Published in final edited form as:

J Mol Evol. 2014 May ; 78(5): 279–292. doi:10.1007/s00239-014-9620-5.

Bayesian inference of local trees along chromosomes by the sequential Markov coalescent

Chaozhi Zheng,

Department of Statistics, Box 354322, University of Washington, Seattle, WA 98115-4322, USA,
Tel.: (206) 543-7237, Fax: (206) 685-7419

Mary K. Kuhner, and

Department of Genome Sciences, Box 355065, University of Washington, Seattle, WA
98115-5065, USA, Tel.: (206) 543-8751, Fax: (206) 685-7301

Elizabeth A. Thompson

Department of Statistics, Box 354322, University of Washington, Seattle, WA 98115-4322, USA,
Tel.: (206) 685-0108, Fax: (206) 685-7419

Chaozhi Zheng: chaozhi@gmail.com; Mary K. Kuhner: mkkuhner@uw.edu; Elizabeth A. Thompson: eathomp@uw.edu

Abstract

We propose a genealogy sampling algorithm, SMARTree, that provides an approach to estimation from SNP haplotype data of the patterns of coancestry across a genome segment among a set of homologous chromosomes. To enable analysis across longer segments of genome, the sequence of coalescent trees is modeled via the modified sequential Markov coalescent (Marjoram and Wall, 2006). To assess performance in estimating these local trees, our SMARTree implementation is tested on simulated data. Our base data set is of the SNPs in ten DNA sequences over 50kb. We examine the effects of longer sequences and of more sequences, and of a recombination and/or mutational hotspot. The model underlying SMARTree is an approximation to the full recombinant-coalescent distribution. However, in a small trial on simulated data, recovery of local trees was similar to that of LAMARC (Kuhner et al., 2000a), a sampler which uses the full model.

Keywords

coalescent; ancestral recombination graph; sequential Markov coalescent; Bayesian inference; Markov chain Monte Carlo

1 Introduction

In the presence of recombination, the genealogy of a population of chromosomes can be represented by an ancestral recombination graph (ARG), with historical coalescence events appearing as merges of branches looking backward in time, and historical recombination events appearing as splits (Figure 1 upper panel). The genealogy relating the alleles at a

Correspondence to: Elizabeth A. Thompson, eathomp@uw.edu.

Present address: C. Zheng, Biometris, Wageningen University and Research Centre, PO Box 100, 6700 Wageningen, The Netherlands

single location is called the local tree (also known as the marginal tree or interval tree) of that location. Along the chromosome, transitions between local trees occur due to the effects of historical recombinations (Figure 1 lower panel). The collection of local trees provides an alternative representation of the genealogy. Local trees can be extracted from the ARG, but not vice versa because details of the historical recombinations are lost in the local-trees representation.

Our goal is to infer from SNP haplotype data the ancestral relationships across a genome segment of interest, among the genomes of a set of diploid individuals sampled from a population. This *identity-by-descent* (IBD) structure is fundamental to analysis of genetic data, whether for genetic mapping, heritability and association analyses of trait data, or inference of population genetic parameters. (See Thompson (2013) for a recent review.) Although simpler models for IBD inference from population data have been widely used (for example, Browning and Browning (2010)), the full specification of the hierarchy of IBD partitions of a set of chromosomes across a segment of genome is that given by the local coalescent trees. Given local trees, likelihood-based methods for linkage disequilibrium mapping are readily implemented (Zollner and Pritchard, 2005; Smith and Kuhner, 2009). In this paper, the focus of interest is therefore on inference of these local trees, and particularly on their topologies and relative branch lengths.

If we adopt the simplifying model of a random sample from a Wright-Fisher population, chromosome histories can be modeled by the coalescent with recombination (CwR) model (Griffiths and Marjoram, 1996). In this model, the sample size is assumed to be much smaller than the population size, and multiple events (coalescences or recombinations) do not occur simultaneously. As a consequence, each local tree is a bifurcating coalescent tree described by the standard coalescent without recombination.

Likelihood inference of the unobserved ARG under the CwR model is extremely challenging. Maximum likelihood estimations have been implemented by importance sampling (Griffiths and Marjoram, 1996; Fearnhead and Donnelly, 2001) and by Markov chain Monte Carlo (MCMC) sampling (Kuhner et al., 2000a). Wang and Rannala (2008) have developed a Bayesian framework via reversible jump MCMC. In these algorithms, population genetic parameters such as recombination rates are of interest, and the ARG is regarded as a nuisance parameter. Larribe and Lessard (2002) sample the ARG under the CwR model by using recursive relations similar to Griffiths and Marjoram (1996), and use it to infer the positions of disease loci. Many of these papers describe difficulty in conducting an adequate search.

Inference of the ARG under the CwR model is limited to relatively short genome segments, mainly because the sequence of trees across the chromosome is not Markovian. Thus the conditional independencies that greatly facilitate MCMC and hidden Markov model (HMM) computations do not hold. The use of a Markov approximation to the CwR model, such as one of those described below, offers the potential to infer the sequence of local trees across much larger genome segments. For example, Li and Durbin (2011) develop an HMM approach to estimation of coalescence times along the genome to infer human population history. However, while scaling to genome-wide analysis of sequence data, their method is

limited to single pairs of homologous chromosomes. In this paper, we build a Bayesian computational and inference framework: Sequential Markov Ancestral Recombination Tree or SMARTree. SMARTree uses a reversible jump MCMC to estimate from SNP haplotype data the local coalescent trees of a genome segment in a set of homologous chromosomes together with allelic typing error rate, scaled mutation rate and scaled recombination rate.

Consideration of the ancestral recombination events is key to modeling and inference of the local coalescent trees of a set of chromosomes. Recombination events can be classified by the concept of ancestral material. A chromosome segment in the ARG is considered ancestral material if it is inherited (regardless of mutations) by any sampled chromosome, and non-ancestral material otherwise (black and gray boxes in the upper panel of Figure 1, respectively). We can divide recombinations into classes depending on the presence or absence of ancestral material. We denote by R_1 a recombination within a tract of ancestral material, R_2 a recombination occurring in non-ancestral “trapped” material that has ancestral material to both sides, and R_3 a recombination with no ancestral material to one side. Recombination events of type R_1 can be further divided into R_{1-0} if the ancestral material has already reached its most recent common ancestor (MRCA), and R_{1-1} otherwise.

