$$\tau(t) = \begin{cases} (t - t_{B1})/(2Na) + (t_{B1} - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t > t_{B1} \\ (t - t_{B2})/(2Nf) + t_{B2}/(2N) & \text{if } t_{B1} < t < t_{B2} \\ t/(2N) & \text{if } t < t_{B2} \end{cases}$$

In addition to scaling t, we must scale μ and ρ, since each contains a factor of N.

When we make these modifications, equation (3) becomes

$$\begin{aligned} p_L(\text{IBD}) &= \int_{\tau=0}^{\tau(t_{B2})} e^{-\tau\cdot(1+L(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B2})}^{\tau(t_{B1})} e^{-\tau\cdot(1+Lf(\mu+\rho))}d\tau + \int_{\tau=\tau(t_{B1})}^{\infty} e^{-\tau\cdot(1+La(\mu+\rho))}d\tau \\ &= \frac{1 - e^{-\tau(t_{B2})(1+L(\mu+\rho))}}{1 + L(\mu+\rho)} + \frac{e^{-\tau(t_{B2})(1+Lf(\mu+\rho))} - e^{-\tau(t_{B1})(1+Lf(\mu+\rho))}}{1 + Lf(\mu+\rho)} \\ &\quad + \frac{e^{-\tau(t_{B1})(1+La(\mu+\rho))}}{1 + La(\mu+\rho)}. \end{aligned}$$

In the same way, we can correct $A_1(L), \ldots, A_L(L)$ for the bottleneck by replacing

$$C_i = \frac{\rho}{i(\mu+\rho)(1 + i(\mu+\rho))}$$

with

$$\begin{aligned} C_i &= \frac{\rho}{i(\mu+\rho)} \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

In terms of these corrected $A_i(L)$, we deduce that

$$\begin{aligned} p_L(\text{IBS}) &= \sum_{i=1}^{L} A_i(L) \left(\frac{1 - e^{-\tau(t_{B2})(1+i(\mu+\rho))}}{1 + i(\mu+\rho)} \right. \\ &\quad \left. + \frac{e^{-\tau(t_{B2})(1+if(\mu+\rho))} - e^{-\tau(t_{B1})(1+if(\mu+\rho))}}{1 + if(\mu+\rho)} + \frac{e^{-\tau(t_{B1})(1+ia(\mu+\rho))}}{1 + ia(\mu+\rho)} \right). \end{aligned}$$

3 The age distribution of maximal IBD segments

Our calculations, along with those in earlier papers, make it clear that IBD segment length is inversely related to age. In [17], Hayes, et al. go as far

as to draw a one-to-one correspondence between the abundance of maximal c-centimorgan IBD segments and the effective size of the population $1/(1+4c)$ generations ago. However, we show here that the mean coalescence time of an L-base IBD tract ($O(1/L)$) is much less than its standard deviation ($O(1/\sqrt{L})$), implying that the segments coalescing at time t have a significant length spread, particularly when t is very ancient. This complicates the effect of population size changes on the distribution of ROH length, particularly for shorter ROHs. While Hayes, et al. studied the distribution of ROHs that were 10^6 to 10^7 base pairs long and found their assumption useful at that length scale, we find that the relationship between effective population size and ROH length is more complicated for shorter ROHs, as we will see corroborated by data in Section 7 (Figures 14, 15, and 16).

As we saw in Section 2, the probability of an L-base ROH being IBD is

$$\int_{t=0}^{\infty} e^{-t(1+L\rho)}dt,$$

while the proability that it will be maximally IBD (i.e. not contained in a larger IBD segment) is

$$\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt.$$

We can use this to calculate a joint distribution between IBD segment length and coalescence time:

$$p_L(t|\text{IBD}) = \frac{e^{-t(1+L\rho)}(1-e^{-t\rho})^2}{\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt}. \quad (8)$$

We compute that

$$\begin{aligned}\int_{t=0}^{\infty} e^{-t(1+L\rho)}(1-e^{-t\rho})^2dt &= \frac{1}{1+L\rho} - \frac{2}{1+(L+1)\rho} + \frac{1}{1+(L+2)\rho} \\ &= \frac{2\rho^2}{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)},\end{aligned}$$

such that

$$p_L(t|\text{IBD}) = \frac{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}{2\rho^2}e^{-t(1+L\rho)}(1-e^{-t\rho})^2. \quad (9)$$

Similarly, we can compute the expected t value $E_t(L)$, measured, as always, in units of $2N$ generations:

$$\begin{aligned} E_t(L) &= \int_{t=0}^{\infty} t p_L(t|\text{IBD}) dt \\ &= \frac{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}{\rho^2} \\ &\quad \cdot \left(\frac{1}{(1+L\rho)^2} - \frac{2}{1+(L+1)\rho)^2} + \frac{1}{(1+(L+2)\rho)^2} \right) \\ &= \frac{3L^2\rho^2 + 6L\rho^2 + 6L\rho + 2\rho^2 + 6\rho + 3}{(1+L\rho)(1+(L+1)\rho)(1+(L+2)\rho)}. \end{aligned}$$

This differs from $1/(1+L\rho)$, the value given by Hayes, et al., because they don't distinguish between maximal and non-maximal IBD.

