paulo dataset, the Dirichlet parameters for the distribution of individual admixture proportions in the population were estimated initially by modeling the entire dataset. These estimates were used to specify the prior on admixture proportions when fitting a three-way admixture model to each individual separately.

Allele frequencies were specified with Dirichlet prior distributions. At each locus, the Dirichlet parameter vector was obtained by adding 0.5 to the observed allele counts in samples from unadmixed modern descendants of the populations that contributed to the admixed population. This is equivalent to updating a Dirichlet $(0, 5, \ldots, 0.5)$ prior distribution with the likelihood given the observed allele counts. For diallelic loci, a Dirichlet $(0.5, 0.5)$ prior is the standard uninformative "reference" prior (11).

At each compound locus consisting of two or more markers, the posterior distribution of haplotype frequencies was generated by running ADMIXMAP with samples from each modern unadmixed descendants of each ancestral continental group. This posterior distribution was approximated by a Dirichlet distribution with the same mean and dispersion, and the parameters of this Dirichlet distribution were used to specify a Dirichlet prior on haplotype frequencies for subsequent studies of admixed individuals.

Using the ADMIXMAP program (7; 4) to sample the posterior distribution of admixture proportions $\boldsymbol{\mu}$ and sum-intensities ρ by Markov chain Monte Carlo (MCMC) simulation, we computed the marginal likelihood of the model given the data as described below. To search for a mode of the joint posterior density of admixture proportions $\boldsymbol{\mu}$ in the softmax (multiple logit) basis and arrival rates ρ on a log scale on each gamete, a stochastic EM algorithm was constructed. In each E step, the posterior expectations of the sufficient statistics described in Appendix C (jump indicators and total number of arrivals) were evaluated by running the MCMC sampler. In each M step, the mode of the conjugate distribution conditional on the prior and the sufficient statistics was calculated. Updates were accepted only if they increased the posterior density.

3.3. Information content of the marker panel

To evaluate the adequacy of a marker panel for measuring individual admixture, we consider a single gamete formed by admixture between two subpopulations, and define the admixture proportion μ as the proportion of the parent's genome that has ancestry from the first subpopulation. We calculate the admixture information content of the marker panel as minus the expectation of the second derivative of the log-likelihood function. This is the Fisher information about μ, asymptotically equal to the variance of the maximum-likelihood estimator. Thus to estimate two-way admixture proportions at $\mu = 0.5$ with a standard error of less than 0.1 requires that the Fisher

information is greater than 100.

The information content for admixture proportions μ of an infinitely dense panel of ancestry-informative markers is $(C + \rho L) / [\mu (1 - \mu)] - 2\rho L$, where C is the number of autosomes, and L is the length of the genome in morgans (see Appendix A for derivation). The information content for $\log \rho$ is $2\mu (1 - \mu) \rho L$. Thus for $C = 22$, $L = 37.75$, $\mu = 0.5$, and $\rho = 6$ (the value estimated for African-American populations) (4), the information content of a infinitely dense marker panel is 541 for admixture proportions μ (equivalent to 135 fully-informative unlinked loci) and 113 for $\log \rho$. Thus even with a maximally informative marker panel and no uncertainty about phase, the standard errors of estimates of gamete admixture proportions and $\log \rho$ cannot be reduced to much less than about 0.05 and 0.1 respectively. The ancestry information content of a single marker can be measured by the f-value (12), defined as the ratio of the information extracted by the marker at $\mu = 0.5$ to the information that would be extracted by a perfectly informative marker locus at which different alleles are fixed in each ancestral subpopulation. The information content at $\mu = 0.5$ of a panel of m unlinked marker loci with information content $f_1, \ldots, f_m$ is $4 \sum f_i$. The admixture information content of a panel of linked markers can be evaluated by averaging the second derivative of the log-likelihood (calculated as described in Appendix B) over simulated datasets.

The ability of a marker panel to detect or exclude small degrees of admixture can be measured by the Kullback-Leibler divergence: the expected log-likelihood ratio in favour of the true model compared with an alternative (false) model. For a panel of linked markers, the expected log-likelihood ratio can be evaluated by averaging over simulated datasets, using the standard hidden Markov model algorithm (13) to calculate the likelihood at given values of the parameters μ and ρ. For these simulations, the arrival rate *rho* was specified as 6 per morgan. each pair of subpopulations, we evaluated the admixture information content, and the expected log-likelihood ratio for a true model of no admixture ($\mu = 0$) versus an alternative model of admixture with $\mu = 0.125$ by averaging over 1000 simulated datasets. The number of simulations was set by increasing the number of runs until stable estimates of admixture information content and expected log-likelihood ratio were obtained.

3.4. Calculation of the marginal likelihood (evidence) for a model

In a Bayesian framework, tests of hypotheses are based on comparing for each model M the probability $P(y \mid M)$ of the observed data y. The terms "prior predictive probability" and "marginal likelihood" have been used for the probability $P(y \mid M)$; more recently the term "evidence" has been proposed (14; 15) and will be used here. The ratio of evidence values between two models is the Bayes factor, reported here in units of natural logarithms (nats). A log Bayes factor of 2 nats between two hypotheses is approximately equivalent to an eightfold ratio of posterior to prior odds in favour of the most likely model.

To test hypotheses about whether one or both of an individual's parents have admixture from two or more continental groups, we compare the evidence values of models with and without such admixture, given the individual's observed genotype data. To calculate the evidence $P\left(y \mid M\right)$ it is not feasible to evaluate directly the expectation of the likelihood $P(y \mid \boldsymbol{\theta}, M)$ for the model parameters $\boldsymbol{\theta}$ over the prior distribution $P\left(\boldsymbol{\theta} \mid M\right)$. For models of the ancestry of the two gametes in a single individual, we have used the candidate's formula (16; 17) to calculate the evidence from Markov chain Monte Carlo output. The candidate's formula rewrites Bayes's theorem to express the evidence $P\left(y \mid M\right)$ for a model M given observed data y as

$$P\left(y \mid M\right) = P\left(y \mid \boldsymbol{\theta}^*, M\right) \times \frac{P\left(\boldsymbol{\theta}^* \mid M\right)}{P\left(\boldsymbol{\theta}^* \mid y, M\right)}$$

where $\boldsymbol{\theta}^*$ is any fixed value of the parameter vector. If $\boldsymbol{\theta}^*$ is the posterior mode, the evidence can be expressed as the product of the best-fit likelihood $P\left(y \mid \boldsymbol{\theta}^*, M\right)$ and the Occam factor $P\left(\boldsymbol{\theta}^* \mid M\right) / P\left(\boldsymbol{\theta}^* \mid y, M\right)$ which is the ratio of prior to posterior density at the posterior mode. The Occam factor can be interpreted as a penalty for "complexity" (15).

For a model of admixture in a single individual, the parameter vector $\boldsymbol{\theta}$ comprises the gamete admixture proportion vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, the arrival rate parameters ρ_1, ρ_2 and the array of allele frequencies $\boldsymbol{\phi} = (\boldsymbol{\phi}_{11}, \ldots, \boldsymbol{\phi}_{TK})$ for K ancestral subpopulations at T loci. The likelihood $p\left(y \mid \boldsymbol{\theta^*}, M\right)$ is calculated from the standard sum-product algorithm for a hidden Markov model (13). The prior density $P(\boldsymbol{\theta}^* \mid M)$ is evaluated directly as the product at $\boldsymbol{\theta}^*$ of the prior densities specified for $\boldsymbol{\mu}, \boldsymbol{\rho}$ and $\boldsymbol{\phi}$. The posterior density $P(\boldsymbol{\theta}^* \mid y, M)$ is computed by averaging over the posterior distribution as described in Appendix C.

