Theoretical Population Biology 75 (2009) 331–345

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

An approximate likelihood for genetic data under a model with recombination and population splitting

D. Davison [a,*,1], J.K. Pritchard [b,c], G. Coop [b,2]

[a] Committee on Evolutionary Biology, University of Chicago, United States
[b] Department of Human Genetics, University of Chicago, United States
[c] Howard Hughes Medical Institute, University of Chicago, United States

ARTICLE INFO

Article history:
Received 26 January 2009
Available online 9 April 2009

Keywords:
Hidden Markov model
PAC likelihood
Haplotype data
Population history

ABSTRACT

We describe a new approximate likelihood for population genetic data under a model in which a single ancestral population has split into two daughter populations. The approximate likelihood is based on the 'Product of Approximate Conditionals' likelihood and 'copying model' of Li and Stephens [Li, N., Stephens, M., 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165 (4), 2213–2233]. The approach developed here may be used for efficient approximate likelihood-based analyses of unlinked data. However our copying model also considers the effects of recombination. Hence, a more important application is to loosely-linked haplotype data, for which efficient statistical models explicitly featuring non-equilibrium population structure have so far been unavailable. Thus, in addition to the information in allele frequency differences about the timing of the population split, the method can also extract information from the lengths of haplotypes shared between the populations. There are a number of challenges posed by extracting such information, which makes parameter estimation difficult. We discuss how the approach could be extended to identify haplotypes introduced by migrants.

1. Introduction

Population structure is a common feature of natural genetic and phenotypic variation (Mayr, 1942). For some applications, summarizing this structure by identifying subgroups and quantifying the extent of differentiation between them may be sufficient (e.g. Pritchard et al., 2000). However, the aim is often to make more explicit statements about the evolutionary history of the populations. While some structured populations may be modeled as a system of populations at equilibrium with respect to gene flow, researchers are often interested in non-equilibrium situations. In particular, at the interface of population genetics and phylogeny we are faced with the challenge of modeling population splitting. Accurate estimates of parameters such as the sizes of the populations, the times at which they separated, and how much subsequent interbreeding there has been would be very valuable. If mechanisms limiting interbreeding between the populations have arisen, then fitting these models may allow us to decide whether this has occurred in parapatry or allopatry (Hey, 2006; Becquet and Przeworski, 2009). Furthermore, the identification of functional loci (such as those involved in reproductive isolation or local adaptation) will be facilitated by knowledge of population history, effective population size and gene flow elsewhere in the genome (Hey and Nielsen, 2004; Becquet and Przeworski, 2009).

Although there is a well-developed theory of population genetic processes that generate data under these types of scenarios (see e.g. Wakeley, 2008), it is often very difficult to compute likelihoods for models of interest. Therefore, in this paper we describe a promising alternative approach that approximates the standard likelihood function. We start by specifying a particular model of population structure, and describing some of the approaches that have been developed for this type of problem.

We focus on the most basic model of population splitting (see for example Wakeley and Hey, 1997), in which an ancestral population splits instantaneously into two daughter populations (see Fig. 1). The parameters of the model are the three effective population sizes (N_a in the ancestral population; N_1 and N_2 in the two daughter populations), the number of generations since the splitting event (G), and the per-generation per-base pair probabilities of mutation and recombination (μ and r) respectively. Since our aim in this paper is to explore the utility of a new approximate likelihood, we focus on a simple version

* Corresponding author.
E-mail address: davison@stats.ox.ac.uk (D. Davison).
[1] Current address: Department of Statistics, University of Oxford, United Kingdom.
[2] Current address: Department of Evolution and Ecology, and Center for Population Biology, University of California, Davis, United States.

0040-5809/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2009.04.001

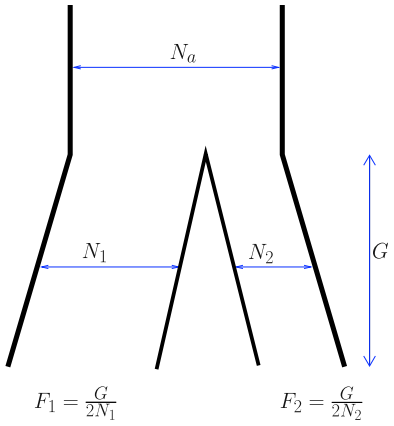

Fig. 1. Our model of population splitting without gene flow. Here, N_a, N_1 and N_2 indicate the effective population sizes in the ancestral population, and in the two daughter populations, respectively. G is the number of generations since the split of populations 1 and 2. The parameters F_1 and F_2 represent the amount of drift in the two daughter populations since the split.

of the model in which all three population sizes are equal ($N_a = N_1 = N_2 = N$). The model with unequal population sizes is a straightforward generalization, described briefly in Section 3.2. As is typically the case in population genetics, the data in fact contain information about the parameters (G, μ, r) on the time-scale on which genetic drift occurs, rather than on a time-scale of generations (see e.g. Wakeley, 2008). Thus our model in fact uses the relative rate parameters $\theta = 2N\mu$ and $\rho = 2Nr$, and the parameter $F = G/2N$ which represents the amount of drift that has occurred in the daughter populations since the split. Note that, as in many coalescent-based models of population genetics, our model of N diploids is equivalent to a model of $2N$ haploids (Wakeley, 2008). Also note that θ and ρ are often defined to be twice the values that we use in this paper.

1.1. Types of data

Historically, non-recombining loci such as mitochondrial DNA have often been used to fit these models, but there is a growing awareness of the statistical and biological limitations of such data sets (e.g. Hey and Machado, 2003). Therefore much research now focuses on using data from multiple unlinked regions of the genome, which reflect multiple realizations of the genealogical process. Often researchers will type a set of completely unlinked markers (e.g. microsatellites or SNPs) scattered around the genome, in which case the data can be summarized without loss of information by the counts of the different alleles in each population at each locus. The expectations of various quantities can be derived under a model without migration (Wakeley and Hey, 1997), and the likelihood of a particular configuration of allele counts at a locus may be computed analytically for a model with no migration and where mutation since divergence can be ignored (e.g. Nielsen and Slatkin, 2000; Nicholson et al., 2002; Roychoudhury et al., 2008), or alternatively can be estimated accurately under more general models using coalescent simulations.

While useful, completely unlinked loci offer only incomplete information about the genealogical process. One consequence is that there is relatively low power to distinguish between isolation models with and without migration (Nielsen and Slatkin, 2000), although multiallelic markers, such as microsatellites, hold additional information. With linked markers, the data are potentially informative about both migration and splitting times, as one can hope to learn about the variability (i.e. the distribution), over loci, of the pairwise coalescence times between the two populations (Wakeley, 1996). The distribution is informative, because under a model with no gene flow, the coalescence times between different populations necessarily predate the time at which the population split; in contrast, if there has been some low rate of gene flow, the coalescence times between populations are more variable, as some lineages migrate and thus coalesce more recently (Wakeley, 1996). However, unlinked sites provide little more information than the expectation of these times. For loosely linked data, information about the timing of the population split and subsequent migration is captured by the lengths and similarities of haplotypes shared between populations, as ancestrally shared or migrant haplotypes are broken up by recombination and diverge by mutation over time (e.g. Pool and Nielsen, 2008). For example, if two individuals in different populations are found to be identical across a large chromosomal region, then this may be strong evidence for recent gene flow, since such data may be unlikely under a pure split model.

1.2. Methods for linked data

Therefore, attention has turned to developing methods that consider a collection of genomic regions with linkage within each region, and free recombination between regions. Such data contain information about the joint distribution of times in the genealogies underlying the data, and thus potentially contain much more information about the parameters of interest. However, statistical inference in this setting is difficult: the likelihoods cannot be computed analytically and are difficult to estimate by simulation since the observed data will be very improbable, or impossible, on the vast majority of genealogies simulated from the coalescent prior (see e.g. Stephens, 2001). This problem has given rise to a large literature on full-, summary- and approximate-likelihood methods for linked data, a very brief overview of which now follows.

Nielsen and Wakeley (2001) and Hey and Nielsen (2004) developed a full likelihood inference scheme for the isolation and migration model, implemented by the IM software which can handle a set of independent fully linked loci. However, these approaches are limited, as extending them to allow intralocus recombination is challenging even under simple demographic models (Fearnhead and Donnelly, 2001; Nielsen, 2000). This requirement of full linkage is a potentially serious drawback. Firstly, low-recombining chromosomal regions may be atypical (Hilton et al., 1994); secondly, authors frequently trim the regions used in order to fit the no-recombination requirement, which may result in bias (Hey and Nielsen, 2004); and finally, this requirement limits the size of contiguous region that can be analyzed and hence the information available.

Summary likelihood methods are based on replacing the data with low-dimensional summary statistics, which allow likelihoods and posterior densities to be estimated, typically by simulation (Pritchard et al., 1999; Beaumont et al., 2002; Cornuet et al., 2008). This approach has been extended to models of population splitting both with gene flow (Becquet and Przeworski, 2007) and without (Leman et al., 2005; Putnam et al., 2007). Intralocus recombination can be incorporated straightforwardly, simply by allowing recombination in the simulated genealogies (e.g. Becquet and Przeworski, 2007). However, the flexibility and relative ease of computation come at the expense of losing information, and none of the existing approaches use summaries of the data that capture detailed information about haplotype structure.

A promising recent development in population genetics inference is the use of approximate likelihood approaches. One such approach was developed by Li and Stephens (2003, henceforth, "LS") for inferring recombination rates. They developed a new model for population genetic data that is simpler–and more computationally

tractable–than standard models. An attractive feature of the LS formulation is that it contains a formal model of haplotype structure, described in detail below.

The LS approach has been applied to inference in a diverse set of problems including estimating the parameters of recombination (Li and Stephens, 2003), mutation (Cornuet and Beaumont, 2007; Roychoudhury and Stephens, 2007), gene conversion (Gay et al., 2007; Yin et al., 2009) and diversifying selection (Wilson and Mcvean, 2006). It has also been used as a tool for modeling haplotype structure – rather than for formal parameter estimation – in genotype imputation and association mapping (Marchini et al., 2007), HLA typing (Leslie et al., 2008) and in models of population admixture (Price et al., 2009) and clustering (Hellenthal et al., 2008).

In this study, we extend the LS approach to estimate the parameters of a model of population splitting. While LS approaches are computationally convenient, they rely on *ad hoc* simplifications of standard population genetic models, such as the coalescent. As such, a key challenge in extending the approach is to develop approximations that are suitable for the new problem. Our approach to this will be a central focus of the paper.

1.2.1. The original Li & Stephens model

There are two main innovations in the LS approach. The first is that it computes a likelihood for the full data by breaking it down into a 'product of approximate conditional' (PAC) probabilities. Consider an observed set of haplotypes, $H = h_1, \ldots, h_n$. The basic inference problem is to compute (and maximize) the likelihood of the haplotype data H, with respect to the parameters ϕ of a population genetic model. The likelihood can be written as a product of conditional probabilities:

$$L(H;\phi) = p_\phi(h_1)p_\phi(h_2|h_1)\ldots p_\phi(h_n|h_1,\ldots,h_{n-1}). \quad (1)$$

However, these conditional probabilities are unknown for most models of interest, and the framework proposed by LS is based instead on the PAC likelihood

$$L_{\text{pac}}(H;\phi) = \hat{p}_\phi(h_2|h_1)\ldots\hat{p}_\phi(h_n|h_1,\ldots,h_{n-1}). \quad (2)$$

Here, terms of the form $\hat{p}_\phi(h_{k+1}|h_1,\ldots,h_k)$ denote the (approximate) conditional probability of the $(k+1)$th haplotype, conditional on the first k haplotypes, as a function of ϕ. We will refer to these as 'AC probabilities'. Note that under neutrality the unconditional likelihood of the first haplotype does not usually depend strongly on the model of population history. Therefore, following LS, we set $\hat{p}_\phi(h_1) = 1$, and omit this term from our notation.

