National Human Genome Research Institute

The Forefront of Genomics®

Mapping Impossible: Aligning long reads through centromeres and satellite arrays

Claudia Chu[1,2], Arang Rhie[1], Sergey Koren[1], Chirag Jain[1], Adam Phillippy[1]

[1]National Human Genome Research Institute, Genome Informatics Section, Bethesda, MD, [2]Georgia Institute of Technology, Atlanta, GA

TELOMERE-TO-TELOMERE CONSORTIUM

Computational and Statistical Genomics Branch

Abstract

Nanopore long read sequences have enabled the first assembly of an entire human chromosome X[1], including the whole centromere. However, these initial drafts contain consensus errors due to the assemblers and the high error rate in long reads. To improve the base-level accuracy, we perform polishing, which requires accurate mapping and alignment of the raw sequencing reads back to the assembly to make the appropriate corrections. Currently, existing mapping and alignment algorithms are not tuned to distinguish between variants and innate base pair errors in nanopore reads and highly repetitive regions such as the centromere.

We propose a weighted unique marker (uniq-mer) approach to calculate similarity scores and identify the correct alignment for a read. We hypothesize this method will be more robust and less sensitive to long read errors in centromere assemblies.

Conclusions

- Nanopore long read sequencing enables assembly through highly repetitive regions
- Current mappers produce biased alignments in centromeric sequences due to chance errors from nanopore reads in otherwise very similar repeat regions
- Using unique markers (uniq-mers) to calculate similarity between two sequences is an alternative method to estimate alignment quality
- Weighted uniq-mers based alignment scores are more effective in repetitive regions, and non-effective in already unique regions
- Post-filtering alignments is dependent on the quality of the alignments from the chosen mappers

Scripts are available on GitHub:

github.com/cchu70/perfect_polish

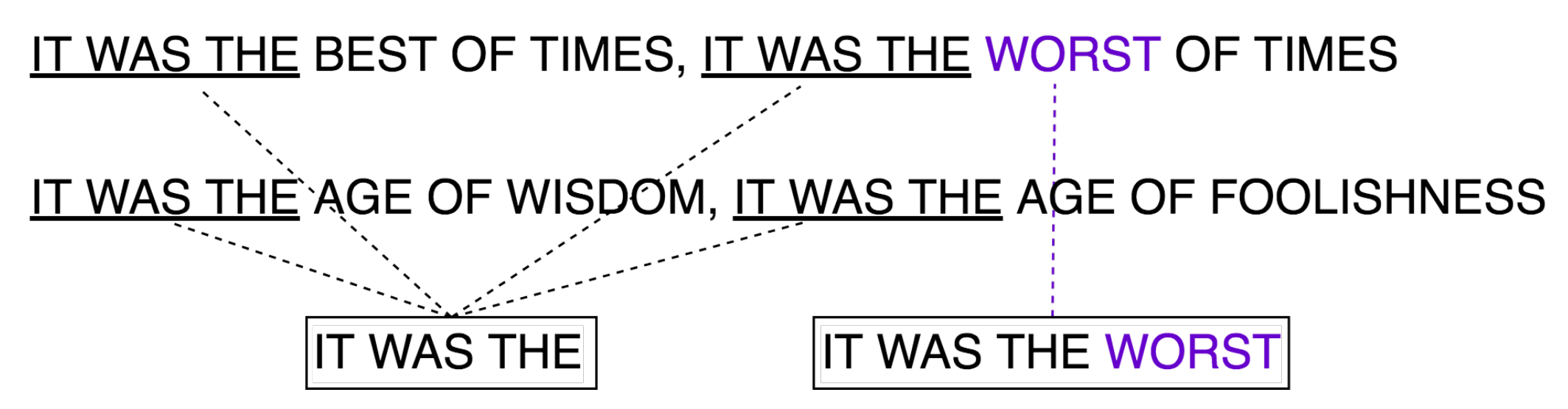
References

1. Telomere-to-Telomere Consortium, https://sites.google.com/ucsc.edu/t2tworkinggroup, Last accessed July 26, 2019
2. Heng Li, Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, 2018
3. Chirag Jain, Sergey Koren, Alexander Dilthey, Adam M Phillippy, Srinivas Aluru, A fast adaptive algorithm for computing whole-genome homology maps, *Bioinformatics*, 2018

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Theory

- ***Reads from repeat regions are difficult to align without unique markers***



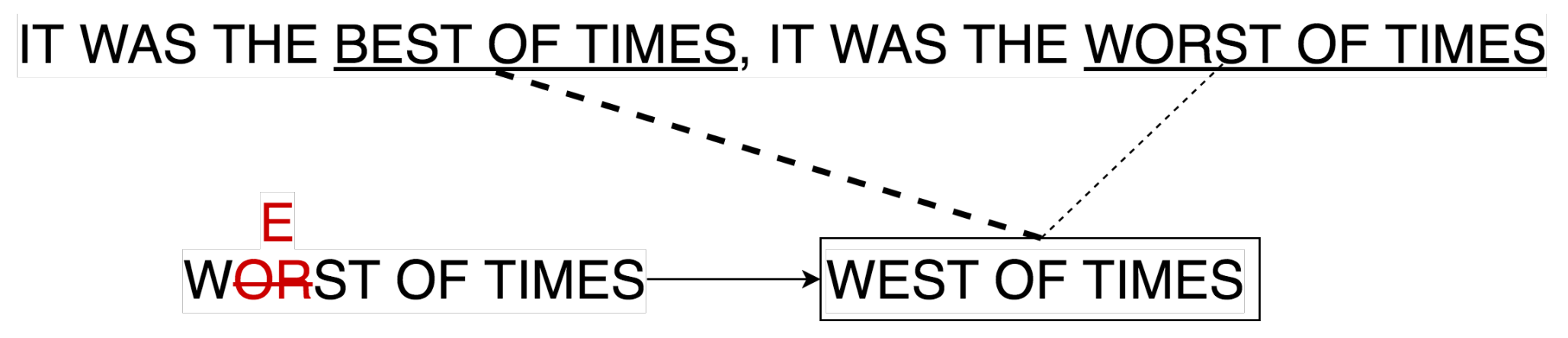
IT WAS THE BEST OF TIMES, IT WAS THE WORST OF TIMES

