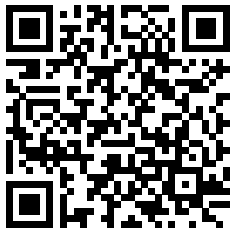
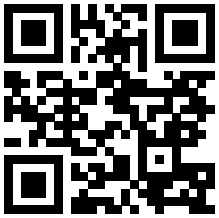
BLEND

A Fast, Memory-efficient and Accurate Mechanism to Find Fuzzy Seed Matches in Genome Analysis

Can Firtina, Jisung Park, Mohammed Alser, Jeremie S. Kim, Damla Senol Cali, Taha Shahroodi, Nika Mansouri Ghiasi, Gagandeep Singh, Konstantinos Kanellopoulos, Can Alkan, Onur Mutlu



Paper (NARGAB)



Source Code

SAFARI

ETH zürich

Carnegie Mellon

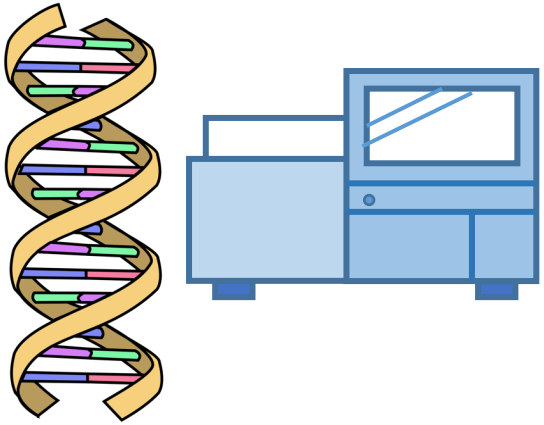
TU Delft

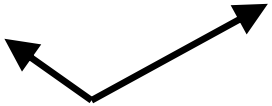
Delft University of Technology

Bilkent University

Genome Analysis

- High-throughput sequencing machines extract smaller fragments of the original DNA sequence, known as **reads**
  - **Challenge:** Perform genome analysis from the small random pieces of genomes

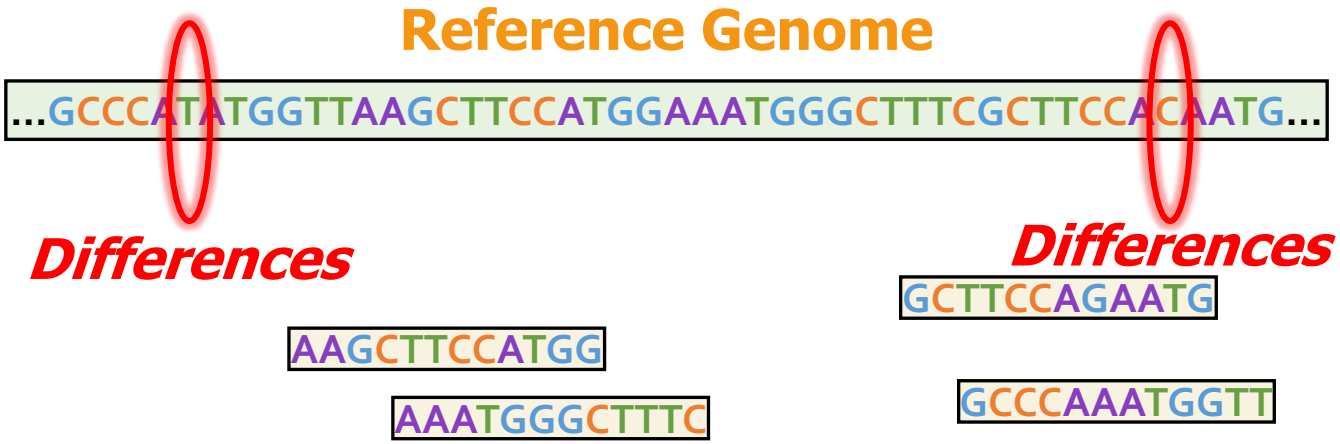




Reads

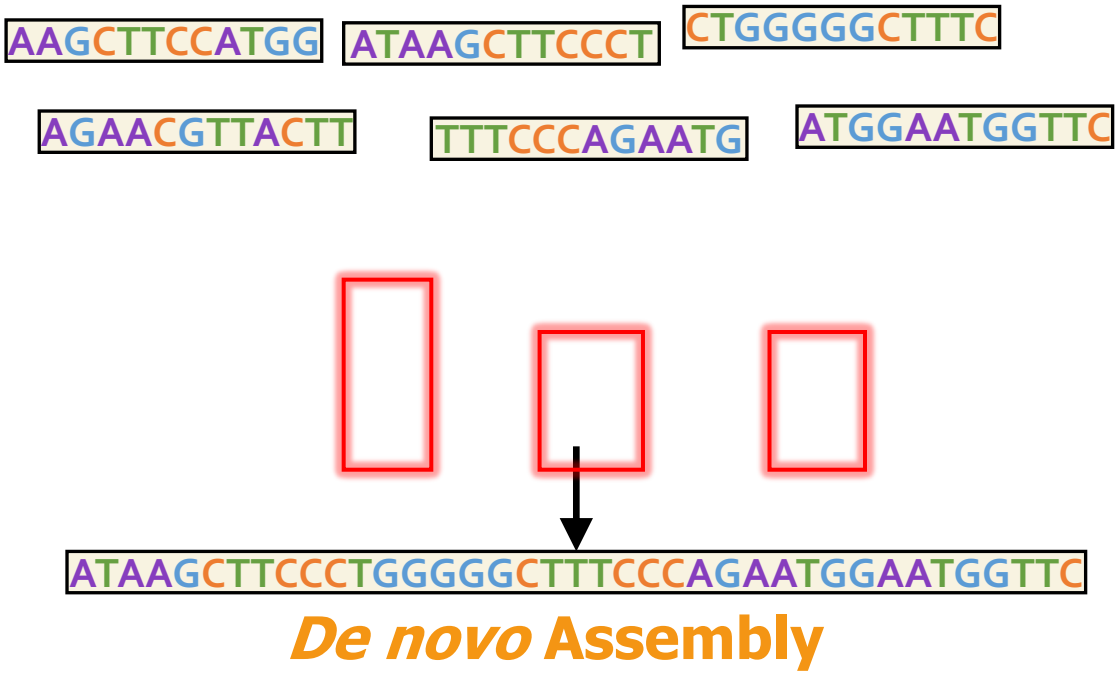
Identifying Sequence Similarities

1. Mapping reads to a **reference genome**
   - Finding the potential matching locations in the genome
   - Identifying differences between a read and a reference genome



Identifying Sequence Similarities

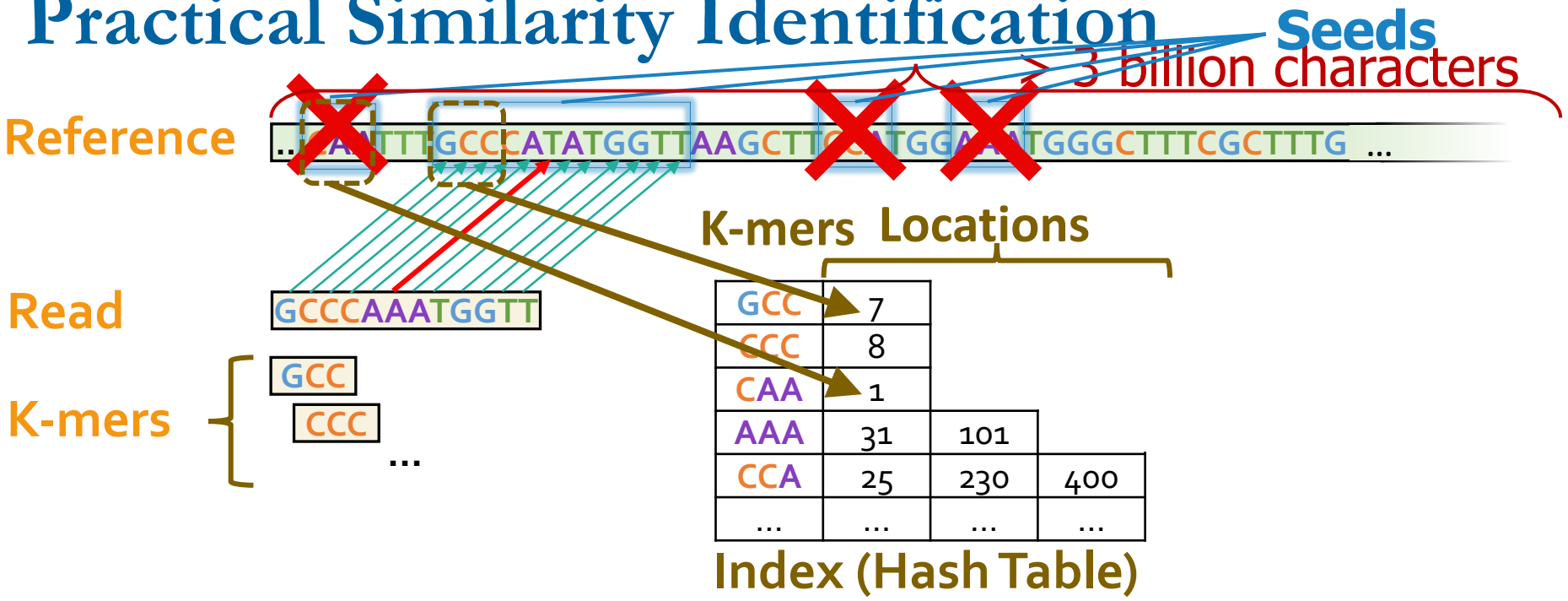
2. Overlapping reads **to each other**
   - Constructing assemblies



***De novo* Assembly**

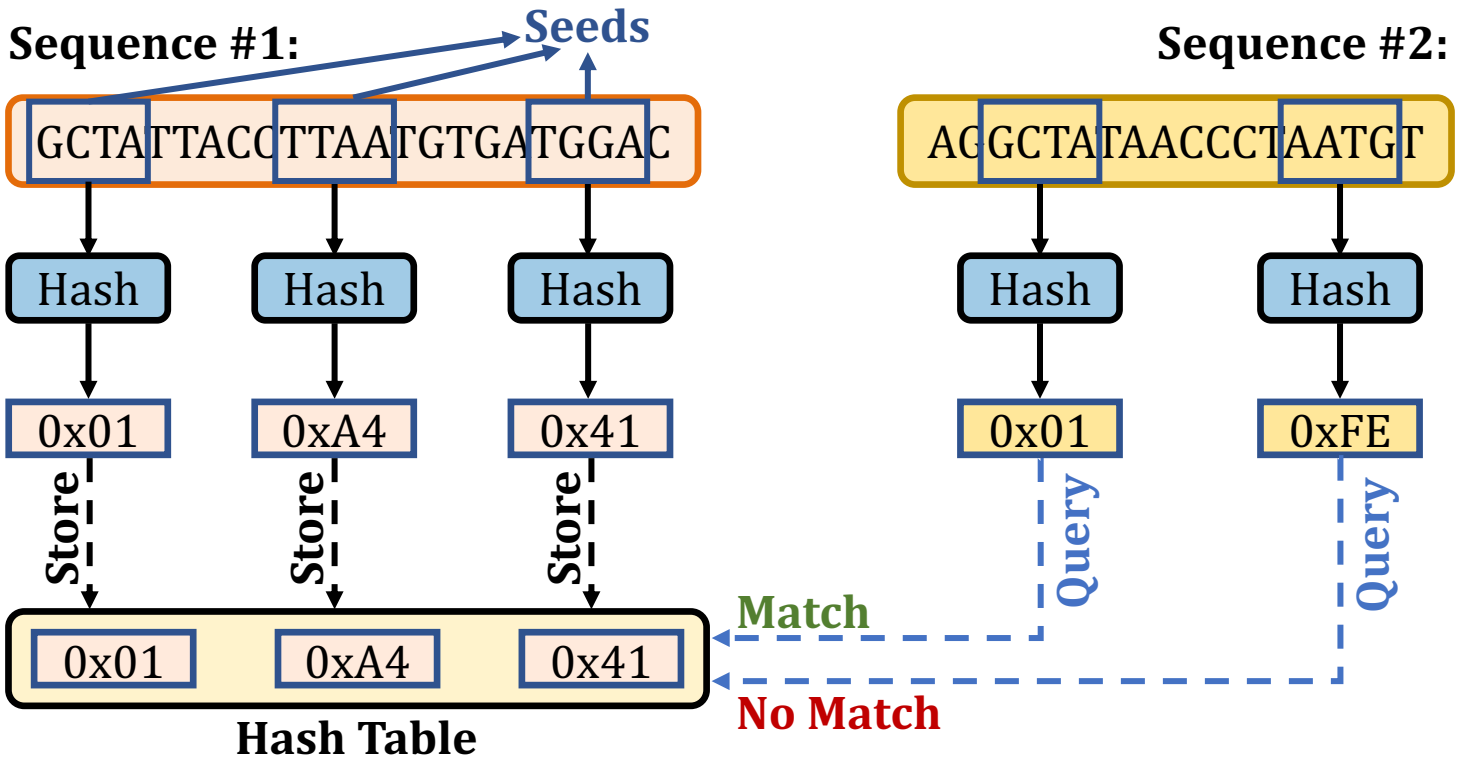
- Similarities must be identified **efficiently**
  - To facilitate a **practical search** among many sequence pairs

