

Inference of population structure using multilocus  
genotype data: Linked loci and correlated allele  
frequencies.

Daniel Falush<sup>1</sup>

Matthew Stephens<sup>2</sup>

Jonathan K. Pritchard<sup>3</sup>

To appear in *Genetics* in August 2003.

July 23, 2003

<sup>1</sup>Department of Molecular Biology, Max-Planck Institut für Infektionsbiologie, 10117 Berlin, Germany

<sup>2</sup>Department of Statistics, University of Washington, Seattle WA 98195, USA

<sup>3</sup>Department of Human Genetics, University of Chicago, Chicago IL 60637, USA

## Abstract

We describe extensions to the method of Pritchard et al. (2000) for inferring population structure from multilocus genotype data. Most importantly, we develop methods that allow for linkage between loci. The new model accounts for the correlations between linked loci that arise in admixed populations (“admixture linkage disequilibrium”). This modification has several advantages, allowing (1) detection of admixture events further back into the past; (2) inference of the population of origin of chromosomal regions; (3) more accurate estimates of statistical uncertainty when linked loci are used. It is also of potential use for admixture mapping. In addition, we describe a new prior model for the allele frequencies within each population, which allows identification of subtle population subdivisions that were not detectable using the existing method. We present results applying the new methods to study admixture in African Americans, recombination in *Helicobacter pylori* and drift in populations of *Drosophila melanogaster*. The methods are implemented in a program, *structure*, version 2.0, that is available at <http://pritch.bsd.uchicago.edu>.

**Keywords:** MALD, hybrid zones, recombination, MCMC, mixture linkage disequilibrium, population genetics

**Running head:** Inferring population structure

**Corresponding author:** Daniel Falush, Max-Planck Institut für Infektionsbiologie, Schumann Strasse 21/22, 10117 Berlin, Germany; [falush@mpiib-berlin.mpg.de](mailto:falush@mpiib-berlin.mpg.de); tel: +49 30 2846 0169; fax: +49 30 2846 0111.

# Introduction

The study of admixed populations arises in many contexts in population genetics: for example, in the study of hybrid zones (Barton and Hewitt, 1989); in methods that use admixed populations for gene mapping (Chakraborty and Weiss, 1988; Stephens et al., 1994; McKeigue, 1998); in studying the ancestry of ethnic groups such as Icelanders and Finns (Thompson, 1973; Guglielmino et al., 1990); and in association mapping (Knowler et al., 1988; Thornsberry et al., 2001). In many of these problems we would like to identify the extent of admixture of individuals, or to infer the origin of particular loci in the sampled individuals (McKeigue et al., 2000). These problems can be difficult if good estimates of the allele frequencies in the parental populations are not available.

In this paper, we develop methods for studying the ancestry of both individuals and specific loci within admixed populations. Much of the previous work on population admixture has aimed to estimate *average* admixture proportions in an entire population (e.g., Long, 1991; Bertorelle and Excoffier, 1998; Chikhi et al., 2001) or to study geographic clines in admixture (e.g., Sites et al., 1995). However, the *distribution* of ancestry proportions can provide additional information about the admixture process (McKeigue et al., 2000; Beaumont et al., 2001; Anderson and Thompson, 2002; Falush et al., 2003). It can also be of interest to determine whether specific parts of the genome are inherited at a higher than normal rate from one of the parental populations, as in admixture mapping or in studying selection across hybrid zones (Rieseberg et al., 1999).

We consider a situation in which we have multilocus genotype data from a sample of individuals collected from a population with (possibly) unknown structure. Pritchard et al. (2000) introduced a method to identify the presence of different subpopulations, if any, and to estimate the ancestry of the sampled individuals. That paper considered two models for the ancestry of individuals. In the first, “no-admixture model”, individuals are assumed to be drawn purely from one of  $K$  populations. In the second, “admixture model”, individuals are allowed to have mixed ancestry: that is, a fraction  $q_k$  of an individual’s genome comes from subpopulation  $k$  (where  $\sum_k q_k = 1$ ). Both of those models assume that all the markers are unlinked and provide independent information on an individual’s ancestry. In this paper we introduce a third model, the “linkage model”, which extends the admixture model to account for the correlations between linked markers that arise as the result of admixture (“admixture linkage disequilibrium”, Stephens et al., 1994). As we will show, the linkage model allows estimation of the origin of chromosomal regions within individuals, and provides better resolution to study the historical process of admixture.

We also discuss a new prior model for the allele frequencies within each population, which can be used in conjunction with any of the three ancestry models. This model, while still relatively simple, is more accurate in many situations, and sometimes allows much more information to be extracted from the data. These, and a number of other extensions to the original model described by Pritchard et al., have been implemented in a computer program,

## Summary of old and new models

Consider a sample of  $N$  individuals, each genotyped at  $L$  loci. Pritchard et al. (2000) began by assuming that the individuals represent a mixture from  $K$  unobserved populations. All of the calculations are conditional on a particular value of  $K$ ; Pritchard et al. suggested performing the analysis for a range of values of  $K$ , and described a heuristic that provides a guide towards the most appropriate value (or values).

In the no-admixture model, each individual comes from one of the  $K$  populations. We let  $z^{(i)}$  denote the population of origin of individual  $i$  and  $Z$  denote the vector  $(z^{(1)} \dots z^{(N)})$ . Each of the  $K$  populations is characterized by a set of allele frequencies at each locus. Let  $p_{klj}$  refer to the frequency of allele  $j$  at locus  $l$  in population  $k$ , and let  $P$  denote the full multi-dimensional vector of allele frequencies for all  $k$ ,  $l$  and  $j$ . A key modelling assumption is that within populations there is linkage equilibrium, and Hardy-Weinberg equilibrium (HWE). Hence, the likelihood of the genotype of individual  $i$ , conditional on its population of origin  $z^{(i)}$ , is simply a product of the frequencies of its alleles in that population.

An obvious limitation of the no-admixture model is that in practice individuals may have recent ancestors in more than one population. To model this, Pritchard et al. introduced an “admixture model”, in which each individual is assumed to have inherited some proportion of its ancestry from each population. Let  $q_k^{(i)}$  denote the proportion of individual  $i$ ’s genome that is derived from population  $k$  (where  $\sum_k q_k^{(i)} = 1$ ), and let  $Q$  be the multi-dimensional vector of ancestry proportions for all the members of the sample. It is now possible for the different allele copies in an individual to come from different populations. (We use the term “allele copy” to refer to an allele carried at a particular locus by a particular individual.) To reflect this, the vector  $Z$  now records the population of origin of every allele copy in each individual, with  $z_l^{(i,a)}$  denoting the origin of the  $a$ th allele copy at locus  $l$  in individual  $i$ . Pritchard et al. (2000) discussed only the case of diploid individuals (so  $a = 1$  or  $2$ ), but the model is easily extended to data of any ploidy, and we have now done this in the updated implementation of *structure* (version 2). Under Pritchard et al.’s admixture model, the  $z$ s for individual  $i$  are assumed to be drawn independently from  $\{1, \dots, K\}$  according to the probability vector  $q^{(i)}$ , so that

$$\Pr[z_l^{(i,a)} = k] = q_k^{(i)}. \quad (1)$$

This admixture model also assumes linkage equilibrium and HWE within populations.

Inference is performed in a Bayesian framework, which offers a number of practical advantages in this context. Among these, it allows a straightforward assessment of the statistical uncertainty in each estimate of interest. It also allows us to make use of any prior information that we might have regarding population membership for some members of the sample. See Pritchard et al. (2000) for further discussion.

The Bayesian approach requires priors for  $P$  and  $Q$ . Following Balding and Nichols (1997), Pritchard et al. assumed that the vector of allele frequencies at locus  $l$  in population  $k$  is drawn from a symmetric Dirichlet distribution parameterized by a single hyperparameter  $\lambda$ , independently for each  $k$ . Some modifications of this prior are described below. The admixture proportions  $q^{(i)}$  for individual  $i$  were also modelled as draws from a symmetric Dirichlet distribution, in this case with a hyperparameter  $\alpha$ . The assumption of symmetry in the prior for the  $qs$  corresponds intuitively to an assumption that the  $K$  populations contribute roughly equal amounts of genetic material to the sample. To better model situations where this is not the case, the updated implementation of *structure* allows different values of  $\alpha$  to be estimated for each population (so  $\alpha$  becomes a vector of  $K$  values, with  $\alpha_k$  representing the relative contribution of population  $k$  to the genetic material in the sample). Otherwise the prior for  $q$  is unchanged. Alternative models for  $q$  are considered by Anderson (2001) and Anderson and Thompson (2002).

In practice we may not know either the allele frequencies  $P$ , or the populations of origin  $Z$  in advance. Pritchard et al. described a Markov chain Monte Carlo (MCMC) scheme that estimates these jointly. This procedure clusters individuals into populations, and estimates the probability of membership (or, for the admixture model, the proportion of membership) in each population for each individual.

A number of related population genetic methods have been described, including Dawson and Belkhir (2001); Sillanpää et al. (2001); Satten et al. (2001). Further, it has recently come to our attention that the admixture model belongs to a class of models known as “Grade of Membership” models, which have been used in other literatures, including medical classification and machine learning (Erosheva, 2002).

**The Linkage Model.** A deficiency of the admixture model is that by assuming that the  $zs$  within each individual are independent, it ignores the correlations in ancestry that one would expect to see along each chromosome. In this context, it is helpful to distinguish between three sources of linkage disequilibrium (LD). The first source is variation in ancestry ( $q$ ) among the sampled individuals. Variation in  $q$  leads to correlations among markers across the genome, even if they are unlinked, because individuals with a large component of ancestry in population  $k$  have an excess of alleles that are common in  $k$ . We call this LD “mixture LD”. The second source is correlations in ancestry along each chromosome, which causes additional LD between linked markers. We visualize this LD as occurring because each chromosome is composed of a set of “chunks” that are derived, as unbroken units, from one or another of the ancestral populations. In our terminology, this second source is “admixture LD”. The third source is “background LD” within populations, which usually decays on a much shorter scale (tens of kilobases in humans). The admixture model in Pritchard et al. (2000) models only mixture LD; here we extend the model to include admixture LD. However, we continue to ignore background LD, and so our model is best suited to data on markers that

are linked, but not so tightly linked that one would expect to see substantial background LD (we return to this point in our discussion).

In order to make inference computationally tractable, we use a simple model that incorporates the notion of discrete chromosomal chunks inherited from ancestral populations. Whereas in the “admixture” model of Pritchard et al. (2000), each allele copy is derived independently from one of the  $K$  populations, in the new “linkage” model chunks of chromosomes are derived as intact units from one or another of the  $K$  populations, and all allele copies on the same chunk derive from the same population. We assume that the breakpoints between successive chunks occur at random (i.e. as a Poisson process), at a rate  $r$  per unit of genetic distance, and that the population of origin of each chunk in individual  $i$  is independently drawn according to the vector  $q^{(i)}$ , which continues to represent the overall (expected) ancestry proportions of the individual.

Formally, the above assumptions translate into replacing the “admixture model” assumption that the  $z$ s along each chromosome are *independent*, with the assumption that  $z$ s along each chromosome are *dependent*, forming a Markov chain. Specifically, for haploid data, independently for each individual  $i$ ,

$$\Pr(z_1^{(i)} = k \mid r, Q) = q_k^{(i)}, \quad (2)$$

and

$$\Pr(z_{l+1}^{(i)} = k' \mid z_l^{(i)} = k, r, Q) = \begin{cases} \exp(-d_l r) + (1 - \exp(-d_l r))q_{k'}^{(i)} & \text{if } k' = k \\ (1 - \exp(-d_l r))q_{k'}^{(i)} & \text{otherwise,} \end{cases} \quad (3)$$

where  $d_l$  denotes the genetic distance from locus  $l$  to locus  $l+1$ , assumed known. For diploid (or polyploid) data, independently for each individual  $i$ , the  $z$ s along each of  $i$ ’s 2 (or more) chromosomes form independent Markov chains satisfying equations (2) and (3).

Note that the linkage model includes the admixture model as a limiting case: as  $r$  tends to infinity in 3, all loci become independent, returning us to the original admixture model (equation 1). Note also that we assume that  $r$  is the same for all individuals, although this assumption could be relaxed at the cost of an increase in the number of parameters.

**Interpretation of the linkage model.** To provide some motivation for the linkage model, consider the following idealised scenario. Suppose that our sample comes from a diploid population that experienced a single “admixture event” followed by  $t_2$  generations of subsequent random mating within post-admixture populations. In the generation of the admixture event, individuals are formed by mating of individuals between two or more ancestral populations. These individuals inherit their DNA intact (i.e. without intervening recombination) from the ancestral populations. In the subsequent generation, the boundaries delineating these intact chunks will correspond to crossover events in a single meiosis, and so (assuming no interference) will form a Poisson process of rate 1 per Morgan. Chromosomes in each

subsequent generation will inherit chunks of DNA from chromosomes in the previous generation in a similar manner, and it follows from standard results on the superposition of Poisson processes that in chromosomes in the current generation, the boundaries between the chunks of DNA inherited intact since the admixture event will form a Poisson process of rate  $t_2$  per Morgan.

This reasoning provides some justification for the form of the transition rates in equation 3. However, it falls short of providing a complete justification for all assumptions of the linkage model, and in particular for the assumption that the ancestral populations of origin of the chunks are independent draws from some (individual-specific) vector  $q$ . Furthermore, in real populations, biological details such as crossover interference, and gene conversion events (or transformation in bacteria) will cause deviations from the assumed model. Nevertheless, the linkage model captures, in a parsimonious and computationally-convenient way, the correlations in ancestry between linked loci that we would expect to see in admixed individuals from real populations.

