DOI: 10.1534/genetics.110.117986

A Principled Approach to Deriving Approximate Conditional Sampling Distributions in Population Genetics Models with Recombination

Joshua S. Paul* and Yun S. Song*,†,[1]

**Computer Science Division and* †*Department of Statistics, University of California, Berkeley, California 94720*

Manuscript received April 20, 2010
Accepted for publication June 24, 2010

ABSTRACT

The multilocus conditional sampling distribution (CSD) describes the probability that an additionally sampled DNA sequence is of a certain type, given that a collection of sequences has already been observed. The CSD has a wide range of applications in both computational biology and population genomics analysis, including phasing genotype data into haplotype data, imputing missing data, estimating recombination rates, inferring local ancestry in admixed populations, and importance sampling of coalescent genealogies. Unfortunately, the true CSD under the coalescent with recombination is not known, so approximations, formulated as hidden Markov models, have been proposed in the past. These approximations have led to a number of useful statistical tools, but it is important to recognize that they were *not derived* from, though were certainly motivated by, principles underlying the coalescent process. The goal of this article is to develop a principled approach to *derive* improved CSDs directly from the underlying population genetics model. Our approach is based on the diffusion process approximation and the resulting mathematical expressions admit intuitive genealogical interpretations, which we utilize to introduce further approximations and make our method scalable in the number of loci. The general algorithm presented here applies to an arbitrary number of loci and an arbitrary finite-alleles recurrent mutation model. Empirical results are provided to demonstrate that our new CSDs are in general substantially more accurate than previously proposed approximations.

THE probability of observing a sample of DNA sequences under a given population genetics model—which is referred to as the sampling probability or likelihood—plays an important role in a wide range of problems in a genetic variation study. When recombination is involved, however, obtaining an analytic formula for the sampling probability has hitherto remained a challenging open problem (see Jenkins and Song 2009, 2010 for recent progress on this problem). As such, much research (Griffiths and Marjoram 1996; Kuhner *et al.* 2000; Nielsen 2000; Stephens and Donnelly 2000; Fearnhead and Donnelly 2001; De Iorio and Griffiths 2004a,b; Fearnhead and Smith 2005; Griffiths *et al.* 2008; Wang and Rannala 2008) has focused on developing Monte Carlo methods on the basis of the coalescent with recombination (Griffiths 1981; Kingman 1982a,b; Hudson 1983), a well-established mathematical framework that models the genealogical history of sample chromosomes. These Monte Carlo-based full-likelihood methods mark an important development in population genetics analysis, but a well-known obstacle to their utility is that they tend to be computationally intensive. For a whole-genome variation study, approximations are often unavoidable, and it is therefore important to think of ways to minimize the trade-off between scalability and accuracy.

A popular likelihood-based approximation method that has had a significant impact on population genetics analysis is the following approach introduced by Li and Stephens (2003): Given a set Φ of model parameters (*e.g.*, mutation rate, recombination rate, etc.), the joint probability $p(h_1, \ldots, h_n \mid \Phi)$ of observing a set $\{h_1, \ldots, h_n\}$ of haplotypes sampled from a population can be decomposed as a product of conditional sampling distributions (CSDs), denoted by π,

$$p(h_1, \ldots, h_n \mid \Phi) = \pi(h_1 \mid \Phi) \times \pi(h_2 \mid h_1, \Phi) \times \ldots \times \pi(h_n \mid h_1, \ldots, h_{n-1}, \Phi), \quad (1)$$

where $\pi(h_{k+1} \mid h_1, \ldots, h_k, \Phi)$ is the probability of an additionally sampled haplotype being of type h_{k+1}, given a set of already observed haplotypes $h_1, \ldots, h_k$. In the presence of recombination, the true CSD π is unknown, so Li and Stephens proposed using an approximate CSD $\hat{\pi}$ in place of π, thus obtaining the following approximation of the joint probability:

$$p(h_1, \ldots, h_n \mid \Phi) \approx \hat{\pi}(h_1 \mid \Phi) \times \hat{\pi}(h_2 \mid h_1, \Phi) \times \ldots \times \hat{\pi}(h_n \mid h_1, \ldots, h_{n-1}, \Phi). \quad (2)$$

Li and Stephens referred to this approximation as the product of approximate conditionals (PAC) model. In

[1]*Corresponding author:* Department of EECS, University of California, 683 Soda Hall No. 1776, Berkeley, CA 94720-1776.
E-mail: yss@cs.berkeley.edu

general, the closer $\hat{\pi}$ is to the true CSD π, the more accurate the PAC model becomes. Notable applications and extensions of this framework include estimating crossover rates (Li and Stephens 2003; Crawford *et al.* 2004) and gene conversion parameters (Gay *et al.* 2007; Yin *et al.* 2009), phasing genotype data into haplotype data (Stephens and Scheet 2005; Scheet and Stephens 2006), imputing missing data to improve power in association mapping (Stephens and Scheet 2005; Li and Abecasis 2006; Marchini *et al.* 2007; Howie *et al.* 2009), inferring local ancestry in admixed populations (Price *et al.* 2009), inferring human colonization history (Hellenthal *et al.* 2008), inferring demography (Davison *et al.* 2009), and so on.

Another problem in which the CSD plays a fundamental role is importance sampling of genealogies under the coalescent process (Stephens and Donnelly 2000; Fearnhead and Donnelly 2001; De Iorio and Griffiths 2004a,b; Fearnhead and Smith 2005; Griffiths *et al.* 2008). In this context, the optimal proposal distribution can be written in terms of the CSD π (Stephens and Donnelly 2000), and as in the PAC model, an approximate CSD $\hat{\pi}$ may be used in place of π. The performance of an importance sampling scheme depends critically on the proposal distribution and therefore on the accuracy of the approximation $\hat{\pi}$. Often in conjunction with composite-likelihood frameworks (Hudson 2001; Fearnhead and Donnelly 2002), importance sampling has been used in estimating fine-scale recombination rates (McVean *et al.* 2004; Fearnhead and Smith 2005; Johnson and Slatkin 2009).

So far, a significant scope of intuition has gone into choosing the approximate CSDs used in these problems (Marjoram and Tavaré 2006). In the case of completely linked loci, Stephens and Donnelly (2000) suggested constructing an approximation $\hat{\pi}_{\mathrm{SD}}(h_{k+1}|h_1, \ldots, h_k, \Phi)$ by assuming that the additional haplotype h_{k+1} is an imperfect copy of one of the first k haplotypes, with copying errors corresponding to mutation. Fearnhead and Donnelly (2001) generalized this construction to include crossover recombination, assuming that the haplotype h_{k+1} is an imperfect mosaic of the first k haplotypes (*i.e.*, h_{k+1} is obtained by copying segments from $h_1, \ldots, h_k$, where crossover recombination can change the haplotype from which copying is performed). The associated CSD, which we denote by $\hat{\pi}_{\mathrm{FD}}$, can be interpreted as a hidden Markov model and so admits an efficient dynamic programming solution. Finally, Li and Stephens (2003) proposed a modification to Fearnhead and Donnelly's model that limits the hidden state space, thereby providing a computational simplification; we denote the corresponding approximate CSD by $\hat{\pi}_{\mathrm{LS}}$.

Although these approaches are computationally appealing, it is important to note that they are *not derived* from, though are certainly motivated by, principles underlying typical population genetics models, in particular the coalescent process (Griffiths 1981; Kingman 1982a,b; Hudson 1983). The main objective of this article is to develop a principled technique to *derive* an improved CSD directly from the underlying population genetics model. Rather than relying on intuition, we base our work on mathematical foundation. The theoretical framework we employ is the diffusion process. De Iorio and Griffiths (2004a,b) first introduced the diffusion-generator approximation technique to obtain an approximate CSD in the case of a single locus (*i.e.*, no recombination). Griffiths *et al.* (2008) later extended the approach to two loci to include crossover recombination, assuming a parent-independent mutation model at each locus. In this article, we extend the framework to develop a general algorithm that applies to an arbitrary number of loci and an arbitrary finite-alleles recurrent mutation model.

Our work can be summarized as follows. Using the diffusion-generator approximation technique, we derive a recursion relation satisfied by an approximate CSD. This recursion can be used to construct a closed system of coupled linear equations, in which the conditional sampling probability of interest appears as one of the unknown variables. The system of equations can be solved using standard numerical analysis techniques. However, the size of the system grows superexponentially with the number of loci and, consequently, so does the running time. To remedy this drawback, we introduce additional approximations to make our approach scalable in the number of loci. Specifically, the recursion admits an intuitive genealogical interpretation, and, on the basis of this interpretation, we propose modifications to the recursion, which then can be easily solved using dynamic programming. The computational complexity of the modified algorithm is polynomial in the number of loci, and, importantly, the resulting CSD has little loss of accuracy compared to that following from the full recursion.

The accuracy of approximate CSDs has not been discussed much in the literature, except in the application-specific context for which they are being employed. In this article, we carry out an empirical study to explicitly test the accuracy of various CSDs and demonstrate that our new CSDs are in general substantially more accurate than previously proposed approximations. We also consider the PAC framework and show that our approximations also produce more accurate PAC-likelihood estimates. We note that for the maximum-likelihood estimation of recombination rates, the actual value of the likelihood may not be so important, as long as it is maximized near the true recombination rate. However, in many other applications—*e.g.*, phasing genotype data into haplotype data, imputing missing data, importance sampling, and so on—the accuracy of the CSD and PAC-likelihood function over a wide range of parameter values may be important. Thus, we believe that the theoretical work presented here will have several practical implications; our method can be applied in a wide

range of statistical tools that use CSDs, improving their accuracy.

The remainder of this article is organized as follows. To provide intuition for the ensuing mathematics, we first describe a genealogical process that gives rise to our CSD. Using our genealogical interpretation, we consider two additional approximations and relate these to previously proposed CSDs. Then, in the following section, we derive our CSD using the diffusion-generator approach and provide mathematical statements for the additional approximations; some interesting limiting behavior is also described there. This section is self-contained and may be skipped by the reader uninterested in mathematical details. Finally, in the subsequent section, we carry out a simulation study to compare the accuracy of various approximate CSDs and demonstrate that ours are generally the most accurate.

A GENEALOGICAL FORMULATION

Before delving into mathematical details, we first describe a genealogical interpretation for our proposed CSD. In addition to providing intuition about the underlying mathematics (which is discussed in detail in the following section), the genealogical interpretation suggests a tractable approximation of our CSD. We discuss how some previously proposed CSDs may also be viewed as approximations of our basic scheme.

Preliminary notation: As our basic stochastic process, we consider a finite-sites, finite-alleles version of the coalescent with recombination. In particular, denote the set of loci by $L = \{1, \ldots, k\}$. The following general notation is used hereafter to describe mutation and recombination events in the coalescent:

Mutation: We use E_ℓ to denote the set of allele types at locus $\ell \in L$. Mutation events at locus ℓ occur with rate $\theta_\ell/2$. Going forward in time, given that there is a mutation, a transition from allele $a \in E_\ell$ to allele $a' \in E_\ell$ occurs with probability $P^{(\ell)}_{a,a'}$. By a parent-independent mutation (PIM) model, we mean a model in which $P^{(\ell)}_{a,a'} = P^{(\ell)}_{a'}$ for all a, a', and ℓ.

Recombination: The set of recombination *breakpoints* is denoted by $B = \{(1, 2), \ldots, (k - 1, k)\}$. Given a breakpoint $b = (\ell, \ell + 1) \in B$, recombination events between loci ℓ and $\ell + 1$ occur with rate $\rho_b/2$.

We use $\mathcal{H} = E_1 \times \ldots \times E_k$ to denote the set of k-locus haplotypes. A sample configuration of haplotypes is specified by a vector $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$, with n_h being the number of haplotypes of type h in the sample. The total number of haplotypes in $\mathbf{n}$ is denoted by $|\mathbf{n}| = \sum_{h \in \mathcal{H}} n_h$. Finally, we use $\mathbf{e}_h$ to denote the singleton configuration with a 1 for haplotype h and 0's elsewhere.

Conditional sampling: Recall that a realization of the coalescent with recombination is a random genealogy comprising a series of events (*i.e.*, mutation, recombination, and coalescence), relating a collection of haplotypes. This genealogy results from a continuous-time Markov process, which moves backward through time and takes collections of haplotypes as states; we refer to a haplotype in the current state as a lineage. An event then corresponds to a jump in the continuous-time Markov process and makes a particular modification to the current state. With the initial state being a set of n *unspecified* haplotypes, the following approach may be used to simulate a random genealogy from the process:

Mutation: Locus $\ell \in L$ of each lineage mutates with rate $\theta_\ell/2$.

Recombination: Each lineage undergoes recombination about breakpoint $b \in B$ with rate $\rho_b/2$.

Coalescence: Each pair of lineages coalesces with rate 1.

When a single most recent common ancestor (MRCA) remains, the process terminates. The types of each lineage in the genealogy are then determined by sampling the MRCA haplotype from the stationary distribution of the mutation process and propagating the information forward along the sampled genealogy; the specifics of each mutation event in the sampled genealogy are stochastically determined by the mutation transition matrix P. We refer to the final genealogical history obtained in this way as an ancestry and denote it by $\mathcal{A}_n$. Observe that associated with a randomly sampled ancestry $\mathcal{A}_n$ is a sample configuration $\mathbf{n}$ with $|\mathbf{n}| = n$ *specified* haplotypes generated at the leaves. See Figure 1a for an illustration.

Suppose we now wish to sample a collection of m additional haplotypes *conditioned* on having already observed a sample $\mathbf{n}$ and the true ancestry $\mathcal{A}_n$ that generated $\mathbf{n}$. The above-mentioned sampling scheme can be modified to sample a *conditional ancestry* $\mathcal{C}_m$ relating a collection of m haplotypes to each other and to the sample $\mathbf{n}$. As illustrated in Figure 1b, the conditional sampling scheme would comprise the usual genealogical events (mutation, recombination, and coalescence) involving the lineages in $\mathcal{C}_m$, along with coalescence events involving a lineage in $\mathcal{C}_m$ and a lineage ancestral to $\mathbf{n}$. We refer to the latter coalescence events as "absorption" events. Note that the ancestral lineages of the sample $\mathbf{n}$ completely determine the type of each lineage in $\mathcal{C}_m$ involved in absorption events, and a valid conditional sample configuration $\mathbf{m}$ with $|\mathbf{m}| = m$ is generated at the leaves of $\mathcal{C}_m$.