Recombination events of type R_{1-1} have an unequivocal effect on the inheritance of observed sites. In contrast, recombination events of types R_{1-0} and R_3 do not affect the ancestry of the current chromosomes at all. Most genealogy-sampling algorithms exclude type R_3 events and some also exclude R_{1-0} events. This leaves events of type R_2 , which, as illustrated in Figure 1, affect only the dependence of local trees along chromosomes, and may be only weakly detectable. The sequential Markov coalescent (SMC) model (McVean and Cardin, 2005) is an approximation to the CwR that considers only recombination events of type R_{1-1} , resulting in a Markov structure for local trees along chromosomes. Marjoram and Wall (2006) proposed a modification in which the emerging lineage is allowed to coalesce with the old lineage (an invisible transition of local trees), for example, the recombination event at $x = 8$ (loop) in the upper panel of Figure 1. They called this modified SMC the SMC' model. McVean and Cardin (2005) showed that CwR and SMC produce similar distributions of pairwise linkage disequilibrium statistics. Marjoram and Wall (2006) confirmed this result and extended it to SMC'.

As a Markov approximation to the CwR model, SMC might be preferred to SMC' in that it has fewer recombinations with no loss of explanatory power. As a measure of the difference, note that invisible transitions are not generated by SMC. In ARGs simulated from the SMC' prior, we observed 19% invisible transitions in data sets of 10 sequences, and 12% in data sets of 100 sequences (data not shown). However, calculation of the coalescent probabilities under SMC is more difficult as the probability of a recombination depends on the rootward structure of the ARG: the fewer lineages available, the more likely that a proposed recombination would be resolved by coalescing with its sibling in violation of the SMC, and thus would not occur at all. The SMC' model leads to straightforward calculations in which the probability of a recombination at a given point depends only on the recombination rate and not on the distribution of potential coalescence partners. SMARTree therefore uses the SMC' Markov model in implementing its MCMC approach.

We evaluate the performance of SMARTree on simulated data. While our base data set consists of just 10 chromosomes over a 50kb region, we consider also the same 10 chromosomes over 500 kb, and a set of 40 chromosomes over 50kb. We also examine the impact of a recombination and/or mutation hotspot on the ability of SMARTree to recover local trees from SNP data. Although using SMC' rather than the full CwR distribution, among published samplers SMARTree is most similar to LAMARC (Kuhner et al., 2000a). We have therefore compared its performance with that of LAMARC on simulated data.

2 Methods

2.1 The model

Consider data consisting of n SNP haplotypes from $n/2$ diploid individuals. All SNP sites are assumed to be selectively neutral, and thus the genealogical process can be modeled separately from the mutational process. The coalescent genealogy of the n haplotypes is that generated by a diploid Wright-Fisher population with effective population size N .

We construct a state space model for the data, the latent variables being the local trees $T(x)$ along the chromosome. The genealogical process is modeled by SMC' (McVean and Cardin, 2005; Marjoram and Wall, 2006), where the local tree at a nucleotide site depends only on the tree at the previous site, and the lineage newly produced by a recombination is allowed to coalesce with the lineage it split from (an invisible transition). The Markov process of local trees along the chromosome is reversible, with the stationary distribution being the standard coalescent tree (Supplemental Materials S1). Thus the probability distribution of local trees does not depend on the direction in which the chromosome is considered. The transition probabilities of the SMC' process are given in Supplemental Materials S2.

Given the local trees, both the ancestral alleles at the tree roots and mutations on local trees are assumed to be independent across sites. Since our goal is to model typically available SNP data (for example, HapMap (2005)), we use a two-allele model, although more complex models are possible. The root allele has 50% probability of each of the two possible alleles at the SNP site. Each lineage evolves independently, and mutations on a lineage follow a Poisson process with rate $\theta/2$, where the scaled mutation rate $\theta = 4N\mu$ and μ is the mutation rate per generation per base. The tip alleles are then subject to typing error, toggling between the two alleles, with probability ε per sequence per position. This is a two-state version of the typing-error model of Felsenstein (2004).

Since we model SNP data typically available in association studies, not all variants are considered SNPs. We use a minor-allele frequency cutoff m_c : all variable positions in the region under consideration that have a minor allele count (NMA) of at least m_c are treated as SNPs, and other sites are treated as non-SNPs. We set $m_c = 2$ in our simulation studies. The removal of singleton sites by this cutoff insures that most of the retained SNPs are due to mutation rather than typing error.

At each SNP site, the probability of the observed alleles given the local tree is calculated by the peeling algorithm of Felsenstein (1981). Non-SNP sites contribute a factor of $p(\text{NMA} <$

$m_c|T$) given the tree T at that site. For both SNP and non-SNP sites, the two-allele typing-error model is incorporated into the probability of the observed data.

The model parameters consist of the latent variables $T(x)$ at all chromosomal locations x , allelic typing error probability ε , scaled mutation rate θ , and scaled recombination rate ρ . We assign a uniform prior distribution on the range (0, 0.1) for ε , and uniform prior distributions on the range (0, 0.01) for θ and ρ . The mathematical formulation for the conditional probability $p(NMA < m_c|T)$ is given in Supplemental Materials S3, and the full posterior distribution of the model parameters is given in Supplemental Materials S4. We approximate the conditional probabilities by assuming that at any one site, no more than 1 mutation and 3 typing errors are present.

To allow comparison with the CwR genealogy sampler LAMARC (Kuhner et al., 2000a) we define M1 as the model described above, and M1* as a restricted version. In LAMARC, ε is fixed and θ and ρ are re-parameterized as θ and ρ/θ . Thus in both model M1* and LAMARC, we fix ε to its true value of 0.0025, assign a uniform distribution in the range (0, 0.01) for θ , and a uniform distribution in the range (0, 10) for ρ/θ .

