Intelligent Genome Analysis via Intelligent Algorithms and Architectures

Mohammed Alser, PhD

@mealser

IGGSy 2022

7 July 2022

IGGSY 2022
July 3-7

SAFARI
SAFARI Research Group

ETH zürich

Plenty of Room at the Indexing

Misconception: Indexing step is built only once for each reference genome,

why we should care about its performance!

Fact: Indexing step affects the performance of all steps of read mapping, as it decides on the number of seeds and their locations.

Reducing the size of the index can speed up read mapping

Genome Index Properties

- Seeds can be Strobemers, Syncmers, BLEND, LSH, overlapping, non-overlapping, spaced, adjacent, non-adjacent, minimizers, compressed, …

[illegible]	[illegible]	[illegible]	Indexing	Indexing
				in
r				n
B				in
BWA-MEM2**	2.2.1	17 GB	default	33.36 min

None of the existing methods try to reduce the size of the reference genome.

*Human genome = 3.2 GB (char-encoded) or 1.6 GB (4bit-encoded)

**Its peak memory = 72.3 GB, minimap2 =11.4 GB when building the index.

Compressed Genomic Analyses

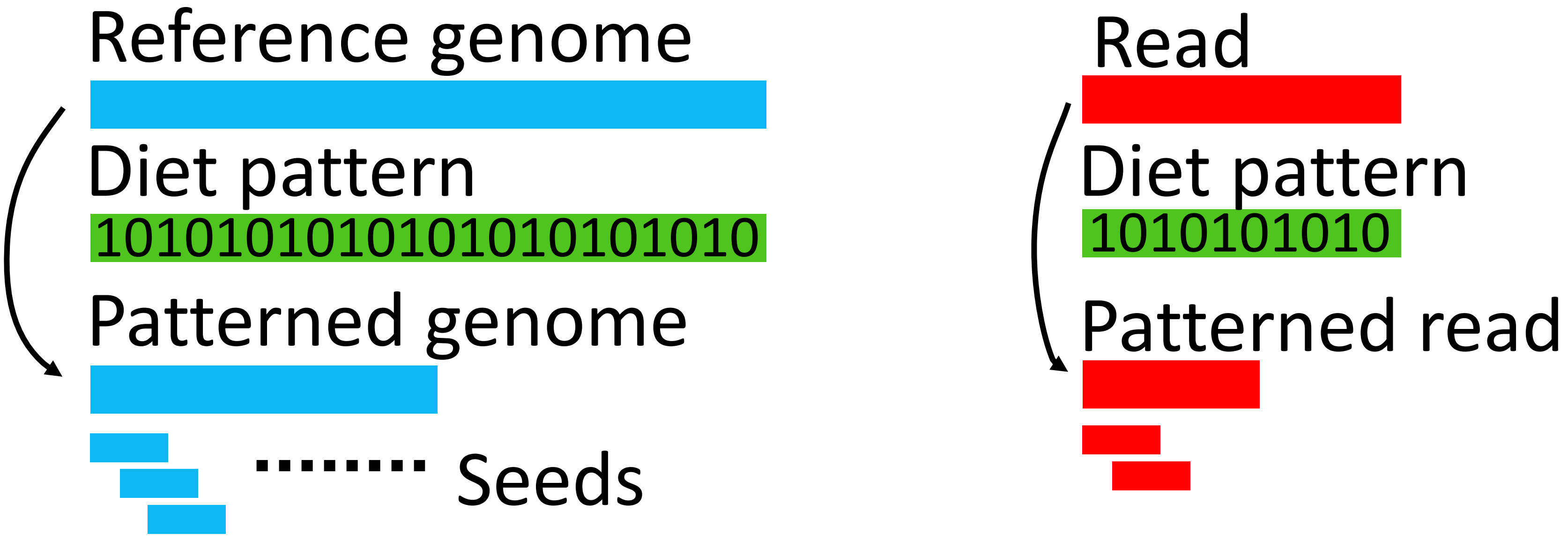
Can we process compressed genomic sequences without decompression?

Image adapted from: https://joinzoe.com/post/genetic-diet-nutrition

Genome-on-Diet Steps

- Compressed Indexing
- Pattern Alignment
- Compressed Seeding
- Location Voting
- Sequence Alignment

