# Mathematical Models for Single-Molecule Sequencing, Error Characterization, Haplotype Classification, and Haplotagging

Pranjal Srivastava

November 2025

## Contents

# 1 Error Quantification

In this section we formalize the objects used to quantify sequencing and basecalling error. We first introduce the mathematical preliminaries, then define refined read- and sequence-level quality variables, and finally give summary metrics that we use to evaluate a basecaller.

## 1.1 Preliminaries (Mathematical)

Let $\Sigma$ denote the nucleotide alphabet (e.g. $\Sigma = \{A, C, G, T\}$). Let

$$\mathcal{S} \subset \Sigma^{<\infty} \qquad \text{and} \qquad \mathcal{R} \subset \Sigma^{<\infty}$$

be the sets of possible input sequences and observed reads, respectively. For any sequence $x \in \mathcal{S} \cup \mathcal{R}$ we write $L(x)$ for its length in bases and $x[\ell] \in \Sigma$ for the base at position $\ell \in \{1, \dots, L(x)\}$.

Each read $r \in \mathcal{R}$ comes with basecaller-provided per–base Phred quality scores

$$Q_{r,\ell} \in \mathbb{R}, \qquad \ell = 1, \dots, L(r),$$

which are related to the basewise error probabilities by $p_{r,\ell} = 10^{-Q_{r,\ell}/10}$.

We will frequently consider *pairs* $(s, r) \in \mathcal{S} \times \mathcal{R}$ consisting of a "true" input sequence $s$ and an observed read $r$ that has been aligned to $s$. On such pairs we define the following joint variables.

- The (integer) Levenshtein edit distance

$$d(s, r) \in \mathbb{N}_0,$$

i.e. the minimum number of substitutions, insertions and deletions required to transform $s$ into $r$.

- An optimal alignment
$$A(s,r),$$
for example a global alignment in the sense of Needleman–Wunsch or an affine-gap variant, used to count how many of the edits in $d(s,r)$ are mismatches versus insertions or deletions.

We consider a finite collection of experiments

$$\mathcal{E} = \{e_1, \ldots, e_N\}.$$

Each experiment $e_i \in \mathcal{E}$ consists of:

- A finite set of input ("ground truth") sequences
$$\mathcal{S}_i = \{s_{ij}\}_{j=1}^{m_i} \subset \mathcal{S},$$
with associated mixture proportions (or purities)
$$\pi_{ij} > 0, \qquad \sum_{j=1}^{m_i} \pi_{ij} = 1.$$

- A finite multiset of reads
$$\mathcal{R}_i = \{r_{ik}\}_{k=1}^{\ell_i} \subset \mathcal{R}.$$
Each $r_{ik}$ carries basewise qualities $Q_{ik\ell} := Q_{r_{ik},\ell}$ for $\ell = 1, \ldots, L(r_{ik})$.

Whenever a read $r_{ik}$ is aligned to an input sequence $s_{ij}$ in experiment $e_i$, the pair $(s_{ij}, r_{ik})$ inherits the joint variables $d(s_{ij}, r_{ik})$ and $A(s_{ij}, r_{ik})$ defined above.

## 1.2 Refined Quality Variables

We now turn the edit distance into empirical error probabilities and corresponding Phred-like quality scores for both reads and sequences.

### Weighted edit distance

Given an alignment $A(s,r)$ of $s$ to $r$, let

$$M(s,r) \quad \text{and} \quad I(s,r)$$

denote, respectively, the number of mismatch events and the total number of insertion and deletion (indel) events in the alignment.

Often we wish to penalize indels more strongly than mismatches. Let $k \geq 0$ be the desired ratio of the indel weight to the mismatch weight; that is, we want one indel to "count" as $k$ mismatches. Set

$$\alpha = \frac{1}{1+k}, \qquad 1-\alpha = \frac{k}{1+k},$$

so that indeed $(1-\alpha)/\alpha = k$. We define the *weighted edit count*

$$E_\alpha(s,r) := \alpha\, M(s,r) + (1-\alpha)\, I(s,r)$$

and the corresponding *normalized weighted edit distance*

$$\tilde{d}_\alpha(s,r) := \frac{E_\alpha(s,r)}{L(s)} \in [0, \infty).$$

**Example.** If we want each indel to weigh three times as much as a mismatch, we take $k = 3$, hence

$$\alpha = \frac{1}{1+3} = 0.25, \qquad 1 - \alpha = 0.75.$$

The unweighted normalized Levenshtein distance is recovered as $\tilde{d}_{1/2}(s, r) = (M(s, r)+I(s, r))/L(s)$.

### Per–read empirical quality

For a fixed aligned pair $(s_{ij}, r_{ik})$, we interpret $\tilde{d}_\alpha(s_{ij}, r_{ik})$ as an empirical estimate of the per–base error probability for that read relative to that sequence. To avoid degenerate values when the observed distance is extremely small or zero, we truncate this estimate to lie in the interval $\big[L(s_{ij})^{-2}, 1\big]$:

$$\hat{p}^{\text{read}}_{ijk} := \min\Big\{1, \max\big(L(s_{ij})^{-2}, \tilde{d}_\alpha(s_{ij}, r_{ik})\big)\Big\}. \tag{1}$$

The corresponding empirical Phred-like quality of read $r_{ik}$ with respect to $s_{ij}$ is

$$Q^{\text{read}}_{ijk} := -10 \log_{10} \hat{p}^{\text{read}}_{ijk}. \tag{2}$$

In particular, smaller normalized edit distances correspond to larger quality values.