Practical Similarity Identification



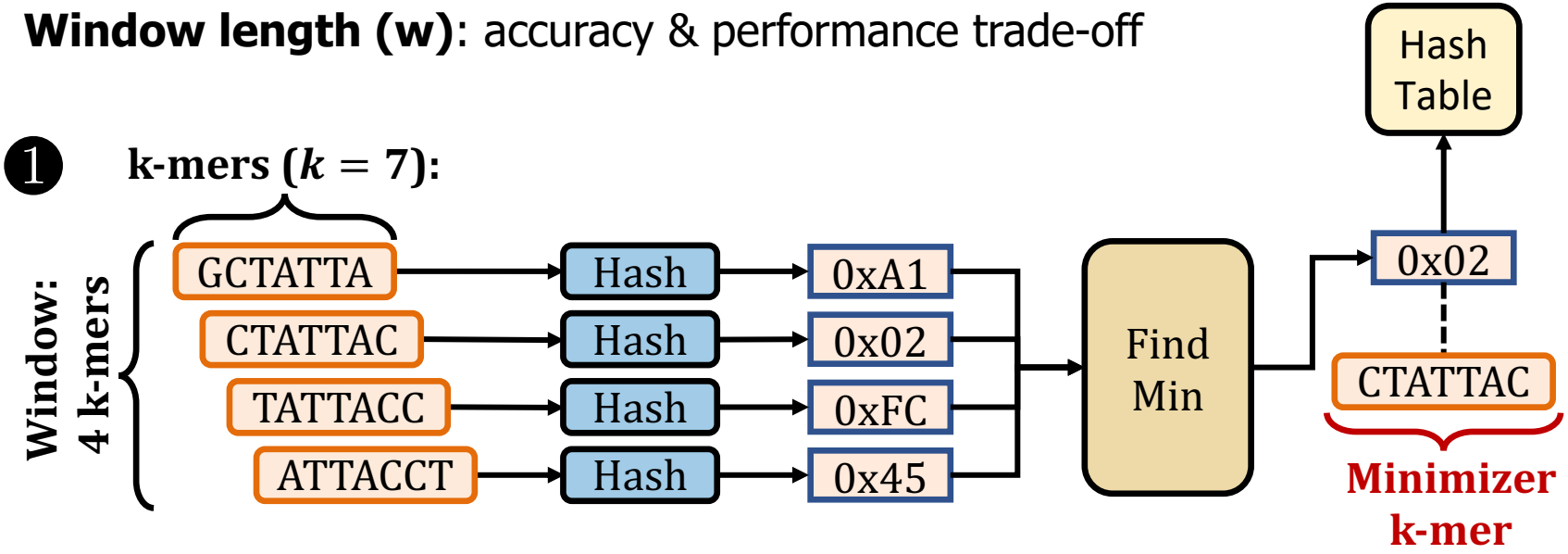
Seeding	Determine potential matching regions (seeds) in the reference genome
Seed Filtering (e.g., Chaining)	Prune some seeds in the reference genome
Alignment	Determine the exact differences between the read and the reference genome

Finding Seed Matches in Hash Tables



Seed Matching Techniques

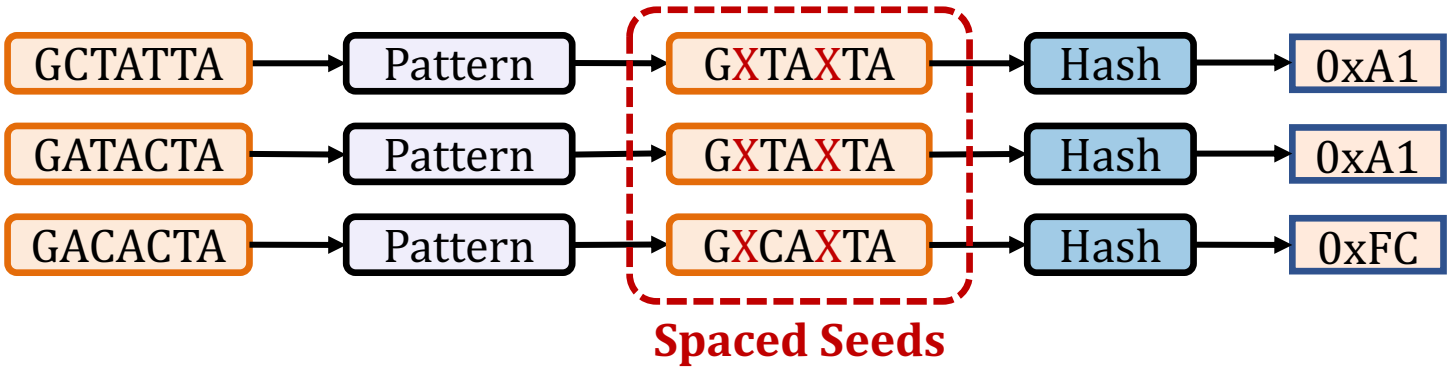
1. Sampling the overlapping k-mers
   - **Minimizers**
   - **Window length (w)**: accuracy & performance trade-off



Seed Matching Techniques

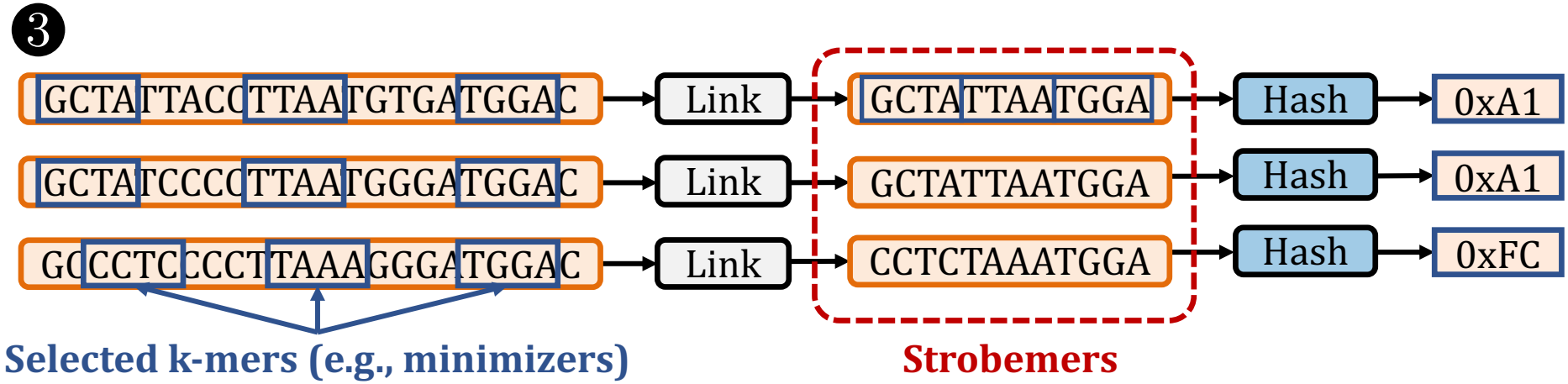
2. Allowing mismatches at **certain positions**
   - **Spaced seeds**
   - **Choice of pattern** is critical for the effectiveness of spaced seeds

❷



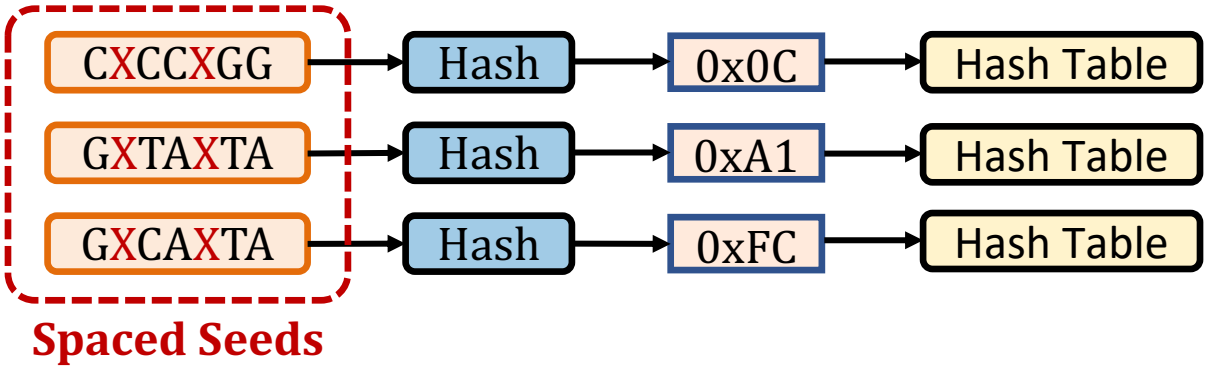
Seed Matching Techniques

3. Allowing insertions and deletions
   - **Linked k-mers** (e.g., strobemers)



Hashing the Seeds

- Seeds are hashed
  - To find seed matches with a **direct lookup of hash values**

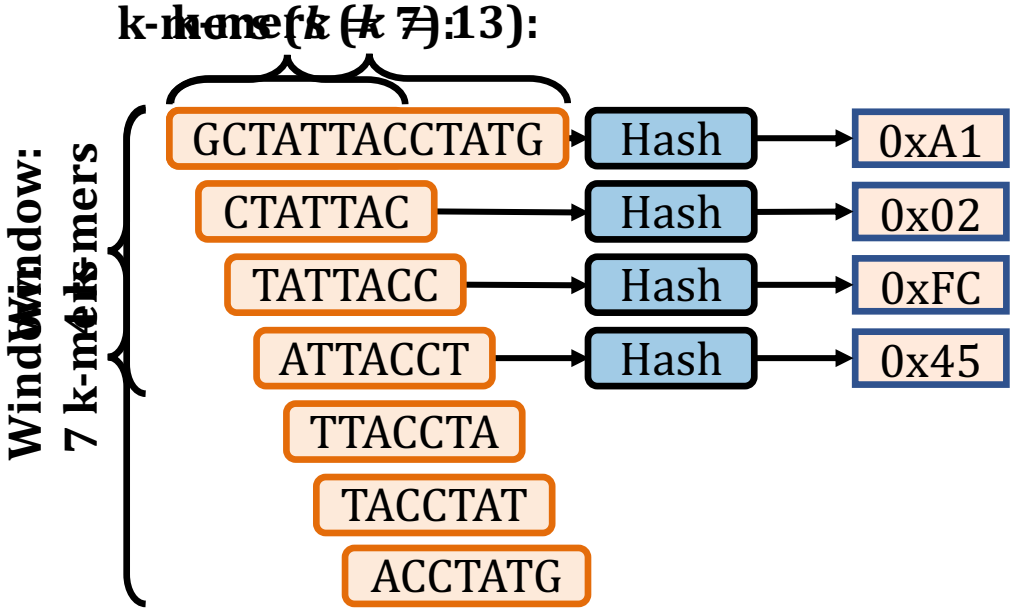


- **Low collision hash functions**
  - Seeds must **match exactly** to generate the **same hash value**
  - **Advantage: Dissimilar seeds** are **unlikely** to match
  - **Limitation: Highly similar seeds (fuzzy seeds)** are **unlikely** to match

Challenges - Exact-matching seeds

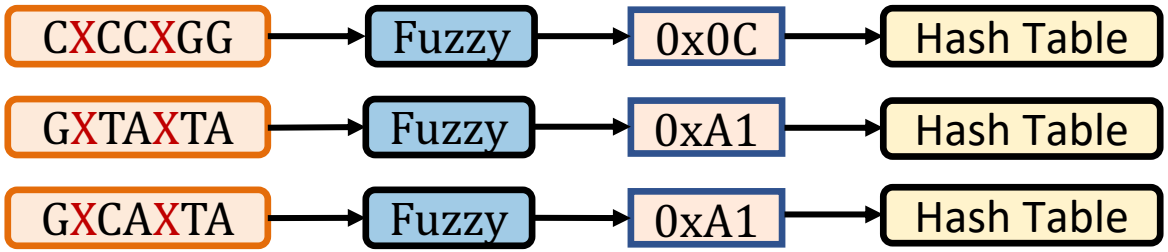
- **Limitations** when adjusting many **seeding parameters**
  - K-mer size
  - Window length (sampling ratio)
  - Determines the content of the hash table

- Trade-off between **performance, memory, and accuracy**



Opportunities - Fuzzy matches of seeds

- A mechanism for **finding fuzzy seed matches** can enable
  - Assigning the **same** hash values to **highly similar seeds**
  - **Different** hash values for **dissimilar seeds**
  - **High performance** (e.g., no distance or similarity calculation) and
  - **Space-efficient** (no multiple hash functions for a single sketch) seed matching



- Finding **useful and novel seed matches** that cannot be identified when finding only exact-matching seeds