There are three sources of complication to the approach just described: (1) The ancestry $\mathcal{A}_n$ associated with a sample $\mathbf{n}$ is usually unknown; (2) although the genealogical process for $\mathcal{C}_m$ is Markov, it is *time inhomogeneous* since the ancestry $\mathcal{A}_n$ is nonconstant in time; and (3) if j lineages in $\mathcal{C}_m$ survive till the time to the MRCA of $\mathcal{A}_n$, then one needs to simulate farther back in time with $j + 1$ lineages, conditioned on one of the lineages being the specified MRCA of $\mathcal{A}_n$. A genealogical approximation, resulting from the diffusion-generator technique described in the subsequent section, avoids

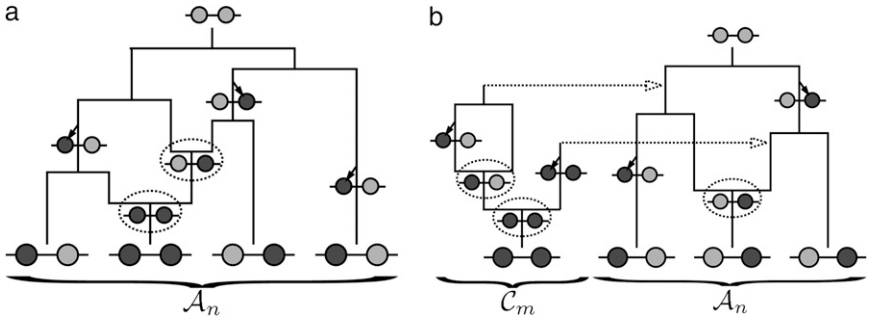

FIGURE 1.—Illustrations of a genealogy and conditional genealogy for a two-locus ($k = 2$), two-allele model. The two loci of a haplotype are each represented by a circle, with the shading (light or dark) indicating the allelic type at that locus. Mutation events, along with the locus and resulting haplotype, are indicated by small arrows. Recombination events (always taking the left loci from the left side and the right locus from the right side), along with the resulting haplotype, are indicated by dotted circles. (a) A genealogy $\mathcal{A}_n$ with $n = 4$. It is easy to verify that, starting with the MRCA and following the genealogy forward in time, the sample configuration n shown at the leaves is obtained. (b) An "observed" genealogy $\mathcal{A}_n$ with $n = 3$ and a conditional genealogy $\mathcal{C}_m$ with $m = 1$. Absorption events are indicated by dotted arrows into $\mathcal{A}_n$. Following the combined genealogy forward in time, it is easy to check that the conditional sample $\mathbf{m}$ shown at the leaf of $\mathcal{C}_m$ is obtained.

all of these difficulties. Assume that $\mathcal{A}_n = \mathcal{A}_n^*$, where $\mathcal{A}_n^*$ is the nonrandom *trunk* ancestry defined as follows: Within $\mathcal{A}_n^*$, the lineages do not mutate, recombine, or coalesce with one another, and instead form a "trunk" extending infinitely into the past. See Figure 2 for an illustration. Note that $\mathcal{A}_n^*$ is an improper ancestry, as there is no MRCA; nonetheless, the above conditional sampling procedure remains well defined. In particular, events within the conditional genealogy $\mathcal{C}_m$ occur at the following rates:

Mutation: Locus $\ell \in L$ of each lineage mutates with rate $\theta_\ell/2$.
Recombination: Each lineage undergoes recombination about breakpoint $b \in B$ with rate $\rho_b/2$.
Coalescence: Each pair of lineages coalesces with rate 1.
Absorption: Each lineage is absorbed into each lineage of $\mathcal{A}_n^*$ with rate 1/2.

Conversely, given a new sample $\mathbf{m}$ and a previously observed sample $\mathbf{n}$, we may wish to compute the *conditional sampling probability* (CSP), denoted $\pi(\mathbf{m} \mid \mathbf{n})$. Although analytic computation of the CSP is impracticable for all but the smallest problems, using our genealogical approximation, namely that $\mathcal{A}_n = \mathcal{A}_n^*$, it is possible to compute an approximate CSP $\hat{\pi}(\mathbf{m} \mid \mathbf{n})$ by decomposing with respect to the unknown conditional genealogy $\mathcal{C}_m$. With $p(\mathcal{C}_m \mid \mathcal{A}_n^*)$ denoting the probability of conditional ancestry $\mathcal{C}_m$, our approximation is

$$\hat{\pi}(\mathbf{m} \mid \mathbf{n}) = \sum_{\mathcal{C}_m} \pi(\mathbf{m} \mid \mathcal{C}_m) p(\mathcal{C}_m \mid \mathcal{A}_n^*),$$

where $\pi(\mathbf{m} \mid \mathcal{C}_m) = 1$ if $\mathbf{m}$ is the configuration of haplotypes generated at the leaves of $\mathcal{C}_m$ and 0 otherwise. Because $\mathcal{A}_n^*$ is invariant in time, $\mathcal{C}_m$ has a time-homogeneous Markov structure, and the above conditioning may be recast as a time-independent recursion. The solution thus obtained is our primary approximation, denoted $\hat{\pi}_{\mathrm{PS}}$. We next examine some computational aspects of $\hat{\pi}_{\mathrm{PS}}$ and consider two genealogical approximations.

Computation and approximation: There is no known general analytic formula for the recursion obtained for $\hat{\pi}_{\mathrm{PS}}$. The procedure for exact computation of $\hat{\pi}_{\mathrm{PS}}$, therefore, is to repeatedly invoke the recursion equation; this yields a closed set of coupled linear equations, which can be solved to provide the desired probability. It is instructive to quantify the size of the linear system that must be generated and solved. Suppose we are interested in the CSP of a single haplotype (*i.e.*, $|\mathbf{m}| = 1$); for simplicity, also assume that $|E_\ell| = s$ for all $\ell \in L$. The number Q_k of equations produced for k loci is

$$Q_k = \sum_{j=1}^{k} \binom{k}{j} B_j s^j,$$

where B_j is the jth Bell number, the number of partitions of a set of cardinality j into nonempty subsets. An algebraic identity involving the Bell numbers implies $Q_k \geq B_{k+1}$ (with equality holding under a PIM model of mutation). Hence, since B_{k+1} is superexponential in k, exact computation of $\hat{\pi}_{\mathrm{PS}}$ is practicable only for $k \leq 12$ loci. For $k > 12$, further approximations (or statistical techniques, which we do not further consider) are required. We describe below two approximations that together lead to an efficient algorithm. We later show empirically that the resulting CSDs have little loss of accuracy in comparison with $\hat{\pi}_{\mathrm{PS}}$.

Approximation 1 (disallowing coalescence): Recall that a conditional genealogy $\mathcal{C}_m$ is composed of mutation, recombination, coalescence, and absorption events. Importantly, within this framework, it is only coalescence events that can *couple* two lineages of $\mathcal{C}_m$ into one (moving backward in time); mutation, recombination, and absorption events have the noncoupling effect of modifying, splitting, and removing lineages, respectively. Intuitively, then, by disallowing coalescence, separate lineages should behave independently; more precisely, given $\mathbf{m} = (\mathbf{e}_{h_1} + \ldots + \mathbf{e}_{h_m})$, and defining $\hat{\pi}_{\mathrm{PS},1}$ to be the CSP obtained from the genealogical process disallowing coalescence, we expect that

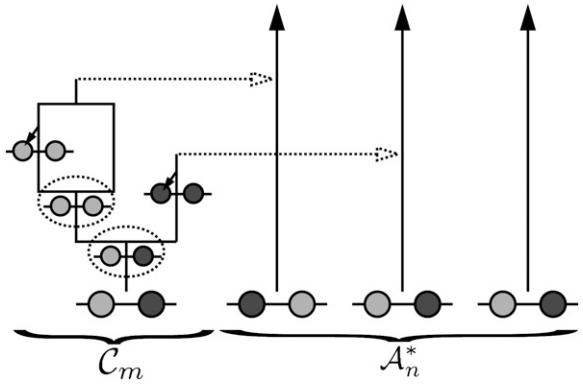

FIGURE 2.—Illustration of a conditional genealogy using the approximation $\mathcal{A}_n = \mathcal{A}_n^*$. Absorption events are indicated by dotted arrows into the "trunk" ancestry $\mathcal{A}_n^*$. Comparing with Figure 1b, observe that $\mathcal{A}_n^*$ is time invariant and extends infinitely into the past.

$$\hat{\pi}_{\mathrm{PS},1}(\mathbf{m} \mid \mathbf{n}) = \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{h_1} + \ldots + \mathbf{e}_{h_m} \mid \mathbf{n}) = \prod_{i=1}^{m} \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{h_i} \mid \mathbf{n}). \quad (3)$$

This is indeed the case, as we prove in the next section. It is worth noting here that disallowing coalescence is not as unreasonable as it first may seem; unlike a normal genealogy, a conditional genealogy does not rely on coalescence events to terminate (absorption events play the analogous role). Although we shall further discuss the merit of this approximation in light of empirical results, for now, it suffices to say that (3) significantly simplifies computation of $\hat{\pi}_{\mathrm{PS},1}$. Assuming a PIM model, a dynamic programming formulation of $\hat{\pi}_{\mathrm{PS},1}$ exists with asymptotic running time $O(2^k k^2)$ (for $|\mathbf{m}| = 1$). Although still exponential in k, this represents a substantial improvement over $\hat{\pi}_{\mathrm{PS}}$, for which constructing and solving a system of equations superexponential in k is required.

Approximation 2 (limiting mutations): We further examine $\hat{\pi}_{\mathrm{PS},1}$, with the objective of finding a sensible polynomial time approximation. Even disallowing coalescence, it is necessary to consider every mutational configuration of the k loci. In a PIM model, there are $O(2^k)$ such configurations, thereby accounting for the exponential running time given above. By artificially limiting the number of mutational configurations, it is again possible to substantially reduce the computational complexity.

In our final approximation $\hat{\pi}_{\mathrm{PS},2}$, we limit the set of mutational configurations to those that are a single mutation away from the original haplotype. Genealogically, this corresponds to disallowing explicit mutation on any lineage that has already mutated; for small values of θ, we expect genealogies that do not conform to this restriction to be relatively unlikely. We shall further discuss the approximation $\hat{\pi}_{\mathrm{PS},2}$ in light of empirical results; for now, it suffices to remark that in a PIM model of mutation, $\hat{\pi}_{\mathrm{PS},2}$ is limited to $k + 1$ mutational states, enabling a modification to the dynamic program with asymptotic running time $O(k^3)$ (for $|\mathbf{m}| = 1$). In principle, this allows the CSP to be computed for a number of loci k on the order of several hundred.

Relation to other approximate CSDs: Several previously proposed approximate CSDs $\hat{\pi}_{\mathrm{SD}}$ (STEPHENS and DONNELLY 2000), $\hat{\pi}_{\mathrm{FD}}$ (FEARNHEAD and DONNELLY 2001), and $\hat{\pi}_{\mathrm{LS}}$ (LI and STEPHENS 2003) are all naturally described as "copying" models, in which a new haplotype is conditionally sampled by making an imperfect copy of one or more haplotypes in an observed sample $\mathbf{n}$. We now describe these copying models and show that each also has a genealogical interpretation; moreover, these interpretations can reasonably be described as approximations of our basic CSD, $\hat{\pi}_{\mathrm{PS}}$.

The copying model for $\hat{\pi}_{\mathrm{SD}}$, applicable when the loci are assumed completely linked (*i.e.*, $\rho = 0$), is as follows: Select a random "source" haplotype h from $\mathbf{n}$ with probability n_h/n and a random copying time t from the exponential distribution with rate $n/2$; having done so, mutate each locus $\ell \in L$ of h a random number m_ℓ of times, with m_ℓ drawn from a Poisson distribution with mean $\theta_\ell t/2$. The resulting haplotype is the conditional sample.

This copying model for $\hat{\pi}_{\mathrm{SD}}$ can be restated as a genealogical process. In particular, set $\rho = 0$ and suppose we wish to conditionally sample a single haplotype. Both $\hat{\pi}_{\mathrm{PS}}$ and $\hat{\pi}_{\mathrm{PS},1}$ are associated with a conditional genealogical process composed of the following events: mutation at locus $\ell \in L$ with rate $\theta_\ell/2$ and absorption into a haplotype lineage of $\mathbf{n}$ with rate $1/2$. By the independence of the mutation and absorption events, this genealogical process coincides with the copying model for $\hat{\pi}_{\mathrm{SD}}$, suggesting that, when $\rho = 0$, $\hat{\pi}_{\mathrm{SD}} = \hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{PS},1}$. This is indeed the case, as we prove in the next section.

The approximate CSD $\hat{\pi}_{\mathrm{FD}}$ extends $\hat{\pi}_{\mathrm{SD}}$ to partially linked loci (*i.e.*, $\rho > 0$). In this case, a new haplotype is sampled in two phases: First, an unspecified haplotype is randomly broken into unspecified fragments under the assumption that a break occurs at each $b \in B$ independently and with probability $\rho_b/(n + \rho_b)$; and second, each fragment is "copied" independently using the $\hat{\pi}_{\mathrm{SD}}$ copying model, restricted to the appropriate set of loci. The specified fragments are then reassembled into a complete haplotype, completing the conditional sample. This copying model is often recast as a hidden Markov model (HMM), with observed states corresponding to the allele at each locus of the sampled haplotype, and hidden states corresponding to the source haplotype in $\mathbf{n}$ and copying time (as in the description of $\hat{\pi}_{\mathrm{SD}}$); the probability of a transition in the hidden states is $\rho_b/(n + \rho_b)$.

