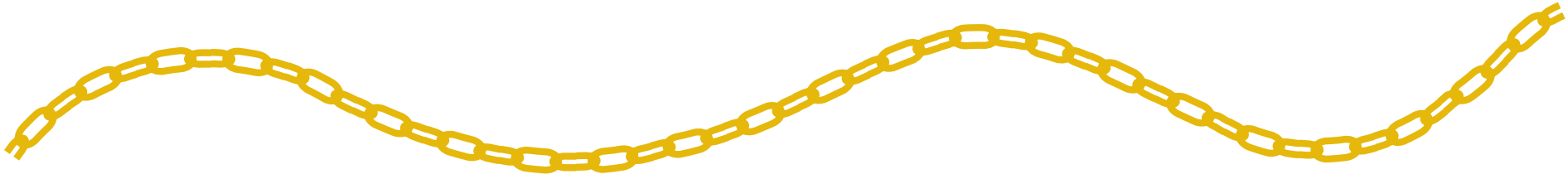
Chaining for Accurate Alignment of Erroneous Long Reads to Acyclic Variation Graphs

Jun Ma, **Manuel Cáceres**, Leena Salmela,
Veli Mäkinen and Alexandru I. Tomescu

June 2022, DSB



Problem & Motivation

Read-to-Pangenome Alignment

Given

- A Pangenome
- A Read

Report

- The *position of the Read* in the Pangenome

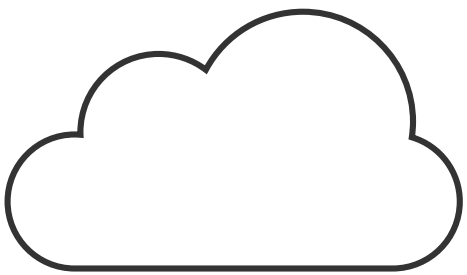
Read-to-Pangenome Alignment

Given

- A Pangenome
- A Read

Report

- The *position of the Read* in the Pangenome



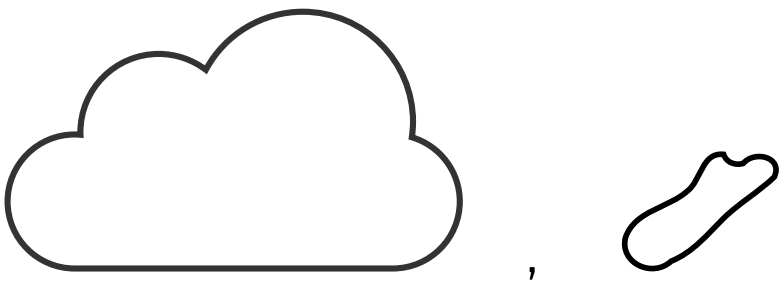
Read-to-Pangenome Alignment

Given

- A Pangenome
- A Read

Report

- The *position of the Read* in the Pangenome



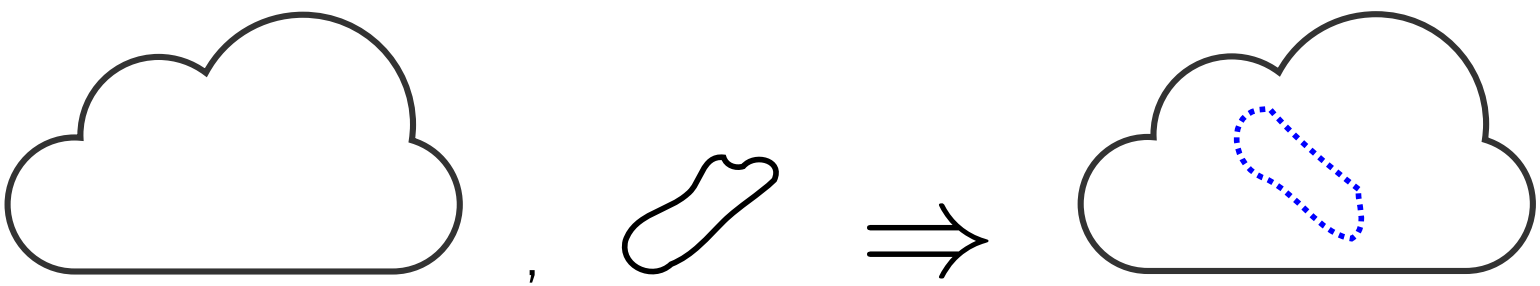
Read-to-Pangenome Alignment

Given

- A Pangenome
- A Read

Report

- The *position of the Read* in the Pangenome



Sequence-to-Graph (S2G) Alignment

Given

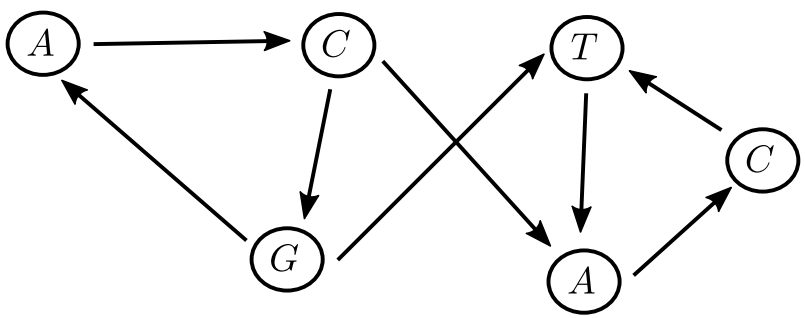
- A graph: $G = (V, E)$
- A function labeling G: $\ell : V|E \to \Sigma|\Sigma^*$