The second innovation of LS was to introduce a simple model for population genetic data under which these AC probabilities may be computed. We will refer to this as their "copying model". The model is an approximation which captures many aspects of the coalescent with recombination, without suffering from the computational difficulties that have hindered attempts to incorporate recombination into coalescent-based MCMC and importance sampling schemes. At a basic level, LS can be thought of as providing a simple model for simulating haplotype data. We will first briefly describe LS in terms of the simulation problem, and then turn to how the LS model can be used for inference.

Consider first the simple case of data at a single SNP. Given k allele copies simulated so far, a new one is simulated by choosing one of the k uniformly at random: the new allele is said to 'copy' the chosen allele and, unless a mutation occurs, it is assigned the same allelic state as the copied allele. (The genotype of the first allele is set arbitrarily; e.g., to carry allele 0.) LS set the probability of mutation to be a decreasing function of k, to reflect the expectation that alleles added later in the order tend to match those already sampled (see Li and Stephens, 2003, for details).

Now consider simulating haplotypes at a set of loosely linked sites. Let X_l be the label of the haplotype copied at site l by the new haplotype. LS extended their model to include recombination by introducing correlation between the Xs at nearby sites: unless a 'switch' occurs, X_{l+1} is the same as X_l. If there is a switch, then the haplotype that is copied at the next site is a draw from the uniform prior on the k previously-sampled haplotypes (including the one that was being copied at site l). Thus the new haplotype is modeled as a mosaic formed of stretches copied from the haplotypes already simulated. Specifically, the probability distribution of the random variables denoting who is copied at each site $(X_1, X_2, \ldots, X_L)$ is a Markov chain along the sites. The switch events were intended to mimic the effects of recombination, and the transition probabilities of this Markov chain are controlled by a recombination parameter ρ. As for mutation, the switch probability is also a decreasing function of k. This reflects the fact that a haplotype added into a large sample tends to be similar to at least one other haplotype over a large genetic distance.

Computing the likelihood. There is a large set of possible values of the sequence $X_1, \ldots, X_L$, i.e. which haplotype is copied by the new haplotype at every site along the sequence, which we will refer to as 'paths' through the missing data. To compute the AC probability of the new haplotype $\hat{p}_\phi(h_{k+1}|h_1,\ldots,h_k)$ under this copying model, the paths are treated as missing data. Thus the AC probability is computed by averaging the data probability over the prior probability distribution on the paths. The Markov chain prior means that this can be done efficiently using standard hidden Markov model (HMM) methods (e.g. Rabiner, 1989). It is then straightforward to calculate the PAC likelihood from these AC probabilities according to Eq. (2).

One drawback of the scheme is that the use of approximate probabilities in Eq. (2) means that the likelihood depends on the order in which the haplotypes are added to the sample. LS found it satisfactory for inference to sum over a small set of random orders, keeping the set of orders the same over the parameter values the likelihood was estimated for.

1.3. Coalescing and copying

Full likelihood approaches for linked data, such as that implemented in the IM software (Hey and Nielsen, 2004, 2007), typically involve explicitly modeling aspects of the ancestral history of the sample such as the genealogical topology, the branch lengths, details of movements of ancestral lineages, or the types of ancestral haplotypes (see e.g. Stephens, 2001). The PAC likelihood and copying model of LS may be seen as an *ad hoc* approximation of such a full likelihood scheme. From this point of view, when modeling the new haplotype, its descent from the ancestral lineages of the existing sample is mimicked by forming it as a mosaic of chunks copied from present-day haplotypes with occasional 'mutations', and the uncertainty regarding its relationships with ancestral lineages of the existing sample is dealt with by averaging over all possible copying paths.

Throughout this paper we make this link between the copying process and the coalescent fairly explicit. That is, we consider copying haplotype X to be analogous to coalescing with the ancestral lineage of haplotype X at some time in the past. With this analogy in mind, we construct our copying model by deriving approximate probabilities and expectations under a coalescent model. This interpretation also lies behind the work of LS, who modeled the probabilities of mutation and switching (i.e. recombining) as decreasing functions of k, reflecting the fact that haplotypes added into an already large sample coalesce rapidly with other haplotypes, leaving little time for mutation or recombination (see Ewens (1990) for a discussion of this sequential sampling construction of the coalescent process).

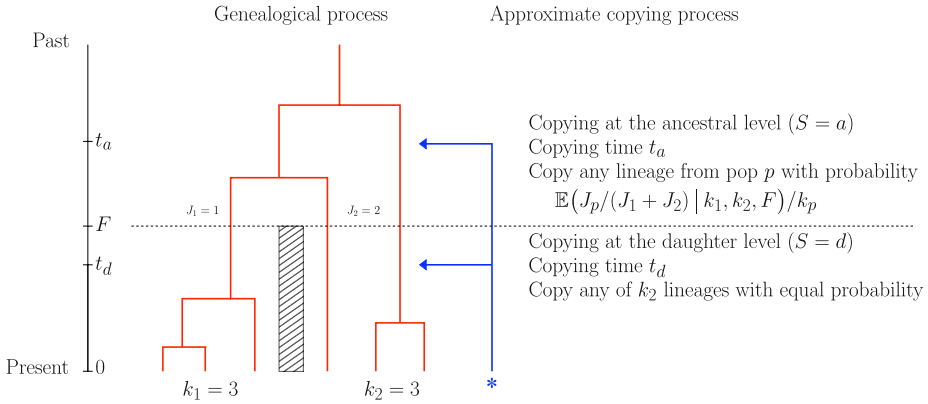

Fig. 2. A schematic depiction of the copying process of our model at a single site. The figure depicts the situation when computing the approximate conditional probability of the seventh haplotype, having already added three haplotypes from each population into the sample ($k_1 = k_2 = 3$). The left side (red) illustrates a possible genealogy of the previously sampled haplotypes (although note that we do not model the genealogy explicitly). The right side (blue) illustrates our approximate copying model for a new haplotype sampled in population 2. With probability $p(S = d)$ the lineage coalesces within the daughter population. In our approximate copying model this occurs at a fixed time $t_d = \mathbb{E}(T_{\text{coal}}|S = d)$, and the new haplotype copies any of the existing k_2 haplotypes with equal probability. Otherwise, the new lineage survives back into the ancestral population (state $S = a$). In that case, it coalesces with a lineage from either population, at fixed time $t_a = \mathbb{E}(t|S = a)$. The copying probabilities are weighted to reflect the different fixation rates in the two populations; the weighting factor involves the expected proportion $\mathbb{E}(J_p/(J_1+J_2))$, where J_p is the unknown number of ancestral lineages entering the ancestral population from population p. This expected proportion will differ from $\frac{1}{2}$ if $k_1 \neq k_2$ (as well as in the asymmetric drift case, $F_1 \neq F_2$). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

However, it should be noted that there is no formal correspondence between the coalescent process and the LS copying process. For example, it will often be the case that the ancestral lineage of the new haplotype coalesces with a lineage that is ancestral to several sampled lineages (haplotypes) at site l. In that case it is less clear which haplotype the new haplotype should be said to be copying at site l. This lack of exact correspondence between the genealogical process and the process under which the approximate likelihood is computed is a feature of the 'copying' approximation in general; it is not specific to the model of population history considered here.

2. A new PAC scheme for the population splitting model

2.1. Overview of the new method

In this paper we develop a PAC copying model for data sampled from two populations that have split from a common ancestral population, using the same framework of 'copying' and 'switching' outlined above. We first give a brief intuitive overview of the method. We then fully specify our model for unlinked sites and present results for unlinked data, and subsequently present our model of linkage and results for linked data. A schematic representation of our model is given in Fig. 2. In order to model the effect of the splitting event, we specify that copying occurs at one of two 'levels' S: ancestral ($S = a$) or daughter ($S = d$). We think of copying at the daughter level as corresponding to the new lineage coalescing at some point prior (measuring time backwards from the present) to the fusion of the populations; copying at the ancestral level corresponds to the new lineage surviving back to the ancestral population before coalescing with another lineage. Note that since there is no gene flow between the daughter populations only haplotypes from the same population may be copied at the daughter level, whereas haplotypes from either population might be copied ancestrally. Recall that in order to simulate a new allele at site l under the unstructured copying model of LS, one generates the latent variable X_l and then copies that haplotype (possibly with mutation). In contrast, under our structured copying model, one first chooses the level of copying S_l, and then conditional on S_l chooses the label X_l of the copied haplotype.

A complete description of our model requires specifying the following three quantities. The values that we use are based on approximating the standard coalescent model of population splitting.

A prior on the hidden copying states (S_l, X_l) **at a single site**. This is based on modeling the probability that, looking backwards in time, the newly sampled lineage survives to the ancestral population without coalescing with any of the other lineages sampled from that population. When F is large, the prior probability of $S_l = a$ is small, which corresponds to an expectation under the prior that allele frequencies in the two populations will have diverged substantially from the ancestral frequency. Conditional on copying at the daughter level ($S_l = d$), the new lineage can copy any of the lineages sampled from the same population, with equal probability. Conditional on $S_l = a$, the new lineage is allowed to copy any of the allele copies sampled from either population; if there are expected to be more ancestral lineages deriving from one population than the other, then the ancestral copying probabilities are weighted accordingly. The latter situation occurs when the number of previously-sampled lineages differs between the populations, and also when drift is asymmetric. The prior probabilities are described in detail in Section 3.1.1, and the extension to asymmetric drift is described briefly in Section 3.2.

The probability of the new allelic state, conditional on the state of the copied allele and the level at which copying occurred. To model the fact that copying an allele at the ancestral level implies a deeper coalescence time, we make the copying fidelity lower for $S_l = a$, by assuming that the time available for mutation is equal to the expected coalescence time, given S_l. This is described in detail in Section 3.1.2.

These two quantities, together with the PAC likelihood (Eq. (2)), specify our model for unlinked sites. In Section 3.2 we study estimators of the scaled split time under this model. Our model for linked variation requires the following third quantity to be specified.

Transition probabilities between the hidden copying states at adjacent sites. There is a large set of possible sequences of hidden states $(S_1, X_1), \ldots, (S_L, X_L)$, which we refer to as 'paths' through the missing data. The transition probabilities between states at consecutive sites specify a Markov chain prior on those paths, which is intended to capture the correlation between sites due to linkage. Our transition probabilities (described in Section 4.1) are parameterized in a way that attempts to capture important features of the genealogical process with recombination under the

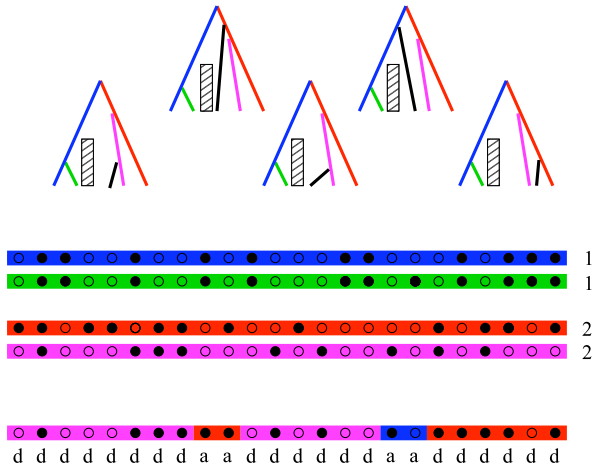

Fig. 3. The copying process in the new PAC model for loosely linked data illustrated with an example path through the missing data. A new haplotype is added to a sample of four ($k_1 = 2$, $k_2 = 2$; population labels are given on the right hand side). At each site along the haplotypes, small circles represent which of the two alleles is present (filled or open). Each of the 4 haplotypes has its own color. The new haplotype at the bottom is made up as a mosaic of these colors, indicating which of the four haplotypes is copied at each site ($X_1, X_2, \ldots, X_L$), and letters (d and a) indicate which level this copying occurs at ($S_1, S_2, \ldots, S_L$). For each of the 5 copied sections, a schematic genealogy is drawn above that might correspond to the state of the copying process below. In the trees, the new lineage is depicted in black. Although the relationships of the colored lines in the genealogies are depicted as remaining the same, note that this is not an assumption of the model.

population splitting model. Thus the new haplotype is modeled as a mosaic of sections of haplotype copied from one of the previous haplotypes, where points at which there is a switch in the haplotype being copied correspond to recombination events.