- **Rethinking the seeding parameters** to achieve better trade-off between
  - Performance, memory, and accuracy

Outline

Background

Goal and Key Ideas

BLEND

Evaluation

Conclusions

Our Goal

Enable finding **fuzzy matches of seeds**
as well as **exact-matching seeds**
with a **single lookup** of hash values of seeds

BLEND

Uses a hashing mechanism, **SimHash**, that can generate **the same hash values** for **similar sets**

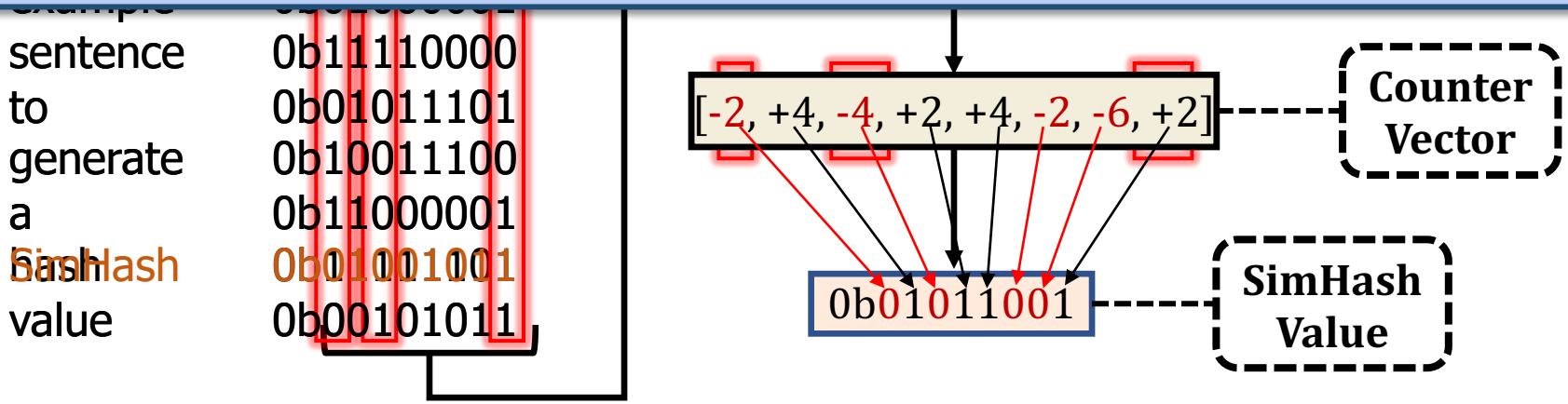
Provides the mechanisms for accurately and efficiently **converting seed sequences into set of items**

The SimHash Technique – Example

- Goal: Generate the **same hash value** for **similar set of items**
  - **Example input:** A sentence (a set of items)
  - **Items:** Words in a sentence (hash values of items)

- Count the net difference between 0s and 1s at each position

Challenge: Efficiently and accurately **convert seeds to set of items** to use with SimHash



Outline

Background

Goal and Key Ideas

BLEND

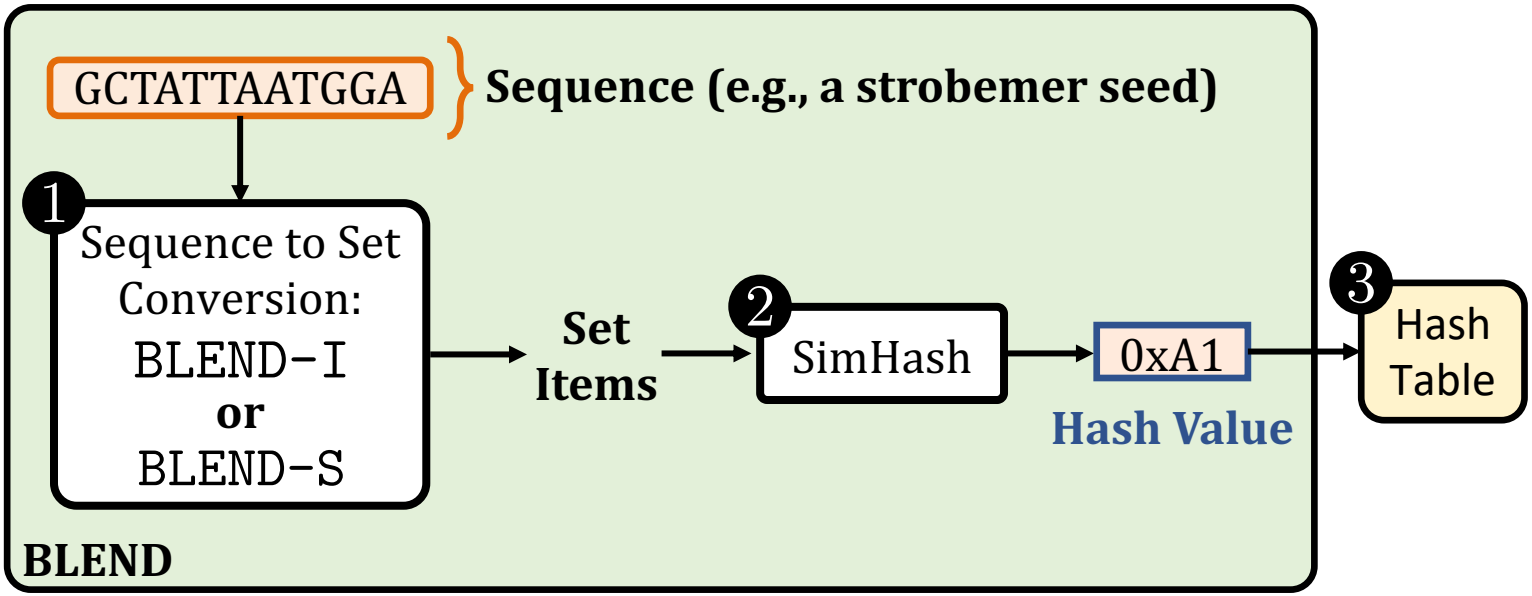
Evaluation

Conclusions

BLEND Overview

- **Goal:** Efficiently find fuzzy seed matches with a single lookup
  - **Input:** A fixed-length sequence (seed sequence)

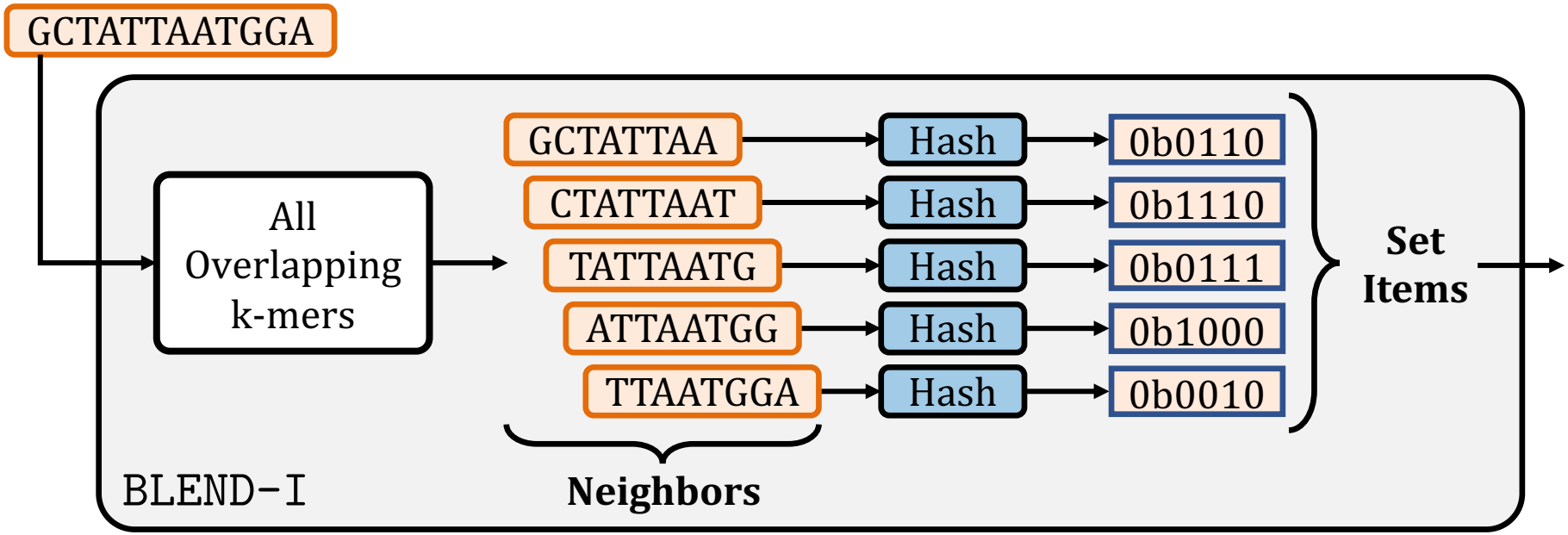
1. Efficiently and accurately **convert the seed** to its **set of items**
   - Two conversion mechanisms: **BLEND-I** and **BLEND-S**
2. Generate the SimHash value of the seed
3. Efficiently identify fuzzy seed matches by matching the SimHash values using a hash table



Sequence-to-Set Conversion (BLEND-I)

- **Goal:** Convert seed sequences into set of items
  - **Input:** A fixed-length sequence (seed sequence)

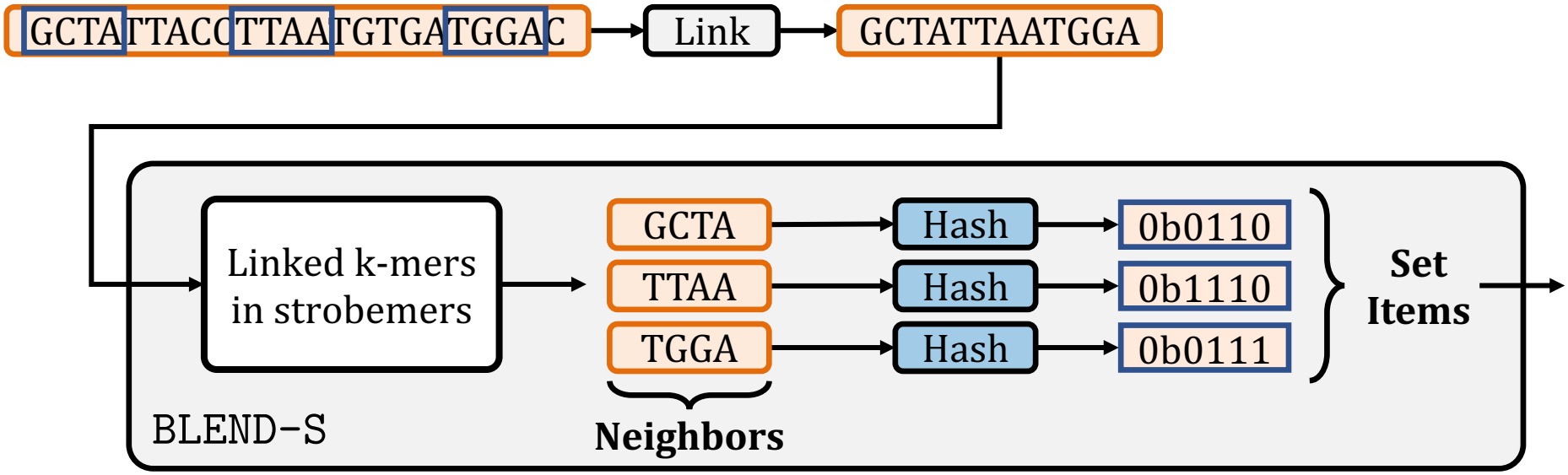
1. Extract **all overlapping k-mers** of the seed **(neighbors)**
2. Generate the **hash values of neighbors** using any hash function
3. **Set items:** Hash values of neighbors



Sequence-to-Set Conversion (BLEND-S)

- **Goal:** Convert seed sequences into set of items
  - **Input:** A fixed-length sequence (seed sequence)