} (Pangenome)

- A string: $S \in \Sigma^m$

Report

- A walk P of G such that $\ell(P)$ *minimizes its distance* to S



$CGTCCCTACT$

Sequence-to-Graph (S2G) Alignment

Given

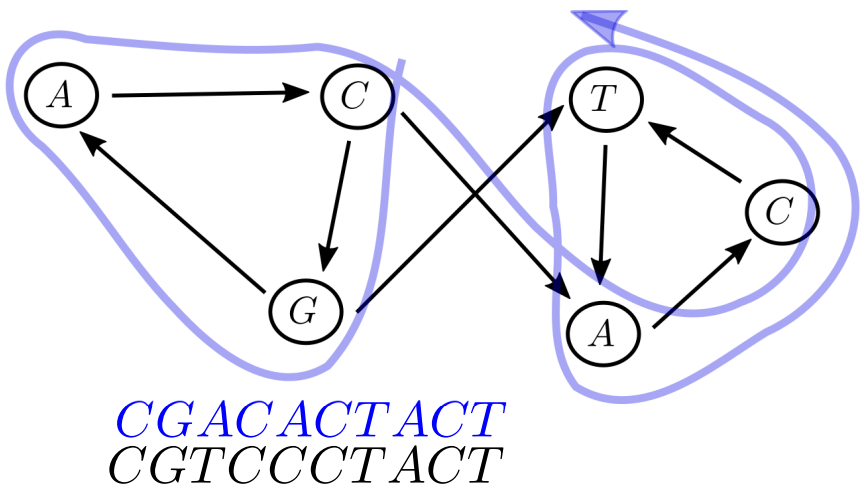
- A graph: $G = (V, E)$
- A function labeling G: $\ell : V|E \to \Sigma|\Sigma^*$

} (Pangenome)

- A string: $S \in \Sigma^m$

Report

- A walk P of G such that $\ell(P)$ *minimizes its distance* to S



Sequence-to-Graph (S2G) Alignment

Given

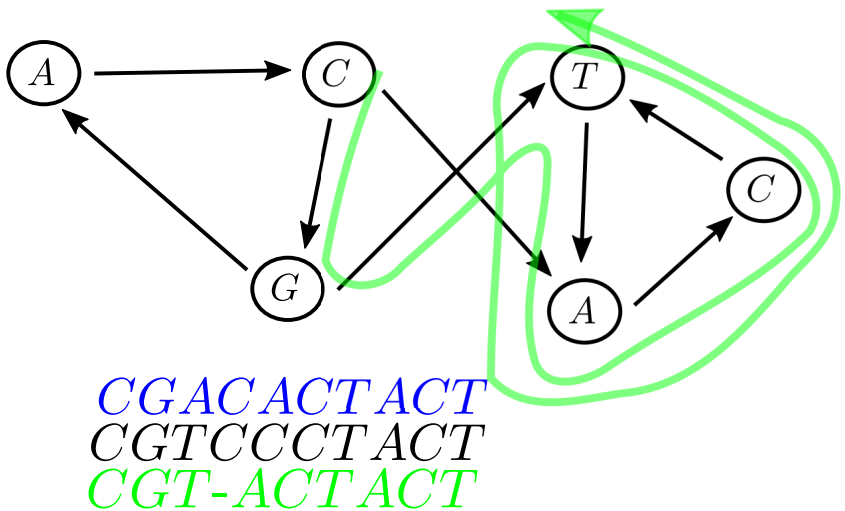
- A graph: $G = (V, E)$
- A function labeling G: $\ell : V|E \rightarrow \Sigma|\Sigma^*$

} (Pangenome)

- A string: $S \in \Sigma^m$

Report

- A walk P of G such that $\ell(P)$ *minimizes its distance* to S



State-of-the-art in S2G Alignment

- Allow edits in $\ell \Rightarrow$ NP-hard [8]

State-of-the-art in S2G Alignment

- Allow edits in $\ell \Longrightarrow$ NP-hard [8]
- Allow edits only in S or $\ell(P) \Longrightarrow$
  - $O(|V| + m|E|)$ [8]
  - No polynomial improvements under SETH [1, 3]

State-of-the-art in S2G Alignment

- Allow edits in $\ell \Rightarrow$ NP-hard [8]
- Allow edits only in S or $\ell(P) \Rightarrow$
  - $O(|V| + m|E|)$ [8]
  - No polynomial improvements under SETH [1, 3]

Quadratic running time still **insufficient** for applications, e.g., pangenomes

State-of-the-art in S2G Alignment

- Allow edits in $\ell \Longrightarrow$ NP-hard [8]
- Allow edits only in S or $\ell(P) \Longrightarrow$
  - $O(|V| + m|E|)$ [8]
  - No polynomial improvements under SETH [1, 3]