To evaluate modeling of SNP ascertainment, we consider an additional model M2 in which non-SNP sites do not contribute to the likelihood at all, and the likelihood at a SNP site is conditioned on ascertainment. That is, the likelihood at a SNP site is the probability of the data on the local tree conditioned on inclusion of only sites with $NMA < m_c$ (Kuhner et al., 2000b). This represents a case in which the number of intervening sites is not known.

2.2 Parameter estimation

We estimate model parameters $T(x)$, ε , θ , and ρ by a Gibbs sampler. The full conditional distributions required for sampling can be derived from the full posterior distribution (Supplemental Materials S4) by keeping only terms involving focal variables (Gilks et al., 1996).

The local trees $T(x)$ along chromosomes are represented by transition locations $\mathbf{x} = \{x_i\}_{i=1}^K$ and local trees $\mathbf{T} = \{T_i\}_{i=1}^K$, where the initial site is set as $x_1 = 1$, $T_i = T(x_i)$, and tree transitions occur between sites $x_i - 1$ and x_i for $i \geq 2$. Parameters K , $\mathbf{x}$, and $\mathbf{T}$ are sampled by reversible jump MCMC (Green 1995) since the number K of tree transitions is not fixed. Implementation details including sampling of ε , θ , and ρ are provided in Supplemental Material S4; here we focus on sampling of K , $\mathbf{x}$, and $\mathbf{T}$.

2.3 Tree Proposals

For any two trees T and T' , we define the *distance* $|T - T'|$ to be the minimum number of tree transitions needed to transform one into the other. This will be important in defining allowable transitions in our algorithm. In the following section, all measures of distance, transition steps, and permitted tree sequences are according to those of the SMC' process.

We first introduce proposal distributions for modifying local trees. Throughout, the superscript * will indicate a new or modified local tree. The distribution $p(T^*|T)$ is the

standard proposal distribution of SMC' as described in Marjoram and Wall (2006) in which the location of a recombination is chosen randomly among all points in local tree T , and the new branch produced by the recombination is then allowed to coalesce according to the non-recombining coalescent distribution to produce local tree T^* . We define a closely related distribution $q(T^*|T)$, called the “basic move”, which chooses the location of the recombination randomly among all branches rather than among all points. This improves the search among tree topologies as $p(T^*|T)$ seldom attempts changes of short branches and thus short-branch portions of the tree change sluggishly.

We define two constrained variants of the basic move which reduce the magnitude of the proposed change by choosing new local trees that are closely related to previous ones. The first variant is $q(T_y^*|T_x, T_y)$ which we call the “same-branch move.” We examine the transition between adjacent local trees T_x and T_y and make a list of all branches at which a recombination in T_x could have generated T_y . We then make the new local tree T_y^* by choosing one of those branches in T_x at random and applying a recombination to it. T_y^* is thus no more than one step from T_y ; without the branch constraint it could be up to two steps away.

The second variant is $q(T^*|T_p, T_x, T_y)$ which we call the “propagation move.” It is used whenever the insertion, deletion or modification of a local tree creates a potential violation of SMC' in subsequent local trees. For example, if we modify a local tree it may have an distance of 2 to its successor, which is not allowed. We will use the propagation move to modify the successor, repeating this process for each successor in turn until a legal sequence results.

The identities of T_p , T_x and T_y depend on whether the propagation move is being used in a modification, insertion or deletion scenario, as shown in Figure 2. We will describe it in general terms here and then note its specific use when each move algorithm is discussed. The algorithm considers two pairs of trees. T_p is a predecessor local tree to the T^* which we are generating. T_x and T_y are two consecutive local trees which establish the “direction” of the desired transition between T_p and T^* ; we will choose T^* to make the T_p to T^* transition as similar as possible to the T_x to T_y transition. Finally, T_p must be a legal predecessor to T_x , so the distance between them must be 1 (if it were 0, the propagation move would not be needed). Figure 2 shows the identities of T_p , T^* , T_x and T_y for different uses of the propagation move.

This move is designed to minimize the differences between the old and new ARGs in order to increase the chance of acceptance; specifically, by making the new transitions resemble the old ones, we increase the chance that propagation will stop before the end of the chromosome and thus reduce the number of redrawn local trees.

We consider three cases in order.

1. If T_x to T_y is an invisible transition, we apply an invisible transition between T_p and T^* (that is, $T^* = T_p$) and continue propagation.
2. If the distance between T_p and T_y is 1, we set $T^* = T_y$ and stop propagation.

3. If the distance between T_p and T_y is 2, we generate T^* so as to satisfy two constraints: the coalescence time chosen for the recombination separating T_p and T^* is the same time as was chosen between T_x and T_y , and the distance between T_x and T^* is 2. (The latter constraint is needed for reversibility.) We then continue propagation. Legal transitions under this move were developed by exhaustive enumeration and are given in Supplemental Materials S5.

This change is propagated along the chromosome until either case 2 occurs or the end of the chromosome is reached. In all cases, as the propagation move is continued along the chromosome T_p will be the previously changed T^* , T_x will be the old local tree which T_p replaced, and T_y will be the successor of T_x . The identities of T_p , T_x and T_y vary among different move types only on the first application of the propagation move.

Sampling $T(x)$ via MCMC—Seven move types are used in the MCMC sampler: (1) update tree transition locations, (2) update coalescent times, (3) update transition trees, (4) insert a tree transition, (5) delete a tree transition, (6) update the scale and origin of coalescent times in T , and (7) update the tree topology of T .