Step 1: Compressed Indexing



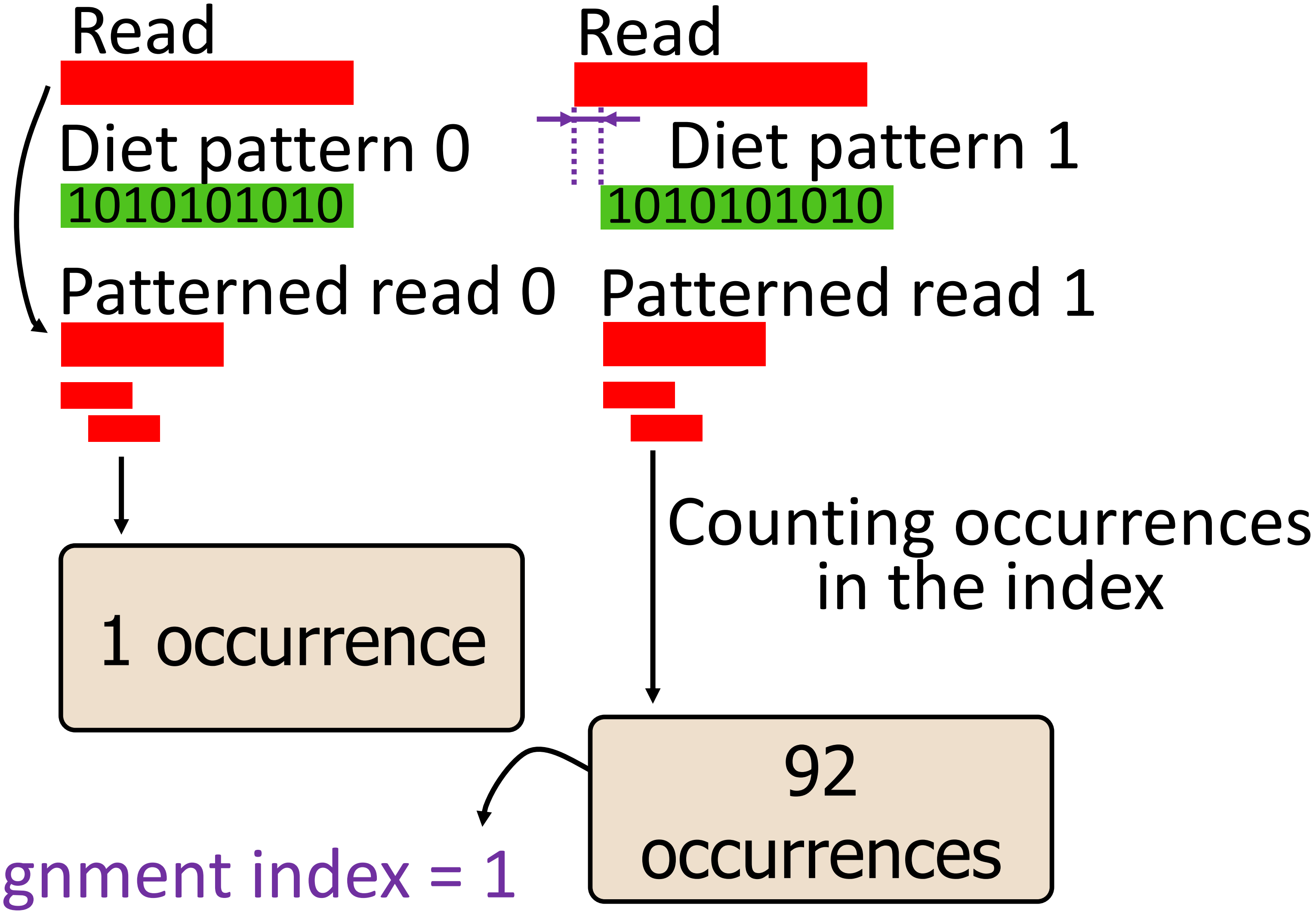
Easy! Isn't it?

Where to Start Applying the Pattern?

```
ACCCTAACCCTAACCCTAACCCTAACCCTAA

A_C_T_A_C_T_A_C_T_A_C_T_A_C_T_A
A_C_T_A_C_T_A_C_
  C_T_A_C_T_A_C_T_
    T_A_C_T_A_C_T_A_
      A_C_T_A_C_T_A_C_

_CCCTAACCCTAACCCTAACCCTAACCCTAA

_C_C_A_C_C_A_C_C_A_C_C_A_C_C_A_
_C_C_A_C_C_A_C_C
  _C_A_C_C_A_C_C_A
    _A_C_C_A_C_C_A_C
      _C_C_A_C_C_A_C_C
```