The discussion above also suggests an interpretation of the parameter  $r$  in terms of the number of generations since admixture first occurred. Specifically, if the genetic distances  $d_i$  between adjacent markers are measured in Morgans, then  $r$  can be interpreted as an estimate of  $t_2$ , the number of generation since the admixture event (although inevitable deviations from the model assumptions mean that it would be wise to treat any such estimate with a degree of caution). Similarly, if the genetic distances are measured in centiMorgans, then  $r$  can be interpreted as an estimate of  $100t_2$ . In some situations the genetic distances between loci may not be known, but a proxy such as physical distance may be available. If the physical distance between loci, measured in nucleotides, is used in place of the genetic distance for  $d_i$ , then  $r$  can instead be interpreted as an estimate of the product of  $t_2$  and the recombination rate (expected number of crossovers per base pair per meiosis). If there is no information on map positions, then the linkage model is not applicable.

For many datasets, we will have little prior knowledge concerning the time since admixture (and perhaps also the recombination rate). We have therefore implemented a uniform prior for  $\log r$ . The bounds of the prior should generally be set to include all biologically-plausible values of  $r$ , which may range over several orders of magnitude (partly explaining the attraction of working with  $\log r$ ).

**Computations for the linkage model.** Because in practice the  $z$ s for each chromosome are not observed, the Markov model for the  $z$ s used by the linkage model (equations 2 and 3) results in a Hidden Markov Model (HMM) for the observed genotype data. Standard HMM methods (see Rabiner, 1989, for a review) can be used to sample efficiently from the conditional distribution of the  $z$ s, and compute the probability of the observed genotype data, given all other parameters. Both these procedures are used in our MCMC algorithm to fit the model, including Metropolis–Hastings updates for  $Q$  and  $r$  (see Appendix for details).

The necessary computations are relatively straightforward when considering haploid data, or if phase relationships among all linked loci are known, as the computations for each chromosome can then be performed separately. If the data are unphased then individual allele copies may be inherited on the same chunk as any one of the allele copies at the preceding locus. This ambiguity leads to a reduction in the amount of information provided by the data on the population of origin of each allele copy, and on the value of  $r$ , in comparison with a completely phased dataset. It also increases the complexity of the algorithm required to fit the model. If relatives have been genotyped, then pedigree information may indicate that particular loci have been inherited together from the same parent, restoring some of the lost information. In the Appendix we describe algorithms that can handle unphased data for diploids, and that can also incorporate simple pedigree information when available (but note that the individuals analyzed by *structure* should all be unrelated).

Although the linkage model was developed with computational tractability in mind, it is nevertheless more computationally-intensive than the admixture model. This can make the linkage model less convenient for particularly large or complicated datasets. For the African American dataset described below (626 diploid individuals and 252 loci) and  $K = 2$  populations, a run consisting of 10,000 burn-in iterations followed by 50,000 further iterations took 3 hours using the admixture model, 7 hours for the linkage model if it was (incorrectly) assumed that the data was fully phased and 11 hours for the linkage model assuming (correctly) that the data was unphased (calculations were performed on a DEC Alpha of 2001 vintage). Performance differentials will increase for larger  $K$ : the computation scales linearly with  $K$  for the admixture model and for the linkage model with phased data, but scales with  $K^2$  for the linkage model with unphased or partially phased data.

**Models of allele frequencies.** As described above, Pritchard et al. (2000) assumed a model in which the allele frequencies in the different populations are independent. Under that model, the allele frequencies in one population provide no information about what the allele frequencies might be in another population. In practice however, we often expect the allele frequencies in closely related populations to be very similar. This suggests that a prior model that accounts for such correlations would be more accurate in many cases, and may therefore lead to improved performance on “difficult” problems, where distinct populations are quite similar. Indeed, Pritchard et al. (2000) remarked briefly that they had found some cases of rather subtle population structure where the model of independent allele frequencies was less successful than an alternative model that allowed for correlations in allele frequencies among different populations. Here, we describe the implementation of an improved model of allele frequency correlations, which has fewer parameters, and is more interpretable, than the model for correlated allele frequencies in Pritchard et al. (2000).

The new model is based on ideas in Nicholson et al. (2002), who present a model for correlated frequencies for biallelic loci (specifically, SNPs) in different populations. The

model assumes that the populations all diverged from a common ancestral population at the same time, but allows that the populations may have experienced different amounts of drift (due to different effective population sizes) since this divergence event. The assumption of simultaneous divergence, while unrealistic, is appealing in its simplicity. One alternative, which might be more realistic in some settings would be to assume that the populations are related to one another by a tree.

The new model for correlated allele frequencies that we describe here is based on the same implicit assumptions as the model of Nicholson et al. (2002), but applies to allele frequencies at multi-allelic loci. (This extension to multi-allelic loci was also suggested independently by Marchini and Cardon (2002), and is related to the parameterization of Balding and Nichols (1997).) We introduce a new (multidimensional) vector,  $P_A$ , which records the allele frequencies in a hypothetical “ancestral” population. With a slight abuse of notation we denote the frequency of allele  $j$  at locus  $l$  in the ancestral population by  $p_{Alj}$ . It is assumed that the  $K$  populations represented in our sample have each undergone independent drift away from the ancestral allele frequencies, at rates parameterised by  $F_1, F_2, F_3, \dots, F_K$ , respectively. The allele frequencies in  $P_A$  are assumed to have Dirichlet priors of the same form as that used in the original model for the (uncorrelated) population frequencies:

$$p_{Al} \sim \mathcal{D}(\lambda_1, \lambda_2, \dots, \lambda_{J_l}), \quad (4)$$

independently for each  $l$ . Here,  $\lambda$  may be fixed or estimated within the MCMC scheme. Conditional on  $P_A$ , the frequencies in each population  $k$  have a prior distribution

$$p_{kl} \sim \mathcal{D}\left(p_{Al1} \frac{1 - F_k}{F_k}, p_{Al2} \frac{1 - F_k}{F_k}, \dots, p_{AlJ_l} \frac{1 - F_k}{F_k}\right), \quad (5)$$

independently for each  $k$  and  $l$ . The size of  $F_k$  tell us about the effective population size of population  $k$  during the time since divergence, with large values of  $F_k$  indicating a smaller effective population size (Nicholson et al., 2002).

We refer to this new model for correlated allele frequencies as the “ $F$ ” model. The name is chosen to reflect the fact that there are close connections between the model and the classical measure of correlations between populations, Wright’s  $F_{ST}$  (Wright, 1951). Nicholson et al. (2002) describe the connections in some detail; here we give a brief summary. Traditionally,  $F_{ST}$  is estimated as a single value that summarizes the average deviation of a collection of populations away from the mean. Although definitions vary, it is commonly written as  $F_{ST} = \text{Var}(p_k) / \bar{p}(1 - \bar{p})$ , where  $p_k$  is the frequency of an allele in population  $k$ , and  $\bar{p}$  is the overall frequency of that allele across all subpopulations (Excoffier, 2001). In our parameterization (equations 4 and 5)  $E(p_{klj}) = p_{Alj}$ , and  $\text{Var}(p_{klj}) = F_k p_{Alj}(1 - p_{Alj})$ . Hence,  $p_{Alj}$  plays a role analogous to that of  $\bar{p}$ , and  $F_k$  plays a role like that of  $F_{ST}$  in the classical model, except that we use a generalized model with different drift rates for each population. Using a different value of  $F$  for each population, rather than a single common value for all populations, introduces a considerable amount of extra flexibility into the model at the expense of only a few additional parameters.

The prior distribution that we have implemented for  $F$  assumes that the  $F_k$  are *a priori* independent, with a density proportional to a gamma distribution truncated at 1 (so that  $\Pr[0 < F_k < 1] = 1$ ). Depending on the parameters of the distribution, the prior can be “harsh”—putting most of its weight on low values of  $F$ , or “permissive”—not discriminating strongly against any value of  $F_k$ . A harsh prior on low values of  $F$  corresponds to strong prior information that the allele frequencies in the different populations are similar to one another, and this seems generally to give the best performance in detecting subtle admixture in problems that are difficult for the independent frequencies model. However, if the values of  $F_k$  are being used to make evolutionary inferences, a permissive prior is more appropriate. In the Appendix, we present Metropolis-Hastings updates for  $P_A$ ,  $P_k$  and  $F_k$ .

## Model results using simulated datasets

In order to assess the uses and limitations of the new *structure* features, we have performed Wright-Fisher simulations, based on the seven demographic scenarios (I-VII) shown in Figure 1. Mutation parameters differ among the simulations and are specified separately for each one. Under each scenario, the goal is first to identify the current populations and secondly to reconstruct elements of their history: for example, the amount of genetic drift, the degree of admixture and the time since admixture occurred.

**Differentiating between closely related populations—The  $F$  Model.** One advantage of the new  $F$  model is that it can sometimes detect population subdivision that is invisible to *structure* when the gene frequencies of the populations are modeled without correlations. An example is shown in Figure 2, where a single random mating population splits into 2 (Scenario II). Eight generations after the split, the uncorrelated model is unable to distinguish between the two populations, while the  $F$  model distinguishes them quite accurately, with the exception of a few individuals that are not assigned with high probability to either population. After 16 generations of separate evolution, the uncorrelated model becomes able to distinguish reasonably accurately between the two populations and the  $F$  model provides little improvement in clustering. Further simulations (not shown) indicate that the  $F$  model is less likely to improve performance when the number of loci is small; rather it can allow accurate clustering of individuals from extremely closely-related populations when large numbers of markers are used (e.g., Rosenberg et al., 2002). However, even with very many markers, it is unlikely to be possible in practice to differentiate between biological populations that have split  $< 20$  generations ago unless there has been little subsequent admixture and genetic drift has been strong.

**Estimation of  $K$ .** Simulations presented by Pritchard et al. (2000), and further experience with other datasets, suggests that the value of  $K$  that maximises the estimated model log-

likelihood  $\log(P(X | K))$  is often a sensible choice for the number of populations. However the value of  $K$  estimated by this procedure can sometimes depend on the model used. The  $F$  model is in general more permissive of additional populations being fitted to a data-set, as it permits the existence of two or more populations with very similar allele frequencies (particularly if the prior on  $F$  is chosen to favour small values). Consequently,  $P(X | K)$  is sometimes maximised for a higher value of  $K$  than under the uncorrelated model. This cuts to the heart of one of the principal reasons why inferring  $K$  is so difficult, and why estimates for  $K$  should be treated with caution: the number of populations supported by the data may depend on how different one would expect allele frequencies in the different populations to be *a priori*, which is often difficult to specify.

For some datasets, higher estimates of  $K$  obtained using the  $F$  model may reflect deviations from random assortment that are not caused by genuine population subdivision. Table 1A shows model likelihoods estimated for a single panmictic population (Scenario I). Whether or not the  $F$  model is used, the highest value of  $P(X | K)$  is given by  $K = 1$ . In Table 1B the evolutionary parameters are identical but there is a 50% selfing rate. In this case, the  $F$  model gives higher probabilities for  $K = 2$ , while the original model continues to give the highest model likelihood for  $K = 1$ . Other situations that might cause additional populations to be inferred by *structure* (with or without the  $F$  model) include significant frequency of inbreeding, cryptic relatedness within the sample, or the presence of null alleles.

**Inference of demographic history.** The  $F$  model can also be used to estimate the amount of genetic drift undergone by the different populations under study. In Figure 3A, estimates of  $F$  are shown for a population that trifurcated (Scenario III). For a substantial time period after the trifurcation, the estimated values of  $F$  are approximately proportional to the time since the split and inversely proportional to the population sizes. When the values of  $F$  start to exceed about 0.2,  $F$  no longer increases linearly but the ranking of the values of  $F$  continues to reflect the relative degrees of drift that the populations have undergone.

The use of the  $F$  model to estimate drift is subject to a caveat, which is that contrary to the model assumption, drift may not have occurred independently in each population. For example, Figure 3B shows results based on Scenario IV, in which a single population divides into two and one of the populations subsequently subdivides. The *structure* algorithm interprets the similarity of the two sub-populations as evidence that their gene frequencies are close to those of the ancestor of all three populations and estimates lower values of  $F$  for them than for their common ancestor prior to subdivision.

In principal it should be possible to generalize the model to allow for the possibility of hierarchical subdivision but we do not attempt this here. Rather, we suggest testing for deviations from the model by estimating values of  $F$  while excluding one or more of the populations in turn. If the assumption that all of the populations evolved independently from a single common ancestral population is correct, then this should leave the  $F$  values

estimated for the other populations approximately unchanged. If the  $F$  values decrease, then this suggests that one of the excluded populations diverged first, so that the remaining populations share a more recent common ancestor than the whole sample considered together. If  $F$  values of one or more of the populations increase, then it may indicate that the original  $F$  values were artificially reduced by the presence of closely related sub-populations in the sample. Other diagnostics are discussed by Nicholson et al. (2002).

**Inference in admixed populations—The linkage model.** Inference of demographic history becomes more difficult if admixture has occurred subsequent to population divergence (e.g., Beaumont et al., 2001). In these situations, it may be very difficult to estimate allele frequencies in the original “pure” populations, as required for standard population genetic analyses. When analyzing data using *structure*, such populations can also be problematic. The algorithm estimates  $P$  and  $Q$  jointly, so the model can either simultaneously underestimate the degree of admixture and the genetic distance between pairs of populations, or simultaneously over-estimate the same quantities.

Linkage information can help to resolve the ambiguity. Informally, admixed individuals contain chromosomal chunks that derive from one population or another. Using closely linked markers, the linkage model aims to detect the chromosomal chunks and can potentially reconstruct the ancestral populations accurately even if no pure members exist.

In order to explore the properties of the new method, we have performed extensive simulations. We consider individuals genotyped at  $L^*$  loci on each of  $C$  chromosomes (i.e., typed at a total of  $CL^*$  loci). The loci are equidistant, with a recombination rate  $R$  per generation between adjacent genotyped sites. The genetic map is assumed known. We analyzed the simulated data using the uncorrelated model for allele frequencies.

**Estimation of allele frequencies.** One measure of whether *structure* is performing well is if it can accurately estimate the population allele frequencies in the ancestral populations. To visualize this, we have constructed Neighbor-Joining trees based on the posterior mean allele frequencies. When the allele frequencies are accurately estimated, the branch tips lie close to the large black dots (which represent the “correct” frequencies).

