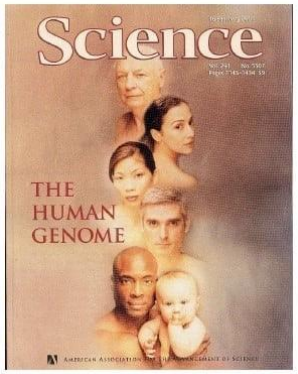
Georgia Tech Institute for Data Engineering and Science

LONG READ MAPPING AT SCALE: ALGORITHMS AND APPLICATIONS

SRINIVAS ALURU
GEORGIA INSTITUTE OF TECHNOLOGY

DATA DELUGE IN GENOMICS

Georgia Tech Institute for Data Engineering and Science



- Human genome project [1990 - 2003]
- First draft of the human genome published

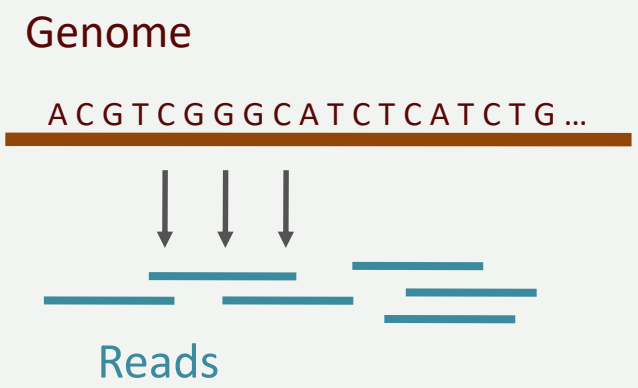
100,000
WHOLE GENOMES SEQUENCED
AT BROAD INSTITUTE

- Broad Institute sequenced its 100,000th human genome [April, 2018]
- 663 Tera bp DNA per month: equivalent to sequencing a human genome every 6 minutes

- RefSeq reference genomes add to 1.3 Tera bp (July 13, 2018) and growing!

LONG READ DNA SEQUENCING

Georgia Tech Institute for Data Engineering and Science



SHORT READS

(since 2006)

Illumina NovaSeq

- 150-bp reads
- Error-rates ~0.1%

LONG READS

(since 2011)

MinION

PromethION

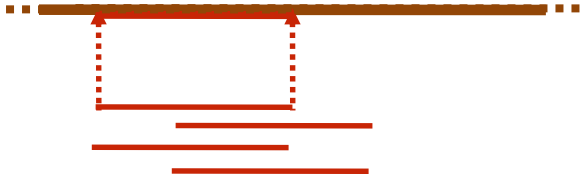
Oxford Nanopore Tech. (ONT)

- Variable lengths
- Long reads (mean >10,000 bp),
- Ultra-long reads (mean >100,000 bp)
- 40 Gb - 2Tb throughput in 48 hour run
- High error rate (10% - 20%)

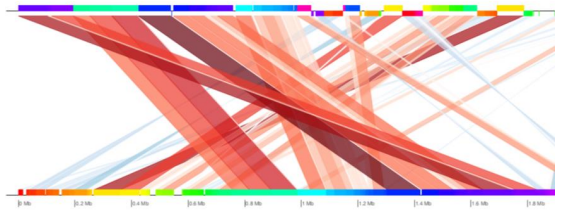
IN THIS TALK…

Georgia Tech Institute for Data Engineering and Science

1. Fast algorithm for mapping long reads that scales to large reference databases
   [RECOMB 2017, JCB 2018]

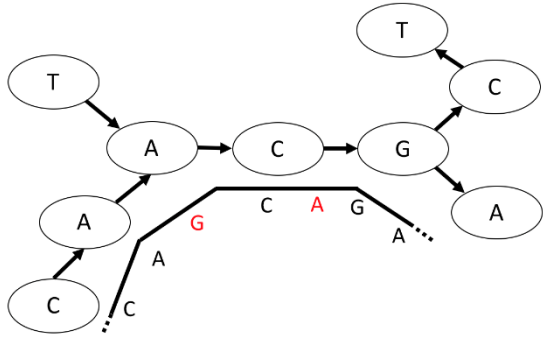


2. Split-mapping and whole-genome homology maps
   [ECCB 2018]

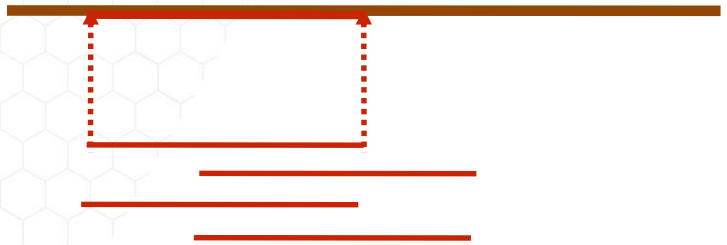


3. Alignment-free computation of genome-wide distance metrics
   [Nature Comm. (in press)]

4. Parallel algorithm to align long reads to graphs
   [IPDPS 2019 (under review)]



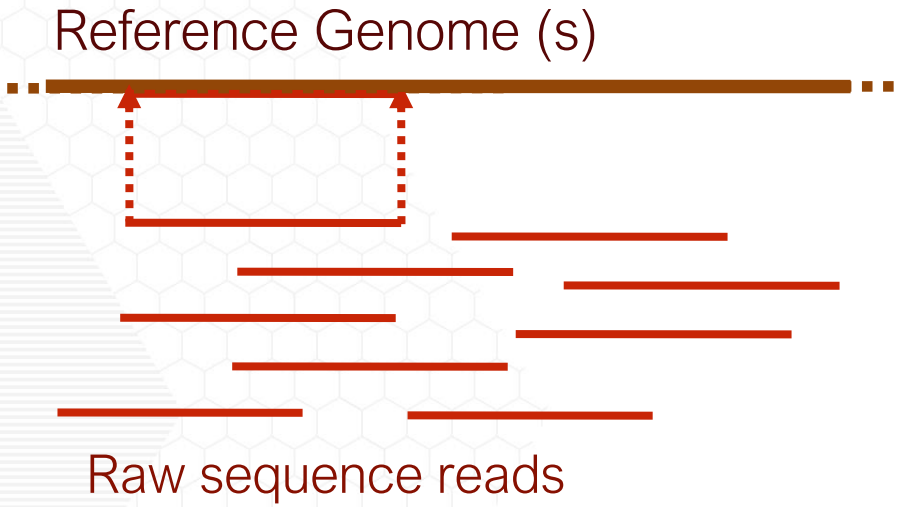
ALGORITHM TO MAP LONG READS TO LARGE DATABASES



LONG READ MAPPING TO REF. GENOME(S)

Georgia Tech Institute for Data Engineering and Science

- Seed and Extend Heuristics - Slow and inefficient
- Lack of theoretical guarantees
- Popular tools- BLASR; BWA-mem, GraphMap, Minimap



Mean length >10K bp

Error rate (10%- 20%)

Goal : Alignment-free approximate algorithm for fast computation of mapping positions and identity estimates for long reads

PROBLEM FORMULATION

Georgia Tech Institute for Data Engineering and Science

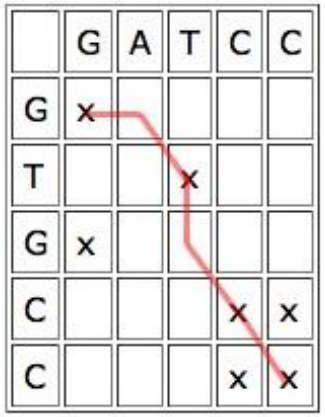
Given a read, identify all target positions in reference where it aligns with $\leq \epsilon_{cutoff}$ error rate.

Reference B

G C C C A T C C G C C G A T C C G G T A T C C T C

Read A

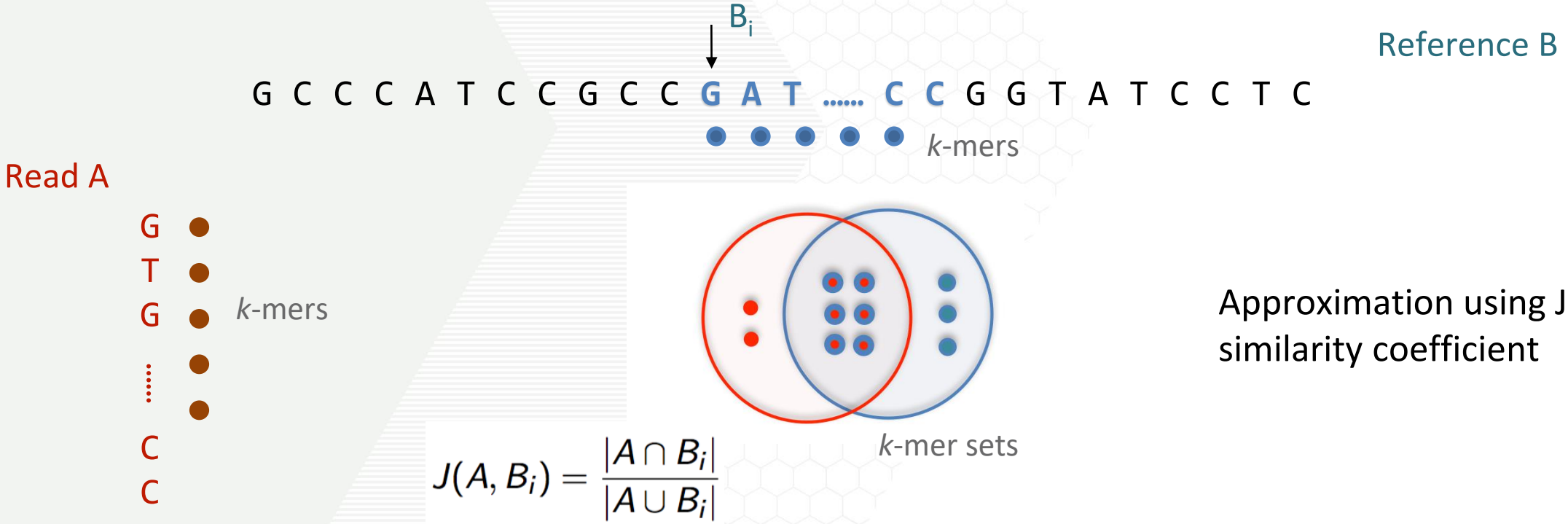
G
T
G
C
C



Exact solution: Semi-global alignment; but computationally prohibitive O(|A||B|)

APPROXIMATION

Given a read, identify all target positions in reference where it aligns with $\leq \epsilon_{cutoff}$ error rate.

