A Renewal Theory Approach to IBD Sharing

Theor. Popul. Biol., **97**, 35-48 (2014)

Shai Carmi[1], Peter R. Wilton[2], John Wakeley[2], Itsik Pe'er[1]

[1] Department of Computer Science, Columbia University [2] Department of Organismic and Evolutionary Biology, Harvard University

Abstract

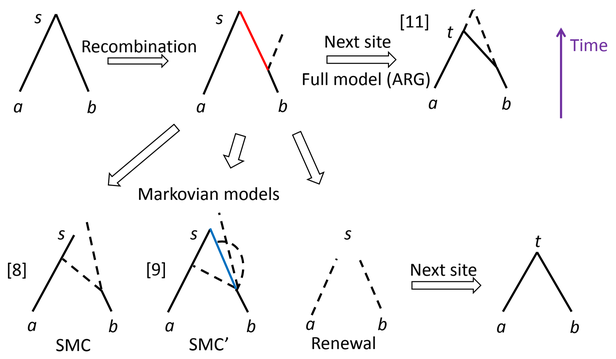
A long genomic segment inherited by a pair of individuals from a single, recent common ancestor is said to be *identical-by-descent* (IBD) [1,2]. Shared IBD segments have numerous applications in genetics, from demographic inference [3,4] to phasing/imputation [5], pedigree reconstruction [6], and disease mapping [7].

We provide a theoretical analysis of IBD sharing under Markovian approximations of the coalescent with recombination (SMC [8] and SMC' [9]). We describe a general framework for the IBD process, as well as introduce the *renewal approximation*, under which lengths of successive segments are independent.

In the infinite-chromosome limit, we recover previous results (for SMC [3,4,10]) 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 expressions (in Laplace space) for the distribution of those quantities. We obtain explicit expressions for the first two moments and demonstrate implications for demographic inference.

Finally, we generalize all results to populations with a variable effective size and to sharing between three chromosomes.

Models

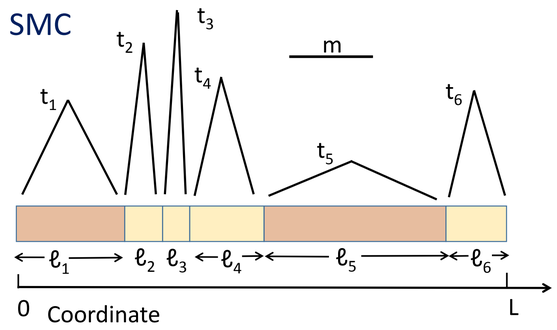


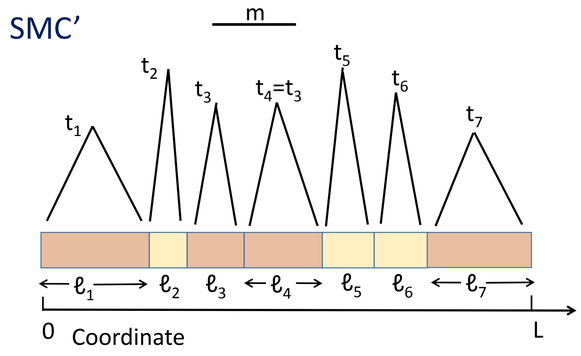
Assumptions:

- Population of size N (haploids) under the coalescent.
- Two chromosomes of length L.
- Recombination is a Poisson process along the chromosome (rate 1 per Morgan)
- A segment is IBD if the two chromosomes share the same TMRCA (time to most recent common ancestor) over sequence of length $>m$.
- Ignoring recent mutations, genotyping errors, etc.

The Renewal Approximation:

- Successive segment lengths are independent of each other.
- The distribution of their lengths is given by the stationary distribution $\psi(\ell)$.
- Justified based on simulations.





An illustration of the IBD process along the chromosome under SMC and SMC'. Recombination events are shown as vertical bars. The TMRCA is shown on top of each segment. Given a TMRCA t_i at segment i, the sequence length until the next recombination event, ℓ_i, is distributed exponentially with rate $2Nt_i$. Under SMC, recombination events necessarily change the TMRCA, and by that terminate the IBD segment. Under SMC', the TMRCA may remain the same after recombination, extending the IBD segment at least until the next recombination event. The minimal segment length, m, is shown as a horizontal bar. IBD segments longer than m are shown in orange (the others are in yellow). In this example, for SMC, we have $n_m = 2$ and $f_m = (\ell_1 + \ell_5)/L$. For SMC', we have $n_m = 3$ and $f_m = [\ell_1 + (\ell_3 + \ell_4) + \ell_7]/L$.

