Journal Title Here, 2022, pp. 1–9

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Skiver: Alignment-free Estimation of Sequencing Error Rates and Spectra using (k, v)-mer Sketches

Zhenhao Gu,[1,2] Puru Sharma,[1] Limsoon Wong[1,*] and Niranjan Nagarajan[1,2,3,*]

[1]Department of Computer Science, School of Computing, National University of Singapore, 117417, Singapore, [2]Genome Institute of Singapore, A*STAR, 138672, Singapore and [3]Yong Loo Lin School of Medicine, National University of Singapore, 117596, Singapore
[*]Corresponding authors. wongls@comp.nus.edu.sg and nagarajann@gis.a-star.edu.sg

Abstract

Background. Quality control of sequencing datasets is an important first step in numerous bioinformatics pipelines such as mapping, variant calling, and assembly. Existing methods typically rely on alignment results or quality scores. However, the reference genome is not always available for mapping, and uncalibrated quality scores may yield biased estimates of error rates.
Results. We present *skiver*, a reference-free and alignment-free framework that estimates sequencing errors using (k, v)-mer sketches. By identifying the consensus through the sketched (k, v)-mers, skiver estimates survival and hazard rates that capture positional information of sequencing errors. Across simulated and real datasets from various sequencing platforms, skiver accurately recovers error rates and spectra. It also reliably handles complex datasets containing multiple strains, alleles, and repetitive regions through an outlier filtering strategy. Skiver is computationally efficient and provides a lightweight solution for error profiling in high-throughput sequencing.
Availability and Implementation. The implementation of skiver is available at `https://github.com/GZHoffie/skiver`, and the dataset and scripts for reproducibility are available at `https://github.com/GZHoffie/skiver-test`.

Key words: sequencing error profiling, survival analysis, alignment-free methods, high-throughput sequencing

Introduction

Given a set of sequenced reads, accurately estimating the overall sequencing error rate and the spectrum of error types (substitutions, insertions, and deletions) is a fundamental task in computational biology. These statistics are central to routine *quality control*. For example, they are used to decide whether a run is usable, to compare runs across different flow cells, chemistries, or library preparations, and to detect systematic failure modes [5]. They also directly affect downstream inference. In particular, error profiles influence alignment scoring and chaining heuristics in read mapping [15], and serve as key priors or likelihood components in sensitive variant calling [36] and strain-aware analysis such as phasing [30]. Because many pipelines implicitly assume a particular error model, biased error-rate estimates can propagate and lead to avoidable miscalls, unstable parameter tuning, and misleading biological conclusions.

The challenge is amplified in *metagenomic* settings. Real samples often contain a mixture of organisms with highly uneven abundance, including low coverage genomes and closely related coexisting strains. These properties can confound error-rate estimation in two ways: (i) true biological variation and strain heterogeneity can be mistaken for sequencing errors, and (ii) reference genomes may be incomplete or unavailable because organisms are missing from databases or differ substantially from available references. Consequently, methods that rely on a single known reference, or assume a single homogeneous genome, can produce biased estimates precisely in the scenarios where robust error profiling is most needed.

Previous work

Existing approaches for estimating sequencing error rates and spectra can be broadly divided into *reference-based* and *reference-free* methods. Reference-based methods infer errors from alignments to an external reference, whereas reference-free methods avoid mapping and instead rely on internal read statistics such as k-mer frequency patterns or basecalling quality scores. Below, we summarize the limitations of each category, with an emphasis on failure modes that are especially pronounced in metagenomic samples.

Reference-based methods.
A common and often effective strategy is to map the reads to a reference genome and compute error statistics from the resulting alignments (e.g., via CIGAR operations) [15, 18]. When a high-quality, closely matching reference is available, this approach can provide accurate estimates. However, it has two major drawbacks. First, mapping and alignment can be computationally expensive for large read sets and for long

references, making it costly as a routine QC step. Second, and more critically for metagenomics, suitable references may be missing, incomplete, or diverge from the sequenced genomes because organisms are not represented in databases or because strains differ substantially from available references. In such cases, true biological differences are conflated with sequencing errors, inflating apparent mismatch and indel rates and biasing the estimated sequencing error rates upward [26]. One possible mitigation is to assemble the reads to obtain a sample-specific consensus and then compare the reads to this consensus [19], but assembly can be computationally intensive and may be unstable for low-coverage components and complex mixtures.

Reference-free methods.
To avoid dependence on an external reference, many approaches estimate error rates directly from read-derived statistics. For example, *shadow regression* [34] and *SequencErr* [6] utilize overlapping (paired-end) short reads to quantify disagreements between reads. Another prominent family of methods leverages k-mer frequency information to infer error-related quantities without explicit alignment [2, 3, 9, 10, 12, 21, 37, 25]. While scalable, these methods are often tailored to specific regimes (e.g., single-genome assumptions or specific sequencing platforms) and can struggle to recover the *full error spectrum* (i.e., distribution of substitutions vs. insertions vs. deletions). Moreover, the severe coverage imbalance typical of metagenomic samples introduces a key pitfall: low-coverage genomes yield sparse and highly variable k-mer counts, making frequency-based inference unstable and causing rare true k-mers to be difficult to distinguish from erroneous k-mers induced by sequencing noise.

Another class of reference-free methods uses Phred quality scores, which theoretically satisfy $Q = -10\log_{10}\Pr[\text{error}]$ and are used by many downstream tools [36, 38]. However, in practice, quality scores can be miscalibrated and vary with sequencing technology [7] and library preparation [28]. Some calibration procedures do exist [17], but require additional assumptions and data. Consistent with these observations, our experiments show that uncalibrated quality scores can substantially underestimate or overestimate true error rates. Moreover, they do not directly provide reliable estimates of the full substitution/insertion/deletion spectrum.

Our contribution

In this work, we propose *skiver*, which uses (k, v)-mer sketches for reference-free profiling of sequencing error rates and spectra. The core idea is to construct sketches that group (k, v)-mers sharing the same *key* and then detect structured variation patterns in the associated *values*; these patterns provide statistical signals for substitutions and indels without requiring alignment to a reference genome. Rather than relying on a single reference or trusting potentially miscalibrated quality scores, we aggregate evidence across key-sharing groups to infer error processes from observable variation events in the read set. Across simulated and real datasets from multiple sequencing platforms, our method yields accurate error-rate estimates while remaining robust in settings where reference genomes are unavailable or incomplete, including metagenomic regimes with uneven coverage and strain heterogeneity.