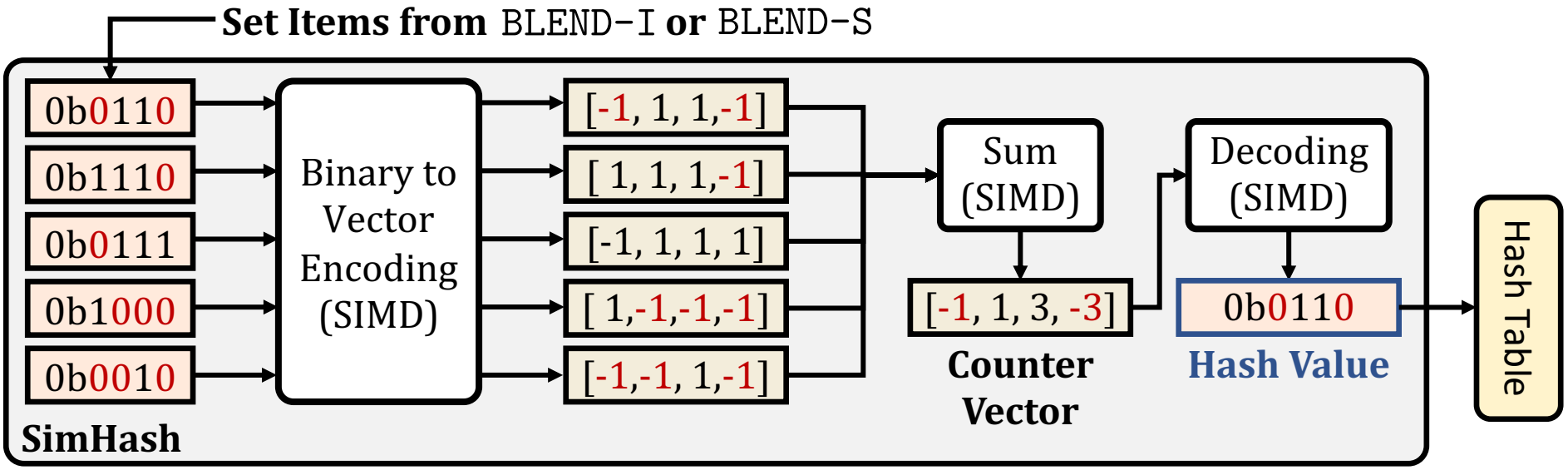
1. Extract **all linked k-mers** of the seed **(neighbors)**
2. Generate the **hash values of neighbors** using any hash function
3. **Set items:** Hash values of neighbors



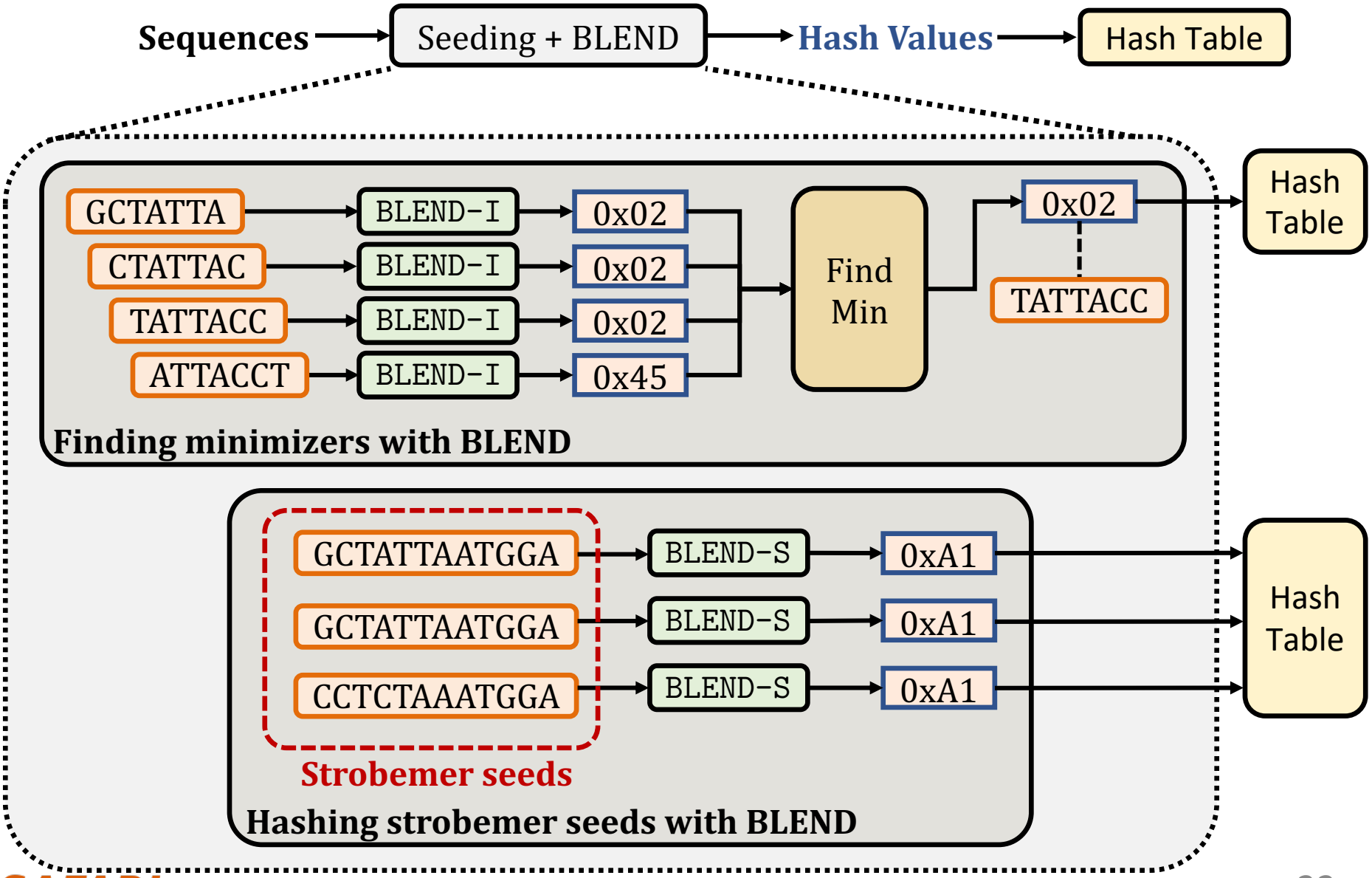
Generating the SimHash values

- **Goal:** Generate the SimHash value of a seed
  - **Input:** Set items from BLEND-I or BLEND-S

1. Encode hash values using vectors of **-1s** and **+1s**
2. Bitwise sum in SimHash: **Vector summation**
3. Decode the counter vector into a **SimHash value for the seed**



Integrating BLEND for Seeding



Outline

Background

Goal and Key Ideas

BLEND

Evaluation

Conclusions

Evaluation Methodology

- **Integrated into minimap2** to perform end-to-end mapping

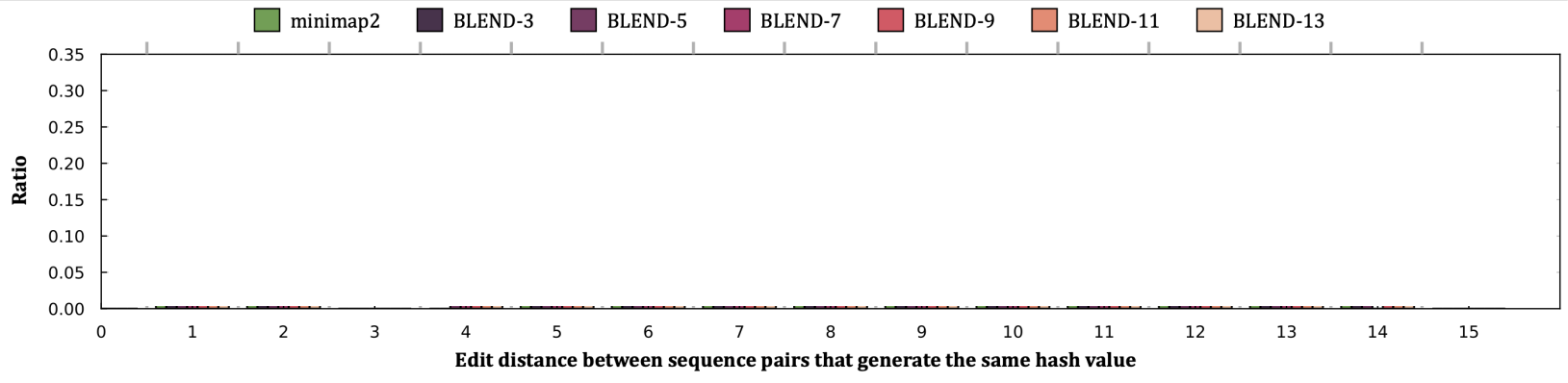
- Real and simulated datasets from
  - PacBio (HiFi and CLR), ONT, and Illumina reads
  - Human CHM13 and HG002, Fruit fly, Yeast, and E. coli genomes

- Use case 1: **Read overlapping** (all-vs-all overlapping)
  - Evaluated the **accuracy, completeness, and contiguity** of *de novo* assemblies generated from overlaps

- Use case 2: **Read mapping** to a reference genome
  - Mapping accuracy from simulated reads
  - Structural variant calling

Emprical Analysis on Fuzzy Seed Matching

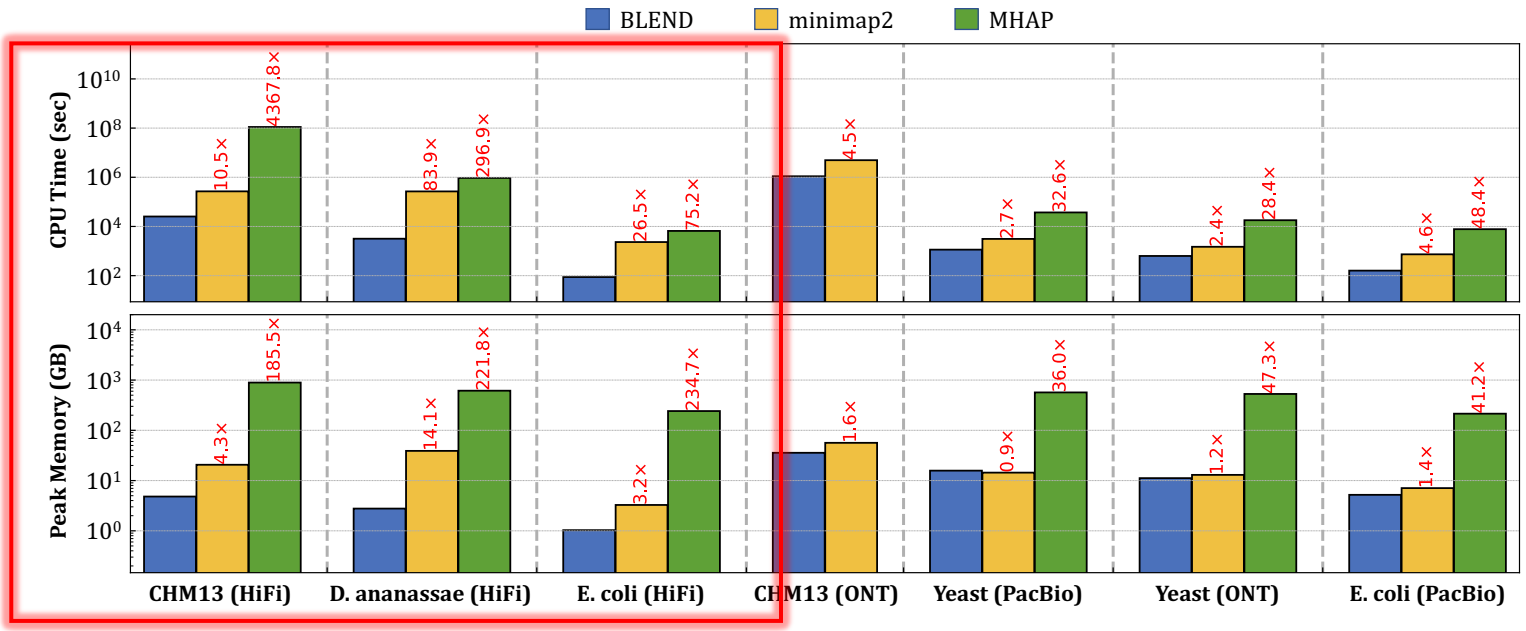
- Non-identical minimizers with the same hash value (collision)
  - Edit distance between minimizer k-mers with the same hash value
  - Ratio of collisions with a certain edit distance using minimap2 and BLEND
  - **Goal:** Increase the collision rate for highly similar seeds



BLEND **increases the collision rate** for highly similar seeds

while keeping a **collision rate similar to minimap2** for dissimilar seeds

Read Overlapping – Performance & Memory



For HiFi: Average **speedup** of 40.3x (minimap2)

Reducing the **memory** footprint by 7.2x

Improving critical parameters without hurting the accuracy:

Window length (200) and **seed length** (31-mers)