1. Update transition locations. A single update changes up to z consecutive transition locations, where z is adjusted during the adaptive phase to give an acceptance ratio of approximately 0.234 (Gelman et al., 2004). A random location is chosen and updates continue from that location until z locations have been updated or the end of the chromosome is reached. This defines a range of affected site positions, flanked by unchanging transition locations or the end of the chromosome. New locations for these z transitions are sampled uniformly and independently from this range (subject to being distinct) and then sorted to preserve transition order.
2. Update a coalescence time. Choose a random local tree and choose a random coalescence time t from it. Update t in the chosen local tree and in all other local trees containing it. The new coalescence time t^* is sampled uniformly between bounds set by the closest coalescences in any of the affected local trees, unless t is the time of the root coalescence in all local trees containing it. In that case, t^* is sampled from an exponential distribution with rate 1 and the new root time is obtained by adding the sampled value to the next-largest time-depth in any of the affected local trees. In cases where this root time is chosen there is a Hastings term equal to $\exp(t^* - t)$.
3. Update a local tree and its subsequent trees. Choose a local tree at random (call it T_A). Resample it using the same-branch move, unless it is the first tree of the chromosome, in which case the basic move is used; this generates T_A^* . Next, resample subsequent local trees using the propagation move until that process terminates. The propagation move $q(T^*|T_p, T_x, T_y)$ is used in the form $q(T_B^*|T_A^*, T_A, T_B)$.
4. Insert a local tree. Choose an interval, and sample a new transition site uniformly among sites within that interval. Call the local trees before this insertion T_A, T_C, T_D etc. Draw a new T_B^* using the basic move $q(T_B^*|T_A)$. Then sample the next local tree

from the propagation move $q(T_C^*|T_B^*, T_A, T_C)$ and continue propagation until the move terminates.

5. Delete a local tree. Choose an interval other than the first one and delete its transition (so that the sites governed by its local tree are now governed by the previous one). Call the local trees before this insertion T_A, T_B, T_C, T_D with T_B being the one to be deleted. Sample T_C^* from $q(T_C^*|T_A, T_B, T_C)$ and continue propagation until the move terminates.
6. Update the time scale of $\mathbf{T}$. This is done randomly by one of two methods. In the first, the overall scale of $\mathbf{T}$ is modified by multiplying all the coalescent times by e^z . This scale transformation is a generalized Gibbs sampler, and the target distribution is obtained by replacing $\mathbf{T}$ with $e^z\mathbf{T}$ in the full conditional posterior distribution of $\mathbf{T}$ and multiplying by the Jacobian factor e^{zd} (Liu and Sabatti, 2000), where d is the number of different coalescent times in $\mathbf{T}$. In the second, $\mathbf{T}$ is modified by adding z to all of the coalescence times. The Jacobian factor of this translation transformation is 1. This second method has the effect of increasing the tip branch lengths while leaving the others unchanged, and is useful to allow effective searching of the error rate ε , which mainly affects the length of the tip branches.
7. Interchange two tree leaves. Choose two leaves at random and examine each local tree to see whether they are nearest neighbors. Independently for each run of local trees in which they are not nearest neighbors, choose randomly whether or not to swap their labels. Since chromosomes are interchangeable in SMC', the acceptance ratio for this proposal is 1 if data are not accounted for. This proposal improves mixing of the tree topologies.

In each iteration of the MCMC, moves of type 1–5 are performed a number of times equal to the length of the chromosome divided by 5 kbp, moves of type 6 are performed once, and moves of type 7 are performed $\frac{n}{2}$ times. Move types (1–5) are used with probabilities $(1-2c)/3$, $(1-2c)/3$, $(1-2c)/3$, c , and c , respectively, where c is a tunable constant (all simulations in this paper used $c=0.35$). To improve mixing of the MCMC, at each iteration the direction of analysis along the chromosome is reversed with probability 0.5.

2.4 Model evaluation

Model performance was evaluated by a simulation study using parameter values appropriate for human genomic data, with effective population size $N = 10^4$, recombination rate $r = 10^{-8}$ per base per generation (so that the scaled recombination rate $\rho = 4Nr = 0.4 \times 10^{-3}$), mutation rate $\mu = 2.5 \times 10^{-8}$ per base per generation (so that $\theta = 4N\mu = 0.001$), and allelic typing error rate $\varepsilon = 0.001$ per base per observation. Local trees for 40 chromosomes over 500 kbp were simulated under the full CwR model using the algorithm of Hudson (1983). For convenience, sequences were simulated under a 2-allele infinite-sites model with sequencing error (rather than the 2-allele finite-sites model assumed by our analysis).

We denote by *Standard* the data set for the first 10 chromosomes over the initial 50 kbp, *Extra-Seqs* the data set for all 40 chromosomes over the initial 50 kbp, and *Extra-BP* the data set for the first 10 chromosomes over the entire 500 kbp. To study the effects of

recombination hotspots, we created a new data set from *Standard* by compressing the middle 15 kbp region by a factor of 10, thus reducing the total length to 36.5 kbp. The inter-site distances in this region were divided by 10 and rounded to the nearest positive integer. In four cases, two SNPs mapped to the same location; the rightward SNP of each such pair was dropped. This data set represents a hotspot for both mutation and recombination (as 15 kbp of mutations have been compressed into 1.5 kbp of sequence) and is denoted as *Rec/Mut-Hotspot*. The possibility of a recombinational hotspot without enhanced mutation was modeled by starting with *Rec/Mut-Hotspot* and randomly deleting 80 of the 90 SNP sites within the compressed region; this is denoted *Rec-Hotspot*. These data sets are summarized in Table 1. It should be noted that these five data sets are highly dependent, sharing the underlying ARG as well as the same mutations on that ARG. The simulated ARG by chance had unusually long local trees within the first 50 kbp, which is reflected in the results.

To explore performance of the methods on a wider variety of underlying ARGs, we simulated 4 additional data sets using the same population parameters as *Standard*, each with an independent underlying ARG. These data sets are denoted *Std1* through *Std4*.