Methods

Problem formulation

The definition of sequencing error rate varies across the literature. It is usually defined as the number of errors divided by the alignment length [27]. In this paper, we borrow concepts from survival analysis. Let T ($T \geq 1$) be the random variable that is the number of bases from a random starting position until the first failure (disagreement between the sequenced base and the true underlying base).

Definition 1 The *hazard rate* at the t-th base from the starting position, denoted by $h(t) := \Pr[T = t \mid T \geq t]$, is the conditional probability of the t-th base disagreeing with the underlying genome given that the previous $(t-1)$ bases agree.

The benefits of using the hazard rate instead of the error rate inferred from alignment include the following. Firstly, alignment can be slow and ambiguous. Different alignment scoring schemes may yield different numbers of aligned bases, resulting in varying estimates using different mappers [8]. Secondly, the hazard rate offers a more comprehensive view of the position-dependent error process. For example, imagine two reads in **Figure 2**, where read A has 4 consecutive errors every 16 bases, and read B has 1 error every 4 bases. The error rates of the two reads are the same and are equal to 1/4, but their hazard rates are drastically different. Specifically, read A has a much lower hazard rate at small t, indicating a higher chance of observing a long stretch of consecutive matches.

Definition 2 The *survival rate* at the t-th base, denoted $S(t) := \Pr[T > t]$, is the probability of the first t bases from the point of observation being free of sequencing errors.

Note that the survival rate $S(k) = \prod_{t=1}^{k}(1 - h(t))$ also indicates the probability of a k-mer in the read being free of sequencing error, and $h(1) = 1 - S(1)$ represents the probability of a random base being erroneous. In this work, we aim to propose an efficient and accurate estimator for hazard rate and survival rate using (k, v)-mer sketches, and show its ability to estimate the frequency of each error type (substitutions, insertions, and deletions) at the same time.

(k, v)-mer sketches

Definition 3 A (k, v)-mer (as shown in **Figure 1.A**) is a segment of DNA of length $k + v$, with the first k bases being the *key* and the last v bases being the *value*.

The structure of (k, v)-mer is similar to the idea of anchor-target k-mer pairs in SPLASH [4, 13], but different in that our key and value must be adjacent for accurate error profiling, whereas the anchor-target pair can be separated in the reads. In this paper, we focus on using this structure for sequencing error rate and spectra estimation.

A (k, v)-mer sketch of a set of sequenced reads is created with the following steps (**Figure 1.B**):

- **Step 1.** We extract all the (k, v)-mers from the reads and their reverse complement. Optionally, only the forward strand of the reads is used.
- **Step 2.** We subsample roughly $1/c$ (k, v)-mers with FracMinHash [11]. In particular, given a random hash

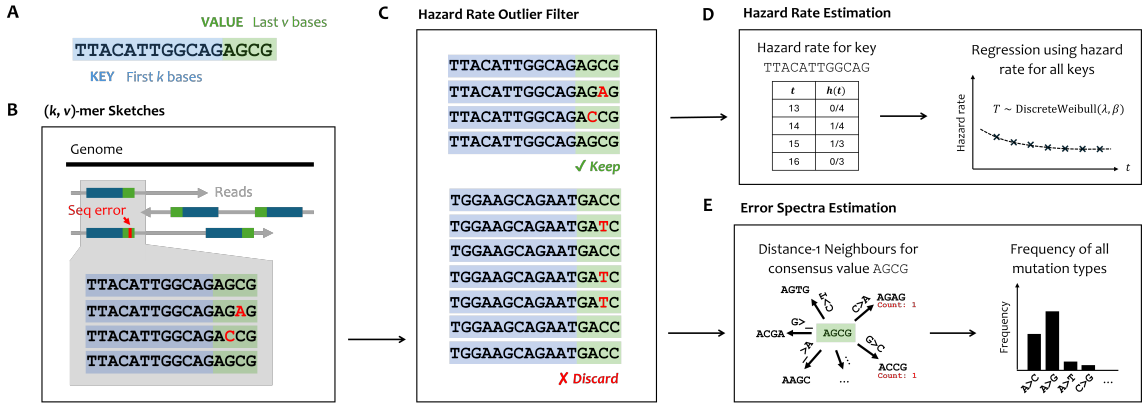

Fig. 1. The workflow of skiver. **A**. The structure of a (k, v)-mer. **B**. Extracting (k, v)-mer sketches from the read set. (k, v)-mers with the same keys are gathered together to infer the consensus value and identify sequencing errors. **C**. The sketched (k, v)-mers go through a hazard rate outlier filter to exclude keys that are associated with multiple values with high counts. **D**. The hazard rate is calculated for each key and used to fit the parameters of a Discrete Weibull distribution. **E**. We find all the distance-1 neighbors of the consensus and infer the frequency of each type of sequencing error (insertion, deletion, and substitution).

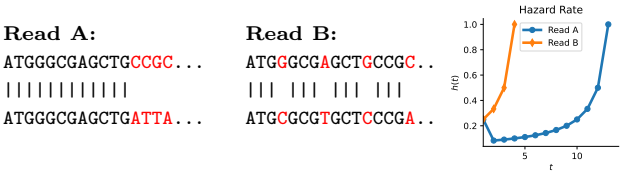

Fig. 2. The alignment of two reads (bottom sequence) to the reference (top sequence). Both reads have an error rate of 1/4 but different hazard rates $h(t)$ (right).

function $H : \Sigma^k \to (0, 1)$, we subsample a (k, v)-mer if the hash value of its key is less than $1/c$ ($c = 1000$ be default).

- **Step 3.** All the subsampled (k, v)-mers are stored in a hash table that maps a key to the set of associated values along with the number of times they appear in the read set. The most frequently appearing value is identified as the *consensus*.

Here, k is chosen to be large ($k = 21$ by default) such that the keys are mostly unique in the set of sequenced genomes. The keys are used as positional identifiers in the genomes. We then identify variation in the associated values to determine possible sequencing errors or mutations.

It is possible to show that given a per-base error rate of ε, if the number of times the keys appear in the read set $N_{\text{key}} = \Omega((1-\varepsilon)^{-2v} \log v)$, the value with the highest count (consensus) matches the true value from the sequenced genome with high probability (**Supplementary Note S1**). In practice, we simply choose a default threshold $N_{\text{key}} = 5$, only use the keys above this coverage for subsequent tasks, and show that this is sufficient for the evaluated datasets.

If a reference genome is provided, we also extract the set of (k, v)-mers from the genome using the same hash function. If a key K is associated with multiple different values in the reference genome, which indicates a repeat sequence, the key is discarded. Otherwise, the unique value associated with the key is regarded as the *consensus*. Since the consensus is obtained from the reference, a lower threshold for key multiplicity is set ($N_{\text{key}} = 1$) to allow profiling of lower coverage read sets.