IT WAS THE AGE OF WISDOM, IT WAS THE AGE OF FOOLISHNESS

IT WAS THE

IT WAS THE WORST

- ***Chance errors in reads can bias alignments in similar regions***



IT WAS THE BEST OF TIMES, IT WAS THE WORST OF TIMES

E

WORST OF TIMES

WEST OF TIMES

- ***Uniq-mers are more informative***
  - **k-mers:** overlapping subsequence of length *k*
  - **uniq-mer:** a *k*-mer that only occurs once in the reference

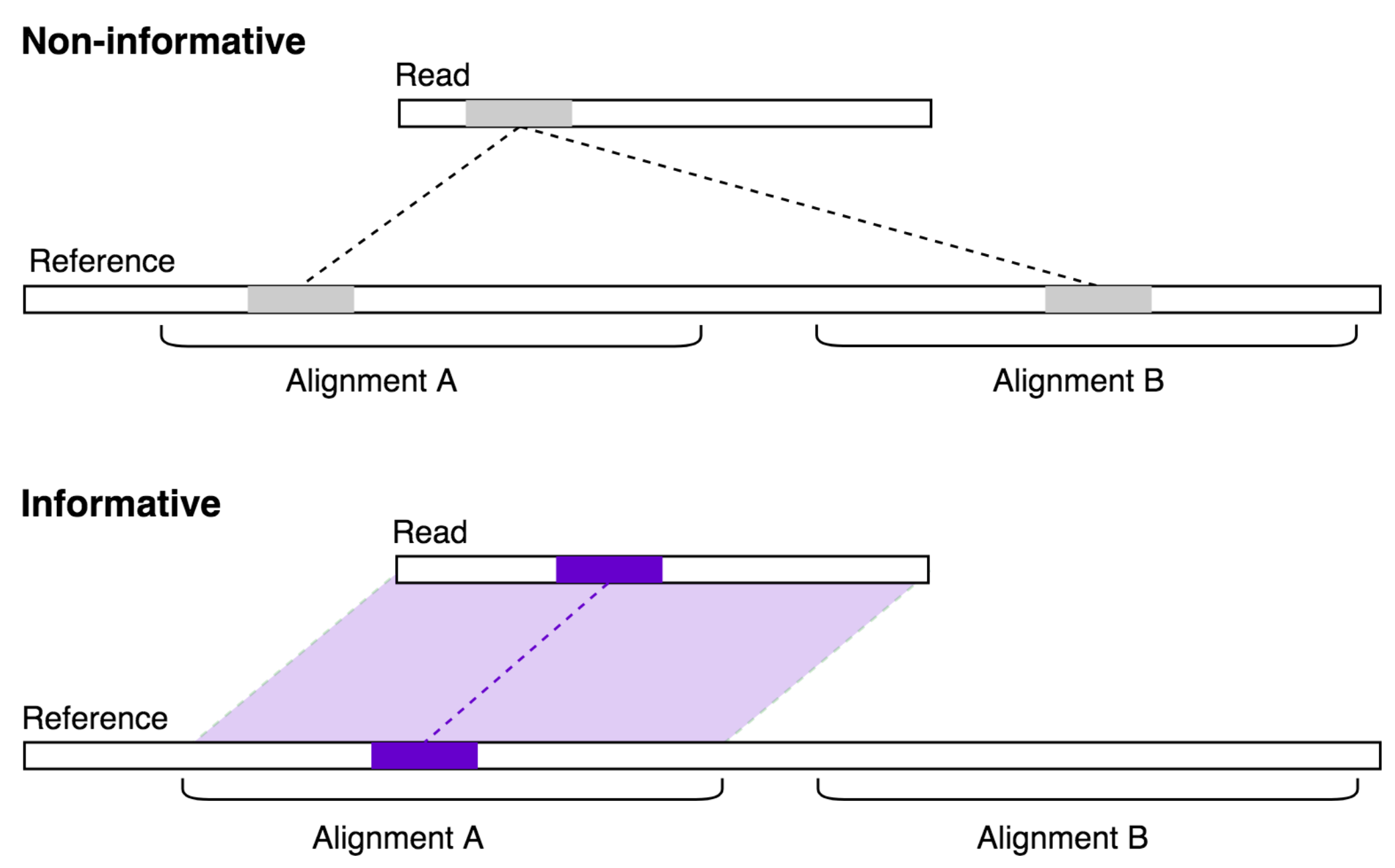
AGCTGATC

TAGCTGAT

CTAGCTGA

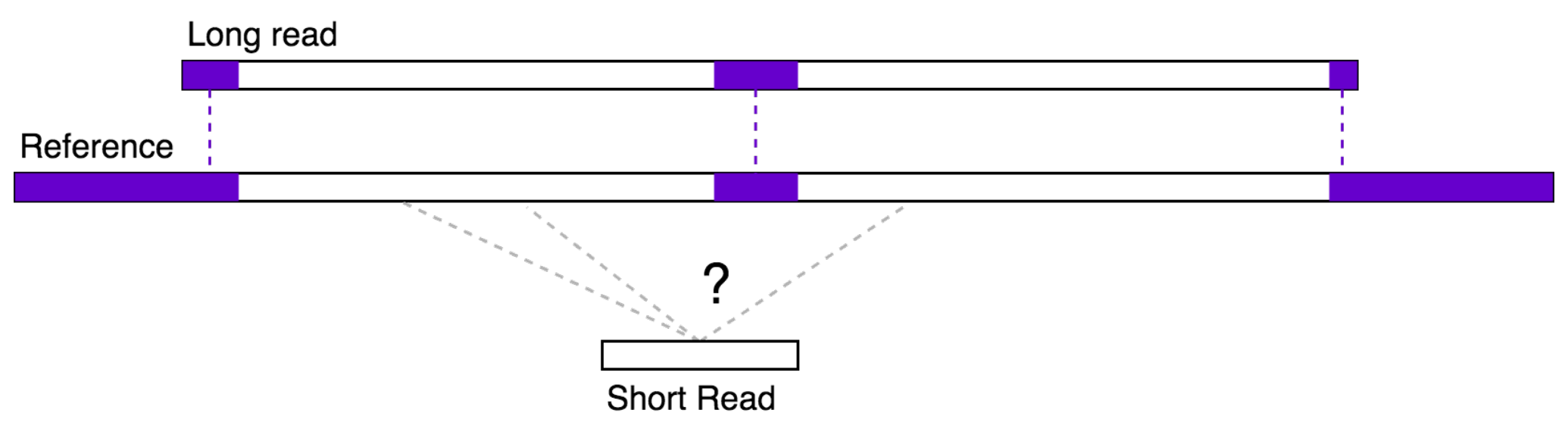
ACTAGCTG

ACTAGCTGATC...