We focus on Scenario VI, “uni-directional” admixture. In 3 out of the 4 cases shown (Figure 4A-C), *structure* is highly consistent in its inference of gene frequencies for ancestral population 1, reflecting the continuous presence of pure individuals in the sample. The accuracy of this inference provides a baseline from which to judge the performance of *structure* in disentangling the gene frequencies of ancestral population 2, which ceases to have any pure descendents a few generations after admixture.

In the first simulated example (Table 2A, Figure 4A), as the number of generations after admixture ( $t_2$ ) increases, the admixture model becomes increasingly biased, under-estimating the divergence between the populations (shown by the intermediate position of the inferred

populations in the gene frequency tree) and under-estimating the amount of admixture ( $H_2$  in Table 2A). In contrast, the linkage model estimates gene frequencies and the degree of admixture accurately for many generations after the admixture event.

The performance of the admixture model is improved by increasing the number of chromosomal regions studied (Table 2B, Figure 4B) but the linkage model continues to prolong the number of generations after admixture for which accurate ancestry estimates can be obtained.

In another example (Table 2C, Figure 4C), the admixture model shows the opposite bias for a number of generations, over-estimating rather than under-estimating admixture and the degree of divergence between ancestral populations. The linkage model again uses linkage information to resolve the ambiguity and it performs well up to 8 generations after the admixture event. However, in this example, the marker density is low enough that in later generations, the linkage information is lost, and the admixture and linkage models produce almost identical (and similarly biased) results.

By contrast, Figure 4D (see also Table 2D) shows the situation where the marker density is very high. In this case, background LD is substantial, leading the linkage model to consistently over-estimate the divergence between the two populations. Further, substantial admixture is estimated for population 1, which is in fact pure. The admixture model actually does rather better in the populations it infers, but a few generations after the admixture event it also produces misleading results. These results illustrate the problems that can arise when there is substantial background LD, a point to which we return in the Discussion.

**Estimating the time since admixture.** In addition to improving estimates of the degree of admixture, the linkage model also provides an indication of the time since admixture. For examples A–C the value of  $r$  (Table 2) provides good estimates of the number of generations since admixture, except immediately after the admixture event (when little admixture has occurred, so that there is not yet much admixture LD and the posterior for  $r$  is uninformative) and more than 100 generations after admixture, when the number of generations is underestimated. The time of admixture may be considerably over-estimated by  $r$  if there is substantial background LD in the sample (2D).

**Population-of-origin assignments for chromosomal regions.** A further advantage of the linkage model is that if the marker density is high enough, it can provide accurate population-of-origin assignments for chromosomal regions, as required in applications such as admixture mapping. For example, Figure 5 shows population of origin assignments for the two allele copies at each locus of a single diploid individual. The Markov structure of the data is clearly evident from the *structure* output, in that nearby loci typically have similar assignment probabilities. When the data are phased, individual loci are often assigned with very high probability to the correct ancestral population, especially in the middle of a large

chunk inherited from one population. At boundaries between chunks, the assignment probabilities typically change rapidly, giving a good indication of the position of the recombination event that brought the chunks from different populations together.

For unphased diploids, the data contain somewhat less information about the population of origin of individual allele copies, particularly in regions of the genome where the two homologous chromosomes are spanned by chunks inherited from different ancestral populations. In these regions, neighbouring loci do not provide information concerning which of the two allele copies at a particular locus comes from one population and which comes from the other. For many problems (such as in admixture mapping) we are mainly interested in inferring the number of allele copies from each population, and this information can be extracted from the data, given sufficiently dense markers (as in Figure 5).

**Coverage properties.** A final advantage of the linkage model is that it gives more accurate estimates of the statistical uncertainty of admixture proportions. This property is illustrated in Figure 6, which shows 90% credibility regions for  $q$  for a sample of individuals from two populations. The two populations partially admixed with each other in an admixture event (Scenario VII) 32 generations before the sample was taken. After 32 generations of random mating within each post-admixture population, the ancestry coefficients of each individual are almost identical (differing by less than 0.001) and are shown by the red horizontal lines in the figure. Ancestry estimates were made by both the admixture and linkage models, for markers at a variety of genetic distances. Tightly linked markers give non-independent information and are therefore less informative about the value of  $q$  for each individual than the same number of unlinked markers, leading to higher variation in estimates between individuals. The admixture model does not take these correlations into account. Consequently, the size of the estimated credibility regions are approximately independent of the actual degree of linkage and are much too narrow for tightly linked markers. Under the linkage model, the credibility regions for  $q$  become wider as linkage between markers increases and continue to reflect the true degree of statistical uncertainty even for tightly linked markers.

## Applications to Data

**Recombination between distinct populations of *Helicobacter pylori*.** The bacterium *Helicobacter pylori* colonizes the human stomach lining. When multiple strains infect the same stomach, they recombine rapidly through the import of fragments of DNA that are typically a few hundred base pairs in length (Falush et al., 2001). Falush et al. (2003) used the no-admixture model on multilocus sequence data from a global collection of strains to show that there are four major modern populations of bacteria, which they named “hpAfrica1”, “hpAfrica2”, “hpEastAsia” and “hpEurope”, on the basis of their geographical distributions.

Here, we use results from that analysis to illustrate site-by-site (in this case nucleotide-by-nucleotide) population-of-origin assignment in haploid organisms.

Figure 7 shows results for a typical isolate from South Africa. South Africa contained isolates from hpAfrica1, hpAfrica2 and hpEurope populations, reflecting the ethnic diversity of the region. The particular isolate we consider here was assigned to the Africa1 (blue) population by the no-admixture model. The upper plot (7A) shows the posterior assignment probabilities for each individual nucleotide based purely on the estimated population allele frequencies (i.e., not using information from  $q$ ). The plot shows that most sites provide little information about ancestry, with assignment probabilities to all four population being around 0.25. The remaining nucleotides are mostly assigned with high probability to Africa1 or, less frequently, Africa2 (red). The Africa2 nucleotides appear to come in runs, suggesting import of specific DNA fragments into a bacterium from the Africa1 population. This conclusion was confirmed by further exploratory analysis (Figure 7B). For each population, the sum of the log of the assignment probabilities was computed within a 100 nucleotide moving window. For most of the sequence, the value of the sum was positive for the Africa1 population (indicating higher probabilities than under random assignment) and negative for the other three populations. However, there are three stretches where it is the sum for the Africa2 population that gives positive values, suggesting DNA import into those regions.

The linkage model provides a formal method to make population-of-origin assignments that take the linkage relationships into account (Figure 7C). Nucleotides in the three regions identified by the exploratory analysis were assigned to the Africa2 population with probabilities close to 1.0, providing statistical support for the conclusion that there have been (at least) three imports of Africa2 DNA into these fragments. This example shows that given highly-differentiated populations and enough informative sites, it can be possible in practice to make accurate population-of-origin assignments for individual loci. Further, because of the large amount of information provided by linkage, it is also possible to reconstruct ancestral populations in the absence of pure individuals, using the linkage model. See Falush et al. (2003) for details of this analysis.

**Admixture LD in African Americans.** We used the new linkage model to study the extent of admixture linkage disequilibrium in a Chicago-based population of African Americans. Previous work on African Americans has shown significant levels of European admixture in the range of around 5% to 25%, with substantial variation across studies and across study populations (summarized by Parra et al., 1998). Since the admixture is quite recent (primarily in the last 200 years or so; Parra et al., 1998; Pfaff et al., 2001), it is likely that admixture LD extends over large distances, and hence could be useful for gene mapping (Chakraborty and Weiss, 1988; Stephens et al., 1994; McKeigue, 1998). Indeed, Parra et al. (1998) detected admixture LD across 22 cM between two markers that have extremely large frequency differences between Africans and Europeans, in several African American populations.

The dataset that we used consists of 247 microsatellites genotyped in samples of unrelated individuals including 210 African Americans (from Maywood, Illinois), 158 European Americans (from Michigan) and 308 Nigerians (Yoruba) (Cooper et al., 2002; Thiel et al., 2003). The data were kindly provided by R. Cooper, A. Chakravarti, N. Schork and A. Weder. We analyzed the data using the linkage model described above, without haplotype phase information. Genetic distances between markers were estimated using the Marshfield linkage map (Broman et al., 1998). The average spacing between adjacent markers on the same chromosome was 13 centiMorgans with a minimum of 2.14 and a maximum of 42.66. While the markers are not especially close together, this density should still be informative about admixture that occurred on the order of 10 generations ago, because under those conditions, adjacent markers would frequently be inherited on the same ancestral chunk.

When run with  $K = 2$ , both the admixture and linkage models gave very similar ancestry estimates and both suggested that Nigerians and American whites were almost pure representatives of the respective pre-admixture populations, with average estimated admixture rates of 1.4% for both populations. The African Americans were substantially admixed, having a mean of 17.8% European ancestry (the range of point estimates for individuals' values of  $q$  was 2% to 59%, using the admixture model). Similar results were obtained if we used the USEPOPINFO=1 option to specify the population of origin of the American whites and Nigerians. Our estimate of 17.8% European ancestry is very similar to the estimate of 18.8% obtained by Parra et al. (1998) for their sample from this population.

The posterior distribution for the parameter  $r$  under the linkage model is shown in Figure 8A. The posterior mean of  $r$  was 0.098 chromosome chunk break-points per centiMorgan, with a 90% credible region of 0.07 to 0.13. Under the simplifying assumption that the African American population was created by a single hypothetical admixture event (Scenario V), this event is estimated to have taken place 7-13 generations ago. This is consistent with what one might expect, based on the history of African Americans, who were mostly exported from Africa during the late 18th century (Parra et al., 1998), and have undergone some degree of continuous admixture with the European American population since then.

We repeated our analysis using information provided by map distances from the recently published deCODE map (Kong et al., 2002), and obtained similar answers (data not shown). In order to assess the potential impact of map misspecification more generally, we applied *structure* to a series of simulated datasets with inaccurately specified maps (Figure 9). The simulations assumed that the estimated distances between markers are unbiased on average, but may be inaccurate. The results indicate that as long as the chromosome chunks are typically larger than the inter-marker distances, then even highly inaccurate maps do not lead to biases in  $r$ .

The posterior distribution of  $r$  (Figure 8A) clearly excludes large values of  $r$ , indicating that we are detecting a significant signal of admixture LD. Recall that as  $r$  gets large the linkage model becomes equivalent to the admixture model, so the fact that the posterior for

$r$  excludes large values shows that the linkage model provides a better fit to the data than does the admixture model in this case. For comparison, Figure 8B shows the posterior for  $r$  for the same data and map distances, but with the order of the loci randomised. The posterior for  $r$  has considerable support all the way up to the maximum value of  $r$  permitted by the prior, and would clearly have extended to still larger values had the prior allowed this. Three further randomizations produced similar results, supporting the effectiveness of the posterior for  $r$  in summarising the extent of admixture LD.

Although we detected a definite signal of admixture LD in our sample, most of the LD present in the African Americans is actually due to variation in  $q$ : i.e. “mixture LD” in the terminology we introduced earlier. To differentiate between mixture and admixture LD, we examined the correlations of ancestry estimates (from the admixture model) between adjacent loci. The first measure that we used (Figure 10A) measures the correlation between the estimated probability of African ancestry (averaged over the two allele copies at each locus) for pairs of neighboring loci. The correlations were positive on average (mean 0.041 with standard error 0.009), albeit with a great deal of variation between different locus pairs. These correlations reflect the total LD in the sample. The second measure (Figure 10B) shows the correlations that remain when variation in  $q$  among African Americans is accounted for. For each individual, at each locus, we computed a “residual” by subtracting the individual’s estimated  $q^{(i)}$  from the estimated probability of African ancestry (averaged over the two allele copies at each locus). The figure shows the correlations of these residuals. The correlations are slightly but not significantly negative on average (mean -0.001 with standard error 0.007), implying that most of the LD in the sample can be accounted for by variation in  $q$  (i.e. mixture LD). Further, a regression of the correlations with the genetic distance between the loci does not have a significant slope under either measure, presumably because the trend has been obscured by the high degree of variation in correlation values at each genetic distance. The fact that linkage model obtains plausible estimates of  $r$ , and rejects large values of  $r$ , indicates that the linkage model extracts much more information from the data than do the pairwise comparisons.

Our results also highlight an important feature of human genetic data, which is that there is a great deal of noise in raw LD estimates for individual locus pairs, even when the admixture involves populations that, by human standards, are relatively highly differentiated. Thus, our description of admixture LD in this population would be enhanced by using a denser set of markers, and for applications such as admixture mapping where one needs to estimate the population-of-origin for the sampled chromosomes, a denser marker set would be critical. In admixed populations where loci have been chosen specifically to have large frequency differences between the putative parental populations (Parra et al., 1998), these loci may be more informative for studying admixture than microsatellites were here.

**Genetic drift in *Drosophila melanogaster*.** In order to illustrate some of the possible ways of using the  $F$  model in historical inference, we have reanalyzed the dataset of Agis and Schlötterer (2001). Their dataset consisted of 48 microsatellite loci typed in samples of flies from Israel (1 sampling location, 58 genotypes), New South Wales, Australia (8 sampling locations, 190 genotypes) and Tasmania (2 sampling locations, 53 genotypes). Some of the genotypes came from isofemale lines, so we haplodized the dataset by randomly discarding an allele copy at each locus. In the original analysis, Agis and Schlötterer found that the Tasmanian population had similar high pairwise  $F_{ST}$  values with both Australian and Israeli populations. In contrast,  $F_{ST}$  between Australian and Israeli flies was quite low (though statistically significant), as is typical for populations of *D. melanogaster*. Tasmanian flies had lower genetic diversity than the other groups, but statistical tests for bottlenecks based on the allelic distribution in individual locations were inconclusive, not providing clear evidence for stronger drift within the Tasmanian lineage. Since their analysis did not establish any particular connection between the Australian and Tasmanian populations, Agis and Schlötterer could not rule out the possibility that Tasmanian flies might have come from an entirely different source population with distinct gene frequencies.