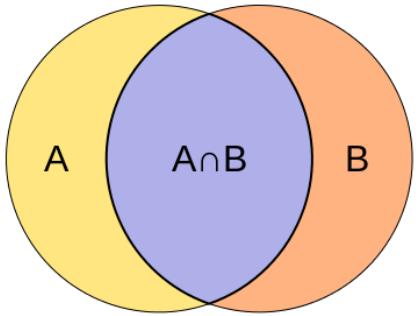


Approximation using Jaccard similarity coefficient

WHY JACCARD?

Georgia Tech Institute for Data Engineering and Science

Jaccard similarity: $$J(A, B_i) = \frac{|A \cap B_i|}{|A \cup B_i|}$$



MinHash

[Broder 1997]

Winnowing

[Schleimer 2003]

[Roberts 2004]

Jaccard similarity ⇤ ⇥ Error rate

[Fan 2015, Ondov 2016]

- Known techniques for efficient estimation of Jaccard similarity
- Originally developed for scalable web document clustering

(Assuming Poisson distribution for errors)
k = *k-mer* size
ϵ = alignment error-rate

$$\epsilon_{\text{mle}} = \mathcal{F}(J, k) = \frac{-1}{k} \times \log\left(\frac{2J}{1+J}\right)$$

WINNOWED-MINHASH ESTIMATOR

Georgia Tech Institute for Data Engineering and Science

A

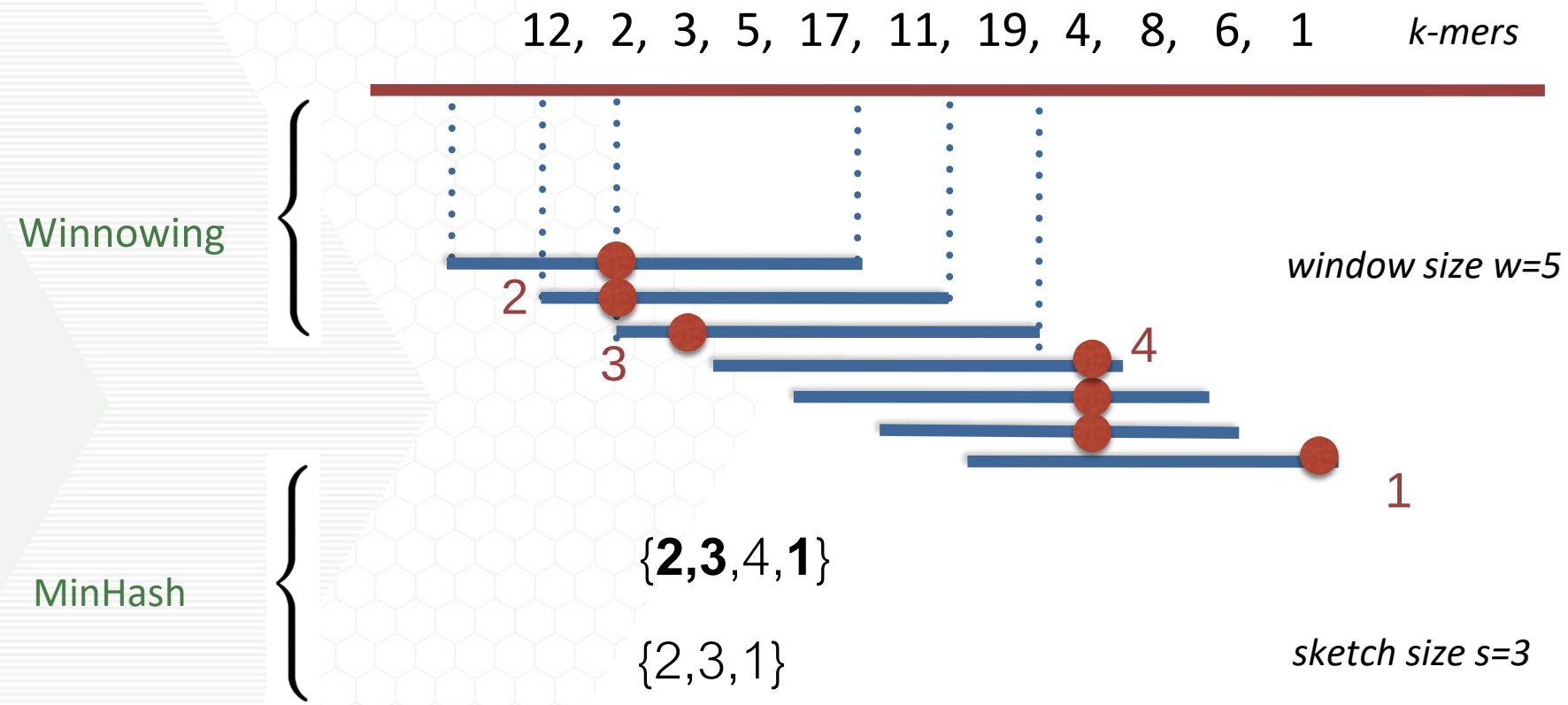
minimizers [Schleimer 2003]

$W(A)$

sketch [Broder 1997]

$S(W(A))$

$S(W(B))$



$$\mathcal{J}(A,B) = \frac{\left|S\left(W(A)\cup W(B)\right)\cap S\left(W(A)\right)\cap S\left(W(B)\right)\right|}{\left|S\left(W(A)\cup W(B)\right)\right|}$$

WINNOWED-MINHASH ESTIMATOR

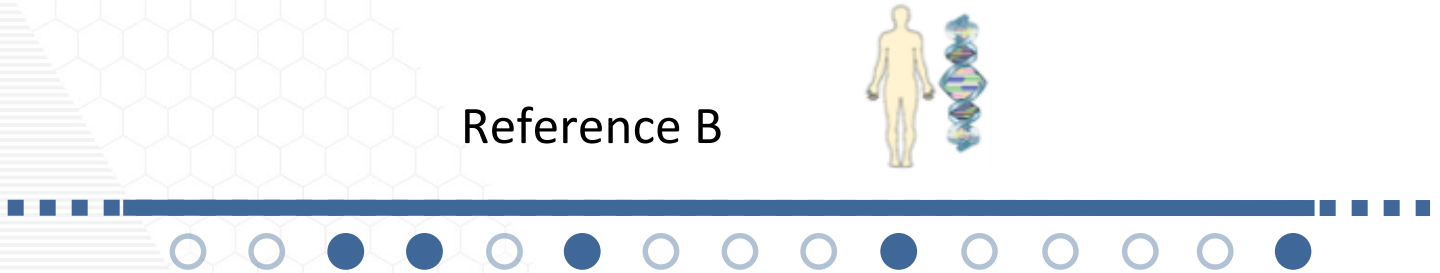
Georgia Tech Institute for Data Engineering and Science

$$J(A, B_i) \geq \tau?$$

$$\mathbb{E}(|W(A)|) = \frac{|A|}{w/2}$$

w = window size [Schleimer 2003]

(storing minimizers is sufficient)



Store all *k*-mers (Memory)

Evaluate all positions (Time)

WINNOWED-MINHASH ESTIMATOR

Georgia Tech Institute for Data Engineering and Science

$$J(A, B_i) \geq \tau?$$

$$\mathbb{E}(|W(A)|) = \frac{|A|}{w/2}$$

w = window size [Schleimer 2003]

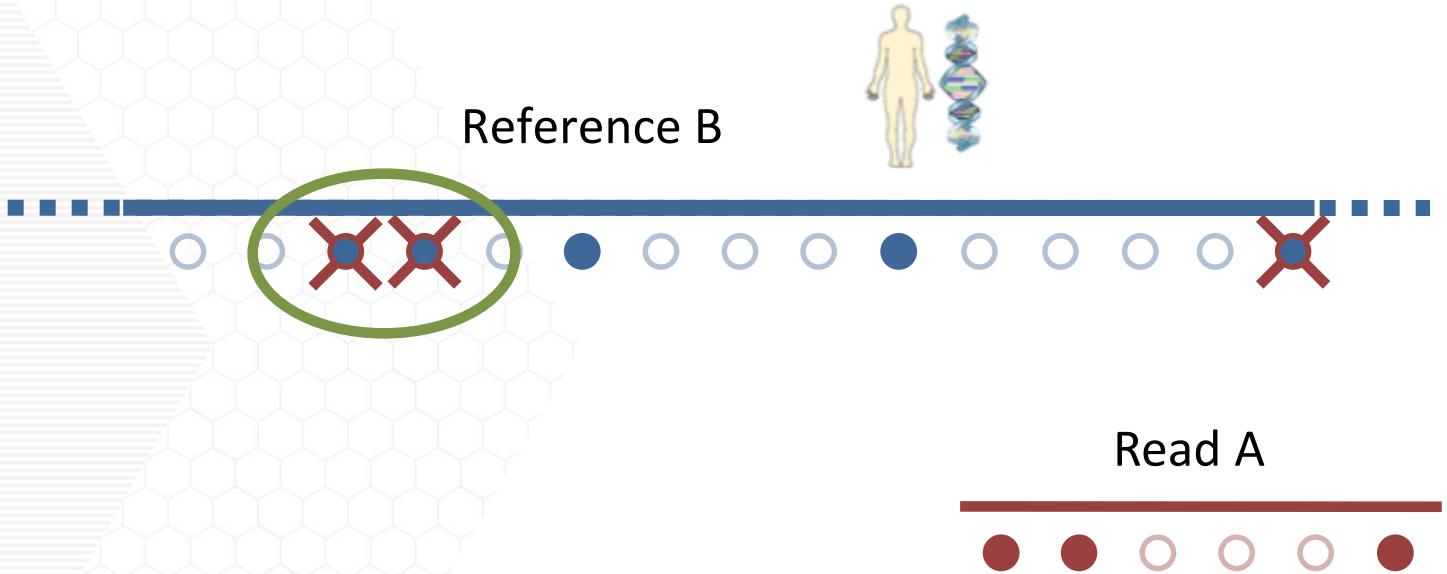
(storing minimizers is sufficient)

Theorem

$$\mathcal{J}(A, B_i) \geq \tau \Rightarrow |W(A) \cap W(B_i)| \geq s \cdot \tau$$

s = sketch size

(sufficient minimizers should match)



Store all *k*-mers ⚠ (Memory)

Evaluate all positions ⚠ (Time)

ADDITIONAL DETAILS

Georgia Tech Institute for Data Engineering and Science

- **Winnowed-MinHash Jaccard Estimator.** Adapt the classic minhash and minimizer techniques for efficiently computing Jaccard similarity along the reference.

- **Choosing k-mer sampling rate.** Automate the choice of an appropriate value of k-mer sampling rate

- **Proof of sensitivity.** Algorithm reports desired mappings (below the specified error-rate) with high probability.