When F is large we have already seen that the prior at a single site places relatively little weight on ancestral copying. Our transition probabilities have the effect that these occasional stretches of ancestral copying are short, since the deep split time gives plenty of time for recombination. Therefore, the haplotype is expected to resemble others from the same population over long stretches. Conversely, when there has been less drift since the split, higher prior probability is associated with copying longer sections of ancestral haplotypes. We capture these properties of the genealogical process with recombination by using expected coalescence times in the daughter and ancestral populations to model the opportunity for recombination (switching) in our approximate copying process.

Fig. 3 illustrates a possible sequence of hidden states. The values of X along the haplotype are indicated by the haplotype coloring, and the values of S are indicated by the sequence of letters underneath. The path that is illustrated is one that might have high prior probability when F is relatively large because, firstly, the sections that are copied at the daughter level are much longer than those copied ancestrally, and secondly, the ancestral-level copying has made more errors (mutations) than the daughter-level copying.

Computing the likelihood. To compute the AC probability we have to consider the set of all possible paths $(S_1, X_1), \ldots, (S_L, X_L)$. The prior probability of one of these paths will depend on the value of the parameters F and ρ, as well as on the numbers k_1 and k_2 of haplotypes added so far from each population. The AC probability, $\hat{p}_\phi(h_{k_1+k_2+1}|h_1, \ldots, h_{k_1+k_2})$, corresponds to the probability of the data averaged over this prior probability distribution on possible paths and, as in LS, can be computed in an efficient manner using the forward algorithm for hidden Markov models (HMMs; see e.g. Rabiner (1989) and Appendix A).

As in LS, the approximations made in our PAC model make the likelihood depend upon the ordering of the haplotypes, and we average the likelihood of a dataset over a small set of random orders of haplotypes. For the results presented here we constrain these random orders to sample a haplotype from each of the two populations in turn.

3. Unlinked data

3.1. Methods: Unlinked

In the unlinked case the likelihood for the alignment of haplotypes may be computed as the product of the likelihoods for individual sites. Therefore, in this section we will use h_i to refer to an allele copy at a single site rather than an entire haplotype. The problem of computing the likelihood is now reduced to computing the approximate probability of a new allele $h_{k_1+k_2+1}$ given the alleles $h_1, \ldots, h_{k_1+k_2}$ observed so far, as a function of the model parameters. To do so we now specify our prior probability distribution on the hidden copying states $S_l \in \{a, d\}$ and $X_l \in \{1, \ldots, k_1 + k_2\}$, as well as the mutation probability conditional on the level at which copying occurs.

3.1.1. The prior on S_l and X_l

In contrast to LS, under our population splitting model the prior probability on X_l is uniform only in the unstructured case $F = 0$. To model the structure we introduce an additional state S_l, the prior probability distribution of which is given by the probability $p(S_l = a)$ that the new lineage coalesces in the ancestral population, given the amount of drift in the population from which it was sampled, and the number of lineages already sampled from that population. The probability is a decreasing function of both the latter two quantities, capturing the idea that few lineages make it back to the ancestral population in populations that have experienced considerable drift, and that lineages added later in the order are *a priori* more likely to coalesce in the daughter population and so resemble alleles already sampled within that population. The probability can be calculated exactly for a coalescent model (Eq. (31), Appendix C.1).

Conditional on S_l, we use the following prior probability distribution on X_l:

$$p(X_l = i|S_l = d) = \begin{cases} \dfrac{1}{k_{z_*}} & \text{if } z_i = z_* \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$p(X_l = i|S_l = a) = \mathbb{E}\left(\frac{J_{z_i}}{J_1 + J_2}\right)\frac{1}{k_{z_i}}. \tag{4}$$

Here, $z_i \in \{1, 2\}$ is the population label of sampled lineage i, and k_{z_i} refers to the number of lineages sampled so far from population z_i. z_* is shorthand notation for $z_{k_1+k_2+1}$, i.e. the label of the population from which the new lineage is sampled. J_{z_i} is the unknown number of distinct ancestral lines that enter the ancestral population, starting with k_{z_i} lines in daughter population z_i. The expectation can be calculated using the transition probabilities in Tavaré (1984) (Appendix C Eq. (32)). Note that although quantities such as $p(S_l = a)$ and $p(X_l = i|S)$ depend on k_1, k_2, and F, this dependence is generally left implicit in our notation.

In words, conditional on copying in the daughter population, the prior on 'who you copy' has zero weight on alleles from the other population, and is uniform over alleles from the same population. The prior, conditional on copying ancestrally, allows any allele to be copied but takes into account the numbers of allele copies sampled so far from each population (k_1 and k_2). This prior extends straightforwardly to the case of unequal drift in the daughter populations, as discussed in Section 3.2.

3.1.2. Mutation probability

Conditional on copying in the daughter population, the probability that the new allele is identical to the copied allele should be (as in LS) an increasing function of k_{z_*}, since when k_{z_*} is large we expect the new lineage to coalesce rapidly into the existing tree. Furthermore, since $S_l = a$ implies a more ancient coalescence time than $S_l = d$ whatever the value of k_1 and k_2, the ancestral copying fidelity is lower. The approach we take is to view the mutation probability in terms of the opportunity for mutation before the copying (coalescence) time t_s. We set this time to its expectation

$$t_s = \mathbb{E}(T_{\text{coal}}|S = s; k_1, k_2, F), \tag{5}$$

where T_{coal} is the coalescence time of the $(k_1 + k_2 + 1)$th lineage.

These expectations can be computed exactly using results from Tavaré (1984) and Fu and Li (1993) (see Appendix C). The mutation probabilities used are based on the assumption that either 0 or 1 mutations occurred on the branch joining the new line to the existing tree and are

$$\begin{aligned} u(h_{k_1+k_2+1}|h_i, s) &= p(h_{k_1+k_2+1}|S_l = s, X_l = i; k_1, k_2, F) \\ &= \begin{cases} 1 - \exp(-\tilde{\theta} t_s) & \text{if } h_{k_1+k_2+1} \neq h_i \\ \exp(-\tilde{\theta} t_s) & \text{if } h_{k_1+k_2+1} = h_i. \end{cases} \end{aligned} \tag{6}$$

An exception occurs when $k = 1$ in which case $2t_a$ is substituted for t_a in order to account for the time on the branch ancestral to the first line ($S = d$ is impossible for the second line sampled, as the random orderings are chosen such that haplotypes are sampled alternately from each population).

The value of $\tilde{\theta}$ depends on the way in which the sampled loci were ascertained. If the loci were selected without regard to whether or not they show polymorphism (resequencing data), then $\tilde{\theta}$ would be treated as a parameter of the model to be estimated. Alternatively, if loci were only included if they were polymorphic in the sample (SNP data) then, following LS, an arbitrary small value of $\tilde{\theta}$ is used. For the results presented in this paper we used $\tilde{\theta} = 1/\mathbb{E}(\mathrm{T}_{\text{total}})$, where $\mathrm{T}_{\text{total}}$ is the expected total length of the full genealogy with n tips, given the sample configuration and the model parameters. This can be calculated exactly (Wakeley and Hey, 1997) but in the implementation described here we used an approximation (Appendix C.3).

3.1.3. Calculating the likelihood of unlinked data

The AC probability under the isolation model is obtained by averaging the mutation probability over the prior probability distribution on the missing data:

$$\begin{aligned} \hat{p_\phi}(h_{k_1+k_2+1}|h_1, \ldots, h_{k_1+k_2}) = \sum_{s\in\{d,a\}} \sum_{i=1}^{k_1+k_2} u(h_{k_1+k_2+1}|h_i, s) \\ \times p(S = s, X = i). \end{aligned}$$

Let $\bar{u}(h_{k_1+k_2+1}|s) = \sum_{i=1}^{k_1+k_2} u(h_{k_1+k_2+1}|h_i, s)p(X = i|S = s)$ denote the emission probability averaged over which allele is copied, given that $S = s$. Then the AC probability under isolation without gene flow is

$$\begin{aligned} \hat{p_\phi}(h_{k_1+k_2+1}|h_1, \ldots, h_{k_1+k_2}) &= p(S = a)\bar{u}(h_{k_1+k_2+1}|S = a) \\ &+ (1 - p(S = a))\bar{u}(h_{k_1+k_2+1}|S = d). \end{aligned} \tag{7}$$

Eqs. (2), (6) and (7), together with the expressions for $p(S = a)$ and t_s given in Appendix C, specify an algorithm for computing the PAC likelihood under the isolation without gene flow model, for a single ordering of alleles. The likelihoods that we actually use are an average of these quantities over a random sample of orderings, subject to taking alleles alternately from the two populations. For the results in this paper we have used 10 random orderings. We implemented the algorithms described in this paper in R (R Development Core Team, 2008).

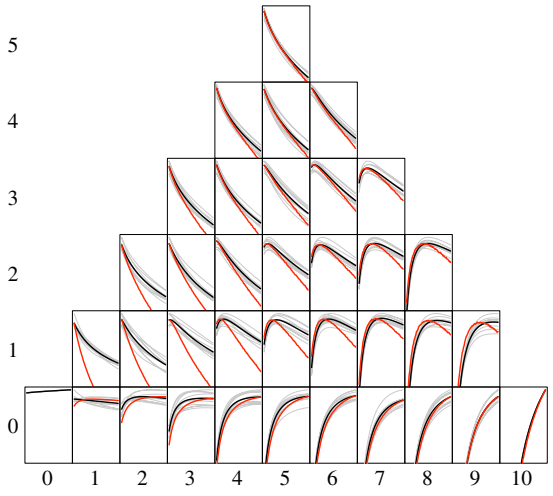

Fig. 4. A comparison of log likelihood curves for F between the PAC (black) and coalescent (red) models. The set of panels show results for all distinct allele count configurations at a single SNP. Each panel shows log likelihood surfaces for a data set at a single SNP, with 10 allele copies sampled from each population. Within each panel, the x axis ranges from $F = 0$ to $F = 0.7$; y-axis values range upwards from 2 log-likelihood units below the maximum. Average PAC likelihood surfaces are in black (individual orderings in grey); coalescent likelihood surfaces are in red. The integers along the bottom and left-hand side of the plot are minor-allele counts in the two populations, specifying the data which were used to compute the likelihood surfaces in the corresponding panel. For example, the panel which lies in the row labelled 2 and in the column labelled 4 corresponds to a data set in which there are 4 copies of the minor allele out of 10 in population 1, and 2 copies out of 10 in population 2. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Results: Unlinked

For unlinked data, likelihood surfaces for F under the standard coalescent model can be estimated accurately by simulation, and our first check on the performance of the new method is to compare these with likelihood surfaces computed under the new PAC model. In order to compute these "coalescent" likelihoods for SNP data, we first estimate the probabilities of all possible single-site allele-count configurations, conditional on the occurrence of a single mutation, as the average lengths of branches (relative to the total length of the genealogy) that would lead to those configurations in 10^6 genealogies simulated from the prior. The likelihood for a data set comprising multiple sites is, under the assumption of no linkage, given by multinomial sampling with cell probabilities equal to the allele-count configuration probabilities, and cell counts equal to the number of sites observed to have those allele-count configurations. An alternative is to compute the allele-count configuration probabilities using the results in Wakeley and Hey (1997). Fig. 4 shows log likelihood surfaces for multiple sets of data at a single diallelic locus, with a sample of 10 allele copies from each daughter population. As described in Section 3.1.2, these PAC likelihoods for SNP data are computed using a fixed, small value of θ. The PAC model does not use information about the ancestral and derived states of the alleles, and the panels correspond to the 36 possible non-equivalent allele-count configurations. The shapes of the surfaces and the locations of the maxima are broadly similar, although it is evident that for several of the configurations the PAC likelihood surface is less tightly curved around the maximum, indicating that the data are slightly less informative about F under the PAC than under the coalescent model. In Fig. 5, likelihood surfaces are shown for two data sets of 60 SNPs, simulated with different values of F. Encouragingly, there is close agreement between the coalescent and PAC likelihood curves.

