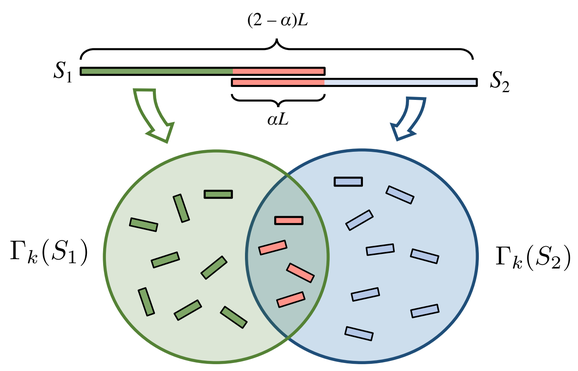
DAMN fast: DNA Alignment using Multi-Armed Bandits
Spectral Jaccard Similarity for long-read alignment

Tavor Baharav[1], Govinda Kamath[1], Ilan Shomorony[2], David Tse[1]
[1]Stanford University and [2]University of Illinois, Urbana-Champaign

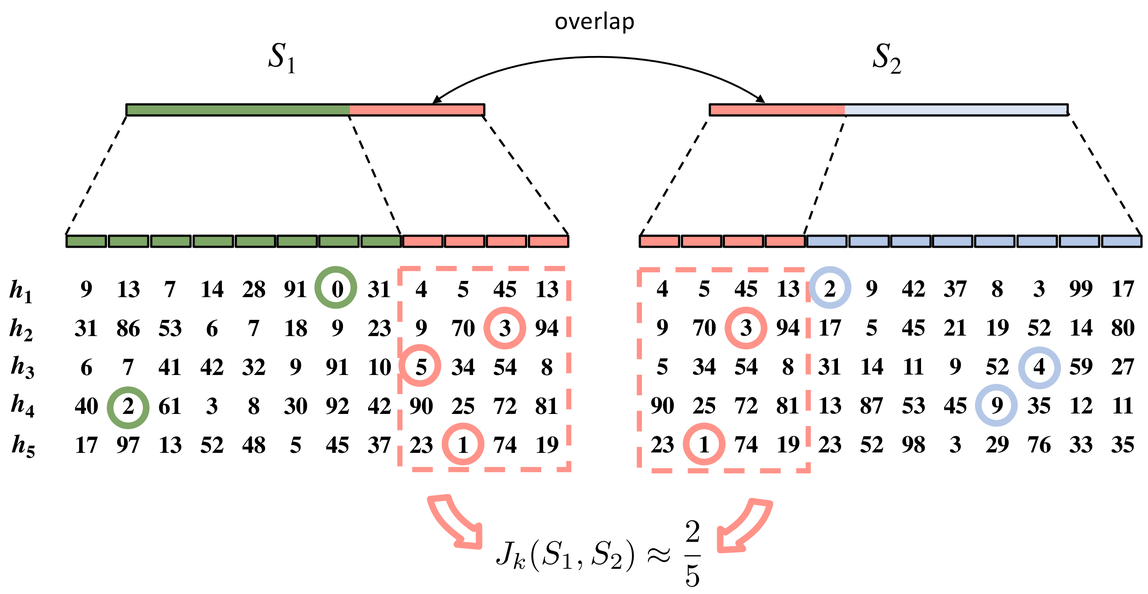
Motivation: efficient long-read pairwise alignment

- For long-read sequencing technologies, pairwise alignment is a computational bottleneck [1]
- Identifying good candidate pairs prior to full alignment is desirable
- k-mer Jaccard Similarity is a common proxy for alignment size

$$J_k(S_1,S_2) = \frac{|\Gamma_k(S_1)\cap\Gamma_k(S_2)|}{|\Gamma_k(S_1)\cup\Gamma_k(S_2)|} \approx \frac{\alpha L}{(2-\alpha)L} = \frac{\alpha}{2-\alpha}$$

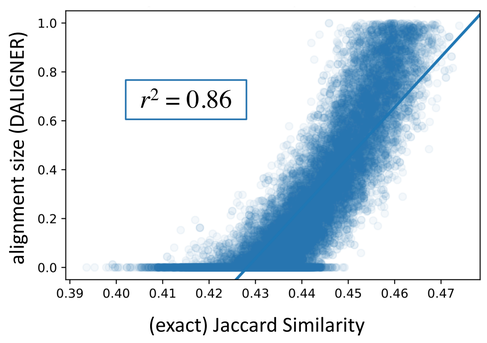


- *Min-hashes* provide a fast, probabilistic method to estimate the Jaccard similarity [2,3]