[Jain et al. RECOMB 2017]

RESULTS

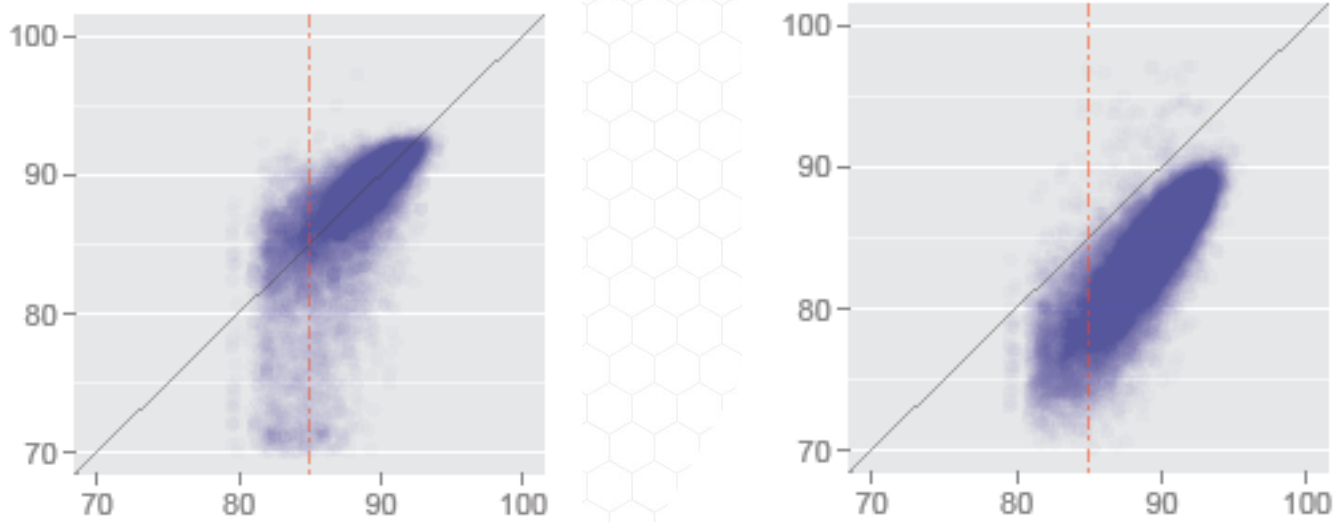
Georgia Tech Institute for Data Engineering and Science

Jaccard similarity of k-mer sets yields high quality alignment identity estimates

Pacbio (P1)

Nanopore (N1)

Smith-Waterman identity



Mashmap's estimated alignment identity

RESULTS

Georgia Tech Institute for Data Engineering and Science

Datasets

Id	Query data	#Reads	Mean Length	Reference
N1	E. *coli* K12 (MinION)	30K	14.0 Kbp	K12
P1	Human (Pacbio)	18K	14.5 Kbp	GRCh38

Mapping Time

(using AMD Opteron 2376 CPU)

Mashmap parameters
Error threshold = 15%,
read length ≥ 5K bp

Dataset	Mashmap	Minimap	BWA-mem	BLASR
N1	54s	**37s**	5h 39m	10h 17m
P1	**1m 24s**	1m 56s	6h 46m	20h 40m

Accuracy

Filtering: BWA-mem and minimap (qcov ≥ 80%)

Recall: against BWA-mem mappings (with <15% error)

(Recall)

Dataset	Mashmap	Minimap
N1	**100%**	99.87%
P1	96.8%	**98.7%**

Precision: Validation using Smith-Waterman alignments
(<25% error, ≥ 80% qcov)

(Precision)

Dataset	Mashmap	Minimap
N1	**94.39%**	94.32%
P1	**84.59%**	30.34%

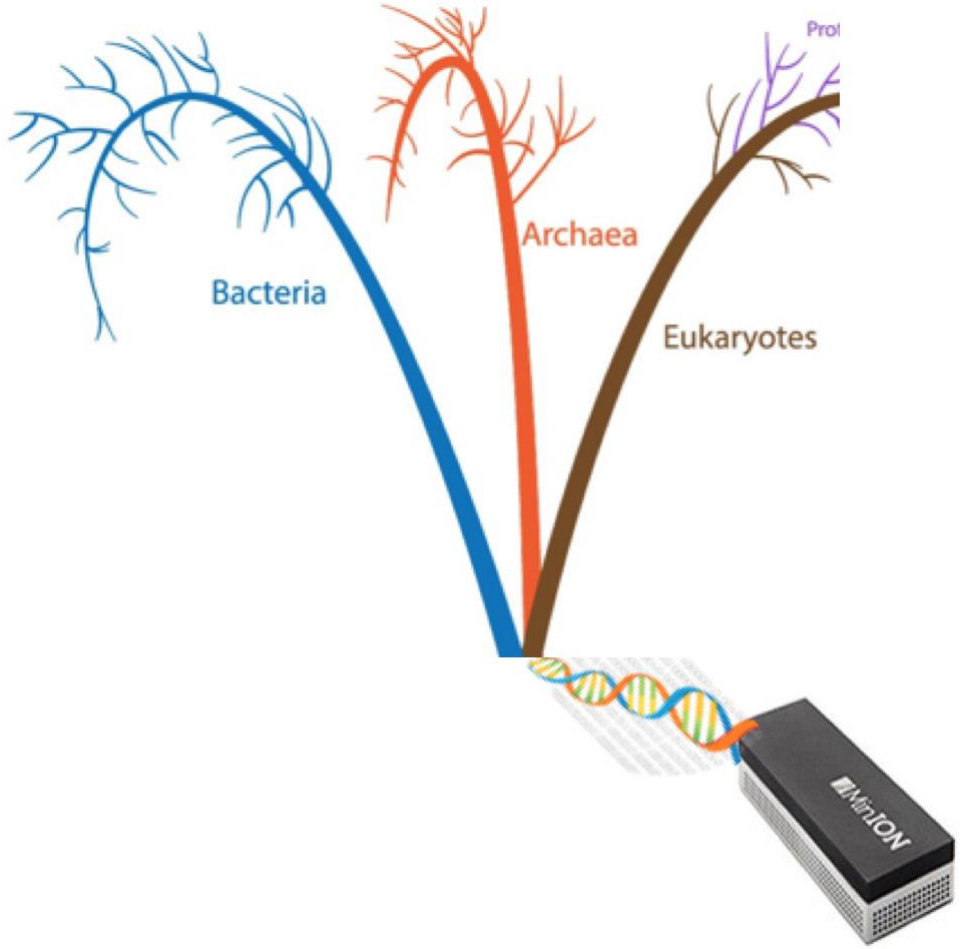
RESULTS

Georgia Tech Institute for Data Engineering and Science

First algorithm to scale to complete RefSeq database

- Reference : RefSeq (838 Gbp, >60K genomes)
- Query : Pacbio sequences (130 K) from mock Human Microbiome Project sample
- 29 CPU hours for index, and 16 CPU hours for mapping using 660 GB memory (BWA-mem, minimap, BLASR require more than a TB memory)

  (Parameters: error threshold = 15%, read length ≥ 5K bp)
- Recall against BWA-mem mappings to 20 known genomes ranges from 97.7% to 99.1%



SPLIT READ MAPPING OF LONG READS AND WHOLE-GENOME COMPARISON

CONTRIBUTIONS

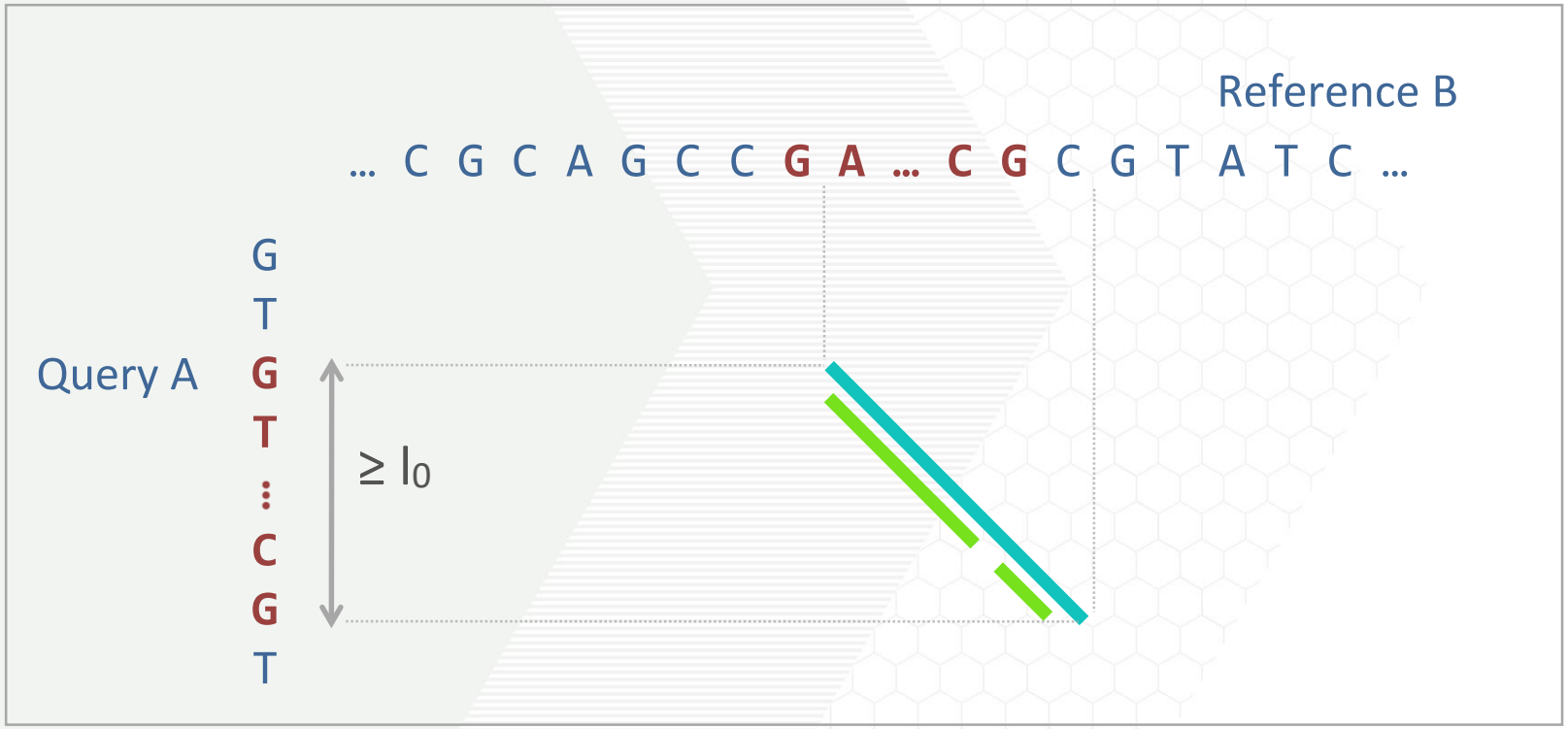
Georgia Tech Institute for Data Engineering and Science