Quadratic running time still **insufficient** for applications, e.g., pangenomes

- PO alignment, CDSs, Minimizers, Snarls, Seed-and-extend, Parallel computation, Bit-parallel, Improved graph search, ...
  - `vg` [4, 13], `PaSGAL` [7], `AStarix` [5, 6], `GraphAligner` [11]

State-of-the-art in S2G Alignment

- Allow edits in $\ell \Rightarrow$ NP-hard [8]
- Allow edits only in S or $\ell(P) \Rightarrow$
  - $O(|V| + m|E|)$ [8]
  - No polynomial improvements under SETH [1, 3]

Quadratic running time still **insufficient** for applications, e.g., pangenomes

- PO alignment, CDSs, Minimizers, Snarls, Seed-and-extend, Parallel computation, Bit-parallel, Improved graph search, ...
  - `vg` [4, 13], `PaSGAL` [7], `AStarix` [5, 6], `GraphAligner` [11]

... For chromosome-level graphs, and long and erroneous reads ...

- `GraphAligner` [11]

The Challenge

The graph:

- chromosome-level: Cannot afford to find optimal alignment

The Challenge

The graph:

- chromosome-level: Cannot afford to find optimal alignment

The reads:

- long: Most tools are tailored to short reads
- erroneous: More solutions, location versus distance

The Challenge

The graph:

- chromosome-level: Cannot afford to find optimal alignment

The reads:

- long: Most tools are tailored to short reads
- erroneous: More solutions, location versus distance

`GraphAligner` [11], fast and accurate tool based on seed-and-extend

- **Find** seed hits with minimizers
- **Extend** seed hits (clusters) with banded and bit-parallel DP

The Challenge

The graph:

- chromosome-level: Cannot afford to find optimal alignment

The reads:

- long: Most tools are tailored to short reads
- erroneous: More solutions, location versus distance

`GraphAligner` [11], fast and accurate tool based on seed-and-extend

- **Find** seed hits with minimizers
- **Extend** seed hits (clusters) with banded and bit-parallel DP

... but the accuracy can still be improved

Our Rationale

Replace the "seed-and-extend" strategy by co-linear chaining on acyclic variation graphs, while still being fast

Our Contributions

- Algorithm to solve co-linear chaining on string-labeled acyclic graphs
  - Linearithmic running time parameterized in the width k
  - Acyclic variation graphs of human chromosomes have small k
  - Allows one-vertex overlaps

Our Contributions

- Algorithm to solve co-linear chaining on string-labeled acyclic graphs
  - Linearithmic running time parameterized in the width k
  - Acyclic variation graphs of human chromosomes have small k
  - Allows one-vertex overlaps

Implemented into ...

- `GraphChainer`: A new S2G aligner built on top of `GraphAligner`
  - Acyclic graphs
  - Within the capabilities of modern high-performance computer
  - Significantly more accurate

Our Contributions

- Algorithm to solve co-linear chaining on string-labeled acyclic graphs
  - Linearithmic running time parameterized in the width k
  - Acyclic variation graphs of human chromosomes have small k
  - Allows one-vertex overlaps

Implemented into ...

- `GraphChainer`: A new S2G aligner built on top of `GraphAligner`
  - Acyclic graphs
  - Within the capabilities of modern high-performance computer
  - Significantly more accurate

Our experiments show that ...

- `GraphChainer` aligns $95 - 99\%$ of simulated and real PacBio reads
- `GraphChainer` aligns $\geq 12\%$ more real reads than `GraphAligner`

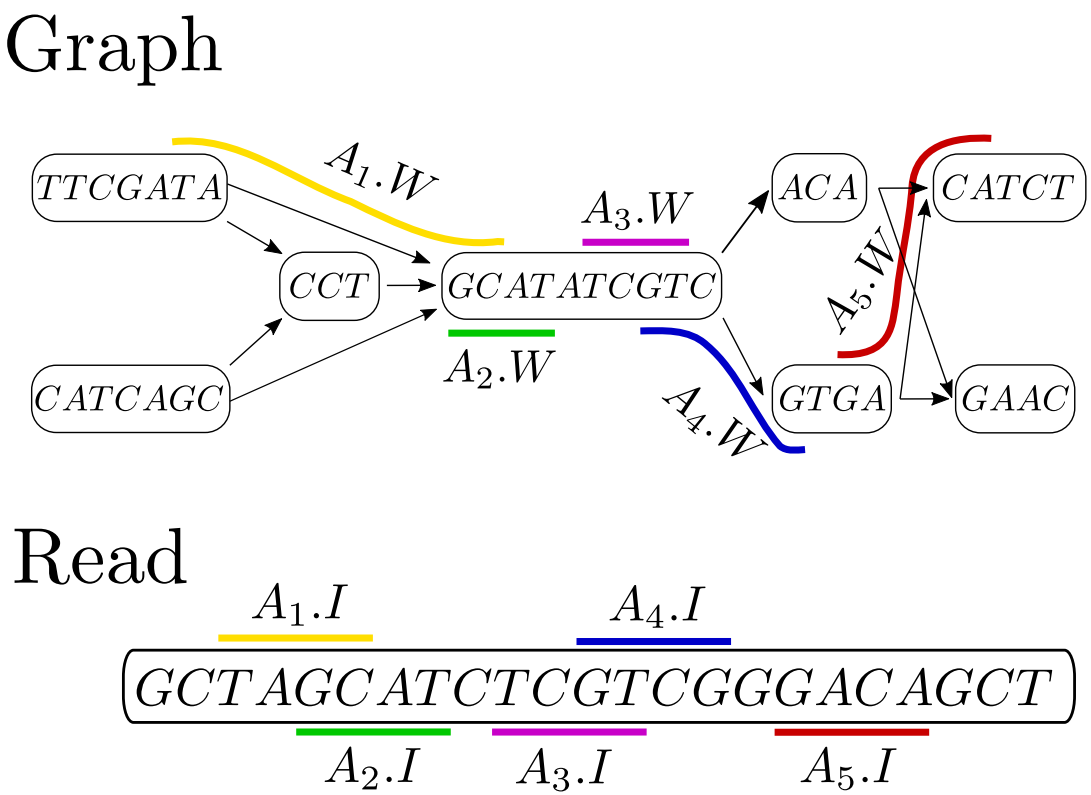
Co-Linear Chaining (CLC)