As in the case of $\hat{\pi}_{\mathrm{SD}}$, the copying model for $\hat{\pi}_{\mathrm{FD}}$ can be restated as a genealogical process. In particular, consider the conditional genealogical process associ-

ated with $\hat{\pi}_{\mathrm{PS},1}$, artificially divided into an initial "recombination phase," wherein an unspecified haplotype is randomly broken into fragments, and a "nonrecombination phase," wherein these fragments are subject to the normal genealogical events, conditioned on no additional recombinations occurring. In the recombination phase, each breakpoint is used independently, and with probability $\rho_b/(n+\rho_b)$, corresponding to the *marginal* probability of the breakpoint being used in the usual genealogical process for $\hat{\pi}_{\mathrm{PS},1}$. In the nonrecombination phase, each fragment maintains independence by virtue of $\hat{\pi}_{\mathrm{PS},1}$ disallowing coalescence. This two-phase genealogical process coincides with the copying model for $\hat{\pi}_{\mathrm{FD}}$. We conclude that the approximate CSD $\hat{\pi}_{\mathrm{FD}}$ can be considered an approximation of $\hat{\pi}_{\mathrm{PS},1}$.

Finally, the approximate CSD $\hat{\pi}_{\mathrm{LS}}$ is a computational simplification of $\hat{\pi}_{\mathrm{FD}}$ in which the copying process for each fragment is assumed to have $t = 2/n$, rather than t drawn from an exponential distribution. This corresponds, in the associated genealogical process for $\hat{\pi}_{\mathrm{FD}}$, to the assumption that each fragment absorbed into some haplotype of $\mathbf{n}$ in time $t = 2/n$. We do not say anything further about $\hat{\pi}_{\mathrm{LS}}$ since it is closely related to $\hat{\pi}_{\mathrm{FD}}$.

A MATHEMATICAL FORMULATION

In this section, we provide a mathematical derivation of our conditional sampling distribution. Rather than formalizing the genealogical interpretation/approximation discussed in the previous section, we extend the diffusion-generator approximation technique (De Iorio and Griffiths 2004a,b; Griffiths *et al.* 2008) and demonstrate equivalence. We also prove several useful limiting results and provide concrete mathematical statements for the approximations (disallowing coalescence and limiting mutations) mentioned in the previous section.

Notation: To describe our mathematical formulation for an arbitrary number of loci, we need to introduce more notation. In what follows, we build on the notation defined in the previous section. Given a haplotype $h \in \mathcal{H}$ and a locus $\ell \in L = \{1, \ldots, k\}$, we use $h[\ell] \in E_\ell$ to denote the allele at locus ℓ of h. Given any two haplotypes $h, h' \in \mathcal{H}$, we define the following operations:

Substitute: Given a locus $\ell \in L$ and an allele $a \in E_\ell$, define $\mathcal{S}^a_\ell(h) \in \mathcal{H}$ as the haplotype derived from h by substituting the allele at locus ℓ with a.

Recombine: Given a breakpoint $b = (\ell, \ell+1) \in B$, define $\mathcal{R}_b(h, h') \in \mathcal{H}$ as the mosaic haplotype derived by concatenating $h[1], \ldots, h[\ell]$ and $h'[\ell+1], \ldots, h'[k]$.

We also require partially specified haplotypes, in which the alleles at some loci are unspecified. Denote such an unspecified allele by $\bullet$ and define the space of partially specified haplotypes as $\mathcal{G} = (E_1 \cup \{\bullet\}) \times \ldots \times (E_k \cup \{\bullet\})$. For $g \in \mathcal{G}$, let $L(g)$ denote the set of loci at which g has specified (*i.e.*, not $\bullet$) alleles. Then, for $g, g' \in \mathcal{G}$, we say that g and g' are *compatible* and write $g \curlywedge g'$, if $g[\ell] = g'[\ell]$ for all $\ell \in L(g) \cap L(g')$. We define an operation for combining two compatible partially specified haplotypes:

Coalesce: If $g \curlywedge g'$, define $\mathcal{C}(g, g')$ as the haplotype constructed as follows: For $\ell \in L$, $\mathcal{C}(g,g')[\ell] = g[\ell]$ if $g'[\ell] = \bullet$, $\mathcal{C}(g,g')[\ell] = g'[\ell]$ if $g[\ell] = \bullet$, and $\mathcal{C}(g,g')[\ell] = g[\ell] = g'[\ell]$ otherwise.

Given a partially specified haplotype $g \in \mathcal{G}$, we use $B(g)$ to denote the set of breakpoints between the leftmost and the rightmost loci in $L(g)$ and define the following operation for breaking up g into parts:

Break: Given a breakpoint $b = (\ell, \ell+1) \in B(g)$, we use $\mathcal{R}^-_b(g) = (g[1], \ldots, g[\ell], \bullet, \ldots, \bullet)$ to denote the haplotype obtained from g by replacing $g[j]$ with $\bullet$ for all $j \geq \ell + 1$ and $\mathcal{R}^+_b(g) = (\bullet, \ldots, \bullet, g[\ell+1], \ldots, g[k])$ to denote the haplotype obtained from g by replacing $g[j]$ with $\bullet$ for all $j \leq \ell$.

To illustrate the above definitions, consider a three-locus model, setting $E_\ell = \{0, 1\}$ for each locus $\ell \in L = \{1, 2, 3\}$. Suppose $g_1 = (\bullet, \bullet, 1)$, $g_2 = (0, \bullet, 1)$, and $g_3 = (1, 1, \bullet)$. The loci with specified alleles are $L(g_1) = \{3\}$, $L(g_2) = \{1, 3\}$, and $L(g_3) = \{1, 2\}$, and the valid breakpoints are $B(g_1) = \{\emptyset\}$, $B(g_2) = \{(1, 2), (2, 3)\}$, and $B(g_3) = \{(1, 2)\}$. Furthermore, $g_1 \curlywedge g_2$ with $\mathcal{C}(g_1, g_2) = (0, \bullet, 1)$ and $g_1 \curlywedge g_3$ with $\mathcal{C}(g_1, g_3) = (1, 1, 1)$.

A general strategy for computing $\hat{\pi}$: We begin by briefly reviewing the neutral multilocus diffusion process. Within this framework, we formally state the problem and outline the general strategy we use to solve it.

The neutral multilocus diffusion process: Dual to the coalescent is a forward-in-time diffusion process. The state space of the multilocus diffusion process is

$$\Delta = \left\{ \mathbf{x} = (x_h)_{h \in \mathcal{H}} \mid x_h \geq 0 \text{ for all } h \in \mathcal{H} \text{ and } \sum_{h \in \mathcal{H}} x_h = 1 \right\},$$

where x_h corresponds to the population-wide frequency of haplotype h. Being continuous in both time and space, diffusion processes possess many useful mathematical properties. In particular, associated with a diffusion process is a fundamental differential operator $\mathscr{L}$, called the *generator*, with the following property: For any bounded, twice-differentiable function f with continuous second derivatives, the generator satisfies $\mathbb{E}[\mathscr{L} f(\mathbf{X})] = 0$, where $\mathbb{E}$ denotes expectation with respect to the stationary distribution of the diffusion process. The diffusion generator for the neutral model with crossover recombination is $\mathscr{L} = \sum_{h \in \mathcal{H}} \mathscr{L}_h(\partial/\partial x_h)$, where

$$\mathcal{L}_h = \frac{1}{2}\Big\{ x_h \sum_{h' \in \mathcal{H}} (\delta_{h,h'} - x_{h'}) \frac{\partial}{\partial x_{h'}} + \sum_{\ell \in L} \theta_\ell \sum_{a \in E_\ell} x_{\mathcal{S}^a_\ell(h)} \big[P^{(\ell)}_{a,h[\ell]} - \delta_{h,\mathcal{S}^a_\ell(h)}\big] + \sum_{b \in B} \rho_b \sum_{h' \in \mathcal{H}} \big[x_{\mathcal{R}_b(h,h')} x_{\mathcal{R}_b(h',h)} - x_h\big] \Big\},$$

with $\delta_{h,h'}$ denoting the Kronecker delta symbol. Denote by $q(\mathbf{n})$ the probability of obtaining an *ordered* sample with configuration $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$. Making reference to the diffusion process, $q(\mathbf{n}) = \mathbb{E}(q(\mathbf{n} \,|\, \mathbf{X}))$, where $q(\mathbf{n} \,|\, \mathbf{X})$ is the conditional probability of obtaining $\mathbf{n}$ given the population frequencies $\mathbf{X} = (X_h)_{h \in \mathcal{H}}$; more precisely, $q(\mathbf{n} \,|\, \mathbf{X}) = \prod_{h \in \mathcal{H}} \mathbf{X}_h^{n_h}$.

Now let $\mathbf{m} = (m_h)_{h \in \mathcal{H}}$ with $|\mathbf{m}| = m$. Denote by $\pi(\mathbf{m} \,|\, \mathbf{n})$ the conditional probability that, having already observed sample configuration $\mathbf{n}$, the next m sampled haplotypes have configuration $\mathbf{m}$. By the definition of conditional probability, the distributions π and q satisfy the following key identity:

$$\pi(\mathbf{m} \,|\, \mathbf{n}) = \frac{q(\mathbf{m} + \mathbf{n})}{q(\mathbf{n})}. \tag{4}$$

The diffusion-generator formulation: It is our objective to use the diffusion characterization of $q(\mathbf{n})$ along with the above conditioning identity (4) to find a distribution $\hat{\pi}$ approximating π. Shown below is an outline of the diffusion-generator approximation technique for computing $\hat{\pi}$:

1. At stationarity, instead of $\mathbb{E}[\mathscr{L} f(X)] = 0$, assume that a distribution exists with expectation operator $\hat{\mathbb{E}}$ such that the vanishing condition holds componentwise; *i.e.*, for each $h \in \mathcal{H}$,

$$\hat{\mathbb{E}}\Big[\mathcal{L}_h \frac{\partial}{\partial X_h} f(\mathbf{X})\Big] = 0. \tag{5}$$

2. Define the *approximate* sampling distribution $\hat{q}(\mathbf{n}) = \hat{\mathbb{E}}(q(\mathbf{n} \,|\, \mathbf{X}))$ and, motivated by the conditioning identity (4), define the *approximate* CSD $\hat{\pi}(\mathbf{m} \,|\, \mathbf{n}) = \hat{q}(\mathbf{m} + \mathbf{n}) / \hat{q}(\mathbf{n})$.
3. Use an appropriate set of functions $f(\mathbf{X})$ and haplotypes $h \in \mathcal{H}$ in (5) to derive a recursion for $\hat{\pi}$ that does not include $\hat{q}$ terms.

Applying this general strategy, De Iorio and Griffiths (2004a,b) were able to reproduce formally the widely used one-locus CSD introduced by Stephens and Donnelly (2000); in a similar vein, Griffiths *et al.* (2008) were able to devise an approximate CSD in the case of two loci with a restricted mutation model. Our present goal is to apply this diffusion-generator formulation yet again to derive a recursion for an arbitrary number of loci and an arbitrary finite-alleles mutation model. This will be our approximate CSD, which we denote $\hat{\pi}_{\mathrm{PS}}$. After deriving the recursion for $\hat{\pi}_{\mathrm{PS}}$, we show that it coincides with the genealogical formulation of the previous section and provide some intuition for the above approximation.

The main recursion: Using the diffusion-generator approximation formulation described above, we obtain the following theorem, which is proved in the appendix:

Theorem 1. *Let $\mathbf{m} = (m_h)_{h \in \mathcal{H}}$ with $|\mathbf{m}| = m$ and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ with $|\mathbf{n}| = n$. Then the approximate conditional sampling distribution $\hat{\pi}_{\mathrm{PS}}(\mathbf{m} \,|\, \mathbf{n})$ satisfies the following recursion:*

$$\begin{aligned} m\Big[(n + m - 1) + \sum_{\ell \in L} \theta_\ell + \sum_{b \in B} \rho_b\Big] \hat{\pi}_{\mathrm{PS}}(\mathbf{m} \,|\, \mathbf{n}) &= \sum_{h \in \mathcal{H}} m_h \Big[(n_h + (m_h - 1)) \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_h \,|\, \mathbf{n}) \\ &\quad + \sum_{\ell \in L} \theta_\ell \sum_{a \in E_\ell} P^{(\ell)}_{a,h[\ell]} \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_h + \mathbf{e}_{\mathcal{S}^a_\ell(h)} \,|\, \mathbf{n}) \\ &\quad + \sum_{b \in B} \rho_b \sum_{h' \in \mathcal{H}} \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)} \,|\, \mathbf{n}) \Big]. \end{aligned} \tag{6}$$

Although we consider the recursion stated in Theorem 1 to be our primary result, explicit evaluation is not possible since the number of states that must be explored is infinite. To establish a practicable formulation, we extend this result to partially specified haplotypes.

