of IBD sharing.

2 Computing the probability of identity by state

Since

$$p(\text{IBD}|\text{IBS}) = \frac{p(\text{IBD\&IBS})}{p(\text{IBS})}, \tag{2}$$

where p(IBD&IBS) is easy to compute (see equation (2.1)), the crux of our approach will be calculating p(IBS) given sequence length and the history of the effective population size. In section 2.1, we treat the case of constant effective population size, while section 2.3 describes how to condition on any locally constant population size history.

2.1 Constant effective population size

Let L be the length of an alignment between two haplotypes sampled at random from a diploid population of effective size N. Assume that the DNA undergoes m mutations per base per generation and r recombinations per base per generation, letting $\mu = 4Nm$ and $\rho = 4Nr$. We will hereafter measure time in units of $2N$ generations.

The alignment will coalesce at time t, both IBD and IBS, if and only if the following events coincide:

1. The leftmost locus coalesces at time t without mutating (probability $e^{-t(1+\mu)}dt$)
2. No other base in either sequence undergoes a mutation or a recombination between time zero and time t (probability $e^{-t(L-1)(\mu+\rho)}$)

From this observation, it follows that

$$p(\text{IBD\&IBS}) = \int_{t=0}^{\infty} e^{-t(L-1)(\mu+\rho)} \cdot e^{-t(1+\mu)}dt = \frac{1}{1 + L\mu + (L-1)\rho}. \tag{3}$$

In an analogous way, we will derive the probability $p_L(\text{IBS}|t)dt$ that the alignment coalesces IBS with its rightmost base coalescing at time t. We proceed

by induction on the length variable L, claiming that

$$\begin{aligned} p_L(\text{IBS}|t)dt &= p_{L-1}(\text{IBS}|t)dt \cdot e^{-t(\mu+\rho)} \\ &+ \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} p_{L-1}(\text{IBS}|t_0)e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dtdt_rdt_0 \\ &+ \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} p_{L-1}(\text{IBS}|t_0)e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dtdt_rdt_0. \end{aligned}$$

The dummy variable t_0 is the coalescence time of the base next to the rightmost one. The first term is the probability that no recombination occurs between the rightmost base of the alignment and the base next to it, while the second term (the first integral) is the probability that a recombination occured at some time t_r, and that t is greater than t_0. The third term accounts for the remaining possibilities, integrating over times t_0 that are greater than t.

It will be convenient to write

$$p_L(\text{IBS}|t) = \sum_{i=1}^{L} A_i(L)e^{-t(1+i\mu+(i-1)\rho)}dt \tag{4}$$

and solve for the coefficients $A_1(L), \cdots, A_L(L)$, which will not depend on t.

Since

$$p_1(\text{IBS}|t) = e^{-t(1+\mu)}dt$$

and

$$\begin{aligned} & e^{-t_0(1+i\mu+(i-1)\rho)}\cdot e^{-t(\mu+\rho)} + \\ & \int_{t_0=0}^{t}\int_{t_r=0}^{t_0} e^{-t_0(1+i\mu+(i-1)\rho)}e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dt_rdt_0 + \\ & \int_{t_0=t}^{\infty}\int_{t_r=0}^{t} e^{-t_0(1+i\mu+(i-1)\rho)}e^{-\mu t-\rho t_r}\cdot\rho e^{-(t-t_r)}dt_rdt_0 \\ = & \frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho)}e^{-t(\mu+1)} + \\ & \left(1-\frac{\rho}{i(\mu+\rho)(1+i\mu+(i-1)\rho)}\right)e^{-t(1+(i+1)\mu+i\rho))}, \end{aligned}$$

we can let

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

and conclude that

$$A_1(L) = \sum_{i=1}^{L-1} C_i A_i(L-1), \tag{5}$$

while

$$A_i(L) = (1 - C_{i-1}) A_{i-1}(L-1) \tag{6}$$

for $i > 1$.

Integrating equation (4) with respect to time, we find that

$$p_L(\text{IBS}) = \sum_{i=1}^{L} \frac{A_i(L)}{1 + i\mu + (i-1)\rho}. \tag{7}$$

Although it is time-intensive to compute $A_1(L), \ldots, A_L(L)$ for $L \gg 10^4$, the run time can be decreased by picking an appropriate constant c and substituting $(c\mu, c\rho, L/c)$ for (μ, ρ, L). This approximation reduces the run time c^2-fold, and Figure 2 records its modest effect on the computation accuracy.

The reader may prefer to think about $p_L(\text{IBS})$ using matrix algebra rather than recursion, seeing that

$$p_L(\text{IBS}) = \begin{pmatrix} \frac{1}{1+\mu} & \frac{1}{1+2\mu+\rho} & \cdots & \frac{1}{1+L\mu+(L-1)\rho} \end{pmatrix} \begin{pmatrix} C_1 & C_2 & \cdots & C_{L-1} & C_L \\ 1-C_1 & 0 & \cdots & 0 & 0 \\ 0 & 1-C_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1-C_{L-1} & 0 \end{pmatrix}^L \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It is important to note that we have been talking about ROHs that are *at least* L bases long; when we compare our results to real genome data in Section 7, we will need to know the frequency of ROHs that are *exactly* L bases long. The following is the probability $p_{L\max}(\text{IBS})$ of observing an L-base IBS stretch ending with a het:

$$\begin{aligned} p_{L\max}(\text{IBS}) &= \sum_{i=1}^{L} A_i(L) \int_{t=0}^{\infty} e^{-t(1+i\mu+(i-1)\rho))}(1 - e^{-(\mu+\rho)t})dt \\ &= \sum_{i=1}^{L} \frac{A_i(L)(\mu+\rho)}{(1+i\mu+(i-1)\rho)(1+(i+1)\mu+i\rho)}. \end{aligned}$$