Estimating hazard rate and survival rate

In this work, we assume that T follows a discrete Weibull distribution with parameters λ and β, which is often used for discrete-time survival analysis. This assumption comes from the empirical observation that hazard rates $h(t)$ in real sequencing datasets decrease with t (**Figure 5.A**), and the survival rate $S(t)$ fits well to the curve $S(t) = \exp(-\lambda t^{\beta})$ (**Supplementary Figure S1**).

Our model differs from existing error models which typically assume a constant error rate [21], essentially assuming $\beta = 1$. In real datasets, especially Nanopore and Illumina, the best fitted parameter $\hat{\beta}$ is much smaller than 1. This indicates a decreasing hazard rate, and is often interpreted as evidence for heterogeneity in the failure process [33], or a heterogeneous sequencing error rate across reads in our case.

Given a (k, v)-mer sketch of the sequenced reads, we count $N_{K,t}$, the number of (k, v)-mers that have key K and match with the consensus up to the t-th base. For example, in the example of **Figure 1.B**, 4 (k, v)-mers share the key $K =$ `TTACATTGGCAG`. The consensus value is taken as the value with the highest count, in this case, `AGCG`. Two of the (k, v)-mers differ from the consensus in the 14th and 15th base, respectively. We therefore have $N_{K,12} = N_{K,13} = 4$, $N_{K,14} = 3$, $N_{K,15} = N_{K,16} = 2$.

This process is repeated for all keys. Assuming the hazard rate at the t-th base is $h(t)$, we should have $N_{K,t} \sim \text{Binomial}(N_{K,t-1}, 1 - h(t))$. We take the maximum likelihood estimate,

$$\hat{h}(t) = 1 - \frac{\sum_K N_{K,t}}{\sum_K N_{K,t-1}} \qquad \text{for } k < t \le k + v, \tag{1}$$

This allows us to estimate $h(t)$ in a small interval between k and $k + v$. Under the assumption that T follows a discrete Weibull distribution, we have

$$\begin{aligned} & h(t) = 1 - \exp\left(-\lambda\left(t^{\beta} - (t-1)^{\beta}\right)\right) \\ \Rightarrow \quad & \log(1 - h(t)) = -\lambda\left(t^{\beta} - (t-1)^{\beta}\right) \approx -\lambda\beta t^{\beta-1} \\ \Rightarrow \quad & \log(-\log(1 - h(t))) \approx \log(\lambda\beta) + (\beta - 1)\log t \end{aligned}$$

This transformation (complementary log-log) is widely used in discrete survival analysis [1]. We can then perform a ridge

regression with Huber loss of $\log(-\log(1-h(t)))$ vs. $\log t$ in the range $[k+1, k+v]$. Let a and b be the estimated slope and intercept. Then, we have

$$\hat{\beta} = a + 1, \qquad \hat{\lambda} = \exp(b)/\hat{\beta}.$$

The survival rate can then be estimated by directly plugging the estimated parameters, $\hat{S}(t) = \exp(-\hat{\lambda}t^{\hat{\beta}})$, and the sequencing error rate is estimated to be $\hat{h}(1) = 1 - \hat{S}(1) = 1 - \exp(-\hat{\lambda})$. This process takes time and space that are linear in v and the number of keys in the sketch.

Estimating error spectra

To estimate the composition of errors (types of substitutions, insertions, and deletions), we make an additional assumption that the error composition is roughly constant and is independent of t. In other words, for a given single base error type e (such as substitution `C` → `A`), the hazard rate of this type of error happening can be expressed as

$$h_e(t) = \pi_e h(t),$$

where π_e is a constant value. The probability of the type of error e happening exactly once in the value, while no other error is present, is

$$\begin{aligned}&\Pr[\text{error } e \text{ happens exactly once in the value}]\\ =\ &\sum_{t=k+1}^{k+v} S(t-1)h_e(t)S(k+v-t) \propto \pi_e,\end{aligned}$$

given that k and v are fixed. Under this assumption, we can estimate π_e by the frequency of e happening exactly once in the value.

After identifying the consensus value for each key, we find the set of all pairs `(value, edit_type)`, where the `value` can be obtained from the consensus via one edit of `edit_type` (such as `A` → `C`). An example of the neighbor set can be found in **Supplementary Table S1**. If a neighbor can be reached via multiple types of edit, we mark the `edit_type` to be `Ambiguous`.

We then count the number of times each distance-1 neighbor appears in the (k, v)-mer sketch. In the example in **Figure 1.E**, the neighbors corresponding to the single base substitutions `C` → `A` and `G` → `C` appear once respectively. This process is repeated for all keys, which gives us the total number of times each `edit_type` has appeared. The frequency of an `edit_type` is estimated to be its count divided by the sum of counts of all the profiled error types. The `Ambiguous` neighbors are not counted.

There are at most $11v$ distance-1 neighbors given a consensus value. As a result, the time needed to count the error frequencies is only $\mathcal{O}(v \cdot \#\text{keys} + \#\text{values})$, where #keys and #values are the total number of keys and values in the (k, v)-mer sketch. This is much more efficient than the case in which we align all values to the consensus, which can take time $\mathcal{O}(v^2 \cdot \#\text{values})$.

Dealing with repeats, multiple alleles or strains

In real datasets, it is common that the sequenced genomes contain long repetitive regions and heterozygous sites in the case of the human genome, or multiple co-existing strains of the same species in the case of metagenomic samples. In these cases, a key can be associated with multiple values in the sequenced genome, causing overestimation of the hazard rate (**Figure 3.E**).

Algorithm 1: Outlier filter in hazard rate estimation

Input: $N_{K,t}$ for all keys K and $k < t \le k + v$.
Output: A set of keys that pass the filter.

```
/* Initialization of the frontier                    */
keep_key ← {};
foreach key K do
    keep_key[K] ← true;
end
/* Exclude outliers                                  */
for t ← k + 1 to k + v do
    hazard_rates ← [];
    foreach key K do
        h_K(t) ← 1 − N_{K,t}/N_{K,t−1};
        if h_K(t) > 0 then
            hazard_rates.append(h_K(t));
        end
    end
    IQR ← interquartile_range(hazard_rates);
    foreach key K do
        if h_K(t) > median(hazard_rates) + 3 × IQR then
            keep_key[K] ← false;
        end
    end
end
return {K : keep_key[K] is true};
```

A key observation is that if a key K is associated with multiple values that have high counts, the estimated hazard rate $h_K(t) = 1 - N_{K,t}/N_{K,t-1}$ of that key at one of $k < t \le k + v$ is going to be significantly higher. We therefore use a simple outlier filter as shown in **Algorithm 1** that filters all K that have a significantly higher $h_K(t)$.

