A renewal theory approach to IBD sharing

Shai Carmi [a,*], Peter R. Wilton [b], John Wakeley [b], Itsik Pe'er [a]

[a] *Department of Computer Science, Columbia University, New York, NY, 10027, USA*
[b] *Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA*

ARTICLE INFO

ABSTRACT

A long genomic segment inherited by a pair of individuals from a single, recent common ancestor is said to be *identical-by-descent* (IBD). Shared IBD segments have numerous applications in genetics, from demographic inference to phasing, imputation, pedigree reconstruction, and disease mapping. Here, we provide a theoretical analysis of IBD sharing under Markovian approximations of the coalescent with recombination. We describe a general framework for the IBD process along the chromosome under the Markovian models (SMC/SMC'), as well as introduce and justify a new model, which we term the *renewal approximation*, under which lengths of successive segments are independent. Then, considering the infinite-chromosome limit of the IBD process, we recover previous results (for SMC) and derive new results (for SMC') for the mean number of shared segments longer than a cutoff and the fraction of the chromosome found in such segments. We then use renewal theory to derive an expression (in Laplace space) for the distribution of the number of shared segments and demonstrate implications for demographic inference. We also compute (again, in Laplace space) the distribution of the fraction of the chromosome in shared segments, from which we obtain explicit expressions for the first two moments. Finally, we generalize all results to populations with a variable effective size.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

IBD sharing of a genomic segment between a pair of individuals is traditionally defined in terms of recent co-ancestry, no more remote than some time depth t (Thompson, 2013). In population samples, the time of the common ancestor is unknown, and in practice, IBD segments are often identified as long stretches that are nearly or fully identical-by-state (IBS), to an extent distinguishable from population-level LD. The decision whether a segment is called IBD is either rule-based (e.g., using a certain length cutoff) or model-based, using an underlying hidden Markov model for the IBD state (Thompson, 2013). In this paper, we define an IBD segment shared between two chromosomes as the maximal sequence over which the chromosomes have the same most recent common ancestor (MRCA). Recent mutations (or genotyping errors) separating the two sequences do not disqualify the segment from being IBD. On the other hand, we require the segment to be longer than an (arbitrary) cutoff m. This definition enables a theoretical treatment, while largely capturing the way in which some methods (and, for sufficiently large m, virtually all methods) discover IBD segments in real data.

Much attention has recently been devoted to efficient algorithms for IBD detection in large samples (e.g., Purcell et al., 2007; Gusev et al., 2009; Browning and Browning, 2011; Brown et al., 2012; Browning and Browning, 2013a, to give a few examples). Detected segments have found numerous applications, for example, characterization of relationships between populations (Atzmon et al., 2010; Bray et al., 2010; Moorjani et al., 2013; Gauvin et al., 2014; Botigué et al., 2013; Ralph and Coop, 2013), detection of positive selection (Han and Abney, 2013), estimation of heritability (Browning and Browning, 2013b), mapping haplotypes associated with a trait (Gusev et al., 2011; Browning and Thompson, 2012; Lin et al., 2013), phasing and imputation (Kong et al., 2008; Palin et al., 2011), and pedigree reconstruction (Huff et al., 2011; Henn et al., 2012). See Browning and Browning (2012) and Thompson (2013) for up-to-date reviews.

In parallel, theory has been developed for the expected amount of IBD sharing in model populations, with implications for demographic inference. Palamara et al. (2012) and Palamara and Pe'er (2013) computed, under the coalescent and for complex demographies, the moments of the fraction of the chromosome found in shared segments of a given length. Palamara et al. (2012) and Carmi et al. (2013) then approximated the distribution of this quantity, assuming a Poisson distribution for the number of segments (see also Huff et al., 2011). Ralph and Coop (2013) computed the

expected number of shared segments of a certain length given an arbitrary demographic history. However, certain theoretical problems of interest have remained open.

Here, we introduce a general framework for the analysis of the IBD process along the chromosome, based on a renewal approximation. Renewal theory is the study of processes in which events are separated by independent waiting times, and where each waiting period or event may be associated with a value (Karlin and Taylor, 1975). Under certain conditions, consecutive shared segments along the chromosome can be approximated as independent. Then, interpreting segments with shared ancestry as waiting times, renewal theory can be applied to compute, for example, the distribution of the number of segments of a certain length.

A renewal approach to the IBD process has been considered in the past (e.g., Stam, 1980; Chapman and Thompson, 2003, with initial contributions already by Fisher, 1954), in a model where the population has been recently founded by individuals of heterogeneous genetic types. Alternatively, in those works, IBD is defined with respect to a given time depth (Thompson, 2013). The IBD segment lengths were either assumed exponential or fitted. In contrast, we consider a model that can be applied without reference to a particular time point. In our model, two chromosomes can trace their common ancestor, at each locus, to any time in the past, and IBD segments are defined with respect to a length cutoff.

According to our renewal approximation for a pair of chromosomes, the time to the common ancestor is drawn, at a recombination event, independently of the previous time and from a position-independent stationary distribution. The distribution has been derived for the pairwise Sequentially Markov Coalescent (SMC) by Li and Durbin (2011), and we derive it here for the more accurate, yet tractable SMC' model (Marjoram and Wall, 2006). Under this approximation, the distribution of segment lengths emerges naturally. Using renewal theory, we are then able to derive new results, such as the distribution of the number of shared segments, as well as recover previous results as special cases.

Our results are organized as follows. In Section 2, we introduce the renewal approximation in the context of successively simplified approximations of the coalescent with recombination. We then describe the IBD process under the different models and present numerical evidence to justify the renewal approach. In Section 3, we show how simple quantities, such as the mean number of shared segments and the mean fraction of the chromosome in shared segments, emerge naturally from our definition of the IBD process by taking the infinite-chromosome limit. Specifically, we recover previous results for SMC and obtain new results for SMC'. In Section 4, we derive results for finite chromosomes. Specifically, we derive an expression, in Laplace space, for the distribution of the number of shared segments and consider implications for demographic inference. Additionally, we derive, again in Laplace space, the distribution of the fraction of the chromosome found in shared segments, from which we obtain explicit expressions for the first two moments, recovering and extending previous results. Finally, in Section 5, we generalize our results to populations with variable size. We summarize and discuss the results in Section 6.

2. The IBD process

2.1. Overview of the coalescent with recombination and its Markovian approximations

We consider a sample of two chromosomes of length L (Morgans) in a population of a constant effective size N (haploid chromosomes) and with recombination modeled as a Poisson process along the chromosome. The ancestral process can be described by the coalescent with recombination (Hudson, 1983; Griffiths and
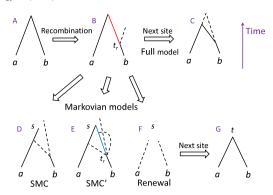

Fig. 1. An illustration of the coalescent with recombination for two chromosomes, and the associated Markovian approximations. Part **A** shows the coalescent tree at a random site. The two extant chromosomes are denoted a and b. Part **B** is indicating a recombination event occurring at time t_r. The old branch connecting the breakpoint and the MRCA is colored red, and the branching lineage is shown as a dashed line. Under the full model of the coalescent with recombination (the ARG; Wiuf and Hein, 1999; **C**), the new lineage can coalesce with any branch in the existing tree (in this example, earlier than the previous TMRCA), and both the old lineage (which is not ancestral to the sample anymore) and the new lineage are carried over to the next site. The 'marginal' tree at the new site is shown in solid lines; the remainder of the ARG is in dashed lines. The Markovian approximations are presented in parts **D–G**, where the current TMRCA is denoted as s and the new as t. In SMC (McVean and Cardin, 2005; **D**), the old branch (red in **B**) is deleted, and the branching lineage can coalesce only with the lineage corresponding to the other chromosome (either earlier or later than the previous TMRCA; corresponding to the two dashed lines). In SMC' (Marjoram and Wall, 2006; **E**), the branching lineage can coalesce with the old branch (blue), but any non-ancestral material is deleted once the new tree is formed. Under the renewal approximation (**F**), the new tree height is drawn independently of the previous tree height. In all Markovian approximations, the new tree (**G**) contains only the lineages ancestral to the sample at that position. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Marjoram, 1997). In that model, looking backwards in time, lineages can either coalesce (at rate 1 per pair of lineages, when the time is scaled by N) or recombine at a random position along the chromosome (split into two, at rate $\rho = 2Nr$, where r is the recombination probability per generation). The resulting structure is called the *ancestral recombination graph* (ARG). Wiuf and Hein (1999) described an alternative but equivalent formulation, where the ARG is obtained by walking along the chromosome. In that model, a coalescent tree is first formed at the leftmost end of the chromosome (Fig. 1A). Recombination then occurs at a genetic distance distributed exponentially with rate equal to the total branch length of the tree; the position of the breakpoint (t_r) is randomly and uniformly distributed along the tree (Fig. 1B). The branching lineage then coalesces with any of the existing branches of the ARG, and the process is repeated until reaching the end of the chromosome (Fig. 1C). The model is non-Markovian, in the sense that the tree formed at a given position depends on all preceding trees.

McVean and Cardin (2005) proposed a Markovian approximation to the coalescent with recombination (the *Sequentially Markov Coalescent*, or SMC). At each recombination event in SMC, the branch leading from the breakpoint to the most recent common ancestor (MRCA) is deleted, and the branching lineage is allowed to coalesce only with the lineage ancestral to the other individual (Fig. 1D,G). Once the MRCA is reached, the process is continued with the newly formed tree. Marjoram and Wall (2006) suggested a more accurate approximation, called SMC', in which the branching lineage is allowed to coalesce with the branch it had split from, but once the tree has formed, any branch not ancestral to the sample is again deleted (Fig. 1E,G). See Hobolth and Jensen (2014) for the joint distribution of tree heights for two sequences at two loci under the ARG and the Markovian approximations.

We propose the *renewal approximation*, which is a further simplification of SMC. According to our approximation, at a recombination event, the new tree height is drawn, *independently* of

the previous tree height, from the stationary distribution of tree heights under SMC (Fig. 1F,G). The stationary distribution was derived by Li and Durbin (2011) (see the next section). While the independence assumption is strong, the fact that we use the SMC stationary distribution guarantees that for sufficiently long sequences (see simulations in Section 2.5), the statistical properties of SMC and the renewal process are similar.

In the following subsections, we define the IBD process under the three models: SMC, renewal, and SMC' (Tables 1–3, respectively).

2.2. The IBD process under SMC

Recently, Li and Durbin (2011) derived the probability density function (PDF) of the tree height for a pair of chromosomes (equivalently, time to MRCA or TMRCA; and scaled by N) at a recombination site, given the TMRCA of the preceding tree. The result is given in their supplementary Eq. (6),

$$q_{\text{SMC}}(t|s) = \begin{cases} \frac{1}{s}\left(1 - e^{-t}\right) & t < s, \\ \frac{1}{s}e^{-(t-s)}\left(1 - e^{-s}\right) & t > s, \end{cases} \quad (1)$$

where s and t are the previous and new TMRCA, respectively. Note that $t \neq s$ by definition and that $q_{\text{SMC}}(t|s)$ is normalized. At a recombination site, and for a given new tree height t, the sequence length to the next recombination event is distributed exponentially with rate $2Nt$, the total branch length of the tree (in generations; Wiuf and Hein (1999)). The sequence between recombination sites is a *shared segment*, because the common ancestor of the two chromosomes is fixed throughout the segment. In SMC, the MRCA necessarily changes at recombination sites; therefore, segments are terminated by recombination events. With these preliminaries, and imposing a minimal segment length cutoff, m, we define in Table 1 the IBD process along the chromosome (see also Fig. 2).