We go on to compute the variance

$$\begin{aligned} E_{t^2}(L) - E_t(L)^2 &= \int_{t=0}^{\infty} t^2 p_L(t|\text{IBD}) dt - \left(\int_{t=0}^{\infty} t p_L(t|\text{IBD}) dt \right)^2 \\ &= \frac{12(L^3 + \rho^2 L^2 + 2\rho L + \rho^2 + \rho + 1)(3\rho^2 L^2 + 6\rho^2 L + 10\rho^2 + 6\rho + 3)}{(1+\rho L)^2(1+\rho(L+1))^2(1+(L+2))^2} \\ &\quad - \frac{12(\rho L + 1)(\rho L + \rho + 1)}{(1+\rho(L+1))^2(1+\rho(L+2))^2} \\ &\quad - \frac{(3L^2\rho^2 + 6L\rho^2 + 6L\rho + 2\rho^2 + 6\rho + 3)^2}{(1+L\rho)^2(1+(L+1)\rho)^2(1+(L+2)\rho)^2}. \end{aligned}$$

Looking at the leading terms, we note that

$$E_t(L) \approx \frac{3}{\rho L} \ll \sqrt{E_{t^2}(L) - E_t(L)^2} \approx \frac{6}{\rho^2 \sqrt{L}},$$

meaning that the standard deviation of $E_t(L)$ is much greater than its mean.

Figure 3 shows the length distribution of IBS segments that coalesce $0.2N$ generations ago, while Figure 4 plots the length distribution of segments that coalesce $0.3N$ generations ago. Even if IBD were the same as IBS and segments coalesced at only these two times, it would not be straightforward to look at a sum of plots like this and quantify an excess of one type of segment. Hayes, et al. track recent population growth by assuming that a dearth of L-base IBD segments means a larger population at time $1/(1+L\rho)$, but it would seem that this approach must be modified for shorter L where the length and coalescence

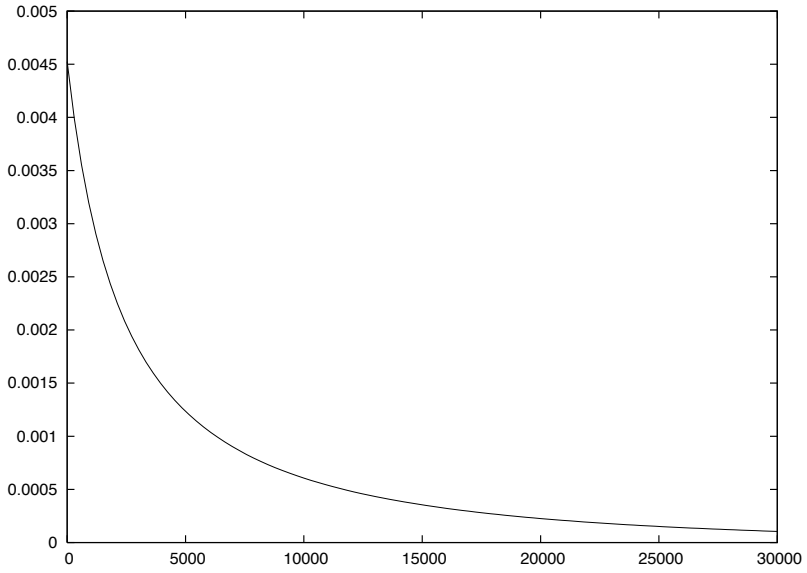

Figure 3: This plot shows the length spread of IBD segments that coalesce $0.2N$ generations ago. Comparing this to Figure 4, we see that it will be difficult to tell these segments apart from segments that coalesced $0.3N$ generations ago.

time are related so inexactly. We will see in Section 7, that precisely calculated IBS probabilities make it possible to use the distribution of shorter ROHs to estimate the effective population size at earlier points in history.

4 The probability of IBD given diploid IBS with uncertain haplotype phasing

In [31], Kong, et al. find IBS haplotypes by looking for diploid sequences L_1, L_2 with the property that $\text{IBS}(L_1, L_2) \geq 1$ at every base in the sequence, i.e. that the alignment contains no locus for which L_1 and L_2 are homozygous for different alleles. However this condition does not guarantee that a haplotype of L_1 is