We started the analysis without making any assumption about geographical clustering. Using the no-admixture and  $F$  models,  $K = 3$  gave the highest model likelihood. The three inferred populations correlated well with the land masses Israel, Tasmania and Australia, although many flies were not clearly allocated to one population and 50 Australian flies, 2 Tasmanian flies and 7 Israeli flies were assigned to their home population with less than 50% probability. These inconsistencies are due to limited statistical power rather than identifiable admixture events because an analysis under the migration model with the USEPOPINFO option (Pritchard et al., 2000), in which each genotype was assigned a 5% prior probability of being admixed, did not identify any evidence for admixture. The value of  $F$  inferred for the Tasmanian population was high, consistent with a strong episode of genetic drift (Figure 11A). The much lower values of  $F$  for the other two populations and the weaker correlation of assignments with geography, indicate that there has been no comparable episode of genetic drift separating Israeli and Australian flies.

Where did the Tasmanian flies come from? We estimated  $F$  values when analyzing Tasmanian and Israeli flies together (Figure 11C) and Tasmanian and Australian flies together (Figure 11D). The value of  $F$  inferred for the Australian population was close to zero, much lower than for the Israeli population, while a high value of  $F$  was estimated for the Tasmanian population in both cases. This analysis suggests that Tasmanian and Australian flies share a more recent common ancestor with each other than with the Israeli flies. Further, the amount of drift that the Australian flies have undergone since splitting with the Tasmanian flies is very low, implying that Tasmania was colonized from Australia and underwent a bottleneck in the process. A possible technical objection to our analysis is that flies were sampled in several locations in Australia and that this might somehow account for the par-

ticularly low estimated value of  $F$ . We tested for this possibility using the five sampling locations in Australia which had more than 25 genotypes. We ran *structure* separately for each one, using all of the Tasmanian genotypes in every case. The analysis gave a consistently high value of  $F$  for the Tasmanian population (0.100 to 0.125) and a low value for the Australian population (0.004 to 0.023)—lower than for the Israeli population in the equivalent analysis (0.039). Our inference therefore appears to be robust to the exact combination of populations chosen for analysis.

The approach we have taken is similar to that used by Nicholson et al. (2002), who calculated drift parameters for populations that were defined on the basis of geographical labels (this type of analysis can be performed using *structure* by using the program option USE-POPINFO). For this dataset utilizing geographical labels does not change the conclusions concerning the origin of the Tasmanian flies, although it does lead to different estimates for the extent of genetic drift, estimating higher values of  $F$  for the Israeli population (0.055 vs 0.026) and lower values of  $F$  for the Australian one (0.015 vs 0.022) when all three populations are included in the analysis. These results show that when clustering is inaccurate due to limited divergence between populations, this can significantly affect  $F$  estimates. The advantage of the approach taken here is that no assumption is made at the outset about the likely patterns of geographical partitioning. If the actual population structure is either independent of geography, for example because of cryptic speciation, or the boundaries to gene flow are unexpected, then this will be picked up by the analysis. In the *Drosophila* example, the most likely cause of differentiation is one which might have been expected at the outset - a founder effect during the colonization of an island from a mainland source.

## Discussion

We have presented two major modifications of the *structure* approach, namely the linkage model and the  $F$  model. Both can significantly improve the technical quality of the inference, giving better clustering, more realistic confidence limits and more accurate admixture estimates. For some datasets, these improvements are critical, allowing the detection of population substructure or admixture that was invisible using the earlier algorithm. The new models are also an important step in making *structure* into a tool for performing detailed historical inference.

The linkage model allows *structure* to analyze datasets containing markers with admixture LD between them, significantly expanding the range of datasets for which it is an appropriate tool. For datasets with weak admixture LD (e.g., the African American example), the linkage model gives similar clustering and ancestry estimates to the admixture model, but also estimates a chunk size parameter  $r$  that provides information on the average rate of decay of admixture LD in the sample. The rate of decay reflects the amount of time that has elapsed since populations admixed. For very informative datasets (e.g. the *Helicobacter*

*pylori* data), the linkage model also allows accurate assignment of chromosomal chunks to ancestral populations. One consequence is that ancestral populations can be reconstructed even if admixture has been so extensive that no pure individuals remain in the sample. In highly informative datasets, it might also be useful to relax the assumption of a single value for  $r$ . Possible ways forward include using a different  $r$  for each individual or geographical location, and/or allowing the expected chunk size to depend on the population of origin of the chunk. In this way, it might be possible to extract additional information about the timing of admixture between different sub-populations.

Although the linkage model takes into account the correlations between markers that occur due to admixture, *structure* will always need data from several unlinked or weakly linked genetic regions from each individual in order to make meaningful inferences. We strongly recommend against using the model for a dataset consisting of human Y chromosome or mitochondrial haplotypes, for example. As well as having markers from several genomic regions, it is also important that none of the markers be too strongly linked. Background linkage disequilibrium arises through genetic drift within populations, and in some scenarios this can lead *structure* to produce misleading results.

Background LD causes problems when particular allele combinations are overrepresented in two or more of the populations before admixture takes place. Such LD can arise through genetic drift before the populations separate (i.e. during the burn-in phase in Scenarios V, VI, and VII in Figure 1). If the markers are tightly linked then this LD can persist throughout the period of divergence and subsequent admixture. Suppose that for this reason, allele combinations  $ab$  and  $AB$  are overrepresented in both of the populations, compared to  $Ab$  and  $aB$ . The linkage model tries to attribute this LD to admixture and hence tends to overestimate the frequency of alleles  $a$  and  $b$  in one population, and the frequency of alleles  $A$  and  $B$  in the other. In this manner, *structure* can both overestimate the divergence between ancestral populations and infer spurious admixture. Because background LD decays over short distances, this also leads to over-estimation of the time since admixture.

In designing a dataset to be used by the linkage model, it is therefore desirable to ensure that the markers are sufficiently closely linked to allow for admixture LD, yet sufficiently far apart that there is not substantial background LD between them. Historical information about the likely time of admixture, combined with knowledge of inter-marker recombination rates, can be used to help select an appropriate marker spacing.

An alternative, post-hoc, approach is to rely on inspection of *structure* output. One observation that we have made from simulations (based on a variety of demographic scenarios; data not shown), is that when background LD is a problem, *structure* infers that *all individuals* from *both* populations are admixed. Genuine admixture, by contrast, is often asymmetrical, affecting some populations much more than others. For example, if there is a pre-defined sub-population in the sample that has substantial ancestry from one *structure* population without corresponding ancestry from a second, then this is an indication that

background LD is not a serious bias. Further when  $K$  is greater than 2, background LD causes admixture to be inferred preferentially between the most closely related *structure* populations; thus if admixture is inferred between distantly related populations, this result is unlikely to result from background LD.

In future, it should be possible to specifically model the background LD present in the data. Full coalescent approaches to this type of problem are computationally daunting, and so models that simplify the structure of the data may be suitable. One approach would be to use a Markov formulation of haplotype structure, perhaps like that used by Daly et al. (2001). This model would split each chromosome into segments that contained background LD, so that any LD between segments would be due to admixture. Another, more ambitious, approach would be to approximate patterns of LD using the kinds of ideas applied in a quite different setting by Stephens et al. (2001).

This paper also introduced the new  $F$  model for correlated allele frequencies. This update improves clustering for some datasets where populations are very weakly differentiated and also allows inference of the pattern of drift. We used the model to show that Tasmania was almost certainly colonized by Australian *Drosophila melanogaster*, and that a significant bottleneck occurred in the process. A weakness of the current model is that it assumes that all populations have evolved by independent drift from a single ancestral source population which may not always be a good approximation. This framework could potentially be generalized to incorporate, and differentiate between, more complicated scenarios, with populations splitting from each other sequentially.

The new implementation of *structure* (version 2) also contains other new options, that can be useful for some datasets. In the earlier version, the gene frequencies  $P$  and admixture proportions  $Q$  were drawn from symmetric Dirichlet distributions with parameters  $\lambda$  and  $\alpha$  respectively.  $\lambda$  was normally fixed at 1.0 which corresponds to a uniform prior on allele frequencies. For some types of markers (e.g. sequence polymorphisms and SNPs, depending on the process of ascertainment), the frequency spectrum may be skewed towards rare alleles. In this situation, the data are better modelled by smaller values of  $\lambda$  (smaller values of  $\lambda$  place more prior weight on configurations where all but one allele at a locus is rare). Therefore, we have implemented a Metropolis-Hastings update for  $\lambda$ . We have found that for data where most alleles are rare, updating  $\lambda$  can lead to more accurate estimates of  $P$ . We now also allow different values of  $\alpha$  for each population. This generalized prior for  $Q$  allows for the fact that in practice not all populations are equally represented in the sample, which may lead to more accurate ancestry estimates, particularly in situations of highly asymmetric admixture. The no-admixture model assumes a prior probability that each individual is drawn from one of the  $K$  populations is  $1/K$ . It would also be possible to generalize this prior to allow for differences in frequency between populations but this option has not yet been implemented.

The large number of model options that have now been implemented might lead to concerns about over-parameterization. However, provided that there are a reasonable number

of loci ( $> 10$ ), the number of model parameters added by each model option is small in comparison with both the number of elements in the dataset and the total number of parameters estimated by *structure* (which include all the elements of  $P$  and  $Q$ ). Indeed, although the  $F$  model may appear to increase the number of parameters to be estimated by the model, in some sense it *decreases* the effective number of parameters, by introducing correlations among the allele frequencies in the different populations.

Nevertheless, there may not be enough information in a particular dataset to estimate all of the parameters, in which case a simpler model would be more appropriate. The linkage model adds only a single parameter,  $r$ , but for some datasets there may be little or no admixture LD present (indicated by posterior support for large values of  $r$ ), in which case the admixture model should provide essentially the same answers (and be faster to run). It is also possible that certain parameter *combinations* could lead to problems. For example, our experience with simulated data suggests that it is sometimes difficult to estimate both  $\lambda$  and  $F$  jointly (perhaps because estimated values for  $\lambda$  are sensitive to the estimates for the ancestral allele frequencies  $P_A$ , which in turn are sensitive to estimates of  $F$ ). In such cases we have found it useful to first estimate  $\lambda$  using the old independent model for population allele frequencies, and then to fix  $\lambda$  at its estimated value while using the  $F$  model for correlated allele frequencies.

In many cases, symptoms that might be attributed to over-parameterization in fact provide useful indicators of genuine biological uncertainty. For example, the model used for gene frequencies sometimes has a large effect on the estimates of admixture proportions that are obtained. In Figure 2 the estimates of admixture between the populations (in fact 0) are much more accurate under the  $F$  model than under the uncorrelated model. However, neither model makes accurate estimates for the proportion of European ancestry of the African Americans if Nigerians are excluded from the sample. The uncorrelated model overestimates admixture (28%) while the  $F$  model under-estimates it (3%). For both of these datasets, the large difference in admixture estimates between models provides an indication that there is limited information on the true degree of admixture. In contrast, when Nigerians are added to the latter dataset, both models give very similar estimates (18% and 17%) for the European ancestry of African Americans, reflecting the additional information about African gene frequencies provided by the un-admixed individuals.

Despite the many complexities in the demography of species, their genetic structure can often be approximated by a division into discrete (or semi-discrete) populations. This approximation greatly simplifies statistical inference, allowing assignment of individuals to populations, with modest computational requirements even for large datasets. Here, we have extended the approach to gain greater information on the history of the populations themselves. In effect, we are approximating a complex history by a narrative of population fissions and fusions. The resulting algorithms remain computationally tractable and can potentially help resolve the major events in the history of the species under study.

## Acknowledgements

We thank Aravinda Chakravarti, Richard Cooper, Nicholas Schork and Alan Weder for kindly allowing us to use the African American dataset (and Richard Cooper in particular for help interpreting the data), Christian Schlötterer for the *D. melanogaster* data and comments on the manuscript, Sebastian Suerbaum and Mark Achtman for permission to use the *H. pylori* data, Peter Donnelly and David Balding for helpful discussions of  $F_{ST}$ , Martin Vingron for use of a computer cluster at the Max Planck Institute for Molecular Genetics, Giovanni Montana for comments, and William Wen for help with data manipulation. Two anonymous reviewers read the manuscript carefully and made many useful suggestions. The work was supported by grants from the National Science Foundation (BIR-9807747 to MS), Burroughs Wellcome Fund (JKP), and from the National Heart, Lung and Blood Institute (#54512 and 54485 for collection of the African American data).

## References

- Agis, M. and Schlötterer, C. (2001). Microsatellite variation in natural *Drosophila melanogaster* populations from New South Wales (Australia) and Tasmania. *Molecular Ecology*, 10:1197–1205.
- Anderson, E. C. (2001). *Monte Carlo methods for inference in population genetic models*. PhD thesis, University of Washington.
- Anderson, E. C. and Thompson, E. A. (2002). A model-based method for identifying species hybrids using multilocus genetic data. *Genetics*, 160:1217–1229.
- Balding, D. J. and Nichols, R. A. (1997). Significant genetic correlations among Caucasians at forensic DNA loci. *Heredity*, 78:583–589.
- Barton, N. H. and Hewitt, G. M. (1989). Adaptation, speciation and hybrid zones. *Nature*, 341:497–503.
- Beaumont, M., Gottelli, D., Barratt, E. M., Kitchener, A. C., Daniels, M. J., Pritchard, J. K., and Bruford, M. W. (2001). Genetic diversity and introgression in the Scottish wildcat. *Molecular Ecology*, 10:319–336.
- Bertorelle, G. and Excoffier, L. (1998). Inferring admixture proportions from molecular data. *Mol. Biol. Evol.*, 15:1998–1311.
- Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. (1998). Comprehensive human genetic maps: individual and sex-specific variation in recombination. *Am J Hum Genet*, 63:861–9.
- Chakraborty, R. and Weiss, K. M. (1988). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc. Natl. Acad. Sci. USA*, 85:9119–9123.
- Chikhi, L., Bruford, M. W., and Beaumont, M. A. (2001). Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics*, 158:1347–1362.
- Cooper, R. S., Luke, A., Zhu, X., Kan, D., Adeyemo, A., Rotimi, C., Bouzekri, N., and Ward, R. (2002). A genome-wide scan among Nigerians linking blood pressure to regions on chromosomes 2, 3, and 19. *Hypertension*, 40:629–633.
- Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet*, 29:229–32.