Table 1 The IBD process under SMC

1: **Initialize**
2: $\quad x \leftarrow 0 \quad \triangleright$ The position along the chromosome
3: $\quad n_m \leftarrow 0 \quad \triangleright$ The number of shared segments longer than m
4: $\quad f_m \leftarrow 0 \quad \triangleright$ The fraction of the chromosome in shared segments longer than m
5: $\quad$ Draw TMRCA: $t \sim \text{Exp}(1)$
6: **while** $x < L$
7: $\quad$ Draw segment length: $\ell \sim \text{Exp}(2Nt)$
8: $\quad$ **if** $(x + \ell) > L \quad \triangleright$ If the new position exceeds the chromosome length
9: $\quad\quad \ell \leftarrow (L - x)$
10: $\quad$ **if** $\ell > m \quad \triangleright$ The segment is longer than the cutoff
11: $\quad\quad n_m \leftarrow n_m + 1$
12: $\quad\quad f_m \leftarrow f_m + \ell/L$
13: $\quad s \leftarrow t$
14: $\quad$ Draw new TMRCA t with PDF $q_{\text{SMC}}(t|s)$ (Eq. (1))
15: $\quad x \leftarrow x + \ell$

Steps 8 and 9 are needed in case the new position exceeds the chromosome length. In simulations, step 14 is implemented by drawing a random recombination time, t_r, uniform in $[0, s]$, and then a random coalescence time t_c, exponential with rate 1. The new TMRCA is then set to $t \leftarrow t_r + t_c$ (Fig. 1D).

2.3. The IBD process under the renewal approximation to SMC

Eq. (1) for $q_{\text{SMC}}(t|s)$ can be interpreted as the transition probability for a Markov chain whose states are the tree heights
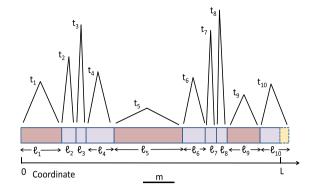

Fig. 2. An illustration of the IBD process along the chromosome under SMC. Segments are broken by recombination events (vertical bars). The TMRCA is shown on top of each segment. Given a TMRCA t_i at segment i, the segment length, ℓ_i, is distributed exponentially with rate $2Nt_i$, and the TMRCA at the next segment, t_{i+1}, is distributed according to Eq. (1). The minimal segment length, m, is shown as a horizontal bar under the chromosome. Segments longer than m are shown in dark pink. In this example, there are three such segments; hence $n_m = 3$ and the fraction of the chromosome in shared segments is $f_m = (\ell_1 + \ell_5 + \ell_9)/L$. Segments shorter than m are in light pink. The last segment exceeds the chromosome length; the excess length (yellow) is ignored. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

at consecutive recombination sites. Li and Durbin (2011), who derived Eq. (1), further computed the stationary distribution of the chain,

$$\pi_{\infty}^{\text{SMC}}(t) = te^{-t}. \quad (2)$$

Note that this stationary distribution is not the same as the 'marginal' coalescence distribution, $P_c(t) = e^{-t}$, which would apply to the tree height at a pre-specified site, such as the end of a chromosome (Wiuf and Hein, 1999), or to a randomly chosen site. In fact, $\pi_{\infty}^{\text{SMC}}(t)$ is identical to the distribution at a site conditional on a recombination event having occurred at that site when the recombination rate per site is very small. It thus has mean equal to 2 (Griffiths and Marjoram, 1996, Eq. (9)), as for example is the case for tree heights around rare insertions in the human genome (Huff et al., 2010). In other words, $\pi_{\infty}^{\text{SMC}}(t)$, may be interpreted as the PDF of the TMRCA of a randomly chosen segment (rather than site).

To test the convergence to the stationary distribution, we numerically computed the PDFs of successive tree heights, as follows,

$$\pi_1^{\text{SMC}}(t) = e^{-t},$$
$$\pi_{n+1}^{\text{SMC}}(t) = \int_0^{\infty} q_{\text{SMC}}(t|s)\pi_n^{\text{SMC}}(s)ds; \quad n \geq 1. \quad (3)$$

The resulting PDFs for the first 10 trees are shown in Fig. 3, demonstrating fast convergence to the stationary PDF (Eq. (2)). For typical (human) parameters ($N \approx 10^4$, $L \approx 1$ Morgan), the average number of recombination events along the chromosome is $2NL \sim 10^4 \gg 1$ (Griffiths and Marjoram, 1997). Therefore, the vast majority of trees are expected to have the stationary PDF.

Using the stationary PDF, segment lengths are therefore distributed as (see also Li and Durbin, 2011; Palamara et al., 2012)

$$\psi_{\text{SMC}}(\ell) = \int_0^{\infty} \pi_{\infty}^{\text{SMC}}(t) \cdot 2Nte^{-2Nt\ell}dt = \frac{4N}{(1 + 2N\ell)^3}. \quad (4)$$

The mean segment length is $\langle\ell\rangle_{\text{SMC}} = 1/2N$, but no higher moments exist. The distribution of $\rho = 2N\ell$, the scaled recombination rate, is $\psi(\rho) = 2/(1 + \rho)^3$, which is, as expected, independent of N (a property that holds generally; see Section 5).

Having the distribution of segment lengths, we can now invoke the assumption of independence between successive segments and define the IBD process in the renewal approximation (Table
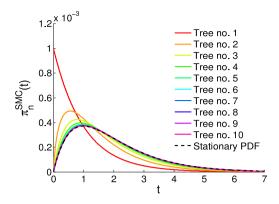
Fig. 3. Convergence of the distribution of tree heights in the SMC model. The first tree is distributed as e^{-t}, according to the standard coalescent. Subsequent trees are distributed according to Eqs. (1) and (3). The integrals were solved numerically. The stationary PDF (dashed line; Eq. (2)) is reached quickly.

2). To generate numbers from $\psi_{SMC}(\ell)$, we used the transformation method: let u be uniform in [0, 1]; we set $\ell = \left(1 - \sqrt{u}\right) / \left(2N\sqrt{u}\right)$.

Table 2 The IBD process under the renewal approximation to SMC

1: **Initialize**
2:     $x \leftarrow 0$ ▷ The position along the chromosome
3:     $n_m \leftarrow 0$ ▷ The number of shared segments longer than m
4:     $f_m \leftarrow 0$ ▷ The fraction of the chromosome in shared segments longer than m
5: **while** $x < L$
6:     Draw segment length ℓ with PDF $\psi_{SMC}(\ell)$ (Eq. (4))
7:     **if** $(x + \ell) > L$
8:       $\ell \leftarrow (L - x)$
9:     **if** $\ell > m$
10:      $n_m \leftarrow n_m + 1$
11:      $f_m \leftarrow f_m + \ell/L$
12:     $x \leftarrow x + \ell$

2.4. The IBD process under SMC'

In SMC', the PDF of the new TMRCA, t, given the previous TMRCA, s, is given by (see also Zheng et al. (2014))

$$q_{SMC'}(t|s) = \begin{cases} \int_0^s \frac{1}{s} \left[\int_{t_r}^s e^{-2(t_c - t_r)} dt_c \right] dt_r & t = s, \\ \int_0^t \frac{1}{s} e^{-2(t - t_r)} dt_r & t < s, \\ \left[\int_0^s \frac{1}{s} e^{-2(s - t_r)} dt_r \right] e^{-(t-s)} & t > s. \end{cases} \quad (5)$$

To understand Eq. (5), consider how the new TMRCA, t, is drawn in simulations. First, a random recombination time, t_r, is drawn uniformly in $[0, s]$, as in SMC. But then, the random coalescence time, t_c, is drawn from an exponential distribution with rate 2, since the branching lineage can coalesce with either the other chromosome or the lineage it had branched from (Fig. 1E). If $t_r + t_c < s$, the new TMRCA is set to either $t \leftarrow s$ (coalescence with the lineage it had branched from) or $t \leftarrow t_r + t_c$ (coalescence with the other chromosome) with probability 1/2 each. If $t_r + t_c > s$, a new coalescence time, τ_c, is drawn from an exponential distribution with rate 1 (since after time s, there is only one other lineage), and the new TMRCA is set to $t \leftarrow s + \tau_c$. The upper limit of the integral for $t < s$ is t, not s, since the recombination time, t_r, cannot be greater than the new coalescence time, t. For the case $t = s$, the density is implicitly assumed to be multiplied by Dirac's delta function ($\delta(t - s)$), omitted for notational simplicity. The integrals

in Eq. (5) can be solved, yielding

$$q_{SMC'}(t|s) = \begin{cases} \dfrac{2t + e^{-2t} - 1}{4t} & t = s, \\ \dfrac{1 - e^{-2t}}{2s} & t < s, \\ \dfrac{e^{-(t-s)} - e^{-(t+s)}}{2s} & t > s. \end{cases} \quad (6)$$

Note that $q_{SMC'}(t|s)$ is normalized. Curiously, the stationary distribution of the chain is $\pi_\infty^{SMC'}(t) = te^{-t}$, exactly as in SMC (Eq. (2)). This can be proved by validating the detailed balance equation, $\pi_\infty^{SMC'}(t)q_{SMC'}(s|t) = \pi_\infty^{SMC'}(s)q_{SMC'}(t|s)$, which also shows that SMC' is reversible (Zheng et al., 2014).

To define the IBD process (Table 3), we note that in the case $t = s$, the common ancestor of the two chromosomes does not change, and therefore, the shared segment extends until (at least) the next recombination event.

Table 3 The IBD process under SMC'

1: **Initialize**
2:     As in Table 1
3: **while** $x < L$
4:     $\ell \leftarrow 0$ ▷ The current *total* segment length
5:     **repeat**
6:        ▷ Draw distance to next recombination; not necessarily a new segment
7:       Draw $\ell_0 \sim \text{Exp}(2Nt)$
8:       $\ell \leftarrow \ell + \ell_0$
9:       $s \leftarrow t$
10:      Draw new TMRCA t with PDF $q_{SMC'}(t|s)$ (Eq. (6))
11:     **until** $t \neq s$
12:     **if** $(x + \ell) > L$
13:       $\ell \leftarrow (L - x)$
14:     **if** $\ell > m$
15:      $n_m \leftarrow n_m + 1$
16:      $f_m \leftarrow f_m + \ell/L$
17:     $x \leftarrow x + \ell$

We now derive the stationary distribution of segment lengths. Given the TMRCA t at the beginning of a segment, the rate at which the segment terminates is the product of the recombination rate ($2Nt$) and the probability that the segment does not extend beyond the recombination site ($1 - q_{SMC'}(t|t)$). Therefore, given t, segment lengths are exponential with rate

$$\lambda(t) = 2Nt[1 - q_{SMC'}(t|t)] = \frac{N}{2}\left(2t + 1 - e^{-2t}\right). \quad (7)$$

Note that this also implies that for two loci distance ℓ apart, and given t at the left locus, the probability of the right TMRCA to remain t is $\exp[-\lambda(t)\ell] = \exp\{-\rho t[1 - q_{SMC'}(t|t)]\}$, as in the small ρ limit of Eq. (30) in Harris and Nielsen (2013).