### Sequence-level empirical quality

Fix a sequence $s_{ij}$ in experiment $e_i$, and suppose that $r_{ik}$, $k = 1, \ldots, \ell_i$, are the reads aligned to $s_{ij}$. Under a simple model in which each base in each read is independently wrong with probability $p_{\text{he}}$ ("per–base error"), the total number of errors across all alignments is approximately

$$X_{ij} \sim \text{Binomial}\big(N_{ij}, p_{\text{he}}\big), \qquad N_{ij} := \ell_i\, L(s_{ij}).$$

An unbiased and maximum-likelihood estimator of $p_{\text{he}}$ is

$$\hat{p}^{\text{seq}}_{ij} := \frac{1}{\ell_i L(s_{ij})} \sum_{k=1}^{\ell_i} E_\alpha\big(s_{ij}, r_{ik}\big) = \frac{X_{ij}}{N_{ij}}. \tag{3}$$

We then define the empirical sequence-level quality of $s_{ij}$ as

$$Q^{\text{seq}}_{ij} := -10 \log_{10} \hat{p}^{\text{seq}}_{ij}. \tag{4}$$

### Basecaller-derived read quality

The basecaller provides per–base Phred scores $Q_{ik\ell}$ for read $r_{ik}$. Converting these to per–base error probabilities $p_{ik\ell} = 10^{-Q_{ik\ell}/10}$ and averaging over the read gives the basecaller's implied per–base error probability for that read,

$$\hat{p}^{\text{bc}}_{ik} := \frac{1}{L(r_{ik})} \sum_{\ell=1}^{L(r_{ik})} 10^{-Q_{ik\ell}/10}. \tag{5}$$

The corresponding read-level quality reported by the basecaller is

$$Q^{\text{bc}}_{ik} := -10 \log_{10} \hat{p}^{\text{bc}}_{ik}. \tag{6}$$

Here the $Q_{ik\ell}$ are the per–base $q$-values produced by the basecaller.

### 1.3 Quality Metrics of the Basecaller

Finally, we define summary statistics that compare the empirical qualities above and thereby quantify the performance of a basecaller.

For each aligned pair $(s_{ij}, r_{ik})$ we have three associated qualities:

$$Q_{ij}^{\text{seq}}, \qquad Q_{ijk}^{\text{read}}, \qquad Q_{ik}^{\text{bc}}.$$

We consider the following random differences over all such triples $(i, j, k)$:

$$\Delta_{\text{seq-read}} := Q_{ij}^{\text{seq}} - Q_{ijk}^{\text{read}}, \tag{7}$$

$$\Delta_{\text{bc-read}} := Q_{ik}^{\text{bc}} - Q_{ijk}^{\text{read}}. \tag{8}$$

**Precision.** We define the *precision* of our empirical quality estimates as the empirical standard deviation of $\Delta_{\text{seq-read}}$:

$$\text{Precision} := \text{sd}\big(\Delta_{\text{seq-read}}\big). \tag{9}$$

Small values indicate that the per–read empirical qualities $Q_{ijk}^{\text{read}}$ are tightly concentrated around the sequence-level qualities $Q_{ij}^{\text{seq}}$.

**Accuracy.** We define the *accuracy* (or calibration error) of the basecaller as the empirical standard deviation of $\Delta_{\text{bc-read}}$:

$$\text{Accuracy} := \text{sd}\big(\Delta_{\text{bc-read}}\big). \tag{10}$$

A perfectly calibrated basecaller would yield both small variance and a mean of $\Delta_{\text{bc-read}}$ close to zero; in practice we report Accuracy together with its empirical mean to summarize both dispersion and bias.

These definitions provide a mathematically consistent framework for quantifying sequencing error and for comparing basecaller-reported quality values with empirically observed error rates.

## 2 Overview

This document merges and harmonizes the mathematical content for:

- single-molecule sequencing and basecalling,
- Phred quality scores and alignment-based quality metrics,
- sequence classification and confusion-matrix error models,
- haplotype / diplotype classification (including polyploidy),
- read-level haplotagging with known haplotypes,
- plasmid replication and purity bounds, and
- dual Cas9 cutting efficiency.

Each section reintroduces its own notation so that variable definitions remain consistent with the original sources.

## 3 Definitions and Mathematical Priors

We work over the DNA alphabet $\mathcal{A} = \{A, C, G, T\}$. A sequence $s$ of length $L_s$ is written

$$s = s_1 s_2 \dots s_{L_s} \in \mathcal{A}^{L_s}.$$

### 3.1 Solution purity and generated reads

In a given experiment, the sequence $s$ is present in solution with purity $\pi \in (0, 1]$, meaning that a fraction $\pi$ of molecules in solution are equal to $s$ (and the remaining molecules may be arbitrary other sequences).

Sequencing this solution with a given basecaller yields a set of reads

$$\mathcal{R}(\pi) = \{r_1, r_2, \ldots, r_{n_\pi}\},$$

where each read $r \in \mathcal{R}(\pi)$ has length $L(r) = |r|$ (which may differ from $L_s$).

For any read $r$ and the target sequence $s$, we define $d(r, s)$ to be the global alignment edit distance (number of insertions, deletions, and substitutions).

### 3.2 Per-read Q-score

For an individual read $r$ aligned to $s$, we define the normalized error rate

$$\hat{e}(r \mid s) \;=\; \max\left(\frac{1}{L(r)^2}, \frac{d(r, s)}{L(r)}\right),$$

where the floor term $1/L(r)^2$ prevents the error rate from being exactly zero.

The per-read Q-score is then

$$Q_{\text{read}}(r \mid s) \;=\; -10 \log_{10} \hat{e}(r \mid s) \;=\; -10 \log_{10} \left(\max\left(\frac{1}{L(r)^2}, \frac{d(r, s)}{L(r)}\right)\right),$$

which matches our current transformation

$$Q_{\text{read}} \;=\; -10 \log_{10}\left(\max\left(\frac{1}{L^2}, \frac{\text{Edit Distance}}{L}\right)\right),$$

with $L$ interpreted as the length of the *read.*

### 3.3 Sequence-level Q-score at purity $\pi$

We now define a single sequence-level quality score that summarizes the current basecaller performance on sequence $s$ when it is present in solution at purity $\pi$.