Key quantities

- The total number of IBD segments: $\boldsymbol{n_0}$.
- The number of IBD segments longer than m: $\boldsymbol{n_m}$.
- The fraction of the chromosome found in IBD segments longer than m: $\boldsymbol{f_m}$.
- The distribution of IBD segment lengths: $\boldsymbol{\psi(\ell)}$.

Deriving the segment length distribution

- Given the TMRCA at the previous IBD segment, s, the TMRCA at the new segment, t, is given by $q(t|s)$, which can be seen as the transition probability of a Markov chain. The stationary distribution of the chain (i.e., the distribution of TMRCAs at IBD segments) is $\pi(t)$.
- Under SMC', the probability that a recombination event will change the TMRCA is $p_{change} = (2t + 1 - e^{-2t})/4t$. Integrating over all t, exactly 1/3 of the recombination events do not change the TMRCA.
- Given the TMRCA t, the IBD segment length, $\psi(\ell|t)$, is exponential with rate $2Nt$ under SMC, or $2Ntp_{change} \equiv N\lambda(t)$ under SMC'.
- The unconditional distribution of segment lengths, $\psi(\ell)$, is obtained by integrating over all times. The mean segment length is $\langle\ell\rangle$. Segments in SMC' are longer than in SMC.

SMC

$$q_{SMC}(t|s) = \begin{cases} (1-e^{-t})/s & t < s, \\ \left(e^{-(t-s)} - e^{-t}\right)/s & t > s. \end{cases} \quad [12]$$

$$\pi_{SMC}(t) = te^{-t}$$

$$\psi_{SMC}(\ell|t) = 2Nte^{-2Nt\ell}$$

$$\psi_{SMC}(\ell) = \int_0^\infty \pi_{SMC}(t)\psi_{SMC}(\ell|t)dt = \frac{4N}{(1+2N\ell)^3}$$

$$\langle\ell\rangle_{SMC} = (2N)^{-1} \quad [3]$$

SMC'

$\lambda(t) = (2t + 1 - e^{-2t})/2$

$$q_{SMC'}(t|s) = \begin{cases} (1-e^{-2t})/\lambda(s) & t < s, \\ \left(e^{-(t-s)} - e^{-(t+s)}\right)/\lambda(s) & t > s. \end{cases}$$

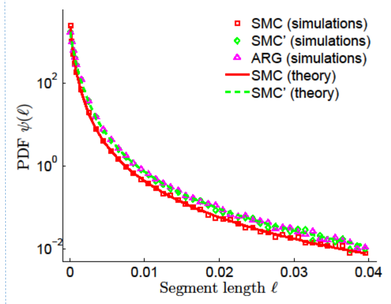
$$\pi_{SMC'}(t) = \frac{3}{4}e^{-t}\lambda(t)$$

$$\psi_{SMC'}(\ell|t) = N\lambda(t)e^{-N\lambda(t)\ell}$$

$$\psi_{SMC'}(\ell) = \int_0^\infty \pi_{SMC'}(t)\psi_{SMC'}(\ell|t)dt$$

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

Simulations



Infinite-chromosome means

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

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

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

$$\langle n_0\rangle_{SMC} = 2NL \quad [3]$$

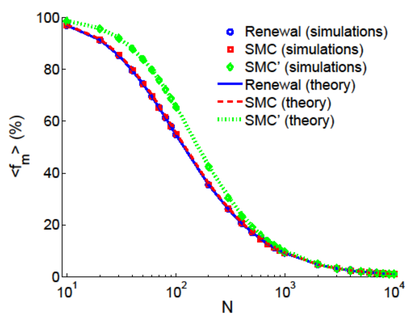
$$\langle n_m\rangle_{SMC} = \frac{2NL}{(1+2mN)^2}$$

$$\langle f_m\rangle_{SMC} = \frac{1+4mN}{(1+2mN)^2}$$

$$\langle n_0\rangle_{SMC'} = 4NL/3$$

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

$$\langle f_m\rangle_{SMC'} = \int_0^\infty [1+Nm\lambda(t)]e^{-t-Nm\lambda(t)}dt$$