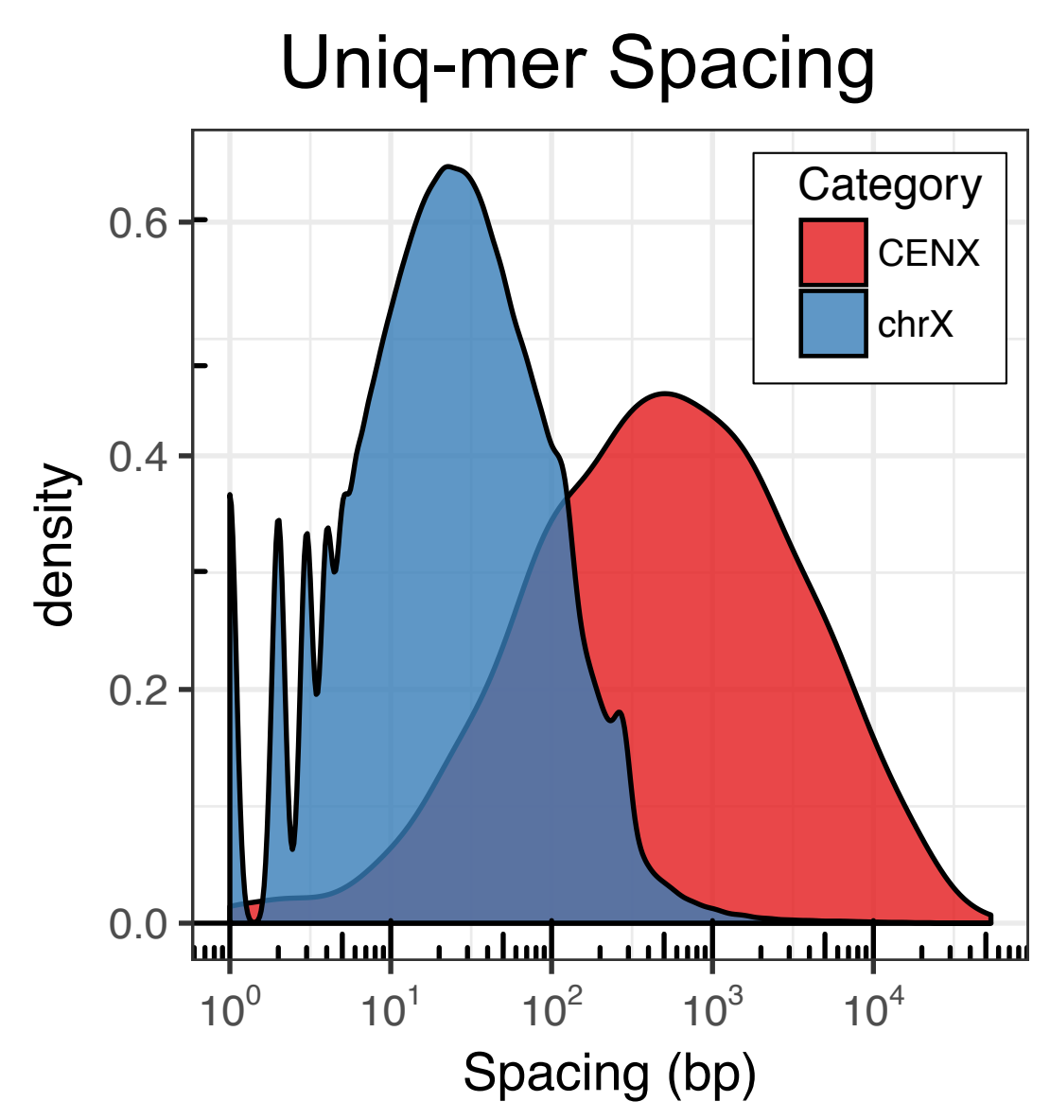


Feasibility

- ***Long reads likely contain unique regions***

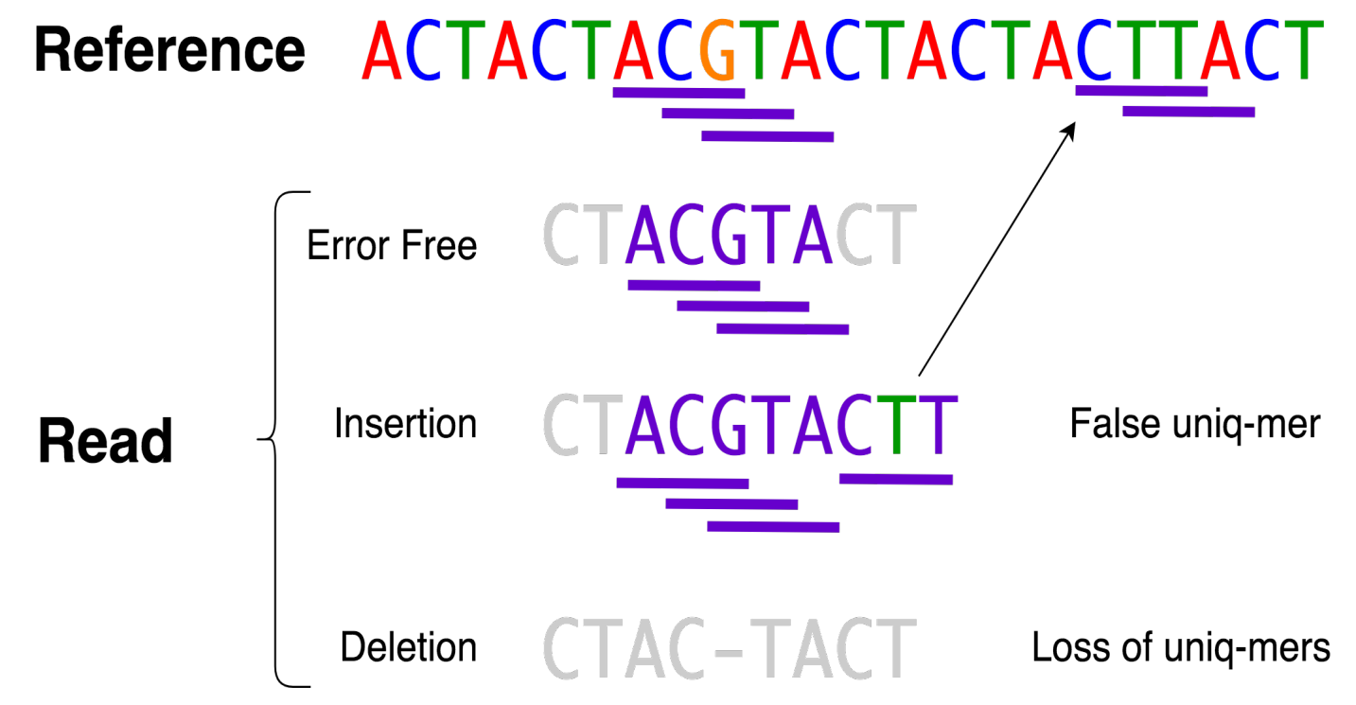


- ***Uniq-mers exist in centromere***



- Centromere region has larger spaces between unique regions
- Average nanopore read ~10kbp can span most of these non-unique regions