If the reference genome is provided, this filter is disabled by default as keys that are associated with multiple values are already discarded.

Results

Baselines and datasets

To test the ability of our algorithm to estimate $h(t)$ and $S(t)$, we benchmarked against popular state-of-the-art baselines.

For mapping-based tools, we used BEST [18] to profile the BAM output of Minimap2 [15], assuming that the correct references are known. We used the field `matches_per_kbp` to infer the chance of observing a match in the read, and

$$\hat{\varepsilon}_{\text{BEST + Minimap2}} = 1 - \frac{\texttt{matches\_per\_kbp}}{1000}.$$

For reference-free methods, we used GenomeScope2.0 [25] to fit the k-mer histogram output of KMC 3.2.4 [14], and used the `Read Error Rate` field in the summary file. For quality-score-based methods, we used seqtk [16] and the `ErrQ` field to compute sequencing error rates,

$$\hat{\varepsilon}_{\text{seqtk}} = 10^{-\texttt{ErrQ}/10}.$$

We selected a wide range of real sequencing datasets from multiple sources and sequencing platforms (**Supplementary Table S2**) [23, 20, 29], including mock bacterial communities, bacterial isolates, and human reads. All of the chosen datasets have well-defined reference genomes for fair comparisons.

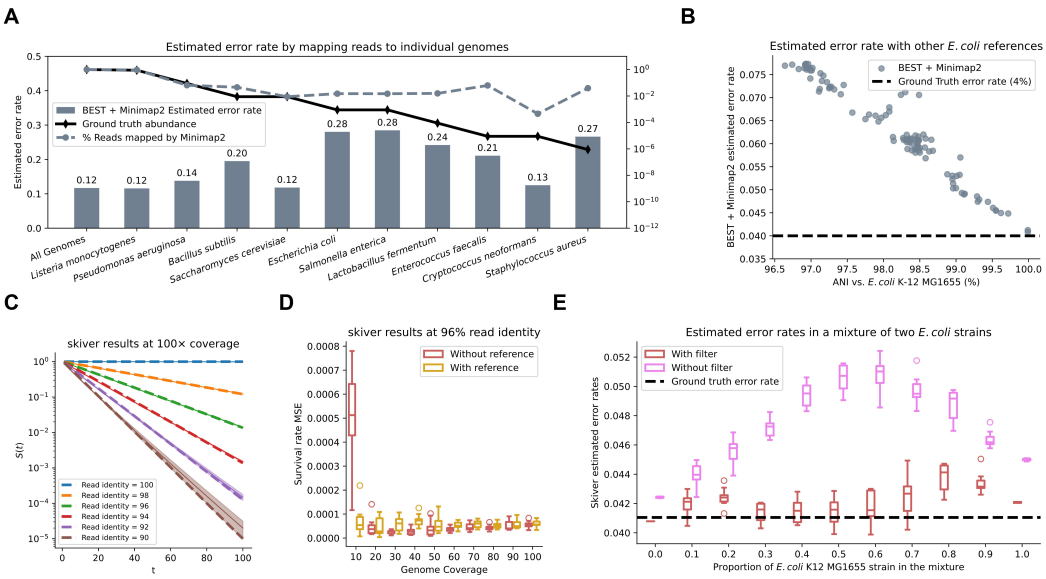

Fig. 3. Performance of the tools under simulated conditions. **A**. Profiled error rate for BEST + Minimap2 on the Zymo Log mock community dataset, if only one bacterial isolate is given as the reference (on the x-axis). The lines representing the abundance and the % reads mapped use the axis on the right of the figure. **B**. Profiled error rate for BEST + Minimap2 for a simulated read from an *E. coli* K-12 MG1655 strain, using different *E. coli* strains as references. **C**. Estimation of survival rates $S(t)$ for simulated datasets of *E. coli* with 100× coverage. The dashed line represents the ground truth $S(t)$ at different read identities (%), the solid line represents the mean of the estimates $\hat{S}(t)$, and the shaded area covers the range between the min and max of $\hat{S}(t)$ in the experiments. **D**. MSE_S of skiver estimates at different coverage values with simulated reads that have 96% read identity. **E**. Estimated error rates for skiver when mixing two *E. coli* strains together at different proportions, with and without the outlier filter.

Evaluation metrics

In all of the experiments, we calculate the ground truth survival rate $S(t)$ by counting the proportion of t-mers in the reads that align with the reference without any error in the BAM output of Minimap2. The ground truth hazard rate is calculated by $h(t) = 1 - S(t)/S(t-1)$ for $t \geq 2$ and $h(1) = 1 - S(1)$.

The main metrics we used to evaluate the tools are the mean squared error (MSE) of the estimated survival rate $\hat{S}(t)$ vs. the ground truth $S(t)$, calculated by

$$\text{MSE}_S := \frac{1}{n}\sum_{t=1}^{n}\left(S(t) - \hat{S}(t)\right)^2,$$

where the upper bound n is chosen to be 100. This is similar to the definition of Brier scores, which are used extensively in the evaluation of survival curve estimation [32]. Intuitively, this metric measures how well the tools predict the proportion of the t-mers in the read set that are error-free for various t. Similarly, we define the MSE for hazard rate to be $\text{MSE}_h := \frac{1}{n}\sum_{t=1}^{n}(h(t) - \hat{h}(t))^2$. Here, $\hat{h}(t) = 1 - \exp(-\hat{\lambda}(t^{\hat{\beta}} - (t-1)^{\hat{\beta}}))$ and $\hat{S}(t) = \exp(-\hat{\lambda}t^{\hat{\beta}})$ for skiver, while $\hat{h}(t) = \hat{\varepsilon}$, $\hat{S}(t) = \exp(-\hat{\varepsilon}t)$ for the other methods that only report one error rate estimate $\hat{\varepsilon}$.

Results for mapping-based methods may be biased by wrong or missing references

In this section, we evaluate the behavior of mapping-based methods under simulated conditions in which the reference genome is missing or incorrect.

We first consider the scenario of an incomplete reference database, a common situation in metagenomic analyses. Using the *Zymo Log mock community (Nanopore GridION)* dataset, we applied Minimap2 to map the reads under two reference configurations: (i) using all 10 known genomes in the mock community as references, and (ii) using each individual genome alone as the reference (**Figure 3.A**).