To obtain the unconditional distribution of segment lengths, we cannot use $\pi_\infty^{SMC'}(t)$, because we need the distribution of tree heights at *segments ends*, not at recombination sites. We therefore define a new Markov chain with transition probability

$$q_{SMC',seg}(t|s) = \frac{q_{SMC'}(t|s)}{1 - q_{SMC'}(s|s)} = \frac{q_{SMC'}(t|s)}{1 - \frac{2s + e^{-2s} - 1}{4s}}, \quad (8)$$

which is the *conditional* probability density of the new tree height, given that it has changed (i.e., a new segment began). By construction, the stationary distribution of the chain, $\pi_\infty^{SMC',seg}(t)$, is the desired distribution of tree heights at the beginning of
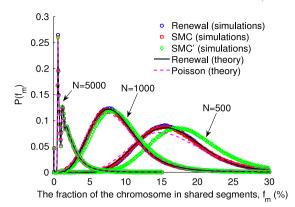
Fig. 4. The distribution of the fraction of the chromosome found in shared segments longer than m, f_m. We simulated the IBD process for three values of the population size ($N = 500, 1000, 5000$), for $L = 2$ and $m = 0.01$ (Morgans), for SMC (the process defined in Table 1, Section 2.2), the renewal approximation (Table 2, Section 2.3), and SMC' (Table 3, Section 2.4), and for 10^6 realizations for each setting. The distribution for $N = 5000$ was divided by 3 for visibility. For all population sizes, SMC and the renewal approximation produced identical results, which also agree well with the renewal theory result (numerical inversion (Brančík, 2011) of Eq. (B.3)). SMC' and the Poisson approximation (Eq. (47)) deviate from SMC/renewal, increasingly for smaller values of N. The fluctuations for $N = 5000$ are due to the sharing of exactly 0, 1, 2, … segments of length very close to m, and were previously described (Carmi et al., 2013).

segments. It is easy to verify by detailed balance that $\pi_\infty^{\text{SMC',seg}}(t) \propto te^{-t}[1 - q_{\text{SMC'}}(t|t)] \propto e^{-t}\lambda(t)$, and then, by normalization,

$$\pi_\infty^{\text{SMC',seg}}(t) = \frac{e^{-t}\lambda(t)}{\int_0^\infty e^{-t'}\lambda(t')dt'} = \frac{3}{8}e^{-t}\left(2t + 1 - e^{-2t}\right). \qquad (9)$$

To obtain the distribution of segment lengths, $\psi_{\text{SMC'}}(\ell)$, we integrate over all t (as in Eq. (4)),

$$\psi_{\text{SMC'}}(\ell) = \int_0^\infty \pi_\infty^{\text{SMC',seg}}(t)\lambda(t)e^{-\lambda(t)\ell}dt$$
$$= \frac{\int_0^\infty \lambda^2(t)e^{-t-\lambda(t)\ell}dt}{\int_0^\infty e^{-t}\lambda(t)dt}. \qquad (10)$$

The integrals in Eq. (10) can be solved in terms of special functions; the final expression is given in Appendix A (Eq. (A.1)). Note that setting $\lambda(t) = 2Nt$ (i.e., setting the probability of $t = s$ to zero) reduces Eq. (10) to the SMC distribution (Eq. (4)). Using the representation of Eq. (10), it is easy to see that $\psi_{\text{SMC'}}(\ell)$ is normalized and that the mean segment length is

$$\langle\ell\rangle_{\text{SMC'}} = \frac{1}{\int_0^\infty e^{-t}\lambda(t)dt} = \frac{3}{4N}. \qquad (11)$$

Segments in SMC' are, by definition, longer than in SMC, and in SMC, $\psi_{\text{SMC}}(\ell)$ had no moments higher than the first. Therefore, $\psi_{\text{SMC'}}(\ell)$ also has no second or higher moments.

It is possible, using Eq. (10), to define a renewal process for SMC' analogous to the process defined in Table 2. However, with the exception of the infinite-chromosome results (Section 3), we do not further investigate the properties of such a model.

2.5. Simulations

To demonstrate the IBD process under SMC and SMC', as well as provide empirical justification to the renewal approximation, we show simulation results for the distribution of the fraction of the chromosome found in shared segments longer than m, $P(f_m)$, (Fig. 4) and the distribution of segment lengths, $\psi(\ell)$ (Fig. 5). Simulations were performed precisely as described in Tables 1–3 above. For all values of N tested, simulation results for $P(f_m)$ were identical between SMC and its renewal approximation. For small values
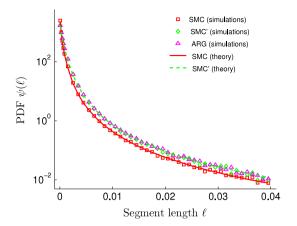
Fig. 5. The distribution of segment lengths, $\psi(\ell)$, under SMC, SMC', and the ARG. Simulations for SMC and SMC' were as described in Fig. 4, but with $N = 1000$ and $L = 0.5$ (Morgan). ARG simulations were performed in *ms*, by outputting the marginal trees and extracting segment lengths. We ran 5000 realizations for each model. Theory for SMC is from Eq. (4) and theory for SMC' is from Eq. (10) (equivalently (A.1)). Interestingly, simulation results for the ARG are indistinguishable from those of SMC'.

of N (or more precisely, as $1/2N$, the average distance between recombination sites, approaches m), there is more sharing in SMC' than in SMC/renewal. This is because in SMC', short segments may extend beyond the first recombination event, and by that exceed the length cutoff. Simulation results for the distribution of segment lengths in SMC and SMC' (Fig. 5) agree well with Eqs. (4) and (10), respectively. As expected, the SMC' distribution has a heavier tail than in SMC and interestingly, is indistinguishable from that of the ARG, reinforcing the importance of the SMC' model.

3. The infinite-chromosome limit of the IBD process

In this section, we derive the mean number of shared segments and the mean fraction of the chromosome in shared segments at the infinite-chromosome limit, under the renewal approximation to SMC and SMC'. Let us first derive some general, model-independent results. Given a segment length distribution $\psi(\ell)$ and using the elementary renewal theorem (Karlin and Taylor, 1975, Theorem 4.2), the mean total number of segments (of any length) for $L \to \infty$ is

$$\langle n_0 \rangle = \frac{L}{\langle\ell\rangle} = \frac{L}{\int_0^\infty \ell\psi(\ell)d\ell}. \qquad (12)$$

Using the elementary renewal theorem for reward processes (Karlin and Taylor, 1975, Chapter 5, Section 7.C.II), the mean number of segments longer than m is, for $L \to \infty$,

$$\langle n_m \rangle = \langle n_0 \rangle \int_m^\infty \psi(\ell)d\ell. \qquad (13)$$

Similarly, the mean fraction of the chromosome found in segments longer than m is

$$\langle f_m \rangle = \frac{\langle n_0 \rangle}{L} \int_m^\infty \ell\psi(\ell)d\ell. \qquad (14)$$

We now turn to specific models, recovering previous results for SMC (Palamara et al., 2012) and obtaining new results for SMC'.

3.1. The SMC model

Under SMC, the distribution of segment lengths is given by Eq. (4). The mean total number of segments is

$$\langle n_0 \rangle_{\text{SMC}} = \frac{L}{\int_0^\infty \frac{4N\ell}{(1+2N\ell)^3}d\ell} = 2NL. \qquad (15)$$
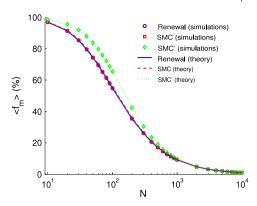
Fig. 6. The mean fraction of the chromosome found in shared segments longer than m, $\langle f_m \rangle$. Simulation details are as in Fig. 4. Simulation results and theory for SMC and the renewal approximation coincide. The renewal theory curve was obtained by numerically inverting Eq. (41). Theory for SMC and SMC′ (infinite-chromosome limits) is from Eqs. (17) and (A.3), respectively.

The mean number of shared segments longer than m is

$$\langle n_m \rangle_{\text{SMC}} = 2NL \int_m^\infty \frac{4N}{(1+2N\ell)^3} d\ell = \frac{2NL}{(1+2mN)^2}. \quad (16)$$

The mean fraction of the chromosome in segments longer than m is

$$\langle f_m \rangle_{\text{SMC}} = 2N \int_m^\infty \frac{4N\ell}{(1+2N\ell)^3} d\ell = \frac{1+4mN}{(1+2mN)^2}. \quad (17)$$

Eq. (17) has been previously derived by Palamara et al. (2012), by studying the distribution of segment lengths surrounding a randomly chosen site. Simulation results for $\langle f_m \rangle_{\text{SMC}}$ (Fig. 6) agree well with Eq. (17). While simulations were shown before (Palamara et al., 2012; Carmi et al., 2013), here we are able to observe perfect agreement even for very small values of N. Eq. (16) was derived by Palamara et al. (2012) using the relation $\langle n_m \rangle = L \langle f_m \rangle / \langle \ell_m \rangle$, where $\langle \ell_m \rangle$ is the mean length of segments longer than m. Eq. (16) can be derived in yet another way, using a result from Ralph and Coop (2013), who showed that for a fixed TM-RCA t, the mean number of segments longer than m is $K(t,m) = e^{-2mNt}[2Nt(L-m)+1]$. Integrating over all t using $P_c(t) = e^{-t}$, we have $\langle n_m \rangle = \int_0^\infty K(t,m)P_c(t)dt = (1+2NL)/(1+2mN)^2$. For $L \gg 1/2N$, we recover Eq. (16). Also note that for a fixed t, the mean number of segments of length in $[\ell, \ell+d\ell]$ is $-\partial K(t,\ell)/\partial \ell\, d\ell$. Integrating over all t as before, this gives $4N(1+2NL)/(1+2N\ell)^3\, d\ell$. Since the total number of segments (of all lengths) is $K(t,0) = (1+2NL)$, the probability of a random segment to be of length in $[\ell, \ell+d\ell]$ is $\psi(\ell)d\ell = 4N/(1+2N\ell)^3\, d\ell$, exactly as in our Eq. (4).

3.2. The SMC′ model

Under SMC′, the distribution of segment lengths is given by Eq. (10). The mean total number of segments is (using Eq. (11))

$$\langle n_0 \rangle_{\text{SMC}'} = \frac{L}{\int_0^\infty \ell \psi_{\text{SMC}'}(\ell)d\ell} = \frac{4NL}{3}. \quad (18)$$

Eq. (18) represents a surprisingly simple result, stating that for long chromosomes, the mean number of segments in SMC′ is precisely 2/3 of the total number of recombination events ($2NL$). To provide an intuitive explanation, we recall (Section 2.4) that the stationary distribution of tree heights at recombination sites in SMC′ is $\pi_\infty^{\text{SMC}'}(t) = te^{-t}$ (as in SMC). At a recombination site, there is probability $1 - q_{\text{SMC}'}(t|t)$ for the TMRCA to change and consequently, for the segment to terminate. Integrating over all t,

$$\int_0^\infty te^{-t}[1-q_{\text{SMC}'}(t|t)]dt = \int_0^\infty te^{-t}\frac{2t+1-e^{-2t}}{4t}dt = \frac{2}{3}. \quad (19)$$

In fact, it can be shown that at stationarity, the new tree has equal probability to be either larger, smaller, or equal to the previous tree. Also note that the probability to change the MRCA at a recombination site is 2/3 also for the ARG (Griffiths and Marjoram, 1997, Theorem 2.4).

Next, using Eqs. (10)–(12), it can be seen that

$$\psi_{\text{SMC}'}(\ell) = \frac{\int_0^\infty \lambda^2(t)e^{-t-\lambda(t)\ell}dt}{\langle n_0 \rangle_{\text{SMC}'}/L}. \quad (20)$$

Using Eqs. (13) and (20), the mean number of segments longer than m is

$$\langle n_m \rangle_{\text{SMC}'} = \langle n_0 \rangle_{\text{SMC}'} \int_m^\infty \psi_{\text{SMC}'}(\ell)d\ell$$

$$= L \int_0^\infty \lambda(t)e^{-t-\lambda(t)m}dt. \quad (21)$$

The final result, which we obtained using MATHEMATICA (Wolfram Research, 2012), is given in Appendix A (Eq. (A.2)).

Finally, using Eqs. (14) and (20), we have

$$\langle f_m \rangle_{\text{SMC}'} = \frac{\langle n_0 \rangle_{\text{SMC}'}}{L} \int_m^\infty \ell \psi_{\text{SMC}'}(\ell)d\ell$$

$$= \int_0^\infty e^{-t-\lambda(t)m}[1+\lambda(t)m]dt. \quad (22)$$

The result of the integral is given in Appendix A (Eq. (A.3)). Numerical evaluation shows perfect agreement with simulation results, for all values of N (Fig. 6).

4. Renewal theory results for finite chromosomes

In this section, we use renewal theory to derive the complete distribution of our quantities of interest: the number of segments longer than m (Section 4.1) and the fraction of the chromosome in segments longer than m (Section 4.2), for a chromosome of a finite size L. In both cases, we derive an expression in Laplace space for the distribution (Eq. (32) for the number of segments and Eq. (38) for the fraction of the chromosome). Those expressions are general for any segment length distribution. We then substitute the specific SMC form, to obtain explicit expressions (Appendix B). As we show, the distributions can be numerically inverted and compared to simulations or be used for demographic inference. Using standard techniques, we also obtain the first two moments (in real space) for long (but finite) chromosomes. Our method in this section is adapted from the physics literature (Godrèche and Luck, 2001).