- ***Errors can cause false uniq-mers***



Reference ACTACTACGTACTACTACTTACT

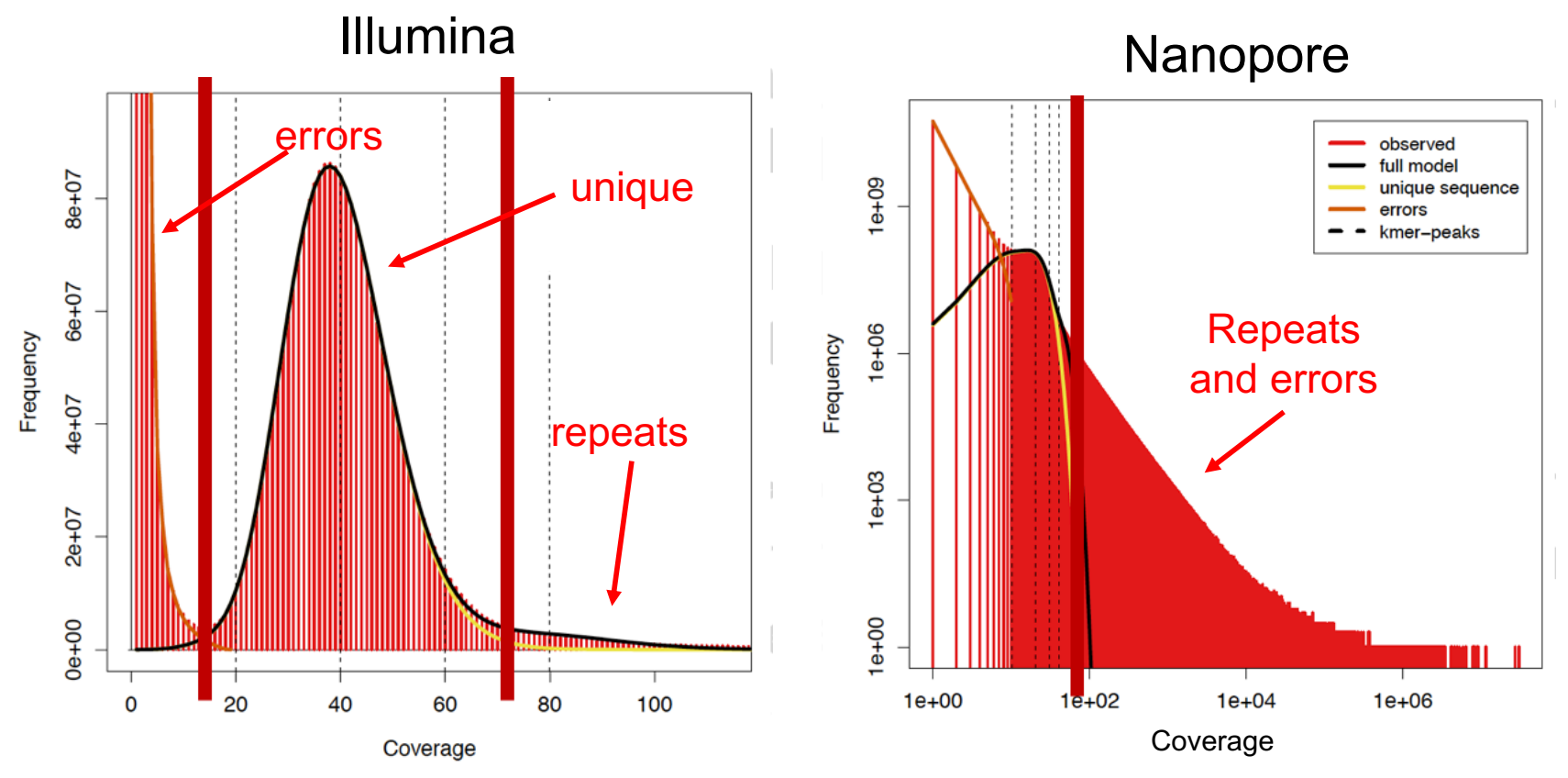
Read

Error Free CTACGTACT

Insertion CTACGTACTT False uniq-mer

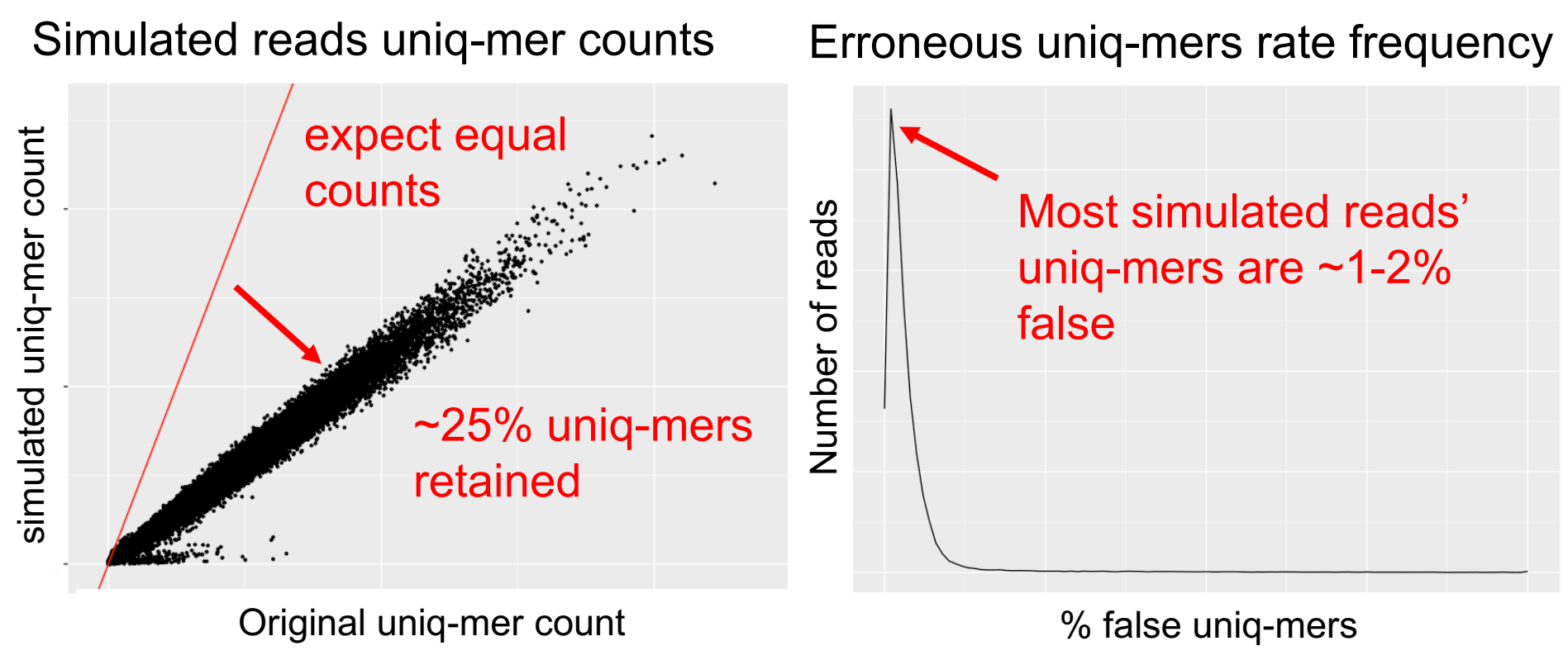
Deletion CTAC-TACT Loss of uniq-mers