- Dawson, K. J. and Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet Res*, 78:59–77.
- Erosheva, E. A. (2002). *Grade of Membership and Latent Structure Models With Application to Disability Survey Data*. PhD thesis, Department of Statistics, Carnegie Mellon University.
- Excoffier, L. (2001). Analysis of population subdivision. In Balding, D., Bishop, M., and Cannings, C., editors, *Handbook of Statistical Genetics*, pages 271–307. Wiley.
- Falush, D., Kraft, C., Taylor, N. S., Correa, P., Fox, J. G., Achtman, M., and Suerbaum, S. (2001). Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proc Natl Acad Sci U S A*, 98:15056–61.
- Falush, D., Wirth, T., Linz, B., Pritchard, J. K., Stephens, M., and others], . (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science*, 299:1582–1585.
- Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- Guglielmino, C. R., Piazza, A., Menozzi, P., and Cavalli-Sforza, L. L. (1990). Uralic genes in Europe. *Am. J. Phys. Anth.*, 83:57–68.
- Knowler, W. C., Williams, R. C., Pettitt, D. J., and Steinberg, A. G. (1988). Gm3-5,13,14 and Type-2 Diabetes-Mellitus—an association in American-Indians with genetic admixture. *Am. J. Hum. Genet.*, 43:520–526.
- Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., Shlien, A., Palsson, S. T., Frigge, M. L., Thorgeirsson, T. E., Gulcher, J. R., and Stefansson, K. (2002). A high-resolution recombination map of the human genome. *Nat Genet*, 31:241–7.
- Kumar, S., Tamura, K., Jakobsen, I. B., and Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics*, 17:1244–5.
- Long, J. C. (1991). The genetic structure of admixed population. *Genetics*, 127:417–428.
- Marchini, J. L. and Cardon, L. R. (2002). Discussion on statistical modelling and genetic data. *J. Roy. Stat. Soc. B*, 64:740–741.
- McKeigue, P. M. (1998). Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am. J. Hum. Genet.*, 63:241–251.

- McKeigue, P. M., Carpenter, J. R., Parra, E. J., and Shriver, M. D. (2000). Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann. Hum. Genet.*, 64:171–186.
- Nei, M. and Li, W.-H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. USA*, 76:5269–5273.
- Nicholson, G., Smith, A. V., Jónsson, F., Gústafsson, O., Stefansson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single nucleotide polymorphism data. *J. Roy. Stat. Soc. B*, 64:695–715.
- Parra, E. J., Marcini, A., Akey, J., Martinson, J., Batzer, M. A., Cooper, R., Forrester, T., Allison, D. B., Deka, R., Ferrell, R. E., and Shriver, M. D. (1998). Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.*, 63:1839–1851.
- Pfaff, C. L., Parra, E. J., Bonilla, C., Hiester, K., McKeigue, P. M., Kamboh, M. I., Hutchinson, R. G., Ferrell, R. E., Boerwinkle, E., and Shriver, M. D. (2001). Population structure in admixed populations: effect of admixture dynamics on the pattern of linkage disequilibrium. *Am. J. Hum. Genet.*, 68:198–207.
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.
- Rieseberg, L. H., Whitton, J., and Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, 152:713–727.
- Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002). Genetic structure of human populations. *Science*, 298:2381–5.
- Satten, G. A., Flanders, W. D., and Yang, Q. (2001). Accounting for unmeasured population structure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.*, 68:466–477.
- Sillanpää, M. J., Kilpikari, R., Ripatti, S., Onkamo, P., and Uimari, P. (2001). Bayesian association mapping for quantitative traits in a mixture of two populations. *Genetic Epidemiology*, 21:692–699.
- Sites, J. W., Barton, N. H., and Reed, K. M. (1995). The genetic-structure of a hybrid zone between 2 chromosome races of the *Sceloporus grammicus* complex (Sauria, Phrynosomatidae) in central Mexico. *Evolution*, 49:9–36.

- Stephens, J. C., Briscoe, D., and O'Brien, S. J. (1994). Mapping admixture linkage disequilibrium in human populations: limits and guidelines. *Am. J. Hum. Genet.*, 55:809–824.
- Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68:978–989.
- Thiel, B. A., Chakravarti, A., Cooper, R. S., Luke, A., Lewis, S., Lynn, A., Tiwari, H., Schork, N. J., and Weder, A. (2003). A genome wide linkage analysis investigating the determinants of blood pressure in Caucasians and African Americans. *Am J Hypertension*, 16:151–153.
- Thompson, E. A. (1973). The Icelandic admixture problem. *Ann. Hum. Genet. Lond.*, 37:69–80.
- Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., and Buckler, E. S. (2001). *Dwarf8* polymorphisms associate with variation with flowering time. *Nature Genetics*, 28:286–289.
- Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.*, 15:323–354.

# Appendix

Here we provide the computational details for the new MCMC updates. Recall that our goal is to sample from the joint posterior distribution of

$$\Pr(P, Q, Z, r, F, \alpha, \lambda \mid X, K). \quad (6)$$

MCMC methods provide an approach for doing this. We start by making arbitrary initial choices for each parameter, and then propose updates that change a subset of these, conditional on the other parameters and the data. One full iteration of our Markov chain proposes changes to each parameter. This algorithm results in a Markov chain whose stationary distribution is the joint posterior distribution of interest. For background on these methods in the present context see Pritchard et al. (2000), or a general text such as Gilks et al. (1996).

Pritchard et al. (2000), described updates for  $P$ ,  $Q$ ,  $Z$  and  $\alpha$ . Here, we describe updates for  $r$ ,  $F$  and  $\lambda$ , as well as modified updates for  $Q$  and  $Z$  under the linkage model, and for  $P$  under the  $F$  model.

**MCMC update for the  $F$  model.** The update for  $p_{kl}$  is similar to that for the original (independent frequencies) model (Eq. A6, Pritchard et al., 2000), but substitute  $p_{Alj}(F_k^{-1}-1)$  in place of  $\lambda_j$ .

The update for the ‘‘ancestral’’ allele frequencies  $P_A$  is more complicated. The algorithm that we have implemented is as follows. Start with initial guesses for  $P_A$  (e.g., based on the overall sample frequencies). Then, at each step of the Markov chain, propose the following update once for each locus. Select at random two alleles  $m$  and  $n$ , such that  $1 \leq m < n \leq J_l$ , where  $J_l$  is the total number of alleles observed at locus  $l$ . Simulate a value  $\delta$  from a normal with mean 0 and some small (fixed) standard deviation (we used 0.05), and propose changing the allele frequencies  $p_{Alm}$  and  $p_{Aln}$  to  $p'_{Alm} = p_{Alm} + \delta$  and  $p'_{Aln} = p_{Aln} - \delta$ , respectively. Reject the proposal if either of the proposed allele frequencies is outside (0,1). Otherwise, accept the updated allele frequencies according to the appropriate Metropolis-Hastings probability, that is, the minimum of 1 and

$$\frac{\Pr(P'_A)\Pr(P_1, \dots, P_K \mid P'_A, F)}{\Pr(P_A)\Pr(P_1, \dots, P_K \mid P_A, F)}, \quad (7)$$

where  $\Pr(P_A)$  is the prior probability of  $P_A$ , assumed to be symmetric Dirichlet with parameter  $\lambda$ . Expression 7 can be rewritten as

$$\left(\frac{p'_{Alm}p'_{Aln}}{p_{Alm}p_{Aln}}\right)^{\lambda-1} \prod_{k=1}^K \frac{\Gamma(f_k p_{Alm})\Gamma(f_k p_{Aln})}{\Gamma(f_k p'_{Alm})\Gamma(f_k p'_{Aln})} p_{klm}^{f_k \delta} p_{kln}^{-f_k \delta}, \quad (8)$$

where  $f_k \equiv (1 - F_k)/F_k$ .

Finally, in order to update each value of  $F_k$ , we start from an initial value  $F_k^{(0)}$  (the prior mean, say), and then update it as follows. Conditional on the current value,  $F_k$ , we

propose a new value  $F'_k$ , from a normal distribution with mean  $F_k$  and some fixed standard deviation (0.05, say). If the new value  $F'_k$  is outside the interval (0,1), we reject the proposal. Otherwise, we accept it with the Metropolis-Hastings probability, namely, the minimum of 1 and

$$\frac{\Pr(F'_k)\Pr(P_k | P_A, F'_k)}{\Pr(F_k)\Pr(P_k | P_A, F_k)}, \quad (9)$$

where  $\Pr(F_k)$  is the prior probability of  $F_k$ , assumed to be proportional to a gamma distribution with mean  $\mu$  and variance  $\sigma^2$  (see main text). Expression 9 can be rewritten as

$$\exp\left[\frac{\mu(F_k - F'_k)}{\sigma^2}\right] \left(\frac{F'_k}{F_k}\right)^{\frac{\mu^2}{\sigma^2}-1} \prod_{l=1}^L \left[ \frac{\Gamma(f') \prod_j \Gamma(f p_{Alj}) p_{klj}^{f' p_{Alj}}}{\Gamma(f) \prod_j \Gamma(f p_{Alj}) p_{klj}^{f p_{Alj}}} \right]. \quad (10)$$

where  $f'$  and  $f$  are  $(F'_k - 1)/F'_k$  and  $(F_k - 1)/F_k$  respectively. If we assume a single value  $F$  for all populations, the update is much the same, except that the posterior ratio is given by a product over both  $k$  and  $l$ .

**Full description of the linkage model, including unphased data.** The following describes an MCMC scheme for simulating from a Markov chain with stationary distribution  $\Pr(P, Z, r, Q | X, K)$ .

1. Sample from  $\Pr(Z | P, r, Q, X)$ .
2. Sample from  $\Pr(P | Z, r, Q, X) = \Pr(P | Z, X)$ .
3. Update  $r$  by Metropolis-Hastings update.
4. Update  $Q$  by Metropolis-Hastings update.

Step 2 is exactly the same as in Pritchard et al. (2000). Step 1 can be done separately for each individual, using the Forwards-Backwards algorithm (e.g., Rabiner, 1989), as follows. We start by assuming that individuals are haploid, or that phase is known. For each chromosome from one individual (and omitting the individual's superscript  $i$  for clarity), we let  $\beta_{lk} = \Pr(x_1, \dots, x_l, z_l = k | P, r, Q)$ . Then, recalling that  $p_{klj}$  is the frequency of allele  $j$  at locus  $l$  in population  $k$ , we have

$$\beta_{1k} = q_k p_{k1x_1} \quad (11)$$

for  $k = 1, \dots, K$ , and

$$\beta_{(l+1)k'} = \sum_{k=1}^K \beta_{lk} \Pr(z_{l+1} = k' | z_l = k) p_{k'(l+1)x_{l+1}}. \quad (12)$$

This allows us to compute  $\beta_{lk}$  for  $k = 1, \dots, K$ , and  $l = 1, \dots, L$  (the ‘‘forwards’’ part of the algorithm). This computation can be made linear in  $K$  rather than quadratic, following a simple rearrangement. After substituting Equation 3 for  $\Pr(z_{l+1} = k' | z_l = k)$ , we get

$$\beta_{(l+1)k'} = \left[ \sum_{k=1}^K \beta_{lk} \right] p_{k'(l+1)x_{l+1}} q_{k'} (1 - \exp(-d_l r)) + \beta_{lk'} p_{k'(l+1)x_{l+1}} \exp(-d_l r). \quad (13)$$

Note that the sum in square brackets is independent of  $k'$  and therefore only needs to be calculated once for each  $l$ .

Having done this, the “backwards” part allows us to simulate  $Z$ , starting from  $z_L$ , using Gibbs sampling based on:

$$\Pr(z_L = k \mid X, P, r, Q) \propto \beta_{Lk} \quad (14)$$

and

$$\begin{aligned} & \Pr(z_l = k \mid z_{l+1}, \dots, z_L, X, P, r, Q) \\ & \propto \Pr(z_l = k, x_1, \dots, x_l \mid P, r, Q) \Pr(z_{l+1}, \dots, z_L, x_{l+1}, \dots, x_L, P, r, Q) \\ & \propto \beta_{lk} \Pr(z_{l+1} \mid z_l = k, r, Q) \end{aligned} \quad (15)$$

where  $\Pr(z_{l+1} \mid z_l = k, r, Q)$  is given by (3).

