Supplementary Methods

Correcting observer bias due to drop-out of reads

Let N reads from K clusters align to a reference sequence of length L. Let the proportion of reads whose 5' and 3' ends align, respectively, to coordinates a and b ($1 \leq a \leq b \leq L$) be η_{ab} (assuming these proportions are equal for all clusters). Let the mutation rate of base j ($1 \leq j \leq L$) in cluster k ($1 \leq k \leq K$) be μ_{jk}. Let the proportion of cluster k in the ensemble be π_k. To express these quantities as probabilities, let C_k be the event that a read comes from cluster k; let E_{ab} be the event that a read aligns with 5' and 3' coordinates a and b, respectively; let S_j be the event that a read contains position j (i.e. its alignment coordinates a and b satisfy $1 \leq a \leq j \leq b \leq L$); let M_j be the event that a read has a mutation at position j; and let G_g be the event that a read has no two mutations separated by fewer than g non-mutated bases.

Deriving mutation rates of reads with no two mutations too close

In terms of these events, the total mutation rates (μ_{jk}) are $P(M_j|S_jC_k)$, i.e. the probability that a read would have a mutation at position j given that it contained position j and came from cluster k; and the observable mutation rates (m_{jk}) are $P(M_j|S_jC_kG_g)$, i.e. the probability that a read would have a mutation at position j given that it contained position j, came from cluster k, and had no two mutations closer than g bases. Using these definitions and Bayes' theorem yields a probabilistic formula for m_{jk}:

$$m_{jk} = P(M_j|S_jC_kG_g) = P(M_j|S_jC_k)\frac{P(G_g|S_jM_jC_k)}{P(G_g|S_jC_k)} = \mu_{jk}\frac{P(G_g|S_jM_jC_k)}{P(G_g|S_jC_k)}$$

The term $P(G_g|S_jC_k)$ is the probability that a read would have no two mutations closer than g bases given that it contained position j and came from cluster k. It can be computed using $P(G_g|E_{ab}C_k)$ (abbreviated d_{abk}): the probability that

a read would contain no two mutations closer than g bases given that its 5' and 3' coordinates are a and b, repectively ($1 \leq a \leq b \leq L$), and that it came from cluster k. If position b were mutated (probability μ_{bk}), then the read would contain no two mutations closer than g bases if and only if none of the g bases preceding b (i.e. positions $b-g$ to $b-1$, inclusive) were mutated (probability $\prod_{j'=\mathsf{max}(b-g,a)}^{b-1}(1-\mu_{j'k})$, abbreviated $w_{\mathsf{max}(b-g,a),b-1,k}$) and two no mutations between positions a and $b-(g+1)$, inclusive, were too close (probability $d_{a,\mathsf{max}(b-(g+1),a),k}$)). If position b were not mutated (probability $1-\mu_{bk}$), then the read would contain no two mutations closer than g bases if and only if no mutations between positions a and $b-1$, inclusive, were too close (probability $d_{a,\mathsf{max}(b-1,a),k}$). These two possibilities generate a recurrence relation:

$$d_{abk} = \mu_{bk} w_{\mathsf{max}(b-g,a),b-1,k} d_{a,\mathsf{max}(b-(g+1),a),k} + (1-\mu_{bk}) d_{a,\mathsf{max}(b-1,a),k}$$

The base case is $d_{abk} = 1$ when $a = b$ because such a read would contain one position and thus be guaranteed to have no two mutations too close. Then, $P(G_g|S_jC_k)$ is the average of d_{abk} over every read that contains position j, weighted by the proportions η_{ab}:

$$P(G_g|S_jC_k) = \frac{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}d_{abk}}{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}}$$

The term $P(G_g|S_jM_jC_k)$ is the probability that a read would have no two mutations too close given that it contained a mutation at position j and came from cluster k. It can be computed using $P(G_g|M_jE_{ab}C_k)$ (abbreviated f_{abjk}): the probability that a read would contain no two mutations too close given that position j is mutated ($1 \leq a \leq j \leq b \leq L$), that its 5' and 3' coordinates are a and b (respectively), and that it came from cluster k. Because position j is mutated, having no two mutations too close requires that none of the g bases on both sides of position j be mutated. The probability that none of the preceding g positions ($j-g$ to $j-1$) is mutated is $w_{\mathsf{max}(j-g,a),j-1,k}$, while that of the following g positions ($j+1$ to $j+g$) is $w_{j+1,\mathsf{min}(j+g,b),k}$. Upstream of the g bases flanking position j (i.e. positions a to

