Supporting Information

Genome-wide fine-scale recombination rate variation in
Drosophila melanogaster

Andrew H. Chan\(^1\)*, Paul A. Jenkins\(^1\)*, Yun S. Song\(^1,2,\)**

\(^1\) Computer Science Division, University of California, Berkeley, CA, USA
\(^2\) Department of Statistics, University of California, Berkeley, CA, USA

* These authors contributed equally to this work
** Corresponding author e-mail: yss@cs.berkeley.edu

Text S1

Two-locus recursion relation

Suppose we sample \(n\) haplotypes, observing their alleles at each of two loci and obtaining configuration \(n = (a, b, c)\). Here \(c = (c_{ij})\) is a matrix of the counts of haplotypes for which both alleles were observed; \(c_{ij}\) is the number of haplotypes with allele \(i\) at the first locus and allele \(j\) at the second locus. We also allow for the possibility that a haplotype had data missing at one locus: \(a = (a_i)_{i=1,..,K}\) is the vector of counts of haplotypes with allele \(i\) observed at the first locus and missing data at the second locus, and \(b = (b_j)_{j=1,..,L}\) is the vector of counts of haplotypes with allele \(j\) observed at the second locus and missing data at the first locus. Further, let:

\[
\begin{align*}
    a & = \sum_{i=1}^{K} a_i, \quad c_i = \sum_{j=1}^{L} c_{ij}, \quad c = \sum_{i=1}^{K} \sum_{j=1}^{L} c_{ij}, \\
    b & = \sum_{j=1}^{L} b_j, \quad c_{ij} = \sum_{i=1}^{K} c_{ij}, \quad n = a + b + c.
\end{align*}
\]

The probability that, when we sample \(n\) haplotypes in some fixed order, we obtain a set consistent with configuration \(n\), is denoted by \(q(n; \theta_A, \theta_B, \rho)\). This probability is a function of \(\theta_A, \theta_B,\) and \(\rho\): the mutation rates at the two loci, and the recombination rate between them. The respective mutation transition matrices at the two loci, which we denote \(P^A\) and \(P^B\), are fixed. A system of equations for \(q(n; \theta_A, \theta_B, \rho)\) is given in [1]. We denote by \(q(n, s_1, s_2; \theta_A, \theta_B, \rho)\) the joint probability of obtaining \(n\) with the events that there were precisely \(s_1\) mutations in the history of the sample at the first locus and \(s_2\) mutations in the history of the sample at the second locus. The corresponding system of equations for \(q(n, s_1, s_2; \theta_A, \theta_B, \rho)\) is:

\[
\begin{align*}
    & [n(n-1) + \theta_A(a+c) + \theta_B(b+c) + \rho c]q((a, b, c), s_1, s_2; \theta_A, \theta_B, \rho) = \\
    & \sum_{i=1}^{K} a_i(a_i - 1 + 2c_i)q((a - e_i, b, c), s_1, s_2; \theta_A, \theta_B, \rho) + \sum_{j=1}^{L} b_j(b_j - 1 + 2c_j)q((a, b - e_j, c), s_1, s_2; \theta_A, \theta_B, \rho) \\
    & \quad + \sum_{i=1}^{K} \sum_{j=1}^{L} [c_{ij}(c_{ij} - 1)q((a, b, c - e_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) + 2a_ib_jq((a - e_i, b - e_j, c + e_{ij}), s_1, s_2; \theta_A, \theta_B, \rho)] \\
    & \quad + \theta_A \sum_{i=1}^{K} \sum_{j=1}^{L} \sum_{k=1}^{K} P^A_{ik}q((a, b, c - e_{ij} + e_{kj}), s_1, s_2; \theta_A, \theta_B, \rho) \\
    & \quad + \theta_B \sum_{j=1}^{L} \sum_{i=1}^{K} \sum_{k=1}^{K} P^B_{jk}q((a, b, c - e_{ij} + e_{kj}), s_1, s_2; \theta_A, \theta_B, \rho)
\end{align*}
\]
Suppose we have one genomic sequence of \( \lambda \) and \( \lambda' \). To compute the distribution on the ancestral allele, say \( \lambda \) and \( \lambda' \) at loci \( A \) and \( B \), respectively. Then we replace the relevant instances of (1) with the following:

\[
q((0, b, e_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} 
q((0, b + e_j, 0), s_1, s_2; \theta_A, \theta_B, \rho) & \text{if } i = \lambda \text{ and } s_1 = 0, \\
0 & \text{otherwise,}
\end{cases}
\]

\[
q((a, 0, e_{ij}), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} 
q((a + e_i, 0, 0), s_1, s_2; \theta_A, \theta_B, \rho) & \text{if } j = \lambda' \text{ and } s_2 = 0, \\
0 & \text{otherwise,}
\end{cases}
\]

\[
q((e_i, 0, 0), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} 
1 & \text{if } i = \lambda \text{ and } s_1 = s_2 = 0, \\
0 & \text{otherwise.}
\end{cases}
\]

\[
q((0, e_j, 0), s_1, s_2; \theta_A, \theta_B, \rho) = \begin{cases} 
1 & \text{if } j = \lambda' \text{ and } s_1 = s_2 = 0, \\
0 & \text{otherwise.}
\end{cases}
\]

(2)

### Padé summation

Modifications to the approach described in [2] are made, following from the boundary conditions given above. These can be converted into modifications of entries of the dynamic programming tables given in [2]. For example, using (2) we have that

\[
q((a, 0, e_{i\lambda}), 1, 0; \theta_A, \theta_B, \rho) = q((a + e_i, 0, 0), 1, 0; \theta_A, \theta_B, \rho)
\]

\[
= q(a + e_i, 1; \theta_A) + \frac{0}{\rho} + \frac{0}{\rho^2} + \ldots,
\]

where \( q(a + e_i, 1; \theta_A) \) is the one-locus solution given by equation (3) in the main text. Notice that this expansion is in fact independent of \( \rho \), from which it follows (by comparison with eq. (3.7) of [2]) that a number of entries in the dynamic programming tables are modified. For example, the second row in the dynamic programming table for the configuration \((a, 0, e_{i\lambda})\) is set to zero. Other boundary conditions may be interpreted in a similar fashion.

### Ancestral allele estimation

Suppose we have one genomic sequence of \( D. \) simulans and \( n \) sequences of \( D. \) melanogaster. Let \( S \) represent the sequence of \( D. \) simulans and \( M^{(k)} \) represent the sequence of the \( k \)th \( D. \) melanogaster, where \( S_l \) denotes the \( l \)th base of the sequence, and \( S_l' \) represents the sequence with the exclusion of the \( l \)th base. Given \((S, M^{(k)})\), let \( T_{l}^{(k)} \) be the time to the most recent common ancestor (TM RCA) at locus \( l \); \( f_{l}^{(k)}(t \mid M_{l}, S_{l}') \) be the density of the TM RCA conditioned on both their sequences but excluding the \( l \)th locus; and \( A_{l}^{(k)} \) be the ancestral allele at the \( l \)th locus, i.e., the allele of the most recent common ancestor (MRCA).

To compute the distribution on the ancestral allele at the \( l \)th locus conditioned on \( M^{(k)} \) and \( S \), we use
Bayes’ theorem to obtain
\[
P(A_{l}^{(k)} = i \mid M^{(k)}, S) = \frac{\int_{0}^{\infty} p(A_{l}^{(k)} = i, M^{(k)}, S, T_{l}^{(k)} = t)dt}{\mathbb{P}(M^{(k)}, S)} = \frac{\int_{0}^{\infty} \mathbb{P}(M_{l}^{(k)} | A_{l}^{(k)} = i, T_{l}^{(k)} = t) \mathbb{P}(S_{l} | A_{l}^{(k)} = i, T_{l}^{(k)} = t) \mathbb{P}(A_{l}^{(k)} = i) \mathbb{P}(T_{l}^{(k)} = t)dt}{\sum_{j} \int_{0}^{\infty} \mathbb{P}(M_{l}^{(k)} | A_{l}^{(k)} = j, T_{l}^{(k)} = t) \mathbb{P}(S_{l} | A_{l}^{(k)} = j, T_{l}^{(k)} = t) \mathbb{P}(A_{l}^{(k)} = j) \mathbb{P}(T_{l}^{(k)} = t)dt}.
\] (3)

