

Motivation

Estimating alignment-free distances btw. sequences of any length:

- Scalable • Accurate & sensitive (e.g., >15%) • Modeling uncertainty

Applications of such a framework:

Metagenomics:

- Read-to-genome distances
- Phylogenetic placement

Comparative genomics:

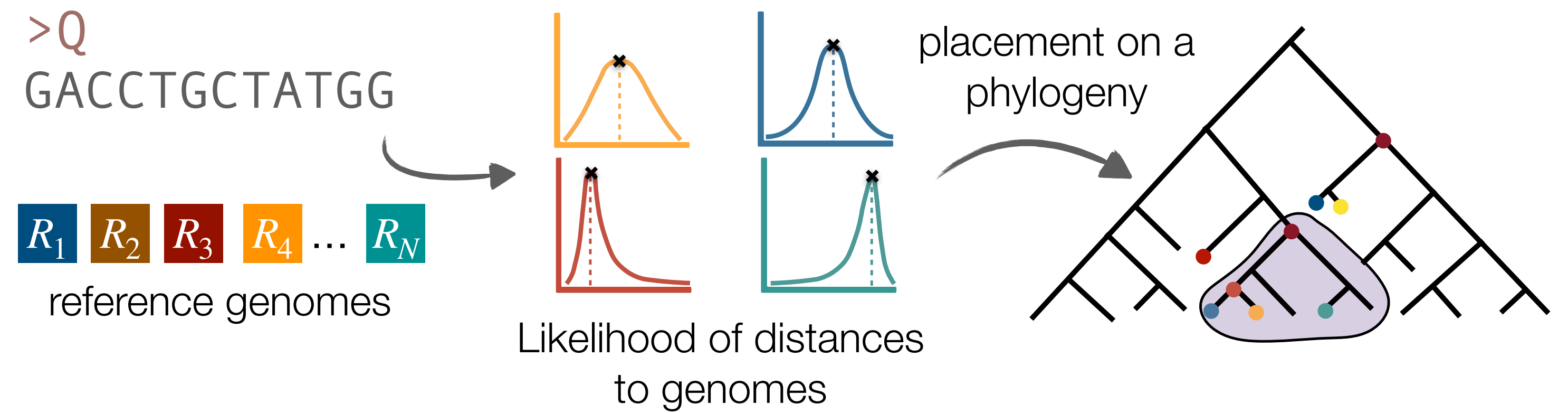
- Horizontal gene transfer
- Ultra-conserved elements

Problem

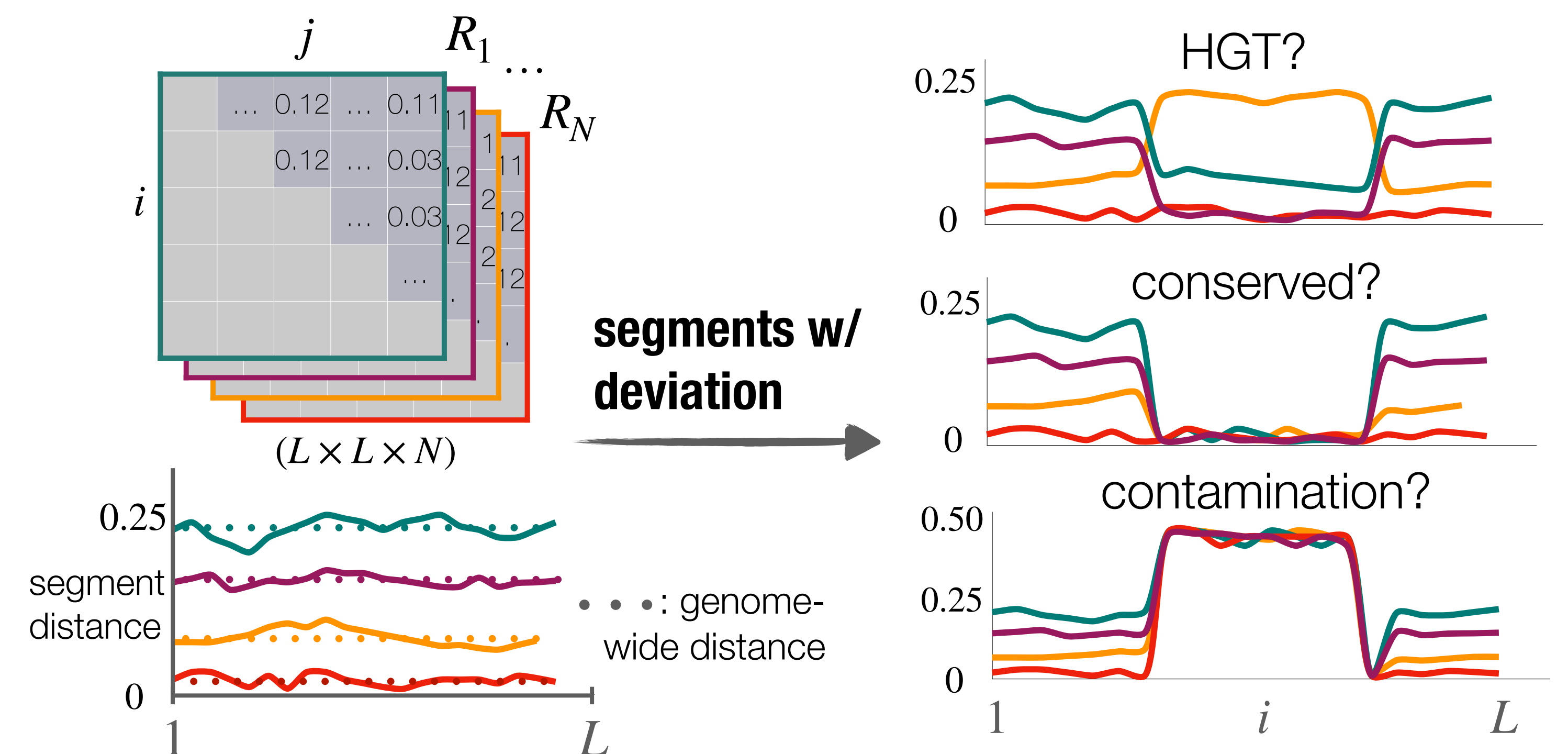
For a query Q and a reference $R \in \mathcal{R}$

