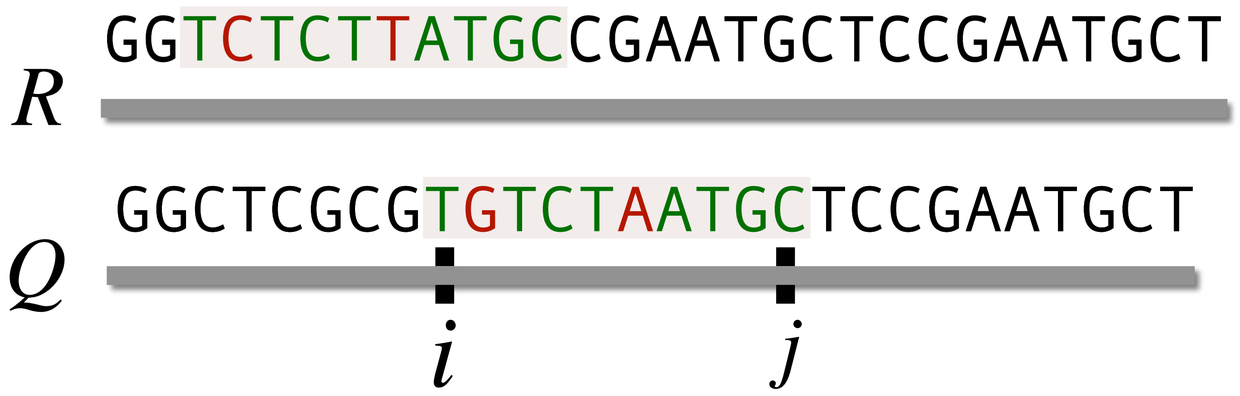
A genome-wide likelihood framework for distance-based pattern matching

Ali Osman Berk Şapcı & Siavash Mirarab
UC San Diego

RECOMB-Seq 2026

My promise for this talk

Given a query genome and a reference, and a distance Δ



2/10 mismatches:

$\Delta = 0.25$ ✓

$\Delta = 0.1$ ✗

P1: For any given interval $Q_{i:j}$, **decide (yes/no)** if R has a *(set of) homologous* counterpart(s) w/ distance $< \Delta$

constant time

P2: Given R, **enumerate** all such maximal intervals of Q

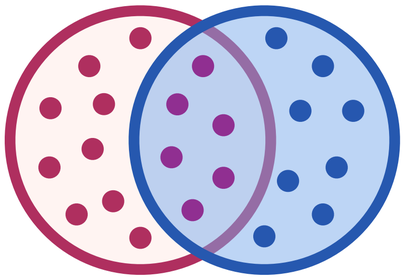
linear time

?!? Quadratic number of intervals ?!?

Genomes are evolutionarily heterogeneous

Genome-wide distances & nucleotide identity:

- comparing genomes and MAGs



Jaccard index of *k*-mers sets

$$D = -\frac{1}{k}\ln\frac{2J}{J+1}$$

Many flavors:
Mash, skani, Skmer, FastANI & orthoANI…

ANI is just a summary statistic…

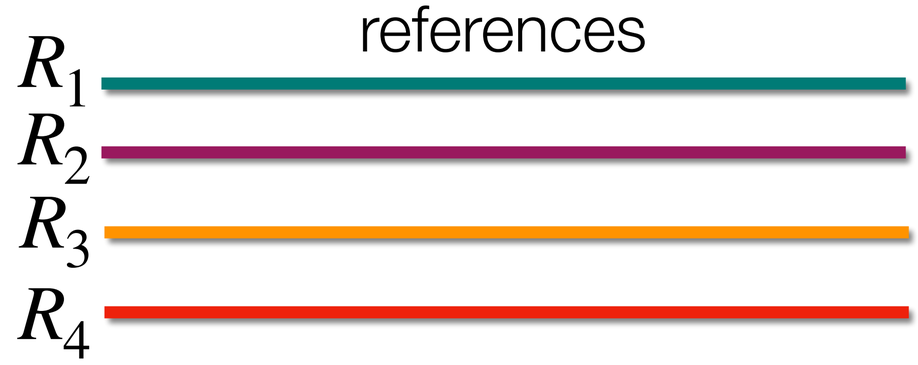
Causes for local deviations:

- Horizontally transferred genes
- Conserved elements
- Contamination in assemblies
- Viral integration