Read Overlapping – Assembly Evaluation

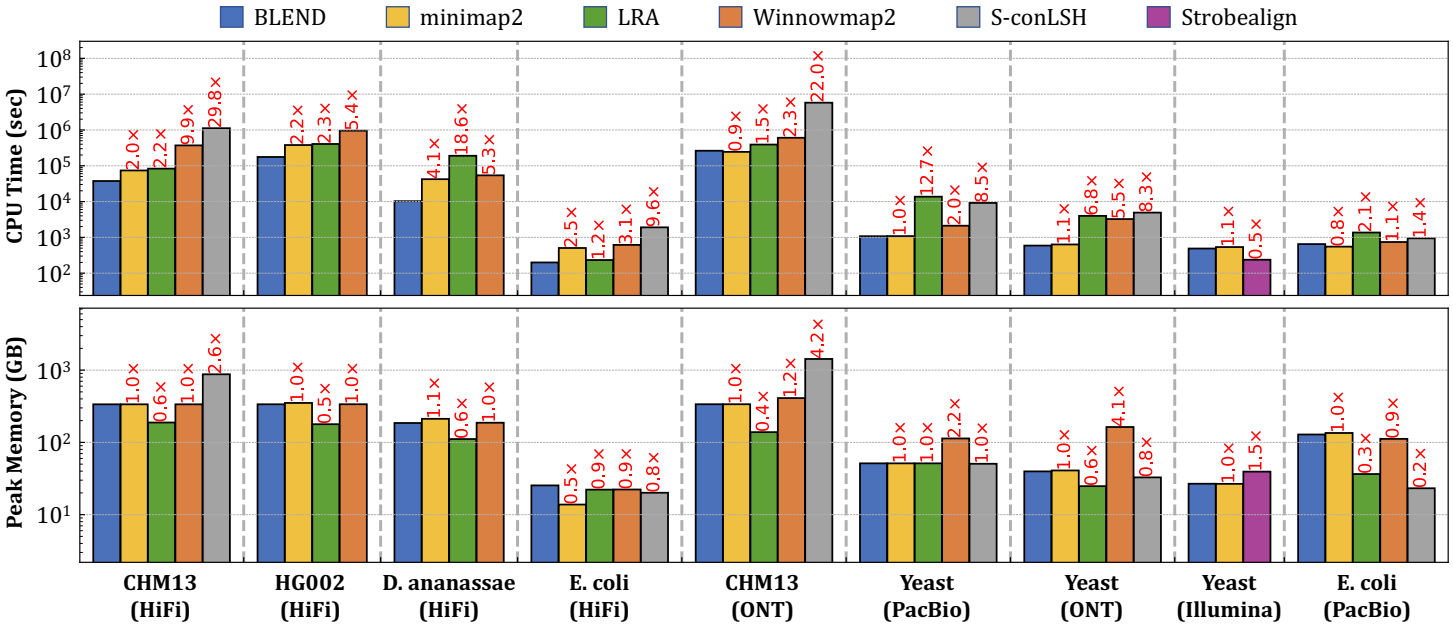
Dataset	Tool	Average Identity (%)	Genome Fraction (%)	*k*-mer Compl. (%)	Largest Contig (Mb)	NGA50 (kb)	NG50 (kb)
CHM13 (HiFi)	BLEND	**99.8526**	**98.4847**	**90.15**	22.8397	5442.25	5442.31
	minimap2	99.7421	97.1493	83.05	**47.1387**	**7133.43**	**7134.31**
	MHAP	N/A	N/A	N/A	N/A	N/A	N/A
	Reference	100	100	100	248.387	154.260	154.260
D. ananassae (HiFi)	BLEND	**99.7856**	**97.2308**	**86.43**	**6.23256**	**792.407**	**798.913**
	minimap2	99.7044	96.3190	72.33	4.43396	273.398	278.775
	MHAP	99.5551	0.7276	0.21	0.028586	N/A	N/A
	Reference	100	100	100	30.6728	26 427.4	26 427.4
E. coli (HiFi)	BLEND	**99.8320**	**99.8801**	**87.91**	**3.41699**	**3416.99**	**3416.99**
	minimap2	99.7064	99.8748	79.27	3.08849	3087.05	3087.05
	MHAP	N/A	N/A	N/A	N/A	N/A	N/A
	Reference	100	100	100	4.94446	4944.46	4944.46

For HiFi: More accurate and complete assemblies

Better contiguity in most cases

Fuzzy seed matches can lead to finding **novel and useful overlaps**

Read Mapping – Performance & Memory



Average **speedup** of 1.7x (minimap2) and similar peak memory usage

The computational cost of sequence alignment slightly hinders the benefits of fuzzy seed matching

Read Mapping – SV Calling

- Structural variant (SV) calling using read mappings from each tool
  - Sniffles2 to call SVs from HG002 long read mappings
  - Truvari to compare the resulting SVs with the benchmarking SV set (Tier 1 set from GIAB)

	HG002 SVs (high-confidence tier 1 SV set)					
Tool	TP (#)	FP (#)	FN (#)	Precision	Recall	F_1
BLEND	**9229**	855	**412**	0.9152	**0.9573**	**0.9358**
minimap2	9222	915	419	0.9097	0.9565	0.9326
LRA	9155	**830**	486	**0.9169**	0.9496	0.9329
Winnowmap2	9170	1029	471	0.8991	0.9511	0.9244

Best overall accuracy in downstream analysis

Outline

Background

Goal and Key Ideas

BLEND

Evaluation

Conclusions

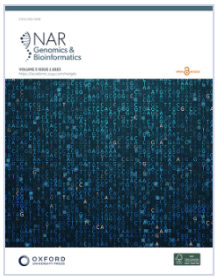
BLEND Summary

Problem	Finding exact-matching seeds introduce limitations in further improving the performance and accuracy of genome analysis
Goal	Enable finding the fuzzy seed matches as well as the exact-matching seeds accurately and efficiently
BLEND	• Provides effective mechanisms for converting seeds into set of items to use with the SimHash technique • Uses the SimHash technique to enable generating the same hash value for highly similar seeds to find fuzzy seed matches with a directly lookup of hash values
Key Results	• Significant speedups and lower memory footprint especially when using **HiFi reads** • Improves the accuracy of important applications in genome analysis

BLEND

- Can Firtina, Jisung Park, Mohammed Alser, Jeremie S. Kim, Damla Senol Cali, Taha Shahroodi, Nika Mansouri Ghiasi, Gagandeep Singh, Konstantinos Kanellopoulos, Can Alkan, and Onur Mutlu,
  "BLEND: A Fast, Memory-Efficient, and Accurate Mechanism to Find Fuzzy Seed Matches in Genome Analysis"
  NAR Genomics and Bioinformatics, March 2023.
  [Online link at NAR Genomics and Bioinformatics Journal]
  [arXiv preprint]
  [bioRxiv preprint]
  [BLEND Source Code]

Paper (NARGAB)



Volume 5, Issue 1
March 2023

JOURNAL ARTICLE

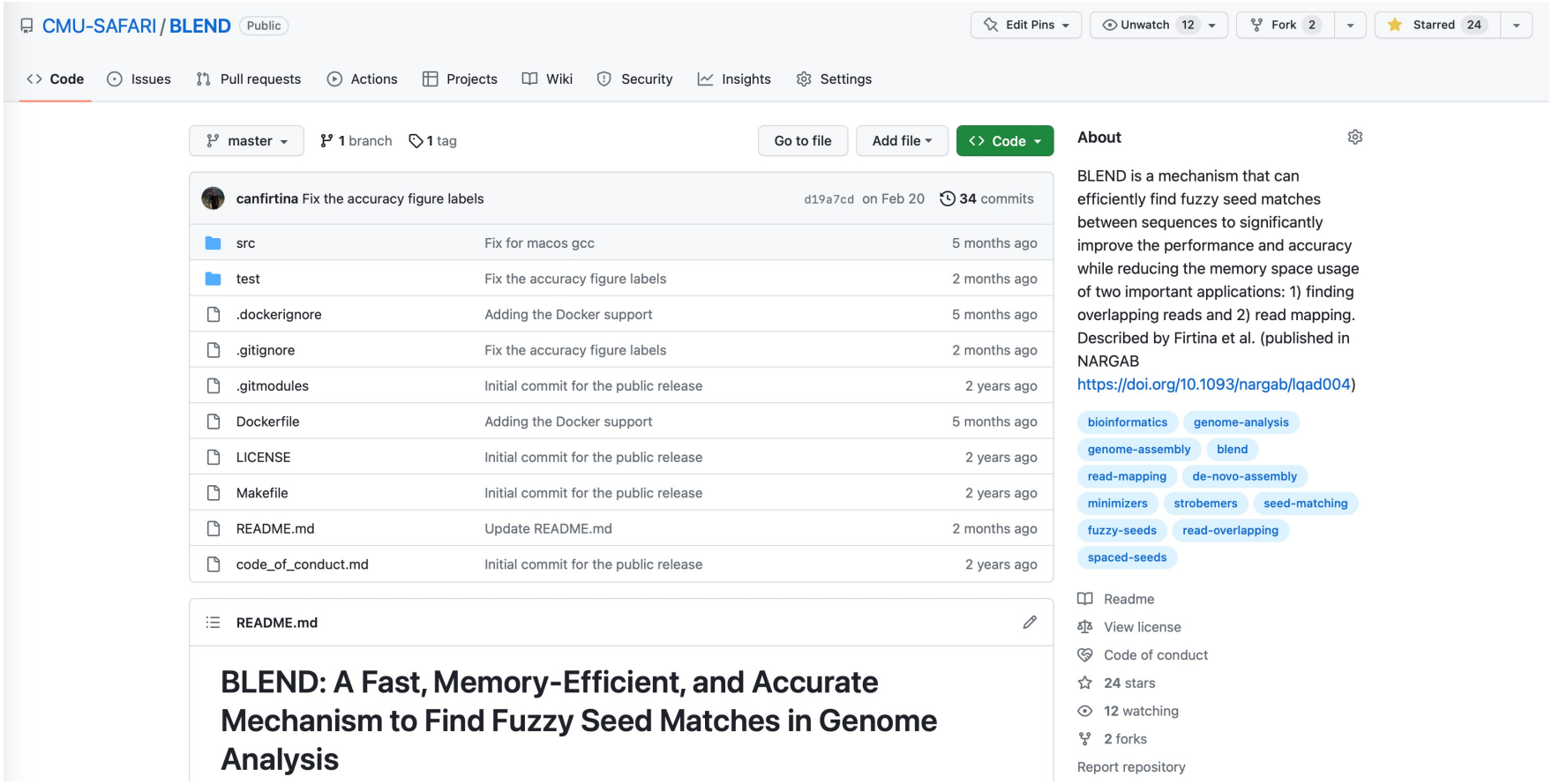
BLEND: a fast, memory-efficient and accurate mechanism to find fuzzy seed matches in genome analysis

Can Firtina, Jisung Park, Mohammed Alser, Jeremie S Kim, Damla Senol Cali, Taha Shahroodi, Nika Mansouri Ghiasi, Gagandeep Singh, Konstantinos Kanellopoulos, Can Alkan, Onur Mutlu

NAR Genomics and Bioinformatics, Volume 5, Issue 1, March 2023, lqad004,

BLEND Source Code

Source Code

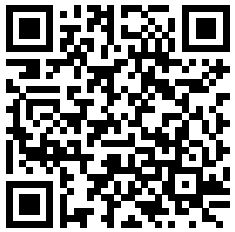


https://github.com/CMU-SAFARI/BLEND

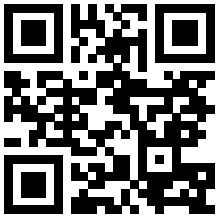
BLEND

A Fast, Memory-efficient and Accurate Mechanism to Find Fuzzy Seed Matches in Genome Analysis

Can Firtina, Jisung Park, Mohammed Alser, Jeremie S. Kim, Damla Senol Cali, Taha Shahroodi, Nika Mansouri Ghiasi, Gagandeep Singh, Konstantinos Kanellopoulos, Can Alkan, Onur Mutlu



Paper (NARGAB)