We investigate the performance of the new maximum PAC likelihood estimator of F by simulating data sets from the infinite-sites coalescent model for a range of true F values, using the

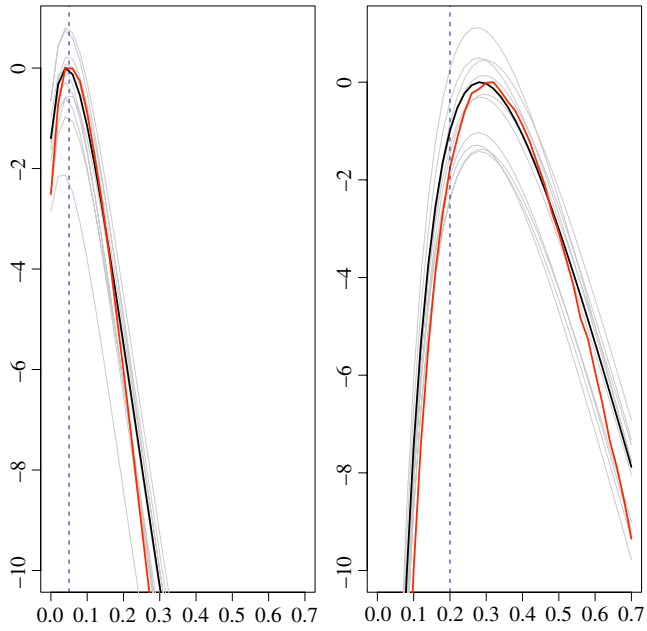

Fig. 5. Inference for F in the unlinked model using the PAC (black) and coalescent (red) models. The plots show relative log likelihood surfaces for two data sets of 60 unlinked SNPs each. The vertical dotted lines indicate the value of F used to simulate data. Results from the PAC likelihood are plotted in black (different orderings in grey); the coalescent log likelihood is in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

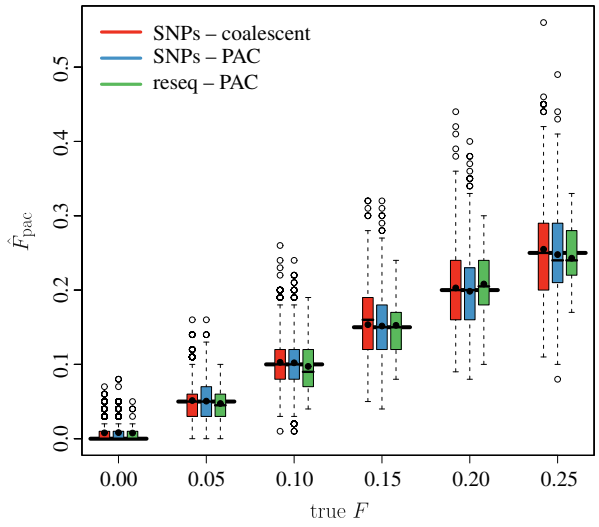

Fig. 6. Unlinked model: Estimation of F. Each panel shows the distribution of MLEs for 1000 data sets simulated with the indicated value of F. Likelihoods were evaluated at points of a grid of F values with spacing 0.01. The boxplots indicate 25%, 50% and 75% quantiles. Long horizontal black bars indicate the location on the y-axis of the true value of F. For resequenced data the model was provided with the per-site value of θ used in the simulations.

software `ms` (Hudson, 2002). In all cases our simulated data sets comprise 10 sampled haplotypes in each population, i.e. we assume that the organisms are haploid or that the haplotype phase was known in advance. For SNP data we simulate datasets with 60 unlinked SNPs for a range of F values. (The SNP number was chosen to match the study design of Conrad et al. (2006)). These data sets were simulated by picking 60 SNPs at random from data generated with a large value of θ. For the resequencing data we simulated data at 17,000 unlinked sites by simulating 17,000 independent genealogies and applying mutations to each at rate $4N\mu = 2\theta = 10^{-3}$ (this value of θ was chosen so that the expected number of segregating sites was approximately 60 at $F = 0$). Likelihoods for resequenced data were computed using the value of θ used in the simulations. For both the coalescent and PAC inference schemes we estimate the likelihood at a grid of F values with grid-spacing 0.01, and take the grid point with the maximum estimated likelihood as the MLE.

Fig. 6 illustrates the behaviour of the $\hat{F}_{\text{pac}}$ estimator under the unlinked model. We construct a confidence interval around the MLE by taking the most extreme grid points whose log likelihood was within 2 of the maximum log likelihood grid point. In Table 1 we give the coverage of this confidence interval, and the root mean square error of the estimators. There is little bias, and the results for the PAC method for the SNP data are very similar to those obtained by using computationally intensive coalescent simulations. Indeed there is a strong linear correlation between the MLE from the PAC and coalescent inference approach (Fig. 7). These results suggest that, despite the approximations, the performance of our PAC copying model at estimating the parameter F is indistinguishable from that of a full likelihood estimator.

For simplicity of presentation we have concentrated on a model in which there has been equal drift in each daughter population. In practice however, investigators frequently need to fit models in which drift has been quite strongly asymmetric (e.g. Hey, 2005; Anderson and Slatkin, 2007). Our method naturally extends to this more general case of unequal population sizes ($N_1 \neq N_2 \neq N_a$). Briefly, in the expression for the prior probability of copying ancestrally (Eq. (31), Appendix C.1), F is replaced by the population-specific parameter F_{z_*}. Conditional on copying ancestrally, the prior on which haplotype is copied is weighted towards haplotypes from the population which has experienced less drift, by conditioning the expectation in Eq. (4) on F_1 and F_2 in addition to k_1 and k_2. ρ and θ are replaced by values scaled by the appropriate population size when considering recombination and mutation events in the different populations. We note that our estimators from unlinked SNPs in the unequal drift case seem to perform reasonably well, although the estimated drift parameters (F_1 and F_2) show some tendency to be more similar than those used to generate the data (results not shown).

Table 1
Coverage rates and Root Mean Square Error of the MLE under the unlinked model for SNP data for a range of values of F. The coverage rates use intervals containing values with log-likelihood within two units of the maximum (see text).

F_{true}	0	0.05	0.1	0.15	0.2	0.25
Coverage						
PAC	0.98	0.96	0.97	0.96	0.97	0.97
Coalescent	0.98	0.95	0.96	0.95	0.95	0.95
RMSE						
PAC	0.01	0.03	0.04	0.04	0.05	0.06
Coalescent	0.01	0.03	0.03	0.04	0.05	0.06

4. Loosely linked data

4.1. Methods: Linked

With loosely linked data, correlation is anticipated between the patterns of polymorphism at nearby sites (over and above that induced by population structure), as a result of limited recombination. Modeling this phenomenon is in general challenging, but failure to do so (i.e by treating the sites as unlinked) will have two undesirable effects: firstly, valuable information that is present in the data about the genealogy along the chromosome will be lost; secondly, confidence intervals can be too narrow, and thus would fail to have the nominal coverage. Since the likelihood can no longer be computed as a product of likelihoods at individual

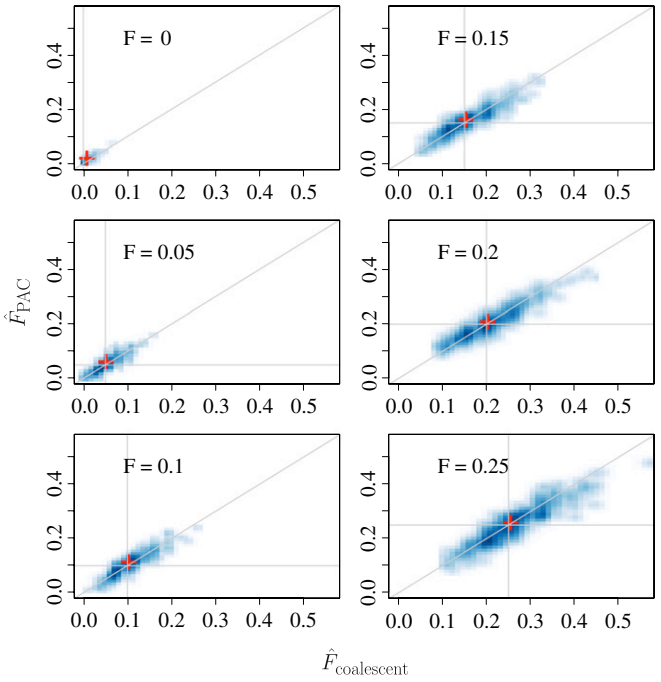

Fig. 7. Unlinked model: Estimation of F. Each panel shows the joint distribution of coalescent (x-axis) and PAC (y-axis) MLEs for 1000 SNP data sets simulated with the indicated value of F. Darker colors indicate higher local density of points. Grey lines indicate the true value of F, and the line $y = x$. Red crosses lie at the mean value of the MLEs. Likelihoods were evaluated at points of a grid of F values with spacing 0.01. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sites, we now switch notation so that h_i refers to haplotype i, as opposed to a single allele copy as it did in the previous section.

Recall that our model involves, at each site l, the unknown copying states S_l (the level at which copying occurs) and X_l (the identity of the haplotype that is copied). Whereas the unlinked case simply required the specification of the prior distribution on (S_l, X_l) at a single site, in the linkage model the copying states at nearby sites are not independent under the prior. Hence it is less straightforward to specify the joint prior distribution on the copying states $(S_1, X_1), \ldots, (S_L, X_L)$ at all sites. As in LS, we use a Markov chain for this prior. Therefore, in addition to the marginal (single-site) prior on (S_l, X_l), we also specify transition probabilities between (S_l, X_l) and (S_{l+1}, X_{l+1}). These govern properties such as the expected length of stretches of ancestral copying, and must be appropriately parameterized by model parameters such as F, as well as by k_1 and k_2. The other parameter that features in these transition probabilities is ρ_l, which corresponds to the population-scaled recombination rate between sites l and $l + 1$. We focus on the case of constant recombination rate along the chromosome, so ρ_l corresponds to the population-scaled, per base-pair rate of recombination (ρ_{bp}) multiplied by the physical distance between site l and $l + 1$. Variation in recombination rates along the chromosome could be incorporated simply by setting the ρ_l individually, according to an estimated genetic map. Since we view $(S_l = s, X_l = i)$ as a statement about the way in which the new lineage coalesces in to the existing genealogy at site l, in order to parameterize the transition probabilities we consider how recombination can result in changes to this genealogical structure.