Interestingly, the estimated error rate obtained using the complete reference set is the lowest (approximately 12%), whereas estimates derived from individual reference genomes are consistently biased upwards. Examination of the mapping results reveals that Minimap2 assigns a disproportionately large number of reads to certain isolate genomes relative to their expected abundances in the mock community. This indicates that many reads originating from other species are incorrectly mapped to these genomes. As a result, the apparent sequencing error rate is substantially inflated. These observations suggest that read mapping with an incomplete reference database is inherently unreliable for error-rate profiling. While it may be possible to reduce spurious mappings by applying more stringent alignment score thresholds, the optimal threshold itself depends on the unknown sequencing error rate, creating a circular dependency that limits the robustness of mapping-based approaches.

Next, we examine the case in which the reference genome is present but does not exactly match the sequenced strain. Reads were simulated from *E. coli* K-12 MG1655 at 96% read identity using Badread [35]. Error rates were then estimated using Minimap2 and BEST, with reference genomes chosen from other *E. coli* strains randomly sampled from GTDB [24].

The average nucleotide identity (ANI) between each reference genome and *E. coli* K-12 MG1655 was computed using skani [31], and all selected references were verified to have ANI values greater than 95%. Despite this high sequence similarity, the estimated error rate almost doubled as the ANI between the reference genome and the true sequenced genome decreased to 96% (**Figure 3.B**). These results demonstrate

that even modest strain-level divergence can severely bias error-rate estimates derived from read mapping. Consequently, accurate profiling using mapping-based methods requires the exact strain, or a near-identical genome, to be present in the reference database, a condition that is rarely met in real metagenomic samples.

Skiver accurately estimates hazard rates and survival rates in simulated reads

Next, we test the behavior of skiver on simulated reads.

We first evaluated skiver under a simplified setting in which the read set contains only a single genome. Reads from the *E. coli* K-12 MG1655 strain were simulated across a range of read identities (90% to 100%) and sequencing coverages (10× to 100×).

Across these conditions, skiver accurately recovered both the survival rate and the hazard rate. For example, at 100× coverage (**Figure 3.C**), the estimated survival curve $\hat{S}(t)$ (solid line) closely matches the ground-truth $S(t)$ (dashed line), exhibiting minimal bias and low variance. When the read identity is fixed (**Figure 3.D**), the mean squared error of the survival rate estimate (MSE_S) remains close to zero as coverage decreases, until coverage drops to 10×. At this depth, the default skiver configuration shows increased variance, whereas providing a reference genome restores accuracy and reduces variance.

Based on the complete set of simulations (**Supplementary Figures S3, S4**), skiver produces generally reliable survival and hazard rate estimates when at least one genome in the read set has coverage exceeding 20×, and the read error rate is below 6%. For lower coverage or higher error rates, a reference genome is required for an accurate estimate.

In addition to accurately estimating survival and hazard functions, skiver is also able to recover the underlying sequencing error spectrum used by the simulator. In this experiment, Badread was configured with a random error model, in which both the type of edit operation (substitution, insertion, or deletion) and the resulting base after each edit are selected uniformly at random. Consistent with this design, the error spectrum profiled by skiver (**Figure 4**) exhibits approximately equal frequencies across all single-base substitution (SBS) types, while the overall frequencies of substitutions, insertions, and deletions are each close to one third. These results closely match the parameters specified in the simulator.

In contrast, when BEST is applied to Minimap2 alignments generated from the same dataset, a higher proportion of substitutions is reported relative to insertions and deletions. This bias is likely attributable to the alignment scoring scheme of Minimap2, which favors mismatches over indels during alignment optimization. By comparison, skiver is an alignment-free method and therefore remains unaffected by alignment heuristics or scoring parameters, enabling a more faithful recovery of the true sequencing error spectra.

Next, we evaluated the effect of the outlier filter in datasets containing multiple closely related strains (**Figure 3.E**). We simulated reads with 64× coverage from *E. coli* K-12 MG1655 and *E. coli* O157:H7 at the same read identity. These two strains share an average nucleotide identity (ANI) of 98% according to skani [31]. The two read sets were then subsampled and mixed to generate datasets with a fixed total coverage of 64× but varying proportions of the two strains. For each mixture, we estimated the sequencing error rate using skiver,

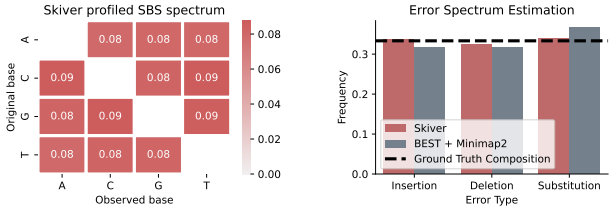

Fig. 4. Estimated sequencing error spectrum of a 128× coverage *E. coli* dataset with 96% read identity simulated using Badread, where all error types are equally likely.

defined as the first-base hazard $\hat{h}(1)$, both with and without the outlier filter enabled.

When the proportion of K-12 MG1655 is 1.0 or 0.0, the dataset effectively represents a single isolate. In these cases, skiver without filtering reports substantially elevated error rates, largely due to repetitive genomic regions that introduce spurious (k, v)-mer groupings. As the mixture proportion approaches 0.5, the estimated error rate from skiver without filtering increases sharply, reflecting the growing ambiguity between homologous regions shared by the two strains. In contrast, skiver with the outlier filter remains relatively robust across all mixture proportions and consistently reports error rates close to the ground truth. These results demonstrate that the outlier filter effectively suppresses confounding signals arising from repetitive and shared regions in strain genomes, enabling accurate error-rate estimation even in multi-strain datasets.

Skiver generalizes well to real reads on various sequencing platforms

We next evaluated the performance of skiver and several baseline methods on real sequencing datasets.

Across all datasets, skiver produces accurate hazard rate estimates (**Figure 5.A**). For Illumina reads, the estimated hazard rates are slightly biased upward, which is likely attributable to known positional biases in Illumina sequencing, where errors are more prevalent at the ends of the reads (**Supplementary Figure S2**). Notably, the estimated $\hat{h}(t)$ curves are nearly identical for skiver with and without a reference genome, indicating that selecting the most frequently observed value as the consensus is a reliable strategy.

For survival rate estimation, skiver consistently outperforms the baseline methods, achieving the lowest MSE_S and MSE_h across almost all datasets (**Figure 5.B; Table 1; Supplementary Tables S4, S6, and S7**). The only dataset for which skiver shows substantial deviation from ground truth is the Human, Nanopore R9.4 dataset, which is characterized by both high sequencing error rate (approximately 15%) and relatively low coverage (approximately 20×).