Source Code

SAFARI

ETH zürich

Carnegie Mellon

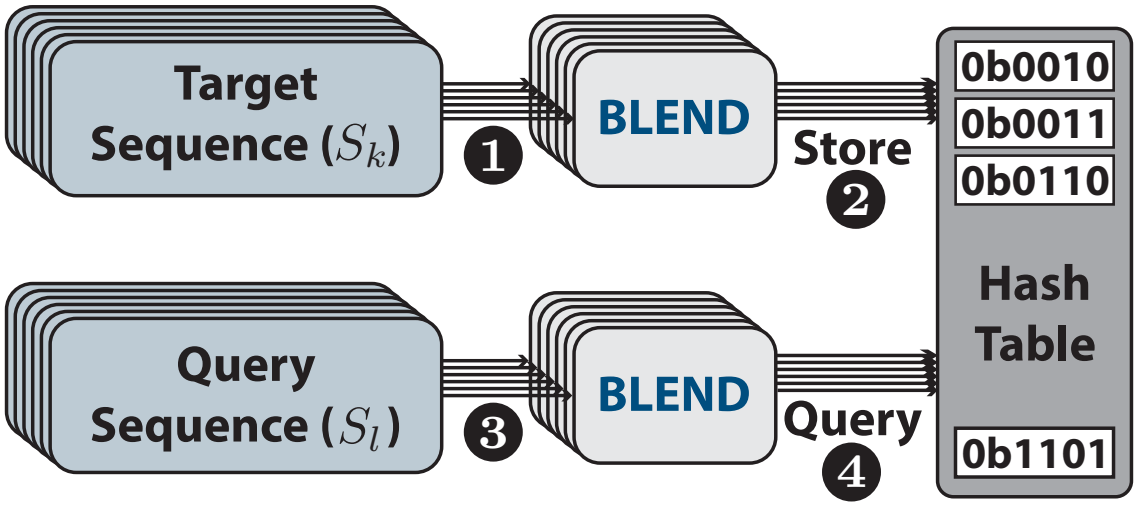
TU Delft

Delft University of Technology

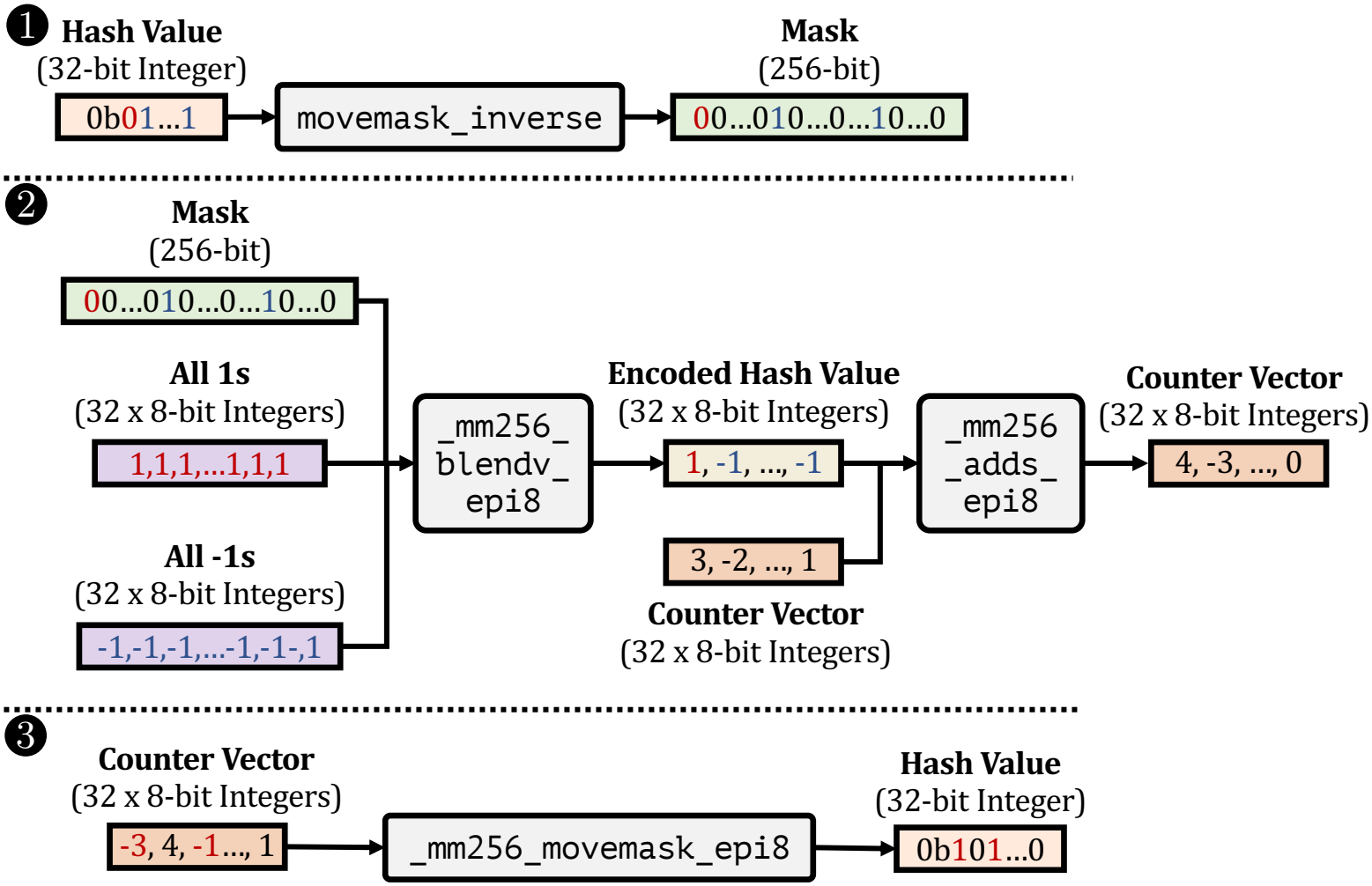
Bilkent University

Backup Slides

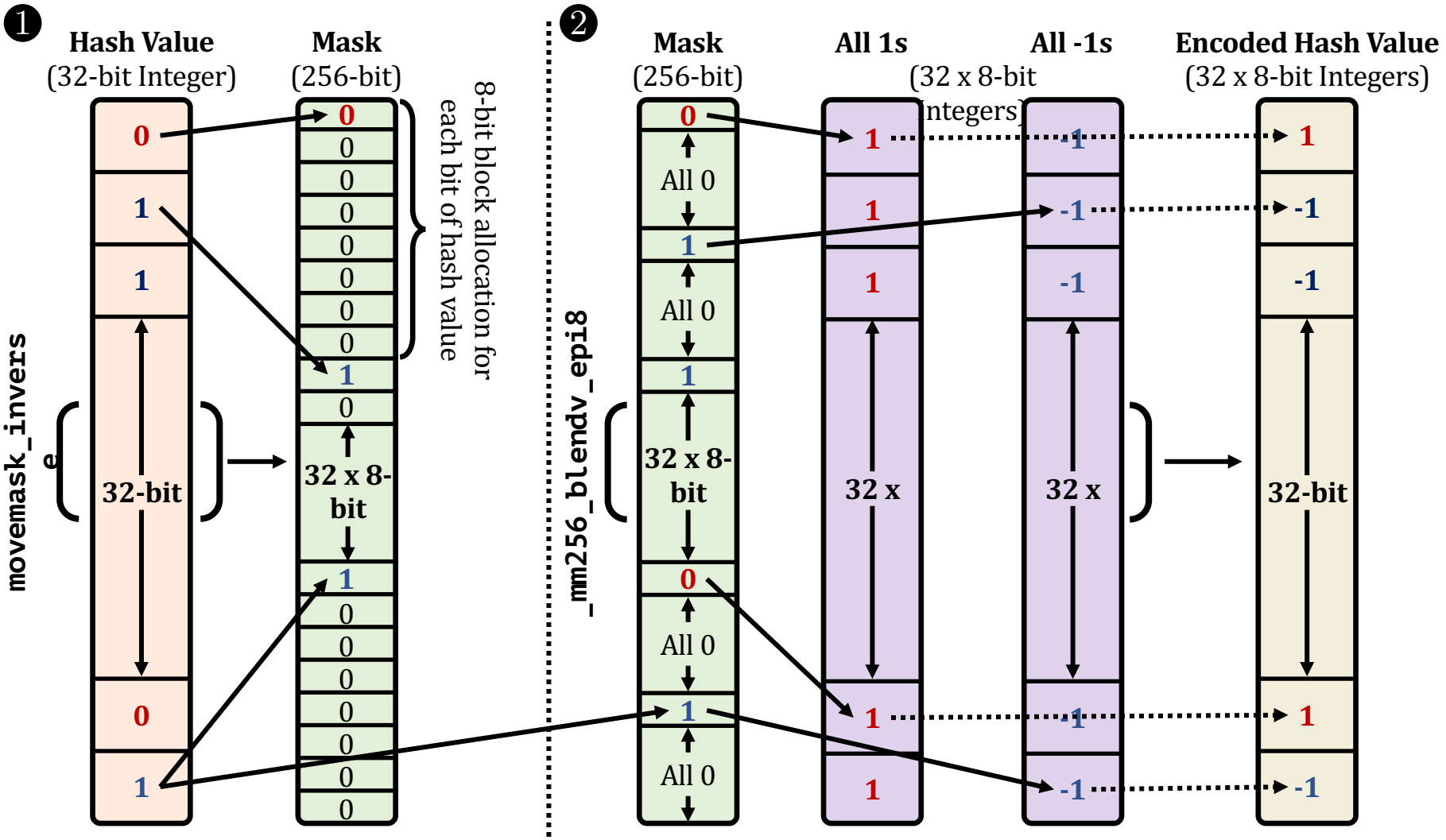
Integrating BLEND in Read Mappers



Calculating SimHash using SIMD



Encoding Hash Values into Vectors using SIMD



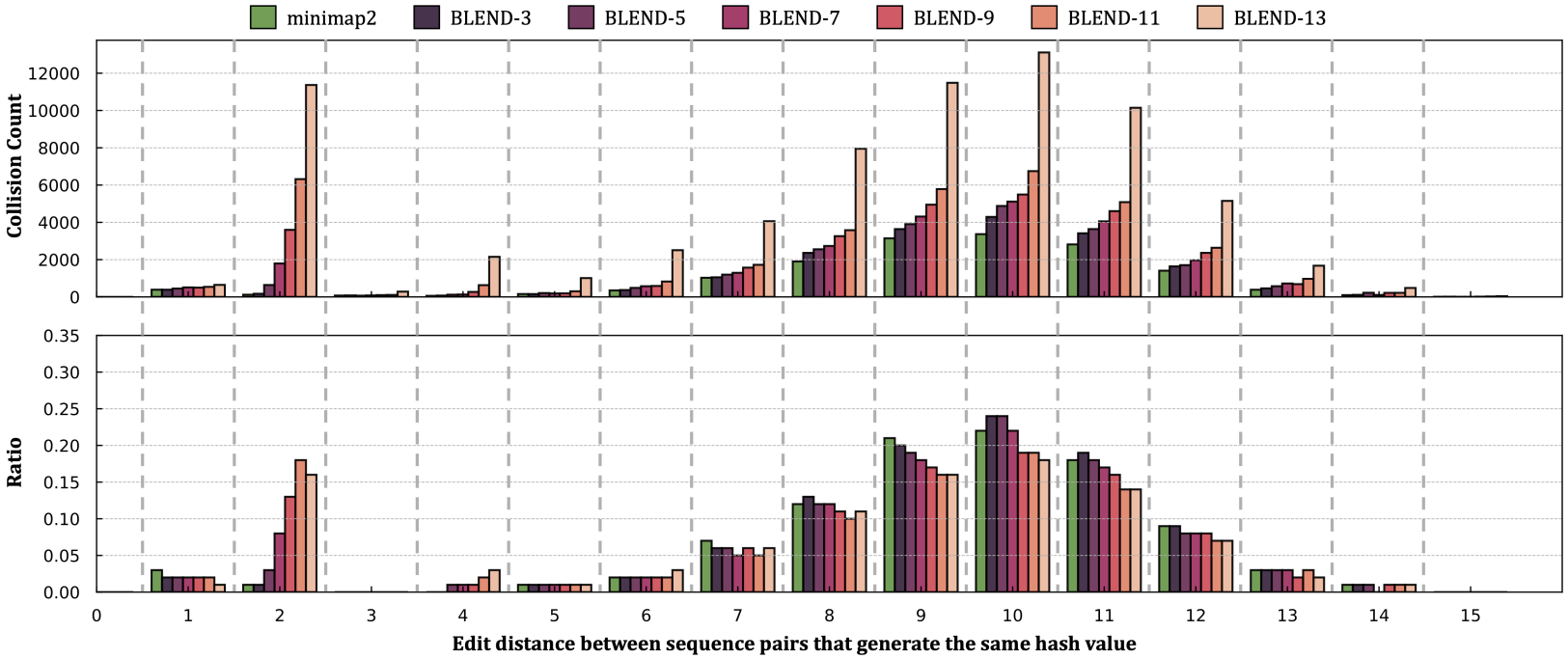
Dataset

Organism	Library	Reads (#)	Seq. Depth	SRA Accession	Reference Genome
Human CHM13	PacBio HiFi	3 167 477	16	SRR11292122-3	T2T-CHM13 (v1.1)
	ONT*	10 380 693	30	Simulated R9.5	T2T-CHM13 (v2.0)
Human HG002	PacBio HiFi	11 714 594	52	SRR10382244-9	GRCh37
D. ananassae	PacBio HiFi	1 195 370	50	SRR11442117	(90)
Yeast	PacBio CLR*	270 849	200	Simulated P6-C4	GCA_000146045.2
	ONT*	135 296	100	Simulated R9.5	GCA_000146045.2
	Illumina MiSeq	3 318 467	80	ERR1938683	GCA_000146045.2
E. coli	PacBio HiFi	38 703	100	SRR11434954	(90)
	PacBio CLR	76 279	112	SRR1509640	GCA_000732965.1

* We use PBSIM2 to generate the simulated PacBio and ONT reads.
We show the simulated chemistry under the SRA Accession column.