With a sample of individuals from the population, we have a hierarchical model, in which prior distributions are specified for the parameters of the Dirichlet distribution of admixture proportions and the gamma distribution of arrival rates on gametes in the population. As the prior on the individual parameters is not available in closed form, it is not computationally feasible to use the candidate's formula to evaluate the evidence $P\left(y \mid M\right)$ (6). Instead we use the method of "thermodynamic integration" (18). We extend the Markov chain Monte Carlo sampling algorithm to sample at "coolness" (inverse temperature) λ from a modified posterior density $P_\lambda\left(\boldsymbol{\theta}\right) \propto \left[P\left(y \mid \boldsymbol{\theta}, M\right)\right]^\lambda P\left(\boldsymbol{\theta} \mid M\right)$ in which the likelihood has been "annealed" by raising it to the power λ. At coolness 0, we are sampling from the prior, and at coolness 1 we are sampling the unmodified posterior. If the prior is reparameterized as a uniform distribution over a large number of discrete microstates indexed by x, this modified posterior density can be written in the form of a Gibbs distribution.

$$P\left(x\right) = \frac{e^{-\lambda E(x)}}{Z\left(\lambda\right)} \tag{1}$$

where the "energy" $E\left(x\right)$ is minus $\log P\left(y \mid x\right)$, and the normalizing constant $Z\left(\lambda\right)$, equal to $\sum_x e^{-\lambda E(x)}$, is the partition function (number of microstates accessible at coolness λ). This formulation of the inference problem allows us to use tools from statistical mechanics, in which the Gibbs distribution arises as the distribution of energy levels of a system in a heat bath at temperature

$1/\lambda$. Thus the mean energy $\langle E_\lambda \rangle$ at coolness λ is equal to minus $\mathrm{d} \log Z(\lambda) / \mathrm{d}\lambda$ (15), and the log evidence is

$$\log P(y \mid M) = \log Z(1) - \log Z(0) = \int_0^1 \frac{\mathrm{d} \log Z(\lambda)}{\mathrm{d}\lambda} = -\int_0^1 \langle E_\lambda \rangle \, \mathrm{d}\lambda$$

At coolness λ, the mean energy $\langle E_\lambda \rangle$ can be evaluated as the mean of minus the log-likelihood $P(y \mid \boldsymbol{\theta}, M)$ over samples from the modified posterior density $P_\lambda(\boldsymbol{\theta})$. The integral over λ can be evaluated by averaging $\langle E_\lambda \rangle$ over values of λ between 0 and 1. The difference between the posterior mean of the log-likelihood $\log P(y \mid \boldsymbol{\theta}, M)$ and the log evidence $\log P(y \mid M)$ is the information (reduction in entropy) obtained by fitting the model to the data. The information, like the Occam factor, can be interpreted as a penalty for complexity, and a plot of $\langle E_\lambda \rangle$ against λ can be interpreted visually as showing how the fit of the model (measured by the mean energy at $\lambda = 1$) is penalized by the adaptation of model parameters to the data (represented by the fall in energy as the system is cooled).

The implementation of annealed sampling in ADMIXMAP is described in Appendix D. To test hypotheses about the country of origin of an unknown individual i, we use genotype data y_s from individuals from each country of origin to generate draws from the posterior distribution $P(\boldsymbol{\theta} \mid y_s)$ of model parameters, and re-use these values in a model that specifies them as draws from the prior distribution $P(\boldsymbol{\theta_i} \mid M)$ on the corresponding parameters $\boldsymbol{\theta_i}$ for the individual under study. At each coolness λ we evaluate the mean energy $\langle E_\lambda \rangle$ as the expectation of minus $\log P(y \mid \boldsymbol{\theta_i})\lambda$ over the modified posterior $P_\lambda(\boldsymbol{\theta_i}) \propto [P(y \mid \boldsymbol{\theta_i}, M)]^\lambda P(\boldsymbol{\theta_i} \mid M)$.

4. RESULTS

4.1. Calculated information content of the DNAPrint marker panel

Table 1 shows the Fisher information for gamete admixture proportions of the marker panel used in this study, calculated from simulations of 1000 gametes formed by two-way admixture with $\mu = 0.5$ and $\rho = 6$. For this analysis, 3 of the 176 SNPs for which allele frequency data were incomplete were excluded, and the model assumed no allelic association between SNPs (which would require haplotypes to be modeled). This shows that the marker panel has adequate information content (> 50) to measure two-way admixture between west Africans and other groups, or between Europeans and East Asians, but not to measure two-way admixture between groups originating on the western side (Europeans, Middle-Eastern and South Asia) of the Eurasian land mass or between groups originating on the eastern side (East Asians, Pacific islanders and Native Americans). For African versus non-African admixture, the information content of the marker panel is of the order of 10% of the theoretical maximum. For admixture between more closely-related populations such as Europeans and South Asians, the admixture information content of the marker panel is lower. If admixture information content of the marker panel were calculated without allowing for linkage, the calculated values would be about 30% higher than the values shown in table 1 which allow for

linkage.

We evaluated also the ability of the marker panel to detect or exclude small degrees of admixture from another continental group. Table 2 shows the expected log-likelihood ratio in favour of the true hypothesis (gamete has 12.5% admixture from the subpopulation indexed by the row) versus the alternative hypothesis (no admixture from this subpopulation). This shows that although the marker panel can easily detect small degrees of non-African admixture on gametes of mainly African ancestry, it has rather less ability to exclude small degrees of Native American admixture on gametes of mainly European ancestry. Thus for a hypothesis of 12.5% Native American ancestry, 87.5% European ancestry, versus a true hypothesis of European ancestry with no Native American admixture ancestry, the expected log-likelihood ratio info in favour of the true hypothesis is only 0.8 nats.

4.2. Inference of individual ancestry

For each individual in the test family, the first model fitted was one with three-way admixture on each parental gamete. The subpopulation with the lowest proportionate contribution to the ancestry of a gamete at the posterior mode was removed from the model for that gamete (by setting the corresponding element of the Dirichlet prior parameter vector to zero), and the analysis repeated until removing subpopulations did not increase the evidence value further. Table 3 shows the results of these analyses. As expected, more complex models (those with more subpopulations ancestral to each gamete) are penalized more by the Occam factor.

For parent 1 (who reports European ancestry), the highest evidence value was for a model with both gametes of European ancestry. However the evidence for a model with one European and one admixed European/Native American was only slightly lower. Even though the evidence slightly favoured a model with no European admixture, the posterior mode in a model with Native American admixture on one gamete was at 25% Native American admixture. For parent 2 (who reports Mexican ancestry), the evidence was highest for models in which one gamete was of unadmixed Native American ancestry and the other gamete represented either by three-way admixture between African, European and Native American, or two-way African/Native American admixture. For each of the three offspring, the results were consistent with the inferred ancestry of the parents. The most likely model for each offspring was one in which one parent was unadmixed European and the other parent was of mixed African/Native American ancestry. However, as with direct analysis of parent 1, models with Native American admixture in the European-ancestry parent had only slightly lower evidence values than models in which this parent was of unadmixed European ancestry. As with direct analysis of parent 2, models with three-way African/European/Native American admixture had only slightly lower evidence values than models with two-way African/Native American admixture.