Among the baselines, Minimap2 + BEST tends to underestimate the survival rate $S(t)$ when sequencing errors cluster spatially along the reads. In such cases, the error rate inferred by counting mismatches and indels in the CIGAR string can exceed the true hazard rate, analogous to the example shown for read A in **Figure 2**. KMC + GenomeScope 2.0 is not designed for metagenomic samples and therefore struggles to identify an appropriate cutoff between erroneous and solid k-mers in the k-mer count histogram. Its performance is nevertheless substantially better on bacterial isolates and human sequencing data. Seqtk, on the other hand, may either

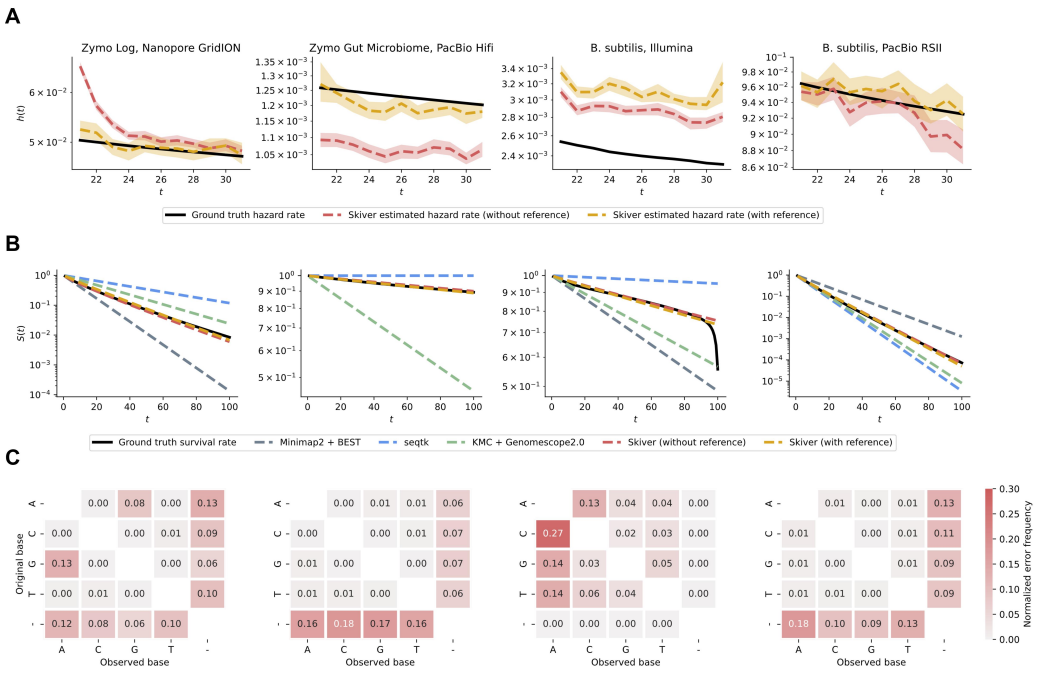

Fig. 5. Performance of skiver on real metagenomic datasets. **A**. Estimated hazard rate using skiver, with and without reference genomes. The shaded areas are the 90% confidence intervals for the hazard rate estimated via bootstrapping. **B**. Estimated survival rate using baselines and skiver. **C**. Estimated error spectra for the datasets. The last row represents insertions, and the last column represents deletions, while the rest of the entries are substitutions.

Table 1. MSE_S for benchmarked tools on real sequencing datasets. The tool with the lowest MSE_S is marked in bold.

Dataset	Minimap2 + BEST	seqtk	KMC + GenomeScope2.0	skiver
Zymo Log, Nanopore GridION	6.29×10^{-3}	6.38×10^{-2}	1.05×10^{-2}	$\mathbf{2.95 \times 10^{-5}}$
Zymo Gut Microbiome, PacBio HiFi	$\mathbf{1.67 \times 10^{-5}}$	4.36×10^{-3}	7.95×10^{-2}	4.51×10^{-5}
B. subtilis, Illumina	2.78×10^{-2}	1.83×10^{-2}	1.24×10^{-2}	$\mathbf{6.76 \times 10^{-4}}$
B. subtilis, PacBio RSII	6.13×10^{-3}	8.53×10^{-4}	3.61×10^{-4}	$\mathbf{8.67 \times 10^{-5}}$
Human, Nanopore R10.4	1.65×10^{-1}	6.69×10^{-2}	1.66×10^{-2}	$\mathbf{1.14 \times 10^{-3}}$
Human, Nanopore R9.4	1.04×10^{-2}	2.18×10^{-1}	$\mathbf{4.95 \times 10^{-3}}$	5.91×10^{-2}
Human, PacBio HiFi	$\mathbf{1.88 \times 10^{-4}}$	2.90×10^{-3}	2.60×10^{-4}	2.25×10^{-4}

under- or overestimate error rates, reflecting the fact that base quality scores in many read sets are not well calibrated.

In addition to estimating error rates, skiver enables profiling of sequencing error spectra (**Figure 5.C**). For nanopore datasets, substitutions are dominated by `G` → `A` and `A` → `G` transitions; PacBio datasets are characterized primarily by insertions and deletions; and Illumina datasets are dominated by substitution errors. These patterns are consistent with previously reported error profiles for the respective sequencing platforms [22, 8, 7]. The estimated spectra produced by skiver are also highly consistent regardless of whether a reference genome is provided (**Supplementary Figure S5**), further supporting the idea that the consensus derived from the most frequently observed value recovers the true reference base with high probability. On the other hand, the error spectra estimated using BEST in combination with Minimap2 differ noticeably from those obtained by skiver (**Supplementary Figure S6**). In particular, substitution rates are frequently inflated, which is again likely a consequence of the alignment scoring scheme employed by Minimap2, where mismatches are often favored over insertions and deletions.

Finally, we compared the memory usage and running times of all evaluated tools (**Figure 6; Supplementary Tables S7 and S8**). Although skiver currently operates in a single-threaded mode, it remains among the fastest and most memory-efficient methods evaluated. This efficiency highlights the suitability of skiver as a lightweight and scalable first step in bioinformatics pipelines, particularly for large-scale analysis of sequencing datasets.

Ablation Studies

Lastly, we assessed whether modeling the time to first sequencing error, T, using a discrete Weibull distribution is appropriate. To this end, we constructed a variant of skiver in which the shape parameter is fixed at $\hat{\beta} = 1$. Under this constraint, the survival function simplifies to $\hat{S}(t) = \exp(-\hat{\lambda}t)$, corresponding to a constant hazard rate $\hat{h}(t) = \hat{\lambda}$.