First, we aggregate the normalized error rates over all reads in $\mathcal{R}(\pi)$:

$$\bar{e}_s(\pi) \;=\; \frac{1}{|\mathcal{R}(\pi)|} \sum_{r \in \mathcal{R}(\pi)} \hat{e}(r \mid s).$$

The sequence-level Q-score is then

$$Q_{\text{sequence}}(s; \pi) \;=\; -10 \log_{10} \bar{e}_s(\pi).$$

We sometimes write $Q_{\text{sequence}}(\pi)$ when the underlying sequence $s$ is clear from context. This scalar $Q_{\text{sequence}}(s; \pi)$ provides a single Phred-like number summarizing the basecaller performance on sequence $s$ at the specified solution purity $\pi$, derived from the distribution of per-read Q-scores.

# 4 Single-Molecule Sequencing and Basecalling Model

## 4.1 Raw signal and segmentation

A single-molecule sequencing instrument outputs a raw signal

$$X = (x_1, x_2, \dots, x_t), \tag{11}$$

where $x_j$ is the measurement at time index $j$ and $t$ is the total number of measurements in a run.

A *single-molecule signal* (one binding event) is a contiguous subsequence of $X$:

$$\mathbf{x}^{(i)} = (x_{a_i}, x_{a_i+1}, \dots, x_{b_i}), \tag{12}$$

with length $\ell_i = b_i - a_i + 1$.

A segmentation model $\mathcal{S}$ is applied to $X$ to identify $n$ binding events:

$$\mathcal{X} = \left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right\} = \mathcal{S}(X), \tag{13}$$

where each $\mathbf{x}^{(i)}$ is intended to correspond to a single molecule interacting with the instrument.

## 4.2 Basecalling

Let $\mathcal{A}$ denote the nucleotide alphabet, typically

$$\mathcal{A} = \{A, C, G, T\} \quad \text{or} \quad \mathcal{A} = \{A, C, G, T, N\}. \tag{14}$$

A basecalling model $f$ maps each single-molecule signal to a sequence of bases (a *read*)

$$r^{(i)} = f\big(\mathbf{x}^{(i)}\big) \in \mathcal{A}^{L_i}, \tag{15}$$

where $L_i$ is the length of read $r^{(i)}$.

For each base $r_j^{(i)}$ in read $r^{(i)}$, $f$ outputs a Phred quality score $Q_j^{(i)}$ encoding an estimated error probability $p_j^{(i)}$:

$$Q_j^{(i)} = -10 \log_{10}\big(p_j^{(i)}\big), \qquad p_j^{(i)} = 10^{-Q_j^{(i)}/10}. \tag{16}$$

Collecting all reads from a run gives

$$R = \left\{r^{(1)}, \dots, r^{(n)}\right\}, \qquad Q = \left\{\mathbf{Q}^{(1)}, \dots, \mathbf{Q}^{(n)}\right\}, \tag{17}$$

where $\mathbf{Q}^{(i)} = (Q_1^{(i)}, \dots, Q_{L_i}^{(i)})$.

## 4.3 Read-level average quality (predicted and empirical)

For read $r^{(i)}$ with basewise error probabilities $p_j^{(i)}$, a predicted *per-read* error probability is

$$\bar{p}_{\text{pred}}^{(i)} = \frac{1}{L_i} \sum_{j=1}^{L_i} p_j^{(i)} = \frac{1}{L_i} \sum_{j=1}^{L_i} 10^{-Q_j^{(i)}/10}. \tag{18}$$

The corresponding average predicted Phred score is

$$\bar{Q}_{\text{pred}}^{(i)} = -10\ \log_{10}\big(\bar{p}_{\text{pred}}^{(i)}\big). \tag{19}$$

Let $s^{(i)}$ denote the (unknown) true sequence that generated signal $\mathbf{x}^{(i)}$. Let $L\big(r^{(i)}, s^{(i)}\big)$ be the Levenshtein edit distance (minimum number of insertions, deletions, and substitutions to transform one sequence into the other). An empirical error rate for read $i$ is

$$\bar{p}_{\text{emp}}^{(i)} = \frac{L\big(r^{(i)}, s^{(i)}\big)}{|s^{(i)}|}, \tag{20}$$

and the corresponding empirical quality score is

$$\bar{Q}_{\text{emp}}^{(i)} = -10\ \log_{10}\!\big(\bar{p}_{\text{emp}}^{(i)}\big). \tag{21}$$

# 5 Sequence Counts, Experiments, and Confusion Matrix

## 5.1 Individual experiments

Consider a single sequencing experiment $E$ generating a set of reads $R = \{r_1, \dots, r_N\}$. Let

$$S_E = \{s_1, \dots, s_{m_E}\} \tag{22}$$

be the set of unique sequences observed in $R$ (after collapsing identical reads).

Define the count vector $\mathbf{c}^{(E)} \in \mathbb{N}^{m_E}$ by

$$c_j^{(E)} = \big|\{r \in R \ : \ r = s_j\}\big|, \qquad j = 1, \dots, m_E. \tag{23}$$

Then

$$N = \sum_{j=1}^{m_E} c_j^{(E)} \tag{24}$$

is the total number of reads in experiment $E$.

### 5.1.1 Additional per-experiment sequence and quality statistics

For an experiment $E$ with observed reads $R = \{r^{(1)}, \dots, r^{(n)}\}$, let the set of unique sequences observed in $R$ be

$$S_E = \{s_1, s_2, \dots, s_{m_E}\}. \tag{15}$$

**Counts of unique sequences:** Let $c^{(E)}$ be the count vector over $S_E$, where $c_j^{(E)}$ is the number of reads equal to sequence $s_j$:

$$c^{(E)} = \left[c_1^{(E)}, c_2^{(E)}, \dots, c_{m_E}^{(E)}\right]. \tag{16}$$

The elements of $c^{(E)}$ sum to the total number of reads:

$$\sum_{j=1}^{m_E} c_j^{(E)} = n. \tag{17}$$

**Average predicted and empirical quality scores:** For each unique sequence $s_j \in S_E$, let $b_{\text{avg},j}$ denote the average predicted Phred quality score over reads matching $s_j$, and let $q_{\text{emp},j}$ denote the empirical read-level quality computed from Levenshtein error rates.