CLC

- Originally defined between two sequences
  - `minimap2` [9], `uLTRA` [12]
- A principled approach to capture the global relation between seed hits
  - `GraphAligner` [11, p.16]

CLC

- Originally defined between two sequences
  - `minimap2` [9], `uLTRA` [12]
- A principled approach to capture the global relation between seed hits
  - `GraphAligner` [11, p.16]



CLC in DAGs

- Solved in $O(k(N \log N + (|V| + |E|) \log |V|))$ by Mäkinen et al. [10]
  - For character-labeled DAGs and no path overlaps allowed

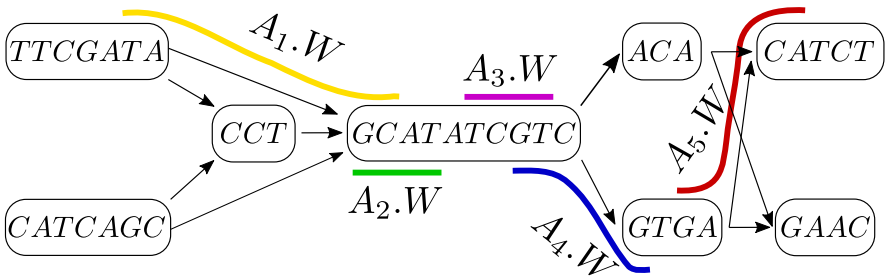
CLC in DAGs

- Solved in $O(k(N \log N + (|V| + |E|) \log |V|))$ by Mäkinen et al. [10]
  - For character-labeled DAGs and no path overlaps allowed

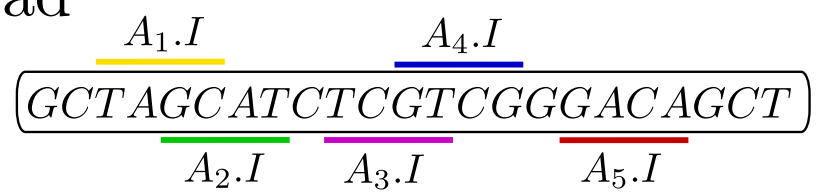
We build on the previous solution [10]

- Decouple to $O(k^3|V| + k|E|)$ pre-processing and $O(kN \log kN)$
- String-labeled DAGs and one-vertex overlaps allowed