4.1. The distribution of the number of segments longer than m under the renewal approximation

4.1.1. Theory

Define $P(n_m = k; L)$ as the probability that two chromosomes share exactly k segments longer than m over a sequence of length L, under the renewal IBD process defined in Table 2 (Section 2.3). We will obtain $\tilde{P}(n_m = k, s)$, the Laplace transform of $P(n_m = k, L)$ with respect to L: $\tilde{P}(n_m = k, s) = \int_0^\infty e^{-sL}P(n_m = k, L)dL$. Let us first define an auxiliary function, $\eta_m(L)dL$, which is the probability that, conditional on recombination at position 0 in the sequence, (a) recombination occurred at position in $[L, L+dL]$; and (b) all intermediate recombination events in $[0, L]$ had terminated segments that were shorter than m. Note that $\eta_m(L)$, as well as $Q_m(k, L)$ below (Eq. (26)), are not PDFs. Then, $\eta_m(L)$ satisfies

$$\eta_m(L) = \delta(L) + \int_0^{\min(m,L)} \psi(\ell)\eta_m(L-\ell)d\ell. \quad (23)$$

In Eq. (23), $\delta(x)$ is Dirac's delta function and $\psi(\ell)$ is the PDF of segment lengths. The derivation will proceed with a general $\psi(\ell)$; we will substitute the explicit SMC form (Eq. (4)) only at the final result. Eq. (23) is explained as follows. The first term ($\delta(L)$) accounts for the case $L = 0$. Otherwise, we condition on the length of the last segment, ℓ, which cannot exceed either m or L. Given ℓ, we require the recombination at $L - \ell$ to end a series of short segments, which happens with probability $\eta_m(L - \ell)$. Note that we made use of the renewal property, namely the independence of successive segment lengths.

We now apply the Laplace transform ($L \to s$) to both sides of Eq. (23),

$$
\begin{aligned}
\tilde{\eta}_m(s) &= 1 + \int_0^\infty e^{-sL} \left[\int_0^{\min(m,L)} \psi(\ell)\eta_m(L-\ell)d\ell \right] dL \\
&= 1 + \int_0^m \left[\int_\ell^\infty e^{-sL}\psi(\ell)\eta_m(L-\ell)dL \right] d\ell \\
&= 1 + \int_0^m e^{-s\ell}\psi(\ell) \left[\int_\ell^\infty e^{-s(L-\ell)}\eta_m(L-\ell)dL \right] d\ell \\
&= 1 + \int_0^m e^{-s\ell}\psi(\ell)d\ell \int_0^\infty e^{-sL'}\eta_m(L')dL' \\
&= 1 + \tilde{\psi}_{<m}(s)\tilde{\eta}_m(s),
\end{aligned}
\tag{24}
$$

where we defined $\tilde{\psi}_{<m}(s) \equiv \int_0^m e^{-s\ell}\psi(\ell)d\ell$. We thus obtained an algebraic equation for $\tilde{\eta}_m(s)$, whose solution is

$$
\tilde{\eta}_m(s) = \left[1 - \tilde{\psi}_{<m}(s) \right]^{-1}.
\tag{25}
$$

Next, we define another auxiliary function, $Q_m(k, L)dL$, which is the probability that (a) recombination occurred at position in $[L, L + dL]$; and (b) that the recombination event of (a) has ended the kth segment *longer than* m. For $k = 0$, $Q_m(0, L) = \delta(L)$. For $k > 0$, we have the following recursion equation,

$$
Q_m(k, L) = \int_m^L \psi(\ell) \left[\int_0^{L-\ell} \eta_m(\ell')Q_m(k-1, L-\ell-\ell')d\ell' \right] d\ell.
\tag{26}
$$

Eq. (26) is explained similarly to Eq. (23). We condition on the length of the last segment, ℓ, which must be longer than m (but shorter than L). Given the preceding recombination at $L - \ell$, we condition on the length of rightmost stretch of short segments, ℓ', which has probability $\eta_m(\ell')$. Note that $\eta_m(L)$ does not depend on the absolute position along the sequence, again, due to the renewal property. Finally, given ℓ and ℓ', there must have been a recombination event at $L - \ell - \ell'$ ending the $(k-1)$th segment longer than m, with probability $Q_m(k-1, L-\ell-\ell')$. We now apply the Laplace transform to Eq. (26),

$$
\begin{aligned}
\tilde{Q}_m(k, s) &= \int_m^\infty e^{-sL} \left\{ \int_m^L \psi(\ell) \right. \\
&\quad \left. \times \left[\int_0^{L-\ell} \eta_m(\ell')Q_m(k-1, L-\ell-\ell')d\ell' \right] d\ell \right\} dL \\
&= \int_m^\infty e^{-s\ell}\psi(\ell)d\ell \int_\ell^\infty e^{-s(L-\ell)} \\
&\quad \times \left[\int_0^{L-\ell} \eta_m(\ell')Q_m(k-1, L-\ell-\ell')d\ell' \right] dL \\
&= \int_m^\infty e^{-s\ell}\psi(\ell)d\ell \int_0^\infty e^{-sL'} \\
&\quad \times \left[\int_0^{L'} \eta_m(\ell')Q_m(k-1, L'-\ell')d\ell' \right] dL'
\end{aligned}
$$

$$
\begin{aligned}
&= \tilde{\psi}_{>m}(s)\tilde{\eta}_m(s)\tilde{Q}_m(k-1, s) \\
&= \frac{\tilde{\psi}_{>m}(s)}{1 - \tilde{\psi}_{<m}(s)}\tilde{Q}_m(k-1, s),
\end{aligned}
\tag{27}
$$

where $\tilde{\psi}_{>m}(s) \equiv \int_m^\infty e^{-s\ell}\psi(\ell)d\ell$, we used the fact that $Q_m(k > 0, L < m) = 0$, and in the last line, we used the convolution theorem and Eq. (25). Using Eq. (27) and the initial condition, $\tilde{Q}_m(0, s) = 1$, we have

$$
\tilde{Q}_m(k, s) = \left(\frac{\tilde{\psi}_{>m}(s)}{1 - \tilde{\psi}_{<m}(s)} \right)^k.
\tag{28}
$$

We next define $\phi(\ell) \equiv 1 - \int_0^\ell \psi(\ell')d\ell' = \int_\ell^\infty \psi(\ell')d\ell'$, the probability that a segment extends for sequence length greater than ℓ. We are now in a position to compute $P(n_m = k, L)$. For $k > 0$,

$$
\begin{aligned}
P(n_m = k, L) &= \int_0^m \phi(\ell) \left[\int_0^{L-\ell} \eta_m(\ell')Q_m(k, L-\ell-\ell')d\ell' \right] d\ell \\
&+ \int_m^L \phi(\ell) \left[\int_0^{L-\ell} \eta_m(\ell')Q_m(k-1, L-\ell-\ell')d\ell' \right] d\ell.
\end{aligned}
\tag{29}
$$

For $P(n_m = k, L)$, we do not require recombination at L. Therefore, we condition on the sequence length ℓ since the rightmost recombination event, with the probability of no recombination since then being $\phi(\ell)$. Then, if $\ell < m$, we require k segments longer than m to be seen by position $L - \ell$, possibly followed by any number of short segments. If $\ell > m$, then the sequence $[L - \ell, L]$ will form a segment longer than m on its own, and we only require $k - 1$ previous segments longer than m. Eq. (29) can be transformed similarly to Eqs. (24) and (27), yielding

$$
\begin{aligned}
&\tilde{P}(n_m = k, s) \\
&= \tilde{\eta}_m(s) \left[\tilde{\phi}_{<m}(s)\tilde{Q}_m(k, s) + \tilde{\phi}_{>m}(s)\tilde{Q}_m(k-1, s) \right],
\end{aligned}
\tag{30}
$$

where $\tilde{\phi}_{<m}(s) = \int_0^m e^{-s\ell}\phi(\ell)d\ell$ and $\tilde{\phi}_{>m}(s) = \int_m^\infty e^{-s\ell}\phi(\ell)d\ell$. For $k = 0$, we have $P(n_m = 0, L) = \int_0^{\min(m,L)} \phi(\ell)\eta_m(L-\ell)d\ell$. Applying the Laplace transform gives

$$
\tilde{P}(n_m = 0, s) = \tilde{\phi}_{<m}(s)\tilde{\eta}_m(s).
\tag{31}
$$

Combining Eqs. (25), (28), (30) and (31), and using $\tilde{\phi}_{<m}(s) + \tilde{\phi}_{>m}(s) = \tilde{\phi}(s) = [1 - \tilde{\psi}(s)]/s$ and $\tilde{\psi}_{<m}(s) + \tilde{\psi}_{>m}(s) = \tilde{\psi}(s)$, we finally obtain

$$
\begin{aligned}
&\tilde{P}(n_m = k, s) \\
&= \begin{cases}
\dfrac{\tilde{\phi}_{<m}(s)}{1 - \tilde{\psi}_{<m}(s)} & k = 0, \\[2ex]
\dfrac{[1 - \tilde{\psi}(s)][\tilde{\psi}_{>m}(s) + s\tilde{\phi}_{>m}(s)]}{s[1 - \tilde{\psi}_{<m}(s)]^2} \left[\dfrac{\tilde{\psi}_{>m}(s)}{1 - \tilde{\psi}_{<m}(s)} \right]^{k-1} & k > 0.
\end{cases}
\end{aligned}
\tag{32}
$$

Eq. (32) is our main result, and is valid for any distribution of segment lengths, $\psi(\ell)$. Due to normalization, we expect $\sum_{k=0}^\infty \tilde{P}(n_m = k, s) = \sum_{k=0}^\infty \int_0^\infty e^{-sL}P(n_m = k, L)dL = \int_0^\infty e^{-sL} \left[\sum_{k=0}^\infty P(n_m = k, L) \right] dL = \int_0^\infty e^{-sL}dL = 1/s$, as can be verified, after some algebra, from Eq. (32).

Our results have so far been general and could apply to any 'IBD process'. We now substitute the SMC segment length PDF, $\psi(\ell) = 4N/(1 + 2N\ell)^3$ (Eq. (4)). The distribution of the number of segments longer than m (Eq. (32)) under SMC is given in Eq. (B.1) (Appendix B). This can be numerically inverted (de Hoog et al., 1982), for each k, to obtain, for a given L, the distribution $P(n_m = k)$. The theoretical prediction compares perfectly to simulation results for both SMC and the renewal approximation (Fig. 7).
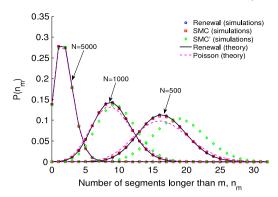
Fig. 7. The distribution of the number of shared segments longer than m, n_m. Simulation details are as in Fig. 4 (specifically, $L = 2$ and $m = 0.01$ (Morgans)). Theory for the renewal approximation was obtained by numerically inverting Eq. (B.1). The Poisson distribution has mean $2NL/(1 + 2mN)^2$ (Eq. (16)).

4.1.2. The mean

The mean number of segments longer than m is $\langle n_m \rangle = \sum_{k=0}^{\infty} k \, P(n_m = k, L)$. Taking the Laplace transform of $\langle n_m \rangle$, using Eq. (32) and the relation $\sum_{k=0}^{\infty} kx^k = x/(1-x)^2$, we obtain, after some algebra,

$$\langle \tilde{n_m} \rangle(s) = \frac{\tilde{\psi}_{>m}(s) + s\tilde{\phi}_{>m}(s)}{s[1 - \tilde{\psi}(s)]}. \tag{33}$$

For SMC, we obtain, using Eq. (33) and MATHEMATICA,

$$\langle \tilde{n_m} \rangle_{\text{SMC}}(s) = \frac{4N^2 e^{-ms}}{s^2(1 + 2mN)^2 \left[se^{\frac{s}{2N}} \text{Ei} \left(-\frac{s}{2N} \right) + 2N \right]}, \tag{34}$$

where Ei is the exponential integral function. Noting that $\lim_{s \to 0} s^2 \langle \tilde{n_m} \rangle_{\text{SMC}}(s) = 2N/(1 + 2mN)^2$, we have $\lim_{L \to \infty} \langle n_m \rangle_{\text{SMC}} / L = 2N/(1 + 2mN)^2$, exactly as in Eq. (16).