- ***Pick reliable uniq-mers***



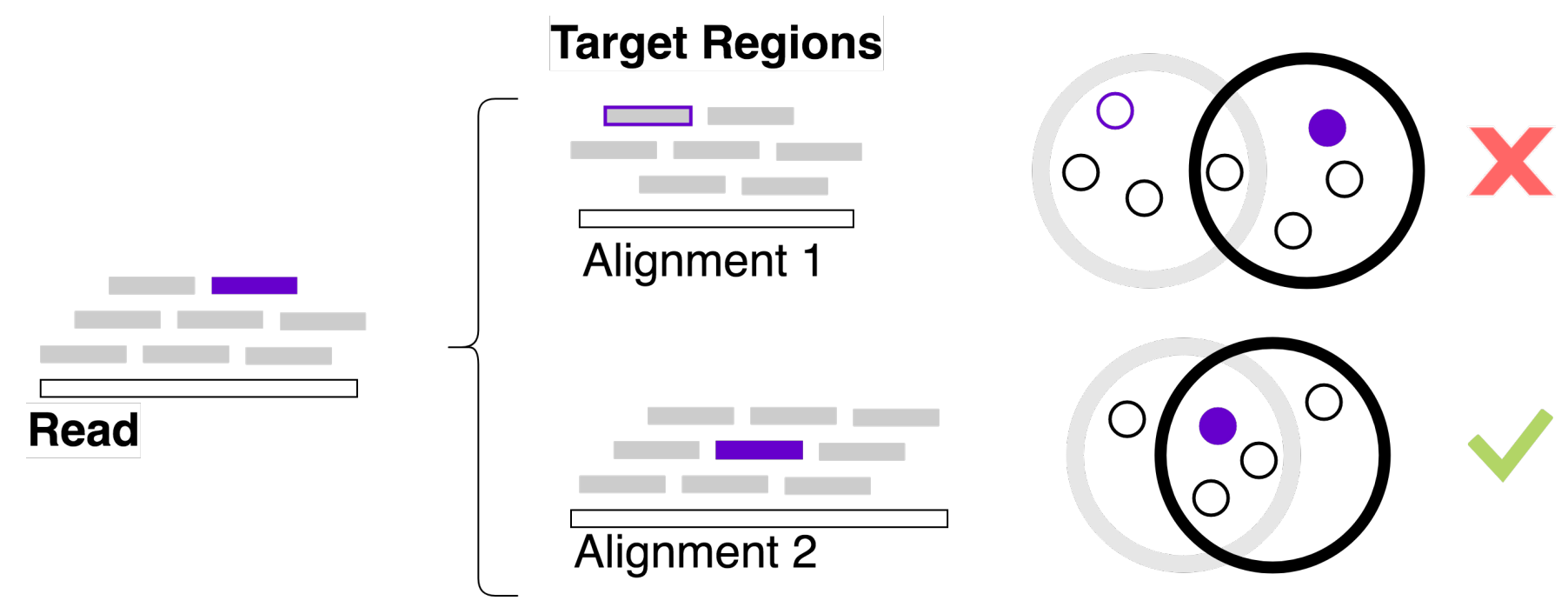
- Illumina datasets have fewer errors
- Pick *k*-mers with frequency = coverage
- **Ideal *k*-mer size = 21**: small enough to find in nanopore read, big enough to be unique