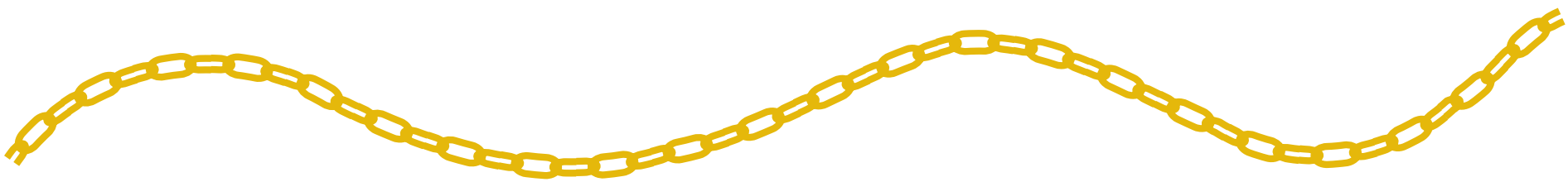
Graph



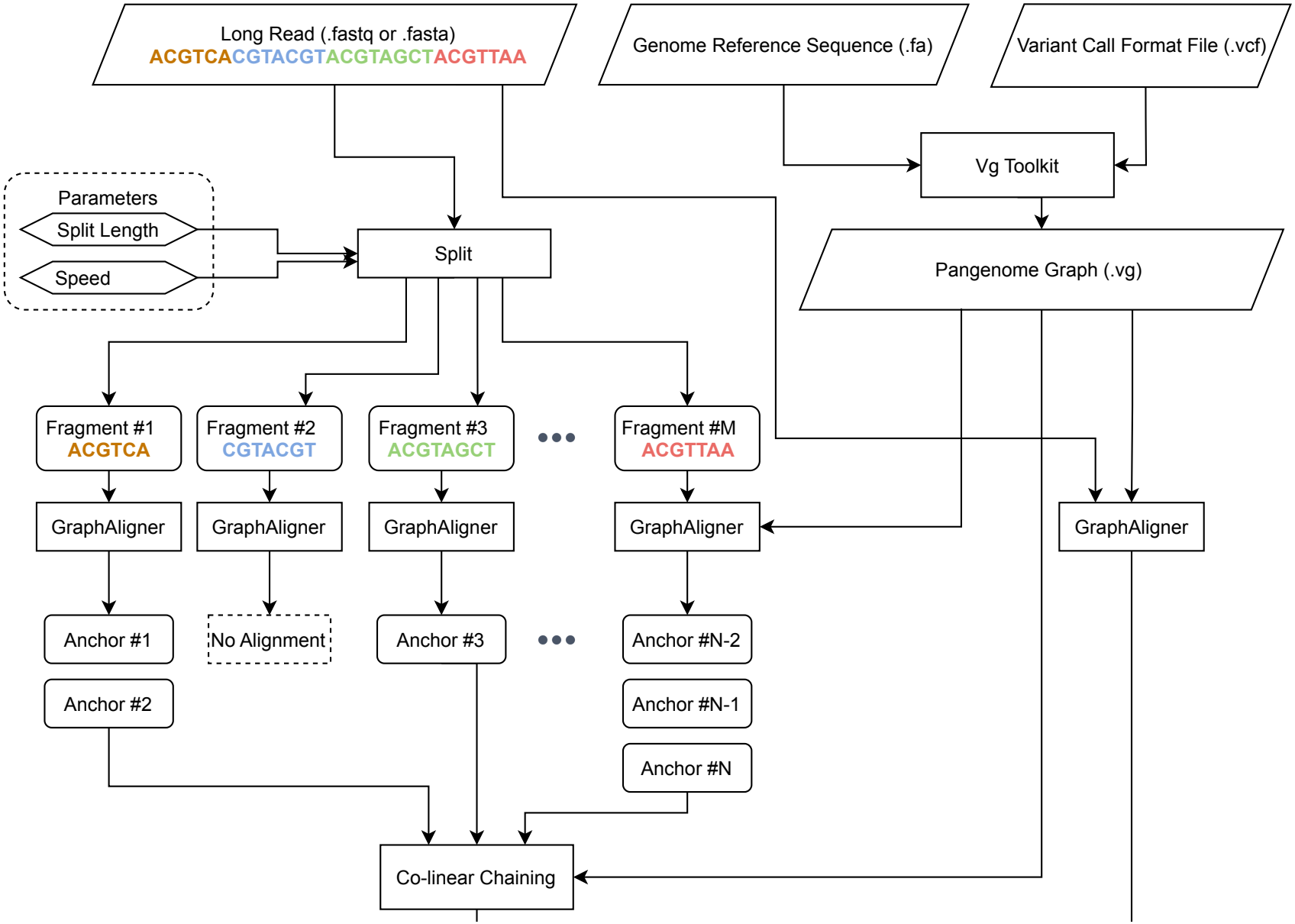
Read



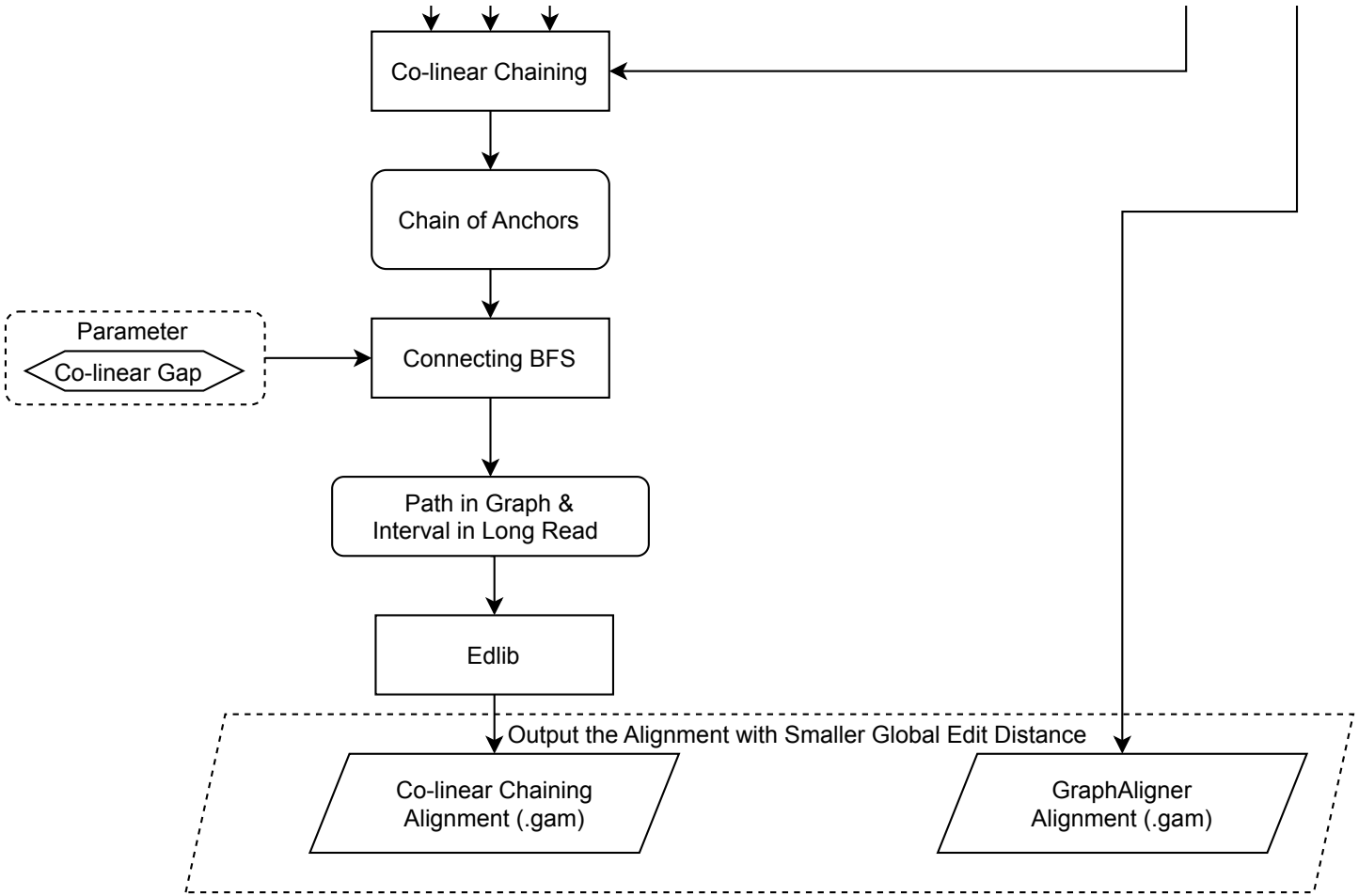
GraphChainer



GraphChainer



GraphChainer



Available at: `github.com/algbio/GraphChainer`

Experiments

Datasets

Variation graphs

- LRC and MHC1 [7]
  - Diverse regions in human genome
- Chr22 and Chr1
  - Built with vg using GRCh37 and variants from 1000Gen [2]

Datasets

Variation graphs

- LRC and MHC1 [7]
  - Diverse regions in human genome
- Chr22 and Chr1
  - Built with `vg` using GRCh37 and variants from 1000Gen [2]

Reads

- Simulated
  - Random path of G + `Badread` [14] (PacBio 15% error rate)
- Real (only Chr)
  - Whole human genome + filter with `minimap2` [9]

Datasets

Graph	#Nodes	Labels bps	k	#Reads	Total read bps	Cov