No Match ☹

Step 2: Pattern Alignment



Read

Diet pattern 0

1010101010

Patterned read 0

1 occurrence

Read

Diet pattern 1

1010101010

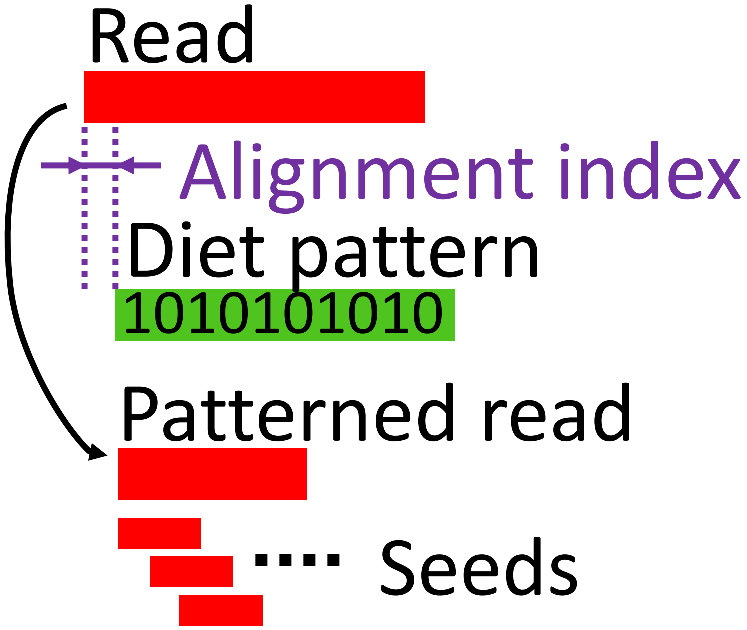
Patterned read 1

Counting occurrences in the index

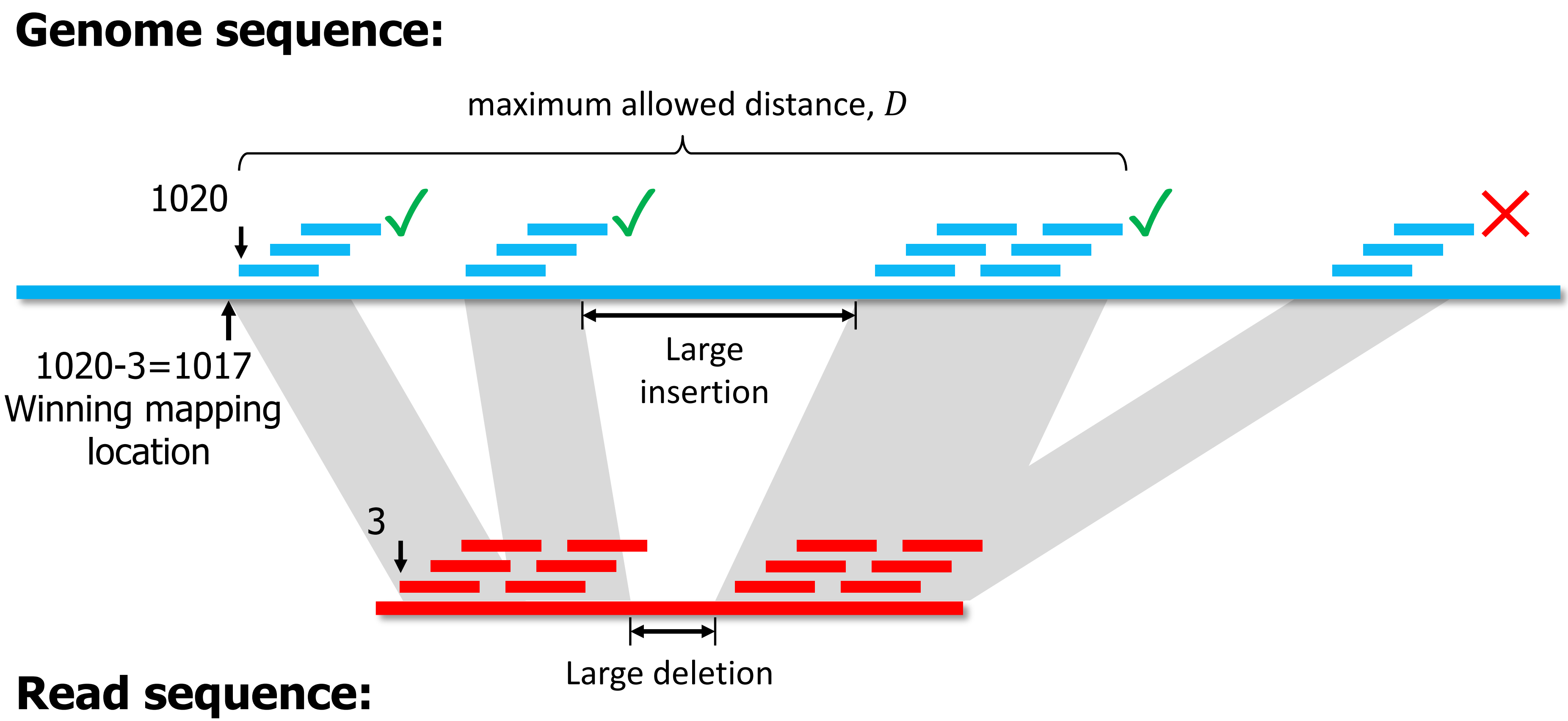
92 occurrences

Alignment index = 1

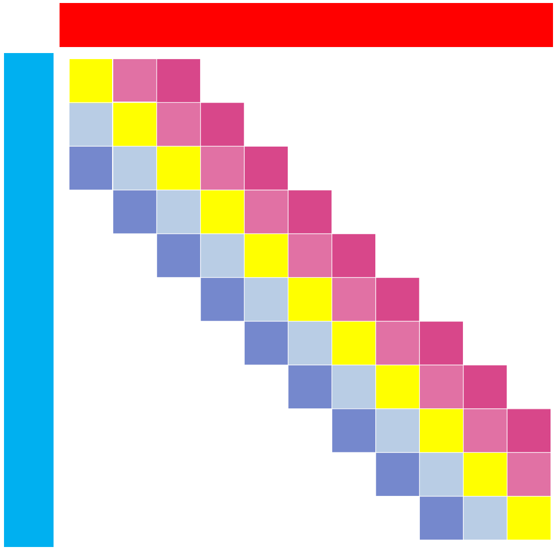
Step 3: Compressed Seeding



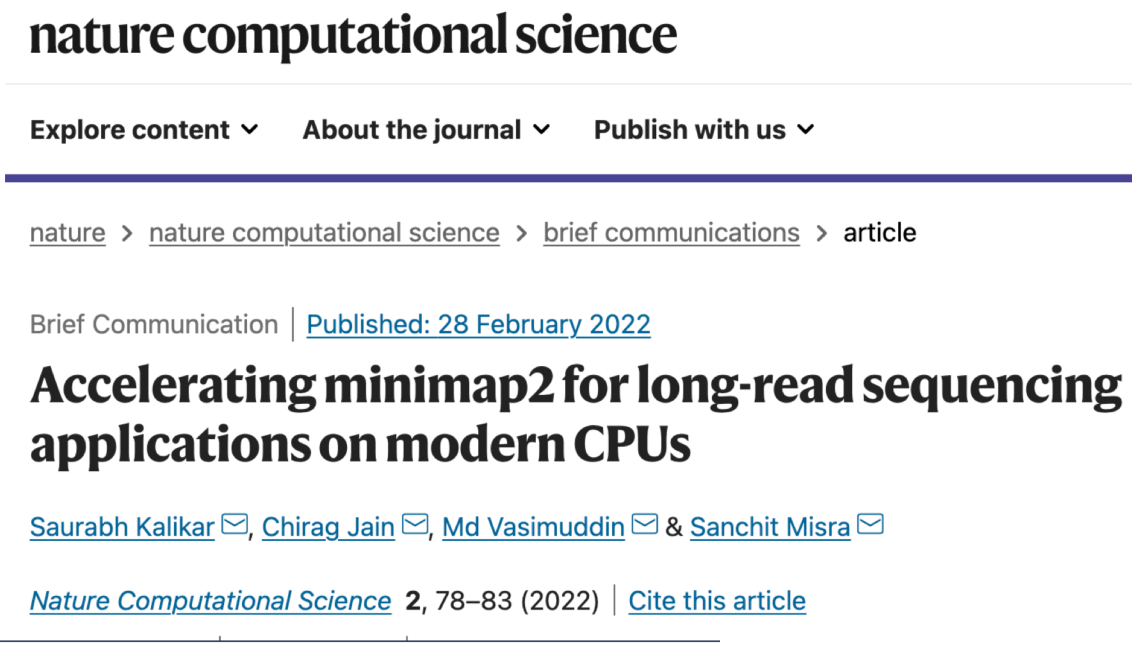
Step 4: Location Voting



Step 5: Sequence Alignment



Dynamic programming matrix



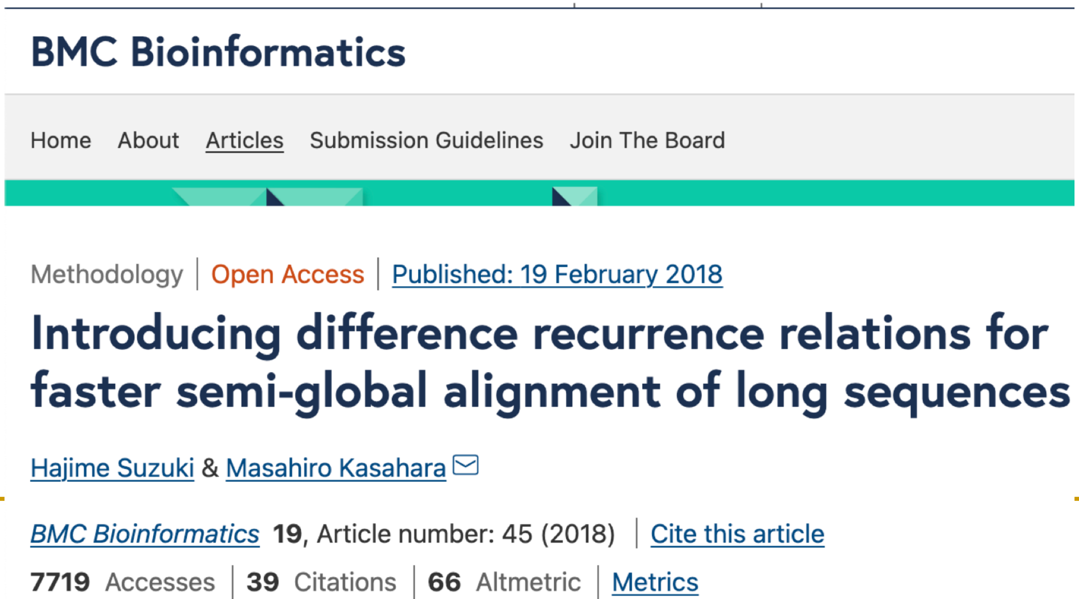
nature computational science
Explore content
About the journal
Publish with us
