recombuddy Mathematical Details

Robert Verity

April 2025

1 Introduction

This document describes the statistical model used by *recombuddy* to simulate recombination blocks along the genome.

2 Basic notation

Samples are simulated individually and independently. For a given sample, we define the multiplicity of infection (MOI), denoted M, which represents the number of *distinct haploid genotypes* comprising the sample. Simulation occurs at the level of these haploid genotypes.

Assume there are C chromosomes indexed by $c \in 1:C$, where $C = 14$ for the *P. falciparum* core genome. Let G_c be the length of chromosome c.

For the m^{th} haploid genotype, simulation output takes the form of a series of contiguous blocks that span the entire genome. Let $B_{m,c}$ denote the number of distinct blocks in chromosome c of the m^{th} haploid genotype. For notational simplicity, we will often condition on a particular m and c, dropping these subscripts to give B blocks indexed via $b \in 1:B$.

Each block has a start position, $x^{(b)}_{\text{start}} \in 1:G_c$, and an end position $x^{(b)}_{\text{end}} \in x^{(b)}_{\text{start}} : G_c$. Each block also has an ancestral index $a^{(b)} \in 1:N$ specifying which of N ancestral genotypes this block inherits from.

3 The ancestral population

Let $\mathbf{z} = [z_1, \ldots, z_N]$ denote the sampling probability vector corresponding to the ancestral population, where $\sum_{i=1}^{N} z_i = 1$. This distribution can either be specified manually by the user or generated using the `rdirichlet_single()` function, which samples from a symmetric Dirichlet distribution with concentration parameter α:

$$\mathbf{z} \sim \text{Dirichlet}(\alpha, \alpha, \ldots, \alpha) \quad \text{(length } N) \tag{1}$$

When taking the Dirichlet distribution approach, the expected correlation between two haploid genotypes sampled from distinct monoclonal individuals is $f = 1/(1 + \alpha)$. We can use this to set a desired level of between-sample relatedness via $\alpha = (1 - f)/f$.

4 The number of serial meioses

We manually define the number of serial meioses for each haploid genotype. Let $\mathbf{k} = [k_1, ..., k_M]$ be the vector of serial meioses, where $k_m \in \mathbb{Z}_{\geq 0}$ for $m \in 1 : M$.

We will treat the non-recombinant and recombinant genotypes differently, and hence introduce some additional notation in the form of index vectors:

$$J^{(0)} = [J_1^{(0)}, J_2^{(0)}, \ldots] = \{m \in 1 : M \mid k_m = 0\} \tag{2}$$

$$J^{(+)} = [J_1^{(+)}, J_2^{(+)}, \ldots] = \{m \in 1 : M \mid k_m > 0\} \tag{3}$$

These vectors contain the indices of genotypes with zero and nonzero meioses, respectively, and are assumed to be ordered in increasing value of m. It follows that the number of non-recombinant genotypes is given by $|J^{(0)}|$ and the number of recombinant genotypes by $|J^{(+)}|$, where $|J^{(0)}| + |J^{(+)}| = M$.

To give a concrete example, if $\mathbf{k} = [0, 1, 0, 1, 2]$ then we would have $J^{(0)} = [1, 3]$ and $J^{(+)} = [2, 4, 5]$.

5 Non-recombinant genotypes

Each non-recombinant genotype is assigned a single parent in the ancestral population. Let $Q = [q_1, ..., q_{|J^{(0)}|}]$ be a vector that specifies these parental assignments, where $q_i \in 1 : N$. To ensure that all genotypes are truly distinct, we sample these parents *without replacement*. Sampling with replacement could result in the same ancestor being selected multiple times, leading to the identical genotype being introduced multiple times. This would violate our definition of M as the number of *distinct* haploid genotypes comprising the sample. By sampling without replacement, we guarantee that each genotype originates from a unique ancestor.

Sampling takes into account the probability vector $\mathbf{z}$. Mathematically, we can write:

$$Q \sim \text{SampleWithoutReplacement}(\{1, \ldots, N\}, |J^{(0)}|, \mathbf{z}) \tag{4}$$

As there is zero recombination, the sampled ancestor applies throughout the genome. Focusing on the i^{th} non-recombinant genotype (equivalent to the element $J_i^{(0)}$ in the original indexing) and the c^{th} chromosome, there is a single recombination block with start position $x_{\text{start}}^{(1)} = 1$, end position $x_{\text{end}}^{(1)} = G_c$, and ancestral index $a^{(1)} = q_i$. This applies for all $c \in 1 : C$ and over all $i \in 1 : |J^{(0)}|$.

6 Recombinant genotypes

Creating recombinant genotypes is a more complex process that involves sampling multiple ancestors per genotype, simulating recombination break-points, and assigning ancestors to the resulting recombination blocks.

6.1 Sampling ancestors

Assume we are focusing on the i^{th} recombinant genotype, equivalent to element $J_i^{(+)}$ in the original indexing. The number of serial meioses for this genotype is $k_{J_i^{(+)}}$, which for convenience we will redefine as $s := k_{J_i^{(+)}}$.

With s serial meioses we need 2^s ancestors. These ancestors are sampled *with replacement* from the ancestral population. However, to avoid the possibility of drawing the same ancestor 2^s times, which would result in a clonal copy of a single ancestral genotype, we impose a constraint: at least two distinct ancestors must be sampled. To enforce this, we first sample $2^s - 1$ ancestors with replacement, and then draw the final ancestor from the set of remaining unsampled genotypes. This ensures that each recombinant genotype is composed of at least two distinct ancestral contributions, preserving the intended number of distinct haploid genotypes (M) in the sample.