LRC	117 787	1 099 856	4	1 093	15 872 214	15×
MHC1	479 531	5 138 362	4	5 091	74 524 274	15×
Chr22	3 197 160	52 423 213	7	52 464	769 238 818	15×
real				136 494	2 858 621 416	56×
Chr1	18 807 963	255 754 179	9	254 251	3 736 803 386	15×
real				907 572	19 617 046 919	79×

Metrics

For simulated reads (overlap criterion δ)

- Correctly aligned if overlaps at least $(100 \cdot \delta)\%$ with the ground truth

Metrics

For simulated reads (overlap criterion δ)

- Correctly aligned if overlaps at least $(100 \cdot \delta)\%$ with the ground truth

For real reads (distance criterion σ)

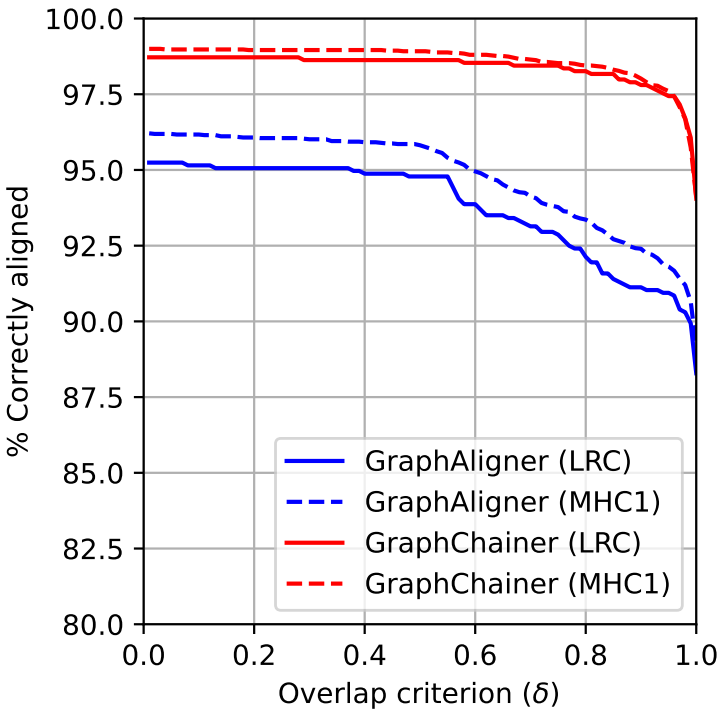
- No ground truth available
- Correctly aligned if its edit distance with the read is at most $(100 \cdot \sigma)\%$ of its length