Figure 1 shows the joint posterior distribution for the proportions of African admixture on

each parental gamete of parent 2 in a model that specifies three-way admixture on both parental gametes. As expected in a model where the two gametes are not identified, this distribution has two symmetrical modes at which one parental gamete has about 30% African admixture proportion and the other has about 5% African admixture. Thus even though no information on phase is available, the program is able to infer that the two parents of parent 2 have different proportions of African admixture.

4.3. Inferring the demographic origin of an unknown individual

To evaluate the ability to test hypotheses about demographic origin, an individual estimated to have a high proportion of Native American admixture was chosen from the Trinidad sample. The log evidence for a model based on origin from each of the five Caribbean regions sampled was then evaluated by thermodynamic integration. Figure 2 shows plots of the gradient of the log partition function for each population of origin. The log evidence for each model is equal to the area under the curve. Table 4 shows the log Bayes factors, relative to the model with highest evidence value (origin from Trinidad). There is very strong evidence against origin from Jamaica, St Vincent or Guyana, with log Bayes factors from 9 to 29 nats. A models with origin from St Lucia has only slightly lower evidence than a model with origin from Trinidad.

4.4. Classifying individuals as admixed or unadmixed

The mean admixture proportions of the São Paulo sample in a model of three-way admixture were estimated as 72% European, 18% African, and 10% Native American. Figure 3 shows for the 1573 individuals in this dataset the relationship of the posterior mode of non-European (African plus Native American) admixture proportions calculated in a model of three-way admixture to the Bayes factor in favour of admixture (calculated by comparing a model based on three-way admixture with a model based on only European ancestry). As one would expect, the estimates of non-European admixture proportions are correlated with the evidence values in favour of admixture, but the relationship is far from perfect. For 8% of these individuals, the log Bayes factor is less than -2 nats (equivalent to a likelihood ratio greater than exp(2) in favour of a model with no admixture). The posterior modes for non-European admixture proportions in these individuals are between 9% and 18%. Bayes factors less than zero, implying that the evidence favours the hypothesis of no admixture, are compatible with estimated admixture proportions as high as 25%.

4.5. Inferring the number of ancestral subpopulations contributing to an admixed population

Table 5 shows the log evidence values for models with one, two and three ancestral subpopulations fitted to the San Luis Valley dataset. There is strong evidence in favour of the model with three subpopulations, with log evidence 3.1 nats higher than for a model with two subpopulations, and 11.4 nats higher than for a single-population model. These results, using uninformative priors on allele frequencies, are consistent with direct estimates of the genetic background of this population using prior information about demographic background and ancestry-specific allele frequencies. Thus using ADMIXMAP to fit a model of three-way admixture with prior information about allele frequencies in Europeans, Native Americans and west Africans, the mean admixture proportions were estimated as 61% European, 34% Native American, and 5% west African. The frequency of null alleles at the FY (Duffy antigen) locus in this sample is 4.9%, supporting a nonzero contribution of African genes to the ancestry of this population and thus a three-way admixture model. For comparison, the results of using an approximate method for calculating the log evidence, proposed by Pritchard et al (6) and implemented in the program STRUCTURE, are given in Table 5: in this example, the approximate method does not correctly rank the models by their evidence values. Figure 4 shows how Pritchard's statistic is based on approximating the area under each energy-coolness curve by the area under a straight line. This is discussed in more detail below.

5. DISCUSSION

Using a panel of at least 10 unlinked markers informative for ancestry, it is straightforward to assign the ancestry of DNA from a single unadmixed individual to one of the three or four main continental groups that have contributed genes to the US population (19). To test hypotheses about the ancestry of an admixed individual requires larger numbers of markers and this in turn requires a more complex statistical model that allows for linkage between marker loci. The calculations of expected log-likelihood ratio for the DNAPrint marker panel show that a marker panel that is adequate to estimate individual admixture proportions is not necessarily adequate to distinguish admixed from unadmixed individuals. Thus, for instance, with the DNAPrint panel of 176 markers the expected log-likelihood ratio between a model specifying unadmixed European ancestry and a model specifying a small proportion of Native American admixture is small. Consistent with these calculations of expected log-likelihood ratio, the Bayes factors obtained in comparisons of models with and without Native American admixture in individuals of European ancestry are close to zero. With a larger panel of markers, such limitations could be overcome; the expected log Bayes factor in favour of the true model should scale arithmetically with the effective number of unlinked markers. To confirm or exclude admixture, an optimal marker set might include more markers with alleles that are unique to one of the putative ancestral populations (20). In principle it is possible to distinguish between recent and long-standing admixture in the genetic background of an individual's parents by inferring the arrival rate parameter on each gamete, but this would

require a larger panel of informative markers than the 176 used in this example.

Although Bayesian and sampling-theory ("frequentist") approaches to estimation of model parameters generally yield similar results in large samples, the Bayesian approach to model comparison differs profoundly from sampling-theory approaches. Classical linkage tests using LOD scores to compare models with fixed parameters are a special case of the Bayesian approach. In the more general Bayesian approach, model parameters are specified with prior distributions and the evidence for each model is the likelihood averaged over the prior. Inference based on Bayes factors is a corollary of using probabilities to represent uncertainty about hypotheses, underpinned by the work of Cox (8), who showed that probability theory is the only set of rules for manipulating uncertainty about propositions that does not violate axioms of logical consistency (14).

Although Bayesian computationally-intensive methods are widely used in statistical genetics, model comparisons based on Bayes factors are rarely undertaken because of the computational difficulties of evaluating the likelihood as an average over the prior on model parameters. Of the two methods described here, the candidate's formula is the most efficient when the prior density is available in closed form, but the thermodynamic algorithm has a wider range of application to complex models. With both methods, the log evidence is evaluated by subtracting a measure of "complexity" from a measure of "fit". In the candidate's formula, minus the log Occam factor is subtracted from the log best-fit likelihood. In the thermodynamic algorithm the information (reduction in entropy) is subtracted from the posterior mean of the log-likelihood (minus the energy).

As others have emphasized (6), a key practical problem in modeling admixture is to infer how many subpopulations (with unspecified allele frequencies) or which subpopulations (where prior information on allele frequencies is available) should be included in the model. **(author?)** (6) have proposed, and implemented in their program STRUCTURE for modeling admixture, an approximate method for calculating the log evidence, in which half the posterior variance of the log-likelihood is subtracted from the posterior mean of the log-likelihood. As this approximate method has been widely used to infer the number of ancestral subpopulations contributing to the gene pool of an admixed or stratified population (21; 22; 23; 24; 25), it is of interest to compare it with direct evaluation of the log evidence by thermodynamic integration. For this comparison we can use a thermodynamic analogy, in which the log-likelihood is minus the "energy" of a system in a heat bath. The variance of the energy is equal to minus the gradient of mean energy with respect to coolness (15). Thus in figure 4, where the log evidence for each model is minus the area under the curve, the method implemented by Pritchard et al. is equivalent to approximating the area under the curve by the area under a tangent to the curve at coolness of 1. Inspection of Figure 4 shows that the area under each curve is poorly approximated by the area under the corresponding tangent, at least in this example where the energy of models with two and three subpopulations falls sharply as the coolness approaches 1. Pritchard's method should not be expected to yield more accurate results with larger sample sizes or more marker loci, as the curve relating energy to coolness does not in general approximate a straight line as the amount of information increases.

