ReSkmer: Modeling Repeats Improves Low-Coverage, Alignment-Free Estimates of Genomic Distance

Eduardo Charvel [1] Isaac Thomas [2] Homère J. Alves Monteiro [4]
Shahab Sarmashghi [5] Glenn Dunshea [6] Vineet Bafna [2] Siavash Mirarab [3]

[1]UC San Diego Bioinformatics and Systems Biology Graduate Program [2]UC San Diego Computer Science and Engineering Department [3]UC San Diego Electrical and Computer Engineering Department
[4]Center for Evolutionary Hologenomics, GLOBE Institute [5]Broad Institute of MIT and Harvard [6]Department of Natural History, Norwegian University of Science and Technology

Reference-Free, Alignment-Free Distance Estimation

- Genomic distance (d) is a direct measure of genetic diversity (θ).
- **Mash** [1] generates an unbiased MinHash sketch of a sequence, uses it to compute the Jaccard index, $J = \frac{|A \cup B|}{|A \cap B|}$, between k-mer sets efficiently, then uses J to solve for distance.
- **Skmer** [2] computes distances between genome skims, using the k-mer frequency histogram of each sample ($j \in \{1, 2, \cdots, n\}$) to estimate:
  - $\lambda^{(j)}$ (k-mer coverage)
  - $\epsilon^{(j)}$ (sequencing error rate)
  - $L^{(j)}$ (genome length)

  Then computes the probability that k-mers are sampled without error.

Repeats Change the *k*-mer MinHash Intersection

- Skmer assumes that genomes are not repetitive.
- r_i = number of k-mers that appear i times in a genome (i.e., its repeat spectrum, R).
- **Uniqueness Ratio** (UR = r_1/L) is the ratio of the number of unique k-mers to the length of the genome.
- UR varies across the tree of life, ranging from $1.0 - 0.28$ [3].
- Repeated k-mers affect the MinHash intersection (I) (Figure 1).

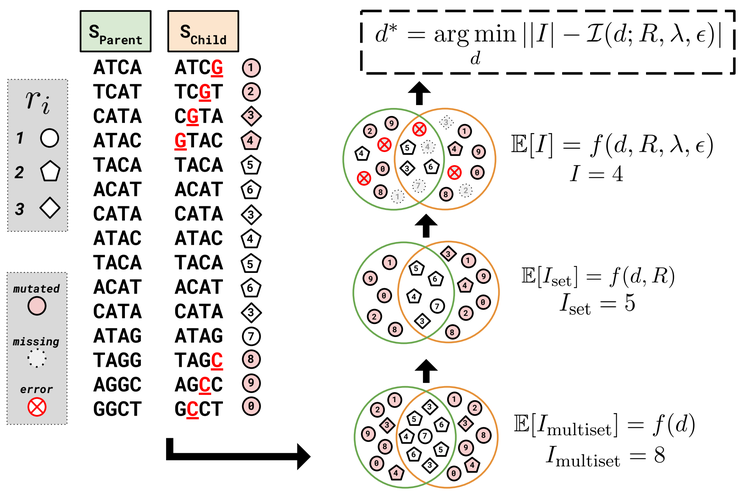

Figure 1. **Repeats change the intersections between *k*-mer sets.**

- k-mers occur at different multiplicities (r_i) in a genome.
- If k-mers are uniquely represented, then I only depends on d **(bottom)**.
- In repetitive genomes, k-mers will occur at different multiplicities (r_i), because MinHash sets I ignores multiplicites, $I_{\text{set}} \neq I_{\text{multiset}}$ **(middle)**.
- The intersection of two genome skims is changed by k-mer coverage (λ) and error rate (ϵ) **(top)**.
- To get d, we model: $\mathbb{E}[|I|] = \mathcal{I}(d; R, \lambda, \epsilon)$ using estimates of λ, ϵ, and R.

Ordinary Least Squares Estimates Error-Free Coverage

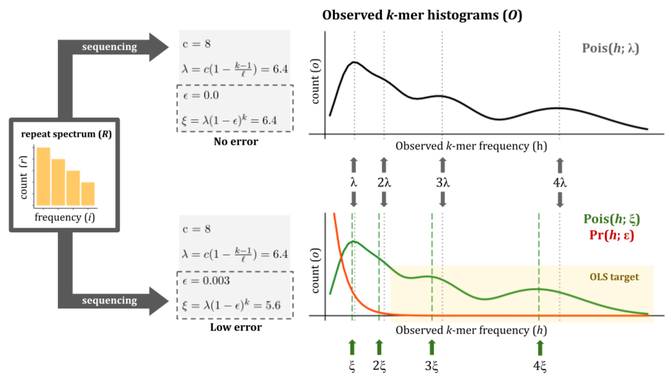

Figure 2. **Observed *k*-mer histograms (O)** are a convolution of the genome repeat spectrum (R) and the sequencing process. If there is no error **(top)**, O's distribution is parametrized by λ. If $\epsilon \neq 0$ **(bottom)**, then O is a sum of two distributions: 1) counts of error-free k-mers parameterized by ξ and, 2) a distribution of erroneous k-mers. For larger k-mer frequency bins (h), we expect no errors. Thus, for a selected range $H_1 \leq h \leq H_2$, we estimate ξ using OLS. In both cases, $\lambda = \sum_h (o_h)/L$.