For SMARTree, initial values of θ and ε were drawn from their priors and the initial value of ρ was set to 0.0004. The initial ARG was generated from the SMC' prior using these parameter values. Each SMARTree run consisted of two independent groups of 16 MCMC chains each. Within each group, the chains were at varying temperature, with the coldest chain having the target probability distribution, and were run as a Metropolis coupled MCMC (Geyer, 1991). The temperature differences among the coupled chains were adaptively changed during the initial burn-in phase to achieve a 0.5 acceptance ratio of chain swaps. After burn-in, 1000 realizations were sampled from the coldest chain, at a spacing shown in Table 2, to give a total of 2000 samples for each data set. Convergence was assessed from traces of the runs. In each run, potential scale reduction factors (PSRF) (Gelman et al., 2004) between the two coldest chains were used to guide the length of burn-in and the sample spacing (Table 2). We computed PSRF for both the natural log of the posterior probability and the natural log of total transitions. Runs were examined periodically and stopped if both PSRFs were < 1.05 .

LAMARC version pre-2.1.7 was run as a single Bayesian analysis of 6×10^7 iterations with the first 2×10^7 discarded as burn-in; the remaining iterations were sampled every 2×10^4 steps yielding 2000 samples. The initial value of θ was 0.00035 and of ρ/θ was 1.0. The initial ARG was generated from the CwR prior using these starting values of the parameters. As LAMARC cannot infer ε it was held constant at the true value. Convergence of LAMARC runs was assessed by visual examination of run traces.

To compare estimated local trees we used the symmetrical distance measure (Robinson and Foulds, 1981) modified to compare rooted trees (abbreviated RF in Figures).

3 Results

3.1 Comparison between SMARTree and LAMARC

Using the *Standard* data set, we compared the SMARTree sampler using model M1* with LAMARC (Kuhner et al., 2000a) run in Bayesian mode. Both algorithms are Bayesian MCMC samplers, but they differ in several ways. First, SMARTree uses SMC' to model the genealogical process along chromosomes and thus ignores recombination events of type R_2 , whereas LAMARC uses the exact CwR model. Second, there are differences in modeling SNP ascertainment. The SMARTree algorithm assumed that inter-SNP sites had no more than one copy of the minor allele, whereas LAMARC assumed that they had none.

The two methods give almost the same posterior estimates for θ and ρ (Supplemental Figure S1). Both methods inferred similar numbers of tree transitions (Figure 3a) but LAMARC inferred somewhat more invisible transitions as LAMARC permits both type R_{1-1} and type R_2 transitions whereas SMARTree permits only type R_{1-1} . As shown in Figure 3b, LAMARC's extra transitions occurred mainly in the center of the sequence, where events of type R_2 (which require ancestral material on both sides) are most probable. Both programs concentrated their inferred visible transitions in the general area of the true visible transitions.

Figures 3c (LAMARC) and 3d (SMARTree) show the total branch length of the inferred local trees. As shown by the black contour lines, the simulated tree structure had much longer local trees near the center than elsewhere. Both programs underestimated total branch length in this region to a similar degree.

Figures 3e (LAMARC) and 3f (SMARTree) show RF distances between the inferred and true local trees. Mean RF distances were in the range 2–6, indicating 1–3 errors in the recovered topology. Interval tree reconstruction was most successful in areas of high SNP density. Again, both programs performed similarly, although LAMARC showed somewhat tighter confidence intervals. Results from the two programs may differ due to the use of the Markov approximation, differences in modeling of SNP ascertainment, or noise in the MCMC process.

To assess the variability among data sets and thus give more perspective on the variability between the two algorithms, we ran 4 additional data sets drawn from the same parameter values as *Standard* but with independent ARGs. Results are summarized in Table 3 and given in full in Supplemental Table S1. Results from SMARTree and LAMARC were similar for all data sets; most of the variability was between data sets rather than between algorithms. In Table 3, the Mean entry indicates the average point estimate and the Min and Max entries are the lowest and highest point estimates among the five data sets. For both SMARTree and LAMARC, the median of the sampled values was used as the point estimate.

As a further comparison between the algorithms, we ran SMARTree and LAMARC on the *Extra-Seqs* data set. The ARG search space of this 40-sequence data set is much larger,

which could reveal differences in search effectiveness. These results are shown in Table 4. Results were again very similar.

3.2 Evaluation of SNP ascertainment models

The results of analyzing the *Standard* data set by models M1 and M2 are shown in Figure 4 and Supplemental Figure S2. Model M1 assumes that all SNPs with $NMA = m_c$ were ascertained and that the number of non-ascertained sites was known. Model M2 did not use the number of non-ascertained sites.

Estimates of θ (Figure 4a) are expected to be strongly affected by ascertainment. Model M1 overestimated θ somewhat due to the unusually long local trees in the center of the sequence. Model M2, in contrast, found almost no information about θ (the posterior distribution is very similar to the prior) as also found by Kuhner et al. (2000b). Both models produced similar and reasonable estimates of ε (Figure 4b) and of ρ (Figure 4c).

Model M1 was able to derive information about the local tree branch lengths, as seen in Supplemental Figure S2, whereas model M2 mainly reproduced its prior. Topological recovery of the local trees was similar between the two models (Figures 4d–4e).

3.3 Effects of recombination hotspots

The results from analyzing data sets *Rec-Hotspot* and *Rec/Mut-Hotspot* under model M1 are shown in Figure 5 and Supplemental Figure S3. In each data set there is one recombination hotspot located from 17.5 to 19 kbp. Data set *Rec/Mut-Hotspot* retains almost all of the SNPs and thus corresponds to a hotspot for both recombination and mutation, whereas *Rec-Hotspot* is a hotspot for recombination only. The posterior estimates of ε obtained from *Standard*, *Rec-Hotspot* and *Rec/Mut-Hotspot* are similar (results not shown), since the error process is assumed to be independent of the mutation and recombination processes.

As expected, compression without SNP removal (*Rec/Mut-Hotspot*) increased the estimated θ , whereas compression with SNP removal reduced it (Figure 5a) due to the decrease in overall SNP density. Relative SNP densities were 1, 0.48, and 1.33 per kbp for *Standard*, *Rec-Hotspot*, and *Rec/Mut-Hotspot* respectively. The *Rec/Mut-Hotspot* data set captures the long interval trees in the hotspot, whereas the *Rec-Hotspot* data set does not have enough variable sites to do so (Figure 5c and Supplemental Figure S3).