Suppose that $\mathbf{n} = (n_g)_{g \in \mathcal{G}}$ is a configuration allowing unspecified alleles. Conditional on $\mathbf{X}$, the sampling probability becomes $q(\mathbf{n} \,|\, \mathbf{X}) = \prod_{g \in \mathcal{G}} Y_g^{n_g}$, where $Y_g = \sum_{h \in \mathcal{H}: h \wedge g} X_h$ is the total proportion of fully specified haplotypes that subsume the partially specified haplotype $g \in \mathcal{G}$. With $\hat{q}$ and $\hat{\pi}$ defined as before with respect to $\hat{\mathbb{E}}$ and the above $q(\mathbf{n} \,|\, \mathbf{X})$, we obtain the following corollary (its proof is deferred to the appendix):

Corollary 2. *Let $\mathbf{m} = (m_g)_{g \in \mathcal{G}}$ with $|\mathbf{m}| = m$ and $\mathbf{n} = (n_h)_{h \in \mathcal{H}}$ with $|\mathbf{n}| = n$. Then the approximate conditional sampling distribution $\hat{\pi}_{\mathrm{PS}}(\mathbf{m} \,|\, \mathbf{n})$ satisfies the following recursion:*

$$\begin{aligned} \sum_{g \in \mathcal{G}} m_g \Big[(n + m - 1) + \sum_{\ell \in L(g)} \theta_\ell + \sum_{b \in B(g)} \rho_b\Big] \hat{\pi}_{\mathrm{PS}}(\mathbf{m} \,|\, \mathbf{n}) &= \sum_{g \in \mathcal{G}} m_g \Big[\Big(\sum_{h \in \mathcal{H}: h \wedge g} n_h\Big) \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_g \,|\, \mathbf{n}) \\ &\quad + \sum_{g' \in \mathcal{G}: g' \wedge g} (m_{g'} - \delta_{g,g'}) \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_g + \mathbf{e}_{\mathcal{C}(g,g')} \,|\, \mathbf{n}) \\ &\quad + \sum_{\ell \in L(g)} \theta_\ell \sum_{a \in E_\ell} P^{(\ell)}_{a,g[\ell]} \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_g + \mathbf{e}_{\mathcal{S}^a_\ell(g)} \,|\, \mathbf{n}) \\ &\quad + \sum_{b \in B(g)} \rho_b \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g)} \,|\, \mathbf{n}) \Big]. \end{aligned} \tag{7}$$

Remark. Determining a simple recursion for $\hat{\pi}(\mathbf{m}\,|\,\mathbf{n})$ in the general case, when $\mathbf{n} = (n_g)_{g\in\mathcal{G}}$ (*i.e.*, haplotypes in $\mathbf{n}$ may contain unspecified alleles), remains an important open problem.

To see that explicit evaluation is possible, suppose $\mathbf{m} = (m_g)_{g\in\mathcal{G}}$ and $\mathbf{n} \in (n_h)_{h\in\mathcal{H}}$ and denote the total number of specified loci in $\mathbf{m}$ by $L(\mathbf{m}) = \sum_{g\in\mathcal{G}} m_g|L(g)|$. Applying (7) for $\hat{\pi}_{\mathrm{PS}}(\mathbf{m}\,|\,\mathbf{n})$, it is evident that each term on the right-hand side is of form $\hat{\pi}_{\mathrm{PS}}(\mathbf{m}'\,|\,\mathbf{n})$ with $L(\mathbf{m}') \le L(\mathbf{m})$. Thus, by induction, only a finite number of states need be explored, and so repeated application of (7) yields a closed set of coupled linear equations, within which $\hat{\pi}_{\mathrm{PS}}(\mathbf{m}\,|\,\mathbf{n})$ is a variable. This system can be solved using standard numerical techniques.

Connection to the genealogical formulation: Recall the conditional genealogical process for constructing $\mathcal{C}_m$ using the approximation $\mathcal{A}_n = \mathcal{A}_n^*$ described in the previous section. Employing this formulation, it is possible to compute $\hat{\pi}(\mathbf{m}\,|\,\mathbf{n})$ by applying the law of total probability with respect to the most recent event (*i.e.*, the usual "forward–backward" argument). We leave it to the reader to verify that doing so will yield the recursion (6) or (7), depending upon whether nonancestral loci are explicitly considered. This establishes the equivalence between our genealogical and mathematical formulations.

This equivalence may appear surprising given that the componentwise vanishing assumption (5) does not have an obvious genealogical interpretation. Griffiths *et al.* (2008) provide some intuition, pointing out that (5) is mathematically equivalent to assuming that, conditioning on sample $\mathbf{n}$, the probability that the most recent event includes haplotype $h \in \mathcal{H}$ is equal to n_h/n. This is precisely the *prior* probability (*i.e.*, the probability if the the haplotypes of $\mathbf{n}$ were unspecified) and therefore furnishes a reasonable and internally consistent approximation. Importantly, this assumption allows us to genealogically restrict attention to a particular haplotype h; we may thus restrict attention to the subconfiguration $\mathbf{m}$ of $\mathbf{m} + \mathbf{n}$. In this way, a genealogy that modifies only lineages associated with $\mathbf{m}$ is constructed, precisely what occurs in our genealogical formulation.

Analytic formulas: In the one-locus case ($k = 1$) with parent-independent mutation, (7) immediately yields a conditional sampling formula that agrees with the exact one-locus CSD π. More precisely, given an additional allele $a \in E = \mathcal{H}$ and a previously observed sample $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$, we obtain

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_a\,|\,\mathbf{n}) = \pi(\mathbf{e}_a\,|\,\mathbf{n}) = \frac{n_a + \theta P_a}{n + \theta}. \tag{8}$$

Both Stephens and Donnelly (2000) and Fearnhead and Donnelly (2001) obtained the same result; as we shall soon see, this is part of a more general result that holds in the limit as $\rho \to 0$.

In the two-locus case ($k = 2$) with parent-independent mutation, it is possible to obtain an analytic formula. Given an additional haplotype $(a_1,\ a_2) \in E_1 \times E_2 = \mathcal{H}$ and a previously observed sample $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$ with $|\mathbf{n}| = n$, we obtain

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{(a_1,a_2)}\,|\,\mathbf{n}) = \frac{1}{\mathcal{N}}\Big[n_{(a_1,a_2)} + \theta_1 P_{a_1}^{(1)}\pi(\mathbf{e}_{a_2}\,|\,\mathbf{n}) + \theta_2 P_{a_2}^{(2)}\pi(\mathbf{e}_{a_1}\,|\,\mathbf{n}) + \rho\cdot\frac{2n+\theta_1+\theta_2}{2(n+1)+\theta_1+\theta_2}\cdot\pi(\mathbf{e}_{a_1}\,|\,\mathbf{n})\pi(\mathbf{e}_{a_2}\,|\,\mathbf{n})\Big],$$

where $\mathcal{N} = n + \theta_1 + \theta_2 + \rho[(2n+\theta_1+\theta_2)/(2(n+1)+\theta_1+\theta_2)]$, and $\pi(\mathbf{e}_a\,|\,\mathbf{n})$ is the exact one-locus CSP (8), with $\mathbf{n}$ appropriately marginalized. This form is quite similar to that derived by Griffiths *et al.* (2008), with the minor differences attributable to a different treatment of "symmetry" conditions.

Although it is theoretically possible to obtain analytic solutions for $k > 2$, little simplification is possible, and solving them is tantamount to generating and solving the coupled system of equations directly. We next show that some algebraic simplification is possible in two limiting cases.

Limiting distributions: For convenience, we set $\rho_b = \rho$, for all $b \in B$, and consider the CSD in both the $\rho \to 0$ and the $\rho \to \infty$ limits. We find that, in the $\rho \to 0$ limit, $\hat{\pi}_{\mathrm{PS}}$ coincides with Stephens and Donnelly's CSD $\hat{\pi}_{\mathrm{SD}}$ and, by extension, Fearnhead and Donnelly's $\hat{\pi}_{\mathrm{FD}}$. In the $\rho \to \infty$ limit, $\hat{\pi}_{\mathrm{PS}}$ coincides with $\hat{\pi}_{\mathrm{FD}}$, with $\hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{FD}} = \pi$ in the case of parent-independent mutation.

The $\rho \to 0$ limit: Set $\rho = 0$, and let $\mathbf{m} = \mathbf{e}_{h'}$ for some $h' \in \mathcal{H}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$. Then (6) yields the following simplified recursion:

$$\left(n + \sum_{\ell\in L}\theta_\ell\right)\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{h'}\,|\,\mathbf{n}) = n_{h'} + \sum_{\ell\in L}\theta_\ell\sum_{a\in E_\ell}P_{a,h'[\ell]}^{(\ell)}\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{\mathcal{S}_\ell^a(h')}\,|\,\mathbf{n}). \tag{9}$$

Recall that Stephens and Donnelly's CSD $\hat{\pi}_{\mathrm{SD}}$, applicable when the loci are completely linked (*i.e.*, $\rho = 0$), is formulated most naturally as a copying model, in which a new haplotype is conditionally sampled by choosing a previously sampled haplotype and stochastically mutating it according to a specified process (see the appendix for details). Despite the disparity of the genealogical description for $\hat{\pi}_{\mathrm{PS}}$ and the copying model description for $\hat{\pi}_{\mathrm{SD}}$, the following proposition (also proved in the appendix) assures us that they are equivalent.

Proposition 3. *Let $\mathbf{m} = \mathbf{e}_{h'}$ for some $h \in \mathcal{H}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$. Then if $\rho_b = 0$ for all $b \in B$, $\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{h'}\,|\,\mathbf{n}) = \hat{\pi}_{\mathrm{SD}}(\mathbf{e}_{h'}\,|\,\mathbf{n})$.*

In addition to providing a genealogical interpretation for $\hat{\pi}_{\mathrm{SD}}$, the above proposition indicates that, when $\rho =$

0, $\hat{\pi}_{\mathrm{PS}}$ may be approximated using the Gaussian quadrature method proposed by Stephens and Donnelly (2000); conversely it provides an exact method for computing $\hat{\pi}_{\mathrm{SD}}$, generalizing similar results to an arbitrary number of loci and mutation model. Finally, when $\rho = 0$, Fearnhead and Donnelly's CSD $\hat{\pi}_{\mathrm{FD}}$ coincides, by construction, with $\hat{\pi}_{\mathrm{SD}}$, and so $\hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{FD}} = \hat{\pi}_{\mathrm{SD}}$.

The $\rho \to \infty$ limit: Let $\mathbf{n} \in (n_g)_{g\in\mathcal{G}}$ and denote the one-locus marginal configuration for $\ell \in L$ by $\mathbf{n}[\ell] = (n_a)_{a\in E_\ell}$, where $n_a = \sum_{g\in\mathcal{G};g[\ell]=a} n_g$ is the number of haplotypes of $\mathbf{n}$ with allele a at locus ℓ. In the appendix, we prove that in the $\rho \to \infty$ limit, $\hat{\pi}_{\mathrm{PS}}$ may be decomposed into a product of one-locus likelihoods:

Proposition 4. *Let $\mathbf{m} \in (m_g)_{g\in\mathcal{G}}$ and $\mathbf{n} \in (n_h)_{h\in\mathcal{H}}$. Then in the limit $\rho \to \infty$,*

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{m} \,|\, \mathbf{n}) = \prod_{\ell\in L} \hat{\pi}_{\mathrm{PS}}(\mathbf{m}[\ell] \,|\, \mathbf{n}[\ell]). \tag{10}$$

Recall that Fearnhead and Donnelly's CSD $\hat{\pi}_{\mathrm{FD}}$ enjoys the same limiting decomposition, and the one-locus $\hat{\pi}_{\mathrm{FD}}$ coincides with the one-locus $\hat{\pi}_{\mathrm{SD}}$, which in turn agrees with the one-locus $\hat{\pi}_{\mathrm{PS}}$ by Proposition 3. In conjunction with Proposition 4, these facts imply that $\hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{FD}}$ in the limit $\rho \to \infty$. It is encouraging that the true CSD π also exhibits this limiting decomposition (this follows directly from the well-known limiting decomposition of the sampling distribution q). Coupled with the fact that the one-locus CSD (8) is exact for PIM models, we may also conclude that for PIM models in the $\rho \to \infty$ limit, $\hat{\pi}_{\mathrm{PS}} = \hat{\pi}_{\mathrm{FD}} = \pi$.

Approximations to $\hat{\pi}_{\mathrm{PS}}$: In the general case, when $0 < \rho < \infty$, computing a CSP using $\hat{\pi}_{\mathrm{PS}}$ requires that a set of coupled linear equations be constructed and solved. In particular, for $|\mathbf{m}| = 1$ in the case of a PIM model, the number of generated equations is the $(k+1)$th Bell number B_{k+1}, where k is the number of loci. Thus, the number of equations is superexponential in k, indicating that computation of $\hat{\pi}_{\mathrm{PS}}$ is intractable with increasing k. We consider two approximations, motivated by the genealogical formulation discussed in the previous section.

Approximation 1 (disallowing coalescence): Modifying (7) by disallowing coalescence—corresponding to removing the second term on the right-hand side and renormalizing the left-hand side—we obtain a recursion for a new approximate CSD, which we denote $\hat{\pi}_{\mathrm{PS},1}$. Some genealogical justification for this approximation was provided for this in the previous section, and empirical justification is provided in the next section. Here, we are interested primarily in the computational aspects, which rely on the following result (proved in the appendix):

Proposition 5. *For $\mathbf{m} = \mathbf{e}_{g_1} + \ldots + \mathbf{e}_{g_m}$, where $g_1, \ldots, g_m \in \mathcal{G}$, and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$, the approximate CSD $\hat{\pi}_{\mathrm{PS},1}$ satisfies*

$$\hat{\pi}_{\mathrm{PS},1}(\mathbf{m} \,|\, \mathbf{n}) = \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{g_1} + \ldots + \mathbf{e}_{g_m} \,|\, \mathbf{n}) = \prod_{i=1}^{m} \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{g_i} \,|\, \mathbf{n}). \tag{11}$$

Resulting from Proposition 5 is a simplified recursion for $\hat{\pi}_{\mathrm{PS},1}$: Letting $g \in \mathcal{G}$,

$$\begin{aligned}
&\left[n + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b \right] \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_g \,|\, \mathbf{n}) \\
&= \left(\sum_{h\in\mathcal{H}:h\wedge g} n_h \right) \\
&\quad + \sum_{\ell\in L(g)} \theta_\ell \sum_{a\in E_\ell} P^{(\ell)}_{a,g[\ell]} \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{\mathcal{S}^a_\ell(g)} \,|\, \mathbf{n}) \\
&\quad + \sum_{b\in B(g)} \rho_b \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{\mathcal{R}^-_b(g)}) \hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{\mathcal{R}^+_b(g)} | \, \mathbf{n}).
\end{aligned} \tag{12}$$

Making use of this recursion, and assuming that $|E_\ell| = s$ for all $\ell \in L$, a system of $O(s^k k^2)$ equations needs to be generated and solved, far fewer than the superexponential number required for $\hat{\pi}_{\mathrm{PS}}$. Moreover, assuming a PIM model of mutation, there is an evident dynamic programming formulation for $\hat{\pi}_{\mathrm{PS},1}$ that runs in $O(2^k \cdot k^2)$ time.