4.1.3. The variance

The second moment of the number of segments longer than m can be computed using $\langle \tilde{n_m^2} \rangle(s) = \sum_{k=0}^{\infty} k^2 \tilde{P}(n_m = k, s)$, from which the variance can be obtained. For SMC and for large L,

$$\text{Var}[n_m]_{\text{SMC}} = \frac{2NL}{(1 + 2mN)^4} [2 \ln(2NL) + 4mN(mN - 1) - 5] + O(\ln^2 L). \tag{35}$$

4.1.4. The Poisson approximation

Palamara et al. (2012), following Huff et al. (2011), proposed that the number of shared segments longer than m is Poisson distributed, with the infinite-chromosome mean, $\langle n_m \rangle_{\text{SMC}} = 2NL/(1 + 2mN)^2$ (Eq. (16)). The Poisson distribution fits the simulation results reasonably (Fig. 7; see also Section 4.2.4). Indeed, for large values of N and L, Eq. (35) gives $\text{Var}[n_m]_{\text{SMC}} \approx \langle n_m \rangle_{\text{SMC}} \approx \frac{L}{2m^2N}$, as expected from a Poisson variable. Deviations appear for small values of N (Fig. 7).

4.1.5. Demographic inference

The results of Section 4.1.1 have attractive implications for demographic inference. While this is not our main focus here, we provide a simple demonstration. For a given population size N (and for $L = 2$ and $m = 0.01$ (Morgans)), we simulated the SMC IBD process $R = 5000$ times and recorded, for each run, the number of shared segments longer than m, n_m. This corresponds, roughly, to the information that will be available from sampling a single chromosome in 50 (diploid) individuals, although we note that in
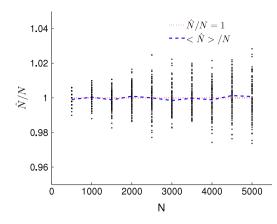

Fig. 8. Inference of the effective population size using the distribution of the number of shared segments longer than m. Simulations for $N = 500, 1000, \dots, 5000$ were performed as in Fig. 4 and for $R = 5000$ pairs of chromosomes, and Eq. (37) was used to compute $\hat{N}$, the estimator of the population size. We then repeated 100 times for each N, and each ratio $\hat{N}/N$ is shown as a dot. The dotted red line represents $\hat{N} = N$ and the blue line shows $\langle \hat{N} \rangle / N$. The estimator in unbiased, with standard deviation as low as $0.003N$ for $N = 500$ and $0.011N$ for $N = 5000$.

reality, pairs of chromosomes in a sample are weakly dependent (see Carmi et al. (2013) and the Discussion). Additionally, the underlying ancestral process is neither SMC nor even the coalescent with recombination, but there is rather a shared underlying pedigree (Wakeley et al., 2012); however, we leave investigation of more complex models to future studies. Given N, m, and L, the log-likelihood of the sample $\{n_m^{(i)}\}$, $i = 1, \dots, R$, is

$$\text{LL}(N) = \sum_{i=1}^{R} \log P \left(n_m = n_m^{(i)}, L \right), \tag{36}$$

where $P(n_m = k, L)$ is given by numerically inverting, $s \to L$, Eq. (B.1). We then computed the *maximum likelihood estimator*,

$$\hat{N} = \arg \max_{N} \text{LL}(N). \tag{37}$$

Simulation results (Fig. 8) show that the estimator performs excellently, with standard deviation $\approx 0.01N$ or lower. The performance of the estimator deteriorates for large values of N, since the number of shared segments longer than m approaches zero (Fig. 7; Eq. (16)). Under our "noise-free" simulations, even the simple-minded estimator, $\hat{N} = 1/(m \langle f_m \rangle) - 3/(4m)$ (Carmi et al., 2013), performs well, although with bias ($\langle \hat{N} \rangle / N \approx 1.02$; see Carmi et al. (2013)) and with $\approx 60\%$ larger standard deviation than the maximum likelihood estimator.

4.2. The distribution of the fraction of the chromosome found in segments longer than m

4.2.1. Theory

Denote $P(f_m)$ as the density of the fraction of the chromosome found in shared segments longer than m. The derivation of $P(f_m)$ uses techniques similar to those used in Section 4.1.1 and is tedious. We therefore omit the details and skip to the analysis of the final result. Let $P(L_m, L)$ be the density of $L_m \equiv Lf_m$, the total sequence length found in shared segments longer than m, given a chromosome of length L, and let $\tilde{P}_{L_m}(u, s)$ be its Laplace transform. This is a double Laplace transform: $L \to s$ and $L_m \to u$, or $\tilde{P}_{L_m}(u, s) = \int_0^{\infty} \int_0^{\infty} e^{-uL_m - sL} P(L_m, L) dL_m dL$. For the renewal IBD process defined in Section 2.3 and with segment length PDF $\psi(\ell)$,
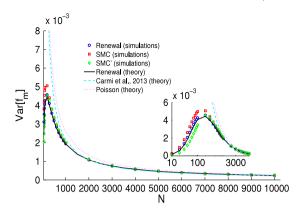
Fig. 9. The variance of the fraction of the chromosome found in shared segments longer than m, Var $[f_m]$. Simulation details are as in Fig. 4. The inset zooms in on the small N region. The renewal theory curve is the large L expansion given in Eq. (43). The line representing Carmi et al. (2013) is from (44), and the Poisson expression is from (50).

it can be shown that

$$
\tilde{P}_{L_m}(u, s)
$$

$$
= \frac{\frac{1}{s} - \frac{1}{s}\tilde{\psi}_{<m}(s) + \phi(m)\left[\frac{e^{-m(s+u)}}{s+u} - \frac{e^{-ms}}{s}\right] - \frac{\tilde{\psi}_{>m}(s+u)}{s+u}}{1 - \tilde{\psi}_{<m}(s) - \tilde{\psi}_{>m}(s+u)}, \quad (38)
$$

where, as in Section 4.1.1, $\phi(\ell) = 1 - \int_0^\ell \psi(\ell')d\ell'$, $\tilde{\psi}_{<m}(z) = \int_0^m e^{-z\ell}\psi(\ell)d\ell$, and $\tilde{\psi}_{>m}(z) = \int_m^\infty e^{-z\ell}\psi(\ell)d\ell$. For $u = 0$, we expect, due to normalization, $\tilde{P}_{L_m}(u = 0, s) = \int_0^\infty e^{-sL}\int_0^\infty P(L_m, L) dL_m dL = \int_0^\infty e^{-sL}dL = 1/s$, as can be verified from Eq. (38).

We then substituted the SMC form, $\psi(\ell) = 4N/(1 + 2N\ell)^3$ (Eq. (4)), and evaluated Eq. (38) in MATHEMATICA. The final result is given in Appendix B, Eq. (B.3). Eq. (B.3) can be numerically inverted with respect to both u and s (Brančík, 2011) to give $P(L_m, L)$, from which we have $P(f_m = L_m/L) = LP(L_m, L)$. The theoretical prediction agrees well with simulations (Fig. 4). Very small deviations may be due to numerical errors in the two-dimensional inversion.

4.2.2. The mean

The mean sequence length in segments longer than m, $\langle L_m \rangle$, can be obtained (in s space) from $\tilde{P}_{L_m}(u, s)$ by

$$
\langle \tilde{L_m} \rangle(s) = -\left.\frac{\partial \tilde{P}_{L_m}(u, s)}{\partial u}\right|_{u=0}. \quad (39)
$$

For a general $\psi(\ell)$, we obtain from Eq. (38),

$$
\langle \tilde{L_m} \rangle(s) = \frac{\phi(m)e^{-ms}(1 + ms) - \tilde{\psi}_{>m}(s)}{s^2\left[1 - \tilde{\psi}(s)\right]}. \quad (40)
$$

For the SMC form of $\psi(\ell)$ (Eq. (4)), this gives

$$
\langle \tilde{L_m} \rangle_{SMC}(s) = e^{-ms}\frac{sC^2 e^{\frac{sC}{2N}}\mathrm{Ei}\left(-\frac{sC}{2N}\right) + 2N(1 + 4mN)}{s^2 C^2\left[se^{\frac{s}{2N}}\mathrm{Ei}\left(-\frac{s}{2N}\right) + 2N\right]}, \quad (41)
$$

where $C = 1 + 2mN$. The prediction of Eq. (41) turns out to be virtually identical (Fig. 6) to the infinite-chromosome SMC expression (Eq. (17)), which can also be obtained by taking the limit $s \to 0$ (corresponding to $L \to \infty$) of Eq. (41).

4.2.3. The variance

The second moment of L_m is given by

$$
\langle \tilde{L_m^2} \rangle(s) = \left.\frac{\partial^2 \tilde{P}_{L_m}(u, s)}{\partial u^2}\right|_{u=0}. \quad (42)
$$

Assuming the SMC form of $\psi(\ell)$ (Eq. (4)), the derivatives can be taken. While the resulting expression (in s space) can be numerically inverted, more insight is gained by looking at the large L limit. Considering only the first order expansion in s and inverting, we obtain $\lim_{L\to\infty}\langle f_m^2 \rangle = \lim_{L\to\infty}\langle f_m \rangle^2$ or $\lim_{L\to\infty}$ Var $[f_m] = 0$, as expected. Expanding to the next order in s and inverting, we find

Var $[f_m]_{SMC}$

$$
= \frac{\ln(1 + 2mN)\left[8mN\left(1 + 2mN - 2m^3N^3\right) + 1\right]}{NL(1 + 2mN)^4}
$$

$$
+ \frac{2mN\left\{8m^3N^3\ln N + mN[4mN[mN(\ln 4 - 1) - 2] - 7] - 1\right\}}{NL(1 + 2mN)^4}
$$

$$
+ \frac{16m^4N^4}{(1 + 2mN)^4}\frac{\ln L}{NL} + O\left(\frac{\ln^2 L}{L^2}\right). \quad (43)
$$

Eq. (43) is compared to simulations in Fig. 9, showing excellent agreement with the renewal process. For large N, Var $[f_m]_{SMC} \approx [\ln(L/m) - 1/2]/(NL)$.

Carmi et al. (2013) computed the variance by approximating, for large N, the probability that two sites lie on shared segments, obtaining

$$
\text{Var } [f_m] \approx \frac{\ln\left(\frac{L}{m}\right) - 1}{NL}. \quad (44)
$$

For $\ln(L/m) \gg 1$, Eq. (44) has the same limit as Eq. (43). Eq. (44) agrees well with simulations for large values of N (Fig. 9); however, the approximation breaks down for small values of N.