The analyses of individuals in the test family demonstrate that even with this limited marker set, the program is able to infer the ancestry of each parental gamete separately. Where the prior on admixture proportions is the same on the two gametes, the two gametes are not identifiable in the model but the program is able to infer two symmetrical modes in the posterior distribution of admixture proportions. For each individual in the test family the model with the highest evidence value is the one consistent with the known ancestry of the individuals under study. However the marker set has only limited ability to exclude Native American admixture in an individual of European ancestry. This is consistent with direct calculations of the expected likelihood ratio based on the marker allele frequencies. If X chromosome markers were included in the analysis, the two gametes would be identifiable in the model. Even without X chromosome data, inference of the continental origin of the mitochondrial and Y haplotypes could be used to restrict the set of models that are compatible with the data.

The usual method of classifying individuals as admixed or unadmixed is to define a cutoff based on the maximum likelihood estimate or posterior mean of individual admixture proportions (6). Our results show, however, that the posterior mode of individual admixture proportions (equivalent, where the prior is flat, to the maximum likelihood estimate) is not necessarily a reliable guide to the evidence that an individual is admixed. Thus, for instance, fitting a model with Native American admixture on one gamete to an individual of European origin without known ancestry from the American continent in the test family yielded a posterior mode of 25% Native American admixture on this gamete. Calculation of the evidence, however, shows that the data give slightly more support to the (presumably correct) model of unadmixed European ancestry than to a model with Native American admixture for this individual. The analyses of the São Paulo case-control dataset again demonstrate that estimates of admixture proportions may be misleading with respect to the strength of evidence that an individual is admixed rather than unadmixed: for about 8% of the sample, there is fairly strong evidence against non-European admixture even though for these individuals the proportion of non-European admixture is typically estimated (as the posterior mode) to be greater than 10%.

The methods described here have practical applications to modeling admixture, and more generally to any Bayesian modeling of genetic data using MCMC simulation. For instance, some studies require restriction of the analysis to unadmixed individuals, as when estimating allele frequencies in continental populations. For admixture mapping, on the other hand, it is desirable to restrict the analysis to admixed individuals before typing genome-wide mapping panels of markers informative for ancestry (4; 26). Where a case-control sample appears to consist of a mixture of admixed and unadmixed individuals, as in the Brazilian case-control dataset examined here, it might be appropriate to classify individuals as unadmixed or admixed so that the two groups can be analysed separately. Another application is to predict the biogeographical ancestry of an unknown individual whose DNA has been recovered from the scene of a crime. In principle, it is possible to infer the ancestry of each parental gamete, and to distinguish recent admixture from admixture that occurred many generations earlier. The objective of such analyses is not, of course,

to establish a match with the identity of a known individual, but rather to infer the demographic background of an unknown individual and thus to narrow the field of enquiry. For instance, if the analysis yields evidence that one or both parents of an individual whose DNA has been recovered from a crime scene has Native American ancestry, then this restricts the field of enquiry to individuals whose demographic origins have some connection with the American continent. Other possible applications are in pharmacogenetics. Where an individual has ancestry from two or more subpopulations that vary in drug response, it may be possible to make useful predictions of drug response from the proportion of the individual's genome that has ancestry from each subpopulation (27). This may help to refine the design and analysis of clinical trials.

Although the thermodynamic algorithm is computationally intensive (requiring about 6 h of CPU time for a dataset based on 22 marker loci typed in 446 individuals, it is amenable to parallel computation on a cluster, simply by running copies of the sampler at different coolnesses. As computation time for the algorithms used in ADMIXMAP scales arithmetically with the number of marker loci and the number of individuals sampled, it is possible to analyse even large datasets where a cluster of a few hundred CPUs is available. Our results demonstrate that it is both feasible and necessary to calculate Bayes factors when testing hypotheses about the genetic background of individuals and populations.

6. Acknowledgements

This work was supported by NIH grant HG002154 to MDS and PMM.

7. Web Resource

The URLs for resources cited herein are as follows:
SNP map position calculator, http://actin.ucd.ie/cgi-bin/rs2cm.cgi
ADMIXMAP program, http://www.ucd.ie/genepi/admixmap/index.html

REFERENCES

1. McKeigue P (2005) Prospects for admixture mapping of complex traits. Am J Hum Genet 76:1–7

2. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67:170–81

3. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–87

4. Hoggart CJ, Shriver MD, Kittles RA, Clayton DG, McKeigue PM (2004) Design and analysis of admixture mapping studies. Am J Hum Genet 74:965–78

5. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, Hauser SL, Smith MW, O'Brien SJ, Altshuler D, Daly MJ, Reich D (2004) Methods for high-density admixture mapping of disease genes. Am J Hum Genet 74:979–1000

6. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–59

7. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG, McKeigue PM (2003) Control of confounding of genetic associations in stratified populations. Am J Hum Genet 72:1492–1504

8. Cox R (1961) The algebra of probable inference. Johns Hopkins University Press, Baltimore, MD

9. Guindalini C, Howard M, Haddley K, Laranjeira R, Collier D, Ammar N, Craig I, O'Gara C, Bubb V, Greenwood T, Kelsoe J, Asherson P, Murray R, Castelo A, Quinn J, Vallada H, Breen G (2006) A dopamine transporter gene functional variant associated with cocaine abuse in a Brazilian sample. Proc Natl Acad Sci U S A 103:4552–7

10. Kong X, Murphy K, Raj T, He C, White P, Matise T (2004) A combined linkage-physical map of the human genome. Am J Hum Genet 75:1143–8

11. Jeffreys H (1961) Theory of Probability, 3rd ed. Oxford University Press

12. Molokhia M, Hoggart C, Patrick AL, Shriver M, Parra E, Ye J, Silman AJ, McKeigue PM (2003) Relation of risk of systemic lupus erythematosus to west African admixture in a Caribbean population. Hum Genet 112:310–8

13. MacDonald I, Zucchini W (1997) Hidden Markov and other models for discrete-valued time series - Monographs on Statistics & Applied Probability. Chapman and Hall/CRC

14. Skilling J (1998) Probabilistic data analysis: an introductory guide. Journal of Microscopy 190:28–36

15. Mackay D (2003) Information theory, inference and learning algorithms. Cambridge University Press, Cambridge, UK

16. Besag J (1989) A candidate's formula: a curious result in Bayesian prediction. Biometrika 76:183

17. Chib S (1995) Marginal likelihood From the Gibbs output. J Am Stat Ass 90:1313–1321

18. Neal R (1993) Probabilistic inference using Markov chain Monte Carlo methods. Technical report CRG-TR-93-1, Department of Computer Science, University of Toronto

19. Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. Am J Hum Genet 60:957–64

20. Chakraborty R, Kamboh M, Ferrell R (1991) 'Unique' alleles in admixed populations: a strategy for determining 'hereditary' population differences of disease frequencies. Ethn Dis 1:245–56