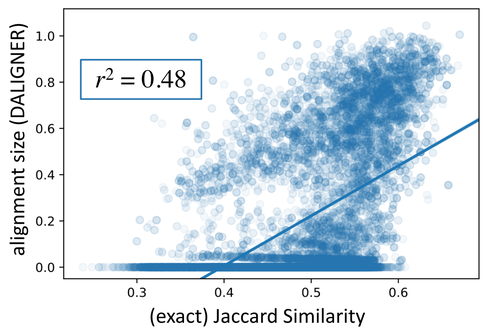


- How good is Jaccard similarity at estimating the alignment length?

Simulated reads from *E. coli*



PacBio reads from *E. coli*



Model: "Some hashes are better than others"

- For a "reference" read S_1, we define a ***min-hash collision matrix*** A:

	h_1	h_2	h_3	h_4	h_5	$\widehat{J_k}(S_1,S_i)$
S_2	0	1	0	0	1	.4
S_3	0	0	0	1	0	.2
S_4	1	0	0	0	0	.2
S_5	0	0	1	0	0	.2
S_6	0	0	0	0	0	0
S_7	0	1	0	0	0	.2

standard estimate of $J_k(S_1,S_i)$ assumes same "reliability" for all hashes

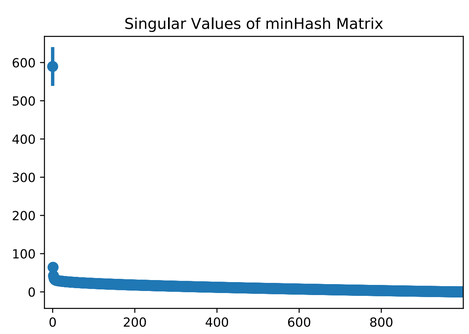
- To allow for different hashes to have different impacts, we assume that

$$A_{ij} \sim \mathrm{Ber}(p_i) \text{ OR } \mathrm{Ber}(q_j)$$

- We then notice that

$$E[A] - \mathbf{1}\mathbf{1}^T = (1-p)(q-1)^T$$

Singular Values of minHash Matrix

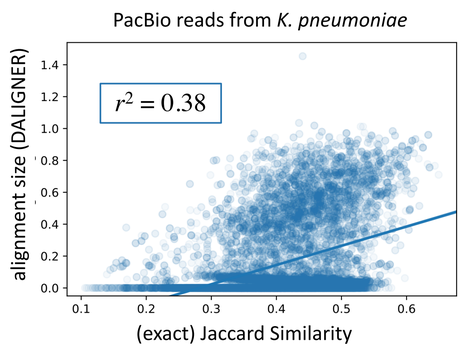


- This allows us to estimate p and q using SVD
- We refer to the resulting p_i's as the *Spectral Jaccard Similarity*

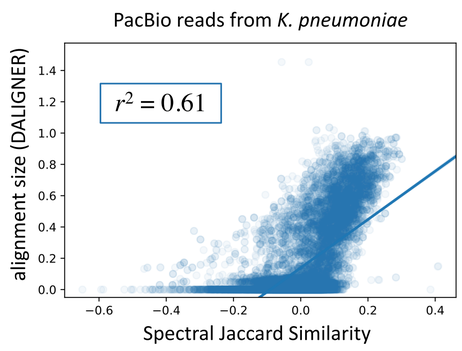
Interpreting the Model

- Spectral Jaccard similarity provides a better estimate for alignment size than standard Jaccard similarity