nature > nature computational science > brief communications > article

Brief Communication | Published: 28 February 2022

Accelerating minimap2 for long-read sequencing applications on modern CPUs

Saurabh Kalikar, Chirag Jain, Md Vasimuddin & Sanchit Misra

Nature Computational Science **2**, 78–83 (2022) | Cite this article



Methodology | Open Access | Published: 19 February 2018

Introducing difference recurrence relations for faster semi-global alignment of long sequences

Hajime Suzuki & Masahiro Kasahara

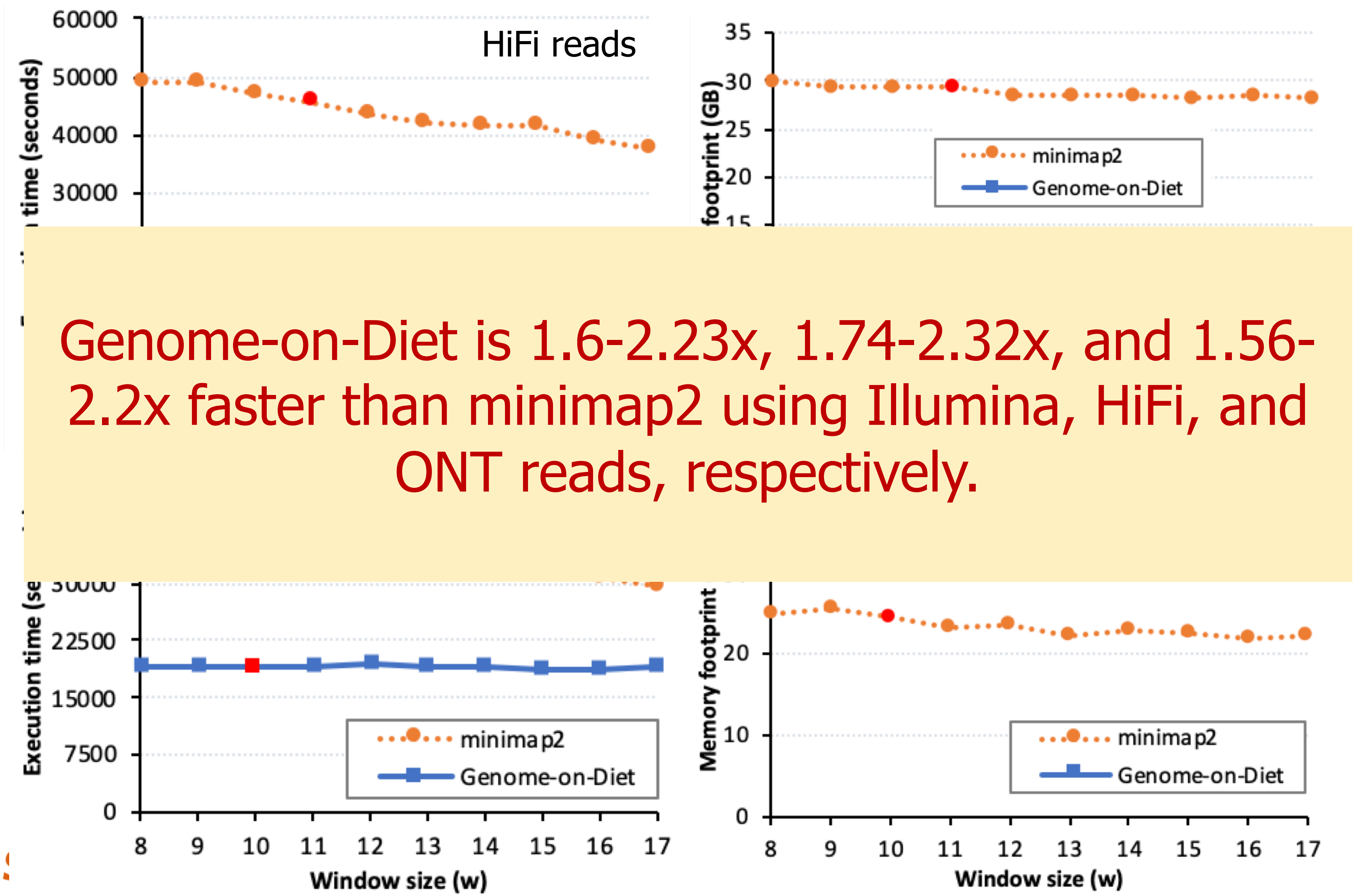
BMC Bioinformatics **19**, Article number: 45 (2018) | Cite this article