When more SNPs were sampled in the hot spot, more tree transitions were detected in that region. This is shown in Figure 5b as a higher estimate of ρ from *Rec/Mut-Hotspot*, and in Figure 5d as a higher peak in the frequency distribution of tree transition locations. While *Rec/Mut-Hotspot* supported increased recombination frequency in the hotspot, this may be due partly to the long local trees found in this region, as similar peaks appeared in the results from *Standard* in the absence of a hotspot (Figure 3b).

In the region of the hotspot, local trees were better estimated based on *Rec/Mut-Hotspot* than based on *Rec-Hotspot* because more SNPs were available for inference (Figure 5e–f). Outside the hotspot region, local trees were more poorly estimated. This may be due to the over-estimation of θ from the *Rec/Mut-Hotspot* data set.

3.4 Number of sites and number of sequences

The three data sets *Standard*, *Extra-Seqs* and *Extra-BP* were analyzed by model M1, and results are shown in Figure 6 and Supplemental Figures S4 and S5.

Adding more sequences improved recovery of the tree topology (Figure 6a), even though with more branches the range of possible topologies was vastly larger. Adding more sites did not consistently improve recovery (Figure 6b), even at the left-hand end of the sequence where the additional sites might have contributed some information. Both addition of sequences and addition of sites produced improvement in recovery of local tree lengths in the first 50 kbp (Figure 6c). This is somewhat unexpected as increased sequence length should not directly affect the interpretation of the first 50 kbp; the improvement is presumably driven by better parameter estimation. Median total branch length tracked the truth fairly closely across the full length of *Extra-BP* as shown in Figure 6d.

A 10-fold increase in number of sites was more effective in tightening the estimate of θ than a 4-fold increase in number of sequences (Supplemental Figure S4A). Increasing either sequence number or length is expected to produce only modest improvements in inference of θ (Felsenstein, 2005). Increasing either sequences or sites tightened inference of ρ considerably (Supplemental Figure S4B). Increasing sequence number was the most effective at tightening estimates of ε (Supplemental Figure S4C).

Supplemental Figure S5 compares tree and breakpoint inference from four non-overlapping 10-sequence subsets of *Extra-Seqs*. Considerable variation among the results is seen, suggesting that 10 sequences, while adequate for parameter inference, do not provide a clear overall view of tree length or topology.

4 Discussion

We present a novel MCMC algorithm for rearranging genealogies under the SMC' approximation to the full CwR model. The use of this Markov approximation provides conditional independence structures that permit new MCMC proposals, and potentially allows for the analysis of much longer genome regions. We present results for sequences up to length 500 Kbp but considerably longer sequences are feasible on a desktop computer.

The SMARTree sampler performed comparably to the full CwR sampler LAMARC in our limited suite of tests, indicating that little is lost by the SMC' simplification. Unfortunately the speed performance of the algorithms cannot directly be compared as SMARTree was implemented in Mathematica and LAMARC in C++; the programs also differ significantly in the use of numerical optimizations. Given comparable implementations, however, it seems likely that the reduced complexity and search space of SMC' should result in more rapid searches. It is noteworthy that the Mathematica implementation of SMARTree, with relatively little optimization, ran substantially faster than LAMARC on most of our data sets. (This was particularly notable for *Std-4*, which took 25 days with LAMARC and under 5 days with SMARTree.) Furthermore the 500 Kbp data set *Extra-BP* was too large to be feasible for LAMARC, whereas it was straightforwardly analyzed by SMARTree.

The comparisons presented here suggest that using SMC' rather than CwR has little or no downside with regard to estimation of local trees. Using an accurate SNP model is important to correct inference of local tree branch lengths, but topologies can be fairly well reconstructed even with an oversimplified model.

Our primary goal in estimating of local tree topologies and relative branch lengths is to infer the structure of coancestry among a set of chromosomes. A key finding is that, despite the greatly increased number of topologies, more sequences can provide more structure information due to the information provided by analysis of the joint patterns of descent. The extension of coancestry inference from pairwise inference between individuals (Browning and Browning, 2010; Brown et al, 2012) to larger sets of chromosomes has the potential to enhance subsequent association and fine-scale linkage analyses. This has already been well recognized in coalescent-based linkage disequilibrium mapping (Zollner and Pritchard, 2005; Smith and Kuhner, 2009), but here we bring this same approach more broadly to human population-based coancestry studies.

Our approach differs from population-genetic approaches in that we do not make use of current population allele and local haplotype frequencies in inferring coancestry. Instead we adopt the approach of the coalescent literature in assigning a uniform prior over the two possible SNP alleles at the root. Thus our approach could be used with rarer alleles, or in situations where good current frequency information is not available. While recent coancestry may be validly inferred using information from current common SNP variation (Brown et al, 2012), for inference of patterns of deep coancestry use of current allele frequencies is questionable.

Our focus is on SNP data widely available in human populations studies, and thus SMARTree has been implemented for SNP data. However, in contrast to the samplers of Griffiths and Marjoram (1996), Fearnhead and Donnelly (2001) and Larribe and Lessard (2002), SMARTree does not include mutation events explicitly on the tree, and thus can readily accommodate a more complex mutational model. In particular, despite its current implementation for SNP data, the SMARTree MCMC sampler is applicable to any form of molecular data for which a mutational model is available.

Despite the focus on SNP data it is important to include information on the number of non-SNP sites. Model M2, which does not use this information, has almost no power to estimate θ (as also noted by Kuhner et al. (2000b)). We also tested a model in which SNPs were analyzed as if they were full DNA sequences; unsurprisingly, the bias in estimation of θ was extreme, and local tree recovery was also worsened (data not shown). Accurate modeling of SNP ascertainment is essential to obtaining accurate parameter estimates from genealogy samplers.

SMARTree is freely available as Mathematica code from the web site <http://www.stat.washington.edu/thompson/Genepi/pangaea.shtml>. This site also contains all input files needed to recreate the simulations used in this paper.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported in part by National Institutes of Health grants GM046255 (E.A.T. and C.Z.) and GM099568 (E.A.T. and M.K.K.). We thank Serge Sverdlov and Jon Yamato for editing assistance.