Neutral scenario: change in distances due to natural variation

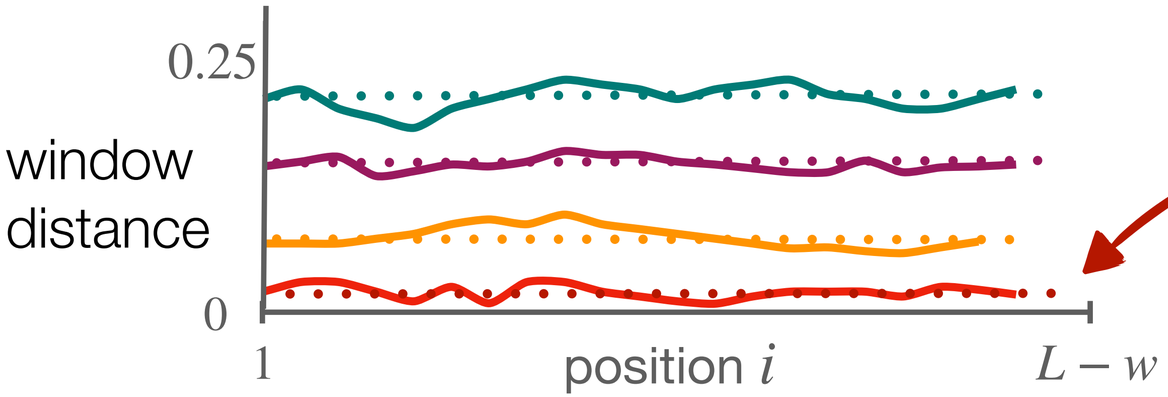
query genome Q

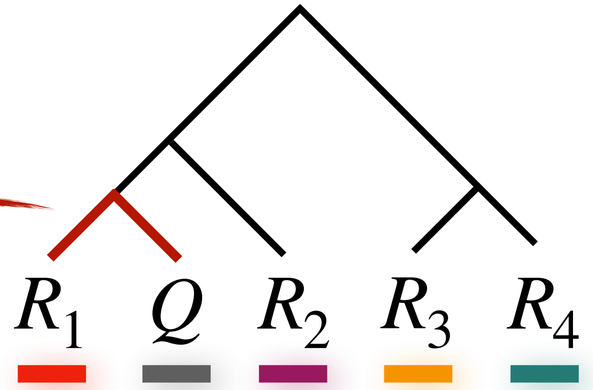
i $\quad$ $i + w$



sliding window

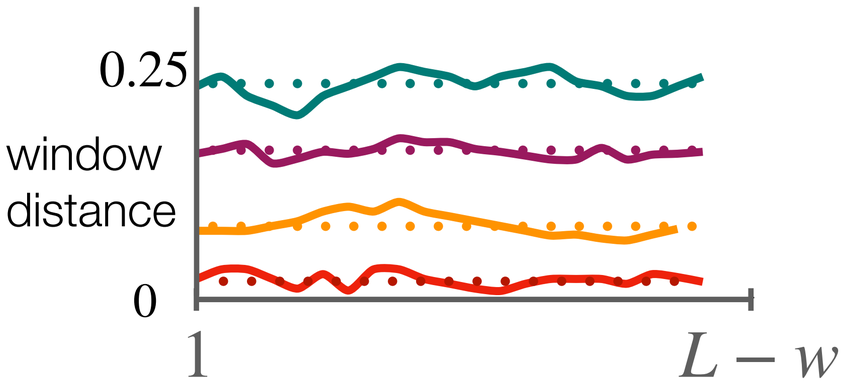
find the best matching window in R





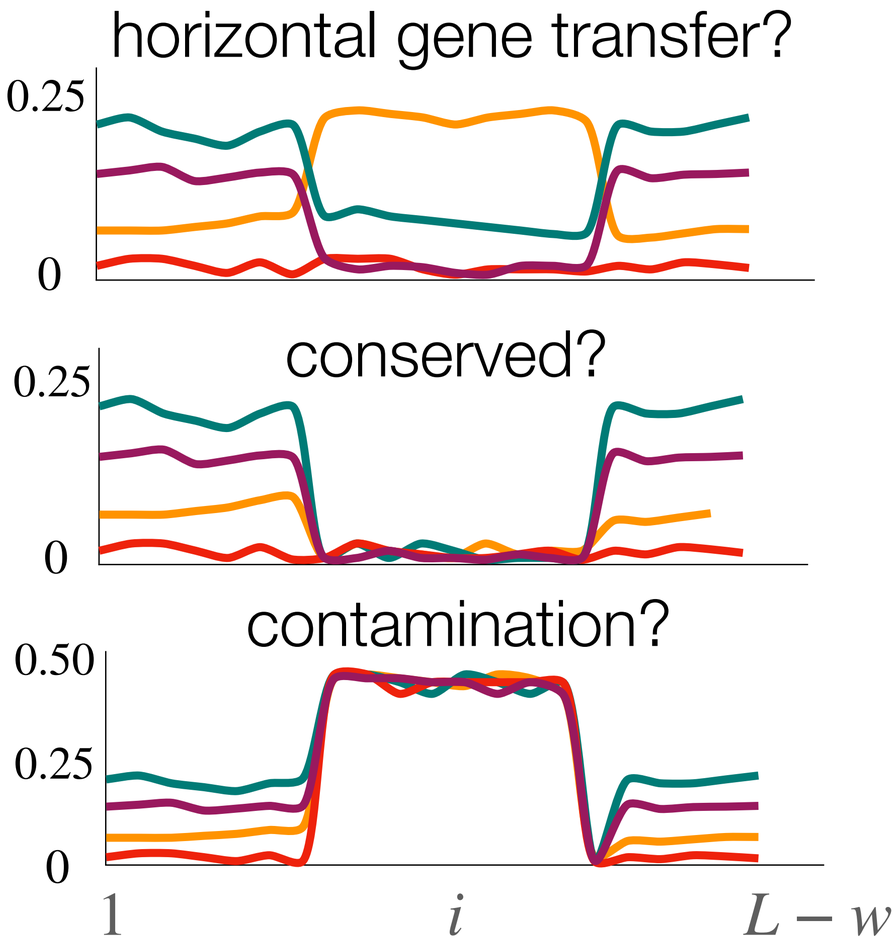
• • • • • • : genome-wide distance

Goal: measure local distances, and detect deviations and outlier regions



• • • : genome-wide distance

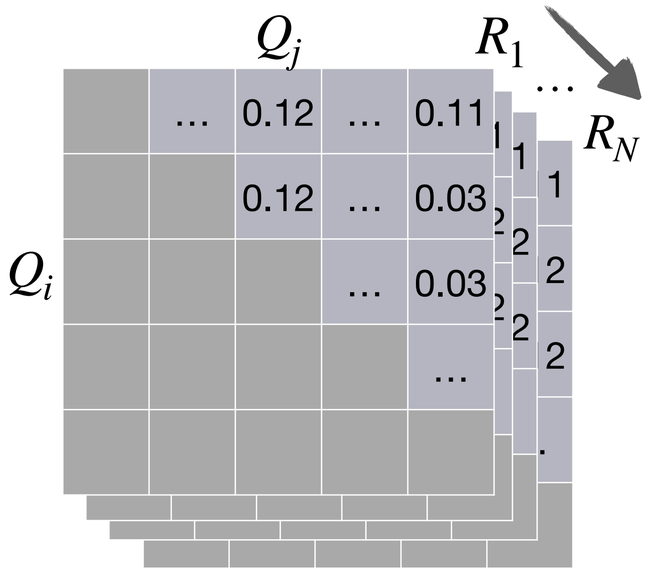
segments with deviation



Ideally, no fixed window size: distance of any interval

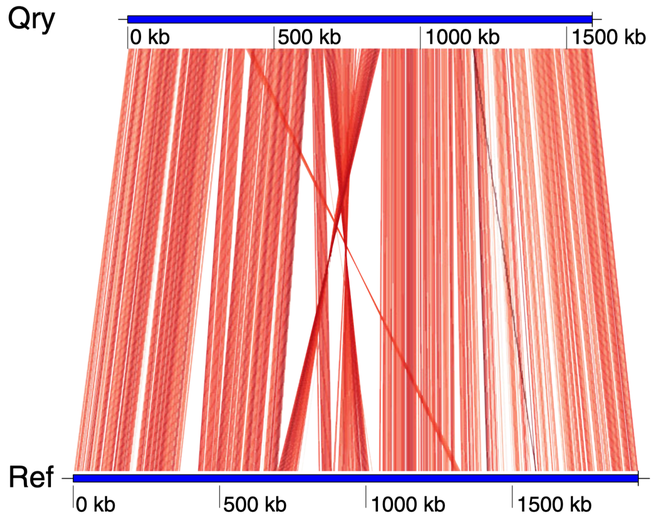
Naive approach: (any window)
(+) parameter free & more precise
(-) $O(|Q|^2)$ intervals, not feasible

Sliding window:
(-) parameterized (window/step size)
(+) $O(|Q|)$ intervals of fixed length



$(|Q| \times |Q| \times N)$

e.g., homology mapping w/ MashMap



FastANI [Jain et al. 2018]

Our focus is outlier regions, not synteny or orthology

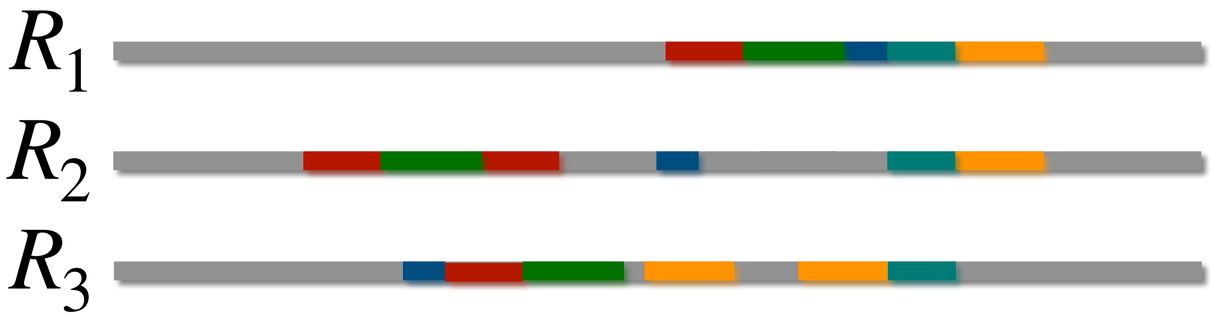
Inadequate for **highly divergent** sequences & **scalibility** is a challenge

We are **indifferent to the positions** of mapped intervals in the reference

Idea: we relax the orthology relations & allow non-contiguity

Suppose all give the same $\mathrm{d}(Q_{i:j}, R)$

set of homologus →
(synteny-free by design)

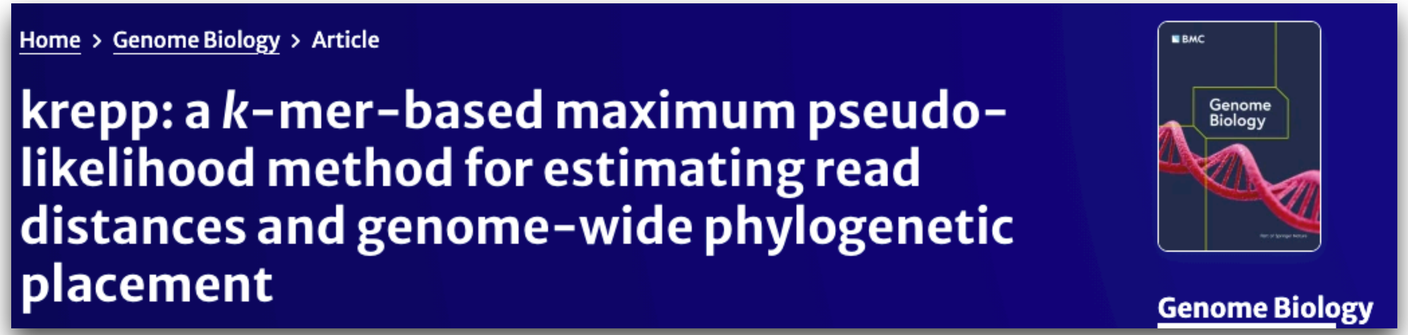


back to k-mer based distances

Computing the Jaccard or Containment index for all intervals is still not feasible