7719 Accesses | **39** Citations | **66** Altmetric | Metrics

Introducing Five Optimization Strategies

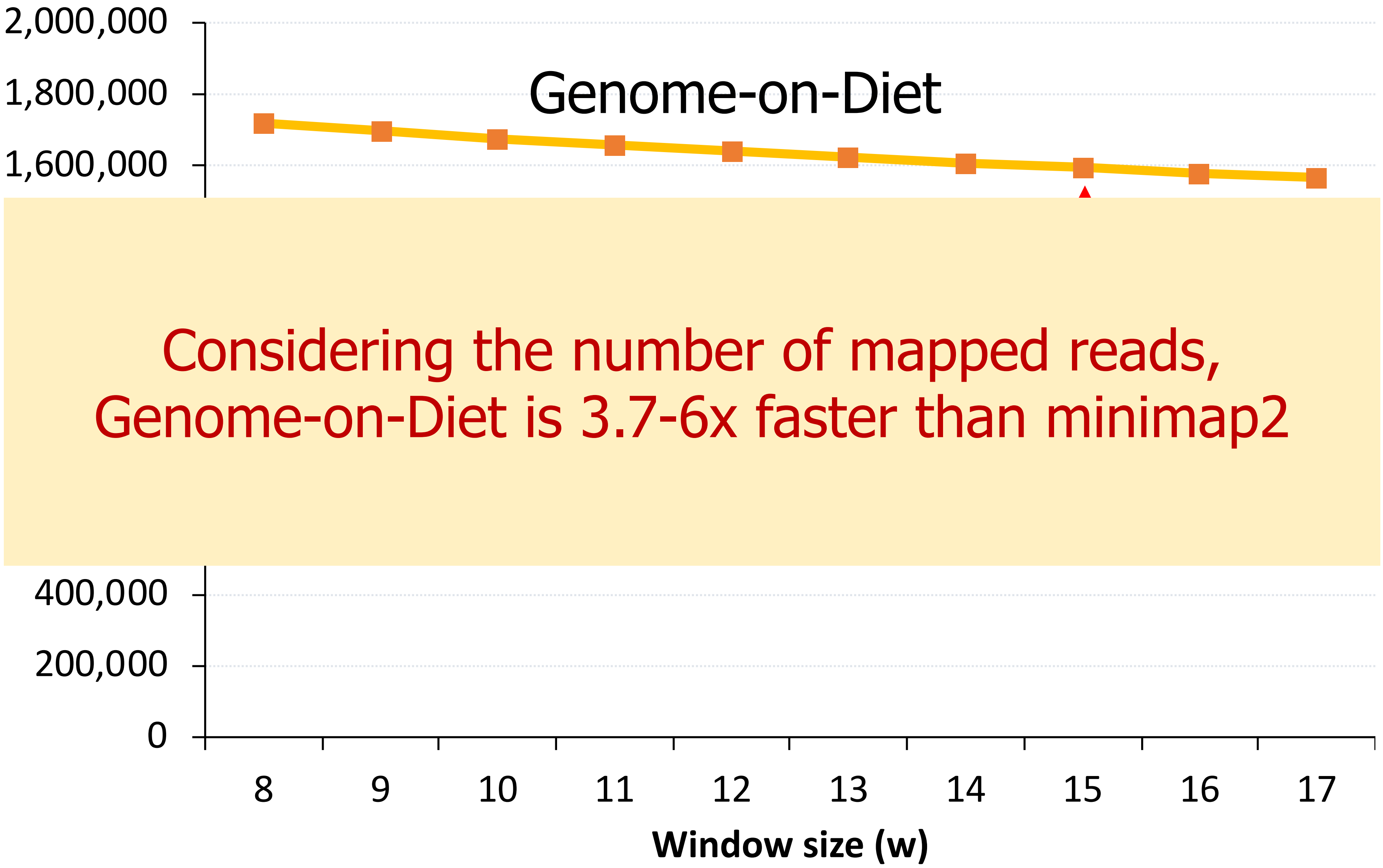
- Accelerating Indexing & Seeding with **SIMD Instructions**
- **Sorting** Seed Locations
- **Progressive** Compressed Seeding
- **Rescuing** Mapping Location
- Handling **Exactly-Matching** Short Reads

Time & Memory Footprint (Long Reads)



Genome-on-Diet is 1.6-2.23x, 1.74-2.32x, and 1.56-2.2x faster than minimap2 using Illumina, HiFi, and ONT reads, respectively.