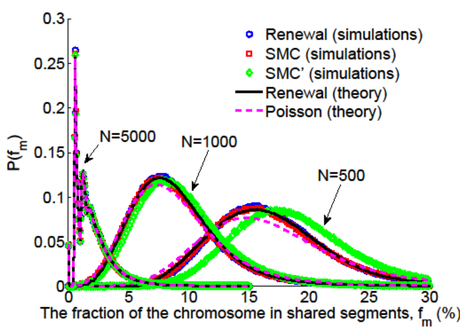
Distributions

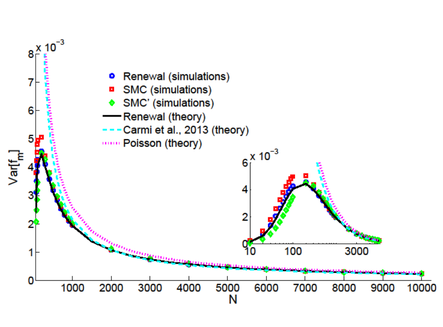
The number of IBD segments longer than m:

- Denote $P(n_m = k, L)$ the distribution of the number of shared segments, n_m.
- Denote $\tilde{P}(n_m = k, s)$ the Laplace transform ($L \to s$) of $P(n_m = k, L)$.
- We have: $\tilde{P}(n_m = k, s) = \begin{cases} \frac{\tilde{\phi}_{<m}(s)}{1-\tilde{\psi}_{<m}(s)} & k = 0, \\ \frac{[1-\tilde{\psi}(s)][\tilde{\psi}_{>m}(s)+s\tilde{\phi}_{>m}(s)]}{s[1-\tilde{\psi}_{<m}(s)]^2}\left[\frac{\tilde{\psi}_{>m}(s)}{1-\tilde{\psi}_{<m}(s)}\right]^{k-1} & k > 0. \end{cases}$
- Definitions: $\tilde{\psi}(s) = \int_0^\infty e^{-s\ell}\psi(\ell)d\ell$ is the Laplace transform of $\psi(\ell)$; $\tilde{\psi}_{<m}(s) = \int_0^m e^{-s\ell}\psi(\ell)d\ell$; $\tilde{\psi}_{>m}(s) = \int_m^\infty e^{-s\ell}\psi(\ell)d\ell$; $\phi(\ell) = \int_\ell^\infty \psi(\ell')d\ell'$ is the probability of a segment length to be longer than ℓ; $\tilde{\phi}_{<m}(s) = \int_0^m e^{-s\ell}\phi(\ell)d\ell$; $\tilde{\phi}_{>m}(s) = \int_m^\infty e^{-s\ell}\phi(\ell)d\ell$.
- For large N and L, $\mathrm{Var}[n_m]_{SMC} \approx L/2m^2N$.

The fraction of the chromosome in segments longer than m:

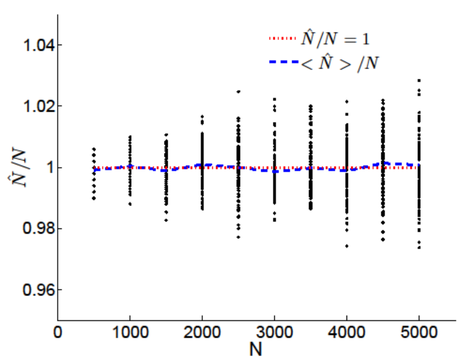
- Denote $L_m = Lf_m$ and $P(L_m, L)$ the density of L_m.
- Denote by $\tilde{P}_{L_m}(u, s)$ the double Laplace transform ($L_m \to u, \mathrm{L} \to s$) of $P(L_m, L)$.
- We have: $\tilde{P}_{L_m}(u,s) = \frac{\frac{1}{s}\frac{\tilde{\psi}_{<m}(s)}{s} + \phi(m)\left[\frac{e^{-m(s+u)}}{s+u} - \frac{e^{-ms}}{s}\right] - \frac{\tilde{\psi}_{>m}(s+u)}{s+u}}{1-\tilde{\psi}_{<m}(s)-\tilde{\psi}_{>m}(s+u)}$. $P(f_m) = LP(L_m, L)$.
- For large N and L, $\mathrm{Var}[f_n]_{SMC} \approx [\ln(L/m) - 1/2]/(NL)$.





Demographic inference:

- For each simulation of 5000 chromosome pairs under SMC, we computed the likelihood of the distribution of the number of IBD segments under different values of the population size N.
- The maximum-likelihood estimator is unbiased and has SD≈0.01N.