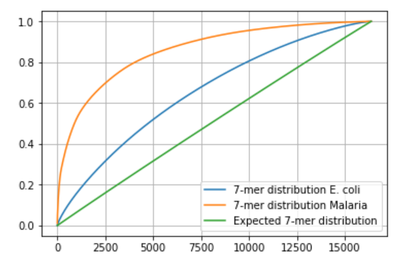
PacBio reads from *K. pneumoniae*



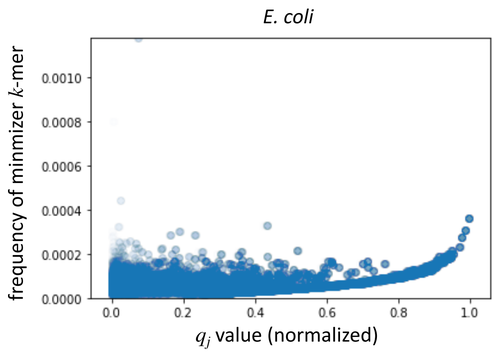
PacBio reads from *K. pneumoniae*



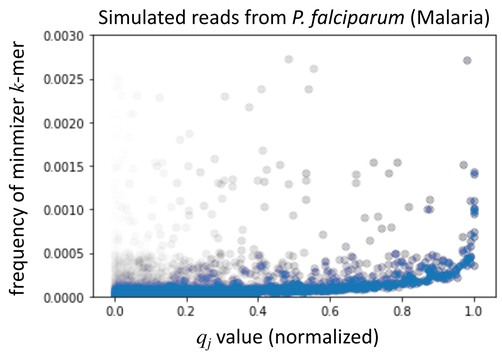
- Hash functions h_j that assign a low value to a common k-mer provide a less reliable estimate of the alignment size
- The q_j parameters capture the unreliability of h_j
- For genomes with highly uneven k-mer distributions (and more common k-mers), q_j's play a more important role



E. coli



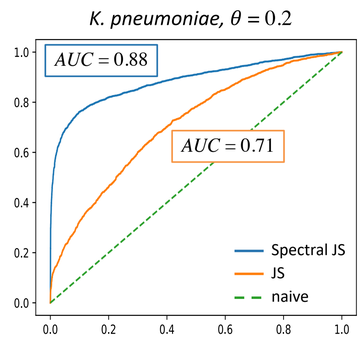
Simulated reads from *P. falciparum* (Malaria)



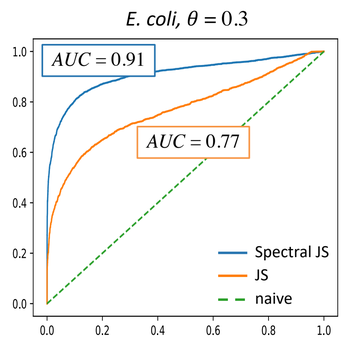
More accurate pre-alignment filter

- Task: identify pairs of reads with alignment larger than a threshold θ
- Spectral Jaccard Similarity performs better than Jaccard Similarity

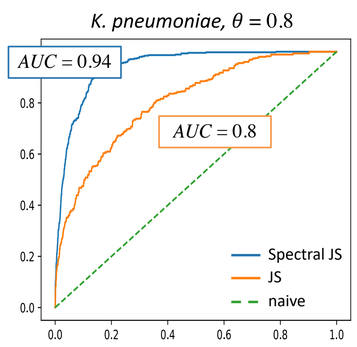
K. pneumoniae, $\theta = 0.2$



E. coli, $\theta = 0.3$



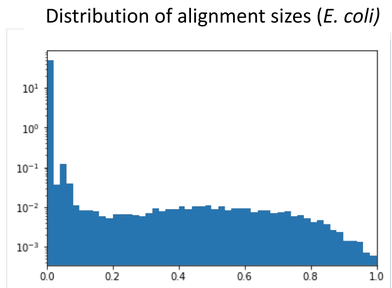
K. pneumoniae, $\theta = 0.8$



Efficient Computation via Multi-Armed Bandits

- While SVD is slow, only a small fraction of pairs of reads have nontrivial alignment (1% for *E. coli)*
- We *adaptively* estimate Spectral Jaccard Similarity using Multi-Armed Bandits [4]
- For pairs of reads with small alignments, only a small number of min-hashes are computed and used to estimate the p_i's

Distribution of alignment sizes (*E. coli)*



[1] Myers, Gene. "Efficient local alignment discovery amongst noisy long reads." 2014.
[2] Heng Li, Minimap2: pairwise alignment for nucleotide sequences, 2018.
[3] Berlin, Konstantin, et al. "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing." 2015
[4] Bagaria, Vivek, Govinda M. Kamath, and David N. Tse. "Adaptive Monte-Carlo Optimization." 2018