- Extend Mashmap algorithm to compute genome-to-genome mapping and split-read mapping
- Mathematically show that all valid local alignment boundaries, that satisfy the user-specified minimum alignment identity and length thresholds are reported with high probability
- New plane-sweep based O(n log n) algorithm to filter mappings; e.g., to identify orthologs or paralogs
- Mashmap2 operates in about a minute and < 4 GB memory to map human genome assembly to a human reference
- deployed for validating genome assemblies in the Vertebrae Genomes Project (VGP)

PROBLEM FORMULATION

Georgia Tech Institute for Data Engineering and Science

Input: query, reference, thresholds for local alignment (min. identity, min. length l_0)



Exact solution: $\Omega(|A||B|)$

Instead, seek mappings (long inexact seeds) along the optimal alignment

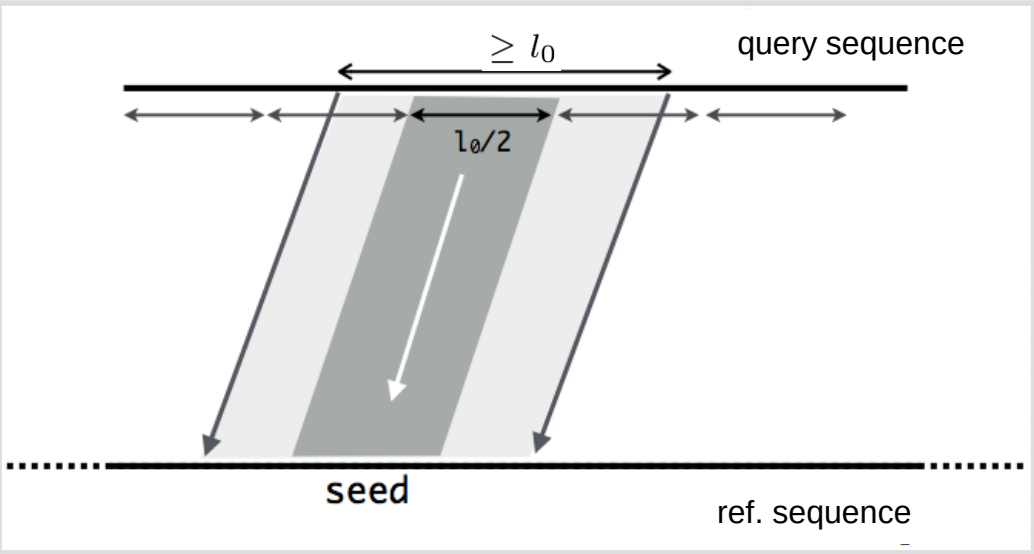
PROPOSED ALGORITHM

Georgia Tech Institute for Data Engineering and Science

Input: query, reference, thresholds for local alignment (min. identity, min. length l_0)

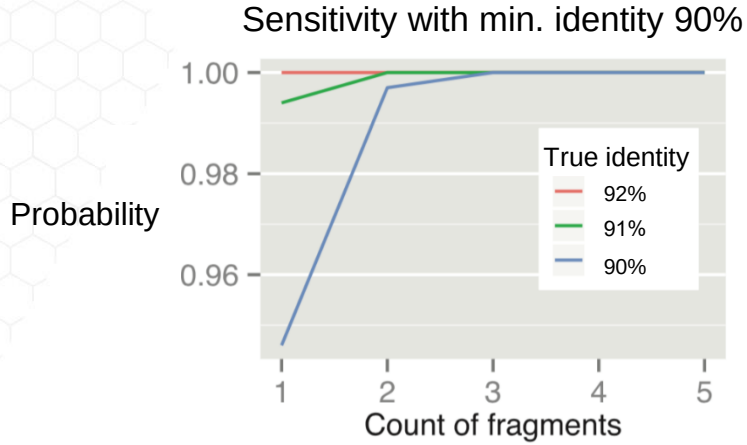
1

- continue to assume the same error model
- split query into $l_0/2$-sized non-overlapping fragments
- map each one using Mashmap routine



2

- high probability of reporting at-least one seed mapping along the optimal alignment



3

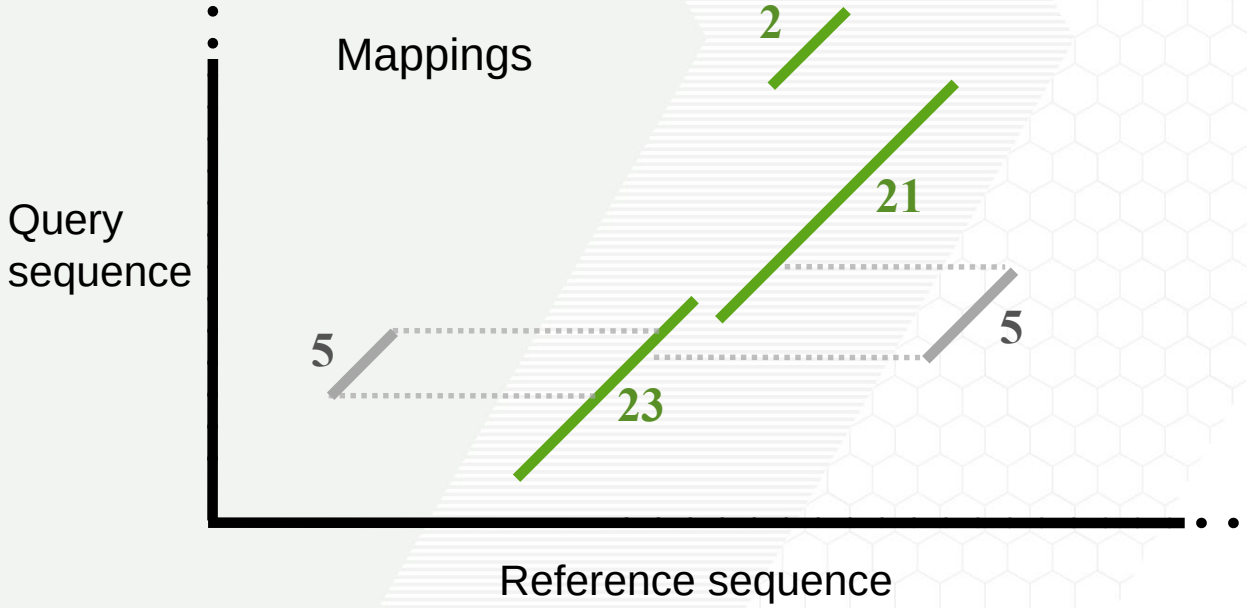
- merge adjacent fragment mappings
- score each mapping by length x identity estimate

FILTERING HEURISTIC

Georgia Tech Institute for Data Engineering and Science

- For analysis of most promising mappings (e.g., orthologs, paralogs)
- Requires extensive filtering in mammalian genomes with high-copy repeats.

Formulation: Discard any mapping segment that is subsumed by higher scoring segments



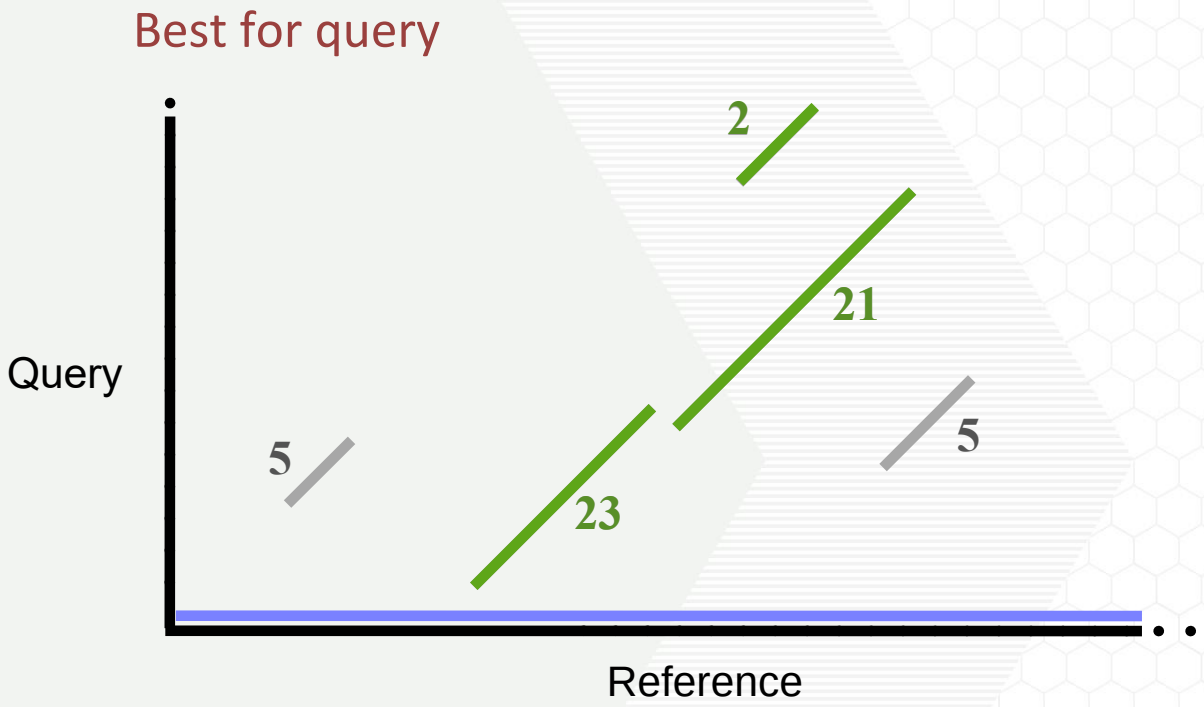
Naive algorithm : $O(n^2)$

Can we do better?