$j-(g+1)$), the probability that no two mutations are too close is $d_{a,\mathsf{max}(j-(g+1),a),k}$; downstream (i.e. positions $j+(g+1)$ to b), the probability is $d_{\mathsf{min}(j+(g+1),b),b,k}$. Since mutations in these four sections are independent, the probability that the read contains no two mutations too close is the product:

$$f_{abjk} = d_{a,\mathsf{max}(j-(g+1),a),k} w_{\mathsf{max}(j-g,a),j-1,k} w_{j+1,\mathsf{min}(j+g,b),k} d_{\mathsf{min}(j+(g+1),b),b,k}$$

Then, $P(G_g|S_jM_jC_k)$ is the average of f_{abjk} over every read that contains position j, weighted by the proportions η_{ab}.

$$P(G_g|S_jM_jC_k) = \frac{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}f_{abjk}}{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}}$$

Combining the above results yields an explicit formula for m_{jk}:

$$m_{jk} = \mu_{jk}\frac{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}f_{abjk}}{\sum_{a=1}^{j}\sum_{b=j}^{L}\eta_{ab}d_{abk}}$$

Deriving end coordinate proportions of reads with no two mutations too close

The total proportions (η_{ab}) of reads aligned to 5' and 3' coordinates a and b, respectively, are $P(E_{ab})$; and the proportions of reads with no two mutations too close that align with coordinates a and b (e_{abk}) are $P(E_{ab}|G_gC_k)$. Note that, while reads are assumed to come from the same distribution of coordinates (η_{ab}) regardless of their cluster k, the observable distribution of coordinates (e_{abk}) varies by cluster because P(G_gC_k) depends on k. Using these definitions and Bayes' theorem yields a probabilistic formula for e_{abk}:

$$e_{abk} = P(E_{ab}|G_gC_k) = P(G_g|E_{ab}C_k)\frac{P(E_{ab}|C_k)}{P(G_g|C_k)} = d_{abk}\frac{\eta_{ab}}{P(G_g|C_k)}$$

The term $P(G_g|C_k)$ is the probability that a read would have no two mutations too close given that it came from cluster k. It can be computed as an average of $P(G_g|E_{ab}C_k)$ (i.e. d_{abk}) over all coordinates a and b (such that $1 \leq a \leq b \leq L$),

weighted by the proportion of each coordinate, $P(E_{ab})$ (i.e. η_{ab}):

$$P(G_g|C_k) = \frac{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}} = \sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}$$

This expression is already normalized because $\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab} = 1$, by definition.

Combining the above results yields an explicit formula for e_{abk}:

$$e_{abk} = \frac{\eta_{ab}d_{abk}}{\sum_{a'=1}^{L}\sum_{b'=a'}^{L}\eta_{a'b'}d_{a'b'k}}$$

Deriving cluster proportions of reads with no two mutations too close

The proportion of total reads in cluster k is $\pi_k = P(C_k)$. The proportion among only reads with no two mutations closer than g bases is

$$p_k = P(C_k|G_g) = P(G_g|C_k)\frac{P(C_k)}{P(G_g)} = \pi_k\frac{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{P(G_g)}$$

The term $P(G_g)$ is the probability that a read from any cluster would have no two mutations closer than g bases and can be solved for by leveraging that the cluster proportions (p_k) must sum to 1:

$$1 = \sum_{k=1}^{K}p_k = \sum_{k=1}^{K}\pi_k\frac{\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{P(G_g)} = \frac{1}{P(G_g)}\sum_{k=1}^{K}\pi_k\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}$$

$$P(G_g) = \sum_{k=1}^{K}\pi_k\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}$$

The result is an explicit formula for p_k:

$$p_k = \frac{\pi_k\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk}}{\sum_{k'=1}^{K}\pi_{k'}\sum_{a=1}^{L}\sum_{b=a}^{L}\eta_{ab}d_{abk'}}$$

Solving total mutation rates and cluster and coordinate proportions

The observed mutation rates (m_{jk}), end coordinate proportions (e_{abk}), and cluster proportions (p_k) can be calculated as weighted averages over the N reads with no two mutations too close:

$$m_{jk} = \frac{\sum_{i=1}^{N} z_{ik} x_{ij}}{\sum_{i=1}^{N} z_{ik} s_{ij}}$$

$$e_{abk} = \frac{\sum_{i=1}^{N} z_{ik} y_{abi}}{\sum_{i=1}^{N} z_{ik}}$$

$$p_k = \frac{\sum_{i=1}^{N} z_{ik}}{N}$$

where s_{ij} is 1 if read i contains position j, otherwise 0; x_{ij} is 1 if read i has a mutation at position j, otherwise 0; y_{abi} is 1 if read i aligns to coordinates a and b, otherwise 0; and z_{ik} is the probability that read i came from cluster k.