We collect these quantities into sets:

$$Q_{\text{avg}} = \{b_{\text{avg},1}, b_{\text{avg},2}, \dots, b_{\text{avg},m_E}\}, \tag{18}$$

$$Q_{\text{emp}} = \{q_{\text{emp},1}, q_{\text{emp},2}, \dots, q_{\text{emp},m_E}\}. \tag{19}$$

**Percentage of overestimated accuracy:** For each unique sequence $s_i$ with count $c_i^{(E)}$, define

$$I_{c_i} = \begin{cases} \sigma\Big(k\big[Q_{\text{pred}}(c_i^{(E)}) - Q_{\text{emp}}(c_i^{(E)}) - \alpha\big]\Big), & \text{if } Q_{\text{pred}}(c_i^{(E)}) > Q_{\text{emp}}(c_i^{(E)}), \\ 0, & \text{otherwise,} \end{cases} \tag{20}$$

where $\sigma(x) = \dfrac{1}{1+e^{-x}}$, and $k > 0$ and $\alpha$ are tunable parameters controlling the slope and offset, respectively.

The percentage of reads whose accuracy is overestimated by the basecaller is then

$$d = \frac{\sum_{i=1}^{m_E} c_i^{(E)} I_{c_i}}{n} \times 100. \tag{21}$$

## 5.2 Set of experiments and global sequence index

Let $\mathcal{E} = \{E_1, \dots, E_K\}$ be a set of experiments performed using the same sequencing technology.

Let the global set of unique sequences across all experiments be

$$S = \bigcup_{k=1}^{K} S_{E_k} = \{s_1, \dots, s_M\}. \tag{25}$$

For each experiment $E_k$, we can construct a count vector $\mathbf{c}^{(k)} \in \mathbb{N}^M$ over the global index $1, \dots, M$, where $c_j^{(k)}$ is the count of sequence $s_j$ in experiment $E_k$. Let $N_k = \sum_{j=1}^{M} c_j^{(k)}$ be the size of experiment $E_k$.

## 5.3 Confusion matrix and empirical error model

Using high-purity standards with known ground-truth sequences, we can construct a confusion matrix $C$ summarizing sequence-level classification performance of the basecaller.

Let $S = \{s_1, \dots, s_M\}$ be the set of possible sequences used in standards. The confusion matrix is an $M \times M$ matrix $C$ with entries

$$C_{ij} = \text{number of times a molecule of true sequence } s_i \text{ was classified as } s_j. \tag{26}$$

Diagonal elements $C_{ii}$ correspond to correct classifications; off-diagonal elements $C_{ij}$ $(i \neq j)$ correspond to misclassifications.

For a given true sequence $s_i$, define

$$N_i = \sum_{j=1}^{M} C_{ij}, \tag{27}$$

the total number of molecules of type $s_i$ used in standards.

**Correct classification (true positive rate).**

$$\text{TPR}_i = P(\hat{s} = s_i \mid s_i) = \frac{C_{ii}}{N_i}. \tag{28}$$

**Misclassification probability for sequence $s_i$.**

$$\epsilon_i = P(\hat{s} \neq s_i \mid s_i) = 1 - \mathrm{TPR}_i = 1 - \frac{C_{ii}}{N_i} = \frac{\sum_{j \neq i} C_{ij}}{N_i}. \tag{29}$$

**Pairwise misclassification probability.** For $i \neq j$,

$$P(\hat{s} = s_j \mid s_i) = \frac{C_{ij}}{N_i}. \tag{30}$$

These probabilities define a sequence-level empirical error model that can be used inside higher-level haplotype and diplotype classification models.

## 6 Mean Phred Score vs. Phred of Mean Error Probability

Let $p_1, \ldots, p_n \in (0, 1]$ be error probabilities corresponding to individual basecalls, and let the Phred quality for base $i$ be

$$Q_i = -10 \log_{10}(p_i). \tag{31}$$

Define the arithmetic means

$$p = \frac{1}{n} \sum_{i=1}^{n} p_i, \qquad Q = \frac{1}{n} \sum_{i=1}^{n} Q_i. \tag{32}$$

**Theorem 1.** *The mean Phred score $Q$ is always greater than or equal to the Phred score of the mean error probability:*

$$Q \;\geq\; -10 \log_{10}(p), \tag{33}$$

*with equality if and only if all $p_i$ are equal (or $n = 1$).*

*Proof.* The base-10 logarithm $\log_{10}(x)$ is concave on $(0, \infty)$. By Jensen's inequality, for any concave $f$,

$$f\left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) \;\geq\; \frac{1}{n} \sum_{i=1}^{n} f(x_i). \tag{34}$$

Take $f(x) = \log_{10}(x)$ and $x_i = p_i$:

$$\log_{10}\left(\frac{1}{n} \sum_{i=1}^{n} p_i\right) \;\geq\; \frac{1}{n} \sum_{i=1}^{n} \log_{10}(p_i). \tag{35}$$

Multiply both sides by $-10$, which reverses the inequality:

$$-10 \log_{10}\left(\frac{1}{n} \sum_{i=1}^{n} p_i\right) \;\leq\; -10 \cdot \frac{1}{n} \sum_{i=1}^{n} \log_{10}(p_i). \tag{36}$$

Recognizing that $Q_i = -10 \log_{10}(p_i)$, we obtain

$$-10 \log_{10}(p) \;\leq\; \frac{1}{n} \sum_{i=1}^{n} Q_i = Q. \tag{37}$$

Equality in Jensen's inequality holds if and only if all $x_i$ are equal, i.e. $p_1 = \cdots = p_n$ (or if $n = 1$). This proves the claim. □

**Interpretation.** Averaging in log-space (Phred) is more optimistic than converting the arithmetic mean error probability to a single Phred score, because the logarithm is concave and therefore gives more weight to smaller error probabilities.