- ***Nanopore reads lose and gain uniq-mers***

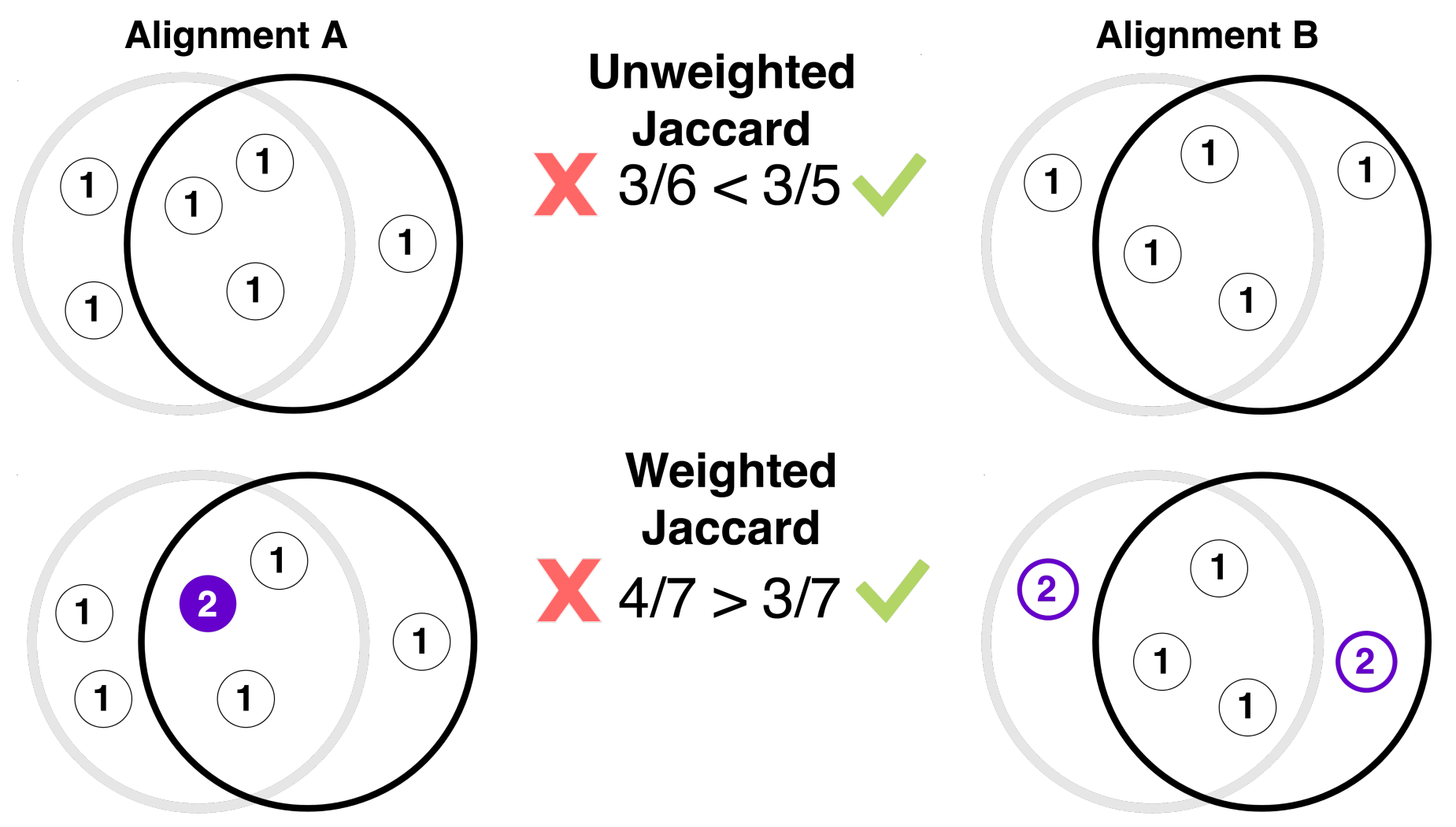


Algorithm

- ***Select best alignment for each read***
  1. Create uniq-mer set from the target
  2. Get multiple candidate alignments to the target for each read with *minimap2*[2]
  3. For each read, use Jaccard (set) similarity[3] of uniq-mers and *k*-mers to pick best alignment



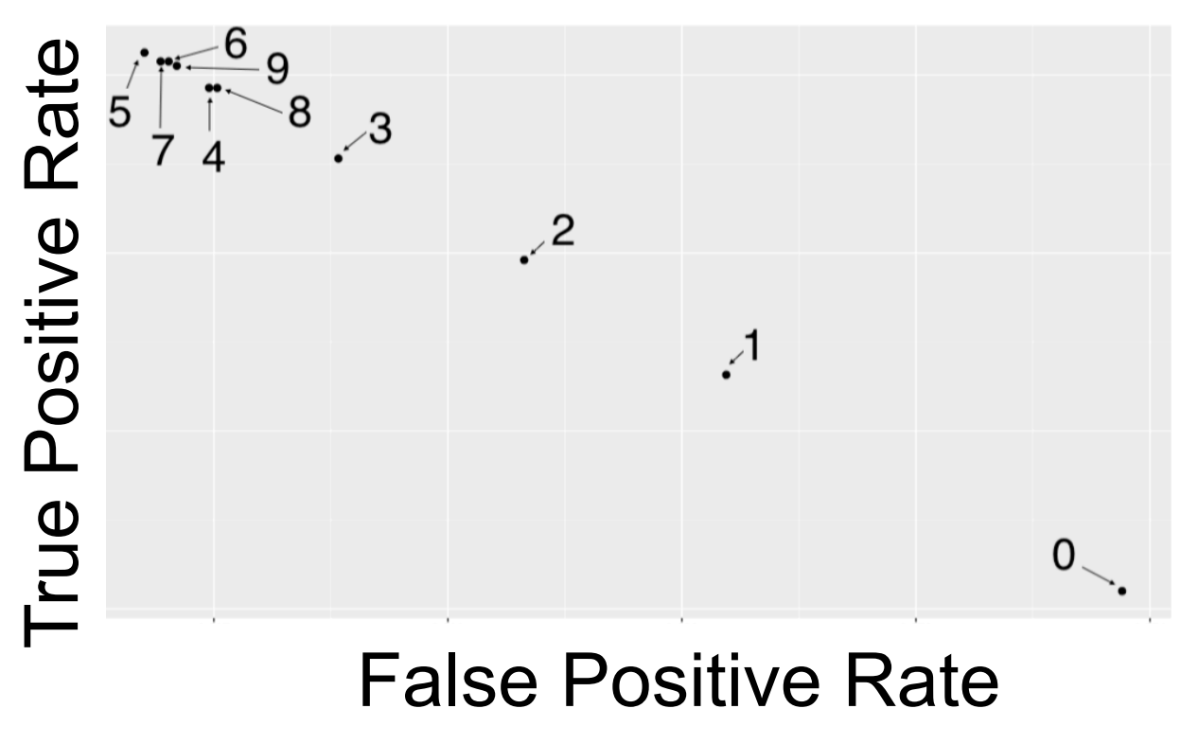
- ***Pick uniq-mer weight to reflect informativeness***