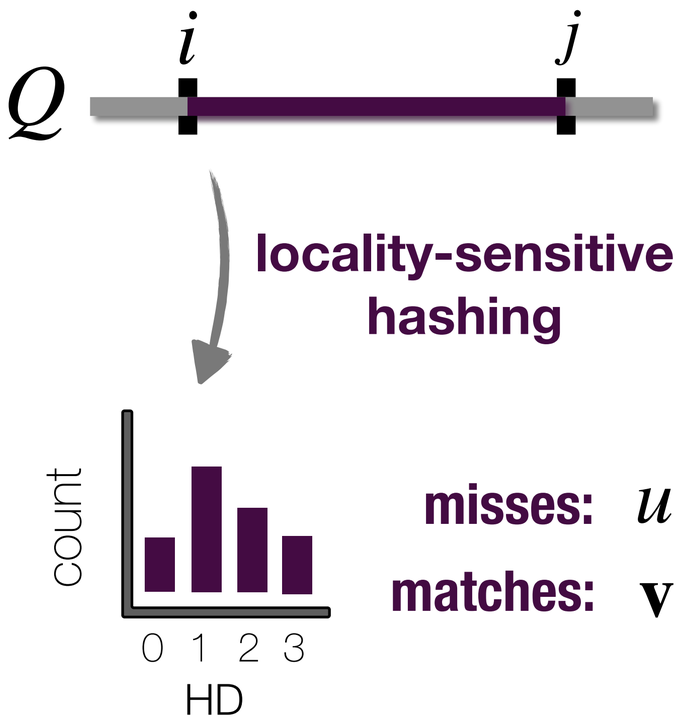
Good news: *krepp* (2025)

- Searching "homologous" k-mers and measuring Hamming distances
- Estimating distances from sequences to reference genomes using k-mers



Ali Osman Berk Şapcı & Siavash Mirarab

Maximum likelihood distances from “homologous” k-mers

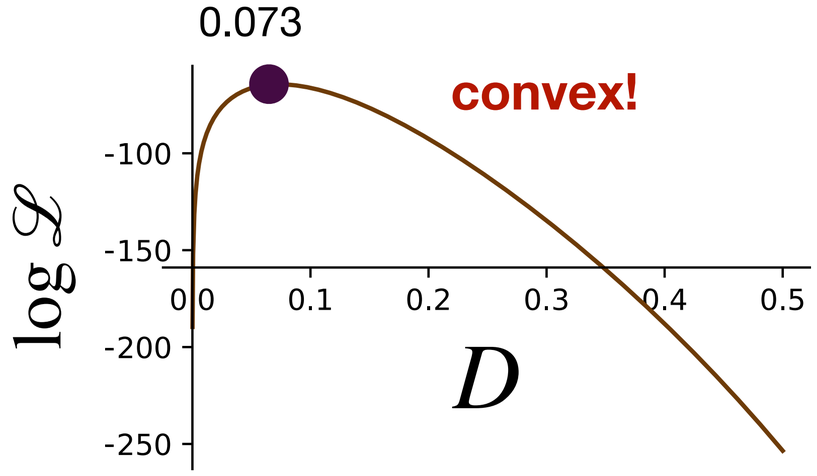


Likelihood of $Q_{i:j}$ having distance D to R:

$$\arg\max_D \; P_{miss}(D;k,h,\delta)^u \prod_{d=0}^{\delta} P_{match}(D;d,k,h)^{v_d}$$

Probability of having u missses in total

Probability of having v_d matches at HD $= d$



Highly accurate in benchmarks: even for short sequences & high distances!

gdiff: distance-based pattern detection using the derivatives

Optimizing the likelihood for all intervals is still not feasible

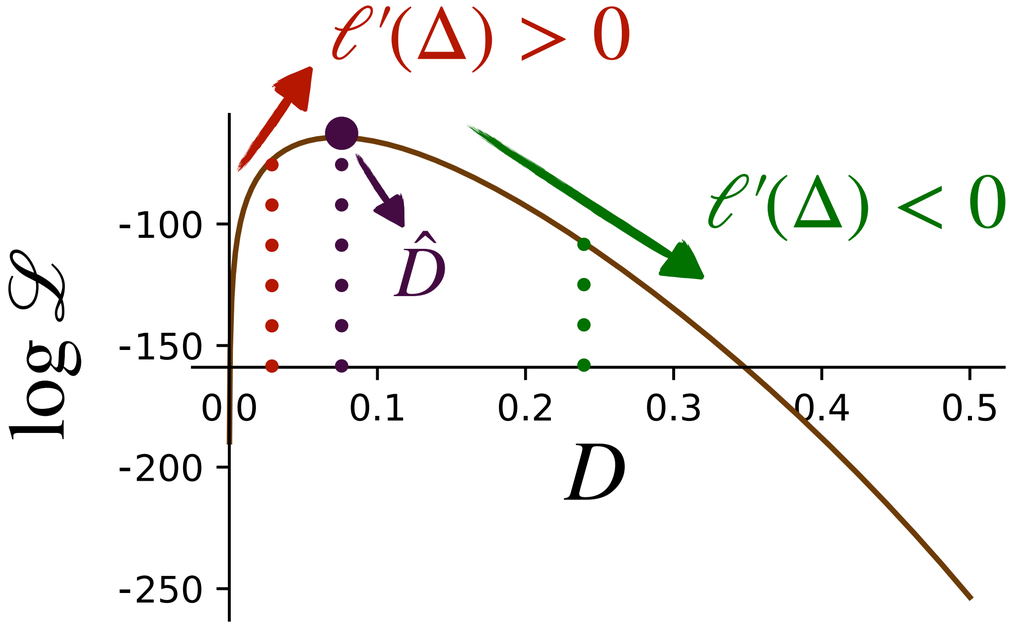
Decide whether an interval satisfies $\hat{\mathrm{d}}(Q_{i:j}, R) < \Delta$ by looking at the derivate at Δ

For MLE distance $\hat{D}$:

If $\hat{D} = \Delta \rightarrow \ell'(\Delta) = 0$

If $\hat{D} < \Delta \rightarrow \ell'(\Delta) < 0$

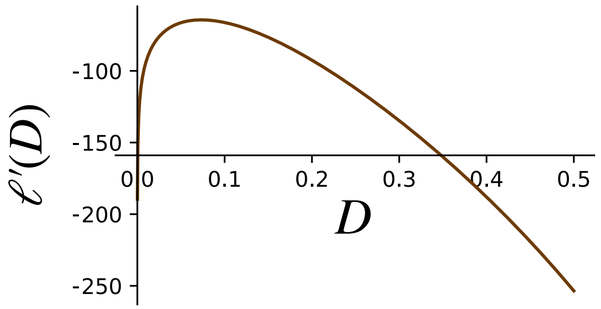
If $\hat{D} > \Delta \rightarrow \ell'(\Delta) > 0$



No optimization: just a linear combination over k-mers

misses: u

matches: $\mathbf{v}$



Each *k*-mer adds a constant to the derivate!

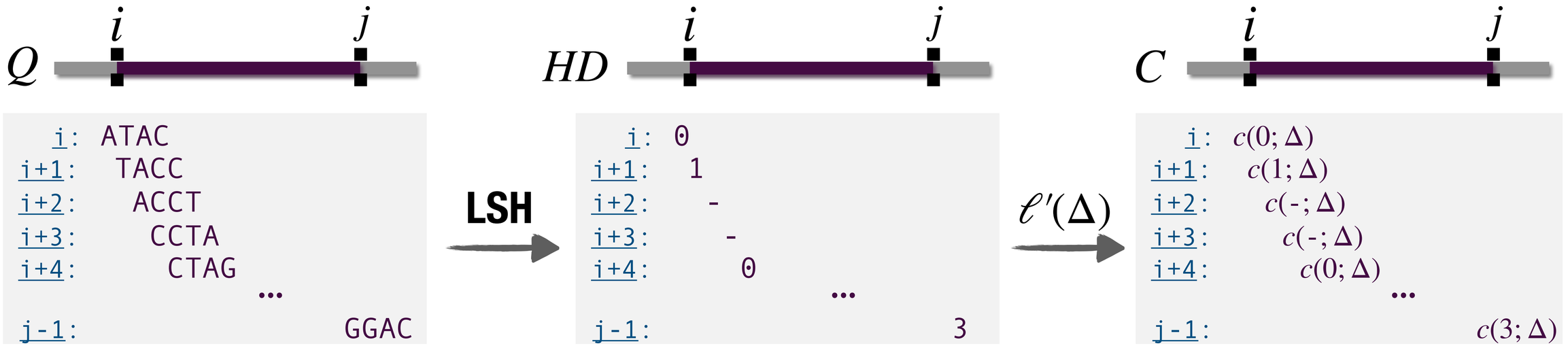
precomputed: $c(-,\Delta)$ $c(d;\Delta)$

$$\ell'(D) = u \frac{\rho\left(\sum_{d=0}^{k} D^d(1-D)^{k-d}\binom{k}{d}\left(\frac{d-kD}{D(1-D)}\right)P_\delta(d)\right)}{1-\rho+\rho\left(\sum_{d=0}^{k} D^d(1-D)^{k-d}\binom{k}{d}P_\delta(d)\right)} + \sum_{d=0}^{\delta} v_d \left(\frac{d-kD}{D(1-D)}\right)$$