## 7 Alignment-Based Quality Metric

Let $G = (g_1, \dots, g_N)$ be a ground-truth sequence and $B = (b_1, \dots, b_N)$ a basecalled sequence aligned to $G$ (gaps are allowed). Each basecall $b_i$ has a Phred quality score $Q_i$, with error probability

$$p_i = 10^{-Q_i/10}. \tag{38}$$

We define a per-alignment-column score $s_i$ as

$$s_i = \begin{cases} 1 - p_i, & \text{if } g_i = b_i, \\ p_i, & \text{if } g_i \neq b_i, \\ 0, & \text{if } g_i = \text{``-''} \text{ or } b_i = \text{``-''}. \end{cases} \tag{39}$$

**Mean correctness score.** Define

$$M = \frac{1}{N} \sum_{i=1}^{N} s_i, \tag{40}$$

so that $M \in [0, 1]$ represents the average correctness score across the alignment, weighted by both basecall accuracy and confidence.

**Phred-like aggregate quality.** Define an *effective* error probability

$$p_{\text{err}} = 1 - M, \tag{41}$$

and a Phred-like score

$$Q_{\text{new}} = -10 \log_{10}(p_{\text{err}}) = -10 \log_{10}(1 - M). \tag{42}$$

Here, $M$ is the mean correctness, $1 - M$ is the effective empirical error probability, and $Q_{\text{new}}$ is a single Phred-scale summary that incorporates both predicted and empirical accuracy.

## 8 Per-Base Variant Likelihood From Basewise Error Rates

Consider a particular haplotype (or reference sequence) $h$ with sequence

$$g = (g_1, \dots, g_L) \in \mathcal{A}^L, \tag{43}$$

and a read

$$r = (r_1, \dots, r_L) \in \mathcal{A}^L \tag{44}$$

aligned to $g$. Let $e_i \in [0, 1]$ be the error rate at base $i$ (e.g. $e_i = p_i$ from its Phred score).

Assume that for each position $i$,

$$P(r_i = g_i \mid g_i, e_i) = 1 - e_i, \tag{45}$$

$$P(r_i = b \neq g_i \mid g_i, e_i) = \frac{e_i}{|\mathcal{A}| - 1}, \qquad b \in \mathcal{A},\, b \neq g_i, \tag{46}$$

and that different positions are conditionally independent given $g$. Then the per-read likelihood is

$$P(r \mid g, \mathbf{e}) = \prod_{i=1}^{L} \Big[ (1 - e_i)\, \mathbf{1}\{r_i = g_i\} + \frac{e_i}{|\mathcal{A}| - 1}\, \mathbf{1}\{r_i \neq g_i\} \Big], \tag{47}$$

where $\mathbf{e} = (e_1, \ldots, e_L)$ and $\mathbf{1}\{\cdot\}$ is the indicator function.

This can be used as a sequence-specific likelihood term inside haplotype or molecule-of-origin likelihood calculations.

# 9 Haplotype Classification (Unknown Haplotype)

## 9.1 Sets and priors

Let

$$\mathcal{H} = \{h_1, \ldots, h_p\} \tag{48}$$

be a set of possible haplotypes for a sample (e.g. a set of known haplotypes from a population). Each $h_i$ is specified as a population of DNA molecules (chromosomes) with known sequences and stoichiometric ratios.

- $P(h_i)$ is the prior probability of haplotype $h_i$, estimated from population frequency data.
- For haplotype $h_i$, let
$$M(h_i) = \{m_{i1}, \ldots, m_{iv_i}\} \tag{49}$$
be the set of DNA molecules in that haplotype (e.g. chromosomes and plasmids).

## 9.2 Cell population and derived sequences

Let

- $C(h_i) = \{c_1, \ldots, c_w\}$ be the set of cells in a sample derived from an original genome with haplotype $h_i$,
- $U(h_i) = \{u_1, \ldots, u_x\}$ be the set of unique DNA sequences present in $C(h_i)$ after mutation accumulation.

Mutations accumulate over $n_{\text{div}}$ cell divisions at mutation rate $\mu$ for a genome (or molecule) of length $L$. Conceptually we write

$$P\big(u \mid m,\, \mu, n_{\text{div}}, L\big) \tag{50}$$

for the probability of generating a particular sequence $u$ from an ancestral molecule $m$.

## 9.3 Fragmentation, labeling, and sequencing

**Fragmentation.** Let

$$D(h_i) = \{d_1, \ldots, d_y\} \tag{51}$$

denote the set of DNA fragments after extraction and fragmentation (from $U(h_i)$). Let $\theta_{\text{frag}}$ be parameters describing the fragmentation process. Then

$$P\big(d \mid u,\, \theta_{\text{frag}}\big) \tag{52}$$

denotes the probability that a fragment $d$ is produced from a sequence $u$.

**Labeling and enrichment.** Let

$$L(h_i) = \{\ell_1, \ldots, \ell_z\} \tag{53}$$

be the set of sequences that successfully receive sequencing-specific adaptors (i.e. are labeled and enriched). Let $\theta_{\text{lab}}$ be parameters describing adapter ligation and enrichment. Then

$$P\big(\ell \mid d,\, \theta_{\text{lab}}\big) \tag{54}$$

is the probability that fragment $d$ becomes labeled sequence $\ell$.

**Sequencing and signal generation.** From the labeled sequences, a single-molecule sequencing experiment produces a set of events and signals

$$X(h_i) = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} \tag{55}$$

and corresponding basecalled reads

$$R = \{r_1, \ldots, r_n\}. \tag{56}$$

Let $\theta_{\text{seq}}$ capture sequencing and basecalling error parameters. Then

$$P\big(r \mid \ell,\, \theta_{\text{seq}}\big) \tag{57}$$

is the probability of obtaining read $r$ from labeled sequence $\ell$, incorporating the empirical error model (e.g. confusion matrix $C$).