Yes, in $O(n \log n)$ time

Plane-sweep technique
[Shamos and Hoey 1976]

PLANE-SWEEP BASED FILTERING

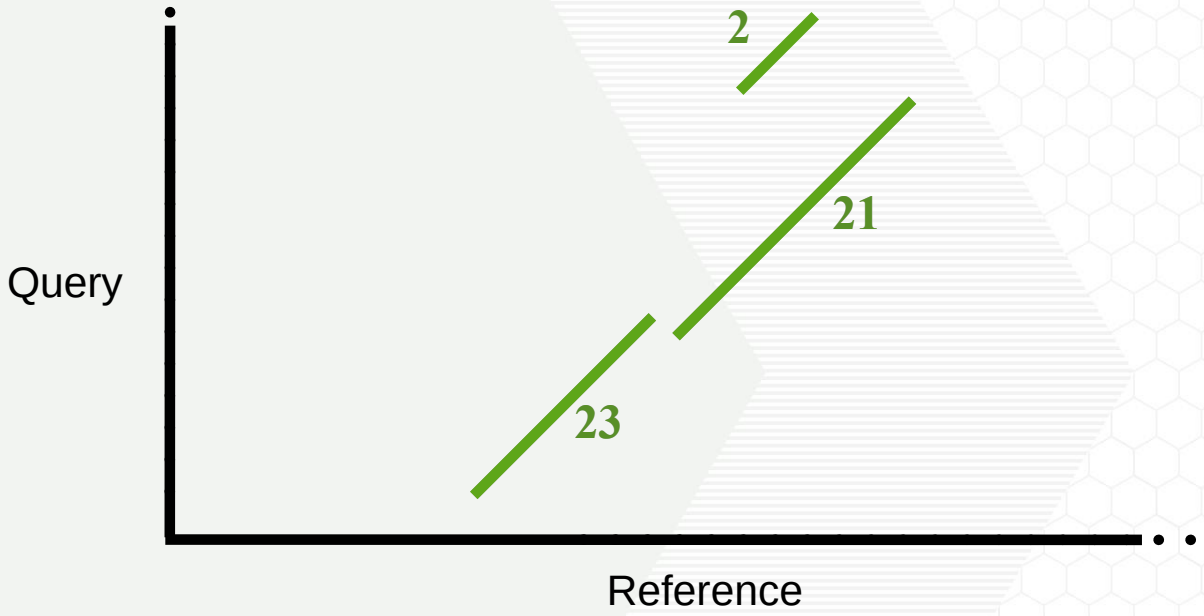


- conceptually sweep a horizontal line from bottom to top

PLANE-SWEEP BASED FILTERING

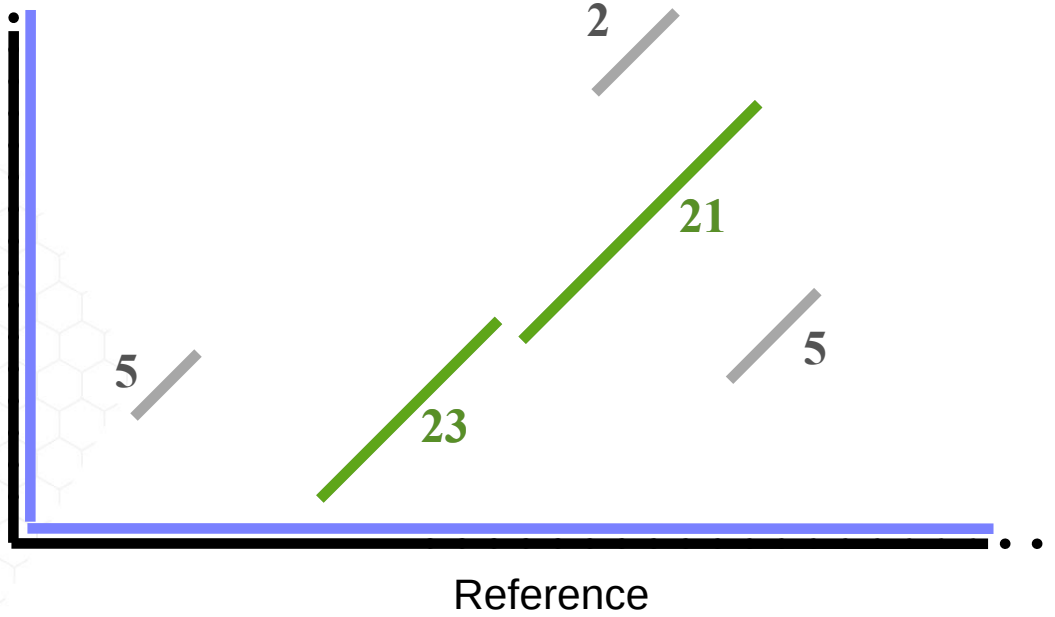
Georgia Tech Institute for Data Engineering and Science

Best for query



➢ conceptually sweep a horizontal line from bottom to top

Best for query and ref. (orthologs only)



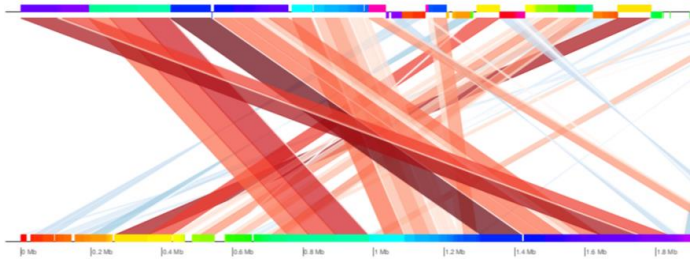
➢ useful for filtering orthologs, paralogs…

- process intersecting segments efficiently to achieve worst-case $O(n \log n)$ bound
- it is easily shown that this alignment filtering algorithm is optimal

WHOLE-GENOME ALIGNMENTS

Georgia Tech Institute for Data Engineering and Science

- Fundamental problem in bioinformatics
- Applications:
  - discover variants
  - detect evolutionarily conserved segments
  - identify large-scale chromosomal rearrangements
  - validate genome assemblies
- Exponential increase in publicly available genome assemblies
  - need faster and memory-efficient algorithms
  - popular tools such as LAST and Nucmer still take >10 CPU hours to compare two human genomes



Source: genoPlotR

VERTEBRATE GENOMES PROJECT

Target: genomes of 66,000 species

Genomics england

Target: 100,000 human genomes

EXPERIMENT-I

Georgia Tech Institute for Data Engineering and Science

Datasets

		Query sequences	Ref. genome	(Truth assumed)
Intra-species	D1	E. coli O157 genome	E. coli K12	Nucmer4 alignments
	D2	NA12878 human genome assembly (polished)	human (hg38)	
	D3	NA12878 human genome assembly		
Inter-species	D4	human genome	gorilla (gorGor5)	UCSC genome browser (BLASTZ)
	D5	chimp genome		

Parameters

- Mashmap2 : alignment length: ≥ 10 Kbp, identity: ≥ 95% (intra), ≥ 90% (inter)
- Minimap2 : *-x asm5*
- Nucmer4 : [default] followed by delta-filter with '-1'

[Li, 2018]
[Marçais et al., 2018]

RUNTIME / MEMORY-USE

Georgia Tech Institute for Data Engineering and Science

Performance comparison

	Mashmap2		Minimap2		Nucmer4	
	Time	Memory	Time	Memory	Time	Memory
D1	0.5s	16M	0.4s	85M	5.2s	138M
D2	1m 26s	4G	3m 3s	17G	5h 01m	53G
D3	6m 33s	4G	3m 11s	16G	2h 10m	53G
D4	27m 33s	9G	15m 06s	27G	33h 04m	57G
D5	25m 40s	8G	5m 54s	26G	24h 58m	56G

Takeaways

- Mashmap2 uses much less memory than other tools
- Alignment-free algorithms yield orders of magnitude speedup

Accuracy

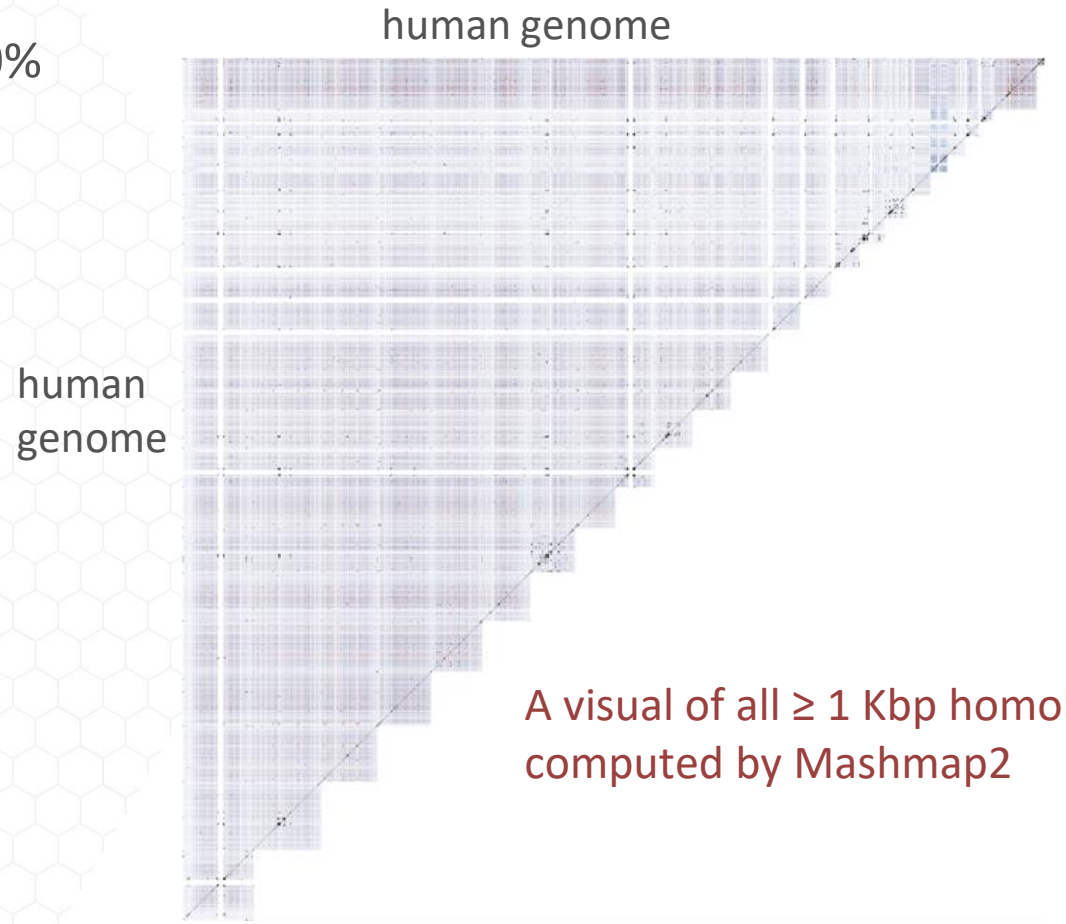
- >97% sensitivity (recall) on all datasets
- Mashmap2's precision (fraction of candidate mappings satisfying thresholds): 35% — 76%.