For unphased or partially phased diploids, we have implemented two algorithms. The two algorithms can incorporate different types of phase information and produce equivalent results for unphased data. The first algorithm is based on a Markov formulation for phase information. For each pair of adjacent linked loci we have an estimate of the probability  $b_l$  that the first allele of loci  $l$  and  $l + 1$  are on the same chromosome. Information that adjacent allele copies were inherited together from the same (unspecified) parent is often available from sib-pair pedigree data. For unphased data the order of the allele copies is random so that,  $b_l = 0.5$  for all loci. We define

$$\beta_{lk^1k^2} = \Pr(x_1^1, x_1^2, \dots, x_l^1, x_l^2, z_l^1 = k^1, z_l^2 = k^2 \mid P, r, Q), \quad (16)$$

where superscript (1) refers to the first allele copy and (2) refers to the second allele copy at each locus. Denote the quantity in the right hand side of equation (3) by  $P_{k'k}$ . In the forwards part of the algorithm we calculate

$$\beta_{1k^1k^2} = q_{k^1} q_{k^2} p_{k^1 1 x_1^1} p_{k^2 1 x_1^2} \quad (17)$$

for  $k^1 = 1, \dots, K, k^2 = 1, \dots, K$ , and

$$\begin{aligned} \beta_{(l+1)k^1k^2} &= \sum_{k^1=1}^K \sum_{k^2=1}^K \beta_{lk^1k^2} p_{k^1 (l+1) x_{l+1}^1} p_{k^2 (l+1) x_{l+1}^2} \\ &\quad \times \{b_l P_{k^1 k^1} P_{k^2 k^2} + (1 - b_l) P_{k^1 k^2} P_{k^2 k^1}\}. \end{aligned} \quad (18)$$

In the backwards part we jointly simulate  $z^1$  and  $z^2$  using:

$$\Pr(z_L^1 = k^1, z_L^2 = k^2 \mid X, P, r, Q) \propto \beta_{Lk^1k^2} \quad (19)$$

and

$$\begin{aligned} & \Pr(z_l^1 = k^1, z_l^2 = k^2 \mid z_{l+1}^1, z_{l+1}^2, \dots, z_L^1, z_L^2, X, P, r, Q) \\ & \propto \beta_{lk^1k^2} \{b_l P_{z_{l+1}^1 k^1} P_{z_{l+1}^2 k^2} \\ & \quad + (1 - b_l) P_{z_{l+1}^1 k^2} P_{z_{l+1}^2 k^1}\} \end{aligned} \quad (20)$$

The model for maternal and paternal phase information is similar; here

$$\beta_{lk^m k^p} = \Pr(x_1^1, x_1^2, \dots, x_l^1, x_l^2, z_l^m = k^m, z_l^p = k^p \mid P, r, Q) \quad (21)$$

where superscripts ( $^m$ ) and ( $^p$ ) refer to the maternal and paternal allele copies at each locus.

$$\beta_{1k^m k^p} = q_{k^m} q_{k^p} \{M_1 p_{k^m 1x_1^1} p_{k^p 1x_1^2} + (1 - M_1) p_{k^m 1x_1^2} p_{k^p 1x_1^1}\} \quad (22)$$

and

$$\beta_{(l+1)k^{m'} k^{p'}} = \sum_{k^m=1}^K \sum_{k^p=1}^K \beta_{lk^m k^p} P_{k^{m'} k^m} P_{k^{p'} k^p} \quad (23)$$

$$\{M_l p_{k^{m'} (l+1)x_{l+1}^1} p_{k^{p'} (l+1)x_{l+1}^2} + (1 - M_l) p_{k^{m'} (l+1)x_{l+1}^2} p_{k^{p'} (l+1)x_{l+1}^1}\}.$$

where  $M_l$  is the probability that allele 1 at locus  $l$  is inherited maternally, assumed known. In the backwards step, we simulate the population of origin of maternal and paternal chromosomes.

$$\Pr(z_L^m = k^m, z_L^p = k^p \mid X, P, r, Q) \propto \beta_{Lk^m k^p} \quad (24)$$

and

$$\Pr(z_l^m = k^m, z_l^p = k^p \mid z_{l+1}^m, z_{l+1}^p, \dots, z_L^m, z_L^p, X, P, r, Q) \quad (25)$$

$$\propto \beta_{lk^m k^p} P_{z_{l+1}^m k^m} P_{z_{l+1}^p k^p}.$$

We then simulate the population of origin for the two allele copies at each locus, conditional on maternal and paternal assignments using:

$$\Pr(z_l^1 = k^m, z_l^2 = k^p \mid z_l^m = k^m, z_l^p = k^p) \propto M_l p_{k^m l x_l^1} p_{k^p l x_l^2}. \quad (26)$$

For both of these diploid models, a rearrangement analogous to that described for the haploid case leads to the computation in the forward part being proportional to  $K^2$  rather than  $K^4$ .

For each of these models, the updates for  $r$  and  $Q$  use the random walk Metropolis algorithm. In haploids, for example, (with the individual superscript  $i$  reinstated),

$$\Pr(x_1^{(i)} \dots x_L^{(i)} \mid P, r, q^{(i)}) = \sum_{k=1}^K \beta_{Lk}^{(i)}, \quad (27)$$

$r$  is updated using a Metropolis-Hastings step, by comparing this sum for proposed and current values of  $r$ . We have implemented a uniform prior for  $\log_{10}(r)$ ; the proposed value of  $\log_{10}(r)$  differs from the current value according to a normal distribution with mean of 0.05.

Lastly, we perform Step 4, the Metropolis-Hastings update for  $Q$ , as follows. For each individual  $i$ , we simulate a proposal value  $q^{(i)*}$  from the prior distribution, which is  $\mathcal{D}(\alpha_1, \alpha_2, \dots, \alpha_K)$ . The proposed value  $q^{(i)*}$  is accepted with probability that is equal to the ratio of the likelihoods.

**MCMC updates for  $\lambda$  and  $\alpha$ .** We begin by placing independent uniform priors in the range (0,10) on  $\lambda_k$  for each population  $k$ . We start the Markov chain at arbitrary initial values,  $\lambda_k^{(0)} = 1$ , for instance, and then update each  $\lambda_k$  as follows. Conditional on the current value of  $\lambda_k$ , we propose a new value  $\lambda'_k$  from a normal distribution with mean  $\lambda_k$  and some standard deviation (0.3, say). If  $\lambda'_k$  is outside (0,10), we reject the new proposal; otherwise we accept it according to the standard Metropolis-Hastings ratio. That is, we accept it with a probability that is the minimum of 1 and

$$\prod_{l=1}^L \frac{\Gamma(J_l \lambda'_k) \Gamma(\lambda_k)^{J_l}}{\Gamma(J_l \lambda_k) \Gamma(\lambda'_k)^{J_l}} (p_{kl1} p_{kl2} \cdots p_{klJ_l})^{\lambda'_k - \lambda_k}. \quad (28)$$

The latter expression is the ratio of the density function for the Dirichlet distribution given the proposed, and current values of  $\lambda_k$ , respectively, multiplied across all loci. If we assume that the same value of  $\lambda$  holds for all  $K$  populations, then the analogous update ratio is

$$\prod_{l=1}^L \prod_{k=1}^K \frac{\Gamma(J_l \lambda') \Gamma(\lambda)^{J_l}}{\Gamma(J_l \lambda) \Gamma(\lambda')^{J_l}} (p_{kl1} p_{kl2} \cdots p_{klJ_l})^{\lambda' - \lambda}. \quad (29)$$

The full update used for  $\alpha$  is closely analogous to that for  $\lambda$ , except that the products are over individuals, not loci, and over ancestry coefficients,  $q$ , not allele frequencies,  $p$ .

## Tables and Figures

K	Uncorrelated model	$F$ model
A		
1	<b>-81404</b>	<b>-81404</b>
2	*	-81819
3	*	-82657
4	*	-82470
B		
1	<b>-76855</b>	-76855
2	-77169	<b>-76780</b>
3	-77059	-77099
4	*	-78171

Table 1: Estimates for model log-likelihood ( $\log P(X | K)$ ) for 100 diploid individuals taken from a simulated random-mating population (Scenario I) with  $N_0 = 2, 500$ . In part B there is a 50% selfing rate. \* indicates that one of the populations was empty in all of the simulated runs. Details: 400 microsatellites were simulated according to the stepwise mutation model with  $\mu = 2 \times 10^{-4}$ . Each model log-likelihood is the average from two *structure* runs with 5,000 generations burn-in and 50,000 iterations. The prior for  $F$  (gamma distribution with mean 0.01 and standard deviation 0.05) puts most of its weight on low values of  $F$ . The highest likelihood run in each group is in bold type.

	Admixture Model				Linkage Model				
A	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$r$
1	0.999	0.999	-4306	0.026	0.995	0.515	-4304	0.0286	3.5
2	0.995	0.512	-5414	0.1042	0.966	0.523	-4694	0.1897	2.4
4	0.991	0.533	-5728	0.1944	0.940	0.522	-4676	0.3711	5.3
8	0.990	0.400	-5965	0.1785	0.932	0.505	-4872	0.4368	7.3
16	0.990	0.419	-5957	0.191	0.937	0.492	-4961	0.4353	14.2
32	0.993	0.221	-6153	0.1144	0.956	0.479	-5428	0.4021	36.4
64	0.990	0.169	-5970	0.0985	0.961	0.476	-5433	0.3886	62.1
128	0.995	0.044	-5935	0.038	0.962	0.450	-5781	0.3977	106.7
256	0.997	0.016	-5836	0.0303	0.995	0.026	-5856	0.0345	114.3
B	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$r$
1	1.000	0.470	-39362	0.0231	1.000	0.471	-39219	0.0247	39.3
2	0.999	0.501	-49602	0.0907	0.998	0.501	-42360	0.0918	2.5
4	0.999	0.531	-55770	0.1786	0.996	0.472	-43881	0.2032	3.9
8	0.999	0.508	-57284	0.16	0.996	0.510	-44938	0.2103	7.4
16	0.999	0.490	-57299	0.1895	0.997	0.507	-46542	0.2047	15.3
32	0.999	0.468	-56969	0.1893	0.997	0.505	-48755	0.1962	32.2
64	0.999	0.452	-56558	0.1879	0.998	0.487	-50787	0.1855	62.2
128	0.998	0.366	-56027	0.1918	0.998	0.443	-52857	0.1805	110.4
256	0.999	0.006	-54832	0.0321	0.998	0.327	-54373	0.1784	199.6
C	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$r$
1	0.998	0.412	-62175	0.03	0.999	0.411	-62141	0.025	0.00
2	0.989	0.493	-64033	0.1669	0.993	0.486	-63491	0.1152	2.46
4	0.987	0.641	-66179	0.2827	0.987	0.570	-65363	0.2661	4.97
8	0.990	0.683	-66145	0.2769	0.984	0.592	-65464	0.3061	8.82
16	0.991	0.667	-66213	0.2459	0.985	0.592	-65878	0.3016	18.3
32	0.989	0.598	-65657	0.2641	0.985	0.578	-65689	0.3097	40.81
64	0.982	0.453	-65429	0.3425	0.985	0.456	-65343	0.3041	58.37
128	0.995	0.024	-64370	0.054	0.997	0.014	-64393	0.0314	97.46
D	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$H_1$	$H_2$	$\log P(X   K)$	$\alpha$	$r$
1	0.993	0.526	-58517	0.0579	0.575	0.450	-53428	6.25	423
2	0.976	0.488	-60928	0.2256	0.573	0.450	-54267	8.65	439
4	0.979	0.639	-61922	0.3278	0.541	0.441	-54570	10.71	373
8	0.986	0.703	-63055	0.3022	0.556	0.455	-54130	10.79	377
16	0.987	0.634	-62297	0.2708	0.605	0.481	-54419	9.62	402
32	0.970	0.566	-62248	0.3924	0.602	0.470	-54324	8.63	426
64	0.980	0.551	-61368	0.348	0.604	0.459	-53901	9.1	437
128	0.969	0.261	-62590	0.3239	0.621	0.440	-53029	7.69	465

Table 2: Estimates for model log-likelihood, ancestry and other parameters for the *structure* runs shown in Figure 4. The first column gives the number of generations after the admixture event,  $T_2$ .  $H_1$  and  $H_2$  are the average estimated ancestry from ancestral population 1 for individuals in the two post-admixture populations (true values are 1.000 and 0.500).  $r$  is measured in generations. The highest log-likelihood in each row gives a heuristic guide to which model fits better.

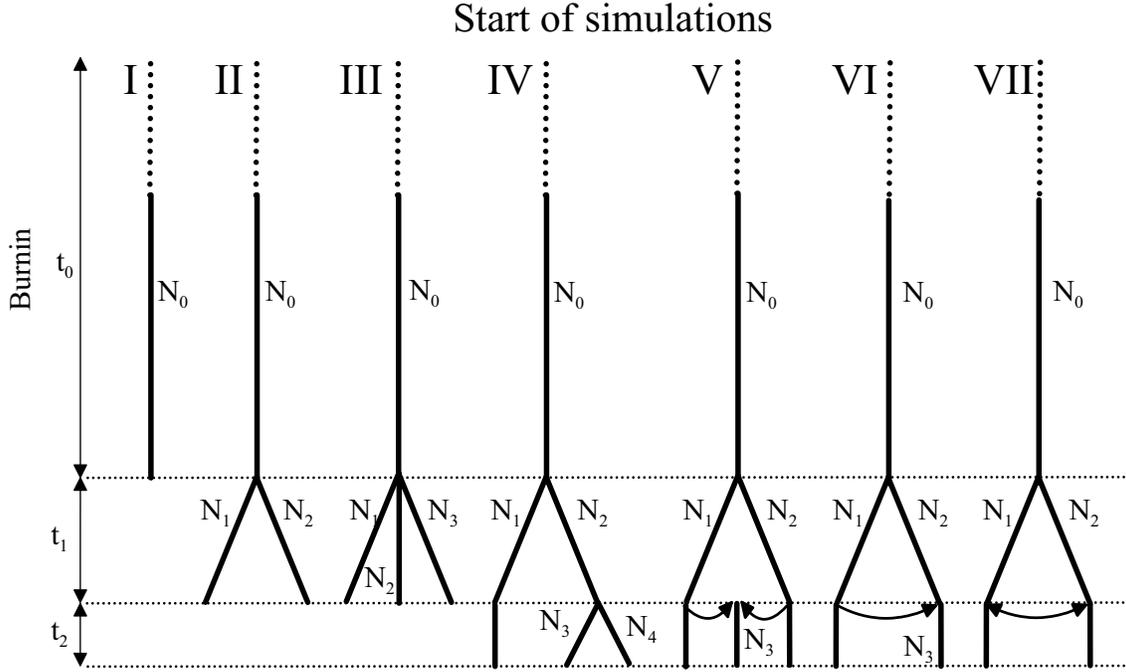


Figure 1: Scenarios simulated using the Wright Fisher model.  $N_0 \dots N_4$  indicate haploid population sizes, while  $t_0$ ,  $t_1$  and  $t_2$  indicate time in generations. In each case  $t_0 > 5N_0$ , so that the population is close to mutation-drift equilibrium at the end of the burn-in. (I) Sample from a single random-mating population. (II) Population bifurcation. (III) Population trifurcation. (IV) Hierarchical subdivision. (V) Admixture creates an additional population. (VI) Unidirectional admixture leaves one population unchanged. (VII) Bidirectional admixture between two populations. In V, VI and VII the diverged populations undergo an admixture event in which individuals are formed by the union of gametes from pre-admixture populations, before undergoing  $t_2$  generations of subsequent evolution. The ancestry of the post-admixture populations is determined by a matrix  $u$ , where  $u_{ij}$  is the proportion of gametes in population  $j$  that come from pre-admixture population  $i$ .