4.2.4. The Poisson approximation

Palamara et al. (2012) approximated the number of shared segments longer than m, n_m, as a Poisson with mean $\langle n_m \rangle_{SMC} = 2NL(1 + 2mN)^2$ (Eq. (16); see also Section 4.1.4). According to that approximation, L_m can be written as a sum of n_m independent random variables, each of which is distributed as $\psi_m(\ell) = \psi_{SMC}(\ell)/\int_m^\infty \psi_{SMC}(\ell)d\ell$. To compute the distribution of L_m under the Poisson approximation, $P_{Poisson}(L_m, L)$, it is again convenient to work in Laplace space (see also Carmi et al. (2013)). Define $\tilde{\psi}_m(u) = \int_0^\infty e^{-u\ell}\psi_m(\ell)d\ell$, the Laplace transform ($\ell \to u$) of $\psi_m(\ell)$, and denote by $\tilde{P}_{L_m,Poisson}(u, L)$ the Laplace transform, $L_m \to u$, of $P_{Poisson}(L_m, L)$. Using the convolution theorem, given n_m,

$$
\tilde{P}_{L_m,Poisson}(u, L|n_m) = \left[\tilde{\psi}_m(u)\right]^{n_m}. \quad (45)
$$

Since n_m is assumed to be Poisson,

$$
\tilde{P}_{L_m,Poisson}(u, L) = \sum_{n=0}^\infty e^{-\langle n_m \rangle_{SMC}}\frac{\langle n_m \rangle_{SMC}^n\left[\tilde{\psi}_m(u)\right]^n}{n!}
$$

$$
= \exp\left[-\langle n_m \rangle_{SMC}\left(1 - \tilde{\psi}_m(u)\right)\right]. \quad (46)
$$

This gives

$$
-\ln\left[\tilde{P}_{L_m,Poisson}(u, L)\right]/L = \frac{u^2 e^{\frac{u}{2N}}\mathrm{Ei}\left[-\frac{u(1+2mN)}{2N}\right]}{2N}
$$

$$
+ \frac{e^{-mu}[2N(e^{mu} + mu - 1) + u]}{(1 + 2mN)^2}, \quad (47)
$$

where Ei is the exponential integral function. Using Eq. (47), $\tilde{P}_{L_m,\text{Poisson}}(u, L)$ can be numerically inverted (de Hoog et al., 1982), showing (Fig. 4) reasonable agreement with simulation results, albeit with deviations for small values of N.

To compute the variance under the Poisson approximation, we redefine $\psi_m(\ell)$ as

$$\psi_m(\ell) = \frac{\psi_{\text{SMC}}(\ell)}{\int_m^L \psi_{\text{SMC}}(\ell)d\ell}; \quad m < \ell < L, \tag{48}$$

imposing an upper limit at L, since otherwise $\langle \ell_m^2 \rangle \to \infty$. Using the law of total variance,

$$\begin{aligned}
\text{Var}[L_m]_{\text{Poisson}} &= \langle \text{Var}[L_m|n_m] \rangle + \text{Var}[\langle L_m|n_m \rangle] \\
&= \langle n_m \rangle \text{Var}[\ell_m] + \text{Var}[n_m] \langle \ell_m \rangle^2 \\
&= \langle n_m \rangle_{\text{SMC}} \langle \ell_m^2 \rangle,
\end{aligned} \tag{49}$$

where we used the fact that a Poisson variable has equal mean and variance. Using Eqs. (13), (15) and (48),

$$\begin{aligned}
\text{Var}[f_m]_{\text{Poisson}} &= \frac{\langle n_m \rangle_{\text{SMC}}}{L^2} \frac{\int_m^L \ell^2 \psi_{\text{SMC}}(\ell)d\ell}{\int_m^L \psi_{\text{SMC}}(\ell)d\ell} \\
&\approx \frac{2N}{L} \int_m^L \ell^2 \psi(\ell)d\ell \\
&= \frac{\frac{2N(m-L)[mN(8NL+3)+3NL+1]}{(1+2mN)^2(1+2NL)^2} + \ln\left(\frac{1+2NL}{1+2mN}\right)}{NL}.
\end{aligned} \tag{50}$$

Here too, for large N, $\text{Var}[f_m]_{\text{Poisson}} \approx \ln(L/m)/(NL)$, which is the same (for $\ln(L/m) \gg 1$) as the renewal theory limit (Eq. (43)). Eq. (50) agrees well with simulations for large values of N, but breaks down already for $N \lesssim 5000$.

5. Variable population size

Many natural populations (including humans) did not maintain a constant population size throughout their history. As we show in this section, our results are generalizable to any arbitrary variable population size, $N(t) = N_0\nu(t)$. The key insight is that all results depend on a single quantity, the PDF of segment lengths, $\psi(\ell)$. This can be seen from Eqs. (12)–(14) (the infinite-chromosome results; Section 3), Eq. (32) (the distribution of the number of shared segments longer than m; Section 4.1), and Eq. (38) (the distribution of the fraction of the chromosome in segments longer than m; Section 4.2). Therefore, we need only show how to compute $\psi(\ell)$ for an arbitrary $\nu(t)$. In Sections 5.1 and 5.2, we compute $\psi(\ell)$ for SMC and SMC', respectively, as well as derive the infinite-chromosome means.

5.1. The SMC model

Define $h(t) \equiv 1/\nu(t)$. Li and Durbin (2011) derived the stationary distribution of tree heights at a recombination site (their supplementary Eq. (7)),

$$\pi_\infty^{\text{SMC}}(t) = \frac{th(t)e^{-\int_0^t h(\tau)d\tau}}{\int_0^\infty e^{-\int_0^{t'} h(\tau)d\tau}dt'}. \tag{51}$$

Eq. (51) reduces to $\pi_\infty^{\text{SMC}}(t) = te^{-t}$ (Eq. (2)) for a constant population size, where $h(t) = 1$. For a given tree height t, the sequence length between recombination events is distributed exponentially with rate $2N_0t$. Therefore (see also Eq. (4)),

$$\begin{aligned}
\psi_{\text{SMC}}(\ell) &= \int_0^\infty \pi_\infty^{\text{SMC}}(t) \cdot 2N_0te^{-2N_0t\ell}dt \\
&= 2N_0 \frac{\int_0^\infty t^2 h(t)e^{-\int_0^t h(\tau)d\tau - 2N_0t\ell}dt}{\int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt}.
\end{aligned} \tag{52}$$

We can now evaluate Eqs. (12)–(14) for the infinite-chromosome means. The mean segment length is

$$\begin{aligned}
\langle \ell \rangle_{\text{SMC}} &= 2N_0 \frac{\int_0^\infty t^2 h(t)e^{-\int_0^t h(\tau)d\tau} \left[\int_0^\infty \ell e^{-2N_0t\ell}d\ell\right]dt}{\int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt} \\
&= \frac{\int_0^\infty h(t)e^{-\int_0^t h(\tau)d\tau}dt}{2N_0 \int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt} = \frac{1}{2N_0 \int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt}.
\end{aligned} \tag{53}$$

Hence (see Eq. (12)),

$$\langle n_0 \rangle_{\text{SMC}} = 2N_0L \int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt. \tag{54}$$

Note that we implicitly assumed that $\lim_{t\to\infty} \nu(t) < \infty$. Eq. (54) can also be obtained using Corollary 3 in Li and Durbin (2011). For the mean number of segments longer than m, we obtain, using techniques similar to those used in Eq. (53),

$$\langle n_m \rangle_{\text{SMC}} = 2N_0L \int_0^\infty th(t)e^{-\int_0^t h(\tau)d\tau - 2N_0mt}dt. \tag{55}$$

Finally,

$$\langle f_m \rangle_{\text{SMC}} = \int_0^\infty h(t)e^{-\int_0^t h(\tau)d\tau - 2N_0mt}(1 + 2N_0mt)dt. \tag{56}$$

Eq. (56) was also derived by Palamara et al. (2012). It can be verified that substituting $h(t) = 1$ in Eqs. (54)–(56), we recover the results of Section 3.1 (Eqs. (15)–(17), respectively).

5.2. The SMC' model

For SMC', we need to recompute $q_{\text{SMC'}}(t|s)$, the probability that the new tree height at a recombination site is t, given that the previous height was s (see Eq. (5)),

$$q_{\text{SMC'}}(t|s)$$
$$= \begin{cases}
\int_0^s \frac{1}{s}\left[\int_{t_r}^s h(t_c)e^{-2\int_{t_r}^{t_c} h(\tau)d\tau}dt_c\right]dt_r & t = s, \\
\int_0^t \frac{1}{s}h(t)e^{-2\int_{t_r}^t h(\tau)d\tau}dt_r & t < s, \\
\left[\int_0^s \frac{1}{s}e^{-2\int_{t_r}^s h(\tau)d\tau}dt_r\right]h(t)e^{-\int_s^t h(\tau)d\tau} & t > s.
\end{cases} \tag{57}$$

Eq. (57) is explained similarly to Eq. (5), once we recognize that coalescence occurs at (absolute) time t at rate $h(t)$, and that the probability of *no* coalescence between $[s, t]$ is $e^{-\int_s^t h(\tau)d\tau}$ (Griffiths and Tavare, 1994). It can be shown that Eq. (57) is normalized ($\int_0^t q_{\text{SMC'}}(t|s)dt = 1$), and that, as in the case of a constant population size (Section 2.4), the stationary distribution of tree heights, $\pi_\infty^{\text{SMC'}}(t)$, is identical to that of SMC and is given by Eq. (51). It can also be shown that at stationarity, the new tree has equal probabilities to be either taller, shorter, or equal to the previous tree, as we have seen for a constant population size (Section 3.2).

As in Section 3.2, we define a chain with probabilities $q_{\text{SMC',seg}}(t|s) = q_{\text{SMC'}}(t|s)/[1 - q_{\text{SMC'}}(s|s)]$ (as in Eq. (8)), whose stationary distribution, $\pi_\infty^{\text{SMC',seg}}(t)$, is the distribution of tree heights at segment ends. Using the marginal distribution of tree heights at random sites (Griffiths and Tavare, 1994),

$$P_c(t) = h(t)e^{-\int_0^t h(\tau)d\tau}, \tag{58}$$

and a detailed balance argument, it can be shown that

$$\pi_\infty^{\text{SMC',seg}}(t) = \frac{P_c(t)\lambda(t)}{\int_0^\infty P_c(t)\lambda(t)dt}, \tag{59}$$

where

$$\lambda(t) = 2N_0 t[1 - q_{\text{SMC'}}(t|t)] \tag{60}$$

$$= 2N_0 t \left[1 - \frac{1}{t} \int_0^t \int_{t_r}^t h(t_c) e^{-2 \int_{t_r}^{t_c} h(\tau) d\tau} dt_c dt_r \right]$$

$$= N_0 \left[t + e^{-2 \int_0^t h(\tau) d\tau} \int_0^t e^{2 \int_0^{t'} h(\tau) d\tau} dt' \right]. \tag{61}$$

The distribution of segment lengths is then given by (see also Eq. (10); Section 2.4)

$$\psi_{\text{SMC'}}(\ell) = \int_0^\infty \pi_\infty^{\text{SMC',seg}}(t) \lambda(t) e^{-\lambda(t)\ell} dt$$

$$= \frac{\int_0^\infty P_c(t) \lambda^2(t) e^{-\lambda(t)\ell} dt}{\int_0^\infty P_c(t) \lambda(t) dt}. \tag{62}$$

Note that Eq. (62) depends solely on $N(t)$, and as expected, the distribution of $\rho = 2N_0 \ell$ is independent of N_0.

We now derive the infinite-chromosome means (Section 3). The mean segment length is

$$\langle \ell \rangle_{\text{SMC'}} = \int_0^\infty \ell \psi_{\text{SMC'}}(\ell) d\ell = \left[\int_0^\infty P_c(t) \lambda(t) dt \right]^{-1}, \tag{63}$$

where we used the fact that $\int_0^\infty P_c(t) dt = 1$. Using Eq. (12) and after some algebra,

$$\langle n_0 \rangle_{\text{SMC'}} = L \int_0^\infty P_c(t) \lambda(t) dt$$

$$= N_0 L \int_0^\infty h(t) e^{-\int_0^t h(\tau) d\tau}$$

$$\times \left[t + e^{-2 \int_0^t h(\tau) d\tau} \int_0^t e^{2 \int_0^{t'} h(\tau) d\tau} dt' \right] dt$$

$$= \frac{4N_0 L}{3} \int_0^\infty e^{-\int_0^t h(\tau) d\tau} dt. \tag{64}$$

This is, as expected, exactly 2/3 of the number of recombination events (Eq. (54)).