EXPERIMENT-II: HUMAN V. HUMAN

Georgia Tech Institute for Data Engineering and Science

- Exhaustive search for all ≥ 1 Kbp repeats with ≥ 90% identity in the human genome

- Duplications in genome have implications in
  - Genome evolution and stability
  - Genetic diseases



A visual of all ≥ 1 Kbp homologs computed by Mashmap2

EXPERIMENT-II: HUMAN V. HUMAN

Georgia Tech Institute for Data Engineering and Science

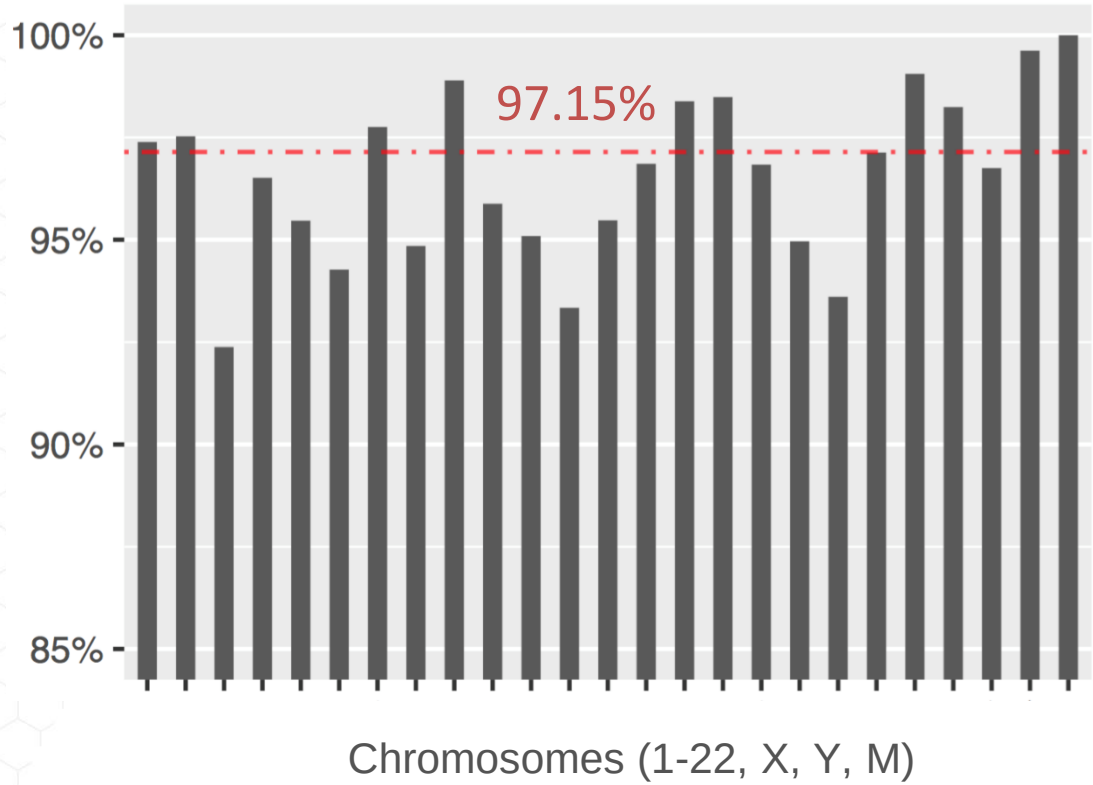
human (hg38) v. human (hg38)

Mashmap2
[≥ 1 Kbp, ≥ 90% identity]
Filter: off

Validation using LAST

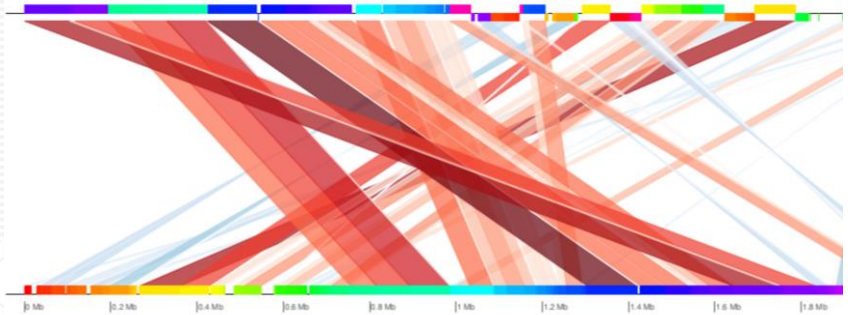
210 million duplications
(10.3% coverage on human genome)

Sensitivity w.r.t. UCSC seg-dup database



2. FAST-ANI: SCALABLE WHOLE-GENOME DISTANCE COMPUTATION WITH 90,000 MICROBIAL GENOMES

Genome A



Genome B

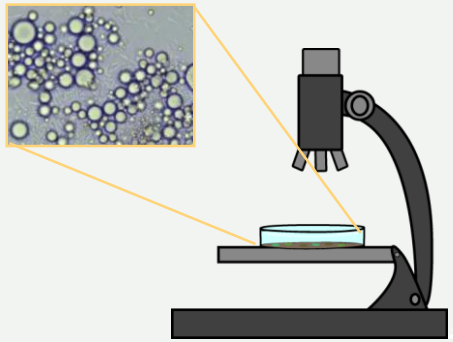
Source: genoPlotR

FAST WHOLE-GENOME DISTANCE COMPUTATION

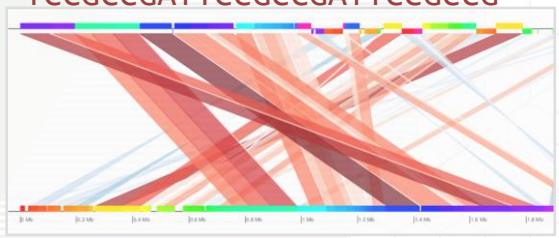
Georgia Tech Institute for Data Engineering and Science

EUKARYOTIC TAXONOMY

PROKARYOTIC TAXONOMY



TCCGCCGATTCCGCCGATTCCGCCG



(use DNA sequences) ✓

ANI: Average Nucleotide Identity
defined as mean alignment identity of conserved genes in the two given genomes

- Popular genome-distance method in microbiology (for bacteria/archaea)
- Useful for taxonomic classification and computing evolutionary distances

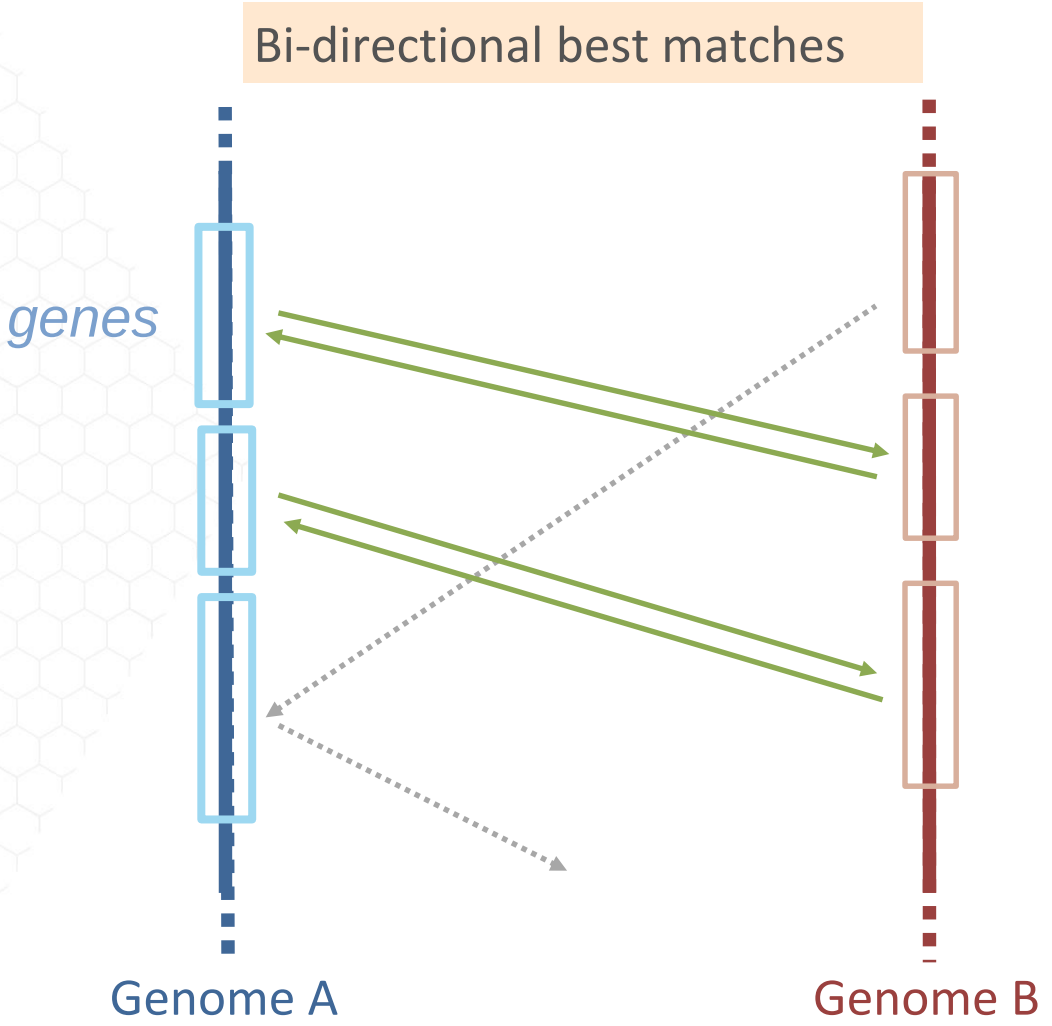
- Existing alignment-based methods don't scale to current public genome databases
- The databases are now big and continue to grow fast
  - isolates
  - metagenomics (MAGs)
  - single-cell genomics (SAGs)

FASTANI ALGORITHM

Georgia Tech Institute for Data Engineering and Science

ANI: Average Nucleotide Identity
[mean alignment identity of conserved genes in given genomes]

- Main computational bottleneck -> mapping sequences to compute conserved genomic segments
- Implement standard heuristics and replace BLAST with our alignment-free mapping framework
- Target:
  ANI within 80%-100% identity range