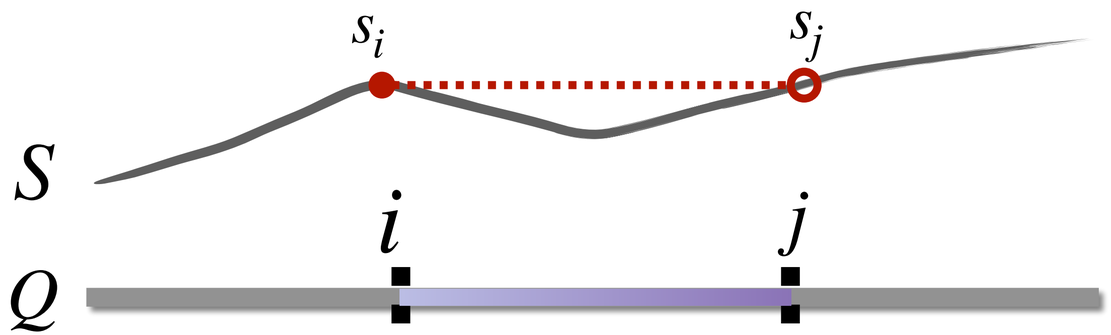
contribution to the derivate per missed k-mer

per matched k-mer

Solving the decision problem after a single pass



Compute the prefix-sum array $S = (s_1, \ldots, s_{N+1})$ where $s_i = \sum_{i'=1}^{i} C(x_{i'})$.



For any given interval $[i, j)$;
$\mathrm{d}(Q_{i:j}, R) < \Delta$ if $s_i - s_j > 0$!

(constant time)

What about enumerating all the intervals?

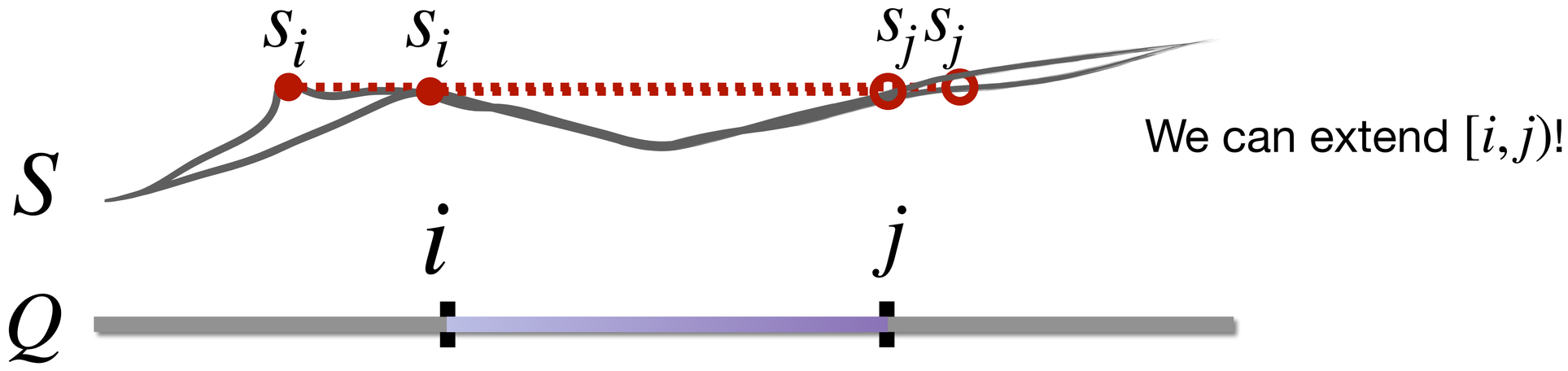
Find all maximal intervals $[i, j)$ with $\mathsf{d}(Q_{i:j}, R) < \Delta$ (or $> \Delta$)

Not contained in any other larger interval

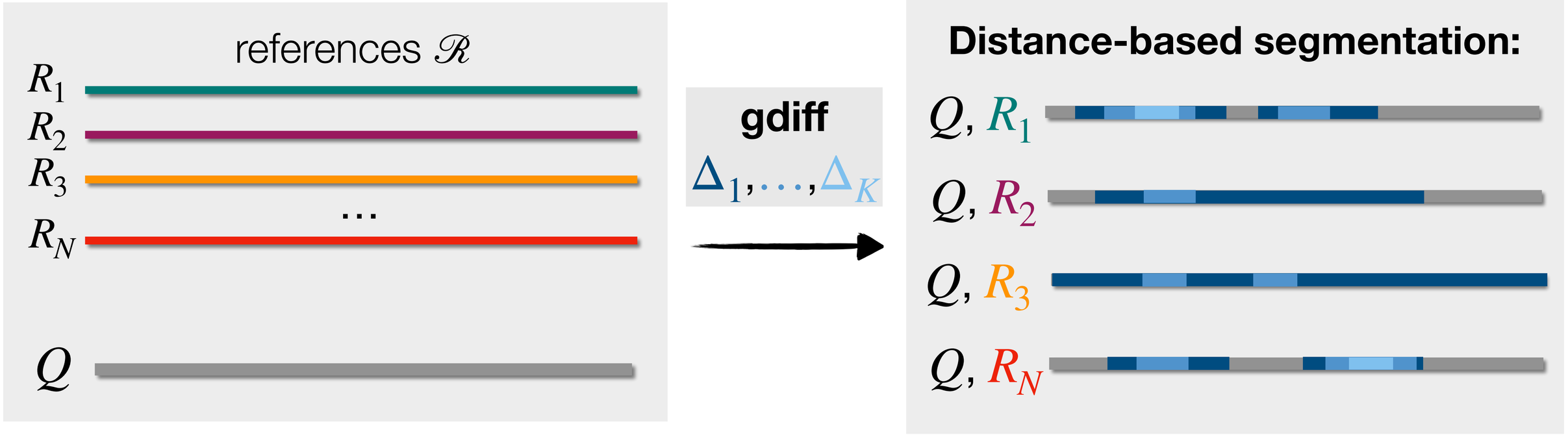
Linear time!

No longer quadratic due to maximality!

1. Left maximal if s_i is a prefix maxima of S
2. Right maximal if s_j is a suffix minima of S



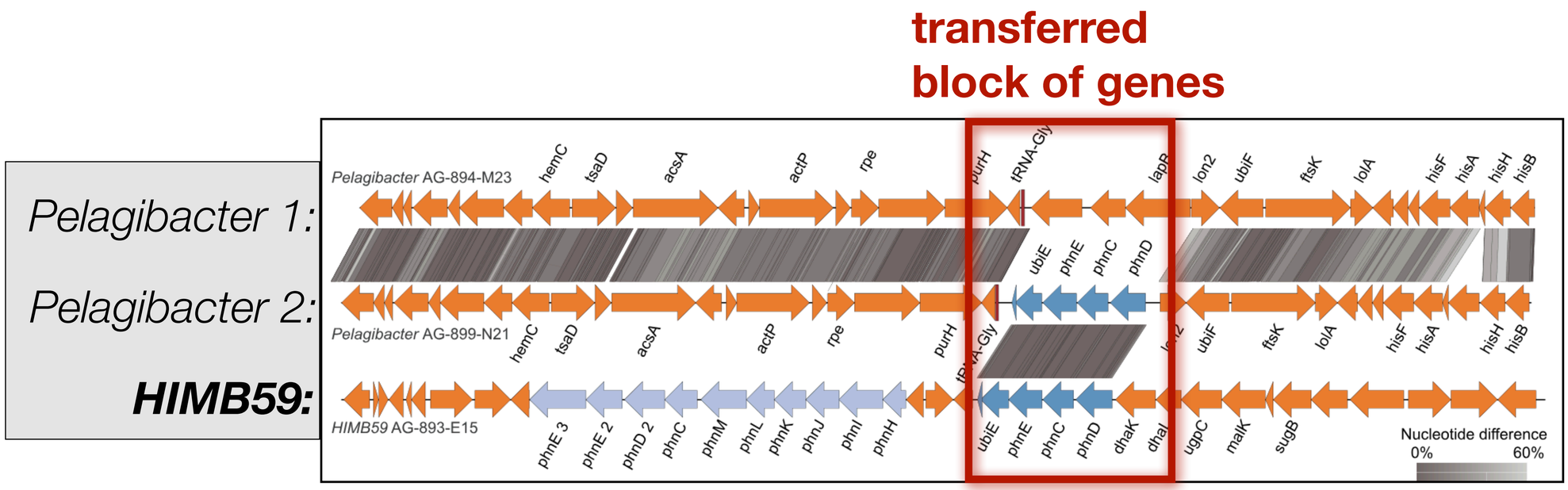
We can extend $[i, j)$!