Number of Mapped Reads



Considering the number of mapped reads, Genome-on-Diet is 3.7-6x faster than minimap2

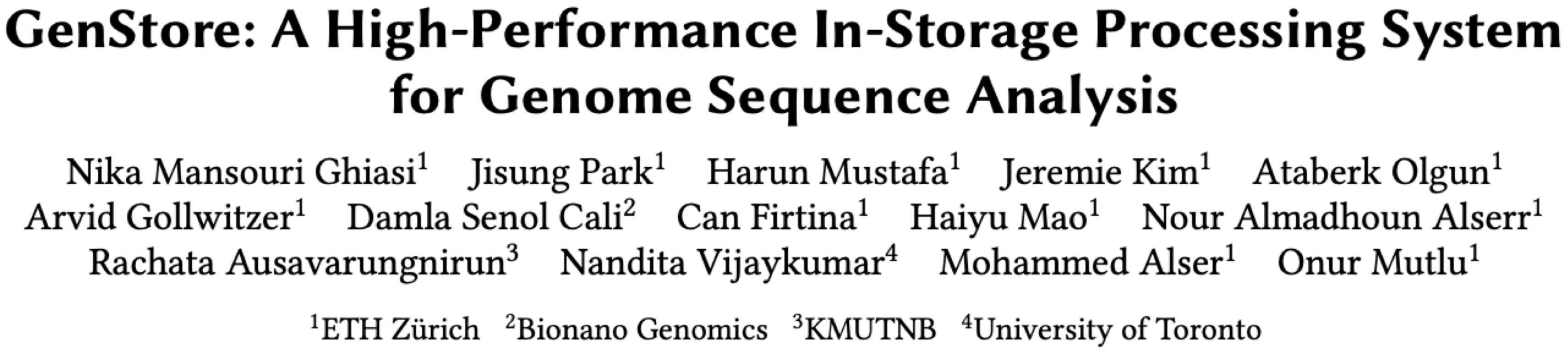
Other Important Results

- Genome-on-Diet provides 3.36x, 10.2x, and 17.53x higher number of mapped reads with the **highest mapping quality** (MAPQ=60) than minimap2.

- Both Genome-on-Diet and minimap2 **agree on the mapping locations** in 91%, 84%, and 73% of high-quality reads (MAPQ=60) mapped by both tools when using Illumina, HiFi, and ONT reads.

https://github.com/CMU-SAFARI/Genome-on-Diet

GenStore (ASPLOS 2022)



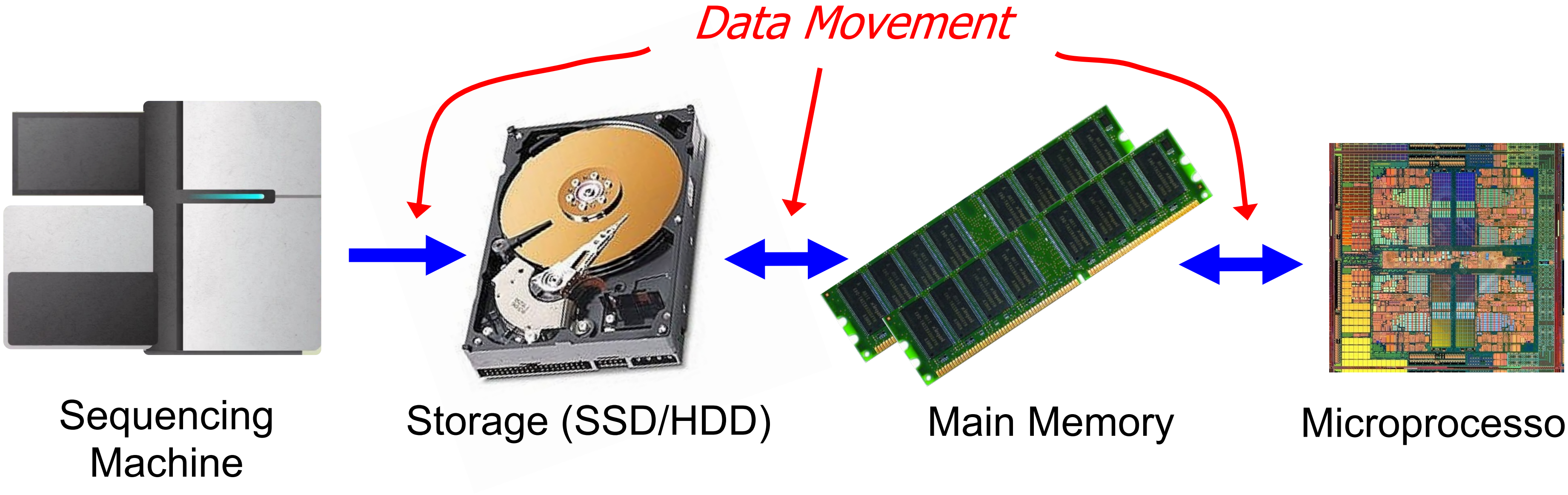
GenStore: A High-Performance In-Storage Processing System for Genome Sequence Analysis