- Compute $\mathcal{L}(D_{ij}; Q, R)$ using k -mers, where $D_{ij} = d(Q[i:j], R)$.
- Estimate $D = d(Q, R)$ and place Q on a phylogeny over \mathcal{R} .
- For all (i, j) and a threshold Δ , decide (in constant time) if $D_{ij} < \Delta$ (or $> \Delta$)
- Given τ , find all “maximal intervals” with $D_{ij} < \Delta$ and $j - i \geq \tau$
 $d(Q[i':j'], R) \geq \Delta$ for $i' \leq i \leq j \leq j'$, $(i, j) \neq (i', j')$

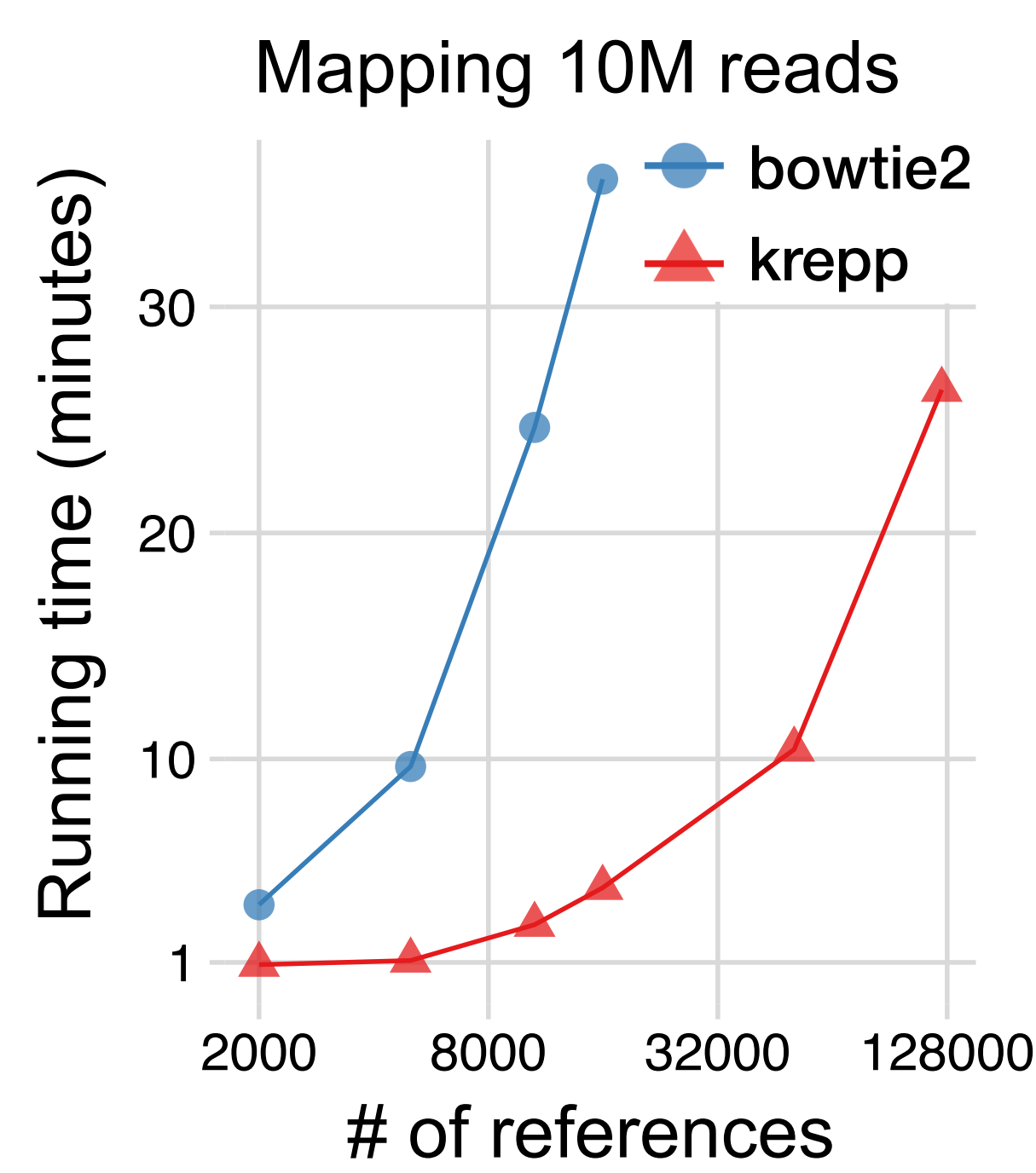
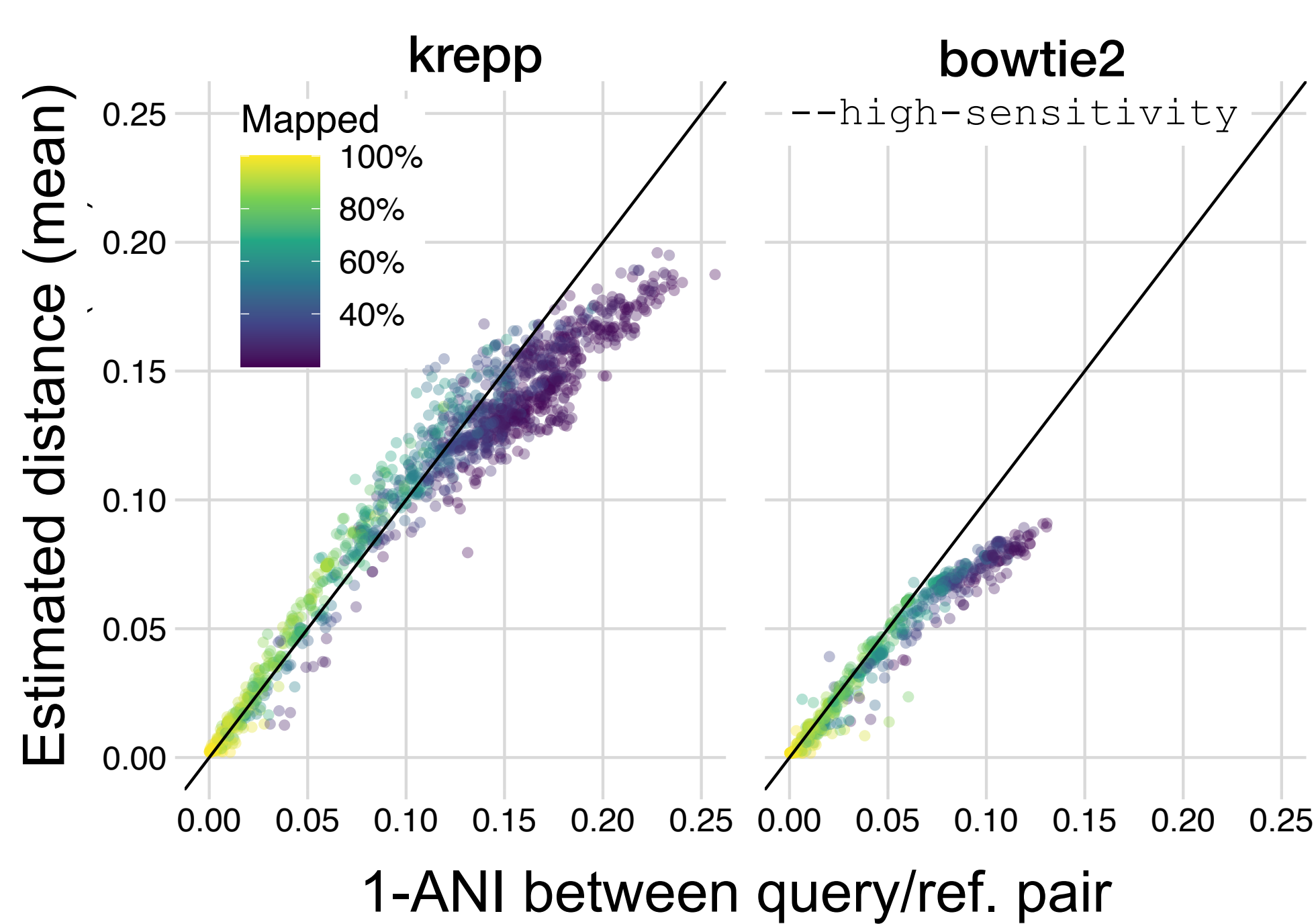
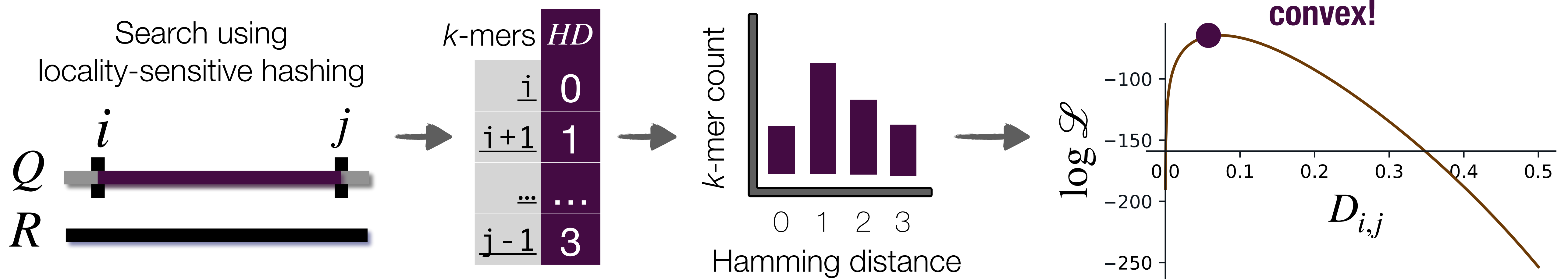
Modeling the distance between Q (e.g., read, contig, interval) & \mathcal{R} :