## 9.4 Likelihood for a haplotype

Conditioning on haplotype $h_i$ and marginalizing over unobserved stages (mutation, fragmentation, labeling), a generic factorization for the likelihood of an observed read $r$ is

$$P(r \mid h_i) = \sum_{u \in U(h_i)} \sum_{d \in D(h_i)} \sum_{\ell \in L(h_i)} P(r \mid \ell, \theta_{\text{seq}}) P(\ell \mid d, \theta_{\text{lab}}) P(d \mid u, \theta_{\text{frag}}) P(u \mid h_i, \mu, n_{\text{div}}, L). \tag{58}$$

Assuming independence of reads given haplotype $h_i$,

$$P(R \mid h_i) = \prod_{r \in R} P(r \mid h_i). \tag{59}$$

## 9.5 Posterior probabilities and classification rule

By Bayes' theorem, the posterior for haplotype $h_i$ given observed reads $R$ is

$$P(h_i \mid R) = \frac{P(R \mid h_i) P(h_i)}{\displaystyle\sum_{j=1}^{p} P(R \mid h_j) P(h_j)}. \tag{60}$$

The posterior probability that the sample corresponds to *any* haplotype other than $h_i$ is

$$P\big(\text{“not } h_i\text{”} \mid R\big) = 1 - P(h_i \mid R). \tag{61}$$

Define the likelihood ratio (LR) for haplotype $h_i$ as

$$\mathrm{LR}_i(R) = \frac{P(h_i \mid R)}{P(\text{“not } h_i\text{”} \mid R)} = \frac{P(h_i \mid R)}{1 - P(h_i \mid R)}. \tag{62}$$

Given a threshold $\tau > 0$, a simple decision rule is

$$\begin{aligned} \text{Accept } h_i \text{ for the sample} \quad & \text{if} \quad \mathrm{LR}_i(R) \geq \tau, && (63)\\ \text{otherwise} \quad & \text{declare the sample “uncertain” or consider resequencing.} && (64) \end{aligned}$$

In a multiclass setting (multiple haplotypes), one may classify the sample as

$$\hat{h} = \arg \max_{1 \leq i \leq p} P(h_i \mid R), \tag{65}$$

optionally requiring that $P(\hat{h} \mid R)$ exceed a minimum probability threshold before accepting the call.

# 10 Diplotypes, Polyploidy, and Cost-Based Decision Rules

## 10.1 Additional notation for diplotypes and polyploidy

Let

- $\mathcal{D}$: set of all possible diplotypes (pairs of haplotypes),
- $K$: number of sets of homologous chromosomes (e.g. $K = 1$ for a single gene, larger for polyploid loci),
- $\gamma$: classification threshold on posterior probabilities,
- $N$: number of reads per sample used for classification,
- $\epsilon_{d \to d'}(N)$: empirical misclassification rate from true diplotype $d \in \mathcal{D}$ to diplotype $d' \in \mathcal{D}$ when $N$ reads are used,
- $C_{d \to d'}$: cost of misclassifying $d$ as $d'$,
- $C_{\mathrm{res},d}$: cost of resequencing a sample whose true diplotype is $d$,
- $\psi_d(\gamma, N)$: probability that a sample with true diplotype $d$ is flagged for resequencing, given threshold $\gamma$ and $N$ reads,
- $\pi_d$: prior probability of diplotype $d$,
- $\eta_{d,s}$: enrichment efficiency for sequence $s$ in diplotype $d$,
- $L_d(R)$: likelihood of observing reads $R$ under diplotype $d$,
- $\ell_{r,s}$: sequence-specific likelihood for signal/read $r$ and sequence $s$ (from basecalling error model).

## 10.2 Expected cost

For a diplotype $d$, the posterior probability given reads $R$ is

$$P(d \mid R) = \frac{L_d(R)\, \pi_d}{\displaystyle\sum_{d' \in \mathcal{D}} L_{d'}(R)\, \pi_{d'}}. \tag{66}$$

A decision policy maps posterior probabilities to one of three actions: *call diplotype $d'$*, *declare uncertain and resequencing*, or *no call.* Under a given policy that depends on $\gamma$ and $N$, define:

$$\epsilon_{d\to d'}(\gamma, N) = P\big(\text{policy calls } d' \neq d \,\big|\, \text{true diplotype } d,\ N,\ \gamma\big), \tag{67}$$

$$\psi_d(\gamma, N) = P\big(\text{policy chooses ``resequencing''} \,\big|\, \text{true diplotype } d,\ N,\ \gamma\big). \tag{68}$$

The expected cost for a given $(\gamma, N)$ is

$$\mathcal{C}(\gamma, N) = \sum_{d\in\mathcal{D}} \pi_d \left[ \sum_{d'\neq d} C_{d\to d'}\, \epsilon_{d\to d'}(\gamma, N) + C_{\text{res},d}\, \psi_d(\gamma, N) \right]. \tag{69}$$

An optimal pair $(\gamma^*, N^*)$ can be defined as

$$(\gamma^*, N^*) = \arg\min_{\gamma, N} \mathcal{C}(\gamma, N), \tag{70}$$

subject to constraints (e.g. minimal acceptable sensitivity, budget limits on resequencing, etc.).

# 11 Read-Level Haplotagging with Known Haplotype

In this section, the haplotype of the genomic source is assumed known: there is a single haplotype $h$ and

$$P(h) = 1. \tag{71}$$

The goal is to assign each read to its most likely *molecule of origin* (e.g. specific chromosome or plasmid) within $h$.