Results

- ***Weights for uniq-mer set similarity score***

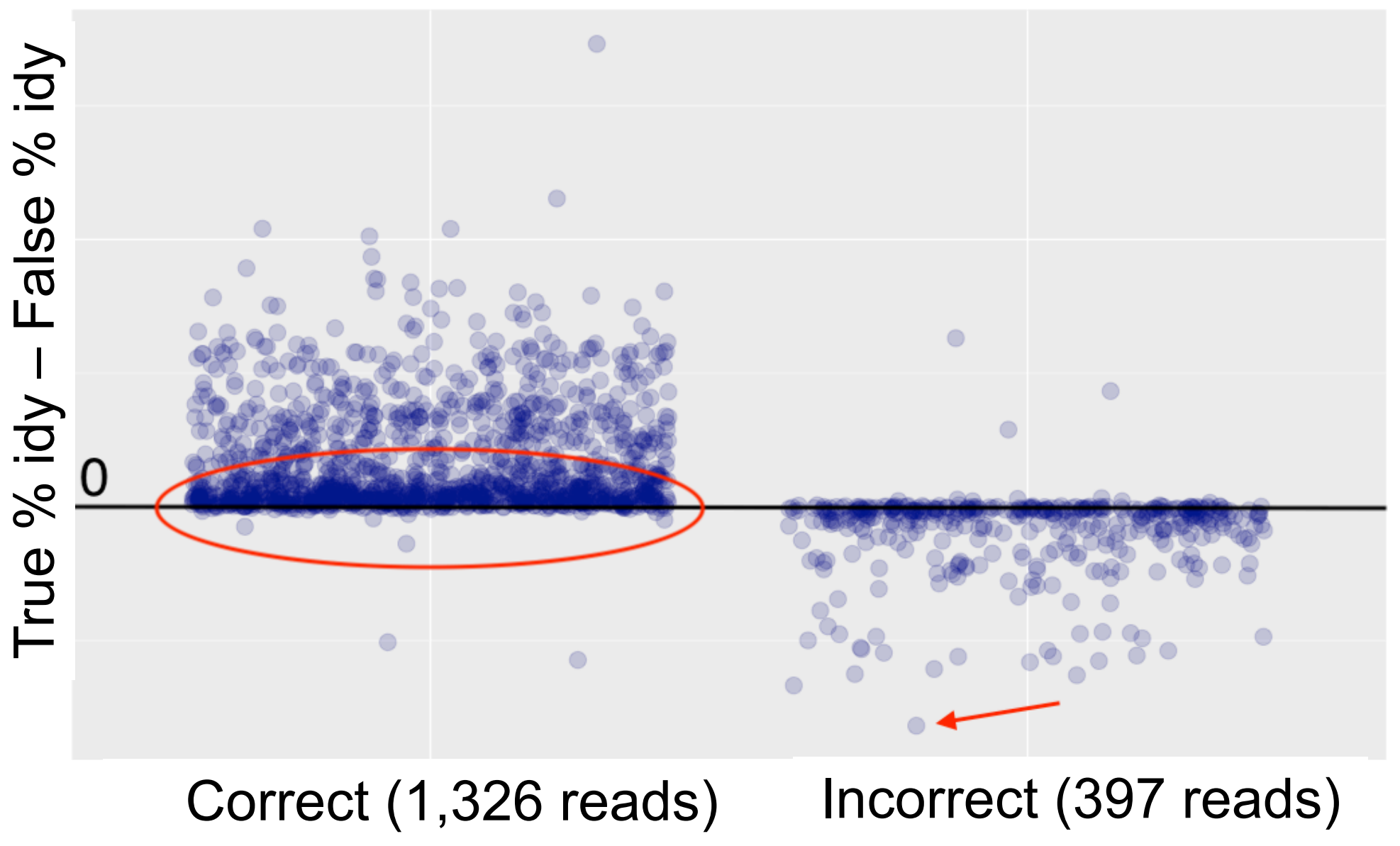
Uniq-mer weight performance in centromere