The original parameters μ_{jk}, η_{abk}, and π_k can be solved by setting the two formulae each for m_{jk}, e_{abk}, and p_k equal to each other, creating a system of equations:

$$\mu_{jk} \frac{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} f_{abjk}}{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} d_{abk}} = m_{jk} = \frac{\sum_{i=1}^{N} z_{ik} x_{ij}}{\sum_{i=1}^{N} z_{ik} s_{ij}}$$

$$\eta_{ab} \frac{d_{abk}}{\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k}} = e_{ab} = \frac{\sum_{i=1}^{N} z_{ik} y_{abi}}{\sum_{i=1}^{N} z_{ik}}$$

$$\pi_k \frac{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}}{\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'}} = p_k = \frac{\sum_{i=1}^{N} z_{ik}}{N}$$

Solving this entire system at once has proven computationally impractical for all but extremely short sequences. A more feasible approach is to first solve for μ_{jk} given an initial guess for η_{ab}, next solve for η_{ab} given the updated μ_{jk}, then solve for π_k given the updated μ_{jk} and η_{ab}, and iterate until all three sets of parameters converge.

Even assuming every η_{ab} is a constant, these equations are still too complex to solve for μ_{jk} analytically because d_{abk} and f_{abjk} also depend on μ_{jk} (as well as on other μ variables). Thus, every μ_{jk} is solved for numerically by rearranging each

equation to

$$\mu_{jk} \frac{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} f_{abjk}}{\sum_{a=1}^{j} \sum_{b=j}^{L} \eta_{ab} d_{abk}} - m_{jk} = 0$$

and applying the Netwon-Krylov method **?** implemented in SciPy **?**.

Once every μ_{jk} has been solved for, every η_{ab} can be updated. Because d_{abk} does not depend on η_{ab} (except indirectly through the μ_{jk} parameters, which are now assumed to be constants), each equation can be rearranged to

$$\eta_{ab} = \frac{e_{ab}}{d_{abk}} \sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k}$$

Leveraging that $\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} = 1$, by definition, leads to

$$\sum_{a=1}^{L} \sum_{b=a}^{L} \frac{e_{ab}}{d_{abk}} \sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k} = 1$$

$$\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \eta_{a'b'} d_{a'b'k} = \frac{1}{\sum_{a=1}^{L} \sum_{b=a}^{L} \frac{e_{ab}}{d_{abk}}}$$

and finally a closed-form expression for each η_{ab} given μ_{jk} (and hence d_{abk}) and e_{abk}:

$$\eta_{ab} = \frac{\frac{e_{ab}}{d_{abk}}}{\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \frac{e_{a'b'}}{d_{a'b'k}}}$$

This equation should theoretically yield the same value of η_{ab} for every k. In practice, the values will differ due to inexactness in floating-point arithmetic. Thus, the consensus value of η_{ab} is taken to be the average η_{ab} over every k, weighted by π_k:

$$\eta_{ab} = \sum_{k=1}^{K} \pi_k \frac{\frac{e_{ab}}{d_{abk}}}{\sum_{a'=1}^{L} \sum_{b'=a'}^{L} \frac{e_{a'b'}}{d_{a'b'k}}}$$

With updated values of μ_{jk} and η_{ab}, π_k can also be solved. The above equations can be rearranged to

$$\pi_k = p_k \frac{\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'}}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}}$$

Given that $\sum_{k=1}^{K} \pi_k = 1$, by definition:

$$\sum_{k=1}^{K} p_k \frac{\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'}}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}} = 1$$

$$\sum_{k'=1}^{K} \pi_{k'} \sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'} = \frac{1}{\sum_{k=1}^{K} \frac{p_k}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}}}$$

which leads to a closed-form expression for each π_k given μ_{jk} (and hence d_{abk}), η_{ab}, and p_k:

$$\pi_k = \frac{\frac{p_k}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk}}}{\sum_{k'=1}^{K} \frac{p_{k'}}{\sum_{a=1}^{L} \sum_{b=a}^{L} \eta_{ab} d_{abk'}}}$$