References

- Brown MD, Glazner CG, Zheng C, Thompson EA. Inferring coancestry in population samples in the presence of linkage disequilibrium. *Genetics*. 2012; 190:1447–1460. [PubMed: 22298700]
- Browning SR, Browning B. High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet*. 2010; 86:526–539. [PubMed: 20303063]
- Fearnhead P, Donnelly P. Estimating recombination rates from population genetic data. *Genetics*. 2001; 159:1299–1318. [PubMed: 11729171]
- Felsenstein J. Evolutionary trees from DNA-sequences - a maximum-likelihood approach. *Journal of Molecular Evolution*. 1981; 17:368–376. [PubMed: 7288891]
- Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates; Sunderland, MA: 2004.
- Felsenstein J. Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Mol Biol Evol*. 2005; 23:691–700. [PubMed: 16364968]
- Gasbarra D, Pirinen M, Sillanpää MJ, Arjas E. Estimating genealogies from linked marker data: a Bayesian approach. *BMC Bioinformatics*. 2007; 8:411. [PubMed: 17961219]
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian data analysis*. Chapman and Hall/CRC; 2004.
- Geyer, CJ. Markov chain Monte Carlo maximum likelihood. In: Keramidas, EM.; Kaufman, SM., editors. *Proceedings of the 23rd Symposium on the Interface*. Interface Foundation of North America; Fairfax Station, VA: 1991. p. 156-163.
- Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. *Markov Chain Monte Carlo in Practice*. Chapman and Hall; 1996.
- Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 1995; 82:711–732.
- Griffiths RC, Marjoram P. Ancestral inference from samples of DNA sequences with recombination. *J Comp Biol*. 1996; 3:479–502.
- Gusev A, Kenny EE, Lowe JK, Salit J, Saxena R, Kathiresan S, Altshuler DM, Friedman JM, Breslow JL, Pe'er I. DASH: A Method for Identical-by-Descent Haplotype Mapping Uncovers Association with Recent Variation. *Am J Hum Genet*. 2011; 88:706–717. [PubMed: 21620352]
- International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005; 237:1299–1319.
- Hein J. A heuristic method to reconstruct the history of sequences subject to recombination. *J Mol Evol*. 1993; 36:396–405.
- Hein, J.; Schierup, MH.; Wiuf, C. *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press; Oxford: 2005.
- Hössjer O, Hartman L, Humphreys K. Ancestral Recombination Graphs under Non-Random Ascertainment, with Applications to Gene Mapping. *Statistical Applications in Genetics and Molecular Biology*. 2009; 8:35.
- Hudson RR. Properties of the neutral allele model with intergenic recombination. *Theor Pop Biol*. 1983; 23:183–201. [PubMed: 6612631]
- Kingman JFC. The coalescent. *Stochastic Processes and Their Applications*. 1982; 13:235–248.
- Kuhner MK, Yamato J, Felsenstein J. Maximum likelihood estimation of recombination rates from population data. *Genetics*. 2000a; 156:1393–1401. [PubMed: 11063710]

- Kuhner MK, Beerli P, Yamato J, Felsenstein J. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*. 2000b; 156:439–447. [PubMed: 10978306]
- Larribe F, Lessard S. Gene mapping via the ancestral recombination graph. *Theor Pop Biol*. 2002; 62:215–229. [PubMed: 12167358]
- Li H, Durbin R. Inference of human population history from individual whole-genome sequences. *Nature*. 2011; 475:493–496. [PubMed: 21753753]
- Liu JS, Sabatti C. Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika*. 2000; 87:353–369.
- Marjoram P, Wall JD. Fast “coalescent” simulation. *BMC Genetics*. 2006; 7:16. [PubMed: 16539698]
- McVean GAT, Cardin NJ. Approximating the coalescent with recombination. *Phil Trans Royal Society B-Biol Sci*. 2005; 360:1387–1393.
- Minichiello MJ, Durbin R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet*. 2006; 79:910–922. [PubMed: 17033967]
- Morris AP, Whittaker JC, Balding DJ. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. *Am J Hum Genet*. 2002; 70:686–707. [PubMed: 11836651]
- Robinson DR, Foulds LR. Comparison of phylogenetic trees. *Math Biosci*. 1981; 53:131–147.
- Smith LP, Kuhner MK. The limits of fine-scale mapping. *Genet Epidemiol*. 2009; 33:344–356. [PubMed: 19048633]
- Thompson EA. Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*. 2013; 194:301–306. [PubMed: 23733848]
- Wang Y, Rannala B. Bayesian inference of fine-scale recombination rates using population genomic data. *Phil Trans Royal Society B-Biol Sci*. 2008; 363:3921–3930.
- Wu YF. New Methods for Inference of Local Tree Topologies with Recombinant SNP Sequences in Populations. *IEEE-ACM Trans Comp Biol and Bioinf*. 2011; 8:182–193.
- Zollner S, Pritchard JK. Coalescent-based association mapping and fine mapping of complex trait loci. *Genetics*. 2005; 169:1071–1092. [PubMed: 15489534]

Fig. 1.

The genealogy of three chromosomes of length 8 bp. The upper panel shows the ARG representation and the lower panel shows the derived local trees, with a schematic showing the transition between local trees T_3 and T_4 . In the upper panel, black boxes denote sampled chromosomes or material ancestral to them, and gray boxes denote non-ancestral material. The pair (R, x) indicates a recombination event of type R occurring between sites $x - 1$ and x . The lower panel shows the derived local trees $T(x)$ and the transition from T_3 to T_4 . The black dot indicates the lineage location of recombination event $(R_1, 3)$, which is chosen uniformly on tree T_3 . The emerging lineage (dotted line) coalesces at a rate based on the number of lineages yielding a new coalescence time t_c at node D in tree T_4 , changing the tree topology from $\{\{1, 2\}, 3\}$ into $\{1, \{2, 3\}\}$.