~2Mbp genomes, $N = 1000$, $K = 8$ [vectorized], 16 threads → **~12 seconds**

gdiff detects a documented horizontal gene transfer

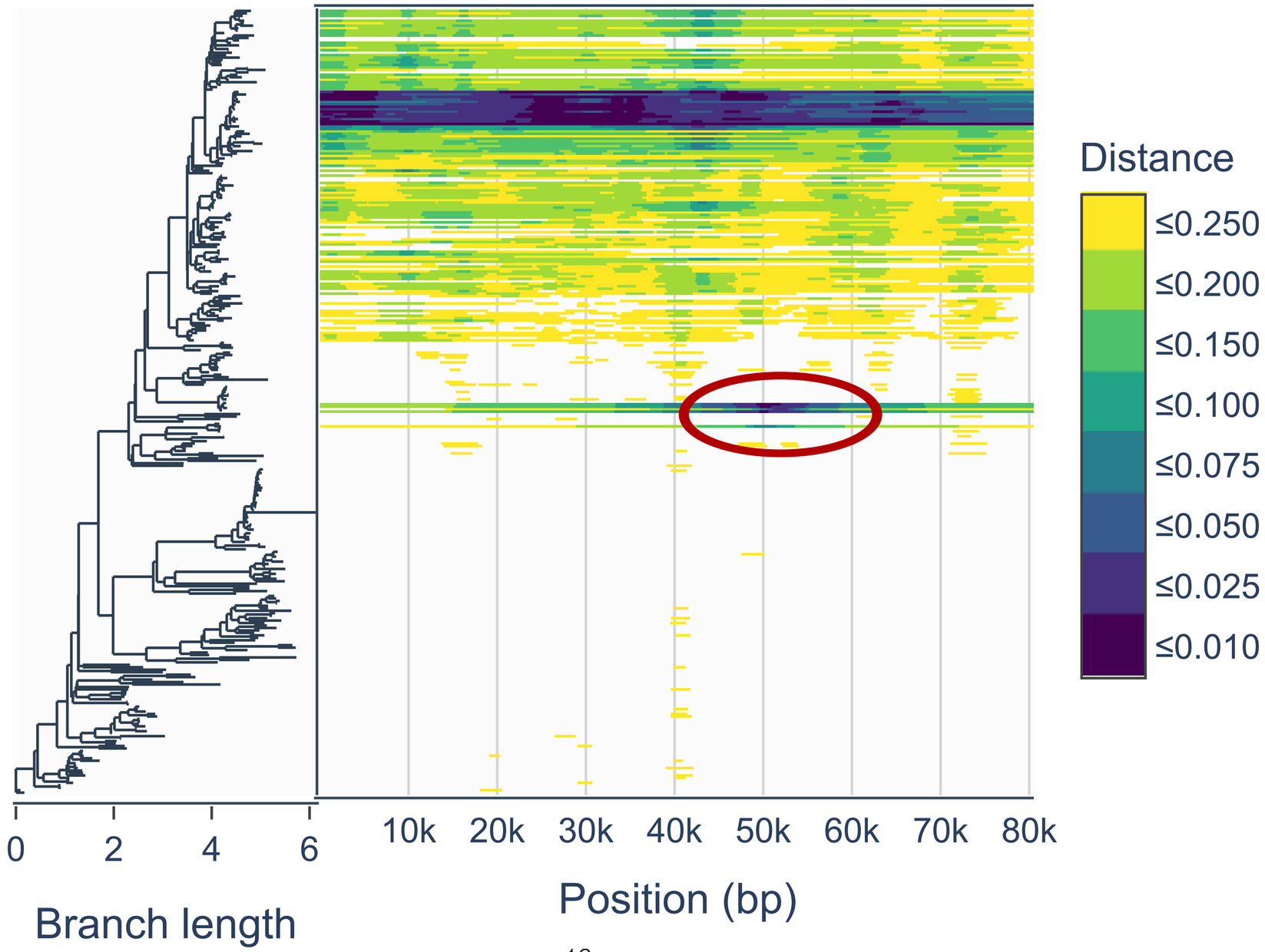
- A horizontally transferred gene btw. *Pelagibacter* & ***Proteobacterium HIMB59***
  → 1.6% regional distance vs. 29.4% genome-wide distance
- Comparing a single-cell assembled (SAG) ***HIMB59*** against 10,000 marine SAG refs.
  → scalable to tens of thousands of genomes!



[Stepanauskas et al. 2025]

A documented HGT event in ocean microbes

500 genomes selected from GORG-Tropical dataset (10,000)



Summary: gdiff

A framework for local distance estimation & genomic “outlier” detection

- detecting contaminations, conserved regions, viral integration, HGT

Future work:

- adding the phylogenetic aspect
- identifying the underlying process
- better benchmarking (e.g., eukaryotes)

Thank you!

Siavash Mirarab

Eduardo Charvel

Funding

RECOMB Travel Fellowship
NSF & ISCB

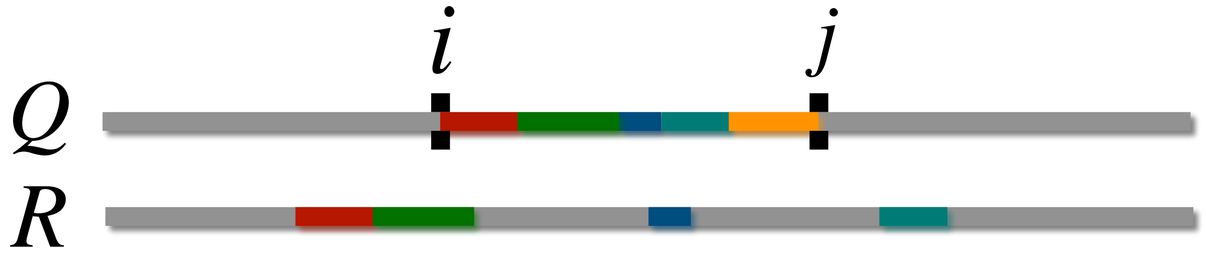
software: github.com/bo1929/gdiff

paper:

Work in progress

Extra Slides

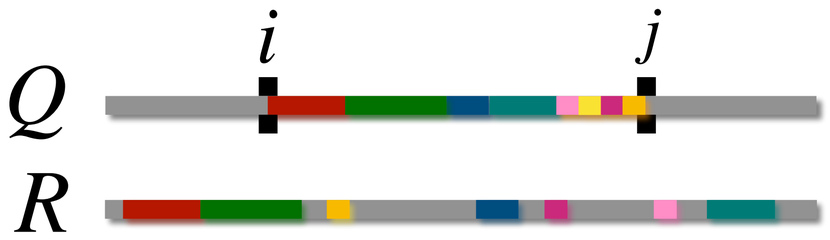
Thinking about it recursively (for a second)



$$\mathsf{d}(Q_{i:j}, R) = \sum_{q \in \{\blacksquare\}} \frac{\text{length}(q)}{j-i} \mathsf{d}(q, R)$$

$\mathsf{d}(\blacksquare, R) = \mathsf{d}(\blacksquare, \blacksquare)$: e.g., Hamming distance, alignment etc.