Nika Mansouri Ghiasi[1] Jisung Park[1] Harun Mustafa[1] Jeremie Kim[1] Ataberk Olgun[1]
Arvid Gollwitzer[1] Damla Senol Cali[2] Can Firtina[1] Haiyu Mao[1] Nour Almadhoun Alserr[1]
Rachata Ausavarungnirun[3] Nandita Vijaykumar[4] Mohammed Alser[1] Onur Mutlu[1]

[1]ETH Zürich [2]Bionano Genomics [3]KMUTNB [4]University of Toronto

Data Movement Dominates Performance

- **Data movement** dominates performance and is a **major** system **energy bottleneck** (accounting for 40%-62%)



Data Movement

Sequencing Machine

Storage (SSD/HDD)

Main Memory

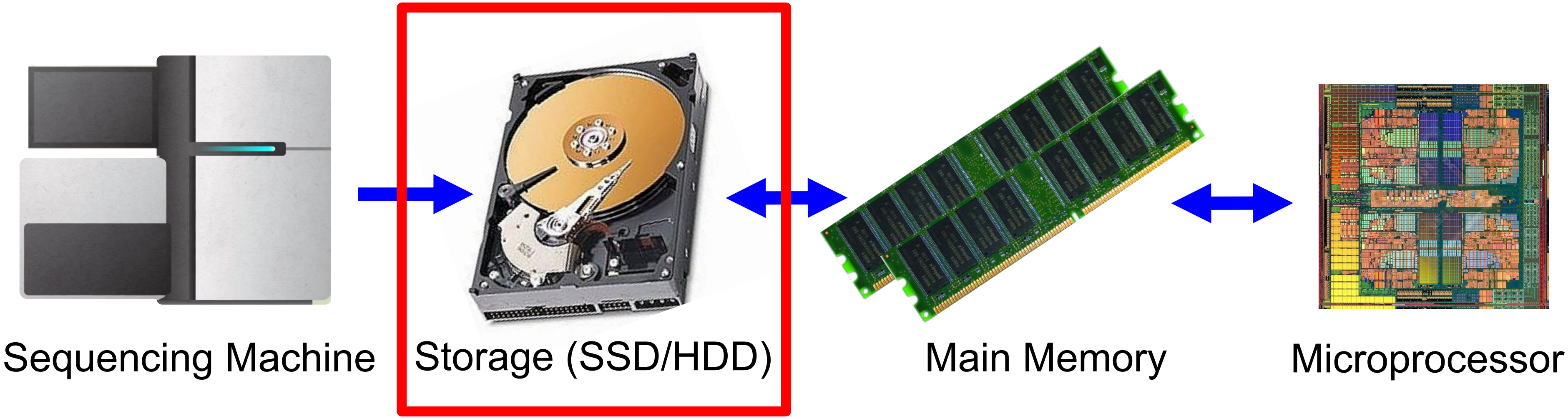
Microprocessor

✻ Boroumand et al., "Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks," ASPLOS 2018
★ Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013
☆ Pandiyan and Wu, "Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms," IISWC 2014

Key Ideas of GenStore (ASPLOS 2022)

GenStore-EM (exactly-matching reads filter): In some cases, a large fraction of reads exactly match to subsequences of the reference genome.

GenStore-NM (non-matching reads filter): In some cases, a large fraction of reads do not match to subsequences of the reference genome.



Sequencing Machine

Storage (SSD/HDD)

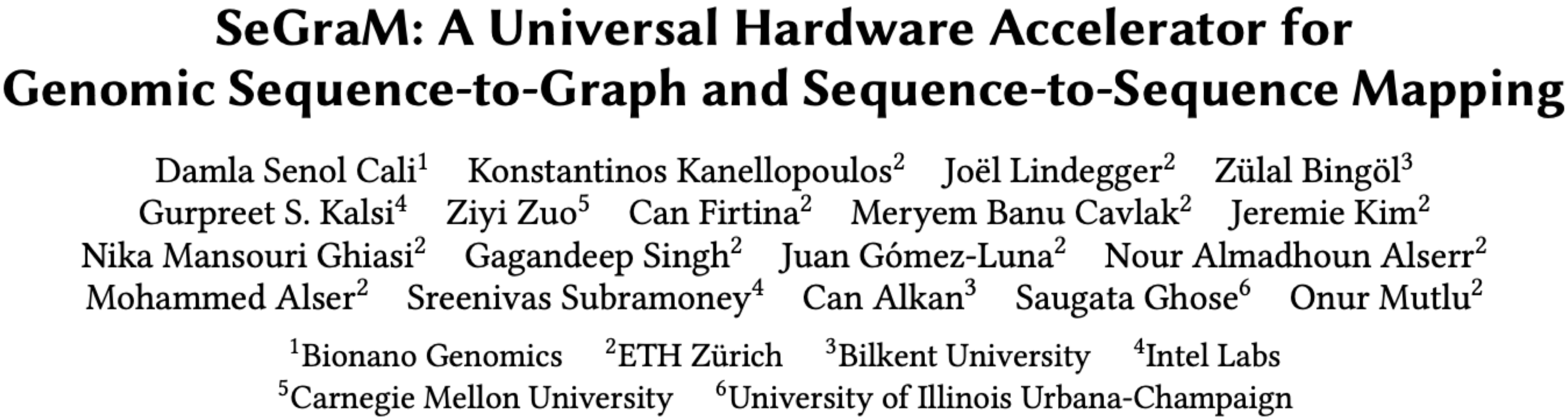
Main Memory

Microprocessor

GenStore-EM: 2.1-6.1× speedup & 3.92x energy saving compared to minimap2.