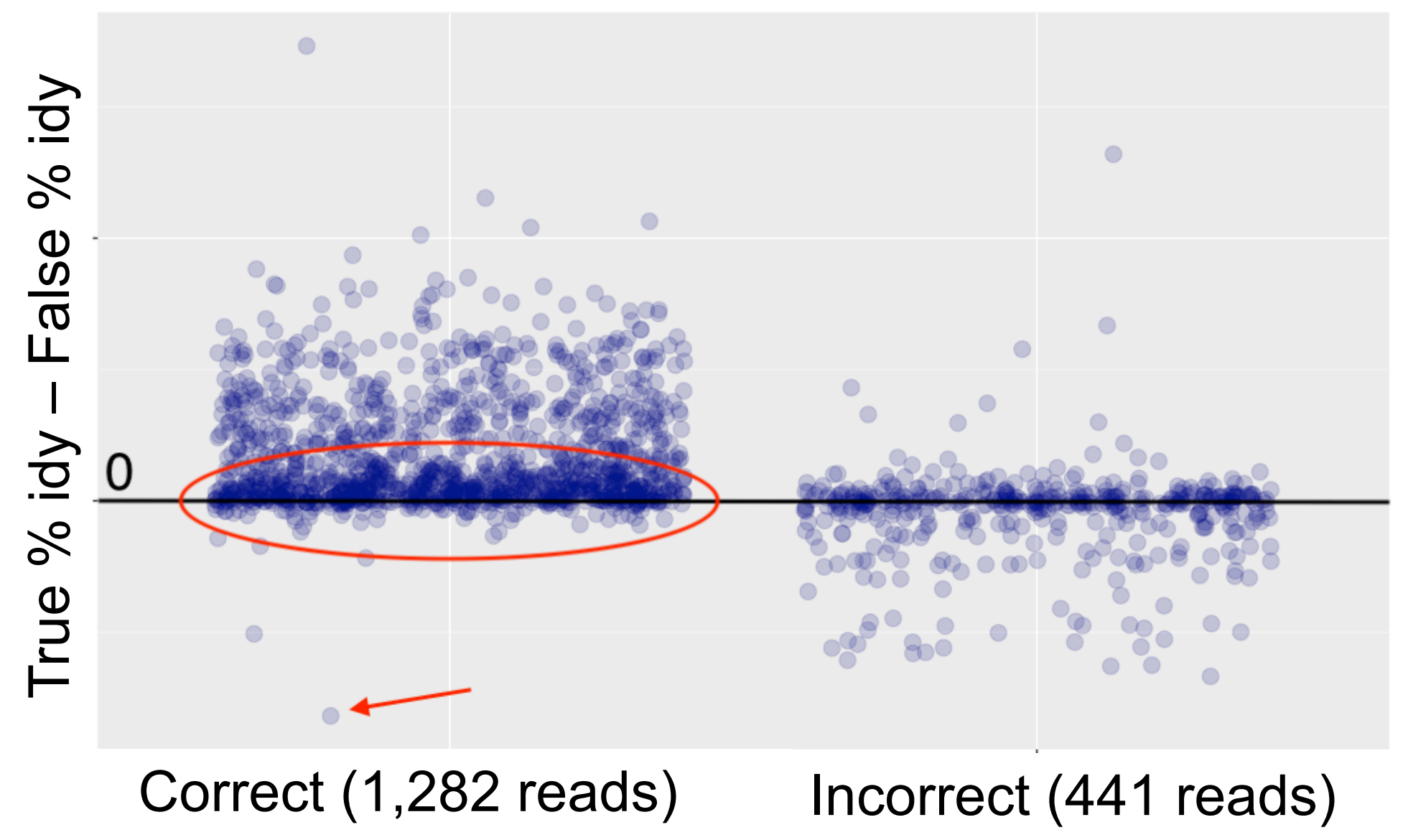
- Non-uniq-mers weight of 1
- Ideal weight higher for repetitive regions, and 1 for non-repetitive regions

- ***Weighted uniq-mer less percent identity bias***
  - **Percent Identity (% idy):** number of matching bases / total alignment length
  - *minimap2* preference to pick primary alignments with higher percent identity compared to uniq-mer method
  - Weighted uniq-mer corrects some of the reads *minimap2* gets incorrect

minimap2 Correct Primary Alignment

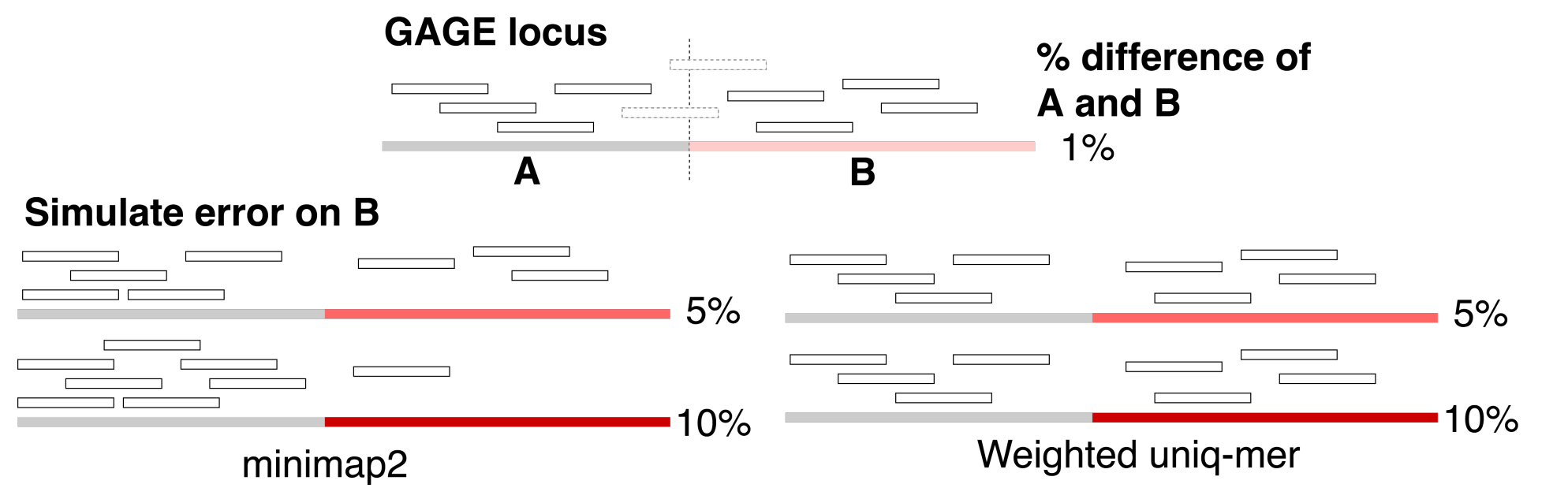


Weighted uniq-mer Correct Primary Alignment



Future Work

- Controlled simulation of inducing errors into centromere to observe percent identity bias



- Varying weights on the non-uniq-mers in addition to uniq-mer weights
- Build uniq-mer based mapper to produce own alignments rather than post-filtering existing alignments from other tools