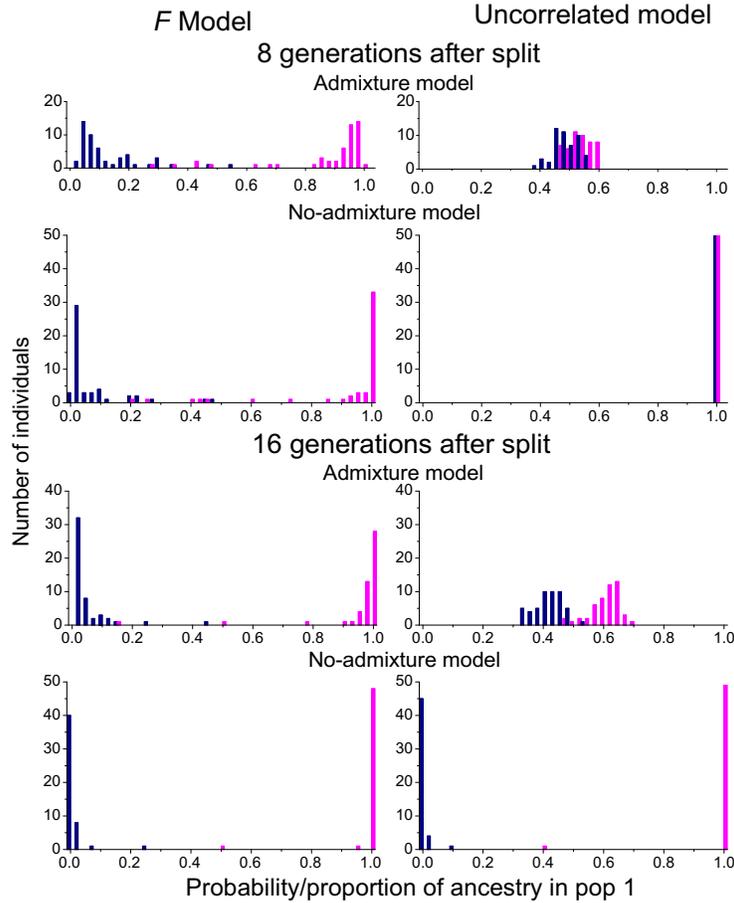


Figure 2: Performance of different models in distinguishing recently diverged populations (Scenario II). Structure was run with 50 diploid individuals sampled from each of two populations. Details: 400 unlinked microsatellite loci were simulated according to the stepwise mutation model with mutation rate  $\mu = 2 \times 10^{-4}$ ; haploid population sizes were  $N_0 = 2N_1 = 2N_2 = 2,500$ . Here, the prior for  $F$  was chosen to put most of its weight on low values (specifically, we used a gamma distribution with mean 0.01 and standard deviation 0.05).

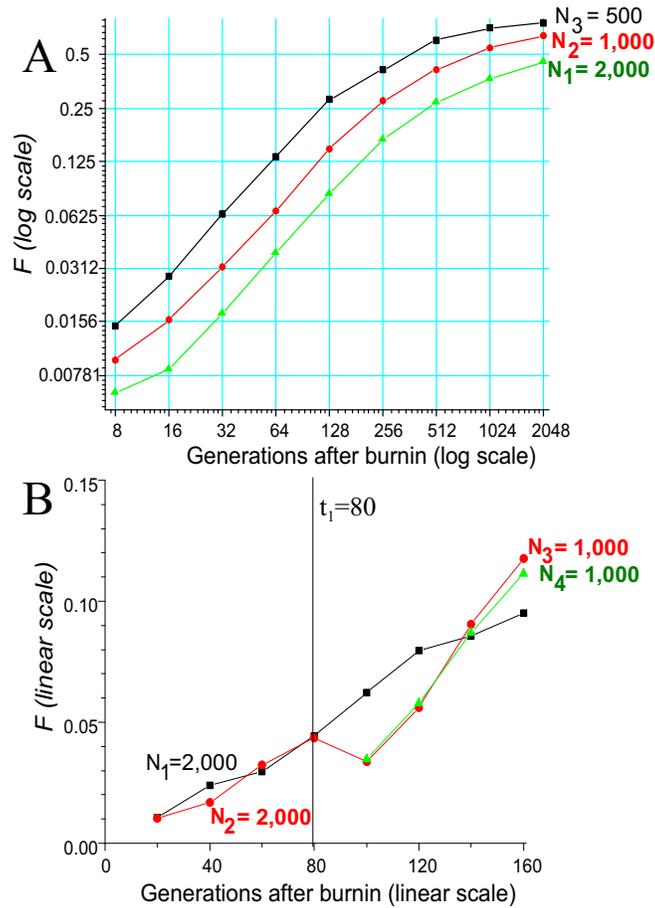


Figure 3: Estimates of  $F$  under two evolutionary scenarios. (A) Scenario III (population trifurcation). (B) Scenario IV (hierarchical subdivision). Structure was run with 50 diploid individuals from each of the populations. The no-admixture model was used throughout. Mutational parameters were as in Figure 2. Population sizes are shown on the graph. The vertical line in B indicates the time at which population 2 split into populations 3 and 4.

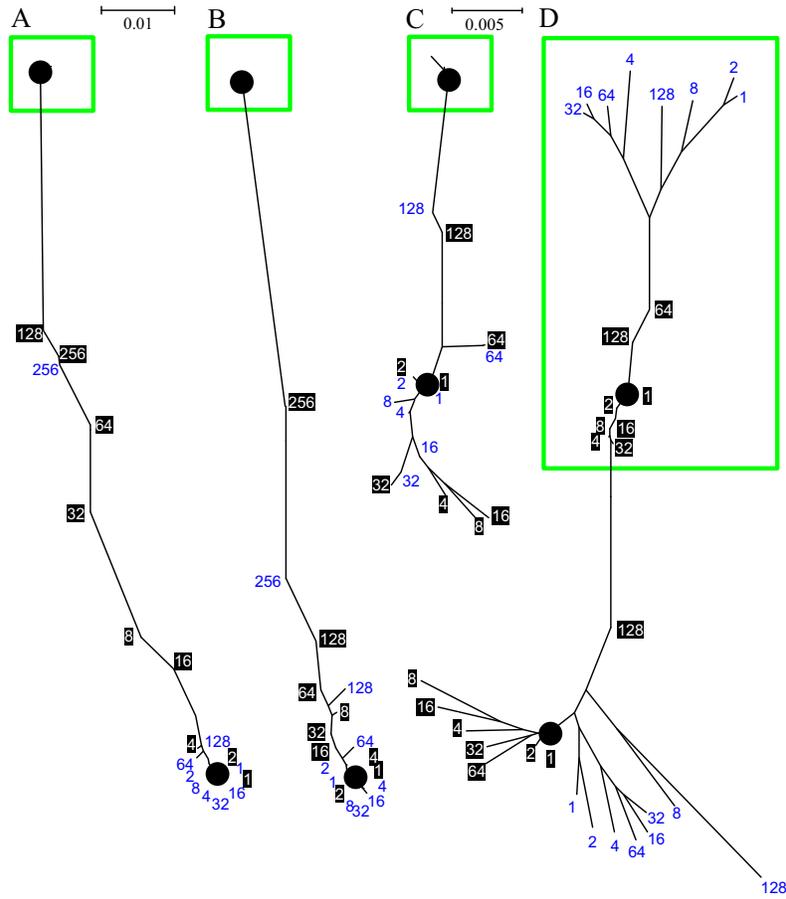


Figure 4: Population allele frequencies estimated using the linkage (blue) and admixture (white on black) models in the generations following an admixture event, as visualized using Neighbor Joining trees of genetic distances (Kumar et al., 2001). The simulations were performed according to Scenario VI, in which ancestral population 2 receives half of its gametes from ancestral population 1 (in boxes) in a single, one-way admixture event. The positions of the true ancestral populations in the trees are indicated by the large black dots. The labels give the number of generations after admixture ( $t_2$ ) at which the sample was taken. When the allele frequencies are estimated accurately, the corresponding labels lie close to the black dots. Labels are not shown for population 1 in parts A-C as they all fall close together on the upper black dot. Details: in each of the simulations,  $L^* = 50$ ,  $N_0 = 2N_1 = N_3 = 5,000$ . (A)  $C = 5$ ,  $t_1 = 5,000$ ,  $R = 0.001$ ,  $N_2 = 2,500$  (B) as A except  $C = 50$  (10-fold more chromosomes). (C)  $C = 50$ ,  $t_1 = 500$ ,  $R = 0.01$ ,  $N_2 = 100$  (resulting in strong drift in the recipient population). (D) as part C except  $R = 0.0001$  (tighter linkage). Biallelic markers were simulated with  $\mu = 10^{-5}$ . In each case, *structure* was run with  $K = 2$ , using 100 haploid individuals from both of the post-admixture populations. In the notation of Figure 1,  $u_{11} = 1, u_{12} = 0.5$ . The trees were constructed using the Nei and Li (1979) distance  $\delta_{P_i P_j}$  between allele frequency vectors  $P_i$  and  $P_j$  as  $\delta_{P_i P_j} = d_{P_i P_j} - (d_{P_i P_i} + d_{P_j P_j})/2$ , where  $d_{P_i P_j} = \frac{\sum_{l=1}^L (1 - \sum_j p_{klj} p_{k'l_j})}{L}$  and  $p_{klj}$  is the frequency of allele  $j$  at locus  $l$  in population  $k$ .

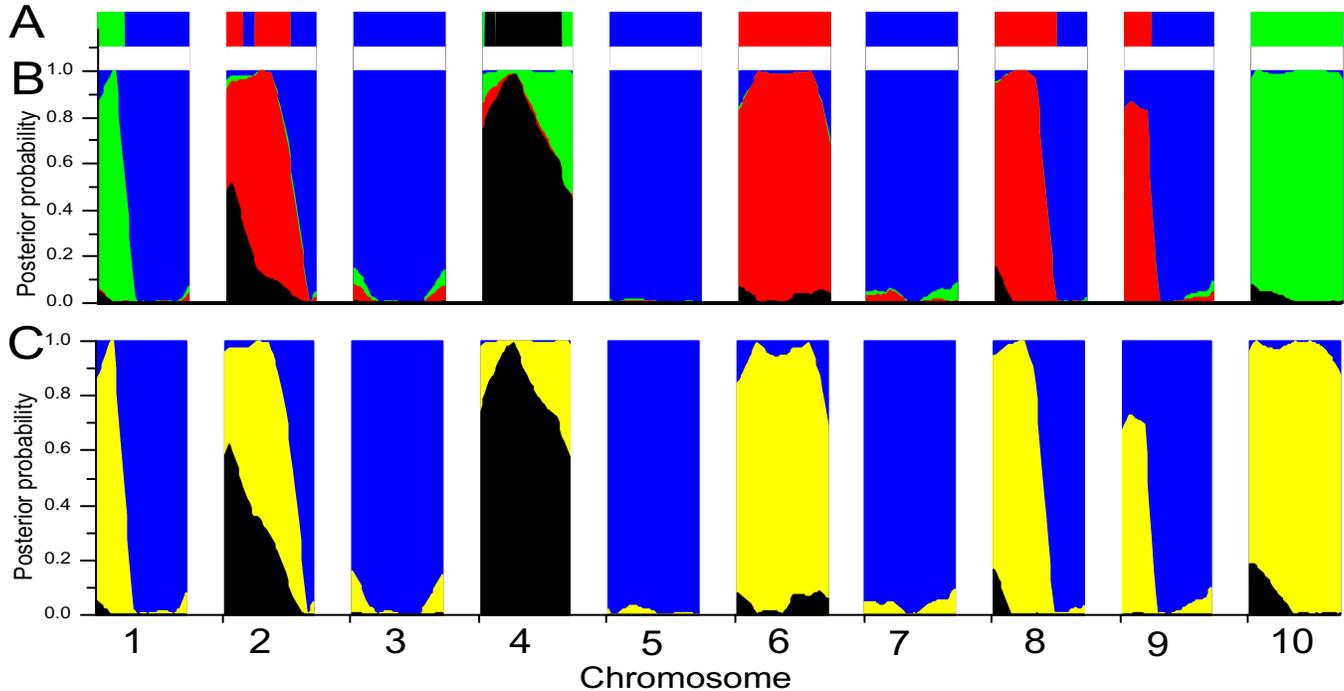


Figure 5: Population-of-origin assignments for maternal and paternal strands from 10 chromosomes of a single diploid individual. The loci are shown in map order, with 50 loci per chromosome. The parts of the figure show (A) true population of origin combinations for each chromosomal region; (B) posterior assignment probabilities (proportion occupied by each colour) calculated using the linkage model with phased data; (C) as in B but for unphased data. The colours indicate different assignment combinations; both allele copies inherited from population 1 (black) or population 2 (blue), maternal from population 1 and paternal from population 2 (red), maternal from population 1 and paternal from population 2 (green) and an unspecified copy from population 1 and the other from population 2 (yellow). In B the assignment probabilities for each strand can be inferred by summing over the different combinations for the other strand. So for example, the probability that the maternal copy is inherited from population 1 is given by the proportion that is either black or red. In C individual allele copies are not assigned separately, so if the true population of origin is either red or green, the correct assignment for that locus is yellow. *structure* was run with  $K = 2$  and 50 individuals from each population. Wright-Fisher simulations were performed according to Scenario VI with  $C = 10$ ,  $L^* = 50$ ,  $R = 0.001$ ,  $t_2 = 32$ ,  $u_{11} = 1$  and  $u_{21} = 0.3$ .  $N_0 = 2N_1 = 2N_2 = 5,000$  and mutational parameters as in Figure 4.