Approximation 2 (limiting mutations): Despite being significantly faster to compute than $\hat{\pi}_{\mathrm{PS}}$, the approximate CSD $\hat{\pi}_{\mathrm{PS},1}$ is still exponential in the number of loci. This remains true even for $\rho = 0$, indicating that the complication is a result of mutation rather than recombination. In particular, looking at the form of (12), it is clear that $\hat{\pi}_{\mathrm{PS},1}$ must be evaluated for every partially specified haplotype $g \in \mathcal{G}$. As discussed in the previous section and empirically justified in the next section, when θ is relatively small, a reasonable approximation to $\hat{\pi}_{\mathrm{PS},1}$ may be obtained by artificially limiting the set of accessible haplotypes.

In particular, denote by $\hat{\pi}_{\mathrm{PS},2}$ the approximate CSD obtained by limiting the "explicitly computed" $\hat{\pi}_{\mathrm{PS},1}$ terms to those haplotypes that are within a single mutational step of the haplotype g of interest. Then,

$$\begin{aligned}
&\left[n + \sum_{\ell\in L(g)} \theta_\ell + \sum_{b\in B(g)} \rho_b \right] \hat{\pi}_{\mathrm{PS},2}(\mathbf{e}_g \,|\, \mathbf{n}) \\
&= \left(\sum_{h\in\mathcal{H}:h\wedge g} n_h \right) \\
&\quad + \sum_{\ell\in L(g)} \theta_\ell \sum_{a\in E_\ell} P^{(\ell)}_{a,g[\ell]} \hat{\pi}_{\mathrm{Alt}}(\mathbf{e}_{\mathcal{S}^a_\ell(g)} \,|\, \mathbf{n}) \\
&\quad + \sum_{b\in B(g)} \rho_b \hat{\pi}_{\mathrm{PS},2}(\mathbf{e}_{\mathcal{R}^-_b(g)}) \hat{\pi}_{\mathrm{PS},2}(\mathbf{e}_{\mathcal{R}^+_b(g)} \,|\, \mathbf{n}),
\end{aligned} \tag{13}$$

where $\hat{\pi}_{\mathrm{Alt}} \neq \hat{\pi}_{\mathrm{PS},2}$ is an alternative approximate CSD. The "canonical" choice for $\hat{\pi}_{\mathrm{Alt}}$ is

$$\left[n + \sum_{b \in B(g)} \rho_b\right] \hat{\pi}_0(\mathbf{e}_g \,|\, \mathbf{n}) = \left(\sum_{h \in \mathcal{H}: h \curlywedge g} n_h\right) + \sum_{b \in B(g)} \rho_b \hat{\pi}_0(\mathbf{e}_{\mathcal{R}_b^-(g)}) \hat{\pi}_0(\mathbf{e}_{\mathcal{R}_b^+(g)} \,|\, \mathbf{n}),$$

which is (13) with further mutation disallowed (*i.e.*, $\theta_\ell = 0$ for all $\ell \in L$). Using $\hat{\pi}_{\text{Alt}} = \hat{\pi}_0$, and again assuming that $|E_\ell| = s$ for all $\ell \in L$, a system of $O(sk^3)$ equations needs to be generated and solved. Further assuming a PIM model of mutation, a dynamic programming formulation can be used, which runs in $O(k^3)$ time. We have found that better results are obtained by using $\hat{\pi}_{\text{Alt}} = \hat{\pi}_{\text{FD}}$, which implicitly does allow for additional mutation. This modification does not change the asymptotic running time.

EMPIRICAL RESULTS

In this section, we evaluate the accuracy of our CSD $\hat{\pi}_{\text{PS}}$, along with the approximations $\hat{\pi}_{\text{PS},1}$ and $\hat{\pi}_{\text{PS},2}$, and compare it with the accuracy of the approximate CSDs $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$, respectively proposed by Fearnhead and Donnelly (2001) and by Li and Stephens (2003). Analytically computing the true CSP is typically not possible, so we rely on importance sampling to provide reference values. Even within this Monte Carlo framework, the size of problems that can be analyzed is modest, thus limiting the scope of our study.

We find that $\hat{\pi}_{\text{PS}}$ and the associated approximations ($\hat{\pi}_{\text{PS},1}$ and $\hat{\pi}_{\text{PS},2}$) are more accurate than $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$ in a variety of circumstances. In addition, we consider the PAC pseudolikelihood framework mentioned in the Introduction and demonstrate that the improved accuracy of our CSDs has a positive impact on PAC-based estimation, generally providing improved accuracy for both likelihood and maximum-likelihood estimates.

Data simulation: For simplicity, we consider a two-allele model and set $\mathbf{P}^{(l)} = \mathbf{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ and $\theta_\ell = \theta$ for all loci $\ell \in L$ and $\rho_b = \rho$ for all breakpoints $b \in B$. Using a coalescent with recombination simulator, with $\rho = \rho_0$ and $\theta = \theta_0$, we may sample a k-locus n-haplotype sample configuration $\mathbf{n}$. Given such a configuration, we may subsample a k'-locus n'-haplotype configuration $\mathbf{n}'$ (for $k' \le k$ and $n' \le n$) by randomly selecting n' haplotypes and restricting attention to a k' subset of the loci. In particular, the k' subset is chosen as follows (method, M):

M1. The central k' loci, when θ_0 is large so that most or all loci segregate.

M2. The central k' *segregating* loci, when θ_0 is small so that few loci segregate. This procedure corresponds to the typical usage of $\hat{\pi}$ on genomic data, in which only segregating sites are considered.

Finally, given a k-locus n-haplotype configuration $\mathbf{n}$, we may subsample a k-locus n-haplotype conditional configuration $C = (\mathbf{e}_h, \mathbf{n} - \mathbf{e}_h)$ by withholding a single haplotype h from $\mathbf{n}$ uniformly at random. For notational simplicity, we define π on such a conditional configuration in the natural way: $\pi_\rho(C) = \pi(\mathbf{e}_h \,|\, \mathbf{n} - \mathbf{e}_h, \rho)$.

CSD accuracy: We evaluate the accuracy of each approximate CSD $\hat{\pi}$ as a function of three parameter values: the number of loci, k; the number of haplotypes in the conditional configuration, n; and the recombination rate, ρ. More precisely, we approximate the expected relative error as

$$\text{CSDErr}_{k,n,\rho}(\hat{\pi}) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{|\,\hat{\pi}_\rho(C^{(i)}) - \pi_\rho(C^{(i)})\,|}{\pi_\rho(C^{(i)})}, \quad (14)$$

where N denotes the number of simulated data sets and $C^{(i)}$ is a k-locus n-haplotype conditional configuration sampled as indicated above, with parameters θ_0 and ρ_0. To keep the requisite computation reasonable, we consider three experiments, each time fixing two parameters and allowing the third one to vary. In all cases, $\theta = \theta_0$ is used to evaluate $\hat{\pi}$. The results for $\hat{\pi}_{\text{PS},1}$ and $\hat{\pi}_{\text{PS},2}$ are very similar, so below we discuss only the latter.

We first consider the case in which $\theta_0 = 1$ and $\rho_0 = 4$. Biologically $\theta_0 = 1$ corresponds to a relatively high mutation rate, not so uncommon in retroviruses (McVean *et al.* 2002). The specific parameter settings and results are shown in Figure 3. Under these circumstances, the CSDErr values of our approximations $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{PS},2}$ are comparable and are smaller than those for both $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$. We remark that these are averaged results and do not imply that the CSP produced by $\hat{\pi}_{\text{PS}}$ is always more accurate than that produced by $\hat{\pi}_{\text{FD}}$ or $\hat{\pi}_{\text{LS}}$.

All of the approximate CSDs become less accurate as the number of loci increases (see Figure 3a). However, there is significant variation in the rate that this loss occurs, and $\hat{\pi}_{\text{FD}}$ and $\hat{\pi}_{\text{LS}}$ lose accuracy more quickly than $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{PS},2}$; this result may have a significant consequence at a genomic scale, in which hundreds of segregating loci (or many more) are often considered. In contrast, all of the approximate CSDs become more accurate as the recombination rate increases (see Figure 3b). The correspondence between $\hat{\pi}_{\text{PS}}$ and $\hat{\pi}_{\text{FD}}$ at $\rho = 0$ may be explained by the theoretical result in Proposition 3 and the surrounding discussion; similarly, Proposition 4 ensures that $\hat{\pi}_{\text{PS}} = \hat{\pi}_{\text{PS},2} = \hat{\pi}_{\text{FD}} = \pi$ in the $\rho \to \infty$ limit, indicating that the values of CSDErr for $\hat{\pi}_{\text{PS}}$, $\hat{\pi}_{\text{PS},2}$, and $\hat{\pi}_{\text{FD}}$ converge to 0. Finally, as the number of haplotypes in the conditional configuration increases, the values of CSDErr for the different CSDs appear to converge (see Figure 3c). Interestingly, as the number n of haplotypes decreases, $\hat{\pi}_{\text{LS}}$ becomes less accurate, while $\hat{\pi}_{\text{PS}}$ becomes more accurate; this result may have an effect on PAC computation, since small conditional configurations are necessarily considered.

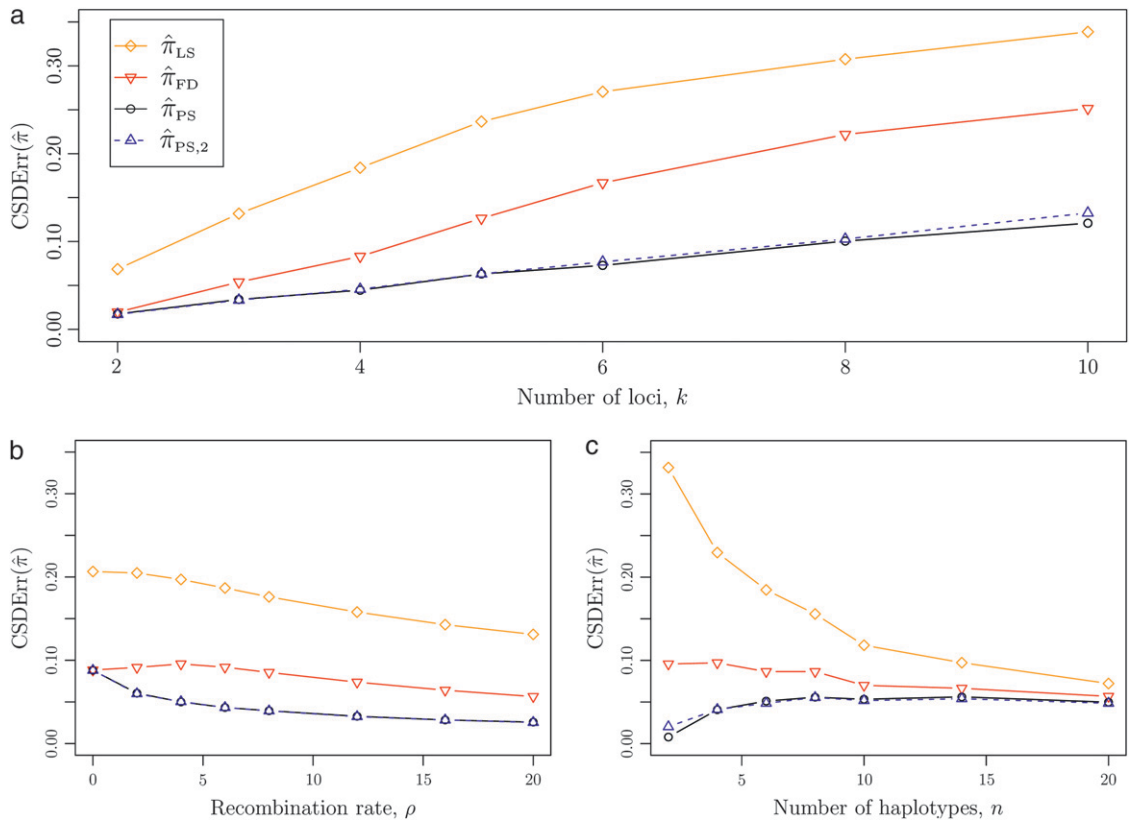

FIGURE 3.—Relative error of CSDs for $\theta_0 = 1$ and $\rho_0 = 4$. See (14) for definition of $\mathrm{CSDErr}_{k,n,\rho}(\hat{\pi})$. With $\theta_0 = 1$ and $\rho_0 = 4$, we used a coalescent simulator to generate 250 data sets, each with 25 haplotypes and 10 loci. Then, requisite k-locus, n-haplotype conditional configurations $\{C^{(i)}\}_{i=1,\ldots,250}$ were obtained using method M1 described in the text. (a) $k \in \{2, 3, 4, 5, 6, 8, 10\}$, $n = 6$, and $\rho = \rho_0$. (b) $k = 4$, $n = 6$, and $\rho \in \{0, 2, 4, 6, 8, 12, 16, 20\}$. (c) $k = 4$, $n \in \{2, 4, 6, 8, 10, 14, 20\}$, and $\rho = \rho_0$.

We next consider the case in which $\theta_0 = 0.01$ and $\rho_0 = 0.1$, corresponding biologically to moderate mutation and recombination rates. The specific parameter settings and results are presented in Figure 4. As in the previous case, the approximations $\hat{\pi}_{PS}$ and $\hat{\pi}_{PS,2}$ are generally more accurate than $\hat{\pi}_{FD}$ and $\hat{\pi}_{LS}$. The accuracy differences among the approximations, however, are less pronounced; the precise cause and degree of this effect (as the parameters, including θ_0 and ρ_0, vary) require further theoretical and empirical investigation.