Strong deviations of “local distances” from genome-wide distances:



Finding homologous k -mers and pseudo-likelihood of distances



$$\text{misses: } u = (j - i) - \sum_{d=0}^{\delta} v_d \quad \text{matches: } \mathbf{v} = [v_0, v_1, \dots, v_{\delta}]$$

$$\mathcal{L}(D; k, h, \delta, u, \mathbf{v}) = P_{\text{miss}}(D; k, h, \delta)^u \prod_{d=0}^{\delta} P_{\text{match}}(D; d, k, h)^{v_d}$$

We define probability of observing u misses & v_d HD = d matches based on **i**) subsampling rate **ii**) number of mutations **iii**) LSH match!

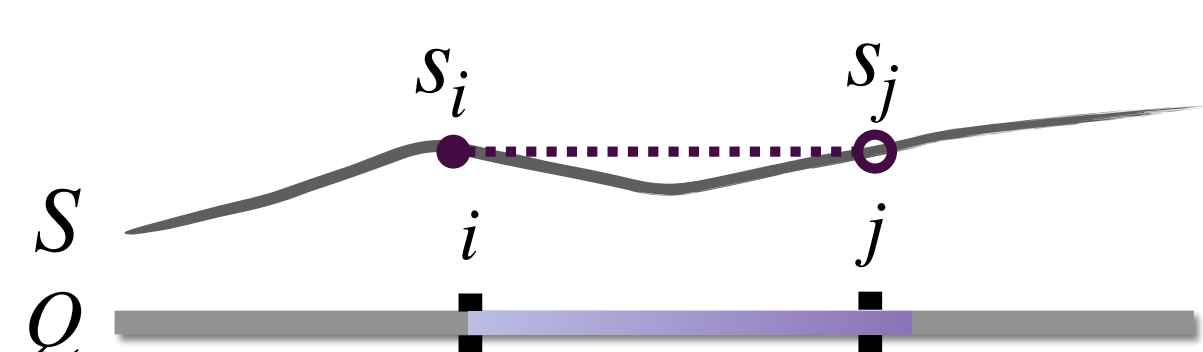
Distance-based segmentation & genome-wide pattern matching