Let $W^{(i)} = [W_1^{(i)}, ..., W_{2^s}^{(i)}]$ be a vector of the 2^s ancestors of the i^{th} recombinant genotype. We draw the first $2^s - 1$ with replacement:

$$[W_1^{(i)}, ..., W_{2^s-1}^{(i)}] \sim \text{SampleWithReplacement}(\{1, \ldots, N\},\ 2^s - 1,\ \mathbf{z}) \tag{5}$$

Then define $\Lambda^{(i)}$ as the set of unsampled ancestors:

$$\Lambda^{(i)} = \{1, \ldots, N\} \setminus [W_1^{(i)}, ..., W_{2^s-1}^{(i)}] \tag{6}$$

Then we define a new probability vector $\tilde{\mathbf{z}}$, which is restricted to the elements of $\Lambda^{(i)}$:

$$\tilde{\mathbf{z}} = \left[\frac{z_j}{\sum\limits_{w \in \Lambda^{(i)}} z_w} \right]_{j \in \Lambda^{(i)}} \tag{7}$$

The final ancestor, $W^{(i)}_{2^s}$, is drawn from $\Lambda^{(i)}$:

$$W^{(i)}_{2^s} \sim \mathrm{SampleWithoutReplacement}(\Lambda^{(i)}, 1, \mathbf{z}) \tag{8}$$

The full vector of ancestors, $W^{(i)}$, has length 2^s and is guaranteed to contain at least two distinct values.

6.2 Sampling recombination breakpoints

Assume we are focusing on the i^{th} recombinant genotype, which results from s serial meioses, and chromosome c. The number of distinct recombination blocks is denoted $B := B_{J_i^{(+)},c}$, with start and end positions $x^{(b)}_{\text{start}}$ and $x^{(b)}_{\text{end}}$ for $b \in 1 : B$.

We model recombination as a Poisson process with constant hazard ρ along the chromosome. The probability of a recombination break-point between adjacent sites is:

$$p = 1 - e^{-s\rho} \tag{9}$$

This uses the superposition property of Poisson processes to model all s recombination events simultaneously, as a Poisson process with rate ρ applied s times is equivalent to a single Poisson process with rate ρs.

The first block starts from position $x^{(1)}_{\text{start}} = 1$. We draw the length from a geometric distribution (supported on $\mathbb{N}_0$), and truncate it at the chromosome length:

$$D \sim \mathrm{Geom}(p), \quad x^{(1)}_{\text{end}} = \min(1 + D,\ G_c) \tag{10}$$

Subsequent recombination blocks are generated iteratively. While $x^{(b)}_{\text{end}} < G_c$:

- Increment the block index $b \leftarrow b + 1$
- Set the next block's start position: $x^{(b)}_{\text{start}} = x^{(b-1)}_{\text{end}} + 1$
- Draw a new block length: $D \sim \mathrm{Geom}(p), \quad x^{(b)}_{\text{end}} = \min(x^{(b)}_{\text{start}} + D,\ G_c)$

This procedure continues until the entire chromosome is covered, i.e. until $x^{(b)}_{\text{end}} = G_c$. The number of blocks B is the final value of b reached.

6.3 Assigning ancestors to recombination blocks

Each recombination block is assigned an ancestor from $W^{(i)}$. For the first block ($b = 1$), the ancestor is chosen uniformly at random from the full set of ancestral lineages:

$$a^{(1)} \sim \mathrm{Uniform}(W^{(i)}) \tag{11}$$

For subsequent blocks ($b = 2, \ldots, B$), the ancestor for block b is also drawn uniformly from $W^{(i)}$, with the constraint that it differs from the ancestor of the previous block:

$$a^{(b)} \sim \text{Uniform}\left(W^{(i)} \setminus \{a^{(b-1)}\}\right) \tag{12}$$

This ensures that the same ancestor is never assigned to two contiguous blocks, enforcing at least one switch at recombination break-points. This process is repeated for all chromosomes.

7 Caveats

While care has been taken in the above scheme to ensure that the final number of distinct haploid genotypes is indeed equal to M, there are some edge cases that are worth mentioning.

For non-recombinant genotypes, there is no risk of sampling the same ancestor multiple times, as we are sampling without replacement.

For recombinant genotypes, while the trivial case of producing a clonal genotype by repeatedly sampling the same parent has been eliminated, there are still other ways that clonal transmission could occur. For example, if ρs is small then entire chromosomes could be copied without recombination, leading to the possibility that a single ancestor applies genome-wide. Similarly, even in the presence of a recombination break-point, there is a technical possibility of simulating the identical recombination break-point and the same pattern of ancestry multiple times. Although these are remote possibilities, it means we cannot strictly give a 100% guarantee of M distinct haploid genotypes.

A more important caveat comes down to the variability in the ancestral genotypes. We have focused on sampling the index of these genotypes between $1 : N$, but in reality these typically point back to a panel of real observed genotypes. Variability in this ancestral population is an important consideration, as deriving from distinct ancestors does not guarantee that there are any genetic differences over a given region. In other words, haploid genotypes could derive from different ancestors, but still appear identical.