As before, all of the CSDs become less accurate as the number of loci increases (see Figure 4a) and more accurate as the recombination rate increases (see Figure 4b). In contrast with the previous case, $\hat{\pi}_{PS,2}$ appears to be somewhat more accurate than $\hat{\pi}_{PS}$; this result is surprising since $\hat{\pi}_{PS,2}$ makes more approximations than $\hat{\pi}_{PS}$. A similar phenomenon appears in the context of PAC accuracy and is explored in more detail below. Finally, as the number of haplotypes in the conditional configuration increases, the values of CSDErr for the different CSDs appear to converge (see Figure 4c); as before, for small numbers of haplotypes $\hat{\pi}_{LS}$ is less accurate than $\hat{\pi}_{PS}$, although the difference is less pronounced.

PAC-likelihood accuracy: We evaluate the accuracy of each approximate CSD $\hat{\pi}$ in the context of the PAC pseudolikelihood framework. Since the true CSD π provides the correct likelihood within this framework, we expect that better approximations $\hat{\pi}$ provide better approximations of the true likelihood. Denote by $q_{\hat{\pi}}(\mathbf{n})$ the ordered PAC likelihood obtained using CSD $\hat{\pi}$ and 100 random permutations of the haplotypes in $\mathbf{n}$. We approximate the mean relative error as

$$\mathrm{PACErr}_{k,n,\rho}(\hat{\pi}) \approx \frac{1}{N}\sum_{i=1}^{N}\frac{|\, q_{\hat{\pi},\rho}(\mathbf{n}^{(i)}) - q_{\rho}(\mathbf{n}^{(i)})\,|}{q_{\rho}(\mathbf{n}^{(i)})}, \quad (15)$$

where N denotes the number of simulated data sets and $\mathbf{n}^{(i)}$ is a k-locus n-haplotype configuration sampled from the coalescent with recombination, with parameters θ_0 and ρ_0. We consider fixing k and n and allowing ρ to vary. In all cases, $\theta = \theta_0$ is used to evaluate $q_{\hat{\pi}}$. The PAC-likelihood accuracy results for $\hat{\pi}_{PS,1}$ and $\hat{\pi}_{PS,2}$ are very similar, and so below we discuss only the latter.

We first consider the case in which $\theta_0 = 1$ and $\rho_0 = 4$. The specific parameter settings and results are presented in Figure 5. Under these circumstances, the approximations $\hat{\pi}_{PS}$ and $\hat{\pi}_{PS,2}$ yield PAC likelihoods that are more accurate than those produced using $\hat{\pi}_{FD}$ or $\hat{\pi}_{LS}$. Moreover, comparing Figure 5a and 5b for $k = 3$ and $k = 5$ loci, respectively, it appears that as the number

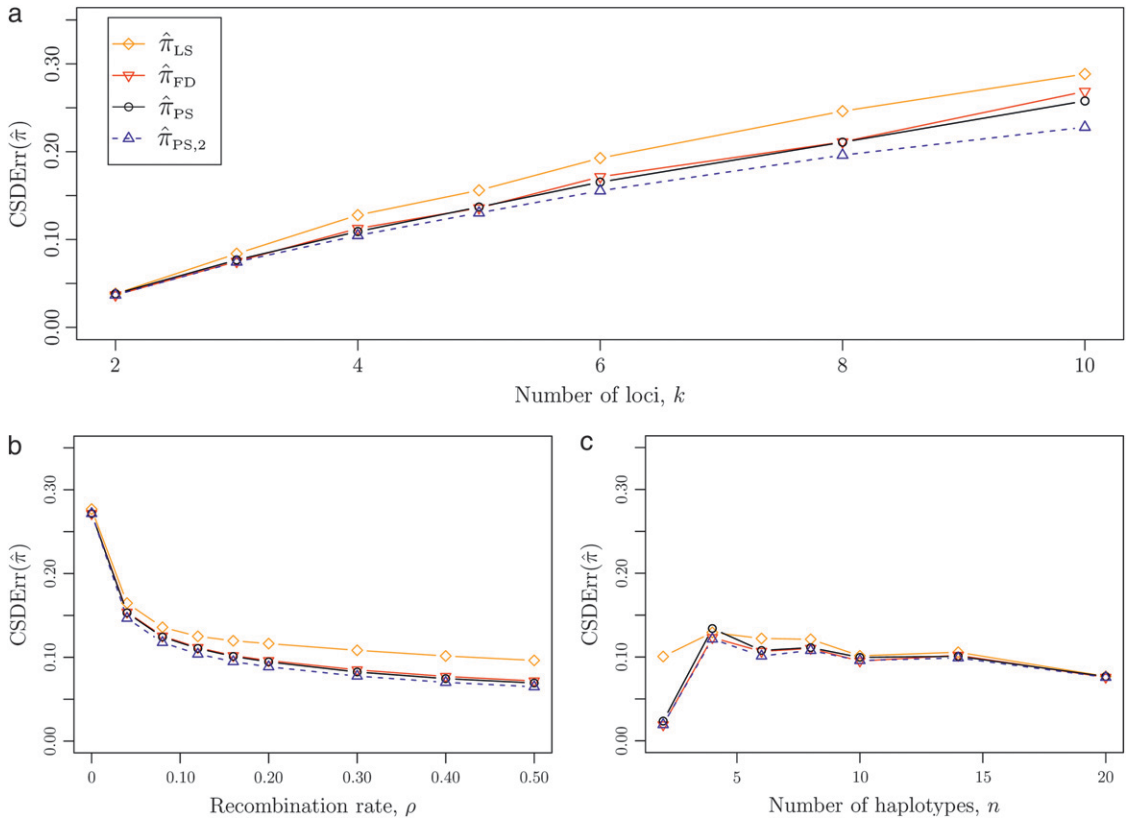

FIGURE 4.—Relative error of CSDs for $\theta_0 = 0.01$ and $\rho_0 = 0.1$. See (14) for definition of $\text{CSDErr}_{k,n,\rho}(\hat{\pi})$. With $\theta_0 = 0.01$ and $\rho_0 = 0.1$, we used a coalescent simulator to generate 250 data sets, each with 25 haplotypes and 500 loci. Then, requisite k-locus n-haplotype conditional configurations $\{C^{(i)}\}_{i=1,\ldots,250}$ were obtained using method M2 described in the text. (a) $k \in \{2, 3, 4, 5, 6, 8, 10\}$, $n = 6$, and $\rho = \rho_0$. (b) $k = 4$, $n = 6$, and $\rho \in \{0, 4, 8, 12, 16, 20, 30, 40, 50\} \times 10^{-2}$. (c) $k = 4$, $n \in \{2, 4, 6, 8, 10, 14, 20\}$, and $\rho = \rho_0$.

of loci increases, the difference in PAC-likelihood accuracy increases; this result might be anticipated from Figure 3a, which shows that the difference in CSD accuracy increases in a similar fashion. Finally, for the range of recombination rates shown, observe that PACErr for $\hat{\pi}_{LS}$ and $\hat{\pi}_{FD}$ notably increases as ρ increases; PACErr for $\hat{\pi}_{PS,2}$ also increases as ρ increases, but only slightly. Contrast this with Figure 3b, which shows that the CSD accuracy decreases as the recombination rate increases. This result is particularly surprising since PACErr $\to 0$ for both $\hat{\pi}_{PS,2}$ and $\hat{\pi}_{FD}$ (because $\hat{\pi}_{PS,2} = \hat{\pi}_{FD} = \pi$) in the $\rho \to \infty$ limit.

We next consider the case in which $\theta_0 = 0.01$ and $\rho_0 = 0.1$. The specific parameter settings and results are presented in Figure 6. As before, the approximations $\hat{\pi}_{PS}$ and $\hat{\pi}_{PS,2}$ yield PAC likelihoods that are more accurate than those produced using $\hat{\pi}_{FD}$ and $\hat{\pi}_{LS}$, and this effect appears to increase with the number of loci. Comparing with CSDErr in Figure 4, there are two interesting observations: First, in contrast to the similar values of CSDErr for $\hat{\pi}_{PS}$ and $\hat{\pi}_{FD}$, the PAC likelihoods using $\hat{\pi}_{PS}$ are significantly more accurate than those using $\hat{\pi}_{FD}$; and second, in concordance with the smaller values of CSDErr for $\hat{\pi}_{PS,2}$ than for $\hat{\pi}_{PS}$, the PAC

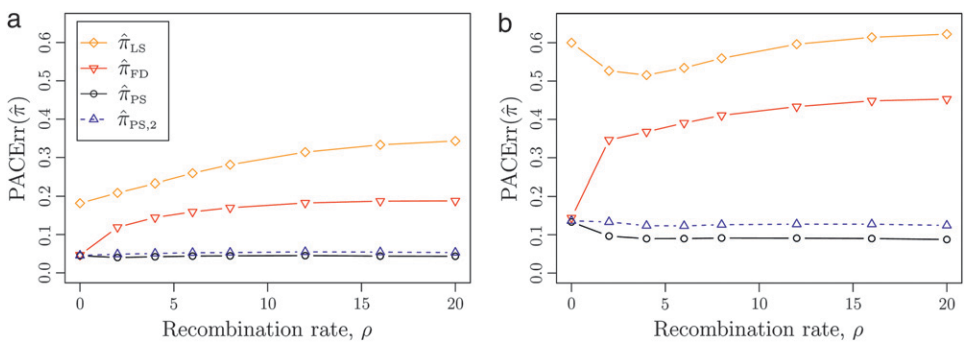

FIGURE 5.—Relative error of PAC likelihoods for $\theta_0 = 1$ and $\rho_0 = 4$. See (15) for definition of $\text{PACErr}_{k,n,\rho}(\hat{\pi})$. With $\theta_0 = 1$ and $\rho_0 = 4$, we used a coalescent simulator to generate 250 data sets, each with 25 haplotypes and 10 loci. Then, requisite k-locus n-haplotype configurations $\{\mathbf{n}^{(i)}\}_{i=1,\ldots,250}$ were obtained using method M1 described in the text. (a) $k = 3$, $n = 25$, and $\rho \in \{0, 2, 4, 6, 8, 12, 16, 20\}$. (b) $k = 5$, $n = 25$, and $\rho \in \{0, 2, 4, 6, 8, 12, 16, 20\}$.

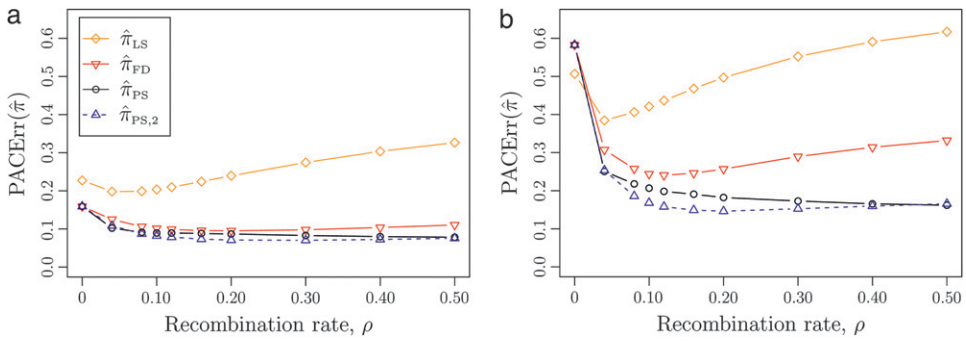

FIGURE 6.—Relative error of PAC likelihoods for $\theta_0 = 0.01$ and $\rho_0 = 0.1$. See (15) for definition of $\text{PACErr}_{k,n,\rho}(\hat{\pi})$. With $\theta_0 = 0.01$ and $\rho_0 = 0.1$, we used a coalescent simulator to generate 250 data sets, each with 25 haplotypes and 500 loci. Then, requisite k-locus n-haplotype configurations $\{\mathbf{n}^{(i)}\}_{i=1,\ldots,250}$ were obtained using method M2 described in the text. (a) $k = 3$, $n = 25$, and $\rho \in \{0, 4, 8, 12, 16, 20, 30, 40, 50\} \times 10^{-2}$. (b) $k = 5$, $n = 25$, and $\rho \in \{0, 4, 8, 12, 16, 20, 30, 40, 50\} \times 10^{-2}$.

likelihoods using $\hat{\pi}_{PS,2}$ are more accurate than those using $\hat{\pi}_{FD}$ for much of the domain.

Thus motivated, we consider the *signed* PACErr, obtained by removing the absolute value from (15); the signed result corresponding to Figure 6b is presented in Figure 7. Observe that the values of the signed PACErr for both $\hat{\pi}_{PS,2}$ and $\hat{\pi}_{PS}$ are initially positive, pass through 0 to become negative, and ultimately must return to 0 in the $\rho \to \infty$ limit; in contrast, values of the signed PACErr for $\hat{\pi}_{FD}$ make a more deliberate descent toward 0. We might expect that such "transient" domains of near unbiasedness demonstrated by $\hat{\pi}_{FD}$ and $\hat{\pi}_{PS,2}$ affect the accuracy of the associated PACErr.

Indeed, comparing with Figure 6b, there is a rough correspondence between the domains in which values of the signed PACErr for $\hat{\pi}_{PS,2}$ and $\hat{\pi}_{FD}$ are very near 0 and the domains in which the PAC likelihoods using $\hat{\pi}_{PS,2}$ and $\hat{\pi}_{FD}$ have the highest accuracy. Within these respective domains, $\hat{\pi}_{PS,2}$ produces a PAC likelihood that is more accurate than $\hat{\pi}_{PS}$, but $\hat{\pi}_{FD}$ does not, an effect that may be due to an increased variance associated with $\hat{\pi}_{FD}$. Finally, recall that $\hat{\pi}_{PS,2}$ is also more accurate than $\hat{\pi}_{PS}$ in terms of CSDErr (see Figure 4). A comparable analysis of signed CSDErr (data not shown) indicates that a similar effect may be at work, although on a significantly larger scale; additional results would need to be collected to make this claim decisively.