21. Randi E, Pierpaoli M, Beaumont M, Ragni B, Sforzi A (2001) Genetic identification of wild and domestic cats (Felis silvestris) and their hybrids using Bayesian clustering methods. Mol Biol Evol 18:1679–93

22. Bamshad M, Wooding S, Watkins W, Ostler C, Batzer M, Jorde L (2003) Human population genetic structure and inference of group membership. Am J Hum Genet 72:578–89

23. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol 14:2611–20

24. Reiner A, Ziv E, Lind D, Nievergelt C, Schork N, Cummings S, Phong A, Burchard E, Harris T, Psaty B, Kwok P (2005) Population structure, admixture, and aging-related phenotypes in African-American adults: the Cardiovascular Health Study. Am J Hum Genet 76:463–77

25. Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet 78:680–90

26. Smith M, Patterson N, Lautenberger J, Truelove A, McDonald G, Waliszewska A, Kessing B, et al. (2004) A high-density admixture map for disease gene discovery in African-Americans. Am J Hum Genet 74:1001–13

27. Nebert D (1999) Pharmacogenetics and pharmacogenomics: why is this relevant to the clinical geneticist? Clin Genet 56:247–58

28. Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J R Statistical Soc, Series B 44:226–232

29. Lystig TC, Hughes JP (2002) Exact computation of the observed information matrix for hidden Markov models. Journal of Computational and Graphical Statistics 11:678–689

30. Atchade YF, Rosenthal JS (2005) On adaptive Markov chain Monte Carlo algorithms. Bernoulli 11:815–828

A. Information content of a dense panel of ancestry-informative markers for individual genetic background

The efficiency of a marker panel for estimating the genetic background of an individual can be evaluated by comparing it with the information extracted by a dense marker panel that allows ancestry states to be inferred without uncertainty at all positions on all chromosomes. This maximum information content can be derived from the missing information principle (28). For a genome of length L and a single gamete, the observed data are the number r of chromosomes that begin in state 1, the number y_1 of arrivals of state 1 that change the ancestry state from 2 to 1, the number y_2 of arrivals of state 2 that change the ancestry state from 1 to 2, and the observed fraction m of the genome that is in state 1. The missing data are the number x_1 of unobserved arrivals of state 1 and the number x_2 of unobserved arrivals of state 2. The complete-data log-likelihood is the sum of two Poisson log-likelihoods with parameters $\mu\rho L$ and $(1-\mu)\,\rho L$, and a binomial log-likelihood with parameter μ. This can be rearranged, ignoring constant terms, as the sum of a Poisson log-likelihood for ρ and a binomial log-likelihood for μ.

$$-\rho L + (x_1 + y_1 + x_2 + y_2)\log\rho + (x_1 + y_1 + r)\log\mu + (x_2 + y_2 + C - r)\log(1-\mu)$$

Given the observed proportion f, x_1, x_2 are distributed as independent Poisson variates with parameters $\mu m\rho L$ and $(1-\mu)(1-m)\,\rho L$ respectively. The observed information is calculated as the difference between the complete information (minus the posterior expectation of the second derivative of the log-likelihood) and the missing information (posterior variance of the score). For μ, the observed information evaluates to

$$\frac{C - r + (1-\mu)(1-m)\,\rho L + y_2}{(1-\mu)^2} + \frac{r + \mu m\rho L + y_1}{\mu^2} - \left(\frac{m}{1-\mu} + \frac{1-m}{\mu}\right)\rho L$$

Over repeated experiments, the expectations of r and m are $C\mu$ and $(1-\mu)$. The expectation of y_1 and y_2 is $\mu(1-\mu))\,\rho L$. The expected observed information is therefore $(C+\rho L)\,/\,[\mu(1-\mu)] - 2\rho L$. For $\log\rho$, the complete information is ρL, the missing information is $[m\mu + (1-\mu)(1-\mu)]\,\rho L$, and the expected observed information is therefore $2\mu(1-\mu)\,\rho L$

B. Calculation of Fisher information content of a marker panel for gamete admixture proportions μ

. The second derivative of the log-likelihood function of a hidden Markov model can be calculated by a recursive algorithm (29). For two-way admixture on a single gamete with parental admixture proportion μ, the probability of ancestry state j at locus $t+1$, given state i at locus t and that the distance between these loci is d_t morgans, is given by $F\delta_{ij} + (1-F)\,\mu$, where $F = \mathrm{e}^{-\rho d_t}$ and δ_{ij} is an indicator variable for $i = j$. The log-likelihood can be calculated as a sum of increments

over loci

$$\mathrm{P}\left(y_1, \ldots, y_T\right) = \sum_{t=1}^{T} \mathrm{P}\left(y_{t+1} \mid y_1, \ldots, y_t\right)$$

We can thus evaluate the second derivative of the log-likelihood as

$$\frac{\mathrm{d}^2 \log \mathrm{P}\left(y_1, \ldots, y_T\right)}{\mathrm{d}\mu^2} = \sum_{t=1}^{T} \frac{\mathrm{d}^2 \log \mathrm{P}\left(y_{t+1} \mid y_1, \ldots, y_t\right)}{\mathrm{d}\mu^2}$$

The incremental contribution of the $(t+1)$th locus to the log-likelihood is

$$\log \mathrm{P}\left(y_{t+1} \mid y_1, \ldots, y_t\right) = \log\left(F\left[\alpha_t p + \left(1-\alpha_t\right)\left(p+\delta\right)\right] + \left(1-F\right)\left[\mu p + \left(1-\mu\right)\left(p+\delta\right)\right]\right)$$

where p is the frequency of the allele observed at locus $t+1$ given ancestry state 1, $p+\delta$ is the frequency of this allele given ancestry state 2, and α_t is the probability of ancestry state 1 at locus t, conditioned on the observations from 1 to t. α_t is calculated from the following recursion from 1 to t.

$$\alpha_{t+1} = \frac{p\left[\left(1-F\right)\mu + F\alpha_t\right]}{\left(p+\delta\right)\left[\left(1-F\right)\left(1-\mu\right) + F\left(1-\alpha_t\right)\right] + p\left[\left(1-F\right)\mu + F\alpha_t\right]}$$

By differentiation we obtain similar recursions for the first and second derivatives of α_{t+1}, which are used in the recursion for the second derivative of $\log \mathrm{P}\left(Y_{t+1} \mid Y_1, \ldots, Y_t\right)$. The Fisher information content of the marker panel is evaluated as the average over repeated simulations of minus the second derivative of the log-likelihood.

C. Calculation of posterior density for the candidate's formula

To calculate the posterior density $p(\boldsymbol{\theta}^* \mid y)$, we sample auxiliary variables $\boldsymbol{z}$ that yield sufficient statistics for $\boldsymbol{\theta}$, allowing the conditional density $p\left(\boldsymbol{\theta}^* \mid \boldsymbol{z}\right)$ to be evaluated directly. The posterior density $p(\boldsymbol{\theta}^* \mid y)$ can then be evaluated by averaging $p\left(\boldsymbol{\theta}^* \mid \boldsymbol{z}\right)$ over the posterior distribution of z given y.

$$p\left(\boldsymbol{\theta}^* \mid y\right) = \int p(\boldsymbol{\theta}^* \mid \boldsymbol{z}) p(\boldsymbol{z} \mid y)\, d\boldsymbol{z}$$

For a model with K ancestral subpopulations and T marker loci, the auxiliary variables z comprise a $2 \times K \times T$ array $\boldsymbol{A}$ of indicator variables for locus ancestry states on each gamete, a $2 \times K \times T$ array of indicator variables $\boldsymbol{\xi}$ which take value 1 if there has been at least one arrival between locus t and locus $t+1$ on each gamete, the total numbers N_1, N_2 of arrivals on each gamete, and an array $\boldsymbol{H}$ of realized counts of each haplotype at each compound locus on each gamete. The ancestry states at loci where $\xi = 1$ are sufficient for $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$. The numbers N_1, N_2 of arrivals are sufficient for ρ_1, ρ_2, and H is sufficient for $\boldsymbol{\phi}$.

