TEXT S1  Painting Algorithm

Li and Stephens (2003) described a likelihood based model that captures key features of the genealogical process with recombination while remaining computationally tractable for large datasets. Under the model, a chromosome is generated chunk-by-chunk by ‘copying’ from a conditional set of fixed haplotypes. In our notation, every individual consists of two haploids, each consisting of a single phased haplotype per chromosome. The \( L \) total SNPs in each haploid are listed one chromosome at a time, in order within each chromosome.

Suppose that we wish to generate a particular haploid \( h^* = \{h^*_1, ..., h^*_L\} \), with \( h_{st} \) the observed allele of \( h_s \) at site \( l \), using \( j \) pre-existing donor haploids \( h_1, ..., h_j \). Let \( \vec{\rho} = \{\rho_1, ..., \rho_{L-1}\} \) be a vector of genetic distances, with \( \rho_l \) the population-scaled genetic distance between sites \( l \) and \( l+1 \) (i.e. \( \rho_l = N_e g_l \), where \( N_e \) is analogous to the “effective population size” and \( g_l \) is the genetic distance in Morgans between sites \( l \) and \( l+1 \)). (Between chromosomes, the genetic distance between the last site of the previous chromosome and the first site of the next chromosome is \( \infty \).) Let \( \vec{f} = \{f_1, ..., f_j\} \) be a vector of copying probabilities, with \( f_k \) the probability of copying from haploid \( h_k \) at any site. Let \( \theta \) correspond to a per site mutation (or “imperfect copying”) parameter. The conditional probability \( \Pr(h^*_l \mid h_1, ..., h_j; \vec{\rho}, \vec{f}, \theta) \) is structured as a Hidden Markov model. Let \( \vec{Y} = \{Y_1, ..., Y_L\} \) represent the hidden state sequence vector, with \( Y_l \) the existing haploid from the set \( h_1, ..., h_j \) that haploid \( h^*_s \) copies from at site \( l \). Switches in the haploid being copied between \( Y_l \) and \( Y_{l+1} \) occur as a Poisson process with rate \( \rho_l \). The transition probabilities for \( Y \) between sites \( l \) and \( l+1 \) are as follows (we exclude \( h_1, ..., h_j \) and the parameters from the left side of equations (1) and (2) below for ease of reading):

\[
\Pr(Y_{l+1} = y_{l+1} \mid Y_l = y_l) =
\begin{cases} 
\exp(-\rho_l) + (1 - \exp(-\rho_l)) f_{y_{l+1}} & \text{if } y_{l+1} = y_l; \\
(1 - \exp(-\rho_l)) f_{y_{l+1}} & \text{otherwise},
\end{cases}
\]

(1)

The observed state sequence component of the Hidden Markov Chain, the probability of observing a particular allele given the haploid that \( h^*_s \) is copying from at a given SNP, allows for “imperfect” copying:

\[
\Pr(h^*_l = a \mid Y_l = y) = \begin{cases} 
1.0 - \theta & h_{yl} = a; \\
\theta & h_{yl} \neq a.
\end{cases}
\]

(2)

Here \( h_{kl} \) refers to the allelic type of haploid \( k \) at SNP \( l \). To calculate \( \Pr(D) \equiv \Pr(h^*_s \mid h_1, ..., h_j; \vec{\rho}, \vec{f}, \theta) \), a summation is performed over all permutations of the copying process, i.e. a summation over all possible \( y \), which can be accomplished efficiently using the forward algorithm (e.g. Rabiner 1989).
For all analyses presented here, we fix the mutation parameter $\theta$ to Watterson’s estimate (Watterson 1975), as used by Li and Stephens (2003), i.e.

$$
\theta = \frac{1}{2} \frac{\left(\sum_{i=1}^{j} 1/i\right)^{-1}}{j + \left(\sum_{i=1}^{j} 1/i\right)^{-1}}
$$

for $j$ total haploids. We fix each $g_l$ by taking the build 36 genetic distance estimates from the HapMap website (http://www.hapmap.org), which were calculated using Phase II genotypes and averaging values across the three HapMap populations as described by the International HapMap Consortium (2007). We also fix each $f_k$ to be $1/j$ for $k = 1, \ldots, j$, allowing for equal a priori probability of copying from each conditional haploid.

Calculating expected number of chunks copied:

The average number of chunks copied to a haploid $*$ is a random variable denoted $\hat{x}_i = E_{l=1..L}(X_{il})$, where $X_{il}$ is the probability that a given locus $l$ is a new haplotypic segment copied from individual $i$. To calculate $\hat{x}_1, \ldots, \hat{x}_j$, the posterior expected number of chunks for which haploid $h_*$ copies from each of $h_1, \ldots, h_j$, respectively, we calculate $\hat{f}_{k,l}$, the probability haploid $h_*$ is copying from haploid $h_k$ at site $l$ given at least one “switch” has occurred between $l-1$ and $l$. Again excluding parameters for ease of reading, let $\alpha_{kl} = \Pr(h_1, \ldots, h_L, Y_l = h_k)$ and $\beta_{kl} = \Pr(h_{*(l+1)}, \ldots, h_L | Y_{l+1} = h_k)$. Then

$$
\hat{x}_k = \frac{\alpha_k \beta_k}{\Pr(D)} + \sum_{l=1}^{L-1} \left(\frac{1}{\Pr(D)} \left[ \alpha_k \beta_{k(l+1)} - \alpha_{kl} \beta_{k(l+1)} \right] \Pr(h_1, \ldots, h_L, Y_{l+1} = h_k) \exp(-\rho_l) \right)
\hat{f}_{k,l}.
$$

Note that we later drop the ‘hat’ notation for convenience, and form the matrix of all haplotype recipients $*$ as $x_{ij}$. Each row of $x_{ij}$ corresponds to the vector $\hat{x}$ calculated above.