Repeat-Aware Distance Equation

With a repeat-aware model of intersection, $\mathcal{I}(d; R^{(1)} \cdots \eta^{(2)})$, we solve for d by minimizing the difference between the observed $|I|$ and our model.

$$d^* = \arg\min_d \left| |I| - \mathcal{I}(d; R^{(1)}, \lambda^{(1)}, \epsilon^{(1)}, \eta^{(1)}, \lambda^{(2)}, \epsilon^{(2)}, \eta^{(2)}) \right| .$$

We model two events ignored by Skmer:

1. **More repetitive k-mers are more likely to be in the intersection.**
   Let $\eta^{(j)}$ be the probability of a k-mer being sampled without error:

$$\mathbb{E}[|I_0|] = \sum_{i=1}^{m} r_i^{(1)} \left(1 - \left(1 - \eta^{(1)}\right)^i\right) \left(1 - \left(1 - \eta^{(2)}(1-d)^k\right)^i\right)$$

2. **Erroneous k-mers artificially inflate the intersection due to error.**
   - Consider two lists containing randomly picked 1-neighbors of u, with replacement.
   - Moreover, we have $r_i^{(1)}$ k-mers repeated i times in $G^{(1)}$, and thus,

$$\mathbb{E}[|I_1|] = \sum_{i=1}^{m} r_i^{(1)} 3k \left(1 - \mathbb{E}[e^{-\frac{M}{3k}}]^i\right) \left(1 - \mathbb{E}[e^{-\frac{M}{3k}}]^i\right) .$$

Finally, We then model how neighbor lists are generated through mutation and/or errors *(not shown here)* and add both intersections:

$$\mathcal{I}\left(d; R^{(1)}, \lambda^{(1)}, \epsilon^{(1)}, \eta^{(1)}, \lambda^{(2)}, \epsilon^{(2)}, \eta^{(2)}\right) = \mathbb{E}[|I_0|] + \mathbb{E}[|I_1|]$$

Simulated and Biological Data Results

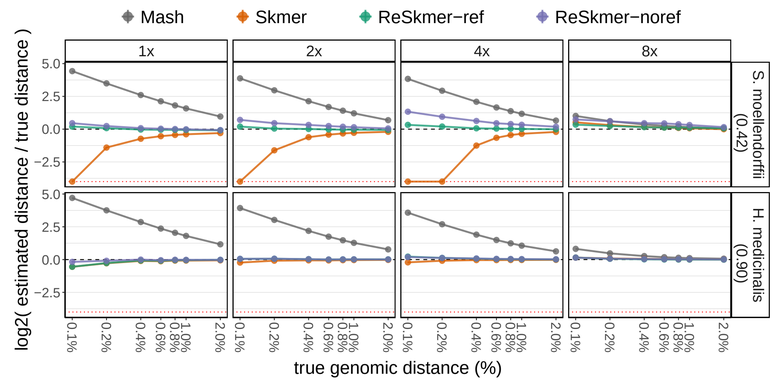

Figure 3. **Simulated Distances Results.** We simulated population-level distances by introducing SNPs to genomes with varying URs (shown on the right). Here is a comparison of alignment-free distance methods. ReSkmer-ref refers to ReSkmer with R obtained from an assembly, while ReSkmer-noref refers to an R estimated with Respect [3].

	Apis mellifera				*Drosophila melanogaster*			
	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)	Mash	Skmer	ReSkmer (ref)	ReSkmer (noref)
1×	0.12	0.61	0.60	**0.65**	0.12	0.58	**0.88**	0.76
2×	0.40	0.69	0.62	**0.71**	0.12	0.61	**0.90**	0.86
4×	0.63	**0.77**	0.58	0.71	0.40	0.69	0.88	**0.90**
6×	0.67	**0.72**	0.57	0.53	0.31	0.40	0.83	**0.90**

Table 1. **Biological Benchmarking.** We test Pearson correlation of θ estimates between alignment-free methods and alignment-based methods (ANGSD for *Apis* and GATK for *Drosophila*). Highest values for each coverage are bolded.

ReSkmer Conclusions

- Modeling the impact of repeats can increase the of alignment-free k-mer-based methods.
- In both simulated and biological data, ReSkmer was able to improve upon repeat-agnostic methods.
- Biodiversity monitoring may decrease cost by eliminating the need for a reference genome and by sequencing at coverage as low as 1×.

References

[1] Brian D andet al Ondov. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, 17(1):132, December 2016.

[2] Shahab Sarmashghi and et al. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology*, 20(1):34, December 2019.

[3] Shahab Sarmashghi and et al. Estimating repeat spectra and genome length from low-coverage genome skims with RESPECT. *PLOS Computational Biology*, 17(11), November 2021.

edcharvel@ucsd.edu