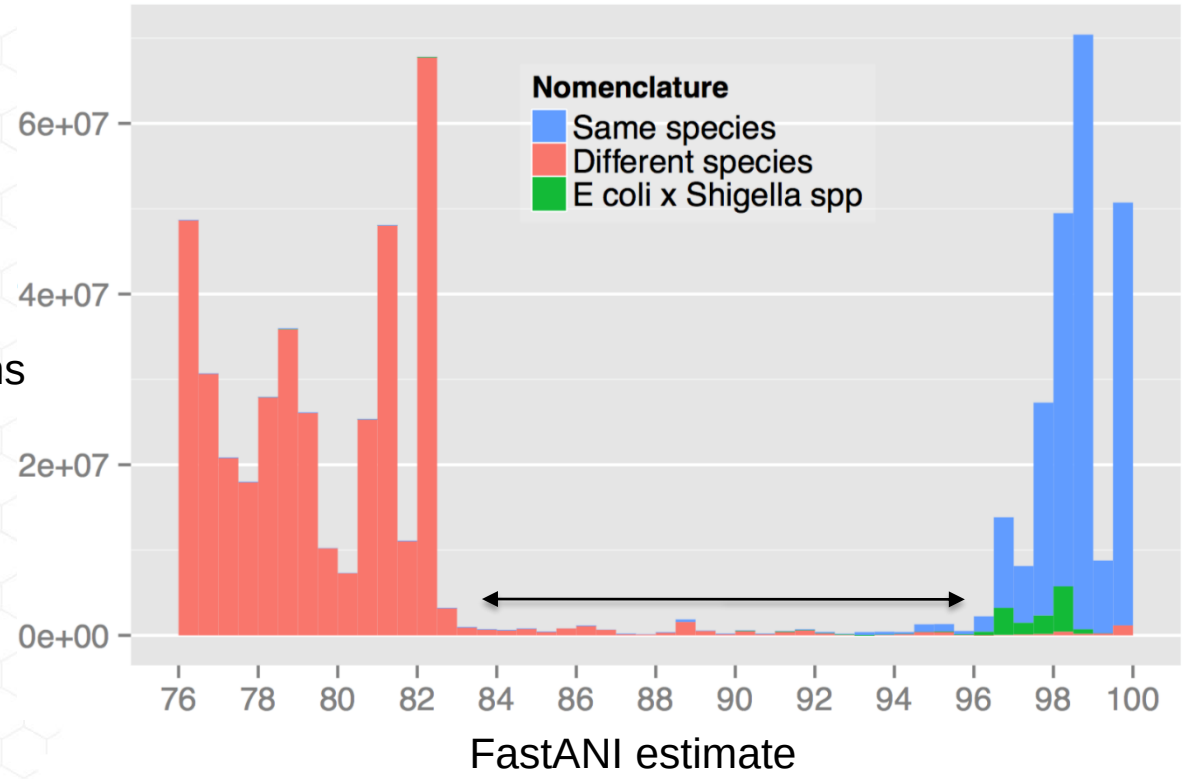


DO SPECIES BOUNDARIES EXIST?

Georgia Tech Institute for Data Engineering and Science

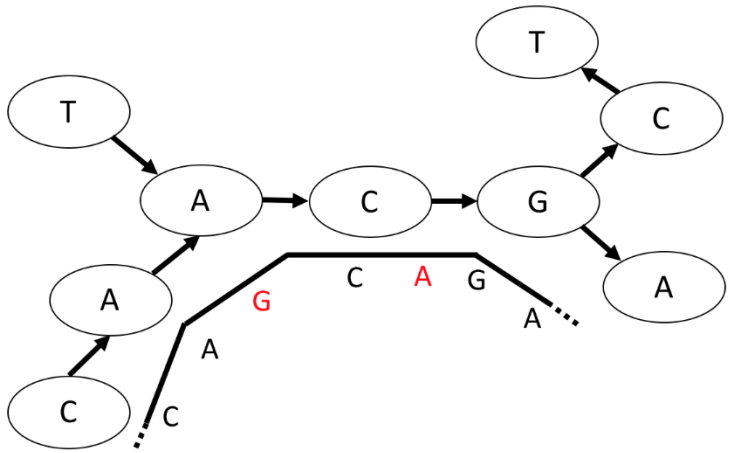
- Scalable to perform all-to-all pairwise comparison among **ALL** 90K genomes in GenBank
- Existence of clusters (species) in prokaryotes is an open and widely debated question
- We show wide species boundaries using 4 billion genome comparisons
    - consistent with small-scale studies [Kim et al. 2014]

Distribution of ANI values among 90,000 genomes



[Jain et al. Nature Comm. 2018 (in press)]

PARALLEL SEQUENCE TO GRAPH ALIGNMENT

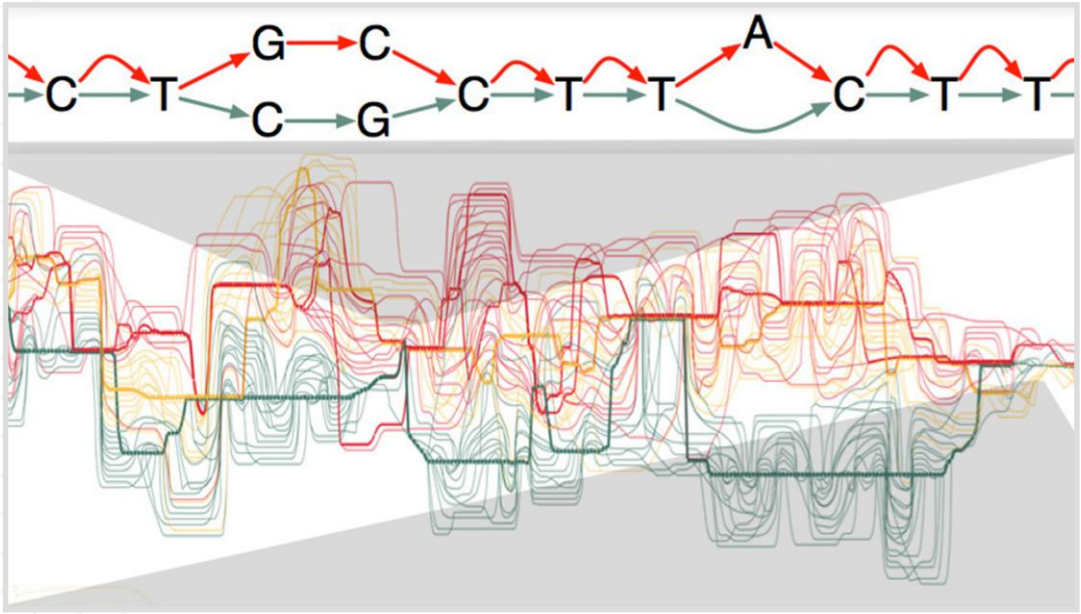


WHY MAP TO GRAPHS?

Georgia Tech Institute for Data Engineering and Science

Graphs are common in genomics:

- Multi-genome reference
  - Variation graphs
- Assembly
  - de Bruijn graphs
  - Overlap graphs
- RNA-seq analysis
  - Splicing graphs



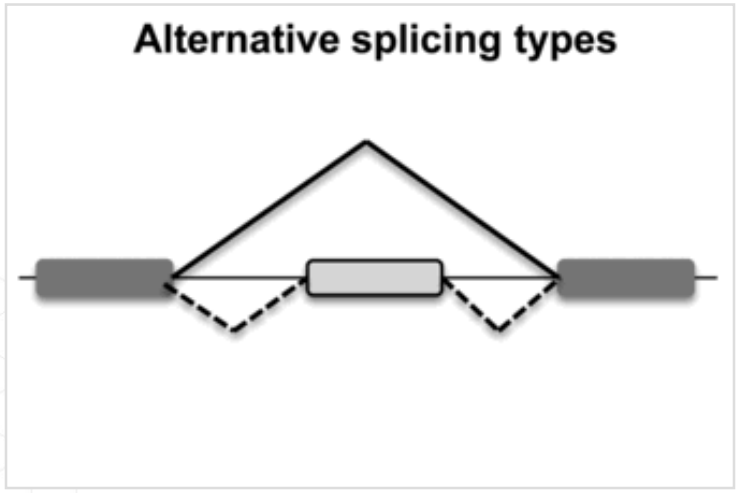
Source: [Genome graphs blog]

WHY MAP TO GRAPHS?

Genotyping using population reference genomes

- Improve mappability
- Detect unknown SNPs, SVs

[Dilthey et al. 2015 Nature Genetics]



Source: Zhang et. al. 2014

RNA-seq alignment to splicing graph

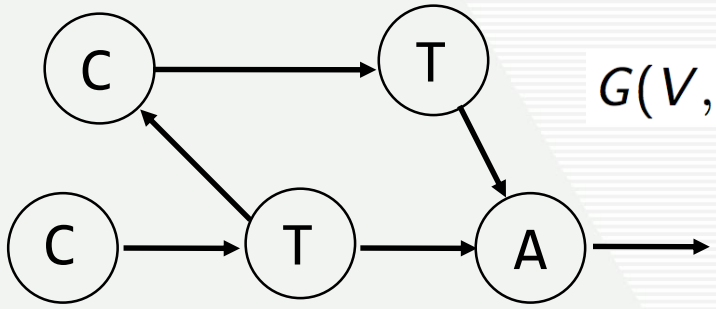
- compute gene expression
- detect novel splicing events w.r.t. reference

[Kuosmanen et al. 2017 Brief. in Bioinf.]

PROBLEM FORMULATION

Georgia Tech Institute for Data Engineering and Science

DAG



$G(V, E)$

READ

Long read of length m

...GCCCCGCCGATCCGCCT...

Local Sequence Alignment to Graph:

Identify a path in the graph s.t. its optimal alignment score with a substring of the read is maximum

Sequential time
(Extension of Smith-Waterman to DAGs)

$O(m\,(|V| + |E|))$ time and $O(|V|)$ memory

[Novarro 2000]

✓ Theoretical guarantee ✗ Compute intensive

OUTCOMES

Georgia Tech Institute for Data Engineering and Science

Algorithmic ideas:

- First parallel algorithm to utilize multi-core SIMD processors
- A three-stage algorithm to keep memory-usage low
- Leverages inter-task parallelism

Key results:

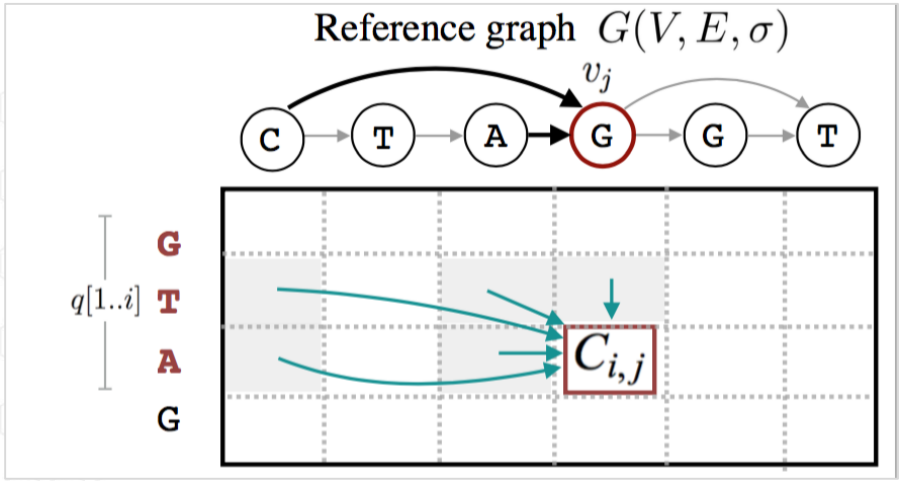
- Makes it possible to optimally align high coverage long read data-sets to large variant graphs (e.g., MHC, LRC)
- Runtime in the order of few minutes or hours; not feasible with prior algorithms.