Emprical Analysis on Fuzzy Seed Matching

- Minimizer collisions and the edit distance between collisions



BLEND **increases the collision rate** for highly similar seeds

while keeping a **collision rate similar to minimap2**

for dissimilar seeds

Finding Fuzzy Seed Matching between Minimizers

Tool	Number of Minimizers	Number of Collisions	Collision/Minimizer Ratio	Avg. Edit Distance Between Minimizers With Collision
minimap2	903,043	15,306	0.016949	9.327061
BLEND-3	1,014,173	18,224	0.017969	9.393437
BLEND-5	1,090,468	20,659	0.018945	9.213660
BLEND-7	1,140,254	23,591	0.020689	8.874698
BLEND-9	1,173,198	28,411	0.024217	8.495301
BLEND-11	1,186,687	35,500	0.029915	8.067549
BLEND-13	1,197,966	72,078	0.060167	8.075918

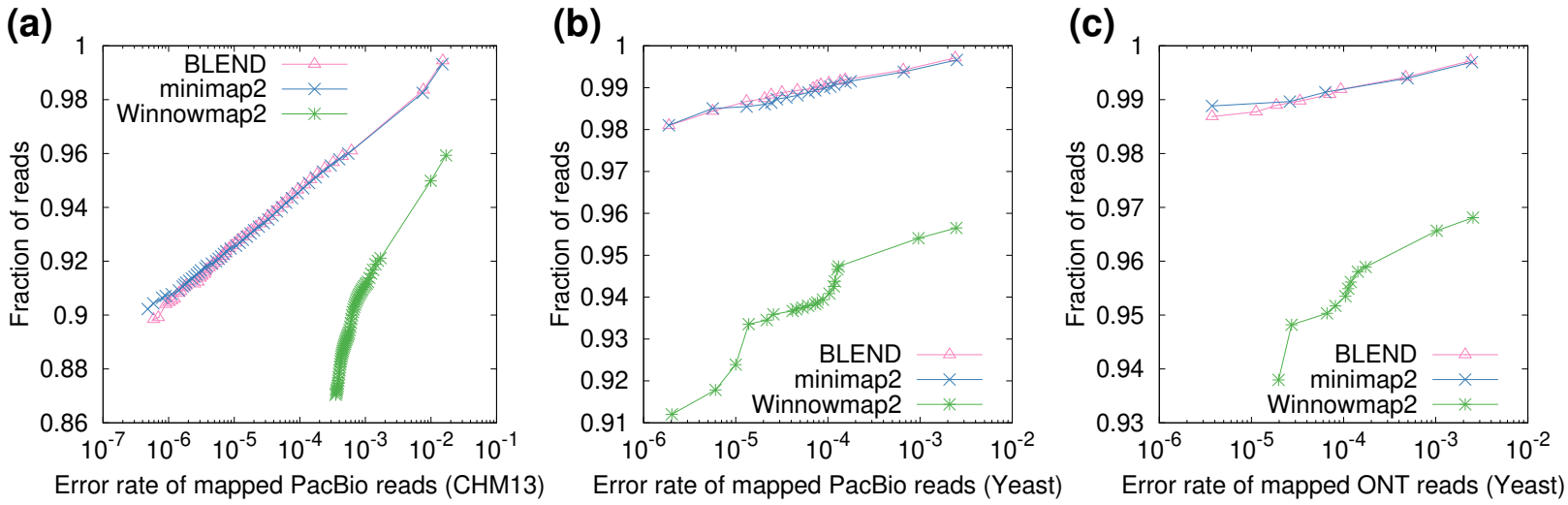
Finding Fuzzy Seed Matching between Similar Sequences

Tool	Number of Sequences	Number of Sequences with Collision	Collision/Sequence Ratio	Avg. Edit Distance Between K-mers With Collision
minimap2	4,130	0	0	N/A
BLEND-3	4,130	0	0	N/A
BLEND-5	4,130	11	0.00263663	1.45455
BLEND-7	4,130	50	0.0119847	1.5
BLEND-9	4,130	77	0.0184564	2.01299
BLEND-11	4,130	273	0.0654362	2.80952
BLEND-13	4,130	329	0.0788591	2.20669

Read Overlapping – Assembly Accuracy

Dataset	Tool	Average Identity (%)	Genome Fraction (%)	*k*-mer Compl. (%)	Aligned Length (Mb)	Mismatch per 100 kb (#)	Average GC (%)	Assembly Length (Mb)	Largest Contig (Mb)	NGA50 (kb)	NG50 (kb)
CHM13 (HiFi)	BLEND	**99.8526**	**98.4847**	**90.15**	3092.54	**22.02**	**40.78**	**3095.21**	22.8397	5442.25	5442.31
	minimap2	99.7421	97.1493	83.05	**3094.79**	55.96	40.71	3100.97	**47.1387**	**7133.43**	**7134.31**
	MHAP	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Reference	100	100	100	3054.83	0.00	40.85	3054.83	248.387	154 260	154 260
D. ananassae (HiFi)	BLEND	**99.7856**	**97.2308**	**86.43**	240.391	**143.13**	**41.75**	**247.153**	**6.23256**	**792.407**	**798.913**
	minimap2	99.7044	96.3190	72.33	**289.453**	191.53	41.68	298.28	4.43396	273.398	278.775
	MHAP	99.5551	0.7276	0.21	2.29	239.76	42.07	2.34951	0.028586	N/A	N/A
	Reference	100	100	100	213.805	0.00	41.81	213.818	30.6728	26 427.4	26 427.4
E. coli (HiFi)	BLEND	**99.8320**	**99.8801**	**87.91**	**5.12155**	**3.77**	**50.53**	5.12155	**3.41699**	**3416.99**	**3416.99**
	minimap2	99.7064	99.8748	79.27	5.09249	19.71	50.47	**5.09436**	3.08849	3087.05	3087.05
	MHAP	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Reference	100	100	100	5.04628	0.00	50.52	5.04628	4.94446	4944.46	4944.46
CHM13 (ONT)	BLEND	N/A	N/A	**29.26**	**2891.28**	**4077.53**	**41.32**	2897.87	25.2071	5061.52	5178.59
	minimap2	N/A	N/A	28.32	2860.26	4660.73	41.36	**2908.55**	**66.7564**	**13 189.2**	**13 820.3**
	Reference	100	100	100	3117.29	0.00	40.75	3117.29	248.387	150 617	150 617
Yeast (PacBio)	BLEND	89.1677	**97.0854**	**33.81**	**12.3938**	2672.37	38.84	12.4176	1.54807	635.966	636.669
	minimap2	88.9002	96.9709	33.38	12.0128	2684.38	38.85	**12.3325**	**1.56078**	**810.046**	**828.212**
	MHAP	**89.2182**	88.5928	32.39	10.9039	**2552.05**	**38.81**	10.9896	1.02375	85.081	436.285
	Reference	100	100	100	12.1571	0.00	38.15	12.1571	1.53193	924.431	924.431
Yeast (ONT)	BLEND	**89.6889**	99.2974	**35.95**	**12.3222**	**2529.47**	**38.64**	**12.3225**	1.10582	793.046	793.046
	minimap2	88.9393	**99.6878**	34.84	12.304	2782.59	38.74	12.3725	**1.56005**	**796.718**	**941.588**
	MHAP	89.1970	89.2785	33.58	10.8302	2647.19	38.84	10.9201	1.44328	118.886	618.908
	Reference	100	100	100	12.1571	0.00	38.15	12.1571	1.53193	924.431	924.431
E. coli (PacBio)	BLEND	**88.5806**	**96.5238**	**32.32**	**5.90024**	**1857.56**	**49.81**	6.21598	2.40671	**769.981**	2060.4
	minimap2	88.1365	92.7603	30.74	5.37728	2005.72	49.66	**6.02707**	**3.77098**	367.442	**3770.98**
	MHAP	88.4883	90.5533	31.32	5.75159	1999.48	49.69	6.26216	1.04286	110.535	456.01
	Reference	100	100	100	5.6394	0.00	50.43	5.6394	5.54732	5547.32	5547.32

Read Mapping – Mapping Accuracy



Read Mapping – Mapping Accuracy

	Overall error rate (%)		
Dataset	BLEND	minimap2	Winnowmap2
CHM13 (ONT)	1.5168427	**1.4914009**	1.7001222
Yeast (PacBio)	**0.2403134**	0.2504307	0.2474206
Yeast (ONT)	**0.2386617**	0.2468770	0.2534777

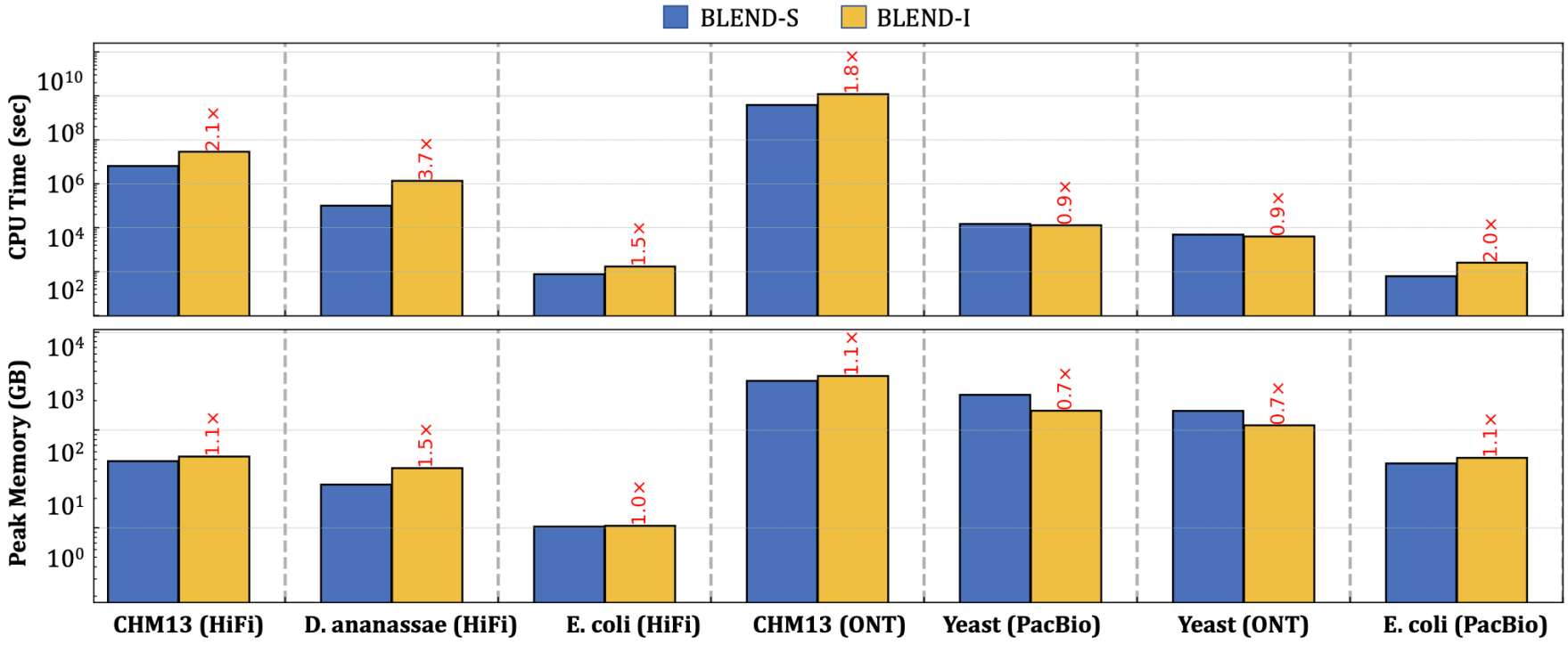
Read Mapping Quality

Dataset	Tool	Average depth of Cov. (×)	Breadth of coverage (%)	Aligned reads (#)	Properly paired (%)
CHM13 (HiFi)	BLEND	**16.58**	**99.991**	3 171 916	NA
	minimap2	**16.58**	**99.991**	**3 172 261**	NA
	LRA	16.37	99.064	3 137 631	NA
	Winnowmap2	**16.58**	99.990	3 171 313	NA
HG002 (HiFi)	BLEND	51.25	92.245	11 424 762	NA
	minimap2	53.08	92.242	12 407 589	NA
	LRA	52.48	**92.275**	**13 015 195**	NA
	Winnowmap2	**53.81**	92.248	12 547 868	NA
D. ananassae (HiFi)	BLEND	57.37	99.662	1 223 388	NA
	minimap2	**57.57**	**99.665**	1 245 931	NA
	LRA	57.06	99.599	1 235 098	NA
	Winnowmap2	57.40	99.663	**1 249 575**	NA
E. coli (HiFi)	BLEND	**99.14**	99.897	39 048	NA
	minimap2	**99.14**	99.897	**39 065**	NA
	LRA	99.10	99.897	39 063	NA
	Winnowmap2	**99.14**	99.897	39 036	NA
CHM13 (ONT)	BLEND	**29.34**	**99.999**	**10 322 767**	NA
	minimap2	29.33	**99.999**	10 310 182	NA
	LRA	28.84	99.948	9 999 432	NA
	Winnowmap2	28.98	99.936	9 958 402	NA
Yeast (PacBio)	BLEND	**195.87**	**99.980**	**270 064**	NA
	minimap2	195.86	**99.980**	269 935	NA
	LRA	194.65	99.967	267 399	NA
	Winnowmap2	192.35	99.977	259 073	NA
Yeast (ONT)	BLEND	**97.88**	**99.964**	**134 919**	NA
	minimap2	**97.88**	**99.964**	134 885	NA
	LRA	97.25	99.952	132 862	NA
	Winnowmap2	97.04	99.963	130 978	NA
Yeast (Illumina)	BLEND	**79.92**	**99.975**	6 493 730	95.88
	minimap2	79.91	99.974	6 492 994	95.89
	Strobealign	**79.92**	99.970	**6 498 380**	**97.59**
E. coli (PacBio)	BLEND	**97.51**	100	83 924	NA
	minimap2	97.29	100	**85 326**	NA
	LRA	93.61	100	80 802	NA
	Winnowmap2	89.78	100	69 884	NA