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

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

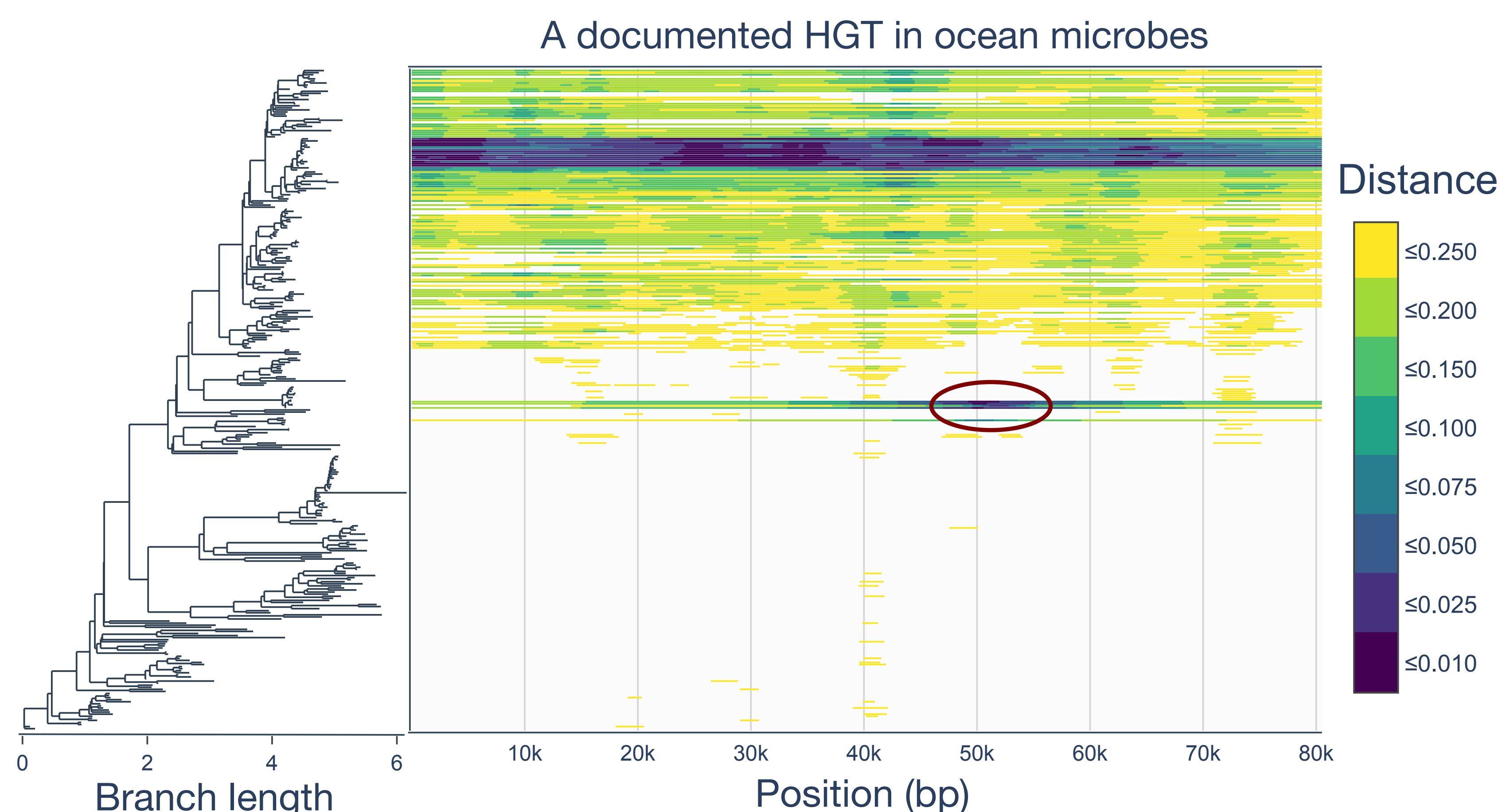
$$\text{If } \Delta > \hat{D} \rightarrow \ell'(\Delta) < 0$$

S : prefix-sum array for $\ell'(\Delta)$



► $\hat{D}_{ij} < \Delta$ iff $s_i > s_j$ (constant time)

► All maximal intervals (linear time)



krepp & gdiff

krepp for read-to-genome distances and placement:

GitHub:

Paper:



gdiff for genome-wide local distances and segmentation:

GitHub:

Paper:



Soon...