We calculate $\alpha_{kl}$ for $k = 1, \ldots, j$ in the following manner (Rabiner 1989):

1. $\alpha_{k1} = \Pr(h_s | Y_1 = h_k)f_k$
2. $\alpha_{kl} = \Pr(h_{*l} | Y_l = h_k) \left( \sum_{i=1}(\alpha_{il}(l-1)) f_k (1 - \exp(-\rho_l)) + \exp(-\rho_l) \alpha_{kl(l-1)} \right)$ for $l = 2, \ldots, L$.

We calculate $\beta_{kl}$ for $k = 1, \ldots, j$ in the following manner (Rabiner 1989):

1. $\beta_{kL} = 1.0$
2. $\beta_{kl} = \left[ \sum_{i=1}^j \beta_i(l+1) f_i \Pr(h_1, \ldots, h_L, Y_{l+1} = h_i) \right] (1 - \exp(-\rho_l)) + \exp(-\rho_l) \Pr(h_1, \ldots, h_L, Y_{l+1} = h_k) \beta_{k(l+1)}$ for $l = 1, \ldots, (L-1)$.  


Calculating expected lengths of copied chunks:

To calculate $\hat{l}_1, ..., \hat{l}_j$, the posterior expected length (in Morgans) of the total genome for which haploid $h_*$ copies from each of $h_1, ..., h_j$, respectively, we calculate the following (let $\Pr_h \equiv \Pr(h_*|Y_{l+1} = h_k)$):

$$\hat{l}_k = \frac{1}{\Pr(D)} \sum_{l=1}^{L-1} g_l \left[ \alpha_{kl} \beta_{k(l+1)} \left( \exp(-\rho_l) + (1.0 - \exp(-\rho_l)) f_k \right) \Pr_h 
+ (1/2) \left[ \alpha_{kl} \beta_{kl} + \alpha_{k(l+1)} \beta_{k(l+1)} - 2 \alpha_{kl} \beta_{k(l+1)} \left( \exp(-\rho_l) + (1.0 - \exp(-\rho_l)) f_k \right) \Pr_h \right] \right].$$

(4)

Note that this involves the approximation that at most only one change point occurs between neighbouring sampled sites. To get the expected length of each chunk copied from donor $h_k$, we divide equation (4) by equation (3) (i.e. $\hat{l}_k/\hat{x}_k$).

Calculating expected number of mutations:

To calculate $\hat{m}_1, ..., \hat{m}_j$, the posterior expected number of SNPs for which haploid $h_*$ copies with mutation (i.e. emission) from each of $h_1, ..., h_j$, respectively, we calculate the following (let $I_{[h_* \neq h_k]}$ be an indicator that the allelic type carried by $h_*$ does not match the allelic type carried by $h_k$ at SNP $l$):

$$\hat{m}_k = \frac{1}{\Pr(D)} \sum_{l=1}^{L-1} \alpha_{kl} \beta_{kl} I_{[h_* \neq h_k]}.$$

(5)

Using the E-M algorithm to estimate the scaling parameter $N_e$:

One can take a fixed $N_e$ for calculating $\bar{\rho}$, or use the Expectation-Maximisation (E-M) algorithm to find a local maximum of $N_e$ in the following manner. Start with an initial value of $N_e$ (we take $N_e \approx 400,000/j$), and at each iteration of the E-M replace $N_e$ with:

$$N_e^{*} = \frac{\sum_{l=1}^{L-1} (\sum_{k=1}^{j} f_{k,l}[|\rho_l|/|1.0 - \exp(-\rho_l)|])}{\sum_{l=1}^{L-1} g_l},$$

(6)

where $\rho_l$ and each $f_{k,l}$ are calculated using the previous value of $N_e$. In analyses presented here, we used 10 iterations of E-M to get our final estimate of $N_e$.

Using the E-M algorithm to estimate the mutation parameter $\theta$

One can take a fixed $\theta$ for calculating $\bar{\rho}$, or use the E-M to find a local maximum of $\theta$ in the following manner. Start with an initial value of $\theta$ (we start with Watterson’s estimate of $\theta$), and at each iteration of the E-M replace $\theta$ with:
\[ \theta^* = \frac{\sum_{i=1}^{L} \left( \sum_{i=1}^{j} \alpha_{il} \beta_{il} I_{[h_{il} \neq h_{il}]} / \Pr(D) \right)}{L}. \] (7)

Here \( I_{[h_{si} \neq h_{si}]} \) is an indicator that the allele \( h_{si} \) carried by the recipient is not equal to allele \( h_{si} \) carried by donor haploid \( i \) at SNP \( l \), and each \( \alpha_{il}, \beta_{il} \) and \( \Pr(D) \) are calculated using the previous value of \( \theta \).

**References**