The transition probabilities have the following form:

$$p(S_{l+1} = s', X_{l+1} = i' | S_l = s, X_l = i) = p_R(S_{l+1} = s' | S_l = s) \times p(X_{l+1} = i' | S_{l+1} = s') + I(i' = i, s' = s) p(NR | S_l = s). \quad (8)$$

In this expression $I()$ is an indicator function that takes the value 1 if its argument is true and 0 otherwise, and $p(NR|S_l = s)$ is the probability, given that copying is at level s, that there is no recombination between sites l and $l + 1$, resulting in no change of state ($i' = i, s' = s$). $p(X_{l+1} = i'|S_{l+1} = s')$ is the probability of copying haplotype i' given copying at level s', and is the same as in the no-linkage model (Eqs. (3) and (4)). The tricky part here is $p_R(S_{l+1} = s'|S_l = s)$, which is the probability, given copying at level s at site l, that a recombination occurs between sites l and $l + 1$ and that as a result copying occurs at level s' at site $l + 1$. Our copying process is an approximation to the coalescent model and so our aim is that, for example, $p_R(S_{l+1} = a|S_l = d)$ approximates the probability, conditional on coalescence of the new lineage in its own population at site l, that there is recombination between sites l and $l + 1$ and the genealogy is altered in such a way that at site $l + 1$ the same line now coalesces in the ancestral population.

We approximate the transition probabilities by considering different classes of genealogical rearrangements which could give rise to the copying transition in question, as illustrated in Fig. 8. That figure is divided into four panels, corresponding to the four different transitions ($d \to d$, $d \to a$, $a \to d$ and $a \to a$). Within each panel, the different classes of genealogical rearrangement are identified by a number (i)–(v). We now explain the expressions that we use for the quantities $p(NR|S_l = s)$ and $p_R(S_{l+1} = s'|S_l = s)$ in more detail. When describing the genealogical motivation for these expressions, we will refer to the ancestral lineage of the haplotype that is copied at site l as the "copied lineage".

4.1.1. Transition probabilities under the linkage model

Transitions from the daughter population. We consider two possibilities involving recombination: a recombination occurs on either the new lineage or the copied lineage, prior to their coalescence (event R_{dc}; genealogies i and iii in the $d \to a$ and the $d \to d$ panels of Fig. 8). The coalescence time of these lineages is assumed to be the conditional expectation t_d. Therefore for the probability of no recombination—and thus no change in state—we use

$$p(NR|S_l = d) = 1 - p(R_{dc}) \quad (9)$$

where

$$p(R_{dc}) = 1 - \exp(-2\rho_l t_d). \quad (10)$$

If there is a recombination, the new lineage subsequently coalesces into the tree at the next site either in the daughter ($S' = d$) or ancestral ($S' = a$) population. Following from Eq. (9), in principle the probability of the transition as a result of recombination should be

$$p_R(S_{l+1} = s'|S_l = d) = p(R_{dc}) p(S_{l+1} = s'|R_{dc}). \quad (11)$$

However, evaluating e.g. $p(S_{l+1} = a|R_{dc})$ requires averaging the probability of surviving back to the ancestral population over the unknown number of uncoalesced lines at the unknown time of the recombination event; in practice we substitute the marginal (prior) probability $p(S = s')$ (Eq. (31), Appendix C.1), which is larger than $p(S = s'|R_{dc})$ in the case $s' = d$ and smaller in the case $s' = a$.

Transitions from the ancestral population. In the case of $a \to a$ and $a \to d$ we classify the recombination events according to whether they occur on the new lineage in the daughter population (event R_d^n), on the copied lineage in the daughter population (event R_d^o) (genealogies i and iii in Fig. 8) or on either line in the ancestral population, before they coalesce (event R_{ac}) (genealogies ii and iv). The length of time available for recombination in a daughter population is F, and in the ancestral population we approximate it by $t_a - F$. Thus we approximate the probability of no recombination (genealogy v) by

$$p(NR|S_l = a) = (1 - p(R_d^n \cup R_d^o \cup R_{ac})) \quad (12)$$

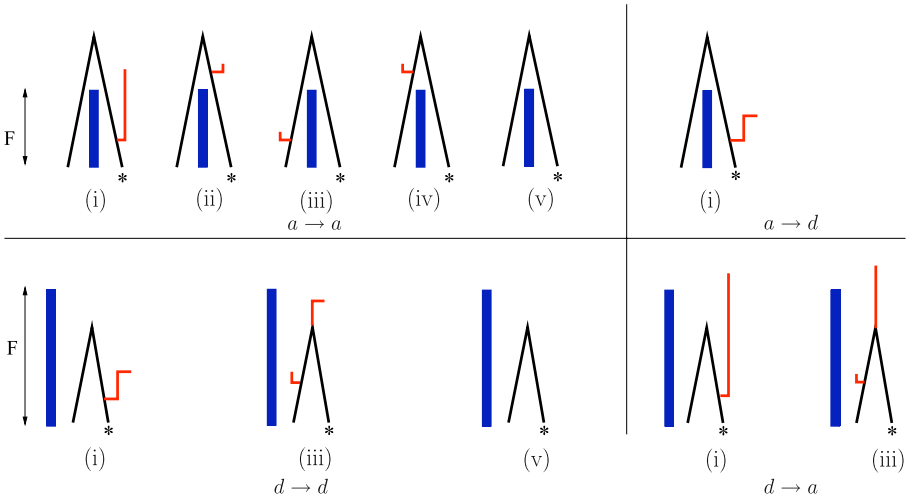

Fig. 8. Transitions between daughter and ancestral copying states. The four panels correspond to the four possible transitions between copying levels ($d \to d$, $d \to a$, $a \to d$ and $a \to a$). Within each panel, we illustrate the various classes of genealogical rearrangement that we consider when approximating the probability of that panel's copying transition. Each class of genealogical rearrangement is illustrated by a diagram of a genealogy of two lineages (in black): the new lineage (marked with an asterisk), and the lineage that it copies at site l. In each genealogy diagram, a thick blue line represents the barrier to gene flow separating the daughter populations. At site $l+1$, the lineage that is copied may be different as a result of recombination in the history of the two samples between sites l and $l+1$. Red lines represent lineages at site $l+1$, and the way they are drawn reflects the way in which the probability of the event being depicted depends on their fate (i.e. on when they coalesce into the rest of the genealogy). Short red rising lines indicate that the transition probability depends only on the occurrence of the recombination event, and not otherwise on the fate of the recombinant line. Long red rising lines indicate that the lineage must remain distinct and enter the ancestral population, prior to its eventual recoalescence. A horizontal terminus to the red line indicates that the line must recoalesce in the daughter population. Red lines without an initial horizontal section do not require a recombination to have occurred (i.e. they already existed at site l). The five types of event are, (i) recombination on the new lineage in the daughter population, (ii) recombination on the new lineage in the ancestral population, (iii) recombination on the copied lineage in the daughter population, (iv) recombination on the copied lineage in the ancestral population, (v) no interrupting event (note that this last event can only contribute to the probability of 'transitions' to the same haplotype at the same level ($s' = s$, $i' = i$)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where

$$p(R_d^n \cup R_d^o \cup R_{ac}) = 1 - \exp(-2\rho_l t_a). \tag{13}$$

The case $a \to d$ is relatively straightforward as only a recombination event on the new lineage in the daughter population (genealogy i), followed by recoalescence in the daughter population, can effect this transition. Thus the expression that we use for the $a \to d$ transition probability as a result of recombination is

$$p_R(S_{l+1} = d | S_l = a) = p(R_d^n) p(S = d | R_d^n), \tag{14}$$

where $p(R_d^n) = 1 - \exp(-\rho_l F)$. Again, we substitute the unconditional prior probability $p(S = d)$ for $p(S = d|R_d^n)$.

The case $a \to a$ is more complex, as this transition can result from any of the following (mutually exclusive) combinations of events:

- a recombination on the new lineage in the daughter population (genealogy i), followed by ancestral recoalescence; probability $p(R_d^n)p(S = a|R_d^n)$;
- no recombination on the new lineage in the daughter population, but recombination on the copied lineage in the daughter population, (genealogy iii); probability $(1 - p(R_d^n))p(R_d^o)$
- recombination on neither the new or copied lineage in the daughter population but an ancestral recombination on one or other lineage, (genealogies ii and iv); probability $(1 - p(R_d^n \cup R_d^o))p(R_{ac})$.

Thus the expression we use for the $a \to a$ transition probability as a result of recombination is

$$p_R(S_{l+1} = a | S_l = a) = p(R_d^n)p(S = a) + (1 - p(R_d^n))p(R_d^o) + (1 - p(R_d^n \cup R_d^o))p(R_{ac}) \tag{15}$$

where

$$p(R_d^o) = p(R_d^n) = 1 - \exp(-\rho_l F)$$
$$p(R_d^n \cup R_d^o) = 1 - \exp(-2\rho_l F)$$
$$p(R_{ac}) = 1 - \exp(-2\rho_l (t_a - F)), \tag{16}$$

and again we have substituted the marginal (prior) probability $p(S = a)$ instead of $p(S = a|R_d^n)$.

4.2. Results: Linked

We now explore the properties of our PAC scheme for loosely linked regions. Again, we focus on the symmetric case in which the amount of drift has been the same in the two daughter populations, and in which the ancestral population had the same effective size as the daughter populations. In keeping with LS and much of the subsequent development of PAC copying models (e.g. Yin et al., 2009; Marchini et al., 2007; Leslie et al., 2008; Price et al., 2009; Hellenthal et al., 2008), we focus here on estimation under the model for SNP data rather than resequencing data (methodological aspects of fitting the linkage model to resequencing data are discussed in Appendix B).

To illustrate how our method uses the information in linkage, and to provide a comparison with LS, we begin by investigating the ability of our method to estimate the recombination parameter ρ (where ρ is ρ_{bp} multiplied by the length of the genomic region simulated). Fig. 9 illustrates the relationship between the maximum PAC likelihood estimator $\hat{\rho}_{pac}$, and the true value of ρ. Since this relationship may vary with the value of F used in the simulation, and when fitting the model, the figure displays results for data simulated using four different values of F. In each case the true value of F was used when fitting the model. When the data are simulated without recombination, $\hat{\rho}_{pac}$ is generally close to zero. As can be seen in Fig. 9, our estimates of ρ are biased, but there is a linear relationship between ρ and $\hat{\rho}_{pac}$, the slope of which varies between 0.25 at $F = 0$ and 0.45 at $F = 0.2$.

Whereas LS were primarily concerned with estimating the recombination rate, our focus here is on estimating the drift parameter F. The linkage model incorporates information about the lengths of haplotypes shared within and between populations into the estimation of F, and therefore these estimates are influenced by the value of ρ used when fitting the model. Fig. 10 illustrates the dependence of $\hat{F}_{pac}$ on F, for some different values of ρ. It is evident that if the value of ρ is increased, the model responds by lowering the estimate of F. A partial explanation of

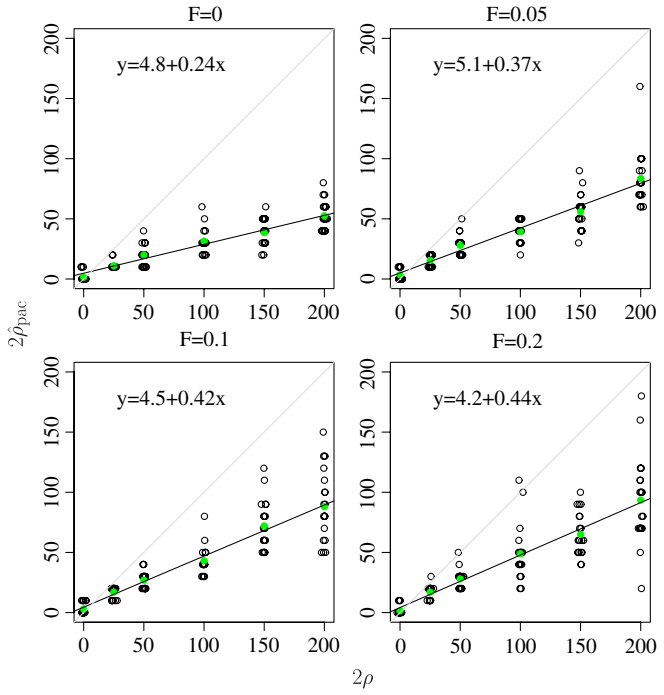

Fig. 9. Dependence of $\hat{\rho}_{pac}$ on ρ. 60 SNPs were simulated using the specified value of ρ for the region, for four different values of F. When fitting the model to estimate $\hat{\rho}_{pac}$ for each region, F was fixed at its true value. The line $y = x$ is shown in light gray. The results of a linear regression of $\hat{\rho}_{pac}$ on ρ are shown as a black line and an equation in each panel.