Experiments conducted on the same datasets (**Table 2; Supplementary Table S6; Supplementary Figure S8**) show that this constant-hazard model still outperforms all baseline methods in terms of both MSE_S and MSE_h, and provides a reasonable single-number estimate of the sequencing

Table 2. MSE_S across different datasets for variants of the skiver algorithm. "With reference": if the reference genome is given to the skiver algorithm; "constant hazard rate": if $\hat{\beta}$ is set to 1; "no filter": if the outlier filter is disabled during hazard rate estimation. The tool with the lowest MSE_S is marked in bold.

Dataset	Default skiver	with reference	constant hazard rate	no filter
Zymo Log, Nanopore GridION	$\mathbf{2.95 \times 10^{-5}}$	7.79×10^{-4}	1.13×10^{-3}	8.95×10^{-5}
Zymo Gut Microbiome, PacBio HiFi	4.51×10^{-5}	$\mathbf{3.46 \times 10^{-6}}$	4.74×10^{-5}	2.16×10^{-3}
B. subtilis, Illumina	6.76×10^{-4}	$\mathbf{5.49 \times 10^{-4}}$	6.73×10^{-4}	5.72×10^{-4}
B. subtilis, PacBio RSII	$\mathbf{8.67 \times 10^{-5}}$	1.13×10^{-4}	7.29×10^{-4}	1.17×10^{-3}
Human, Nanopore R10.4	$\mathbf{1.14 \times 10^{-3}}$	2.47×10^{-3}	2.40×10^{-3}	3.12×10^{-2}
Human, Nanopore R9.4	5.91×10^{-2}	7.68×10^{-2}	7.32×10^{-2}	$\mathbf{3.09 \times 10^{-2}}$
Human, PacBio HiFi	2.25×10^{-4}	$\mathbf{2.11 \times 10^{-6}}$	1.49×10^{-4}	5.87×10^{-2}

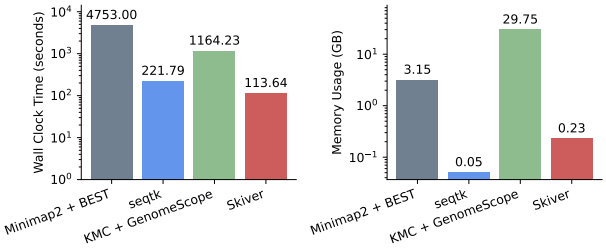

Fig. 6. The wall clock time (left) and peak memory usage (right) of the profiling tools on the *Zymo Gut Micrbiome, PacBio HiFi* dataset.

error rate. However, across all datasets, both MSE_S and MSE_h are consistently similar to or worse than those obtained using the full skiver model. The performance gap is particularly pronounced for Nanopore datasets, where sequencing errors are known to be non-uniform and tend to cluster along reads. These results indicate that the discrete Weibull distribution more closely reflects real sequencing error processes than a constant-hazard assumption.

We further performed the same ablation experiments using skiver with the outlier filter disabled. In this setting, the estimated survival rates are substantially lower (**Supplementary Figure S8**), particularly for the Zymo Gut Microbiome mock community, where multiple *E. coli* strains coexist, and for human datasets, which are diploid and contain significant amounts of repetitive regions. The improved agreement between the default skiver estimates and the ground truth in these challenging datasets further demonstrates that the proposed outlier filter is robust to genomic heterogeneity, enabling accurate survival and hazard rate estimation even in the presence of multiple strains, alleles, or repetitive sequences.

Discussion and Conclusion

In this study, we present skiver, a reference-free and alignment-free framework for profiling sequencing errors by using (k, v)-mer sketches. By modeling the time to first error as a discrete Weibull distribution, skiver provides direct estimates of survival and hazard rates, enabling a more expressive characterization of sequencing error processes than conventional single-number error metrics.

Across both simulated and real datasets, skiver consistently produces accurate survival and hazard rate estimates under a wide range of sequencing conditions. Skiver further enables reliable profiling of sequencing error spectra. The inferred substitution, insertion, and deletion patterns across Illumina, PacBio, and Nanopore datasets closely match previously reported platform-specific error characteristics. The proposed outlier filter substantially improves robustness in datasets containing multiple strains, alleles, or repetitive regions, enabling accurate estimation of error rates of reads coming from complex microbial communities and diploid genomes.

Finally, skiver is computationally efficient, achieving low runtime and memory usage, making it suitable as a lightweight preprocessing step in sequencing analysis pipelines.

There remains room for further improvement in the skiver pipeline. In particular, performance on datasets with high error rates or low sequencing coverage could potentially be enhanced through adaptive selection of the parameters k and v. The results of skiver can also be further applied in existing quality control pipelines, aiding quality score calibration and sequencing bias detection.

Beyond applications to sequencing reads, the skiver framework may also be extended to collections of genomes. In this setting, the hazard rate can be interpreted as the probability that the next genomic position differs due to mutation. Under this formulation, skiver could enable efficient estimation of mutational spectra across large genome collections, providing a scalable approach for large-scale mutational spectra analysis.

References

1. Paul D Allison. Discrete-time methods for the analysis of event histories. *Sociological methodology*, 13:61–98, 1982.
2. Jörn Bethune, April Kleppe, and Søren Besenbacher. A method to build extended sequence context models of point mutations and indels. *Nature Communications*, 13(1):7884, 2022.
3. Antonio Blanca, Robert S Harris, David Koslicki, and Paul Medvedev. The statistics of k-mers from a sequence undergoing a simple mutation process without spurious matches. *Journal of Computational Biology*, 29(2):155–168, 2022.
4. Kaitlin Chaung, Tavor Z Baharav, George Henderson, Ivan N Zheludev, Peter L Wang, and Julia Salzman. Splash: A statistical, reference-free genomic algorithm unifies biological discovery. *Cell*, 186(25):5440–5456, 2023.
5. Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, 34(17):i884–i890, 2018.
6. Eric M Davis, Yu Sun, Yanling Liu, Pandurang Kolekar, Ying Shao, Karol Szlachta, Heather L Mulder, Dongren Ren, Stephen V Rice, Zhaoming Wang, et al. Sequencerr: measuring and suppressing sequencer errors in next-generation sequencing data. *Genome Biology*, 22(1):37, 2021.