In equation (3), the prior on the ancestral allele at locus \(l\), \(P(A_{l}^{(k)} = i)\), is given by the stationary distribution of the allele frequencies from the mutation matrix \(P\). (In the above, \(p\) denotes a joint probability of discrete events together with the density for \(T_{l}^{(k)}\).) The density on the TMRCA, \(f_{l}^{(k)}(t \mid M_{l}^{(k)}, S_{l})\), is estimated using Li & Durbin’s \texttt{psmc} [3]. In practice, we use \texttt{psmc} to compute \(f_{l}^{(k)}(t \mid M_{l}^{(k)}, S_{l})\) and assume \(f_{l}^{(k)}(t \mid M_{l}^{(k)}, S_{l}) \approx f_{l}^{(k)}(t \mid M_{l}^{(k)}, S_{l})\).

The remaining two probabilities, \(P(M_{l}^{(k)} | A_{l}^{(k)} = i, T_{l}^{(k)} = t)\) and \(P(S_{l} | A_{l}^{(k)} = i, T_{l}^{(k)} = t)\), are computed as follows. For the computation of \(P(M_{l}^{(k)} | A_{l}^{(k)} = i, T_{l}^{(k)} = t)\), let \(P = (P_{ij})\) denote the mutation matrix, and let \(r_{l}^{(k)}\) specify the number of mutations that have occurred at the \(l\)th locus of the \(k\)th \(D.\ melanogaster\) sequence during time \(T_{l}^{(k)}\). Then we have

\[
P(M_{l}^{(k)} = j \mid A_{l}^{(k)} = i, T_{l}^{(k)} = t) = \sum_{s=0}^{\infty} \mathbb{P}(r_{l}^{(k)} = s \mid T_{l}^{(k)} = t)(P^{s})_{ij}
\]

\[
= \sum_{s=0}^{\infty} \left( \frac{\theta t}{2} \right)^{s} \frac{e^{-\theta t/2}}{s!} (P^{s})_{ij}
\]

\[
= \sum_{s=0}^{\infty} \left[ \left( \frac{\theta t}{2} P \right)^{s} \right]_{ij} e^{-\theta t/2}
\]

\[
= \left[ e^{-\frac{1}{2} (P^{s} - I)} \right]_{ij},
\]

where \(I\) is the identity matrix with the same dimensions as \(P\). The computation for \(P(S_{l} | A_{l}^{(k)} = j, T_{l}^{(k)} = t)\) is analogous.

After computing \(P(A_{l}^{(k)} = i \mid M^{(k)}, S)\) for every \(k\) and given \(l\), we heuristically aggregate these pairwise probabilities to estimate \(P(A_{l}^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S)\) as follows. Let \(\hat{t}_{l}^{(k)}\) be the posterior mean of \(f_{l}^{(k)}(t \mid M^{(k)}, S)\), i.e.:

\[
\hat{t}_{l}^{(k)} = \int_{0}^{\infty} tf_{l}^{(k)}(t \mid M^{(k)}, S)dt,
\]

and define \(\tau_{l} = \max_{k} \hat{t}_{l}^{(k)}\). We approximate \(P(A_{l}^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S)\) as

\[
P(A_{l}^{(k)} = i \mid M^{(1)}, \ldots, M^{(n)}, S) \approx \frac{\sum_{k=1}^{n} \mathbb{P}(A_{l}^{(k)} = i \mid M^{(k)}, S) f_{l}^{(k)}(\tau_{l} \mid M_{l}^{(k)}, S_{l})}{\sum_{j} \sum_{k=1}^{n} \mathbb{P}(A_{l}^{(k)} = j \mid M^{(k)}, S) f_{l}^{(k)}(\tau_{l} \mid M_{l}^{(k)}, S_{l})},
\]

which is a weighted average of \(P(A_{l}^{(k)} = i \mid M^{(k)}, S)\) over \(k\), weighted by the density of the TMRCA evaluated at \(\tau_{l}\) for each \(k\). This averaging ought to mitigate effects such as genotyping errors and incomplete lineage sorting in individual \(D.\ melanogaster\) genomes.
References