BLEND-I vs. BLEND-S (Accuracy)

Dataset	Tool	Average Identity (%)	Genome Fraction (%)	K-mer Compl. (%)	Aligned Length (Mbp)	Mismatch per 100Kbp (#)	Average GC (%)	Assembly Length (Mbp)	Largest Contig (Mbp)	NGA50 (Kbp)	NG50 (Kbp)
CHM13 (HiFi)	BLEND-I	99.7535	96.7203	83.65	3,054.49	48.49	**40.79**	**3,059.29**	**41.8342**	**8,507.53**	**8,508.92**
	BLEND-S	**99.8526**	**98.4847**	**90.15**	**3,092.54**	**22.02**	40.78	3,095.21	22.8397	5,442.25	5,442.31
	Reference	100	100	100	3,054.83	0.00	40.85	3,054.83	248.387	154,260	154,260
D. ananassae (HiFi)	BLEND-I	99.6890	97.2290	77.85	**270.218**	233.18	41.95	280.388	5.01099	356.745	356.745
	BLEND-S	**99.7856**	**97.2308**	**86.43**	240.391	**143.13**	**41.75**	**247.153**	**6.23256**	**792.407**	**798.913**
	Reference	100	100	100	213.805	0.00	41.81	213.818	30.6728	26,427.4	26,427.4
E. coli (HiFi)	BLEND-I	99.6902	**99.8824**	79.36	5.04157	17.92	**50.52**	**5.04263**	**4.94601**	**4,025.48**	**4,946.01**
	BLEND-S	**99.8320**	99.8801	**87.91**	**5.12155**	**3.77**	50.53	5.12155	3.41699	3,416.99	3,416.99
	Reference	100	100	100	5.04628	0.00	50.52	5.04628	4.94446	4,944.46	4,944.46
CHM13 (ONT)	BLEND-I	N/A	N/A	**29.26**	**2,891.28**	4,077.53	**41.32**	**2,897.87**	**25.2071**	**5,061.52**	**5,178.59**
	BLEND-S	N/A	N/A	0	0.010546	**3,250.70**	51.30	0.010548	0.010548	0	0
	Reference	100	100	100	3,117.29	0.00	40.75	3,117.29	248.387	150,617	150,617
Yeast (PacBio)	BLEND-I	89.1677	**97.0854**	**33.81**	12.3938	**2,672.37**	38.84	**12.4176**	**1.54807**	**635.966**	**636.669**
	BLEND-S	**90.3347**	83.8814	33.17	**22.9473**	4,795.58	**38.71**	22.9523	0.265118	114.125	116.143
	Reference	100	100	100	12.1571	0.00	38.15	12.1571	1.53193	924.431	924.431
Yeast (ONT)	BLEND-I	89.6889	**99.2974**	**35.95**	**12.3222**	2,529.47	38.64	**12.3225**	**1.10582**	**793.046**	**793.046**
	BLEND-S	**91.0865**	7.9798	4.90	0.898565	**2,006.91**	**38.35**	0.899654	0.043321	0	0
	Reference	100	100	100	12.1571	0.00	38.15	12.1571	1.53193	924.431	924.431
E. coli (PacBio)	BLEND-I	88.5806	**96.5238**	**32.32**	**5.90024**	1,857.56	**49.81**	**6.21598**	**2.40671**	**769.981**	**2,060.4**
	BLEND-S	**90.3551**	36.6230	17.07	2.10137	**1,299.50**	48.91	2.10704	0.095505	0	0
	Reference	100	100	100	5.6394	0.00	50.43	5.6394	5.54732	5,547.32	5,547.32

BLEND-I vs. BLEND-S (Performance and Memory)



Read Overlapping Parameters

Tool	Dataset	Parameters
BLEND	*CHM13 (HiFi)*	-x ava-hifi -t 32
BLEND	*D. ananassae (HiFi)*	-x ava-hifi -t 32
BLEND	*E. coli (HiFi)*	-x ava-hifi -t 32
BLEND	*CHM13 (ONT)*	-x ava-ont -t 32
BLEND	*Yeast (PacBio)*	-x ava-pb -t 32
BLEND	*Yeast (ONT)*	-x ava-ont -t 32
BLEND	*E. coli (PacBio)*	-x ava-pb -t 32
minimap2	*CHM13 (HiFi)*	-x ava-pb -Hk21 -w14 -t 32
minimap2	*D. ananassae (HiFi)*	-x ava-pb -Hk21 -w14 -t 32
minimap2	*E. coli (HiFi)*	-x ava-pb -Hk21 -w14 -t 32
minimap2	*CHM13 (ONT)*	-x ava-ont -t 32
minimap2	*Yeast (PacBio)*	-x ava-pb -t 32
minimap2	*Yeast (ONT)*	-x ava-ont -t 32
minimap2	*E. coli (PacBio)*	-x ava-pb -t 32
minimap2-Eq	*CHM13 (ONT)*	-x ava-ont -k19 -w10 -t 32
minimap2-Eq	*Yeast (PacBio)*	-x ava-pb -k23 -w10 -t 32
minimap2-Eq	*Yeast (ONT)*	-x ava-ont -k19 -w10 -t 32
minimap2-Eq	*E. coli (PacBio)*	-x ava-pb -k23 -w10 -t 32
MHAP	*CHM13 (HiFi)*	–store-full-id –ordered-kmer-size 18 –num-hashes 128 –num-min-matches 5 –ordered-sketch-size 1000 –threshold 0.95 –num-threads 32
MHAP	*D. ananassae (HiFi)*	–store-full-id –ordered-kmer-size 18 –num-hashes 128 –num-min-matches 5 –ordered-sketch-size 1000 –threshold 0.95 –num-threads 32
MHAP	*E. coli (HiFi)*	–store-full-id –ordered-kmer-size 18 –num-hashes 128 –num-min-matches 5 –ordered-sketch-size 1000 –threshold 0.95 –num-threads 32
MHAP	*Yeast (PacBio)*	–store-full-id –num-threads 32
MHAP	*Yeast (ONT)*	–store-full-id –num-threads 32
MHAP	*E. coli (PacBio)*	–store-full-id –num-threads 32

BLEND Parameter Definitions

Parameter	Definition
–strobemers	Use the `BLEND-S` mechanism when generating the list of k-mers of a seed
–immediate	Use the `BLEND-I` mechanism when generating the list of k-mers of a seed
-H	Use homopolymer-compressed k-mers
-w INT	Window size used when finding minimizers.
-k INT	k-mer size used when generating the list of k-mers of a seed
–neighbors INT	Number of k-mers included in the list of seeds. Combination of both -k (k) and –neighbors (n) determines the seed length. Seed length in `BLEND-S` is calculated as: $k \times n$ Seed length in `BLEND-I` is calculated as: $k+(n-1)$
–fixed-bits INT	Bit length of hash values that BLEND generates for each seed. Setting it to $2 \times k$ is the default behavior.
-t INT	Number of CPU threads to use.
-x STR	Preset for setting the default parameters given the use case (STR)
-x map-ont	Preset for mapping ONT reads. It uses the following parameters: –immediate -w 10 -k 9 –neighbors 7 –fixed-bits 30
-x map-pb	Preset for mapping erroneous PacBio reads. It uses the following parameters: –immediate -H -w 10 -k 13 –neighbors 7 –fixed-bits 32
-x map-hifi	Preset for mapping accurate long (HiFi) reads. It uses the following parameters: –strobemers -w 50 -k 19 –neighbors 5 –fixed-bits 38
-x sr	Preset for mapping short reads. It uses the following parameters: –immediate -w 11 -k 21 –neighbors 5 –fixed-bits 32
-x ava-ont	Preset for overlapping ONT reads. It uses the following parameters: –immediate -w 10 -k 15 –neighbors 5 –fixed-bits 30
-x ava-pb	Preset for overlapping erroneous PacBio reads. It uses the following parameters: –immediate -H -w 10 -k 19 –neighbors 5 –fixed-bits 38
-x ava-hifi	Preset for overlapping accurate long (HiFi) reads. It uses the following parameters: –strobemers -w 200 -k 25 –neighbors 7 –fixed-bits 50

Read Mapping Parameters

Tool	Dataset	Parameters
BLEND	*CHM13 (HiFi)*	-ax map-hifi -t 32 –secondary=no
BLEND	*HG002 (HiFi)*	-ax map-hifi -t 32 –secondary=no
BLEND	*D. ananassae (HiFi)*	-ax map-hifi -t 32 –secondary=no
BLEND	*E. coli (HiFi)*	-ax map-hifi -t 32 –secondary=no
BLEND	*CHM13 (ONT)*	-ax map-ont -t 32 –secondary=no
BLEND	*Yeast (PacBio)*	-ax map-pb -t 32 –secondary=no
BLEND	*Yeast (ONT)*	-ax map-ont -t 32 –secondary=no
BLEND	*Yeast (Illumina)*	-ax sr -t 32
BLEND	*E. coli (PacBio)*	-ax map-pb -t 32 –secondary=no
minimap2	*CHM13 (HiFi)*	-ax map-hifi -t 32 –secondary=no
minimap2	*HG002 (HiFi)*	-ax map-hifi -t 32 –secondary=no
minimap2	*D. ananassae (HiFi)*	-ax map-hifi -t 32 –secondary=no
minimap2	*E. coli (HiFi)*	-ax map-hifi -t 32 –secondary=no
minimap2	*CHM13 (ONT)*	-ax map-ont -t 32 –secondary=no
minimap2	*Yeast (PacBio)*	-ax map-pb -t 32 –secondary=no
minimap2	*Yeast (ONT)*	-ax map-ont -t 32 –secondary=no
minimap2	*Yeast (Illumina)*	-ax sr -t 32
minimap2	*E. coli (PacBio)*	-ax map-pb -t 32 –secondary=no
Winnowmap2	*CHM13 (HiFi)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-pb -t 32
Winnowmap2	*HG002 (HiFi)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-pb -t 32
Winnowmap2	*D. ananassae (HiFi)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-pb -t 32
Winnowmap2	*E. coli (HiFi)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-pb -t 32
Winnowmap2	*CHM13 (ONT)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-ont -t 32
Winnowmap2	*Yeast (PacBio)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-pb-clr -t 32
Winnowmap2	*Yeast (ONT)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-ont -t 32
Winnowmap2	*E. coli (PacBio)*	meryl count k=15 meryl print greater-than distinct=0.9998 -ax map-pb-clr -t 32
LRA	*CHM13 (HiFi)*	align -CCS -t 32 -p s
LRA	*HG002 (HiFi)*	align -CCS -t 32 -p s
LRA	*D. ananassae (HiFi)*	align -CCS -t 32 -p s
LRA	*E. coli (HiFi)*	align -CCS -t 32 -p s
LRA	*CHM13 (ONT)*	align -ONT -t 32 -p s
LRA	*Yeast (PacBio)*	align -CLR -t 32 -p s
LRA	*Yeast (ONT)*	align -ONT -t 32 -p s
LRA	*E. coli (PacBio)*	align -CLR -t 32 -p s
S-conLSH	*CHM13 (HiFi)*	–threads 32 –align 1
S-conLSH	*E. coli (HiFi)*	–threads 32 –align 1
S-conLSH	*CHM13 (ONT)*	–threads 32 –align 1
S-conLSH	*Yeast (PacBio)*	–threads 32 –align 1
S-conLSH	*Yeast (ONT)*	–threads 32 –align 1
S-conLSH	*E. coli (PacBio)*	–threads 32 –align 1
Strobealign	*Yeast (Illumina)*	-t 32

Versions of each Tool

Tool	Version	GitHub or Conda Link to the Version
BLEND	1.0	https://github.com/CMU-SAFARI/BLEND
minimap2	2.24	https://github.com/lh3/minimap2/releases/tag/v2.24
MHAP	2.1.3	https://anaconda.org/bioconda/mhap/2.1.3/download/noarch/mhap-2.1.3-hdfd78af_1.tar.bz2
LRA	1.3.2	https://anaconda.org/bioconda/lra/1.3.2/download/linux-64/lra-1.3.2-ha140323_0.tar.bz2
Winnowmap2	2.03	https://anaconda.org/bioconda/Winnowmap/2.03/download/linux-64/Winnowmap2-2.03-h2e03b76_0.tar.bz2
S-conLSH	2.0	https://github.com/anganachakraborty/S-conLSH-2.0/tree/292fbe0405f10b3ab63fc3a86cba2807597b582e
Strobealign	0.7.1	https://anaconda.org/bioconda/strobealign/0.7.1/download/linux-64/strobealign-0.7.1-hd03093a_1.tar.bz2

BLEND Backup Slide