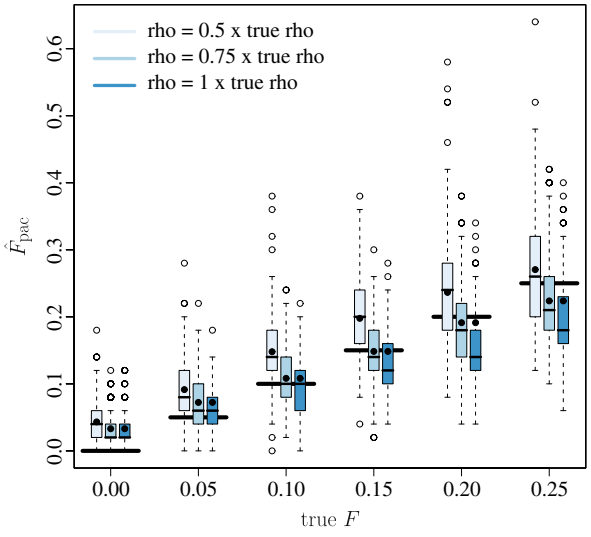

Fig. 10. Linkage model: Estimation of F (symmetric drift, SNP data). The x-axis indicates the value of F used to simulate the data. For each value of F, 200 data sets of 60 SNPs were simulated with $4Nr = 2\rho = 50$. Above each value of F, distributions of $\hat{F}_{pac}$ MLEs are illustrated with boxplots. The 3 boxplots correspond to different values of ρ used when fitting the model. The boxplots indicate 25%, 50% and 75% quantiles of the MLEs and the mean MLEs are indicated by a solid black dots. Horizontal black bars indicate the location on the y-axis of the true value of F.

this phenomenon is provided by considering stretches of similarity between haplotypes from different populations. The distribution of lengths of such stretches under the prior is determined by the product $\rho F = rG$. Therefore if ρ is increased, then the model responds by decreasing the estimate of F.

Although no value of ρ results in unbiased estimation of F across the range of F values we investigated, it is possible to almost

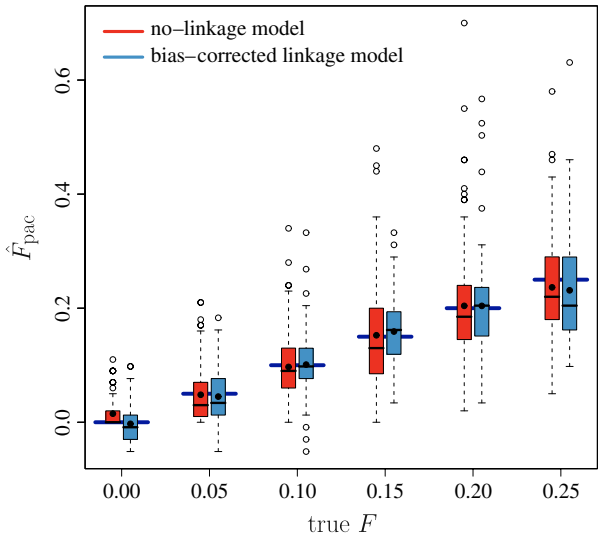

Fig. 11. Bias correction under the linkage model. The figure shows the distribution of MLEs for 100 data sets simulated with the value of F indicated along the x-axis. The boxplots indicate 25%, 50% and 75% quantiles of the MLEs and the mean MLEs are indicated by solid black dots. Horizontal blue bars indicate the location on the y-axis of the true value of F. MLEs from the bias-corrected linkage model (see text) are shown in blue. MLEs resulting from analysing the same data under the no-linkage model are shown in red. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

completely correct this bias. To do this we fit a linear model

$$\hat{F}_{pac} \sim a + bF_{true} \tag{17}$$

for the $\hat{F}_{pac}$ MLE estimates, using a fixed value of ρ across a range of simulations with different F_{true}. We then construct a new estimate with reduced bias by removing the linear component of the bias $(\hat{F}_{pac} - a)/b$. We did this for the values of ρ shown in Fig. 10, using 100 simulations for each F_{true} to estimate the coefficients a and b and then using these to correct MLEs from another 100 simulations. We chose to use the corrected MLEs from $0.5 \times \rho$ as these produced the estimates with the least remaining bias in the mean and median (results not shown). Fig. 11 illustrates the performance of our bias-corrected estimator, alongside the estimator which ignores linkage. The variance of the linkage-model estimators are slightly lower than those ignoring linkage for all values of $F > 0$, demonstrating that the linkage model is successful in extracting the extra information from the data. Encouragingly, the bias-corrected linkage-model estimator has both a lower variance than the no-linkage estimator, and is approximately unbiased. This bias correction also suggests an *ad hoc* method of constructing confidence intervals for an estimated value of $\hat{F}$. This involves recording the range of values of F (on our grid of F) whose log-likelihood fall within 2 of the maximum log-likelihood, representing a confidence interval for our biased estimator, and applying the linear bias correction (Eq. (17)) to this range to obtain a bias-corrected confidence interval. This interval has upward of 80% coverage for the range of F values in the simulations presented in Fig. 11. It is likely that the form of the bias will depend on features of the data such as the sample size and spacing of SNPs, as was found by LS when estimating ρ. Therefore the bias correction would have to be tailored to the data set used by performing simulations, matched to the data in various ways (e.g. according to the number of segregating sites and plausible recombination rate) for a range of F values to estimate a linear form for the bias correction (and to confirm approximate linearity). The complexity of this procedure is a drawback of our method in its current form.

5. Discussion

In principle, the data sets of choice for learning about the evolutionary history of populations are those with physical linkage. This preference stems from the value of information about local genealogical structure (i.e. along the chromosome) when trying to learn about such things as the timing of population splitting, gene flow and admixture events. As a result, even in non-intensively studied species many data sets of linked variation from the nuclear genome have now been assembled with the aim of learning about population history, and for the last few years there has been a pressing need for statistical methods capable of extracting much of the information that they contain.

Here, we have introduced an approximate likelihood method to estimate the parameters of a population split model from recombining data. The method is an extension of the PAC copying model of Li and Stephens (2003) to two populations. The main idea is to allow 'copying' to occur at two temporal 'levels': within the same daughter population, and in the ancestral population. In the former case the copied haplotype is necessarily from the same population, whereas in the latter case a haplotype from either population may be copied. The prior on these possibilities, as well as the mutation probabilities and the transition probabilities between copying states at adjacent sites, are based on the standard coalescent model of a population split.

5.1. The no-linkage model

It is encouraging that the PAC estimator under our no-linkage model performs comparably to the coalescent estimator (Figs. 6 and 7, and Table 1). The latter, as long as it is based on sufficiently many simulations, is optimal in the sense that it is a maximum likelihood estimator under the same model as that which generated the data, and yet it does not differ appreciably in bias or variance from the PAC estimator. These results are encouraging for two reasons: firstly because, while allowing for mutation, they provide a computationally efficient alternative to the simulation-based method for unlinked data; and secondly because they demonstrate that the joint prior distribution on copying states at all sites used in the linkage model is built on solid marginal (single-site) foundations.

5.2. The linkage model

The extension of the no-linkage model to model linkage also follows the work of Li and Stephens (2003): we specify transition probabilities between the copying states at adjacent sites, parameterized appropriately by the recombination rate parameter between the two sites and the other model parameters (Section 4.1); this gives rise to a hidden Markov model under which the likelihood (and other useful quantities) can be computed using standard algorithms (Appendices A and B).

However, whereas our approximations to the marginal coalescent model evidently worked well in the unlinked case, the transition probabilities in the linkage model involve approximating the coalescent-with-recombination under the population splitting model (Section 4.1), which is more challenging. As a result there are some related problematic issues regarding estimation of the model parameters. Firstly, the recombination rate parameter is underestimated across a range of true ρ values (Fig. 9). Since the linkage model is making use of the lengths of shared haplotypes, estimates of the drift parameter F are necessarily influenced by the value of ρ (Fig. 10). However, the value of ρ which gives rise to an unbiased estimator of F depends on the value of F that was used to generate the data, and thus is neither necessarily the true value of ρ, nor the ML estimate of ρ that results when F is fixed at its true value. As a result, joint estimation of all the model parameters is challenging. This problem could potentially be circumvented by obtaining an unbiased estimate of ρ using another approach. With ρ held fixed, bias in the estimation of F varies with the value of F used to generate the data (Fig. 10). However, one effective way forward in this situation is to make a correction to the estimates based on the simulation results (Fig. 11). Thus a reasonably unbiased estimator of F can be constructed that utilizes the extra information contained in linked sites.

It is hard to pinpoint the source of the bias in our estimates of F utilizing linkage. We explored a number of other forms for the transition probabilities that were more accurate descriptions of the coalescent with recombination but these did not result in substantial reduction in the bias. One of the most noticeable features of the estimator $\hat{F}_{\mathrm{pac}}$ with linkage is the upward bias even when the data are simulated from an unstructured model (Fig. 10). This bias is absent when no attempt is made to model linkage. Naively, one would expect that the introduction of structure into a model of data from an unstructured population would result in a decrease in likelihood because of penalties associated with copying haplotypes in the 'other population'. A possible explanation is that this cost is overcome by an increase in likelihood resulting from the better ability of the structured model to fit variation in coalescence times around their expectations. For example, when modeling an unstructured population, Stephens and Scheet (2005) found it advantageous to increase the dimension of the hidden state space in a way that can be viewed as allowing two different 'copying times' as opposed to the single 'copying time' of Li and Stephens (2003). Since the unlinked model does not appear to benefit in this way, perhaps it is the time for recombination, rather than the time for mutation, which is being better fit. While additional work will be needed to understand the sources of bias in our estimates under the linkage model, our approach may also be of use in applications where there is a need to model haplotype structure, but where estimation of population history parameters is not the primary goal.

5.3. Gene flow between daughter populations

While we have concentrated here on a model in which there is no gene flow after the population split, learning about the extent of migrant ancestry in a population is a very important challenge (Hey, 2006). Therefore, one of the prime uses of the information contained in haplotype patterns may be to robustly identify haplotypes contributed to the population by dispersal events. Although a migrant contribution to ancestry in the past couple of generations can be detected using unlinked markers (Rannala and Mountain, 1997), dating older migrant contributions is aided by modeling how a migrant's genome is broken down by recombination as it is passed down through the generations (e.g. Falush et al., 2003). Migration events in the past tens of generations can be detected using markers which are unlinked in the parental populations (e.g. Falush et al., 2003; Pool and Nielsen, 2008), however, older migrant haplotypes will be on a similar length scale as background linkage disequilibrium, and may be difficult to distinguish from shared ancestral haplotypes.