## 11.1 Model and notation

Let

- $M(h) = \{m_1, \ldots, m_v\}$ be the set of DNA molecules in haplotype $h$,
- $P(m_j \mid h)$ be the prior probability that a randomly selected molecule in the sample is of type $m_j$ (stoichiometric ratio),
- $P(r \mid m_j, h)$ be the likelihood of read $r$ given that its molecule-of-origin is $m_j$ (this incorporates the basecalling error model, e.g. confusion matrix and per-base error rates).

## 11.2 Posterior for molecule-of-origin of a single read

By Bayes' theorem, the posterior probability that read $r$ originated from molecule $m_j$ is

$$P(m_j \mid r, h) = \frac{P(r \mid m_j, h)\, P(m_j \mid h)}{\displaystyle\sum_{k=1}^{v} P(r \mid m_k, h)\, P(m_k \mid h)}. \tag{72}$$

The probability that $r$ originated from any other molecule is

$$P(\text{``other''} \mid r, h) = 1 - P(m_j \mid r, h). \tag{73}$$

Define the per-read likelihood ratio

$$\text{LR}_j(r) = \frac{P(m_j \mid r, h)}{P(\text{``other''} \mid r, h)} = \frac{P(m_j \mid r, h)}{1 - P(m_j \mid r, h)}. \tag{74}$$

### 11.3 Decision rule for haplotagging a read

Given a threshold $\tau > 0$, a basic decision rule for assigning read $r$ to molecule $m_j$ is

$$\text{Assign read } r \text{ to } m_j \quad \text{if} \quad \mathrm{LR}_j(r) \geq \tau, \tag{75}$$

$$\text{Declare } r \text{ “unphased” (unassigned)} \quad \text{if} \quad \mathrm{LR}_j(r) < \tau \quad \text{for all } j = 1, \dots, v. \tag{76}$$

### 11.4 Unphased reads and cost-based optimization

Let $N$ be the total number of reads from the sample. For a given threshold $\tau$, define

- $P_{\mathrm{unph}}(\tau)$: probability a randomly selected read is unphased (i.e. no molecule's LR exceeds $\tau$),
- $P_{\mathrm{mis}}(\tau)$: probability a read is assigned to the wrong molecule.

Then the expected number of unphased reads is

$$U(\tau, N) = N\, P_{\mathrm{unph}}(\tau). \tag{77}$$

Let $C_{\mathrm{mis}}$ be the cost of misclassifying a read and $C_{\mathrm{unph}}$ be the cost associated with leaving a read unphased. An example cost function is

$$\mathcal{C}(\tau, N) = C_{\mathrm{mis}}\, N P_{\mathrm{mis}}(\tau) + C_{\mathrm{unph}}\, U(\tau, N). \tag{78}$$

An optimal threshold $\tau^*$ may be defined as

$$\tau^* = \arg\min_{\tau} \mathcal{C}(\tau, N). \tag{79}$$

In practice, $P_{\mathrm{mis}}(\tau)$ and $P_{\mathrm{unph}}(\tau)$ are estimated from standards or simulations, using the same empirical error model (confusion matrix, per-base error rates) that defines $P(r \mid m_j, h)$.

## 12 Plasmid Replication, Mutation, and Purity Bounds

### 12.1 Basic assumptions

- A plasmid is an extragenomic circular dsDNA molecule that can be maintained at (on average) single-copy per bacterium if it contains a functional single-copy origin of replication (ORI).
- An *E. coli* chromosome is a single circular dsDNA molecule of length on the order of $10^6$ bp.
- A typical plasmid is two orders of magnitude shorter than the chromosome, often between $3 \times 10^3$ and $10^4$ bp.
- The per-base replication error rate of the *E. coli* replisome is on the order of $10^{-10}$ errors per base per replication.

Consider a single isolated *E. coli* cell containing:

- one genomic chromosome,
- one single-copy plasmid.

After one replication (cell division), both DNA molecules replicate and segregate into two daughter cells, yielding two chromosomes and two plasmids in total, with the possibility of replication errors.

## 12.2 Theoretical purity under replication errors

Let

- $r$ be the per-base replication error rate,
- $L$ be the plasmid length in bp,
- $k$ be the number of replication cycles.

The probability that a single base remains error-free in a single replication is $(1 - r)$. After $k$ replications, the probability that this base is still identical to the original is $(1 - r)^k$.

Assuming independence across bases, the probability that *all* $L$ bases of the plasmid remain error-free after $k$ replications is

$$P_{\text{pure}}(k) = (1 - r)^{Lk}. \tag{80}$$

This $P_{\text{pure}}(k)$ serves as an upper bound on the fraction of plasmids that remain identical to the original sequence after $k$ replication cycles (i.e. an upper bound on purity).

For small $r$ and large $Lk$, the approximation

$$P_{\text{pure}}(k) \approx \exp(-rLk) \tag{81}$$

is often useful.

## 12.3 Purity Q-value

Let $P = P_{\text{pure}}(k)$ be the theoretical purity. Define the *mutated fraction* (fraction of molecules that contain at least one replication error) as

$$P_{\text{mut}} = 1 - P. \tag{82}$$

A Phred-like *purity Q-value* can be defined by treating $P_{\text{mut}}$ as an error probability:

$$Q_{\text{pur}} = -10 \log_{10}(P_{\text{mut}}) = -10 \log_{10}(1 - P_{\text{pure}}(k)). \tag{83}$$

Higher $Q_{\text{pur}}$ indicates a smaller mutated fraction (higher purity). This quantity can be plotted as a function of $k$, $L$, and $r$ to visualize how purity decays with replication cycles and plasmid length.

## 12.4 Lower-bound purity estimate from capillary electrophoresis

Let

- $C_{\text{major}}$: concentration of the major (intended) plasmid sequence,
- $C_{\text{other}}$: concentration of all other sequences (variants) in the sample.

Capillary electrophoresis (CE) size distributions plus standards of known concentration and length can be used to estimate $C_{\text{major}}$ and $C_{\text{other}}$.