Using Eqs. (58), (62) and (64), we can write

$$\psi_{\text{SMC'}}(\ell) = \frac{\int_0^\infty P_c(t) \lambda^2(t) e^{-\lambda(t)\ell} dt}{\langle n_0 \rangle_{\text{SMC'}} / L}$$

$$= \frac{\int_0^\infty h(t) \lambda^2(t) e^{-\int_0^t h(\tau) d\tau - \lambda(t)\ell} dt}{\frac{4N_0}{3} \int_0^\infty e^{-\int_0^t h(\tau) d\tau} dt}. \tag{65}$$

The mean number of segments longer than m is

$$\langle n_m \rangle_{\text{SMC'}} = \langle n_0 \rangle_{\text{SMC'}} \int_m^\infty \psi_{\text{SMC'}}(\ell) d\ell$$

$$= L \int_0^\infty P_c(t) \lambda(t) e^{-\lambda(t)m} dt. \tag{66}$$

Finally, the mean fraction of the chromosome in segments longer than m is

$$\langle f_m \rangle_{\text{SMC'}} = \frac{\langle n_0 \rangle_{\text{SMC'}}}{L} \int_m^\infty \ell \psi_{\text{SMC'}}(\ell) d\ell$$

$$= \int_0^\infty P_c(t) e^{-\lambda(t)m} [1 + \lambda(t)m] dt. \tag{67}$$

It can be verified that all the results of this section reduce to the SMC results (Section 5.1) for $\lambda(t) = 2N_0 t$ and to the constant population size results (Section 3.2) for $h(t) = 1$.

6. Summary and discussion

In summary, we introduced a general framework for the IBD process in Markovian approximations of the coalescent with recombination (SMC and SMC'), as well as a new renewal approximation, in which tree heights on both sides of a recombination site are independent (Section 2). We showed how previous results for the mean number of segments and the mean shared sequence length in SMC emerge naturally from our framework in the infinite-chromosome limit; we then derived these quantities under SMC' (Section 3). Using renewal theory, we derived expressions for the distributions of the number of shared segments (Section 4.1) and the fraction of the chromosome in shared segments (Section 4.2). Finally, we generalized our results to populations with variable size (Section 5).

Our main contributions are (a) providing a unified framework for the IBD process, depending exclusively on a single distribution (that of segment lengths), in which previous and new results are coherently derived and easily generalized; (b) new results for SMC': the distribution of tree heights at recombination sites (both conditional on the previous tree and at stationarity), the stationary distribution of tree heights at segment ends, the distribution of segment lengths, the mean number of shared segments, and the mean fraction of the chromosome in shared segments; and (c) introducing a novel renewal approximation, under which distributions of key quantities were obtained.

Our results rely on a number of simplifying assumptions, beyond the standard postulates of coalescent theory. First, our model considers segments shared between haploid chromosomes and does not incorporate any model for shared segments detection errors. In reality, genotyping errors, recent mutations, and phase uncertainty do not allow the confident detection of short segments, although this is partly remedied by our theory being entirely specified in terms of a length cutoff (m), which can be tuned for the quality of the data under examination. Next, when computing distributions, we assumed that sharing between each pair of chromosomes is independent, whereas in practice, scans for IBD search for shared segments between all pairs in a cohort. Indeed, as studied in detail by Carmi et al. (2013), while sharing between two pairs in a cohort is only weakly dependent, the cumulative effect increases the observed variance of the amount of overall sharing. Therefore, more work will be needed to understand the distribution of IBD sharing within a cohort. Finally, we derived all results for a single chromosome. To apply the results genome-wide, we must assume inter-chromosome independence, which may not be well justified for the very recent past (Wakeley et al., 2012).

Turning to the quality of the renewal approximation itself, we verified using simulations that for chromosome-wide properties (e.g. the total number of segments), the renewal results are indistinguishable from SMC. We do, however, expect small deviations for very short chromosomes and for very small populations (e.g., see Fig. 9), when segments are few and long compared to the chromosome length and the distribution of tree heights does not reach stationarity. We also note that as opposed to SMC and SMC' (Marjoram and Wall, 2006), the renewal approximation introduces an asymmetry between the two ends of the chromosome: while the segment at the left end has distribution $\psi(\ell)$, the segment at the right end has the distribution of the 'age' of the process (see Karlin and Taylor, 1975 for more details). As we explained in Section 2.2, the number of segments is typically so large that this has a negligible effect. However, one can also formulate a *stationary* renewal process, which begins at coordinate $-\infty$, while observations begin at the origin (Karlin and Taylor, 1975). With some effort, we could rederive all results under the stationary process (not shown).

Our results have consequences for demographic inference. Current approaches rely on the assumption that recombination events terminate shared segments, as in SMC (Palamara et al., 2012; Ralph and Coop, 2013). Using our results, the more accurate SMC' can now be used, particularly for small populations. The distribution of the number of shared segments is also expected to be useful, as we briefly demonstrated (Fig. 8). The case we studied is simple, and would have been easily solved by other methods (e.g., Palamara et al., 2012; Carmi et al., 2013). Nevertheless, our approach has the attractive feature of providing a maximum-likelihood estimator (under the assumptions discussed above). Of course, for either large populations or for the very remote past, long IBD segments are scarce and our method, like any other IBD-based estimator, will have limited power.

Another drawback of our method is that it requires a numerical Laplace transform inversion, and for complex demographies, even the Laplace space solution will have to be numerically computed. Nevertheless, computationally, this is not very different from any method based on results specified as integrals or sums. The inverse transform (at least for the distribution of the number of shared segments) was simple to compute and reliable, as we validated by simulations (e.g., Fig. 7), as well as by comparing a number of inversion methods (not shown). Running time was reasonably short, at ≈2.5 s for each N on a standard machine. We anticipate that using the results for the fraction of the chromosome in shared segments (Section 4.2) will have more limited applications, due to the need for a double Laplace transform inversion. But we also note that, as we showed in Sections 4.1.2, 4.1.3, 4.2.2 and 4.2.3, standard Laplace transform techniques allow insight into the moments of the examined distributions. The Laplace transform method is ideal for problems of Markovian evolution in time or sequence space that are otherwise difficult (e.g., Lohse et al., 2011), and is therefore expected to be of future interest in population genetics.

We foresee a number of future directions and potential extensions. First, it would be useful (e.g., for demographic inference) to have analytical forms for simple non-constant demographies, such as exponential expansions and bottlenecks. Second, while we provided an equation for $\psi_{\text{SMC'}}(\ell)$, the PDF of segment lengths in SMC' (Eq. (10)), we did not investigate the corresponding renewal approximation (beyond the infinite-chromosome means), which should be feasible, since all of our renewal-based results are given in terms of a general segment length distribution. This is expected to rise in importance with the increasing popularity of SMC' (e.g., Harris and Nielsen, 2013) and the emerging understanding that it provides a much better approximation to the coalescent with recombination than SMC (e.g., Hobolth and Jensen (2014)). Another potential future application is pedigree reconstruction using IBD segments (Huff et al., 2011; Henn et al., 2012). For example, for (half-) cousins separated by $2k$ meioses, the segment length distribution will be a superposition of an exponential with rate $2k$, with probability 2^{-2k}, and $\psi_{\text{SMC}}(\ell)$ or $\psi_{\text{SMC'}}(\ell)$ otherwise (Eqs. (4) and (10), respectively). Finally, a challenging extension will be to sharing between more than two chromosomes. Potentially interesting applications are awaiting, as methods for the detection of such segments have been developed (Gusev et al., 2011; Moltke et al., 2011; He, 2013), and the resulting information is expected to improve the accuracy of demographic inference, natural selection detection, and disease mapping.

Acknowledgments

Appendix A. Full expressions for SMC' results

In this section, we provide full expressions for a number of SMC' quantities that were expressed as integrals in the main text.

The full expression for the distribution of segment lengths (Eq. (10), Section 2.4) is

$$
\psi_{\text{SMC'}}(\ell) = \frac{\int_0^\infty e^{-t}\lambda^2(t)e^{-\lambda(t)\ell}dt}{\int_0^\infty e^{-t}\lambda(t)dt}
$$

$$
= 3e^{-\frac{q}{2}(1+2\pi i)}q^{-\frac{q+1}{2}}\left[64\ell q\Gamma\left(\frac{1-q}{2}\right)^2\right]^{-1}
$$

$$
\times\left\{\pi^2 q^2 e^{i\pi q}q^{\frac{q+1}{2}}\sec^2\left(\frac{\pi q}{2}\right)\right.
$$

$$
\times\left[4\,{}_2\tilde{F}_2\left(\frac{q+1}{2},\frac{q+1}{2};\frac{q+3}{2},\frac{q+3}{2};\frac{q}{2}\right)\right.
$$

$$
-(q+1)^2\,{}_2\tilde{F}_2\left(\frac{q+3}{2},\frac{q+3}{2};\frac{q+5}{2},\frac{q+5}{2};\frac{q}{2}\right)
$$

$$
+4\Gamma\left(\frac{q+1}{2}\right){}_3\tilde{F}_3\left(\frac{q+1}{2},\frac{q+1}{2},\frac{q+1}{2};\right.
$$

$$
\left.\left.\frac{q+3}{2},\frac{q+3}{2},\frac{q+3}{2};\frac{q}{2}\right)\right]+i2^{\frac{q+3}{2}}e^{\frac{i\pi q}{2}}\Gamma\left(\frac{1-q}{2}\right)
$$

$$
\times\left[\Gamma\left(\frac{1-q}{2}\right)\left[q^2\Gamma\left(\frac{q+1}{2},-\frac{q}{2}\right)\right.\right.
$$

$$
\left.\left.+4q\Gamma\left(\frac{q+3}{2},-\frac{q}{2}\right)+4\Gamma\left(\frac{q+5}{2},-\frac{q}{2}\right)\right]
$$

$$
\left.\left.-\pi\sec\left(\frac{\pi q}{2}\right)\left(4q^2+6q+3\right)\right]\right\}, \tag{A.1}
$$

where $q = N\ell$ and ${}_a\tilde{F}_b$ is the regularized generalized hypergeometric function (Weisstein, 2014). See simulation results in Fig. 5. Note that $\psi_{\text{SMC'}}(\ell)$ is necessary real (and similarly below).

The full expression for the mean number of shared segments longer than m (Eq. (21), Section 3.2) is

$$
\langle n_m\rangle_{\text{SMC'}} = L\int_0^\infty \lambda(t)e^{-t-\lambda(t)m}dt
$$

$$
= LNi\frac{\left(\frac{-eM}{2}\right)^{-\frac{M}{2}}}{2\sqrt{2M}}
$$

$$
\times\left\{\Gamma\left(\frac{M+1}{2},-\frac{M}{2}\right)+\frac{2}{M}\Gamma\left(\frac{M+3}{2},-\frac{M}{2}\right)\right.
$$

$$
+\Gamma\left(\frac{M+1}{2}\right)\left[\psi^0\left(\frac{M+1}{2}\right)-2-i\pi\right.
$$

$$
\left.\left.-\ln\frac{M}{2}-\frac{1}{M}\right]-G_{2,3}^{3,0}\left(-\frac{M}{2}\left|\begin{matrix}1,1\\0,0,\frac{M+1}{2}\end{matrix}\right.\right)\right\}, \tag{A.2}
$$

where $M = mN$, G is the Meijer G-function (Weisstein, 2014), Γ (with two arguments) is the incomplete Gamma function, and ψ^0 is the digamma function.

The full expression for the mean fraction of the chromosome in segments longer than m (Eq. (22), Section 3.2) is

$$
\langle f_m\rangle_{\text{SMC'}} = \frac{\left(\frac{-eM}{2}\right)^{-\frac{M}{2}}}{2\sqrt{2M}}\left\{\frac{M^{3/2}}{\sqrt{2}}G_{2,3}^{3,0}\left(-\frac{M}{2}\left|\begin{matrix}\frac{1}{2},\frac{1}{2}\\-\frac{1}{2},-\frac{1}{2},\frac{M}{2}\end{matrix}\right.\right)\right.
$$

$$
+i(M+2)\Gamma\left(\frac{M+1}{2},-\frac{M}{2}\right)
$$

$$+ 2i\Gamma\left(\frac{M+3}{2}, -\frac{M}{2}\right)$$

$$+ iM\Gamma\left(\frac{M+1}{2}\right)\left[\psi^0\left(\frac{M+1}{2}\right)\right.$$

$$\left.\left. - \ln\frac{M}{2} - 2 - i\pi - \frac{3}{M}\right]\right\}, \qquad (A.3)$$

where $M = mN$. See simulations results in Fig. 6.