The method developed here provides both a null model for the lengths of shared haplotypes when no migration occurs, and a natural framework for learning about migration events via the identification of haplotypes shared between populations. To illustrate this we simulated a sample in which two of the chromosomes have stretches of migrant ancestry (Fig. 12). Modeling each haplotype in turn conditional on all the others, we estimated the posterior probability of copying ancestrally along

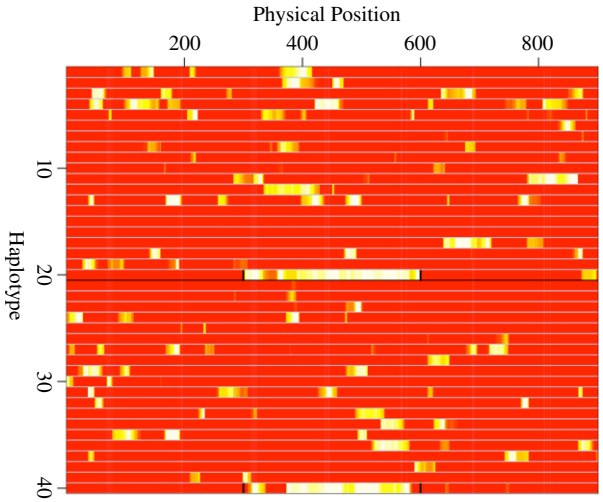

Fig. 12. Visualizing migrant chunks of chromosome. Each row in the figure represents a single haplotype, with lighter colors indicating higher posterior probability that copying is ancestral ($S_l = a$) when that haplotype is added as the final haplotype in the sample. 20 haplotypes were simulated from each of two populations that separated $F = 0.15$ units of drift-scaled time ago. We simulated 900 SNPs across a 900 kb region with a population-scaled recombination rate of $4Nr = 1$ per kb. To create a stretch of migrant haplotype (marked by short black vertical lines) the middle third of the first haplotype in each population was replaced by a haplotype simulated from the other population.

each chromosome. This was calculated by summing probabilities associated with copying a particular haplotype ancestrally:

$$P(S_l = a|h_1, \ldots, h_{n-1}) = \sum_{i=1}^{n-1} P(S_l = a, X = i|h_1, \ldots, h_{n-1}). \quad (18)$$

$P(S_l = A, X = i|h_1, \ldots h_{n-1})$ can be calculated using the forward and backward algorithms (see Appendix A). For haplotypes that do not have migrant ancestry, stretches of ancestral copying are observed due to shared haplotypes from the ancestral population. The migrant regions are evident as long stretches of ancestral copying, due to the migrant haplotype being more closely related to those in the other population than to those from the population it was sampled from.

A natural way to extend our model to include migration would be to add an extra copying state that allows the current haplotype from a population to copy from the other population at the daughter level. These migrant copying events would tend to persist along the chromosome for longer than ancestral copying events, potentially permitting inference for parameters describing the history of gene flow. Such modeling would be useful for judging the timing of migration events (e.g. Pool and Nielsen, 2008) and for understanding the history of particular loci and alleles.

Analysing data sets of loosely linked variation data is probably the way forward for fitting models of population history. Furthermore, the ability to model haplotype structure in a multi-population setting may be useful in contexts other than traditional parameter inference. While some challenges remain in adapting the LS copying model to a multi-population setting, we believe that doing so is one of the most promising approaches to making use of the information contained in haplotypes from large contiguous regions of the genome.

Acknowledgments

The authors would like to thank the Pritchard and Przeworski labs for helpful discussions. DD was supported in the final stages of this work by grant no. 072974/Z03/Z from the Wellcome Trust. JKP and GC were supported by grants from the David and Lucile Packard Foundation, and the Howard Hughes Medical Institute. GC was also supported by funds from UC Davis. Two anonymous reviewers provided many helpful comments.

Appendix A. The 'forward' and 'backward' algorithms

When modeling loosely linked data, the approximate conditional sampling distributions of haplotypes have the form of a 'hidden Markov model' and so evaluation of the probability of the observation sequence (the haplotype), and evaluation of the posterior probability distribution on hidden states at each site, are standard procedures, making use of the 'forward' and 'backward' algorithms (see e.g. Rabiner, 1989). However, since it is important to avoid certain computational inefficiencies, we describe the computations here as they apply to the PAC model.

Define the 'forward probability' for each hidden state at site l to be the joint probability of the hidden state and the data up to and including site l, conditional on the haplotypes sampled so far, as a function of model parameters ϕ:

$$\alpha_l(s, i) = p_\phi(h_{k_1+k_2+1,1}, \ldots, h_{k_1+k_2+1,l}, S_l = s, X_l = i|h_1, \ldots, h_{k_1+k_2}), \quad (19)$$

where $h_{i,l}$ refers to the allele copy present at site l on haplotype i. We will leave the dependence on ϕ implicit hereafter. The approximate conditional probability $p(h_{k_1+k_2+1}|h_1, \ldots, h_{k_1+k_2})$ is obtained as usual by summing the forward probabilities at the last site,

$$p(h_{k_1+k_2+1}|h_1, \ldots, h_{k_1+k_2}) = \sum_{i,s} \alpha_L(s, i), \quad (20)$$

and the PAC likelihood based on a single ordering of all the haplotypes is computed using (2).

The forward algorithm is initialized by setting the forward probabilities at the first (say leftmost) locus to

$$\alpha_1(s, i) = \tilde{p}(S = s)p(X = i|S = s)u(h_{k_1+k_2+1,1}|h_{i,1}, S = s) \quad (21)$$

for each hidden state pair (s, i). For resequenced data the physical spacing of consecutive pairs of marker loci is equal (1 bp) and therefore, under the assumption of homogeneous recombination rates along the chromosome, so is the rate of recombination between them. In this case, the transition probabilities between the hidden states are the same for all consecutive pairs of loci and the ergodic Markov chain specified in Section 4.1.1 has a stationary distribution $\pi(s, i)$. We evaluate this using the normalized first eigenvector of the transition matrix, and use $\tilde{p}(S = s) = \sum_i \pi(s, i)$ in the initialization, thus ensuring that the prior distribution on the hidden states does not depend on the chromosomal location. For irregularly-spaced SNPs however, the rates of recombination between consecutive marker loci vary, and therefore so do the transition probabilities, and the Markov chain on hidden states has no stationary distribution. In this case we use $\tilde{p}(S = s) = p(S = s)$ and the prior therefore differs along the chromosome.

The forward probabilities at sites to the right are computed recursively, using the values at the adjacent site to the left, according to

$$\alpha_{l+1}(s', i') = u(h_{k_1+k_2+1,l+1}|h_{i',l+1}, s') \times \sum_{s,i} \alpha_l(s, i)p(S_{l+1} = s', X_{l+1} = i'|S_l = s, X_l = i). \quad (22)$$

For computational efficiency it is important to avoid an unnecessary extra loop over haplotypes by storing the quantities

$$f_l^{(d)} = \sum_i \alpha_l(d, i)$$

and

$$f_l^{(a)} = \sum_i \alpha_l(a, i)$$

and performing the computation in (22) instead as

$$\begin{aligned}\alpha_{l+1}(s', i') = \; & u(h_{k_1+k_2+1,l+1}|h_{i',l+1}, s') \\ & \times \Big[f_l^{(d)} p(S_{l+1} = s'|S_l = d)p(X = i'|s') \\ & + f_l^{(a)} p_R(S_{l+1} = s'|S_l = d)p(X = i'|s') \\ & + \alpha_l(s', i')p(NR|S_l = s)\Big]. \end{aligned} \tag{23}$$

We note that this form comes about because conditional on a recombination and the level s', the choice of haplotype is specified by the prior $p(X = i'|s')$ which reduces the complexity of the transition probabilities. This makes the computation time for the AC probability of the new haplotype linear in the number of haplotypes added, rather than quadratic.

In order to evaluate the posterior probability on hidden copying states, as in Section 5.3, we need to introduce the backward algorithm. Define the 'backward probability' for each hidden state at site l to be the joint probability of all the data to the right of l, conditional on the hidden state and the haplotypes observed so far, as function of the model parameters:

$$\begin{aligned}\beta_l(s, i) = \; & p_\phi(h_{k_1+k_2+1,l+1}, \ldots, h_{k_1+k_2+1,L}|S_l = s, \\ & X_l = i, h_1, h_2, \ldots, h_{k_1+k_2}). \end{aligned} \tag{24}$$

The posterior probability that site l is in hidden state (s, i) is proportional to the product of the forward and backward probabilities at that site

$$p(S_l = s, X_l = i|h_1, \ldots, h_{k+1}) = \frac{\alpha_l(s, i)\beta_l(s, i)}{\sum_{s',i'} \alpha_l(s', i')\beta_l(s', i')}. \tag{25}$$

The backward algorithm is initialized by setting these probabilities to 1 for all hidden states at the last locus (since there is no data to the right of the last locus). The backward probabilities at loci to the left are computed recursively, using the values at the adjacent site to the right, according to

$$\begin{aligned}\beta_l(i, s) = \; & \sum_{i',s'} p(S_{l+1} = s', X_{l+1} = i'|S_l = s, X_l = i) \\ & \times u(h_{k_1+k_2+1,l+1}|h_{i',l+1}, s')\beta_{l+1}(s', i'). \end{aligned} \tag{26}$$

The analogous efficiency measure in the backward algorithm to that described above for the forward algorithm is to store the quantities $b_{l+1}^{(d)} = \sum_i u(h_{k_1+k_2+1,l+1}|h_{i,l+1}, d)\beta_{l+1}(d, i)$ and $b_{l+1}^{(a)} = \sum_i u(h_{k_1+k_2+1,l+1}|h_{i,l+1}, a)\beta_{l+1}(a, i)$ and to perform the computation in (26) instead as

$$\begin{aligned}\beta_l(s, i) = \; & p_R(S_{l+1} = d|S_l = s)b_{l+1}^{(d)} + p_R(S_{l+1} = d|S_l = s)b_{l+1}^{(a)} \\ & + p(NR|S_l = s)u(h_{k_1+k_2+1,l+1}|h_{i,l+1}, s)\beta_{l+1}(s, i). \end{aligned} \tag{27}$$

Appendix B. Computing the PAC likelihood efficiently for resequenced data

Resequenced data may feature blocks of sites in which the k haplotypes sampled so far all have the same allele as the $(k_1 + k_2 + 1)$th haplotype. It is unnecessary to compute the forward and backward probabilities explicitly at each such site because the emission probabilities remain constant, and if the blocks of monomorphic sites are large it may be computationally inefficient to do so. Let the emission probability of observing the same allele i as that on the copied haplotype, conditional on $S = s$, be

$$u_0(s) = u(i|i, s).$$

Let P be an $m \times m$ matrix, where $m = k_{z_*} + k_1 + k_2$, containing the probabilities of all the possible transitions multiplied by the corresponding emission probability, with 'daughter' events preceding 'ancestral' events along each margin. (z_* is shorthand notation for $z_{k_1+k_2+1}$, i.e. the label of the population from which the new lineage is sampled.) That is,

$$P_{i,i'} = \begin{cases} p(S_{l+1} = d, & X_{l+1} = i'|S_l = d, X_l = i)u_0(d) \\ & \text{if } i \le k_{z_*} \text{ and } i' \le k_{z_*} \\ p(S_{l+1} = a, & X_{l+1} = i'|S_l = d, X_l = i)u_0(a) \\ & \text{if } i \le k_{z_*} \text{ and } i' > k_{z_*} \\ p(S_{l+1} = d, & X_{l+1} = i'|S_l = a, X_l = i)u_0(d) \\ & \text{if } x > k_{z_*} \text{ and } x' \le k_{z_*} \\ p(S_{l+1} = a, & X_{l+1} = i'|S_l = a, X_l = i)u_0(a) \\ & \text{if } x > k_{z_*} \text{ and } x' > k_{z_*}. \end{cases}$$

In the forward case, suppose that site l is the leftmost of a block of B monomorphic sites. The forward probabilities at site $l + B - 1$ are required so that those at the polymorphic site $l + B$ can be computed. They are

$$\alpha_{l+B-1} = \alpha_l P^B, \tag{28}$$

where α_l is a row vector containing the forward probabilities in the order corresponding to the margins of P. That is

$$\alpha_l = [\alpha_l(d, 1), \ldots, \alpha_l(d, k_{z_*}), \alpha_l(a, 1), \ldots, \alpha_l(a, k_1 + k_2)]. \tag{29}$$

In the backward case, suppose that l is the site to the left of the rightmost site in a block of B monomorphic sites. The backward probabilities at polymorphic site $l - B + 1$ are required so that the backward probabilities at the site to the left can be computed. They are

$$\beta_{l-B+1} = P^B \beta_l, \tag{30}$$

where β_l is a column vector arrayed in the same way as α_l. The matrix P^B can be computed as usual via an eigenvector decomposition.