Given the auxiliary variables $\boldsymbol{z}$, $\boldsymbol{\mu}$, $\boldsymbol{\rho}$ and $\boldsymbol{\phi_t}$ are conditionally independent. We write $\boldsymbol{\alpha}_g$ for the parameter vector of the prior on admixture proportions $\boldsymbol{\mu_g}$ on the gth gamete, γ_1, γ_2 for the shape

and rate parameters of the gamma prior on the arrival rate ρ_g on the gth gamete, and β_{tk} for the parameter vector of the prior on allele frequencies at the tth locus in the kth subpopulation. The density $p(\boldsymbol{\theta}^* \mid \boldsymbol{z})$ is evaluated as the product of two conjugate Dirichlet densities with parameter vector $\left(\boldsymbol{\alpha_g} + \sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{\xi_{gt.}} \boldsymbol{A_{gt.}}\right)$ for the admixture proportions on each gamete, two conjugate gamma densities with scale parameter $\gamma_1 + N_g$ and rate parameter $\gamma_2 + L$ for the arrival rate parameters on each gamete, and T conjugate Dirichlet densities with parameter vector $\left(\boldsymbol{\beta_{tk}} + \sum_{\boldsymbol{g=1}}^{\boldsymbol{2}} A_{gtk} \boldsymbol{H_{gtk}}\right)$ for the allele frequencies $\boldsymbol{\phi_{tk}}$ at the tth locus in the kth subpopulation.

D. Annealed sampling for thermodynamic integration

The observed genotype data y enter the model only through the probabilities $P(y_t \mid A_t, \boldsymbol{\phi_t})$ of genotype y_t given locus ancestry state A_t and allele frequencies $\boldsymbol{\phi_t}$ at locus t. Unless the genotype is homozygous or the ancestry states on the two gametes are the same, $P(y_t \mid A_t, \boldsymbol{\phi_t})$ is a sum over possible ordered haplotype pairs compatible with the observed genotype. To anneal the likelihood at coolness λ we substitute these genotype probabilities with the (unnormalized) annealed probabilities $[P(y_t \mid A_t, \boldsymbol{\phi_t})]^{\lambda}$. The annealed genotype probabilities can be substituted directly for the unannealed genotype probabilities in a standard hidden Markov model recursion to calculate the forward probabilities $P(y_1, \ldots, y_t \mid A_t)$, which are used to update admixture proportions, arrival rate parameter and locus ancestry states. To use the annealed genotype probabilities $P(y_t \mid A_t, \boldsymbol{\phi_t})^{\lambda}$ to update the allele frequencies $\boldsymbol{\phi_t}$ conditional on locus ancestry, we use a Hamiltonian sampler (18) with allele frequencies transformed to the softmax basis. For each update of the variable x, the state space is augmented with a randomly-drawn "momentum" variable p. A leapfrog algorithm, using the gradient of the log density to simulate Hamiltonian dynamics, is then used to propose a joint update of x and p for a Metropolis accept/reject step. A stochastic approximation algorithm (30) is used to tune the leapfrog step size automatically to achieve a target acceptance rate of 90% in the Metropolis step.

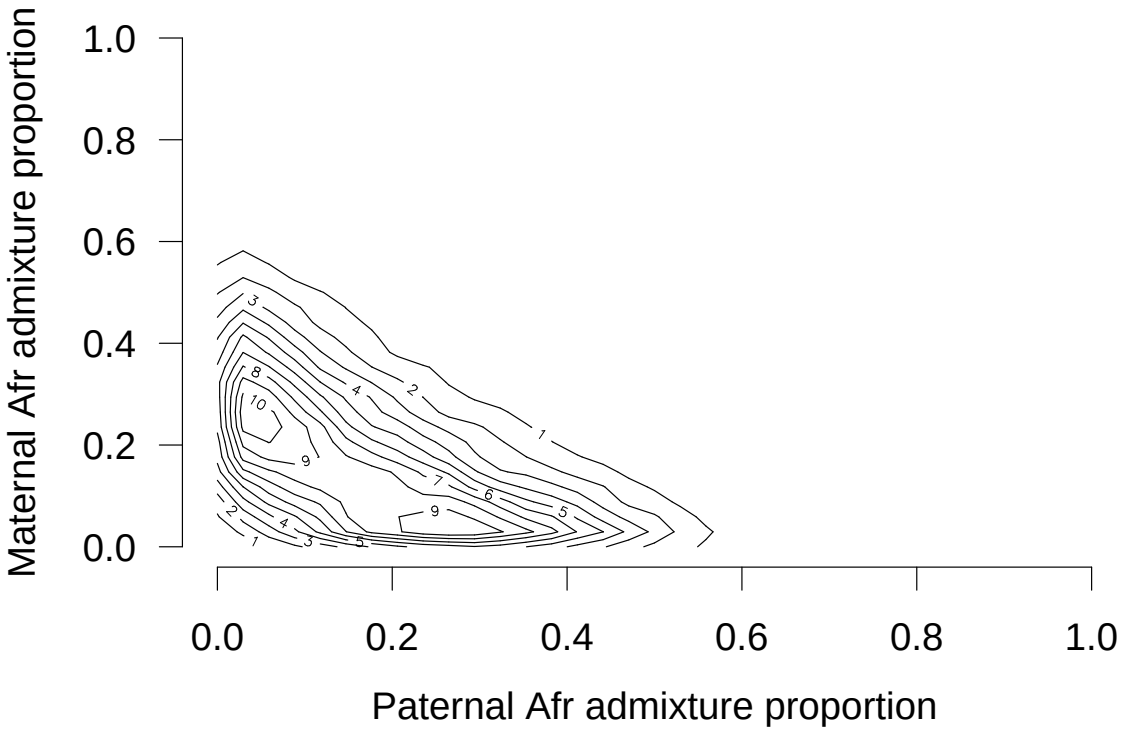

Fig. 1.— Contour plot of posterior density of African admixture proportions for the two parental gametes of individual 2 in the test family, in a model specifying three-way admixture on both gametes. Contours are labelled by their density values