CHALLENGES

Georgia Tech Institute for Data Engineering and Science

$$C_{0,j} = 0$$

$$C_{i,j} = \max \begin{cases} 0 & \\ \Delta_{i,j} & \\ C_{i-1,k} + \Delta_{i,j} & \forall k : (v_k, v_j) \in E \\ C_{i,k} - \Delta_{ins} & \forall k : (v_k, v_j) \in E \\ C_{i-1,j} - \Delta_{del} & \end{cases}$$



- Number and structure of dependencies
- Existing parallel algorithms for local sequence-to-sequence alignment are either inapplicable or inefficient

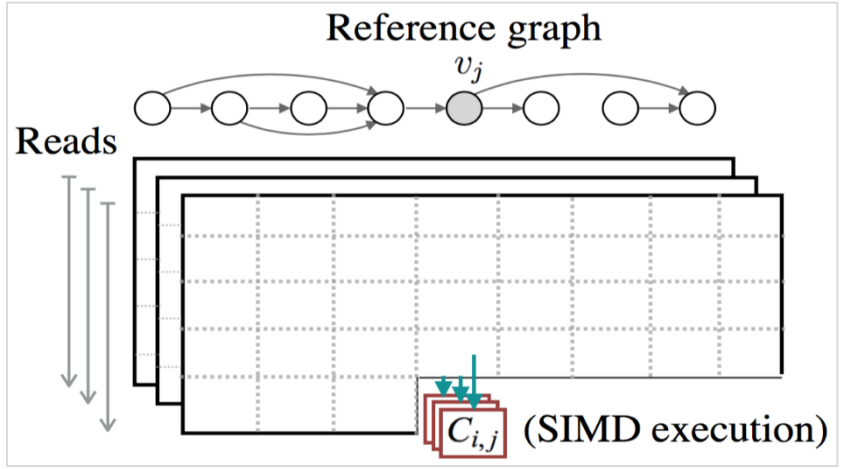
PASGAL: KEY IDEAS

Georgia Tech Institute for Data Engineering and Science

1. **DP-fwd**
   Output: score, ending position and read strand of an optimal alignment for each read
2. **DP-rev**
   Output: optimal alignments' starting positions
3. **Traceback**
   Output: base-to-base alignments

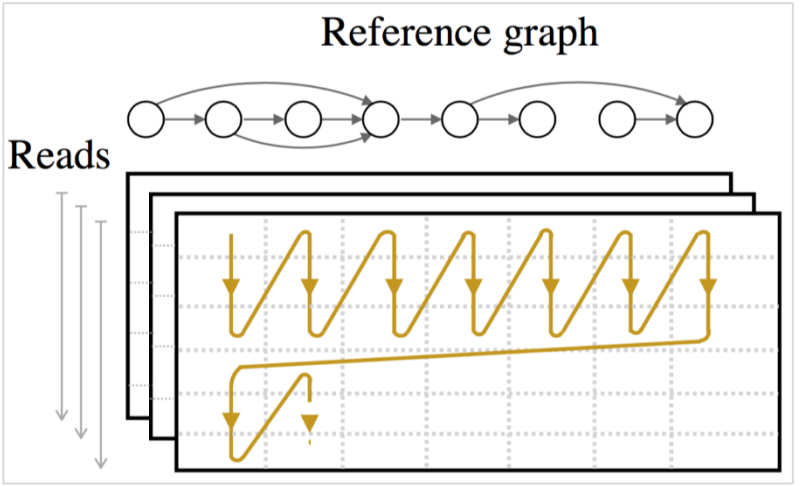
1. Three stages of the parallel algorithm for **low-memory usage** [Huang et al. 1990]

2. **Vectorization** (scope of 64x speedup with wide SIMD lanes)



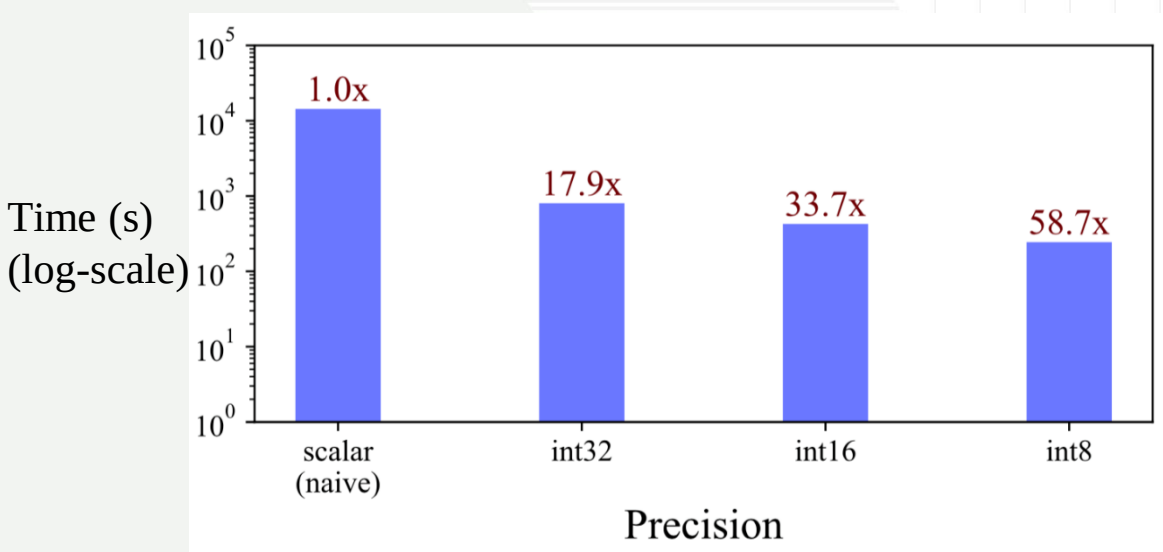
3. Blocked computation instead of row-by-row for better **memory locality**

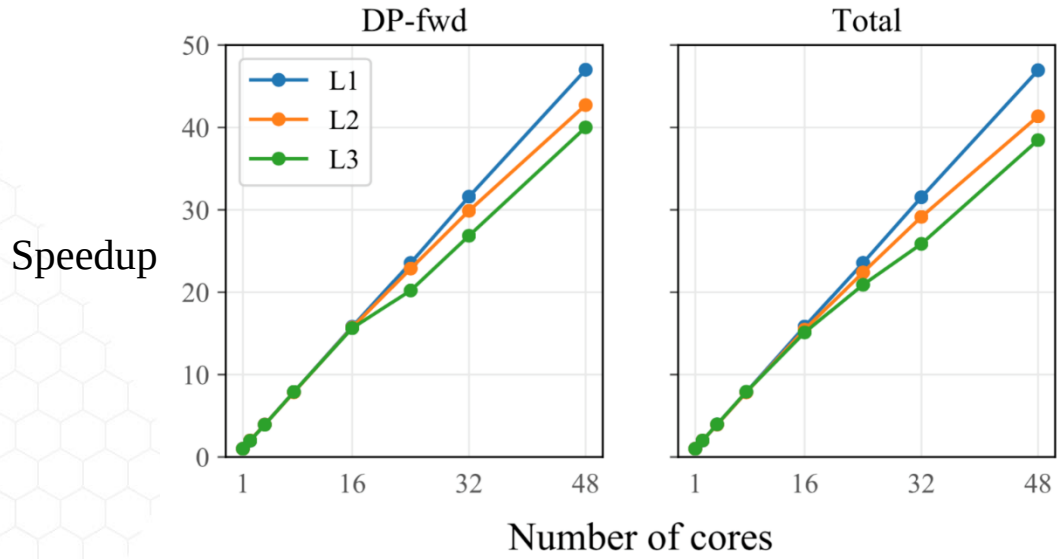
Blocked approach: majority of vertices have 'near' neighbors (SNPs, indels)



RESULTS

Georgia Tech Institute for Data Engineering and Science





- 58.7x speedup using vectorization and better memory locality on single core
- Strong-scaling: near linear speedup using 48 cores

- Alignment of simulated long reads (10x coverage) to MHC variation graph [Dilthey et al. 2016]
  - Time to output base-to-base alignments: < 4 hours (takes multiple days with existing algorithms)
  - Peak performance: 317 billion cell updates per second (GCUPS)

COMPARISONS

Georgia Tech Institute for Data Engineering and Science

- Achieve superior runtime and accuracy than previous exact and heuristic methods respectively.
- Data sets:
  - Variation graph: Leukocyte Receptor Complex (LRC) segment in human genome (|V|, |E| = 1 M)
  - Short read set : L1' (100 bp reads)
  - Long read sets : L2' (mean length = 10 Kbp), L3' (mean length = 25 Kbp)
- Comparison against **exact** algorithms
  - vg [exact]
  - Graphaligner
  - Single threaded execution
- Comparison against **heuristic** algorithm
  - vg [heuristic]

➢ Speedups using PaSGAL

Reference graph	LRC		
Read set	L1′	L2′	L3′
vs. Graphaligner	10.7x	4.2x	3.0x
vs. vg-exact	25.3x	13.3x	-

➢ Memory-usage

Reference graph	LRC		
Read set	L1′	L2′	L3′
PaSGAL (GB)	0.2	0.3	0.8
Graphaligner (GB)	1.1	1.1	1.1
vg-exact (GB)	0.7	108.8	-

➢ Output accuracy using vg (heuristic)

Read set	L1′	L2′	L3′
Fraction of alignments with $> 5\%$ diff. from optimal score (%)	0.04	24.53	100
Fraction of alignments with $> 20\%$ diff. from optimal score (%)	0.00	9.40	100

ACKNOWLEDGEMENTS

Georgia Tech Institute for Data Engineering and Science

Georgia Tech Computational Science and Engineering

- Chirag Jain
- Haowen Zhang

NIH National Human Genome Research Institute

- Alexander Dilthey
- Sergey Koren
- Adam Phillippy

Georgia Tech Civil and Environmental Engineering

- Prof. Kostas Konstantinidis
- Miguel Rodriguez

intel Parallel Computing Lab

- Sanchit Misra

NSF Funding

- IIS-1416259, CCF-1816027