7. Clara Delahaye and Jacques Nicolas. Sequencing dna with nanopores: Troubles and biases. *PloS one*, 16(10):e0257521, 2021.
8. Juliane C Dohm, Philipp Peters, Nancy Stralis-Pavese, and Heinz Himmelbauer. Benchmarking of long-read correction methods. *NAR Genomics and Bioinformatics*, 2(2):lqaa037, 2020.
9. Paul Greenfield, Konsta Duesing, Alexie Papanicolaou, and Denis C Bauer. Blue: correcting sequencing errors using consensus and context. *Bioinformatics*, 30(19):2723–2732, 2014.
10. Mahmudur Rahman Hera, Paul Medvedev, David Koslicki, and Antonio Blanca. Estimation of substitution and indel rates via k-mer statistics. *bioRxiv*, pages 2025–05, 2025.
11. Luiz Irber, Phillip T Brooks, Taylor Reiter, N Tessa Pierce-Ward, Mahmudur Rahman Hera, David Koslicki, and C Titus Brown. Lightweight compositional analysis of metagenomes with fracminhash and minimum metagenome covers. *BioRxiv*, pages 2022–01, 2022.
12. Lauris Kaplinski, Märt Möls, Tarmo Puurand, and Maido Remm. Docestfast and accurate estimator of human ngs sequencing depth and error rate. *Bioinformatics Advances*, 3(1):vbad084, 2023.
13. Marek Kokot, Roozbeh Dehghannasiri, Tavor Baharav, Julia Salzman, and Sebastian Deorowicz. Scalable and unsupervised discovery from raw sequencing reads using splash2. *Nature Biotechnology*, 43(7):1084–1090, 2025.
14. Marek Kokot, Maciej Długosz, and Sebastian Deorowicz. Kmc 3: counting and manipulating k-mer statistics. *Bioinformatics*, 33(17):2759–2761, 2017.
15. Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018.
16. Heng Li. seqtk: Toolkit for processing sequences in FASTA/Q formats. `https://github.com/lh3/seqtk`, 2025. GitHub repository.
17. Peizhou Liao, Glen A Satten, and Yi-Juan Hu. Phredem: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genetic epidemiology*, 41(5):375–387, 2017.
18. Daniel Liu, Anastasiya Belyaeva, Kishwar Shafin, Pi-Chuan Chang, Andrew Carroll, and Daniel E Cook. Best: A tool for characterizing sequencing errors. *bioRxiv*, pages 2022–12, 2022.
19. Yuansheng Liu, Yichen Li, Enlian Chen, Jialu Xu, Wenhai Zhang, Xiangxiang Zeng, and Xiao Luo. Repeat and haplotype aware error correction in nanopore sequencing reads with dechat. *Communications Biology*, 7(1):1678, 2024.
20. Alexa BR McIntyre, Noah Alexander, Kirill Grigorev, Daniela Bezdan, Heike Sichtig, Charles Y Chiu, and Christopher E Mason. Single-molecule sequencing detection of n 6-methyladenine in microbial reference materials. *Nature communications*, 10(1):579, 2019.
21. Páll Melsted and Bjarni V Halldórsson. Kmerstream: streaming algorithms for k-mer abundance estimation. *Bioinformatics*, 30(24):3541–3547, 2014.
22. André E Minoche, Juliane C Dohm, and Heinz Himmelbauer. Evaluation of genomic high-throughput sequencing data generated on illumina hiseq and genome analyzer systems. *Genome biology*, 12(11):R112, 2011.
23. Samuel M Nicholls, Joshua C Quick, Shuiquan Tang, and Nicholas J Loman. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience*, 8(5):giz043, 2019.
24. Donovan H Parks, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. Gtdb: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research*, 50(D1):D785–D794, 2022.
25. T Rhyker Ranallo-Benavidez, Kamil S Jaron, and Michael C Schatz. Genomescope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. *Nature communications*, 11(1):1432, 2020.
26. Michael G Ross, Carsten Russ, Maura Costello, Andrew Hollinger, Niall J Lennon, Ryan Hegarty, Chad Nusbaum, and David B Jaffe. Characterizing and measuring bias in sequence data. *Genome biology*, 14(5):R51, 2013.
27. Kristoffer Sahlin and Paul Medvedev. Error correction enables use of oxford nanopore technology for reference-free transcriptome analysis. *Nature communications*, 12(1):2, 2021.
28. Melanie Schirmer, Umer Z Ijaz, Rosalinda D'Amore, Neil Hall, William T Sloan, and Christopher Quince. Insight into biases and sequencing errors for amplicon sequencing with the illumina miseq platform. *Nucleic acids research*, 43(6):e37–e37, 2015.
29. Kishwar Shafin, Trevor Pesout, Ryan Lorig-Roach, Marina Haukness, Hugh E Olsen, Colleen Bosworth, Joel Armstrong, Kristof Tigyi, Nicholas Maurer, Sergey Koren, et al. Efficient de novo assembly of eleven human genomes using promethion sequencing and a novel nanopore toolkit. *BioRxiv*, page 715722, 2019.
30. Jim Shaw, Jean-Sebastien Gounot, Hanrong Chen, Niranjan Nagarajan, and Yun William Yu. Floria: fast and accurate strain haplotyping in metagenomes. *Bioinformatics*, 40(Supplement_1):i30–i38, 2024.
31. Jim Shaw and Yun William Yu. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nature Methods*, 20(11):1661–1665, 2023.
32. Krithika Suresh, Cameron Severn, and Debashis Ghosh. Survival prediction models: an introduction to discrete-time modeling. *BMC medical research methodology*, 22(1):207, 2022.
33. James W Vaupel, Kenneth G Manton, and Eric Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.
34. Xin Victoria Wang, Natalie Blades, Jie Ding, Razvan Sultana, and Giovanni Parmigiani. Estimation of sequencing error rates in short reads. *BMC bioinformatics*, 13(1):185, 2012.
35. Ryan R Wick. Badread: simulation of error-prone long reads. *Journal of Open Source Software*, 4(36):1316, 2019.
36. Andreas Wilm, Pauline Poh Kim Aw, Denis Bertrand, Grace Hui Ting Yeo, Swee Hoe Ong, Chang Hua Wong, Chiea Chuen Khor, Rosemary Petric, Martin Lloyd Hibberd, and Niranjan Nagarajan. Lofreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic acids research*, 40(22):11189–11201, 2012.
37. Haonan Wu, Antonio Blanca, and Paul Medvedev. A k-mer-based estimator of the substitution rate between repetitive sequences. *bioRxiv*, pages 2025–06, 2025.
38. Y William Yu, Deniz Yorukoglu, Jian Peng, and Bonnie Berger. Quality score compression improves genotyping accuracy. *Nature biotechnology*, 33(3):240–243, 2015.