$\mathsf{d}(\blacksquare, R)$: Recurse? →

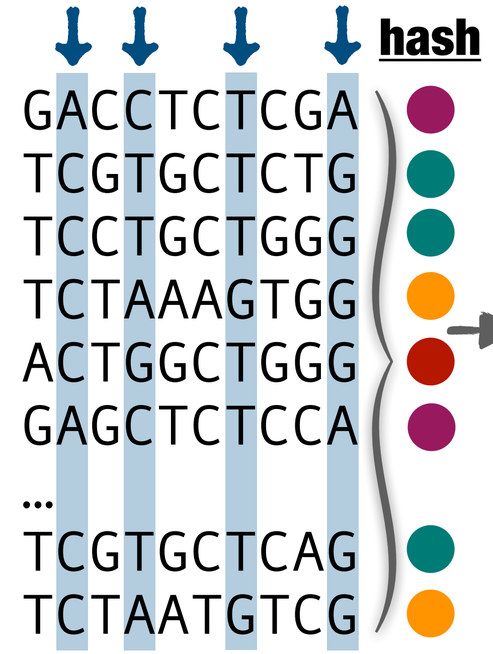


chop into smaller pieces untill smaller segment could be mapped

Approach "inductively" and we are back to k-mers (but for different reasons)…

Computing false negative rates of LSH

Select h random but fixed positions (default h: 14, k: 29)

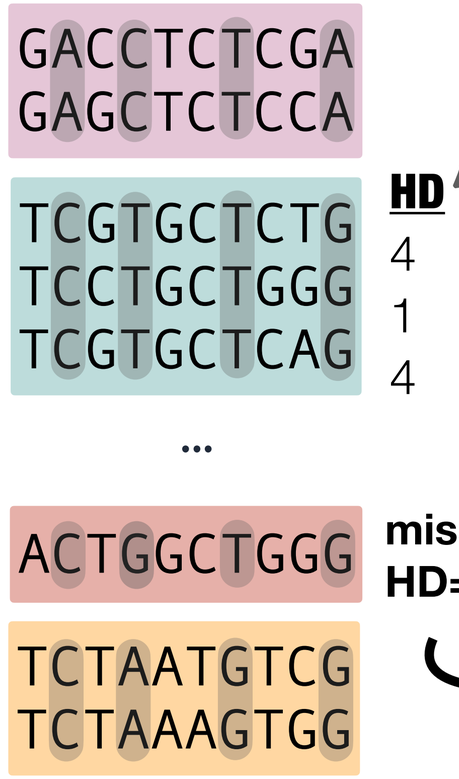


reference *k*-mer set

locality-sensitive hashing

Given a query *k*-mer

ACCTGCTGGG



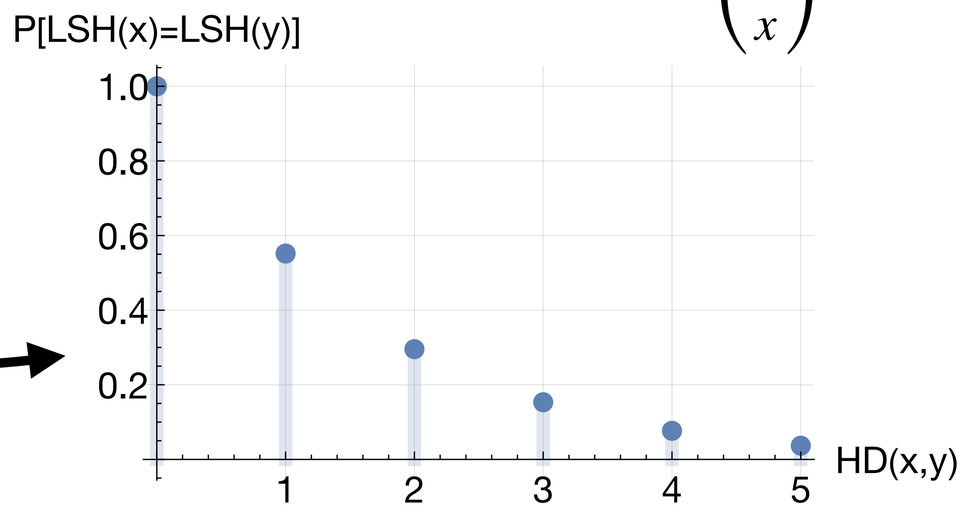
HD
4
1
4

miss at HD=2

4^h LSH buckets

False negative rate:

collides at HD=x with probability $\dfrac{\binom{k-h}{x}}{\binom{k}{x}}$



Observing k-mers matches with varying HDs

$$P_{match}(D; d, k, h) = \rho \; P_{mutate}(D; d, k) \; P_{collide}(d, k, h)$$

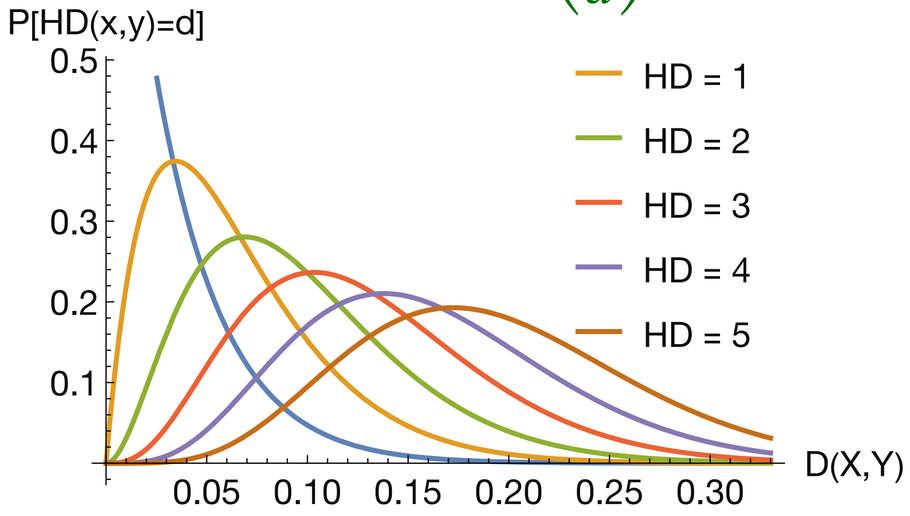
$$\rho = \frac{\text{\# of subsampled}}{\text{\# of distinct}}$$

precomputed for R

→ not all k-mers have to be indexed:

- minimizers
- FracMinHash
- …

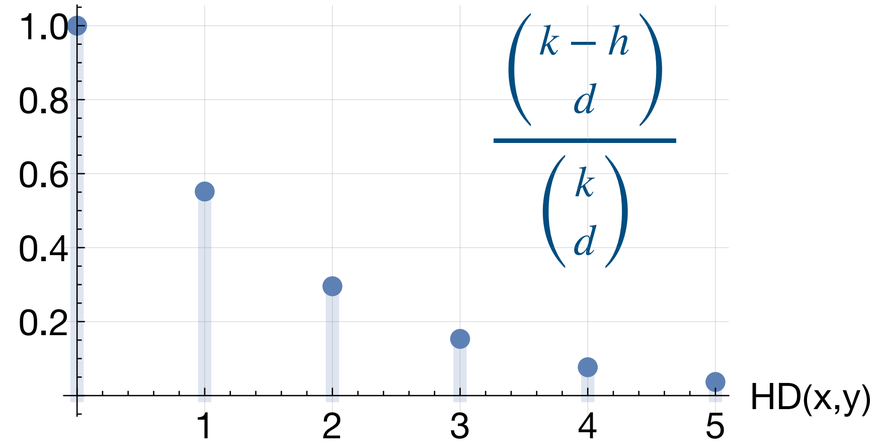
$$D^d(1-D)^{(k-d)}\binom{k}{d}$$



1-FNR of LSH for HD $= d$

FNR

$$\frac{\binom{k-h}{d}}{\binom{k}{d}}$$



Multiple events could lead to a mismatch

$$P_{miss}(D;d,k,h,\delta) = (1-\rho) + \rho\left[\sum_{d=\delta+1}^{k} P_{mutate}(D;d,k) + \sum_{d=0}^{\delta} P_{mutate}(D;d,k)(1-P_{collide}(d,k,h))\right]$$

A mismatch occurs for two *k*-mers (query a and reference b), if

- b is not indexed: $1-\rho$ ***or***
- b is indexed: ρ, **but either:**

i) $\mathrm{HD}(a,b) > \delta$: $\sum_{d=\delta+1}^{k} P_{mutate}(D;d,k)$ ***or***

ii) $\mathrm{HD}(a,b) \leq \delta$ ***and*** $\mathrm{LSH}(a) \neq \mathrm{LSH}(b)$: $\sum_{d=0}^{\delta} P_{mutate}(D;d,k)(1-P_{collide}(d,k,h))$

Can a distance for every interval be computed?

From $\mathcal{O}(L_R^2 L_Q^2)$ distances to $\mathcal{O}(L_Q^2)$ per reference!