A lower bound on purity is

$$P_{\text{low}} = \frac{C_{\text{major}}}{C_{\text{major}} + C_{\text{other}}}. \tag{84}$$

### 12.5 Empirical purity from clonal expansion and Sanger sequencing

For a standard experiment $E$ with nominal sequence $s^\star$:

- Let $a$ be the number of single-colony expansions whose Sanger sequencing matches $s^\star$,
- Let $b$ be the total number of colonies sequenced.

Then an empirical purity estimate is

$$\hat{P}_E = \frac{a}{b}. \tag{85}$$

For all samples derived from a given *E. coli* strain with original sequence $s_0$:

- Let $c$ be the number of colonies whose Sanger sequencing matches $s_0$,
- Let $d$ be the total number of colonies pooled across experiments.

Then an overall empirical purity estimate is

$$\hat{P}_{\text{strain}} = \frac{c}{d}. \tag{86}$$

Comparing $P_{\text{low}}$, $P_{\text{pure}}(k)$, and empirical estimates $\hat{P}_E$ and $\hat{P}_{\text{strain}}$ provides a consistency check between theoretical replication-based purity bounds and experimental measurements.

## 13 Dual Cas9 Cutting: Probability of Isolating a Gene

We model the probability of successfully isolating a gene (or locus) via dual Cas9 cutting, accounting for fragment length distribution and cutting efficiencies.

### 13.1 Definitions

- $G$: length of the gene (or distance between two Cas9 target sites) in base pairs (bp).
- $L$: length of a DNA fragment in bp.
- $f_L(\ell)$: probability density function (pdf) of fragment lengths.
- $F_L(\ell) = P(L \le \ell)$: cumulative distribution function (cdf).
- $p_{\text{frag}}(G) = P(L \ge G)$: probability that a fragment is at least as long as $G$.
- $e_1, e_2$: base cutting efficiencies at Cas9 target sites 1 and 2, respectively.

### 13.2 Probability that a fragment can contain the full gene

$$p_{\text{frag}}(G) = P(L \ge G) = 1 - F_L(G^-), \tag{87}$$

where $F_L(G^-)$ denotes the limit of $F_L(\ell)$ from the left at $\ell = G$ (or simply $F_L(G)$ when the distribution is continuous).

### 13.3 Dual Cas9 cutting probability

Assuming:

- Fragmentation is random and independent of Cas9 cutting.
- Cas9 cutting at each of the two target sites is independent, with probabilities $e_1$ and $e_2$.

Then the probability that both cuts are successful *given* that the fragment is long enough is

$$p_{\text{cut}|\text{frag}} = e_1 e_2. \tag{88}$$

The overall probability of successfully isolating the gene via dual Cas9 cutting is

$$p_{\text{dual}}(G) = p_{\text{frag}}(G) \cdot e_1 e_2 = \big[1 - F_L(G^-)\big] e_1 e_2. \tag{89}$$

### 13.4 Example: Exponential fragment size distribution

If fragment lengths are exponentially distributed with rate $\lambda$ (so mean fragment length $= 1/\lambda$):

$$f_L(\ell) = \lambda e^{-\lambda \ell}, \qquad \ell \geq 0, \tag{90}$$

then

$$F_L(\ell) = 1 - e^{-\lambda \ell}, \qquad P(L \geq G) = e^{-\lambda G}. \tag{91}$$

So

$$p_{\text{dual}}(G) = e^{-\lambda G} e_1 e_2. \tag{92}$$

This explicitly shows the exponential decrease in success probability with gene length $G$ under random fragmentation.

## 14 Summary of Key Notation

For convenience, we collect here several commonly used symbols (many are also redefined locally in their sections):

- $X$ : raw single-molecule signal from an experiment (sequence of current levels).
- $\mathbf{x}^{(i)}$ : single-molecule signal (one binding event).
- $R = \{r_1, \ldots, r_n\}$ : set of basecalled reads.
- $\mathcal{A}$ : nucleotide alphabet (e.g. $\{A, C, G, T\}$).
- $Q_i$ : Phred quality score for base $i$.
- $p_i$ : error probability for base $i$, $Q_i = -10 \log_{10}(p_i)$.
- $M$ : mean correctness score in alignment-based metric.
- $Q_{\text{new}}$ : Phred-style summary from $M$: $Q_{\text{new}} = -10 \log_{10}(1 - M)$.
- $S$ : set of unique sequences across experiments.
- $C_{ij}$ : confusion-matrix entry (true $s_i$, predicted $s_j$).
- $\epsilon_i$ : misclassification probability for sequence $s_i$.
- $\mathcal{H} = \{h_1, \ldots, h_p\}$ : set of possible haplotypes.
- $P(h_i)$ : prior probability of haplotype $h_i$.
- $M(h)$ : set of molecules (chromosomes, plasmids) in haplotype $h$.
- $P(r \mid h_i)$ : likelihood of reads under haplotype $h_i$.
- $P(h_i \mid R)$ : posterior probability of haplotype $h_i$ given reads $R$.
- $\text{LR}_i$ : likelihood ratio for haplotype or molecule $i$.
- $\mathcal{D}$ : set of possible diplotypes.
- $\pi_d$ : prior probability of diplotype $d$.
- $\epsilon_{d \to d'}$ : misclassification rate from diplotype $d$ to $d'$.
- $C_{d \to d'}$ : cost of misclassifying $d$ as $d'$.
- $C_{\text{res},d}$ : cost of resequencing when true diplotype is $d$.
- $r$ : per-base replication error rate (in plasmid model).
- $L$ : plasmid length (bp).
- $k$ : number of replication cycles.
- $P_{\text{pure}}(k) = (1 - r)^{Lk}$ : theoretical purity after $k$ cycles.
- $P_{\text{low}}$ : lower-bound purity from CE and standards.
- $G$ : gene length (distance between dual Cas9 cut sites).
- $L$ (in Cas9 model) : fragment length (bp) with pdf $f_L$ and cdf $F_L$.
- $e_1, e_2$ : Cas9 cutting efficiencies at two sites.