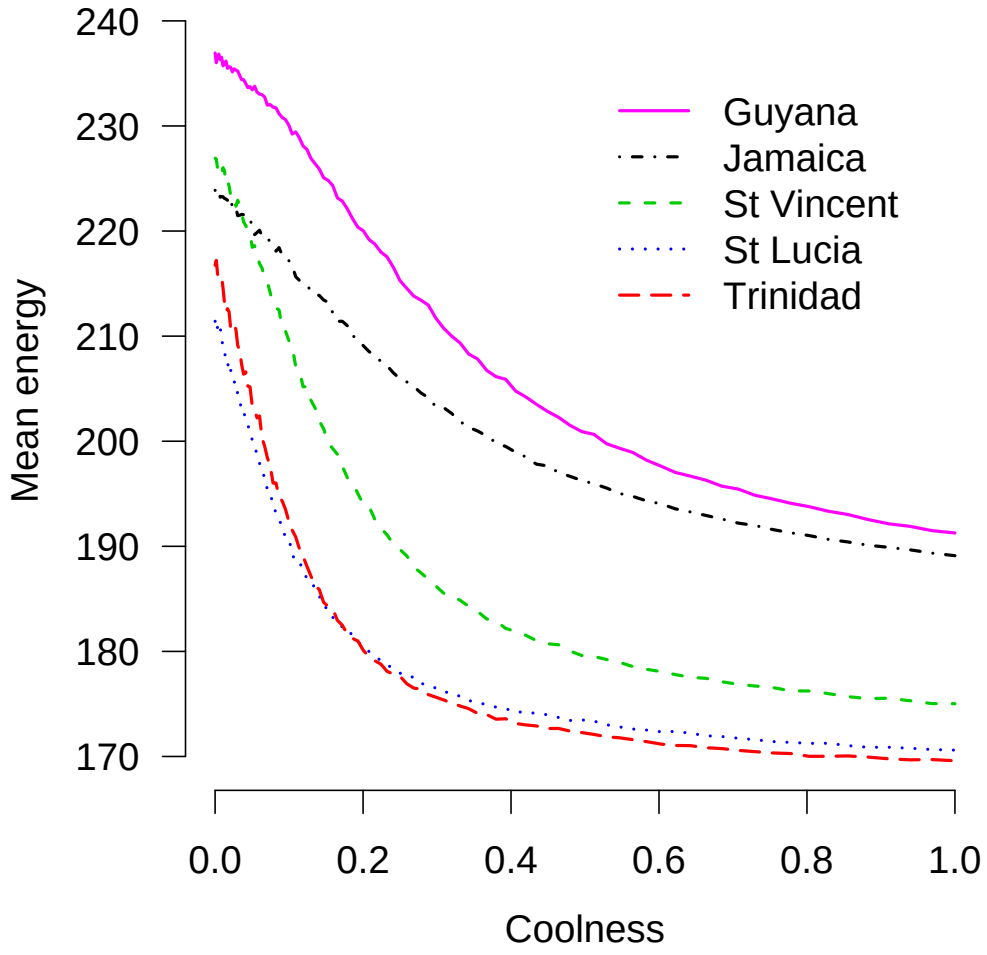

Fig. 2.— Evaluation of log evidence by thermodynamic integration: plot of energy against coolness for models based on different countries of origin for an individual in Trinidad

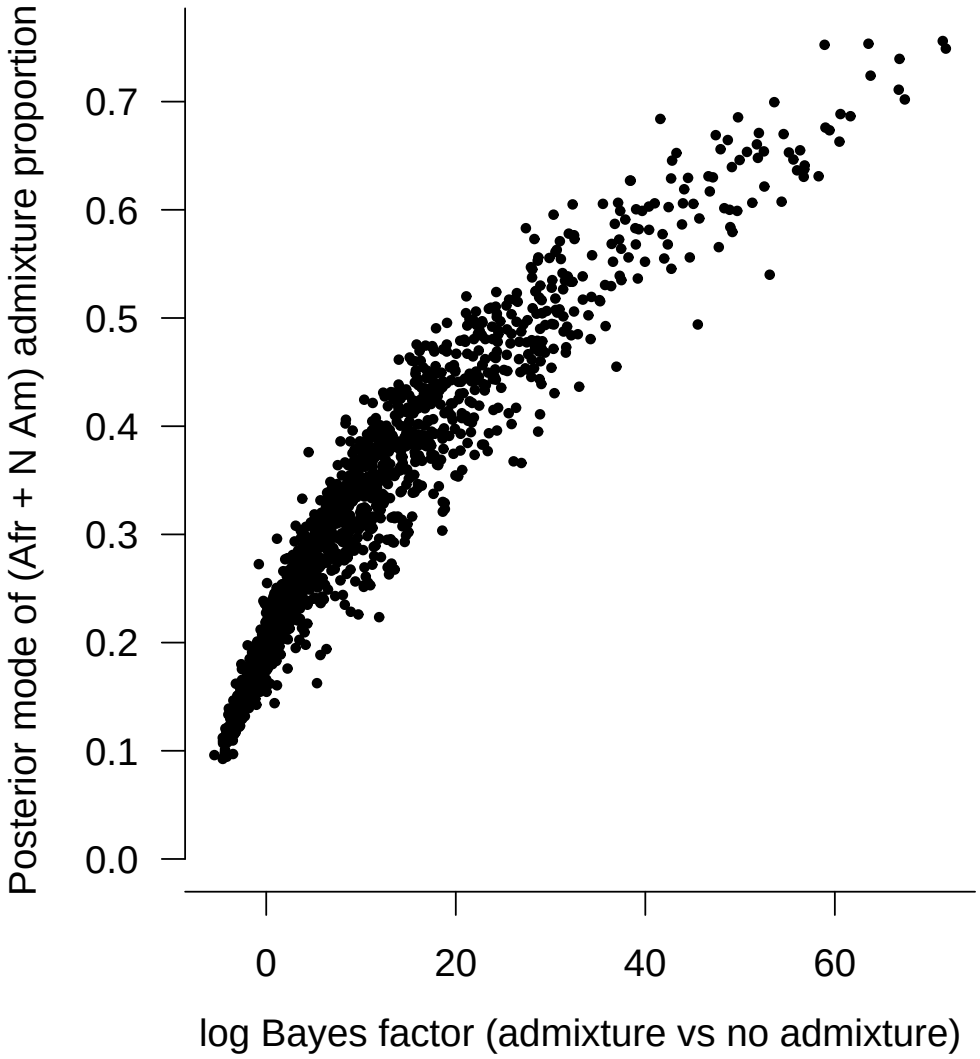

Fig. 3.— São Paulo case-control sample: scatter plot of posterior mode of non-European admixture proportions against log Bayes factor in favour of admixture versus no admixture. A log Bayes factor less than 0 indicates that the evidence value is higher for a model with no admixture than for a model with admixture