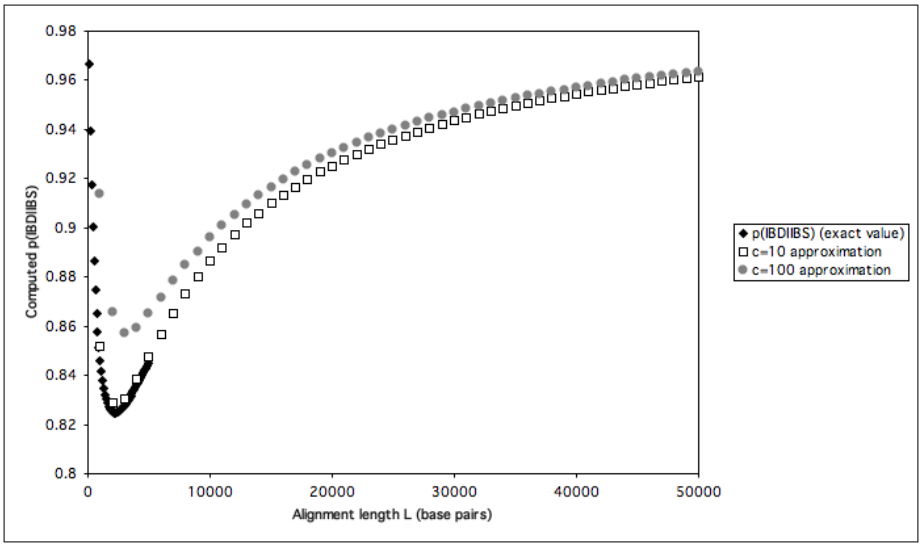

Figure 2: The parameter change $(\mu, \rho, L) \to (c\mu, c\rho, L/c)$ has its greatest effect when L is small. For $L = 1000$, the true value of $p(\text{IBD}|\text{IBS})$ is 0.8459; the calculated value increases to 0.8516 when we let $c = 10$, and increases to 0.9136 when we let $c = 100$. For $L = 50000$, the difference between the $c = 10$ value and the $c = 100$ value is only 0.0131, and taking $c = 100$ makes it practical to compute $p_L(\text{IBS})$ for L in the megabase range.

2.2 Non-uniform mutation and recombination

We have been assuming that μ and ρ are constant throughout the alignment to simplify the formulas as much as possible. However, it is easy to calculate p(IBD—IBS) exactly even when each locus i, $1 \leq i \leq n$, has a distinct mutation rate μ_i and recombination rate ρ_i. If we let $\vec{\mu}$ and $\vec{\rho}$ denote the vectors $(\mu_1, \ldots, \mu_n)$ and $(\rho_1, \ldots, \rho_n)$, it is easy to check (by generalizing the integrals in section 2.1) that

$$p_L(\text{IBS}|\vec{\mu}, \vec{\rho}) = \sum_{i=1}^{L} \frac{A_i(L, \vec{\mu}, \vec{\rho})}{((\mu_1 + \rho_1) + \cdots + (\mu_i + \rho_i))(1 + (\mu_1 + \rho_1) + \cdots + (\mu_{i-1} + \rho_{i-1}) + \mu_i)},$$

where

$$A_i(L, \vec{\mu}, \vec{\rho}) = (1 - C_{i-1}(\vec{\mu}, \vec{\rho}))A_{i-1}(L-1, \vec{\mu}, \vec{\rho})$$

and

$$A_1(L, \vec{\mu}, \vec{\rho}) = \sum_{i=1}^{L-1} C_i(\vec{\mu}, \vec{\rho})A_i(L-1, \vec{\mu}, \vec{\rho})$$

for

$$C_i(\vec{\mu}, \vec{\rho}) = \frac{\rho_i}{((\mu_1 + \rho_1) + \cdots + (\mu_i + \rho_i))(1 + (\mu_1 + \rho_1) + \cdots + (\mu_{i-1} + \rho_{i-1}) + \mu_i)}.$$

2.3 Correcting for changes in effective population size

Because most human populations have undergone growth and/or bottlenecking, we describe how to correct our model for historical changes in effective population size. We work through the example of a simple bottleneck, but the same method can accommodate any locally constant function $N(t)$.

We model a bottleneck following the convention in the coalescent theory reference [18], using a piecewise-constant time transform $t \to \tau(t)$. We suppose that the population began at size aN before the bottleneck, dipped to size fN during the time interval $[t_{B2}, t_{B1}]$, and has existed stably at size N from time t_{B2} to the present. The values t_{B1}, t_{B2}, and t_{B3} are measured in generations before the present, but we must map them to times $\tau(t)$ measured in units of $2N$ generations before the present:

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

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