For a large $\mathscr{R}$, $\mathcal{O}(L_Q^2 |\mathscr{R}|)$ distances is **still not feasible**

Simpler problems: given Q & R, thresholds Δ and τ

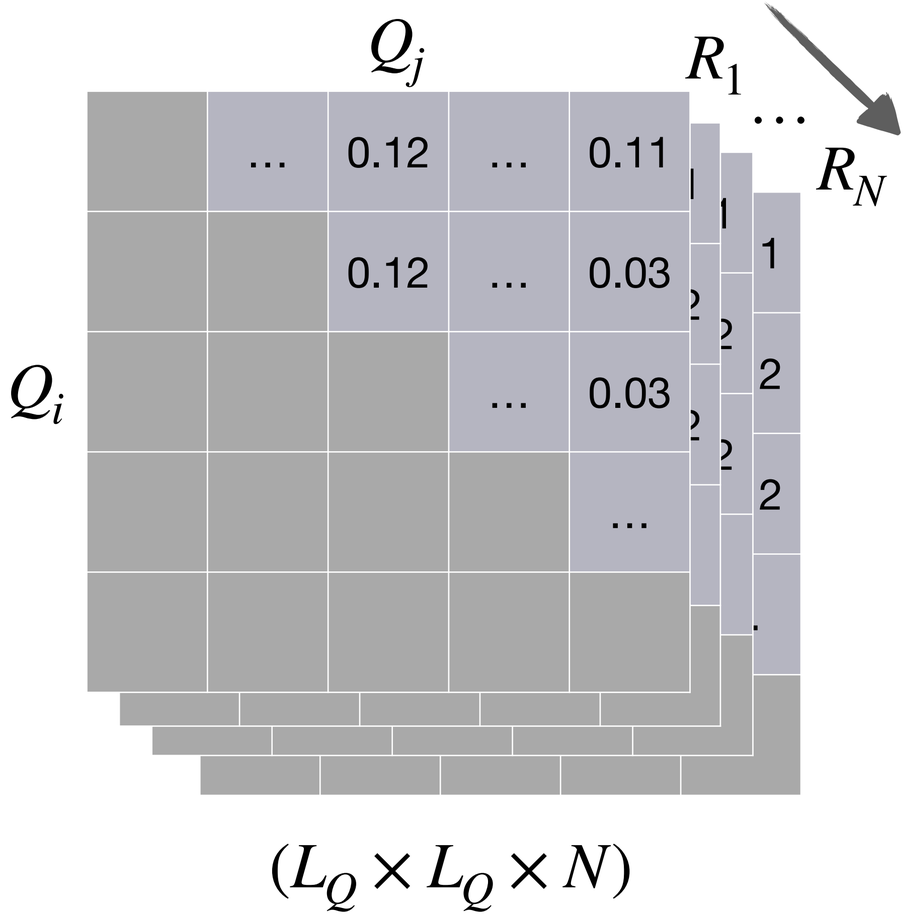
Decide whether an interval (i, j) satisfies $\mathsf{d}(Q_{i:j}, R) < \Delta$.

Enumerate all **maximal intervals** (i, j) such that

$$\mathsf{d}(Q_{i:j}, R) < \Delta \text{ and } j - i \geq \tau,$$

$$\mathsf{d}(Q_{a:b}, R) \geq \delta \text{ for } a \leq i \leq j \leq b, (i, j) \neq (a, b).$$

Query segment to reference genome:



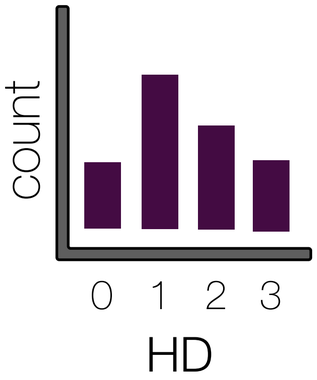
No optimization: just a linear combination over k-mers

Simply a linear combination over the histogram!

Compute the derivative at the threshold $D = \Delta$.

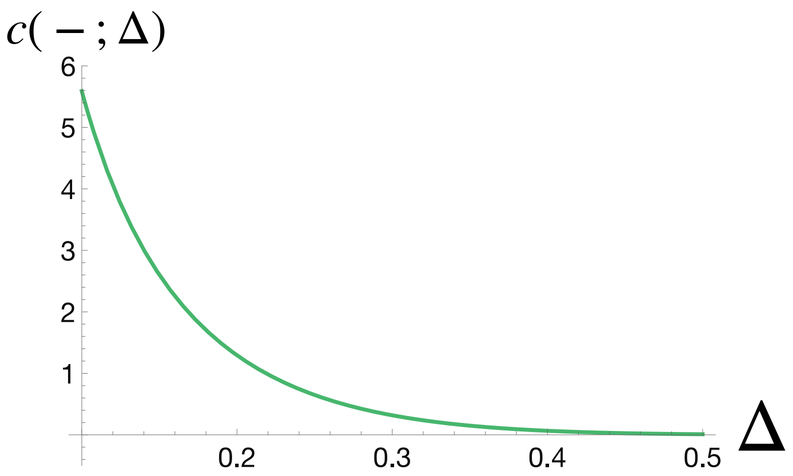
precomputed: $c(-, \Delta)$ $c(d; \Delta)$

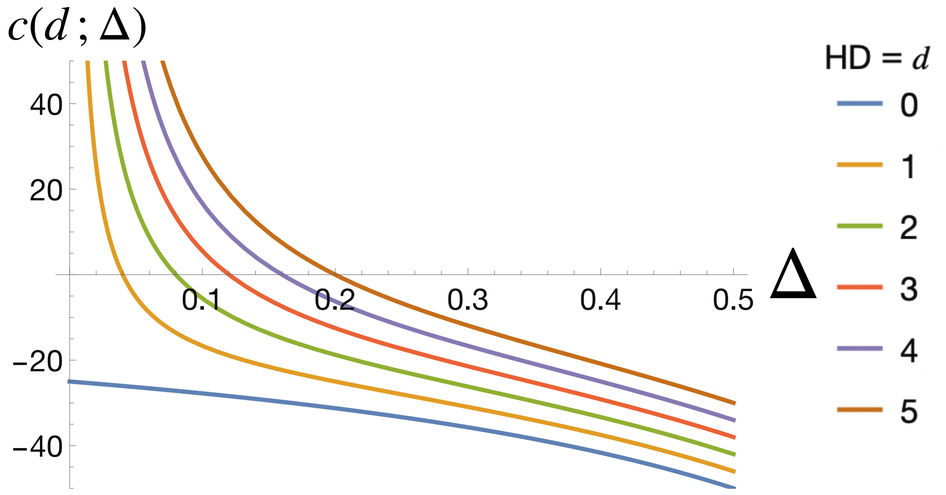
$$\ell'(D) = u \frac{\rho\left(\sum_{d=0}^{k} D^d (1-D)^{k-d} \binom{k}{d} \left(\frac{d-kD}{D(1-D)}\right) P_\delta(d)\right)}{1-\rho+\rho\left(\sum_{d=0}^{k} D^d (1-D)^{k-d} \binom{k}{d} P_\delta(d)\right)} + \sum_{d=0}^{\delta} v_d \left(\frac{d-kD}{D(1-D)}\right)$$