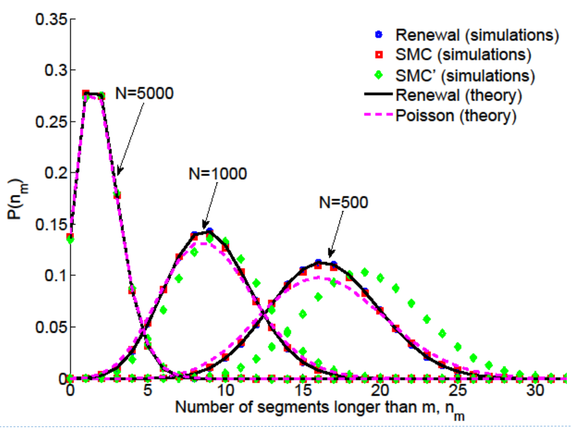


Simulations:

- All simulations were under L=2 and m=0.01 Morgans, with 10^6 repeats per data point.

The Poisson approximation:

- Ref. [3] suggested that n_m is Poisson with mean $\langle n_m\rangle_{SMC} = 2NL/(1+2mN)^2$.
- Good approximation for $N \gtrsim 1000$.



References

1. Thompson, E. A., 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194, 301.
2. Browning, S. R., Browning, B. L., 2012. Identity by descent between distant relatives: Detection and applications. Annu. Rev. Genet. 46, 615.
3. Palamara, P. F., Lencz, T., Darvasi, A., Pe'er, I., 2012. Length distributions of identity by descent reveal ne-scale demographic history. Am. J. Hum. Genet. 91, 809.
4. Ralph, P., Coop, G., 2013. The geography of recent genetic ancestry across Europe. PLoS Biol. 11, e1001555.
5. Palin, K., Campbell, H., Wright, et al., 2011. Identity-by-descent-based phasing and imputation in founder populations using graphical models. Genet. Epidemiol. 35, 853.
6. Huff, C. D., Witherspoon, D. J., Simonson, et al., 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome Res. 21, 68.
7. Browning, S. R., Thompson, E. A., 2012. Detecting rare variant associations by identity-by-descent mapping in case-control studies. Genetics 190, 1521.
8. McVean, G. A. T., Cardin, N. J., 2005. Approximating the coalescent with recombination. Phil. Trans. R. Soc. B 360, 1387.
9. Marjoram, P., Wall, J. D., 2006. Fast "coalescent" simulation. BMC Genetics 7, 16.
10. Carmi, S., Palamara, P. F., Vacic, V., et al., 2013. The variance of identity-by-descent sharing in the wright-Fisher model. Genetics 193, 911.
11. Wiuf, C., Hein, J., 1999. Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248.
12. Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493.

Extensions: variable population size and three chromosomes

- Since all of our results depend on $\psi(\ell)$ alone, they are easily generalizable to variable population size.
- Define $N(t) = N_0 v(t) = N_0/h(t)$.
- We have: $\psi_{SMC}(\ell) = 2N_0 \frac{\int_0^\infty t^2 h(t) e^{-\int_0^t h(\tau)d\tau - 2N_0 t\ell}dt}{\int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt}$ and $\langle n_0\rangle_{SMC} = 2N_0 L\int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt$.
- For SMC', define $\lambda(t) = t + e^{-2\int_0^t h(\tau)d\tau}\int_0^t e^{2\int_0^{t'} h(\tau)d\tau}dt'$.
- We have: $\psi_{SMC'}(\ell) = N_0 \frac{\int_0^\infty [\lambda(t)]^2 h(t) e^{-\int_0^t h(\tau)d\tau - N_0\lambda(t)\ell}dt}{\frac{4}{3}\int_0^\infty e^{-\int_0^t h(\tau)d\tau}dt}$ and $\langle n_0\rangle_{SMC'} = 2\langle n_0\rangle_{SMC}/3$.
- Most results are generalizable to sharing between three chromosomes.
- For example: $\langle n_0\rangle_{SMC} = 3NL$, $\langle n_m\rangle_{SMC} = \frac{NL(3+4mN)}{(1+mN)^2(1+2mN)^2}$, $\langle f_m\rangle_{SMC} = \frac{1+6mN(1+mN)}{(1+mN)^2(1+2mN)^2}$, $\langle n_0\rangle_{SMC'} = 19NL/9$.