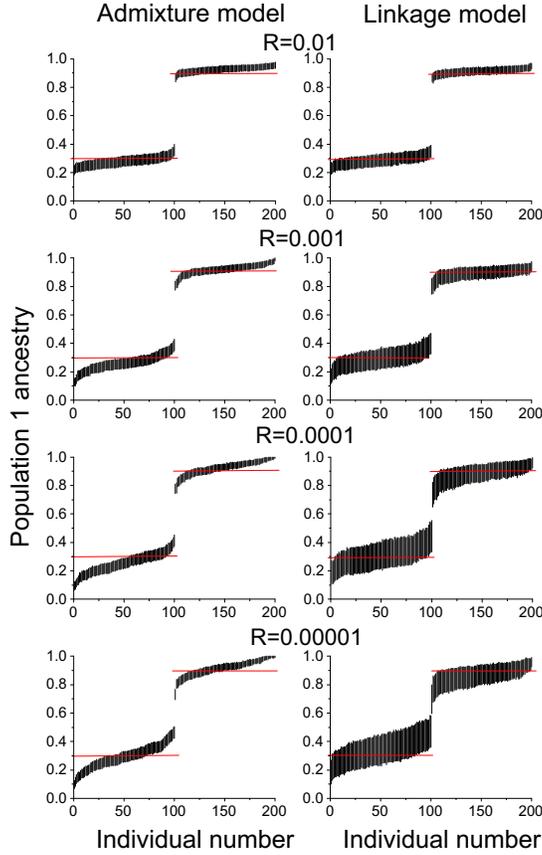


Figure 6: Coverage properties of ancestry estimates for individuals sampled from two post-admixture populations (Scenario VII). In each plot, the 90% credibility region for  $q$  for each individual is drawn as a thin vertical line. Within each population the true ancestry coefficients are essentially the same for all individuals, and are indicated by a red horizontal line. The individuals are ordered according to their point estimates for  $q$ . When the inference is performed correctly, roughly 90% of the individual credibility regions should cross the red line. Details: 100 haploid individuals from each population,  $C = 50$ ,  $L^* = 100$ . Wright-Fisher simulations were performed with  $R$  as shown,  $t_1 = 5000$ ,  $t_2 = 32$ ,  $u_{11} = 0.9$  and  $u_{21} = 0.3$ .  $N_0 = 2N_1 = 2N_2 = 5,000$  and mutational parameters as in Figure 4.

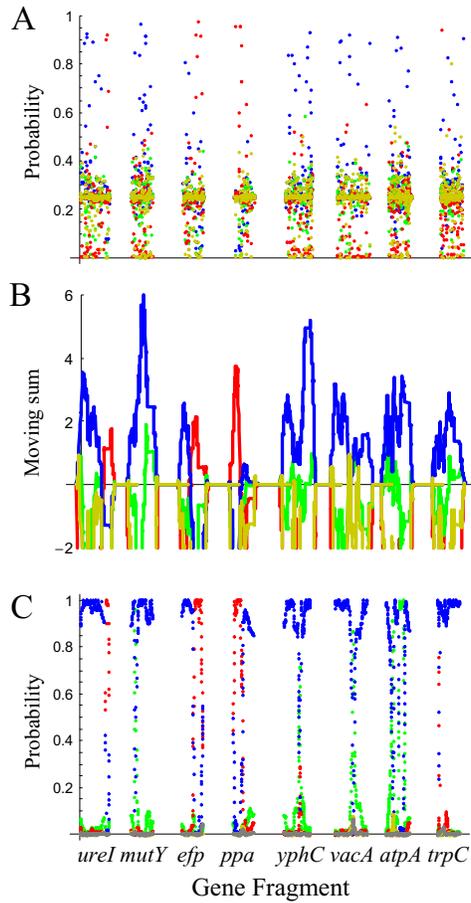


Figure 7: Hybrid ancestry of a *Helicobacter pylori* isolate, showing import of Africa2 fragments (red) into an isolate that is mostly of Africa1 origin (blue). Data for 8 gene fragments of between 398 and 623 nucleotides in length. (A) Posterior assignment probabilities of polymorphic nucleotides to populations, based on the nucleotide frequencies  $P$  estimated by the no-admixture model. (B) 100-nucleotide moving sum of 4 times the log of the assignment probabilities shown in A. (C) Assignment of nucleotides using the linkage model. The colours green and yellow indicate European and East Asian contributions respectively.

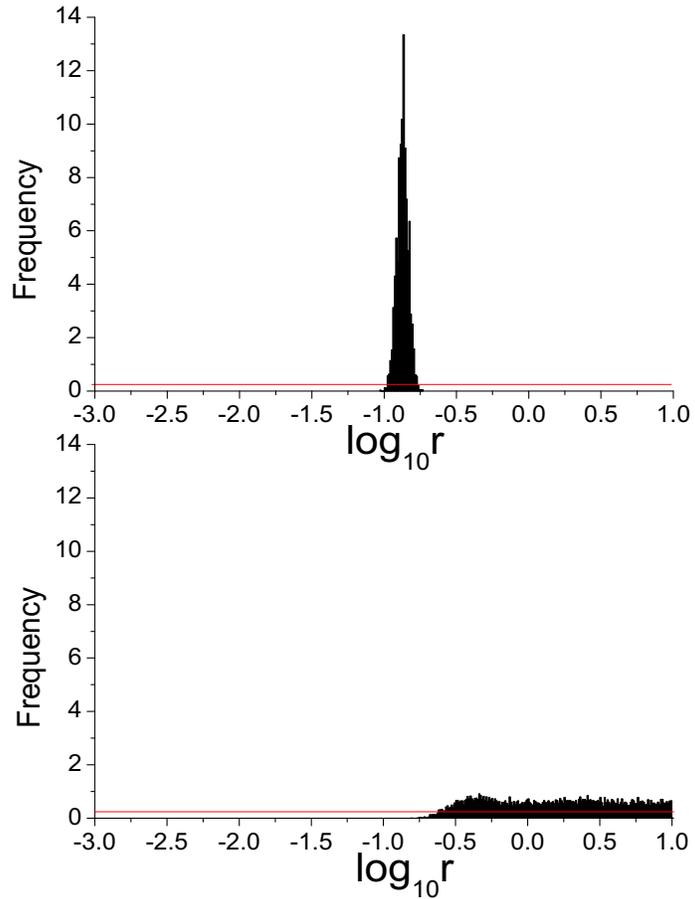


Figure 8: Posterior distribution of the chunk size parameter  $r$ , per centiMorgan, for the dataset of African Americans, European Americans and Nigerians. (A) Genetic distances estimated from the Marshfield map and (B) using Marshfield map distances but with the order of the loci randomized. The prior distribution of  $\log_{10} r$  is uniform on  $(-3,1)$  and is shown by the red line on each graph.

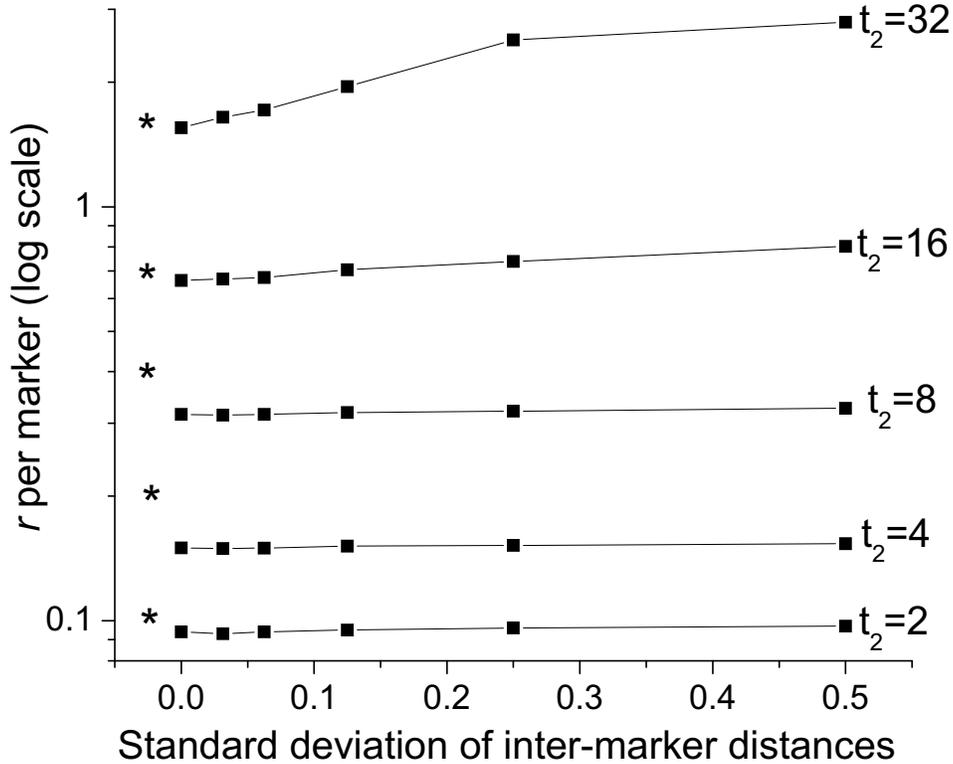


Figure 9: Effect of genetic map-distance errors on estimates of  $r$ . Simulations were performed according to Scenario V, in which an additional population is formed by mixture of two ancestral populations,  $t_2$  generations before sampling. The inter-marker distances used to simulate the data were all equal (5 cM), but *structure* was given distances with errors (mean of 1 unit; standard deviation as shown on x-axis). The stars show the correct estimates of  $r$  for each value of  $t_2$ . Map errors only had a substantial effect on estimates of  $r$  when the average chunk size was smaller than the inter-marker distances. Details:  $L^* = 100$ ,  $C = 25$ ,  $N_0 = 7,500$ ,  $N_1 = N_2 = N_3 = 2,500$ .  $t_1 = 50$ ,  $R = 0.05$ ,  $N_2 = 2,500$ . Markers were simulated according to the stepwise mutation model with  $\mu = 10^{-4}$ . In each case, *structure* was run with  $K = 2$ , using 50 haploid individuals from each of the three populations. The genetic map distances entered into the program were simulated as  $\text{Poisson}(n)/n$ , for a range of values of  $n$  (corresponding to different levels of accuracy). In the notation of Figure 1,  $u_{11} = 1$ ,  $u_{22} = 1$ ,  $u_{31} = 0.5$ ,  $u_{32} = 0.5$ .

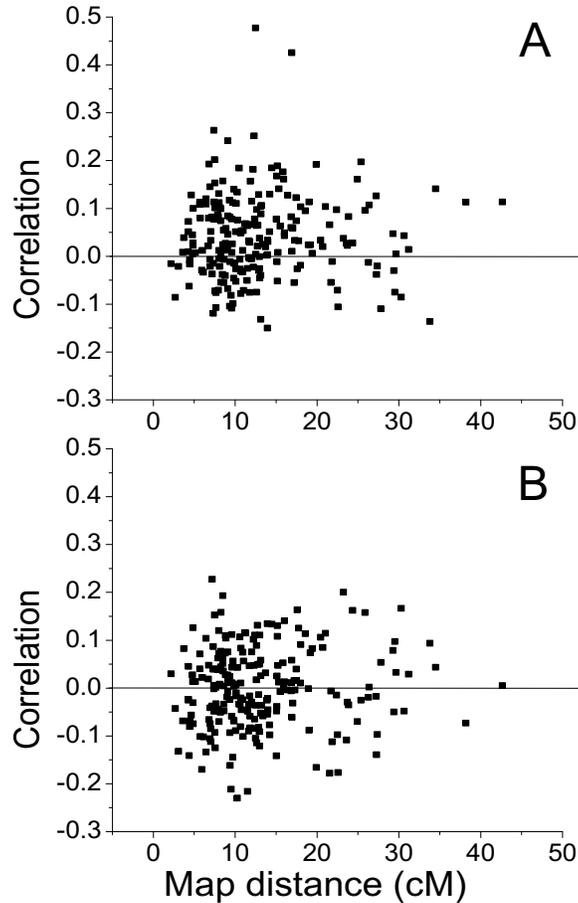


Figure 10: Most of the LD present in African Americans is mixture LD, not admixture LD. The figure shows two different measures of correlation between adjacent markers, plotted as a function of the genetic distance between the markers. (A) Correlation of ancestry probabilities for allele copies at pairs of adjacent markers. (B) Residual correlation of ancestry probabilities, corrected for variation in  $q$  across individuals. The correlations in the upper plot are the result of both mixture, and admixture LD, and are clearly positive on average, albeit noisy. The correlations in the lower plot result from admixture LD only, and the average is not significantly different from zero. However, the more powerful *structure*-based analysis does detect a strong signal of admixture LD. See text for further explanation.

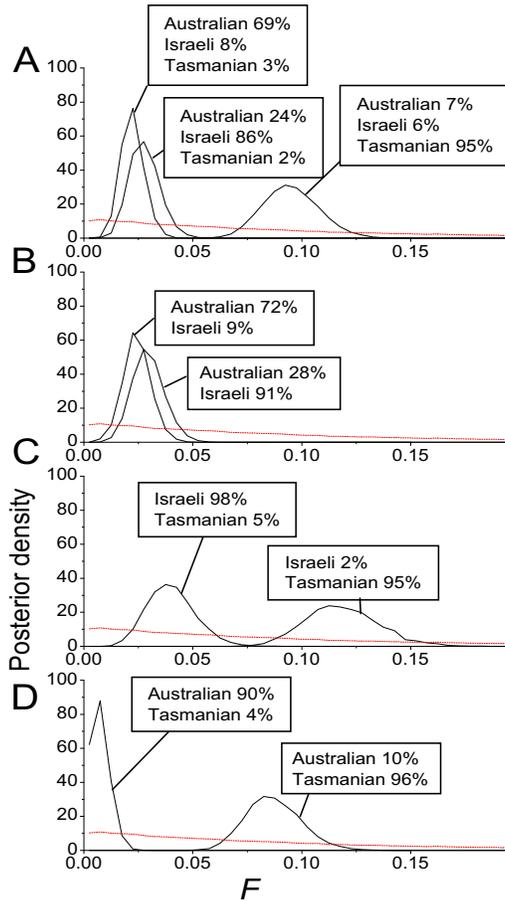


Figure 11: Posterior for  $F$  for *D. melanogaster* populations identified by *structure*. The prior (a gamma distribution with a mean and standard deviation of 0.1) is shown as a red dotted line. (A) Values of  $F$  estimated using whole dataset, with  $K = 3$ . (B-D) Values estimated with  $K = 2$ , excluding data from Tasmania, Australia and Israel respectively. The proportion of genotypes from the three land masses assigned to each *structure* population are shown in boxes.