Modification or continuation: *Deletion:* *Insertion:* **Fig. 2.**

The propagation move. Letters A, B, C and D indicate local trees in the ARG, and superscript * indicates a new (changed) local tree. The identities of T_p , T^* , T_x and T_y are indicated for each type of propagation move.

Fig. 3.

Local tree recovery in SMARTree (black) and LAMARC (gray). **(a)** Inferred numbers of total transitions (heavy lines) and invisible transitions (fine lines). The true ARG had 64 total transitions of which 45 were invisible. **(b)** Inferred visible transition locations. The dashed line indicates the uniform prior, and the vertical bars at the top show the true visible transition locations. **(c, d)** Total branch lengths of the inferred local trees for LAMARC **(c)** and SMARTree **(d)**. Black dots show posterior medians, the stepped black line indicates the true local tree length, vertical lines show 95% support intervals, and gray dashed lines show quantiles 0.025, 0.5, and 0.975 of the random distribution of total branch length under the Kingman coalescent. **(e, f)** Topological similarity between inferred and true local trees for LAMARC **(e)** and SMARTree **(f)**. Symbols as in parts **(c, d)** with the quantiles representing the distribution of RF distances between the true local tree and trees randomly drawn from the Kingman coalescent. (This distribution varies across the sequence as some true local trees are more typical of the coalescent than others. The 0.5 and 0.975 quantiles lie on top of each other at 16 as random genealogies are generally at maximum RF distance.)

Fig. 4. Comparison of models M1 (both SNPs and non-SNPs used) and M2 (only SNPs used). In panels **a–c** black lines indicate model M1 and gray lines indicate model M2; the black dot shows the true simulation value. Panel **a** shows inference of θ , panel **b** shows inference of ε , and panel **c** shows inference of ρ . Panels **d** (model M1) and **e** (model M2) show RF distances between inferred and true local trees, with black dots indicating the posterior median, vertical bars indicating 95% support intervals, and dashed lines showing quantiles as in Figure 3.

Fig. 5.

Inference from data containing a hotspot. Data set *Rec/Mut-Hotspot* is shown in thick gray, data set *Rec-Hotspot* in thin gray, and the non-hotspot data set *Standard* in black. Panels **a** and **b** show inference of θ and ρ , respectively; the horizontal dashed line indicates the uniform prior and black dots indicate the simulation values. Panel **c** shows inferred total length of the local tree; the thin black line indicates true local tree length, and the dashed line indicates the expectation based on the Kingman prior. Panel **d** shows inferred transition locations; black vertical bars above the graph indicate the true transition locations, and the dashed line indicates the Kingman prior. Panels **e** and **f** show RF distances between inferred and true local trees; black dots indicate posterior medians, vertical lines show 95% support intervals, and gray dashed lines indicate quantiles of the expected difference between the true local tree and the expected distribution under the Kingman prior (the 0.5 and 0.975 quantiles lie on top of each other at the maximum). Panel **e** shows *Rec-Hotspot* while Panel **f** shows *Rec/Mut-Hotspot*.

Fig. 6.

Effect of amount of data on SMARTree inference. *Standard* thick black lines, *Extra-Seqs* thin gray lines, *Extra-BP* thick gray lines. Panel **a** shows median RF distance between inferred and true tree for data sets *Standard* and *Extra-Seqs*, and panel **b** shows the same comparison for *Standard* and the first 50 bp of *Extra-BP*. Panel **c** shows median total branch length for the first 50 kb, with the stepped black line showing the true value and the dashed line indicating the uniform prior. Panel **d** compares median total branch length across the entire 500 kb between *Extra-BP* (thick gray line) and the truth (thin black line); the horizontal dashed line indicates the expectation under the Kingman prior.

Table 1

Simulated data sets

Data set	Sequences	Length (kb)	Hotspot type	# SNPs
Standard	10	50	None	130
Extra-Seqs	40	50	None	220
Extra-BP	10	500	None	887
Rec-Hotspot	10	36.5	Recombination only	46
Rec/Mut-Hotspot	10	36.5	Recombination and mutation	126
Std-1, 2, 3, 4	10	50	None	100, 115, 63, 87

Table 2

Run conditions for SMARTree analyses

Data Set	Burn-In	Sample Spacing	Potential Scale Reduction Factor	
			ln(post. prob.)	ln(tot. transitions)
Standard/M1	5568	2	1.000	1.002
Standard/M1*	11936	4	1.002	1.000
Standard/M2	13260	2	1.006	1.010
Std-1/M1*	4492	4	1.000	1.001
Std-2/M1*	5068	4	1.001	1.000
Std-3/M1*	7432	4	1.020	1.030
Std-4/M1*	10682	8	1.001	1.001
Extra-BP	6000	2	1.000	1.000
Extra-Seq/M1	4000	2	1.003	1.001
Extra-Seq/M1*	8532	4	1.001	1.000
RecHotspot	10072	4	1.000	1.009
Rec/MutHotspot	6992	4	1.031	1.004

Table 3

SMARTree and LAMARC comparisons on 5 replicate data sets

Parameter	Truth	SMARTree			LAMARC		
		Mean	Min	Max	Mean	Min	Max
$\theta \times 10^3$	1.0	1.054	0.89	1.38	0.966	0.74	1.34
$\rho \times 10^3$	0.4	0.572	0.26	1.07	0.676	0.23	1.21
Visible Transitions	41.8	58.6	19	118	56.8	16	115
Mean RF Distance	–	4.8	4	6	4.4	2	6

Table 4

SMARTree and LAMARC comparisons on *Extra-Seq*

Parameter	Truth	SMARTree		LAMARC	
		Mean	95% Bounds	Mean	95% Bounds
$\theta \times 10^3$	1.0	1.2	0.86–1.74	0.9	0.63–1.22
$\rho \times 10^3$	0.4	0.23	0.12–0.4	0.22	0.12–0.37
Visible Transitions	39	38	26–54	36	25–52
RF Distance	–	36	26–50	32	22–46