PAC–maximum-likelihood estimate accuracy: Finally, we consider using the PAC pseudolikelihood framework to obtain maximum-likelihood estimates (MLEs) for the recombination rate ρ. Since the true CSD π would provide the true MLE within this framework, we expect that better approximations $\hat{\pi}$ will provide better MLEs. Denote by $\rho_{\hat{\pi}}(\mathbf{n})$ the PAC–MLE obtained using a golden section search on the PAC-likelihood surface associated with the CSD $\hat{\pi}$ and 100 random permutations of the haplotypes in $\mathbf{n}$.

Following LI and STEPHENS (2003), we compute the per-$\mathbf{n}$ error $\text{Err}_{\hat{\pi}}(\mathbf{n}) = \log_2[\rho_{\hat{\pi}}(\mathbf{n})/\rho_0]$, where ρ_0 is the recombination rate under which the $\mathbf{n}$ was generated. Note that $\text{Err}_{\hat{\pi}}(\mathbf{n}) = 0$ indicates that $\rho_{\hat{\pi}}(\mathbf{n}) = \rho_0$; although this is ostensibly a good property, we note here that the true MLE $\hat{\rho}_{ML}(\mathbf{n})$ does not satisfy this property in expectation and may not satisfy it in median. In keeping with our previous empirical results, we believe that a more important comparison is directly between $\rho_{\hat{\pi}}(\mathbf{n})$ and $\hat{\rho}_{ML}(\mathbf{n})$. Unfortunately, such comparisons are difficult for two reasons: First, $\hat{\rho}_{ML}(\mathbf{n})$ can take the values 0 and ∞, making comparisons with $\rho_{\hat{\pi}}(\mathbf{n})$ difficult; and second, $\hat{\rho}_{ML}(\mathbf{n})$ is difficult to compute.

With this caveat in mind, we continue with Li and Stephens' formulation. Treating $\mathbf{n}$ as a random variable, compute the sample median and interquartile range (IQR) of the distribution associated with $\text{Err}_{\hat{\pi}}(\mathbf{n})$. The specific parameter settings used and results are presented in Table 1. Observe that, as the number of loci increases, the IQR generally becomes smaller, indicating that the distribution is becoming more concentrated about the median. In the case that $\theta_0 = 1$ and $\rho_0 = 4$, the results are promising; the approximations $\hat{\pi}_{PS}$, $\hat{\pi}_{PS,1}$, and $\hat{\pi}_{PS,2}$ have medians significantly nearer to 0 than $\hat{\pi}_{FD}$ and $\hat{\pi}_{LS}$. Moreover, this effect becomes more pronounced as the number of loci increases. The results are less clear in the $\theta_0 = 0.01$ and $\rho_0 = 0.1$ case. All of the CSDs demonstrate comparable medians, none particularly close to 0; as the number of loci increases, there appears to be some trend toward a median of 0 for all CSDs. Once again, we urge caution in interpreting these

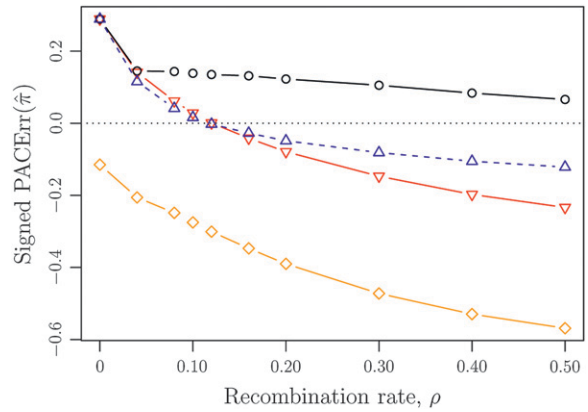

FIGURE 7.—Approximate values of signed $\text{PACErr}(\hat{\pi})$ for $\theta_0 = 0.01$ and $\rho_0 = 0.1$, corresponding to Figure 6b. The correspondence between the symbols and $\hat{\pi}$'s is the same as in previous figures.

TABLE 1

PAC-maximum-likelihood estimate accuracy

	$\theta_0 = 1$, $\rho_0 = 4$						$\theta_0 = 0.01$, $\rho_0 = 0.1$					
	$k = 5$		$k = 7$		$k = 9$		$k = 5$		$k = 7$		$k = 9$	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
$\hat{\pi}_{PS}$	−0.07	3.57	—	—	—	—	−0.74	3.01	—	—	—	—
$\hat{\pi}_{PS,1}$	−0.19	3.58	+0.10	2.05	+0.05	1.64	−0.94	3.14	−0.94	2.14	−0.80	1.55
$\hat{\pi}_{PS,2}$	−0.39	3.75	−0.11	2.17	−0.22	1.79	−0.94	3.10	−0.94	2.14	−0.82	1.58
$\hat{\pi}_{FD}$	−0.79	3.98	−0.80	2.19	−0.96	1.70	−0.99	3.01	−1.00	2.02	−0.87	1.49
$\hat{\pi}_{LS}$	−1.02	4.58	−0.91	2.33	−1.19	1.86	−0.83	3.15	−0.85	1.88	−0.68	1.23

Median and interquartile range (IQR) estimates for the distribution $\mathrm{Err}_{\hat{\pi}}(n) = \log_2[\rho_{\hat{\pi}}(n)/\rho_0]$. Estimates are computed using 250 k-locus 25-haplotype configurations generated from a coalescence simulator using θ_0 and ρ_0.

results, as the nature of the true distribution $\mathrm{Err}_{\pi}(\mathbf{n})$ remains unknown.

DISCUSSION

In this article, we generalized the diffusion-generator approximation technique to derive a novel approximate conditional sampling distribution, $\hat{\pi}_{PS}$, for an arbitrary number of loci and an arbitrary finite-alleles recurrent mutation model. Furthermore, we described a genealogical interpretation for $\hat{\pi}_{PS}$ on the basis of the idea of conditional genealogies. In addition to providing intuition for the mathematical techniques used to derive $\hat{\pi}_{PS}$, the genealogical interpretation motivated us to introduce additional approximations that reduce the asymptotic time complexity of our $\hat{\pi}_{PS}$ from super-exponential in k (the number of loci) to cubic in k. We observed that the approximation of disallowing coalescence in the conditional genealogy $\mathcal{C}_m$ works remarkably well, leading to little loss in accuracy compared with $\hat{\pi}_{PS}$. We note that this is probably because the empirical study we carried out is for the case in which the haplotypes in the conditional sample configuration $\mathbf{m}$ have pairwise disjoint sets of specified alleles. For a more general sample $\mathbf{m}$, we suspect that disallowing coalescence in $\mathcal{C}_m$ may not work as well. Incidentally, note that disallowing coalescence between haplotypes with no overlapping specified alleles is closely related to the so-called sequentially Markov coalescent (McVean and Cardin 2005; Marjoram and Wall 2006; Chen *et al.* 2009), an approximation to the full sequential coalescent formulation introduced by Wiuf and Hein (1999).

In our empirical study, we found that our CSD $\hat{\pi}_{PS}$ and the associated approximations ($\hat{\pi}_{PS,1}$ and $\hat{\pi}_{PS,2}$) are in general more accurate than the previously proposed CSDs. Importantly, this improvement in accuracy gets amplified as the number of loci increases. Moreover, the improvement in CSD accuracy carries over to the PAC framework, for both PAC-likelihood estimation and, to a lesser extent, PAC–MLE estimation. Interestingly, as the mutation rate θ decreases, some improvements in accuracy are attenuated, while others are not. We believe that studying and understanding these effects is an important future research direction.

Approximate CSDs have been fruitfully used in Monte Carlo techniques (*e.g.*, importance sampling) and other approximation strategies (typically via the PAC approximation). In principle, our new CSD may be applied in many of the same situations, potentially providing improved efficiency in the Monte Carlo setting and improved accuracy in the approximation setting. In practice, the details of many algorithms explicitly depend on the CSD used, so we leave as future research adapting such algorithms to the form of $\hat{\pi}_{PS}$. We believe that the work discussed here will have several useful applications in both computational biology and population genetics analysis.

We thank Paul Jenkins for helpful discussions. This research is supported in part by National Institutes of Health grant R00-GM080099, an Alfred P. Sloan Research Fellowship, and a Packard Fellowship for Science and Engineering.

LITERATURE CITED

Chen, G. K., P. Marjoram and J. D. Wall, 2009 Fast and flexible simulation of DNA sequence data. Genome Res. **19:** 136–142.

Crawford, D. C., T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder *et al.*, 2004 Evidence for substantial fine-scale variation in recombination rates across the human genome. Nat. Genet. **36:** 700–706.

Davison, D., J. K. Pritchard and G. Coop, 2009 An approximate likelihood for genetic data under a model with recombination and population splitting. Theor. Popul. Biol. **75**(4): 331–345.

De Iorio, M., and R. C. Griffiths, 2004a Importance sampling on coalescent histories I. Adv. Appl. Probab. **36:** 417–433.

De Iorio, M., and R. C. Griffiths, 2004b Importance sampling on coalescent histories II. Adv. Appl. Probab. **36:** 434–454.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Fearnhead, P., and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rates. J. R. Stat. Soc. B **64:** 657–680.

Fearnhead, P., and N. G. C. Smith, 2005 A novel method with improved power to detect recombination hotspots from polymorphism data reveals multiple hotspots in human genes. Am. J. Hum. Genet. **77:** 781–794.

Gay, J. C., S. Myers and G. McVean, 2007 Estimating meiotic gene conversion rates from population genetic data. Genetics **177:** 881–894.

Griffiths, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19:** 169–186.
Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3**(4): 479–502.
Griffiths, R. C., P. A. Jenkins and Y. S. Song, 2008 Importance sampling and the two-locus model with subdivided population structure. Adv. Appl. Probab. **40**(2): 473–500.
Hellenthal, G., A. Auton and D. Falush, 2008 Inferring human colonization history using a copying model. PLoS Genet. **4:** e1000078.
Howie, B. N., P. Donnelly and J. Marchini, 2009 A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. **5**(6): e1000529.
Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.
Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805–1817.
Jenkins, P. A., and Y. S. Song, 2009 Closed-form two-locus sampling distributions: accuracy and universality. Genetics **183:** 1087–1103.
Jenkins, P. A., and Y. S. Song, 2010 An asymptotic sampling formula for the coalescent with recombination. Ann. Appl. Probab. **20:** 1005–1028.
Johnson, P., and M. Slatkin, 2009 Inference of microbial recombination rates from metagenomic data. PLoS Genet. **5**(10): e1000674.
Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. **13:** 235–248.
Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.
Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.
Li, N., and M. Stephens, 2003 Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics **165:** 2213–2233.
Li, Y., and G. R. Abecasis, 2006 Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am. J. Hum. Genet. **S79:** 2290.
Marchini, J., B. Howie, S. R. Myers, G. McVean and P. Donnelly, 2007 A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. **39**(7): 906–913.
Marjoram, P., and S. Tavaré, 2006 Modern computational approaches for analysing molecular genetic variation data. Nat. Rev. Genet. **7:** 759–770.
Marjoram, P., and J. D. Wall, 2006 Fast "coalescent" simulation. BMC Genet. **7:** 16.
McVean, G., and N. Cardin, 2005 Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B Biol. Sci. **360:** 1387–1393.
McVean, G., P. Awadalla and P. Fearnhead, 2002 A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics **160:** 1231–1241.
McVean, G. A. T., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley *et al.*, 2004 The fine-scale structure of recombination rate variation in the human genome. Science **304:** 581–584.
Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.
Price, A. L., A. Tandon, N. Patterson, K. C. Barnes, N. Rafaels *et al.*, 2009 Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. **5**(6): e1000519.
Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. **78:** 629–644.
Stephens, M., and P. Donnelly, 2000 Inference in molecular population genetics. J. R. Stat. Soc. B **62:** 605–655.
Stephens, M., and P. Scheet, 2005 Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am. J. Hum. Genet. **76**(3): 449–462.
Wang, Y., and B. Rannala, 2008 Bayesian inference of fine-scale recombination rates using population genomic data. Philos. Trans. R. Soc. B **363**(1512): 3921–3930.
Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. **55**(3): 248–259.
Yin, J., M. I. Jordan and Y. S. Song, 2009 Joint estimation of gene conversion rates and mean conversion tract lengths from population SNP data. Bioinformatics **25**(12): i231–i239.