Clustering reads with the expectation-maximization algorithm

Let N reads from K clusters align to a reference sequence of length L. Let the proportion of reads whose 5' and 3' ends align, respectively, to coordinates a and b ($1 \leq a \leq b \leq L$) be η_{ab} (assuming these proportions are equal for all clusters). Let the mutation rate of base j ($1 \leq j \leq L$) in cluster k ($1 \leq k \leq K$) be μ_{jk}. Let the proportion of cluster k in the ensemble be π_k.

Maximization step

The maximization step updates the parameters (μ_{jk}, η_{ab}, and π_k) using the current cluster memberships (z_{ik}). The observed estimates of the parameters m_{jk}, e_{ab}, and p_k are first computed; then, the underlying parameters μ_{jk}, η_{ab}, and π_k are solved for as described in 0.1.4.

Expectation step

The expectation step updates the cluster memberships (z_{ik}) and the likelihood function (L) using the current parameters (μ_{jk}, η_{ab}, and π_k). Each cluster membership is defined as the probability that read i came from cluster k given its

5'/3' end coordinates (E_{ab}) and mutations (M) and given that no two mutations are too close (G_g): $z_{ik} = P(C_k|E_{ab}MG_g)$. The likelihood of the model (L) is the product of the marginal probability (L_i) of observing each read i from any cluster: $L_i = P(E_{ab}M|G_g)$. Both L_i and z_{ik} can be expressed in terms of the joint probability ($L_{ik} = P(E_{ab}MC_k|G_g)$) of observing each read i from each cluster k:

$$L_i = P(E_{ab}M|G_g) = \sum_{k=1}^{K} P(E_{ab}MC_k|G_g) = \sum_{k=1}^{K} L_{ik}$$

$$z_{ik} = P(C_k|E_{ab}MG_g) = \frac{P(E_{ab}MC_kG_g)}{P(E_{ab}MG_g)} = \frac{P(E_{ab}MC_k|G_g)}{P(E_{ab}M|G_g)} = \frac{L_{ik}}{L_i}$$

To derive a formula for L_{ik}, it can be factored into three parts using the chain rule for probability:

$$L_{ik} = P(E_{ab}MC_k|G_g) = \frac{P(E_{ab}MC_kG_g)}{P(G_g)} = P(M|E_{ab}C_kG_g)P(E_{ab}|C_kG_g)P(C_k|G_g)$$

The first part – the probability that a read would have the specific mutations x_{ij} given that its 5'/3' end coordinates are a and b (respectively), it comes from cluster k, and no two mutations are too close – is the product over every position j from a to b of the probability of a mutation (μ_{jk}) if read i is mutated at position j ($x_{ij} = 1$), otherwise ($x_{ij} = 0$) the probability of no mutation ($1 - \mu_{jk}$), normalized by the probability that no two mutations would be too close (d_{abk}):

$$P(M|E_{ab}C_kG_g) = \frac{1}{d_{abk}} \prod_{j=a}^{b} \mu_{jk}^{x_{ij}} (1 - \mu_{jk})^{(1-x_{ij})}$$

The second part, $P(E_{ab}|C_kG_g) = e_{abk}$, can be calculated from the parameters μ_{jk}, η_{ab}, and π_k, as explained in 0.1.2. Likewise, the third part, $P(C_k|G_g) = p_k$, can also be calculated from the parameters, as explained in 0.1.3. Combining all parts yields a formula for L_{ik} in terms of the parameters μ_{jk}, η_{ab}, and π_k and of their derived values d_{abk}, e_{abk}, and p_k:

$$L_{ik} = p_k \frac{e_{abk}}{d_{abk}} \prod_{j=a}^{b} \mu_{jk}^{x_{ij}} (1 - \mu_{jk})^{(1-x_{ij})}$$

The formula for the total likelihood of the model and its parameters follows:

$$L(\mu, \eta, \pi) = \prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \sum_{k=1}^{K} p_k \frac{e_{abk}}{d_{abk}} \prod_{j=a}^{b} \mu_{jk}^{x_{ij}} (1 - \mu_{jk})^{(1-x_{ij})}$$