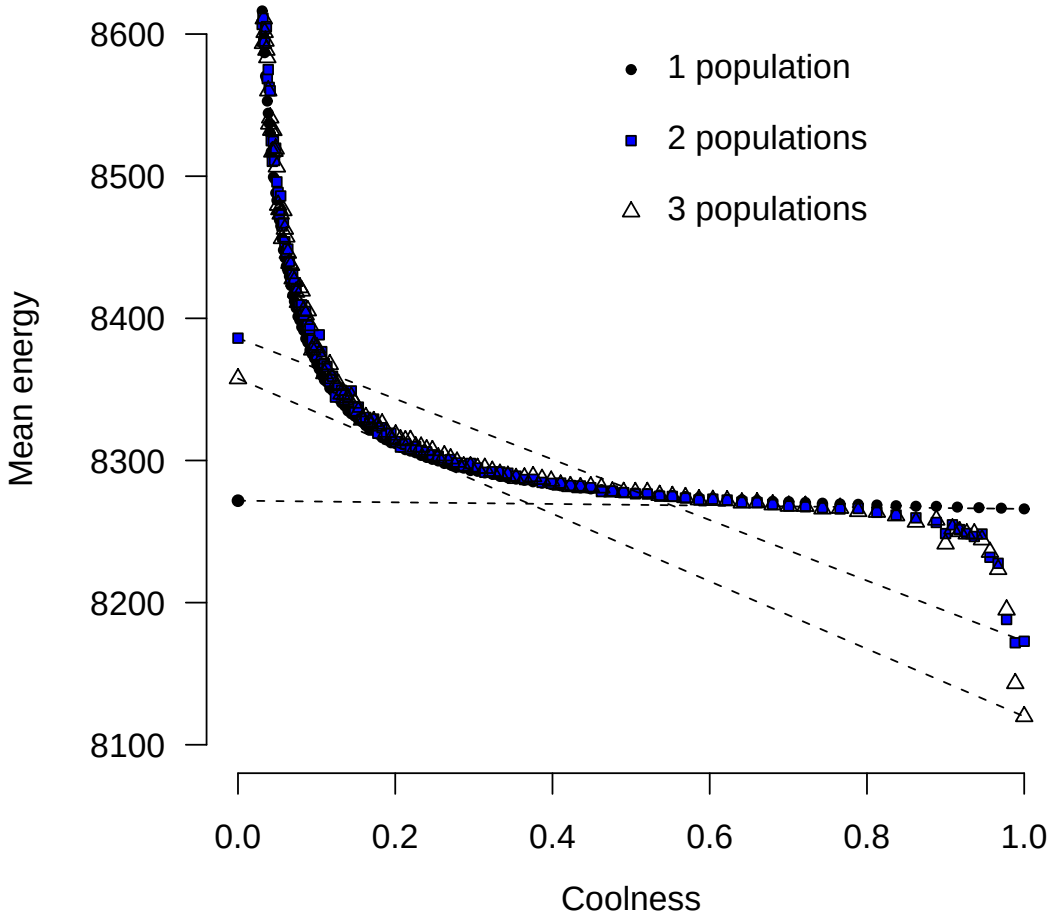

Fig. 4.— San Luis Valley Hispanic-American sample: plot of energy against coolness for models specifying 1, 2 and 3 subpopulations. Broken lines show the approximation to these curves implemented in the STRUCTURE program (6)

Table 1: DNAPrint marker panel with 176 SNPs: total pairwise information content (at admixture proportion 0.5) for gamete admixture of marker panel between continental groups

	African	European	Native Am	E Asian	S Asian	Mid-East	Pacific
African	.						
European	76.9	.					
Native Am	89.4	52.8	.				
E Asian	97.9	61.7	32.5	.			
S Asian	69.8	15.8	35.2	35.6	.		
Mid East	75.1	6.1	50.7	54.8	12.4	.	
Pacific	102.2	64.0	44.2	14.8	42.3	58	.

Table 2: DNAPrint marker panel: expected log-likelihood ratio in favour of no admixture (from population indexed by row) versus 12.5% admixture on single gamete. Thus the cell with row labelled N Am and colum labelled Eur is the expected log-likelihood ratio in favour of no Native American admixture in an individual of European ancestry, and the cell with row labelled Eur and column labelled N Am is the expected log-likelihood ratio in favour of indexed by cell

	Afr	Eur	NAm	EAsian	SAsian	MidEast	Pacific
Afr	.	1.24	1.78	1.75	1.06	1.31	2.15
Eur	1.49	.	1.15	1.05	0.19	0.06	1.20
NAm	1.56	0.80	.	0.35	0.45	0.72	0.57
EAsian	1.77	0.92	0.50	.	0.47	0.80	0.25
SAsian	1.39	0.23	0.72	0.55	.	0.19	0.85
MidEast	1.35	0.06	1.03	0.95	0.14	.	1.07
Pacific	1.86	0.89	0.64	0.19	0.47	0.79	.

Table 3: Comparison of models of ancestry of each individual in the test family: posterior mode for gamete admixture proportions $\boldsymbol{\mu_1}, \boldsymbol{\mu_2}$ and log evidence
(-) indicates subpopulations for which the gamete admixture proportion μ was specified as zero in the model

	Posterior mode						Log best-fit likelihood	Log Occam factor	Log evidence
	$\boldsymbol{\mu_1}$			$\boldsymbol{\mu_2}$					
	Afr	Eur	N Am	Afr	Eur	N Am			
Parent 1	0.05	0.84	0.12	0.07	0.78	0.16	-136.97	-4.70	-141.67
	-	0.91	0.10	0.07	0.81	0.12	-136.42	-3.24	-139.66
	-	0.75	0.25	-	1	-	-135.40	-1.09	-136.49
	-	1	-	-	1	-	-136.31	0.00	-136.31
Parent 2	0.30	0.18	0.51	0.06	0.12	0.82	-140.48	-2.94	-143.42
	0.32	0.27	0.41	-	0.07	0.94	-139.76	-2.88	-142.65
	0.37	0.12	0.51	-	-	1	-139.84	-1.29	-141.13
	0.43	-	0.57	-	-	1	-140.70	-1.44	-142.14
Offspring 1	0.09	0.78	0.13	0.21	0.11	0.68	-139.53	-1.63	-141.16
	-	0.69	0.31	0.27	0.27	0.45	-139.34	-1.14	-140.48
	-	1	-	0.22	0.11	0.67	-138.94	-1.27	-140.21
	-	1	-	0.27	-	0.73	-137.78	- 1.12	-138.70
Offspring 2	0.08	0.19	0.74	0.12	0.79	0.09	-144.62	-2.13	-146.75
	-	0.88	0.12	0.12	0.13	0.75	-143.34	-2.27	-145.61
	-	1	-	0.10	0.16	0.74	-143.43	-1.57	-145.01
	-	1	-	0.18	-	0.82	-142.83	-1.44	-144.27
Offspring 3	0.10	0.43	0.47	0.08	0.64	0.28	-138.53	-2.53	-141.06
	-	0.72	0.28	0.14	0.40	0.46	-138.47	-1.46	-139.93
	-			-			-138.58	-1.14	-139.72
	-	1	-	0.13	0.11	0.75	-137.36	-2.14	-139.50
	-	1	-	0.15	-	0.85	-137.63	-1.80	-139.43

Table 4. Log Bayes factors comparing the evidence for each hypothesis of an island of origin, for a single individual originating from Trinidad

Population	Log-likelihood	Information	Log evidence	Bayes factor
Guyana	-191.27	14.78	-206.05	-29.37
Jamaica	-189.11	10.72	-199.83	-23.16
St Vincent	-175.02	10.80	-185.82	-9.15
St Lucia	-170.60	6.46	-177.06	-0.39
Trinidad	-169.59	7.08	-176.67	0.00

Table 5. Evidence for models with one, two and three ancestral subpopulations given genotypes at 22 ancestry-informative markers for 446 Hispanic-Americans in San Luis Valley, Colorado

Num subpopulations	Log-likelihood	Information	Log evidence	Pritchard approximation
1	-8266.0	69.5	-8335.5	-8271.7
2	-8172.8	154.4	-8327.2	-8386.1
3	-8119.9	204.2	-8324.1	-8357.6