matches: $\mathbf{v} = [v_0, v_1, \ldots, v_\delta]$

misses: $u = (j-i) - \sum_{d=0}^{\delta} v_d$





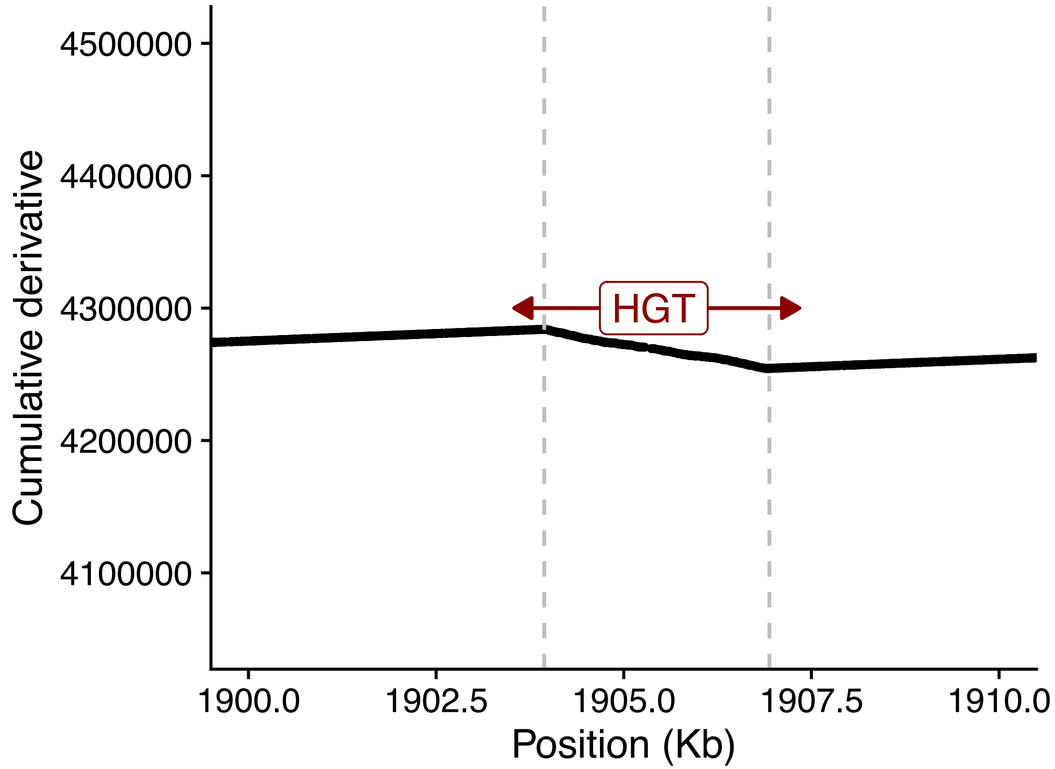
Each *k*-mer adds a constant to the derivate.

An example of horizontal gene transfer

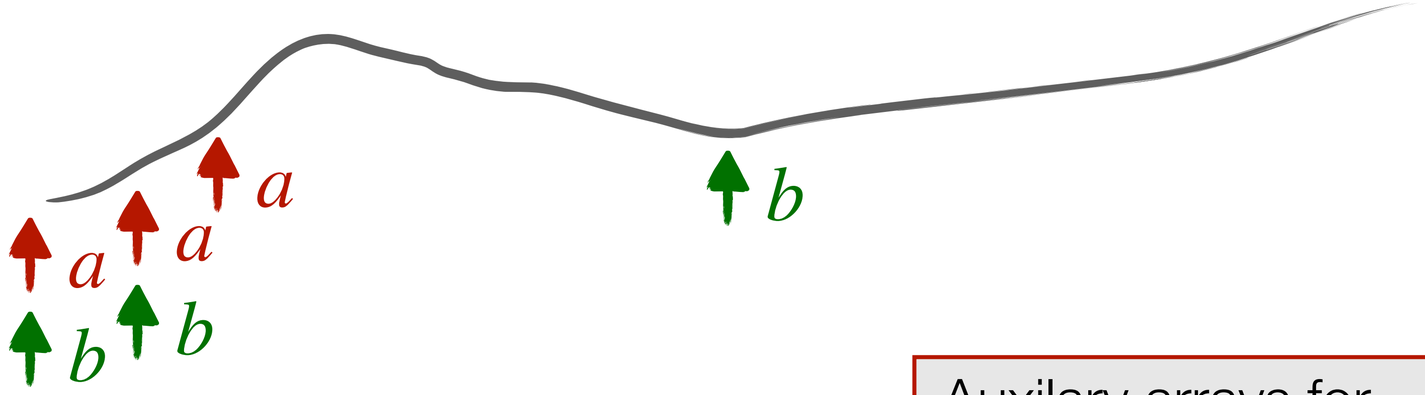
An “archaic” HGT event simulated using Zombi.

Genome-wide distance **32%**, gene distance **6%**

$$\Delta = 10\,\%$$



A linear time algorithm



Auxilary arrays for
$\min\{s_{b+1}, \ldots, s_L\}$, $\max\{s_1, \ldots, s_{a-1}\}$

Two pointers a and b:

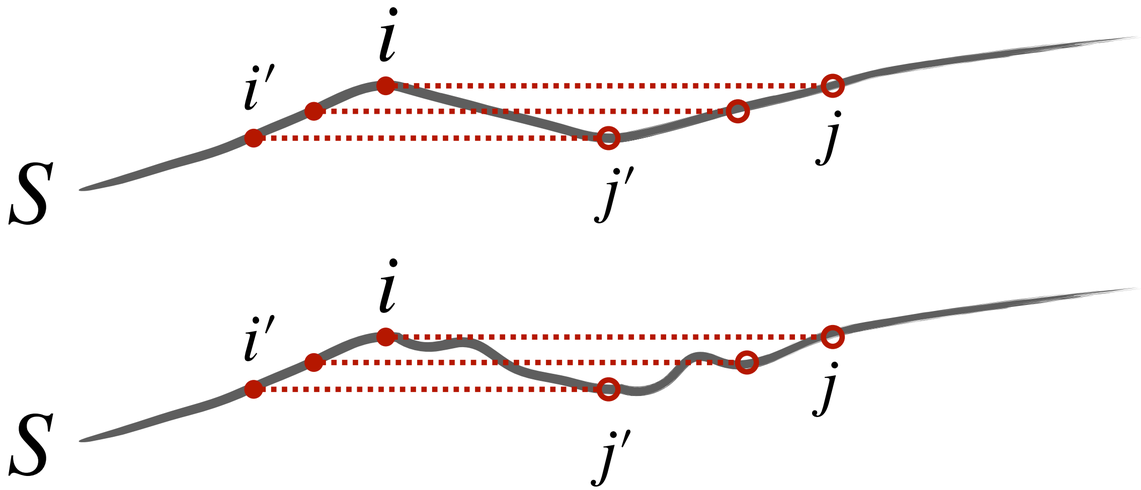
- For each prefix maxima s_a
- Increment b until $\min\{s_{b+1}, \ldots, s_L\} \geq s_a$ **(right maximal)**
- Check if $s_a > s_b$
- Check if $s_b \geq \max\{s_1, \ldots, s_{a-1}\}$ **(left maximal)**

} **report** $[a, b)$

Merging overlapping intervals

Maximal intervals can overlap…

$s_i > s_j$, $s_{i'} > s_{j'}$ but $s_{i'} \leq s_j$



Idea: merge if distance of $[i', j)$ is close to Δ

Can we test the against the maximum likelihood distance without calculating it?

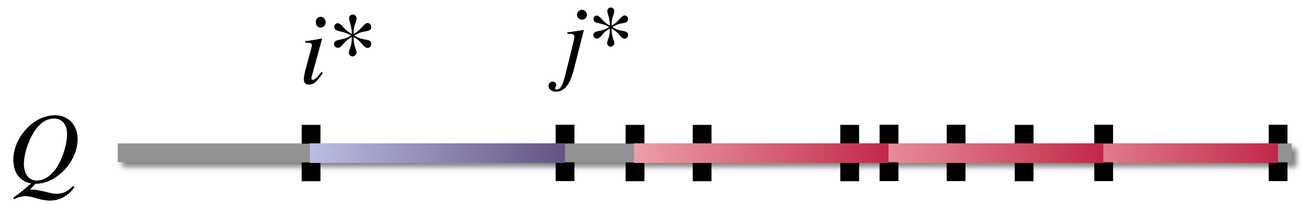
Yes! If $\hat{D} \approx \Delta$, then:

$$\ell(\hat{D}) - \ell(\Delta) \approx \frac{1}{2} \frac{\ell'(\Delta)^2}{-\ell''(\Delta)}$$

Score test

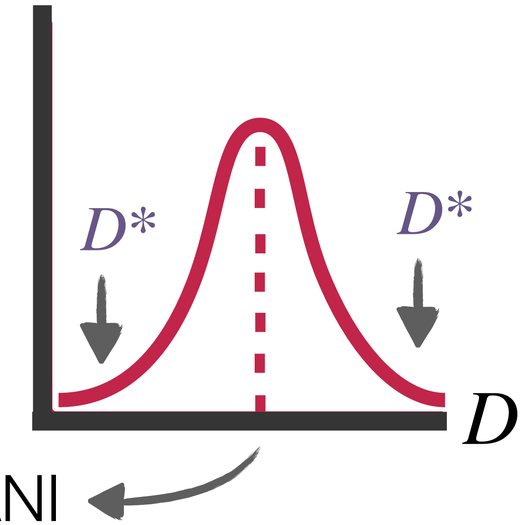
Accounting for rate variation and outliers

Given a candidate segment on Q w.r.t. R, compute its MLE distance D^*



Sample segments across the genome, estimate MLE distances

Fit a Gamma distribution to model the rate variation:



two-sided test for D^*

Obtain a p-value

conserved/HGT?

contamination?