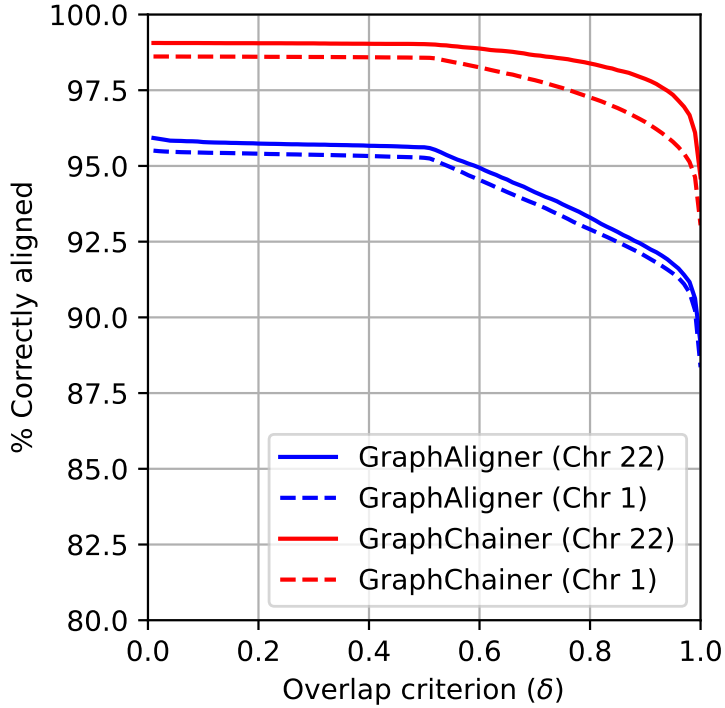
Experimental Results

Accuracy – Simulated Reads, $\delta = 0.1, 0.85$

Graph	Aligner	Correctly aligned	
		$\delta = 0.1$	$\delta = 0.85$
LRC	`GraphAligner`	95.15%	91.22%
	`GraphChainer`	98.72% (+3.75%)	97.90% (+7.32%)
MHC1	`GraphAligner`	96.15%	92.52%
	`GraphChainer`	98.98% (+2.94%)	98.17% (+6.11%)
Chr22	`GraphAligner`	95.78%	92.56%
	`GraphChainer`	99.06% (+3.42%)	98.00% (+5.88%)
Chr1	`GraphAligner`	95.44%	92.26%
	`GraphChainer`	98.61% (+3.32%)	96.68% (+4.79%)