Appendix B. Full expression for the renewal theory results

In the renewal approximation to SMC, the distribution of the number of segments longer than m, in Laplace space (Eq. (32); Section 4.1.1), is

$$\tilde{P}(n_m = k, s) = C^{-2}2^{3-n}N^2 e^{-ms}\left[se^{\frac{s}{2N}}\mathrm{Ei}\left(-\frac{s}{2N}\right) + 2N\right]$$

$$\times \left\{s^2 e^{\frac{s}{2N}}\left[E_1\left(\frac{s}{2N}\right) - E_1\left(\frac{sC}{2N}\right)\right] + 2D - 2Ns\right\}^{-2}$$

$$\times \left\{\frac{s^2 e^{\frac{s}{2N}}\mathrm{Ei}\left(-\frac{sC}{2N}\right) + 2D}{\frac{s^2}{2}e^{\frac{s}{2N}}\left[\Gamma\left(0, \frac{s}{2N}\right) - \Gamma\left(0, \frac{sC}{2N}\right)\right] - Ns + D}\right\}^{n-1} \qquad (B.1)$$

for $k > 0$ and

$$\tilde{P}(n_m = 0, s)$$

$$= \left\{\frac{4N^2 C^{-1}}{e^{ms}C\left[2N - se^{\frac{s}{2N}}\left[E_1\left(\frac{s}{2N}\right) - E_1\left(\frac{sC}{2N}\right)\right]\right] - 2N} + s\right\}^{-1} \qquad (B.2)$$

for $k = 0$, where $C = 1 + 2mN, D = Ne^{-ms}(sC - 2N)/C^2$ and E_1 is related to the exponential integral function ($E_1(x) = -E_i(-x)$) (Weisstein, 2014).

The distribution of the fraction of the chromosome in segments longer than m, in Laplace space (Eq. (38), Section 4.2.1), is

$$\tilde{P}_{L_m}(u, s) = A/(1 - B), \qquad (B.3)$$

where A is given by

$$4C^2 A = \frac{4e^{-mr}}{r} - \frac{4e^{-ms}}{s} + \frac{4}{s} + \frac{2\left(1 - e^{-ms}\right)}{N}$$

$$+ \frac{e^{-mr}}{N^2 r}\left[C^2 r^2 e^{\frac{Cr}{2N}}\mathrm{Ei}\left(-\frac{Cr}{2N}\right) + 2N(Cr - 2N)\right]$$

$$+ \frac{4e^{-ms}}{Ns}\left\{\frac{s}{2}\left(e^{ms} - 1\right) + 2m^2 N^2 s e^{ms}\right.$$

$$\left. + N\left[e^{ms}(2ms - 1) + 1 - ms\right]\right\}$$

$$+ \frac{sC^2 e^{\frac{s}{2N}}}{N^2}\left[\Gamma\left(0, \frac{sC}{2N}\right) - \Gamma\left(0, \frac{s}{2N}\right)\right], \qquad (B.4)$$

B is given by

$$4N^2 C^2 B = 4N^2[4m^2 N^2 + 2mN(2 - ms)$$

$$+ 1 - 2ms + e^{-ms}\left(ms - 1\right)] + 2Ns\left(e^{-ms} - 1\right)$$

$$+ C^2 s^2 e^{\frac{s}{2N}}\left[\Gamma\left(0, \frac{s}{2N}\right) - \Gamma\left(0, \frac{sC}{2N}\right)\right]$$

$$- e^{-mr}\left[C^2 r^2 e^{\frac{Cr}{2N}}\mathrm{Ei}\left(-\frac{Cr}{2N}\right) + 2N(Cr - 2N)\right], \qquad (B.5)$$

$C = 1 + 2mN$, and $r = s + u$.

References

Atzmon, G., Hao, L., Pe'er, I., Velez, C., Pearlman, A., Palamara, P.F., Morrow, B., Friedman, E., Oddoux, C., Burns, E., Ostrer, H., 2010. Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters with shared Middle Eastern ancestry. Am. J. Hum. Genet. 86, 850–859.

Botigué, L.R., Henn, B.M., Gravel, S., Maples, B.K., Gignoux, C.R., Corona, E., Atzmon, G., Burns, E., Ostrer, H., Flores, C., Bertranpetit, J., Comas, D., Bustamante, C.D., 2013. Gene flow from North Africa contributes to differential human genetic diversity in Southern Europe. Proc. Natl. Acad. Sci. USA 110, 11791–11796.

Brančík, L., 2011. Numerical Inverse Laplace Transforms for Electrical Engineering Simulation, MATLAB for Engineers − Applications in Control, Electrical Engineering, IT and Robotics. InTech.

Bray, S.M., Mulle, J.G., Dodd, A.F., Pulver, A.E., Wooding, S., Warren, S.T., 2010. Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. Proc. Natl. Acad. Sci. USA 107, 16222–16227.

Brown, M.D., Glazner, C.G., Zheng, C., Thompson, E.A., 2012. Inferring coancestry in population samples in the presence of linkage disequilibrium. Genetics 190, 1447–1460.

Browning, B.L., Browning, S.R., 2011. A fast, powerful method for detecting identity by descent. Am. J. Hum. Genet. 88, 173–182.

Browning, S.R., Browning, B.L., 2012. Identity by descent between distant relatives: detection and applications. Annu. Rev. Genet. 46, 615–631.

Browning, B.L., Browning, S.R., 2013a. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194, 459–471.

Browning, S.R., Browning, B.L., 2013b. Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. Hum. Genet. 132, 129–138.

Browning, S.R., Thompson, E.A., 2012. Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190, 1521–1531.

Carmi, S., Palamara, P.F., Vacic, V., Lencz, T., Darvasi, A., Pe'er, I., 2013. The variance of identity-by-descent sharing in the wright-fisher model. Genetics 193, 911–928.

Chapman, N.H., Thompson, E.A., 2003. A model for the length of tracts of identity by descent in finite random mating populations. Theor. Popul. Biol. 64, 141–150.

de Hoog, F.R., Knight, J.H., Stokes, A.N., 1982. An improved method for numerical inversion of Laplace transforms. SIAM. J. Sci. Stat. Comput. 3, 357–366. code by K. J. Hollenbeck, INVLAP.M: A Matlab function for numerical inversion of Laplace transforms by the de Hoog algorithm, 1998.

Fisher, R.A., 1954. A fuller theory of "junctions" in inbreeding. Heredity 8, 187–197.

Gauvin, H., Moreau, C., Lefebvre, J.F., Laprise, C., Vezina, H., Labuda, D., Roy-Gagnon, M.H., 2014. Genome-wide patterns of identity-by-descent sharing in the French Canadian founder population. Eur. J. Hum. Genet. 22, 814–821.

Godrèche, C., Luck, J.M., 2001. Statistics of the occupation time of renewal processes. J. Stat. Phys. 104, 489–524.

Griffiths, R.C., Marjoram, P., 1996. Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3, 479–502.

Griffiths, R.C., Marjoram, P., 1997. An ancestral recombination graph. In: Donnelly, P., Tavaré, S. (Eds.), Progress in Population Genetics and Human Evolution. In: IMA Volumes in Mathematics and its Applications, vol. 87. Springer-Verlag, Berlin, pp. 257–270.

Griffiths, R.C., Tavare, S., 1994. Sampling theory for neutral alleles in a varying environment. Phil. Trans. R Soc. B 344, 403–410.

Gusev, A., Kenny, E.E., Lowe, J.K., Salit, J., Saxena, R., Kathiresan, S., Altshuler, D.M., Friedman, J.M., Breslow, J.L., Pe'er, I., 2011. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. 88, 706–717.

Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M., Pe'er, I., 2009. Whole population, genome-wide mapping of hidden relatedness. Genome Res. 19, 318–326.

Han, L., Abney, M., 2013. Using identity by descent estimation with dense genotype data to detect positive selection. Eur. J. Hum. Genet. 21, 205–211.

Harris, K., Nielsen, R., 2013. Inferring demographic history from a spectrum of shared haplotype lengths. PLoS Genet. 9, e1003521.

He, D., 2013. IBD-Groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise IBD relationships. Bioinformatics 29, i162–i170.

Henn, B.M., Hon, L., Macpherson, J.M., Eriksson, N., Saxonov, S., Pe'er, I., Mountain, J.L., 2012. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. PLoS One 7, e34267.

Hobolth, A., Jensen, J.L., 2014. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theor. Popul. Biol..

Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23, 183–201.

Huff, C.D., Witherspoon, D.J., Simonson, T.S., Xing, J., Watkins, W.S., Zhang, Y., Tuohy, T.M., Neklason, D.W., Burt, R.W., Guthery, S.L., Woodward, S.R., Jorde, L.B., 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. 21, 768–774.

Huff, C.D., Xing, J., Rogers, A.R., Witherspoon, D., Jorde, L.B., 2010. Mobile elements reveal small population size in the ancient ancestors of Homo sapiens. Proc. Natl. Acad. Sci. USA 107, 2147–2152.

Karlin, S., Taylor, H.M., 1975. A First Course in Stochastic Processes, second ed.. Academic Press.

Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D.F., Stefansson, H., Stefansson, K., 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nat. Genet. 9, 1068–1075.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496.

Lin, R., Charlesworth, J., Stankovich, J., Perreau, V.M., Brown, M.A., Taylor, B.V., 2013. Identity-by-descent mapping to detect rare variants conferring susceptibility to multiple sclerosis. PloS One 8, e56379.

Lohse, K., Harrison, R.J., Barton, N.H., 2011. A general method for calculating likelihoods under the coalescent process. Genetics 189, 977–987.

Marjoram, P., Wall, J.D., 2006. Fast "coalescent" simulation. BMC Genet. 7, 16.

McVean, G.A.T., Cardin, N.J., 2005. Approximating the coalescent with recombination. Philos. Trans. R. Soc. B 360, 1387–1393.

Moltke, I., Albrechtsen, A., Hansen, T.V., Nielsen, F.C., Nielsen, R., 2011. A method for detecting IBD regions simultaneously in multiple individuals–with applications to disease genetics. Genome Res. 21, 1168–1180.

Moorjani, P., Patterson, N., Loh, P.R., Lipson, M., Kisfali, P., Melegh, B.I., Bonin, M., Kadasi, L., Riess, O., Berger, B., Reich, D., Melegh, B., 2013. Reconstructing Roma history from genome-wide data. PloS One 8, e58633.

Palamara, P.F., Lencz, T., Darvasi, A., Pe'er, I., 2012. Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. 91, 809–822.

Palamara, P.F., Pe'er, I., 2013. Inference of historical migration rates via haplotype sharing. Bioinformatics 29, i180–i188.

Palin, K., Campbell, H., Wright, A.F., Wilson, J.F., Durbin, R., 2011. Identity-by-descent-based phasing and imputation in founder populations using graphical models. Genet. Epidemiol. 35, 853–860.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575.

Ralph, P., Coop, G., 2013. The geography of recent genetic ancestry across Europe. PLoS Biol. 11, e1001555.

Stam, P., 1980. The distribution of the fraction of the genome identical by descent in finite random mating populations. Genet. Res. 35, 131–155.

Thompson, E.A., 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194, 301–326.

Wakeley, J., King, L., Low, B.S., Ramachandran, S., 2012. Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent. Genetics 190, 1433–1445.

Weisstein, E.W., 2014. MathWorld — a Wolfram web resource.
URL http://mathworld.wolfram.com.

Wiuf, C., Hein, J., 1999. Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248–259.

Wolfram Research, I., 2012. Mathematica, version 9.0. Wolfram Research, Inc., Champaign, Illinois.

Zheng, C., Kuhner, M.K., Thompson, E.A., 2014. Bayesian inference of local trees along chromosomes by the sequential Markov coalescent. J. Mol. Evol. 78, 279–292.