Appendix C. Results from coalescent theory used in the PAC likelihood

C.1. The prior on the missing data (S, X)

This prior takes the form

$$p(S = s, X = i) = p(S = s)p(X = i|S = s),$$

for $s \in \{a, d\}$ and $i \in \{1, \ldots, k_1 + k_2\}$. $p(S = a) = p(S = a; k, F)$ is the marginal (single site) probability that a newly sampled chromosome coalesces into the existing tree in the ancestral population (i.e. more than F units of scaled time in the past), when k chromosomes have already been sampled from the same population.

Let $H_k(t)$ be the probability that a newly sampled $(k+1)$th line has not coalesced by scaled time t, so that the desired quantity is $p(S = a) = H_k(F)$, and let $q(n, j)$ be the probability that a particular one of n lines is not involved in any coalescence events as n lines coalesce to j:

$$q(n, j) = \begin{cases} 1 & \text{if } n = 1 \text{ and } j = 1 \\ \prod_{i=j+1}^{n} \frac{\binom{i-1}{2}}{\binom{i}{2}} = \frac{j(j-1)}{n(n-1)} & \text{if } n > 1 \text{ and } 0 < j \le n \\ 0 & \text{otherwise.} \end{cases}$$

$p(S = a)$ can be obtained by averaging $q(k+1, J)$ over the unknown number J of distinct lines at scaled time F:

$$p(S = a; k, F) = H_k(F) = \sum_{j=1}^{k+1} q(k+1, j) g_{k+1,j}(F), \tag{31}$$

where $g_{ij}(t)$ is the probability that i lines coalesce to $j \leq i$ over scaled time t. Exact expressions for these transition probabilities of the coalescent process are given by Tavaré (1984).

The second term $p(X = i|S = s)$ is given in Eqs. (3) and (4). The expectation in Eq. (4) is computed as

$$\mathbb{E}\left(\frac{J_{z_i}}{J_1 + J_2}\right) = \sum_{j_1=1}^{k_1} \sum_{j_2=1}^{k_2} \left(\frac{j_{z_i}}{j_1 + j_2}\right) g_{k_1,j_1}(F) g_{k_2,j_2}(F). \tag{32}$$

C.2. Expected coalescence times

t_d is the expected coalescence time of the new line conditional on coalescence in the daughter population (i.e. before F, looking backwards in time). Since $1 - H_k(t)$ is the cumulative density function of the coalescence time, t_d can be obtained as

$$t_d = \int_0^F H_k(t) \mathrm{d}t. \tag{33}$$

We evaluate the integral numerically in *R*.

t_a is the expected coalescence time of the new line conditional on coalescence in the ancestral population (i.e. after F, looking backwards in time). Conditional on the numbers J_1 and J_2 of distinct ancestral lines entering the ancestral population from the two daughter populations, this is simply F plus the expected time to coalescence of the $(J_1 + J_2)$th line under panmixia, which is $2/(J_1 + J_2)$ (Fu and Li, 1993). t_a can therefore be calculated by averaging this quantity over the joint distribution on (J_1, J_2). Since J_1 and J_2 are independent, for the case in which the new line was sampled in population 1,

$$t_a = F + \frac{1}{p(S = a)} \sum_{j_1=1}^{k_1+1} g_{k_1+1,j_1}(F) q(k_1 + 1, j_1) \times \sum_{j_2=1}^{k_2} g_{k_2,j_2}(F) \frac{2}{j_1 + j_2}.$$

When $F = 0$, this gives $t_a = \frac{2}{k_1+k_2+1}$, and, from Eq. (13), we have that the copying switch rate is $\frac{4\rho_I}{k_1+k_2+1}$. This differs from Li and Stephens (2003) in which the equivalent quantity is (in our notation) $\frac{2\rho_I}{k_1+k_2}$; as a result our estimates of ρ are systematically lower than those of LS.

C.3. Approximating the expected total length of branches in the genealogy

We approximate the total length of the branches of the genealogy of the entire sample $(n_1 + n_2)$ by

$$\mathbb{E}(\mathrm{T_{total}}) \approx \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} g_{n_1,j_1}(F) g_{n_2,j_2}(F) \times \left(\frac{F}{\frac{1}{j_1} - \frac{1}{n_1}} \sum_{i=2}^{j_1} \frac{1}{i-1} + \frac{F}{\frac{1}{j_2} - \frac{1}{n_2}} \sum_{i=2}^{j_2} \frac{1}{i-1} + 2 \sum_{i=2}^{j_1+j_2} \frac{1}{i-1} \right). \tag{34}$$

This averages the approximate expected total branch length given the number of lineages J_1 and J_2 entering the ancestral population over the exact distribution of the number of remaining lineages (J_1, J_2). Conditional on J_1, J_2 we use the exact expression for the total length of the genealogy in the ancestral population, i.e. Watterson's constant $2 \sum_{i=2}^{j_1+j_2} \frac{1}{i-1}$, while we approximate the total length of the portion of the genealogy in daughter population p by assuming that the final coalescence event $(j_p + 1 \to j_p)$ occurred exactly at F and that each interval contributes its expected total time under the standard neutral model ($i \frac{2}{i(i-1)}$ while there are i lineages), scaling the total height of the tree to fit into time F (the scaling factors $\frac{F}{\frac{1}{j_p} - \frac{1}{n_p}}$). However, it was pointed out to us by reviewers that an exact expression for this quantity is given in Wakeley and Hey (1997) and that the conditional distribution of coalescence times given the number of ancestral lineages at a specific time was studied by Blum and Rosenberg (2007).

References

Anderson, E.C., Slatkin, M., 2007. Estimation of the number of individuals founding colonized populations. Evolution 61 (4), 972–983.

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162 (4), 2025–2035.

Becquet, C., Przeworski, M., 2007. A new approach to estimate parameters of speciation models with application to apes. Genome Research 17 (10), 1505–1519.

Becquet, C., Przeworski, M., 2009. Learning about modes of speciation from computational approaches. Evolution (in press).

Blum, M.G., Rosenberg, N.A., 2007. Estimating the number of ancestral lineages using a maximum-likelihood method based on rejection sampling. Genetics 176, 1741–1757.

Conrad, D., Jakobsson, M., Coop, G., Wen, X., Wall, J., Rosenberg, N., Pritchard, J., 2006. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature Genetics 38, 1251–1260.

Cornuet, J., Beaumont, M., 2007. A note on the accuracy of PAC-likelihood inference with microsatellite data. Theoretical Population Biology 71 (1), 12–19.

Cornuet, J.M., Santos, F., Beaumont, M.A., Robert, C.P., Marin, J.M., Balding, D.J., Guillemaud, T., Estoup, A., 2008. Inferring population history with DIY ABC: A user-friendly approach to approximate Bayesian computation. Bioinformatics 24, 2713–2719.

Ewens, W.J., 1990. Population genetics theory—The past and the future. In: Lessard, S. (Ed.), Mathematical and Statistical Problems in Evolution. Kluwer, pp. 177–227.

Falush, D., Stephens, M., Pritchard, J., 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164 (4), 1567–1587.

Fearnhead, P., Donnelly, P., 2001. Estimating recombination rates from population genetic data. Genetics 159 (3), 1299–1318.

Fu., Y.-X., Li., W.-H., 1993. Statistical tests of neutrality of mutations. Genetics 133, 693–709.

Gay, J., Myers, S., Mcvean, G., 2007. Estimating meiotic gene conversion rates from population genetic data. Genetics 177, 881–894.

Hellenthal, G., Auton, A., Falush, D., 2008. Inferring human colonization history using a copying model. PLoS Genet. 4, e1000078.

Hey, J., 2005. On the number of new world founders: A population genetic portrait of the peopling of the Americas. PLoS Biol 3 (6), e193.

Hey, J., 2006. Recent advances in assessing gene flow between diverging populations and species. Current Opinion in Genetics and Development 16, 592–596.

Hey, J., Machado, C., 2003. The study of subdivided populations—New hope for a difficult and divided science. Nature Reviews. Genetics 4 (7), 535–543.

Hey, J., Nielsen, R., 2004. Multilocus methods for estimating population sizes migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167 (2), 747–760.

Hey, J., Nielsen, R., 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the National Academy of Sciences 104 (8), 2785.

Hilton, H., Kliman, R.M., Hey, J., 1994. Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution 48, 1900–1913.

Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18 (2), 337–338.

Leman, S., Chen, Y., Stajich, J., Noor, M., Uyenoyama, M., 2005. Likelihoods from summary statistics: Recent divergence between species. Genetics 171 (3), 1419–1436.

Leslie, S., Donnelly, P., Mcvean, G., 2008. A statistical method for predicting classical HLA alleles from SNP data. American Journal of Human Genetics 82, 48–56.

Li., N., Stephens, M., 2003. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165 (4), 2213–2233.

Marchini, J., Howie, B., Myers, S., Mcvean, G., Donnelly, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature 906–913.

Mayr, E., 1942. Systematics and the Origin of Species. Columbia University Press.

Nicholson, G., Smith, A., Jónsson, F., Gústafsson, O., Stefánsson, K., Donnelly, P., 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society B 64 (4), 695–715.
Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154 (2), 931–942.
Nielsen, R., Slatkin, M., 2000. Likelihood analysis of ongoing geneflow and historical association. Evolution 54 (1), 44–50.
Nielsen, R., Wakeley, J., 2001. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. Genetics 158 (2), 885–896.
Pool, J.E., Nielsen, R., 2008. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics doi:10.1534/genetics.108.098095.
Price, A., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D., Myers, S., 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations (submitted for publication).
Pritchard, J., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155 (2), 945–959.
Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W., 1999. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. Molecular Biology and Evolution 16, 1791–1798.
Putnam, A., Scriber, J., Andolfatto, P., 2007. Discordant divergence times among Z-chromosome regions between two ecologically distinct swallowtail butterfly species. International Journal of Organic Evolution 61 (4), 912–927.
R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing. ISBN 3-900051-07-0. Vienna, Austria URL http://www.R-project.org.
Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of IEEE 77 (2), 257–286.
Rannala, B., Mountain, J., 1997. Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences 94 (17), 9197–9201.
Roychoudhury, A., Felsenstein, J., Thompson, E., 2008. A two-stage pruning algorithm for likelihood computation for a population tree. Genetics 180, 1095–1105.
Roychoudhury, A., Stephens, M., 2007. Fast and accurate estimation of the population-scaled mutation rate, θ, from microsatellite genotype data. Genetics 176 (2), 1363.
Stephens, M., 2001. Inference under the coalescent. In: Handbook of Statistical Genetics. Wiley, pp. 213–238.
Stephens, M., Scheet, P., 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. American Journal of Human Genetics 76 (3), 449–462.
Tavaré, S., 1984. Line of descent and genealogical processes and their applications in population genetics models. Theoretical Population Biology 26, 119–164.
Wakeley, J., 1996. Distinguishing migration from isolation using the variance of pairwise differences. Theor. Popul. Biol 49 (3), 369–386.
Wakeley, J., 2008. Coalescent Theory. Roberts & Company.
Wakeley, J., Hey, J., 1997. Estimating ancestral population parameters. Genetics 145 (3), 847–855.
Wilson, D.J., Mcvean, G., 2006. Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172 (3), 1411–1425.
Yin, J., Jordan, M., Song, Y., 2009. Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. ISMB. In review.