Accuracy – Simulated Reads

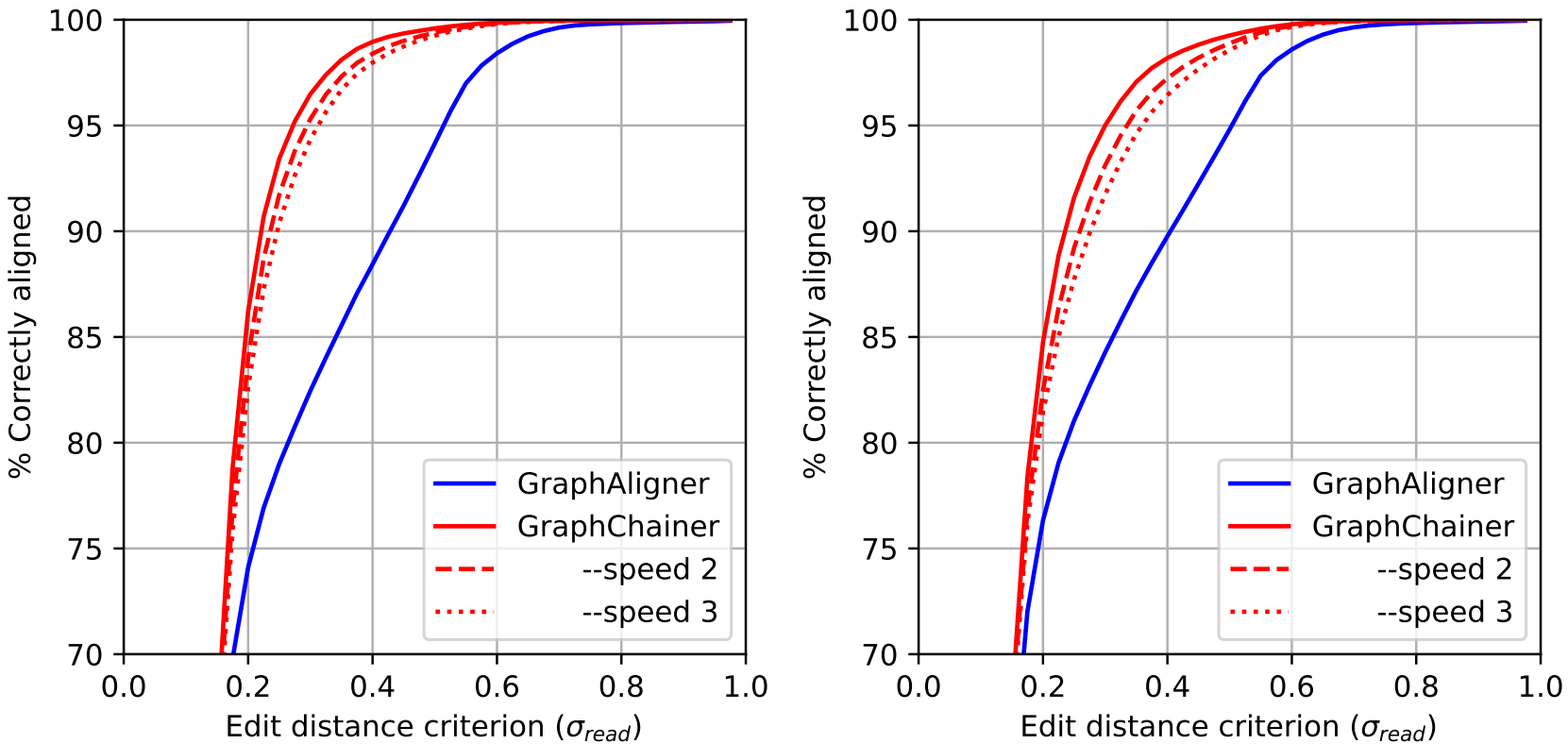




Accuracy – Real Reads, $\sigma = 0.3$

Graph	Aligner	Correctly aligned
Chr22 (real)	`GraphAligner`	82.45%
	`GraphChainer`	96.48% (+17.02%)
	`--speed 2`	95.31% (+15.59%)
	`--speed 3`	94.34% (+14.41%)
Chr1 (real)	`GraphAligner`	84.27%
	`GraphChainer`	95.01% (+12.74%)
	`--speed 2`	93.12% (+10.50%)
	`--speed 3`	91.75% (+8.87%)

Accuracy – Real Reads



Chr22 (left) and Chr1 (right)

Running time – Real Reads

Graph	Aligner	Mem	CPU time	Real time
Chr22 (real)	`GraphAligner`	7.92	02:00:03	00:04:12
	`GraphChainer`	12.58	04:48:09	00:09:56
	`--speed 2`	12.94	03:48:41	00:07:57
	`--speed 3`	13.14	03:30:07	00:07:20
Chr1 (real)	`GraphAligner`	18.66	13:48:46	00:28:42
	`GraphChainer`	58.47	73:05:00	02:28:12
	`--speed 2`	58.59	46:10:51	01:34:21
	`--speed 3`	58.78	37:32:20	01:17:05

Mem in GBs, times in hh:mm:ss

Future Work/Directions

- Getting anchors
- Cyclic graphs
- Gap and overlap costs
- Path gap/overlaps

Funding

European Research Council

Established by the European Commission

References I

[1] Backurs, A., and Indyk, P.
Edit distance cannot be computed in strongly subquadratic time (unless SETH is false).
In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing* (2015), pp. 51–58.

[2] Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tassé, A.-M., and Flicek, P.
The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data.
Nucleic Acids Research 45, D1 (09 2016), D854–D859.

References II

[3] Equi, M., Grossi, R., Mäkinen, V., and Tomescu, A. I.
On the complexity of string matching for graphs.
In *46th International Colloquium on Automata, Languages, and Programming, ICALP 2019, July 9-12, 2019, Patras, Greece* (2019), C. Baier, I. Chatzigiannakis, P. Flocchini, and S. Leonardi, Eds., vol. 132 of *LIPIcs*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 55:1–55:15.

[4] Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., et al.
Variation graph toolkit improves read mapping by representing genetic variation in the reference.
Nature biotechnology 36, 9 (2018), 875–879.

References III

[5] Ivanov, P., Bichsel, B., Mustafa, H., Kahles, A., Rätsch, G., and Vechev, M.
AStarix: Fast and optimal sequence-to-graph alignment.
In *International Conference on Research in Computational Molecular Biology* (2020), Springer, pp. 104–119.

[6] Ivanov, P., Bichsel, B., and Vechev, M.
Fast and optimal sequence-to-graph alignment guided by seeds.
bioRxiv (2021).

[7] Jain, C., Misra, S., Zhang, H., Dilthey, A., and Aluru, S.
Accelerating sequence alignment to graphs.
In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (2019), pp. 451–461.

References IV

[8] Jain, C., Zhang, H., Gao, Y., and Aluru, S.
On the complexity of sequence-to-graph alignment.
Journal of Computational Biology 27, 4 (2020), 640–654.

[9] Li, H.
Minimap2: pairwise alignment for nucleotide sequences.
Bioinformatics 34, 18 (05 2018), 3094–3100.

[10] Mäkinen, V., Tomescu, A. I., Kuosmanen, A., Paavilainen, T., Gagie, T., and Chikhi, R.
Sparse dynamic programming on DAGs with small width.
ACM Transactions on Algorithms (TALG) 15, 2 (2019), 1–21.

[11] Rautiainen, M., and Marschall, T.
GraphAligner: rapid and versatile sequence-to-graph alignment.
Genome biology 21, 1 (2020), 1–28.

References V

[12] Sahlin, K., and Mäkinen, V.
Accurate spliced alignment of long RNA sequencing reads.
Bioinformatics 37, 24 (2021), 4643–4651.

[13] Sirén, J., Monlong, J., Chang, X., Novak, A. M., Eizenga, J. M., Markello, C., Sibbesen, J. A., Hickey, G., Chang, P.-C., Carroll, A., et al.
Pangenomics enables genotyping of known structural variants in 5202 diverse genomes.
Science 374, 6574 (2021), abg8871.

[14] Wick, R. R.
Badread: simulation of error-prone long reads.
Journal of Open Source Software 4, 36 (2019), 1316.