APPENDIX

Proof of Theorem 1. By the componentwise vanishing property (5), for any bounded, twice-differentiable function f with continuous second derivatives,

$$\hat{\mathbb{E}}\left[\sum_{h\in\mathcal{H}} m_h \cdot \mathcal{L}_h \frac{\partial}{\partial x_h} f(\mathbf{X})\right] = \sum_{h\in\mathcal{H}} m_h \cdot \hat{\mathbb{E}}\left[\mathcal{L}_h \frac{\partial}{\partial x_h} f(\mathbf{X})\right] = 0.$$

Setting $f(\mathbf{x}) = q(\mathbf{n} \mid \mathbf{x})$ implies the following relation for $\hat{q}$:

$$\begin{aligned} m\left[(n-1) + \sum_{\ell\in L}\theta_\ell + \sum_{b\in B}\rho_b\right]\hat{q}(\mathbf{n}) = \sum_{h\in\mathcal{H}} m_h \Bigg[& (n_h - 1)\hat{q}(\mathbf{n} - \mathbf{e}_h) \\ & + \sum_{\ell\in L}\theta_\ell \sum_{a\in E_\ell} P^{(\ell)}_{a,h[\ell]}\hat{q}(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{S}^a_\ell(h)}) \\ & + \sum_{b\in B}\rho_b \sum_{h'\in\mathcal{H}} \hat{q}(\mathbf{n} - \mathbf{e}_h + \mathbf{e}_{\mathcal{R}_b(h,h')} + \mathbf{e}_{\mathcal{R}_b(h',h)})\Bigg]. \end{aligned}$$

Substituting $\mathbf{n} \to \mathbf{n} + \mathbf{m}$ and, recalling (4), dividing by $\hat{q}(\mathbf{n})$ produces (6), thereby completing the proof. ■

Proof of Corollary 2. This result follows from Theorem 1. Without loss of generality, let $\mathbf{m} = \mathbf{e}_{g_1} + \ldots + \mathbf{e}_{g_m}$ for $g_1, \ldots, g_m \in \mathcal{G}$ and $\mathbf{n} = (n_h)_{h\in\mathcal{H}}$. Recalling (4) and the appropriate definitions,

$$\sum_{\substack{h_1\in\mathcal{H}\\ h_1\curlywedge g_1}}\cdots\sum_{\substack{h_m\in\mathcal{H}\\ h_m\curlywedge g_m}}\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{h_1}+\ldots+\mathbf{e}_{h_m}\,|\,\mathbf{n})=\frac{1}{\hat{q}(\mathbf{n})}\hat{\mathbb{E}}\left[\left(\prod_{h\in\mathcal{H}}X_h^{n_h}\right)\cdot\sum_{\substack{h_1\in\mathcal{H}\\ h_1\curlywedge g_1}}X_{h_1}\cdots\sum_{\substack{h_m\in\mathcal{H}\\ h_m\curlywedge g_m}}X_{h_m}\right]$$
$$=\frac{1}{\hat{q}(\mathbf{n})}\hat{\mathbb{E}}\left[\left(\prod_{h\in\mathcal{H}}X_h^{n_h}\right)\cdot Y_{g_1}\ldots Y_{g_m}\right]$$
$$=\hat{\pi}_{\mathrm{PS}}(\mathbf{m}\,|\,\mathbf{n}). \tag{A1}$$

Substituting $\mathbf{m}'=\mathbf{e}_{h_1}+\ldots+\mathbf{e}_{h_m}$ for $h_1,\ldots,h_m\in\mathcal{H}$ into (6),

$$m\left[(n+m-1)+\sum_{\ell\in L}\theta_\ell+\sum_{b\in B}\rho_b\right]\hat{\pi}_{\mathrm{PS}}(\mathbf{m}'\,|\,\mathbf{n})$$
$$=\sum_{i=1}^{m}\Bigg[(n_{h_i}+(m'_{h_i}-1))\hat{\pi}_{\mathrm{PS}}(\mathbf{m}'-\mathbf{e}_{h_i}\,|\,\mathbf{n})$$
$$+\sum_{\ell\in L}\theta_\ell\sum_{a\in E_\ell}P^{(\ell)}_{a,h_i[\ell]}\hat{\pi}_{\mathrm{PS}}(\mathbf{m}'-\mathbf{e}_{h_i}+\mathbf{e}_{\mathcal{S}^a_\ell(h_i)}\,|\,\mathbf{n})$$
$$+\sum_{b\in B}\rho_b\sum_{h'\in\mathcal{H}}\hat{\pi}_{\mathrm{PS}}(\mathbf{m}'-\mathbf{e}_{h_i}+\mathbf{e}_{\mathcal{R}_b(h',h_i)}+\mathbf{e}_{\mathcal{R}_b(h',h_i)}\,|\,\mathbf{n})\Bigg]. \tag{A2}$$

Applying (A1) to the left-hand side of (A2) and doing some algebraic manipulation,

$$\sum_{i=1}^{m}\left[(n+m-1)+\sum_{\ell\in L(g_i)}\theta_\ell+\sum_{b\in B(g_i)}\rho_b\right]\hat{\pi}_{\mathrm{PS}}(\mathbf{m}\,|\,\mathbf{n})$$
$$=\sum_{i=1}^{m}\Bigg[\left(\sum_{\substack{h\in\mathcal{H}\\ h\curlywedge g_i}}n_h\right)\hat{\pi}_{\mathrm{PS}}(\mathbf{m}-\mathbf{e}_{g_i}\,|\,\mathbf{n})+\sum_{\substack{g'\in\mathcal{G}\\ g'\curlywedge g_i}}(m_{g'}-\delta_{g_i,g'})\hat{\pi}_{\mathrm{PS}}(\mathbf{m}-\mathbf{e}_{g_i}+\mathbf{e}_{\mathcal{C}(g_i,g')}\,|\,\mathbf{n})$$
$$+\sum_{\ell\in L(g_i)}\theta_\ell\sum_{a\in E_\ell}P^{(\ell)}_{a,g_i[\ell]}\hat{\pi}_{\mathrm{PS}}(\mathbf{m}-\mathbf{e}_{g_i}+\mathbf{e}_{\mathcal{S}^a_\ell(g_i)}\,|\,\mathbf{n})$$
$$+\sum_{b\in B(g_i)}\rho_b\hat{\pi}_{\mathrm{PS}}(\mathbf{m}-\mathbf{e}_{g_i}+\mathbf{e}_{\mathcal{R}^-_b(g_i)}+\mathbf{e}_{\mathcal{R}^+_b(g_i)}\,|\,\mathbf{n})\Bigg].$$

This result is equivalent to (7), completing the proof. ∎

Proof of Proposition 3. Let $\mathbf{n}=(\mathbf{n}_h)_{h\in\mathcal{H}}$ be an observed haplotype configuration. Stephens and Donnelly's CSD $\hat{\pi}_{\mathrm{SD}}$ is formulated by assuming that a new haplotype may be conditionally sampled by choosing a haplotype from $\mathbf{n}$ uniformly at random and mutating the loci using a prescribed scheme dependent on θ_ℓ and $\mathbf{P}^{(\ell)}=(P^{(\ell)}_{a,a'})$ for each locus $\ell\in L$. Letting $h'\in\mathcal{H}$,

$$\hat{\pi}_{\mathrm{SD}}(\mathbf{e}_{h'}\,|\,\mathbf{n})=\sum_{h\in\mathcal{H}}\frac{n_h}{n}\sum_{\mathbf{s}\in\mathbb{N}^m}\binom{s}{\mathbf{s}}F(h,h',\mathbf{s}), \tag{A3}$$

where $\mathbf{s}=(s_1,\ldots,s_m)$ denotes the number of mutations at each locus, $s=\sum_{i=1}^m s_i$, $\binom{s}{\mathbf{s}}$ is the multinomial coefficient, and $F(h,h',s)$ is the probability of h mutating to h' with s_ℓ mutations at each locus $\ell\in L$,

$$F(h,h',\mathbf{s})=\frac{n}{n+\Theta}\prod_{\ell\in L}\left(\frac{\theta_\ell}{n+\Theta}\right)^{s_\ell}\left[\left(\mathbf{P}^{(\ell)}\right)^{s_\ell}\right]_{h[\ell],h'[\ell]},$$

where $\Theta = \sum_{\ell \in L} \theta_\ell$. We show that $\hat{\pi}_{\mathrm{SD}}(\mathbf{e}_{h'} \,|\, \mathbf{n})$ obeys the same recursion as $\hat{\pi}_{\mathrm{PS}}(\mathbf{e}_{h'} \,|\, \mathbf{n})$. By removing the summand with $\mathbf{s} = \mathbf{0} \in \mathbb{N}^m$ in Equation A3, we obtain

$$\hat{\pi}_{\mathrm{SD}}(\mathbf{e}_{h'} \,|\, \mathbf{n}) = \sum_{h \in \mathcal{H}} \frac{n_h}{n} \left[F(h, h', \mathbf{0}) + \sum_{\mathbf{s} \in \mathbb{N}^m} \binom{s}{\mathbf{s}} \sum_{\ell \in L} F(h, h', \mathbf{s} + \mathbf{e}_\ell) \right]. \tag{A4}$$

Additionally, we have that $F(h, h', \mathbf{0}) = \delta_{h,h'}/(n + \Theta)$, and

$$F(h, h', \mathbf{s} + \mathbf{e}_\ell) = \frac{\theta_\ell}{n + \Theta} \sum_{a \in E_\ell} P^{(\ell)}_{a,h'[\ell]} \cdot F(h, \mathcal{S}^a_\ell(h'), \mathbf{s}).$$

Substituting these identities into (A4) yields the recursion

$$\hat{\pi}_{\mathrm{SD}}(\mathbf{e}_{h'} \,|\, \mathbf{n}) = \frac{1}{n + \Theta} \left(n_{h'} + \sum_{\ell \in L} \theta_\ell \sum_{a \in E_\ell} P^{(\ell)}_{a,h'[\ell]} \hat{\pi}_{\mathrm{SD}}(\mathbf{e}_{\mathcal{S}^a_\ell(h')} \,|\, \mathbf{n}) \right),$$

which is identical to the recursion (9) for $\hat{\pi}_{\mathrm{PS}}(h \,|\, \mathbf{n})$, thereby proving the proposition. ∎

Proof of Proposition 4. Define $B(\mathbf{m}) = \sum_{g \in \mathcal{G}} m_g \cdot |B(g)|$ as the total number of *valid* breakpoints in $\mathbf{m}$. Using (7) in the limit that $\rho \to \infty$ and assuming $B(\mathbf{m}) > 0$,

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{m} \,|\, \mathbf{n}) = \frac{1}{B(\mathbf{m})} \sum_{g \in \mathcal{G}} n_g \sum_{b \in B(g)} \hat{\pi}_{\mathrm{PS}}(\mathbf{m} - \mathbf{e}_g + \mathbf{e}_{\mathcal{R}^-_b(g)} + \mathbf{e}_{\mathcal{R}^+_b(g)} \,|\, \mathbf{n}).$$

Repeated application of this equation yields the key identity

$$\hat{\pi}_{\mathrm{PS}}(\mathbf{m} \,|\, \mathbf{n}) = \hat{\pi}_{\mathrm{PS}}(\mathbf{m}^* \,|\, \mathbf{n}), \tag{A5}$$

where $\mathbf{m}^*$ is derived from $\mathbf{m}$ by recombination at every possible breakpoint. More precisely, define $u_{\ell,a} \in \mathcal{G}$ to be the haplotype with allele $a \in E_\ell$ at locus $\ell \in L$ and $\cdot$ elsewhere. Then

$$\mathbf{m}^* = \sum_{\ell \in L} \mathbf{m}_\ell^*, \text{ where } \mathbf{m}_\ell^* = \sum_{a \in E_\ell} (\mathbf{m}[\ell])_a \bullet \mathbf{e}_{u_{\ell,a}}.$$

Observing that $B(\mathbf{m}^*) = 0$, we may apply (A5) to (7) to obtain

$$\begin{aligned}
&\sum_{\ell \in L} (|\,\mathbf{m}[\ell]\,| \cdot (n + \theta_\ell)) \hat{\pi}_{\mathrm{PS}}(\mathbf{m}^* \,|\, \mathbf{n}) \\
&= \sum_{\ell \in L} \sum_{a \in E_\ell} (\mathbf{m}[\ell])_a \cdot \left[((\mathbf{n}[\ell])_a + ((\mathbf{m}[\ell])_a - 1)) \hat{\pi}_{\mathrm{PS}} \left((\mathbf{m}_\ell^* - \mathbf{e}_{u_{\ell,a}}) + \sum_{\ell' \in L : \ell' \neq \ell'} \mathbf{m}_{\ell'}^* \,|\, \mathbf{n} \right) \right. \\
&\qquad \left. + \theta_\ell \sum_{a' \in E_\ell} P^{(\ell)}_{a',a} \hat{\pi}_{\mathrm{PS}} \left((\mathbf{m}_\ell^* - \mathbf{e}_{u_{\ell,a}} + \mathbf{e}_{u_{\ell,a'}}) + \sum_{\ell' \in L : \ell' \neq \ell'} \mathbf{m}_{\ell'}^* \,|\, \mathbf{n} \right) \right].
\end{aligned} \tag{A6}$$

Observe that (A6) is a sum of independent recursions, each for a particular locus $\ell \in L$. It is thus easily verified that the recursion has solution

$$\hat{\pi}_{\mathrm{PS}}(m^* \,|\, n) = \prod_{\ell \in L} \hat{\pi}_{\mathrm{PS}}(m_\ell^* \,|\, n) = \prod_{\ell \in L} \hat{\pi}_{\mathrm{PS}}(\mathbf{m}[\ell] \,|\, \mathbf{n}[\ell]).$$

In conjunction with (A5), this produces the desired result. ∎

Proof of Proposition 5. As described, the $\hat{\pi}_{\mathrm{PS},1}$ is the approximate CSD obtained by removing the second term on the right-hand side of (7) and renormalizing the left-hand side. Writing the resulting recursion for $\mathbf{m} = \mathbf{e}_{g_1} + \ldots + \mathbf{e}_{g_m}$,

$$\sum_{i=1}^{m}\left[n + \sum_{\ell\in L(g_i)}\theta_\ell + \sum_{b\in B(g_i)}\rho_b\right]\hat{\pi}_{\mathrm{PS},1}(\mathbf{m}\,|\,\mathbf{n}) = \sum_{i=1}^{m}\Bigg[\left(\sum_{h\in\mathcal{H}:h\curlywedge g_i} n_h\right)\hat{\pi}_{\mathrm{PS},1}(\mathbf{m}-\mathbf{e}_{g_i}\,|\,\mathbf{n}) + \sum_{\ell\in L(g_i)}\theta_\ell\sum_{a\in E_\ell} P^{(\ell)}_{a,g_i[\ell]}\hat{\pi}_{\mathrm{PS},1}(\mathbf{m}-\mathbf{e}_{g_i}+\mathbf{e}_{\mathcal{S}^a_\ell(g_i)}\mid\mathbf{n}) + \sum_{b\in B(g_i)}\rho_b\hat{\pi}_{\mathrm{PS},1}(\mathbf{m}-\mathbf{e}_{g_i}+\mathbf{e}_{\mathcal{R}^-_b(g_i)}+\mathbf{e}_{\mathcal{R}^+_b(g_i)}\mid\mathbf{n})\Bigg]. \tag{A7}$$

Observe that (A7) is a sum of independent recursions, each for a particular haplotype $g_i \in \mathcal{H}$. It is thus easily verified that the recursion has solution

$$\hat{\pi}_{\mathrm{PS},1}(\mathbf{m}\,|\,\mathbf{n}) = \prod_{i=1}^{m}\hat{\pi}_{\mathrm{PS},1}(\mathbf{e}_{g_i}\,|\,\mathbf{n}),$$

which is our desired result. ■