GenStore-NM: 1.4-33.6x speedup & 27.17x energy saving compared to minimap2.

SeGraM (HPCA 2022)



SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping

Damla Senol Cali[1] Konstantinos Kanellopoulos[2] Joël Lindegger[2] Zülal Bingöl[3]
Gurpreet S. Kalsi[4] Ziyi Zuo[5] Can Firtina[2] Meryem Banu Cavlak[2] Jeremie Kim[2]
Nika Mansouri Ghiasi[2] Gagandeep Singh[2] Juan Gómez-Luna[2] Nour Almadhoun Alserr[2]
Mohammed Alser[2] Sreenivas Subramoney[4] Can Alkan[3] Saugata Ghose[6] Onur Mutlu[2]

[1]Bionano Genomics [2]ETH Zürich [3]Bilkent University [4]Intel Labs
[5]Carnegie Mellon University [6]University of Illinois Urbana-Champaign

SeGraM: Universal Genomic Mapping Accelerator

- ***First universal genomic mapping accelerator*** that can support *both* sequence-to-graph mapping and sequence-to-sequence mapping, for *both* short and long reads

- ***First algorithm/hardware co-design*** for accelerating sequence-to-graph mapping

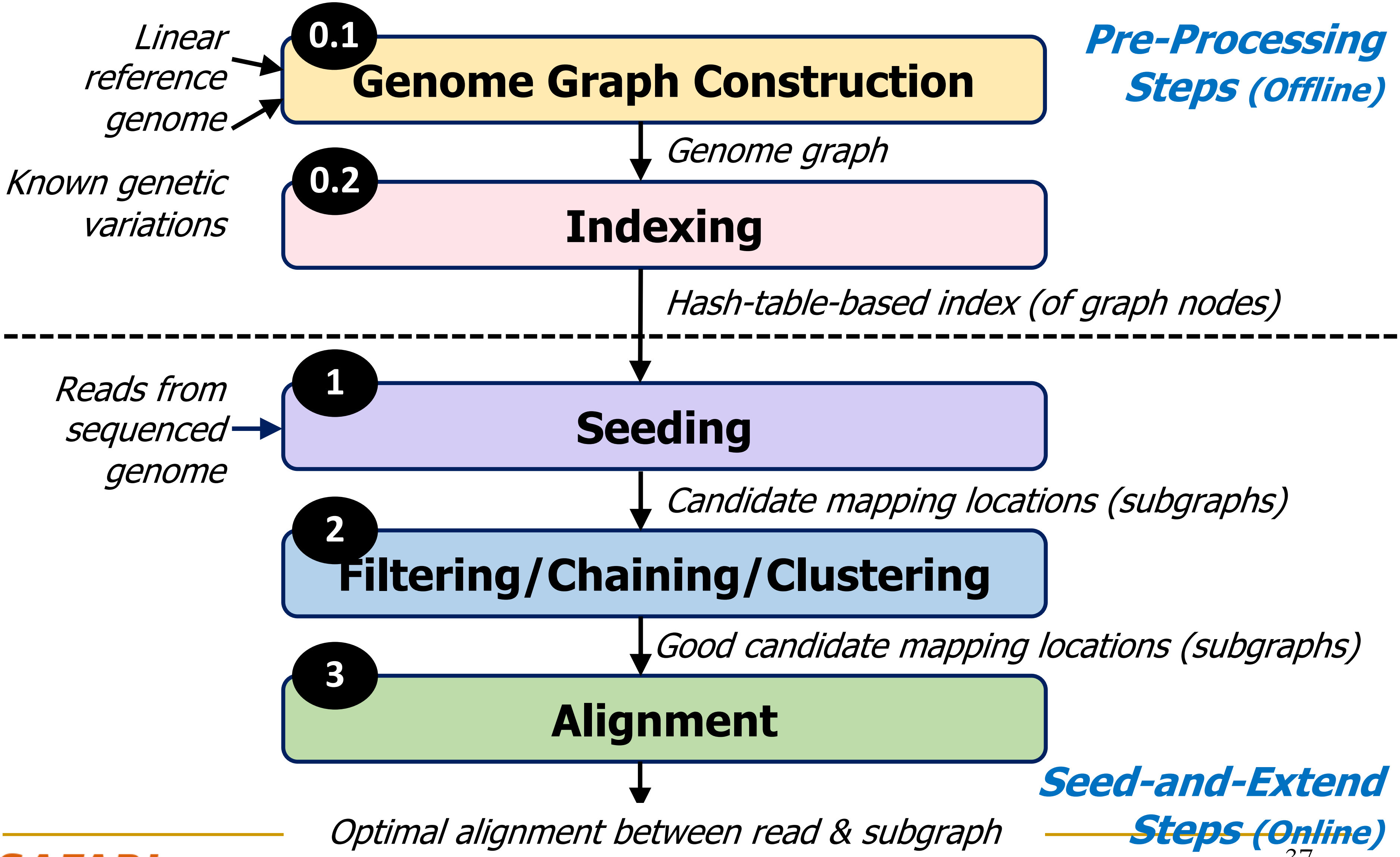
- We base SeGraM upon a minimizer-based seeding algorithm, and
- We propose a novel bitvector-based alignment algorithm to perform approximate string matching between a read and a graph-based reference genome

SW

- We co-design both algorithms with high-performance, scalable, and efficient hardware accelerators

HW

Sequence-to-Graph Mapping Pipeline



Linear reference genome

Known genetic variations

0.1 Genome Graph Construction

Pre-Processing Steps (Offline)

Genome graph

0.2 Indexing

Hash-table-based index (of graph nodes)

Reads from sequenced genome

1 Seeding

Candidate mapping locations (subgraphs)

2 Filtering/Chaining/Clustering

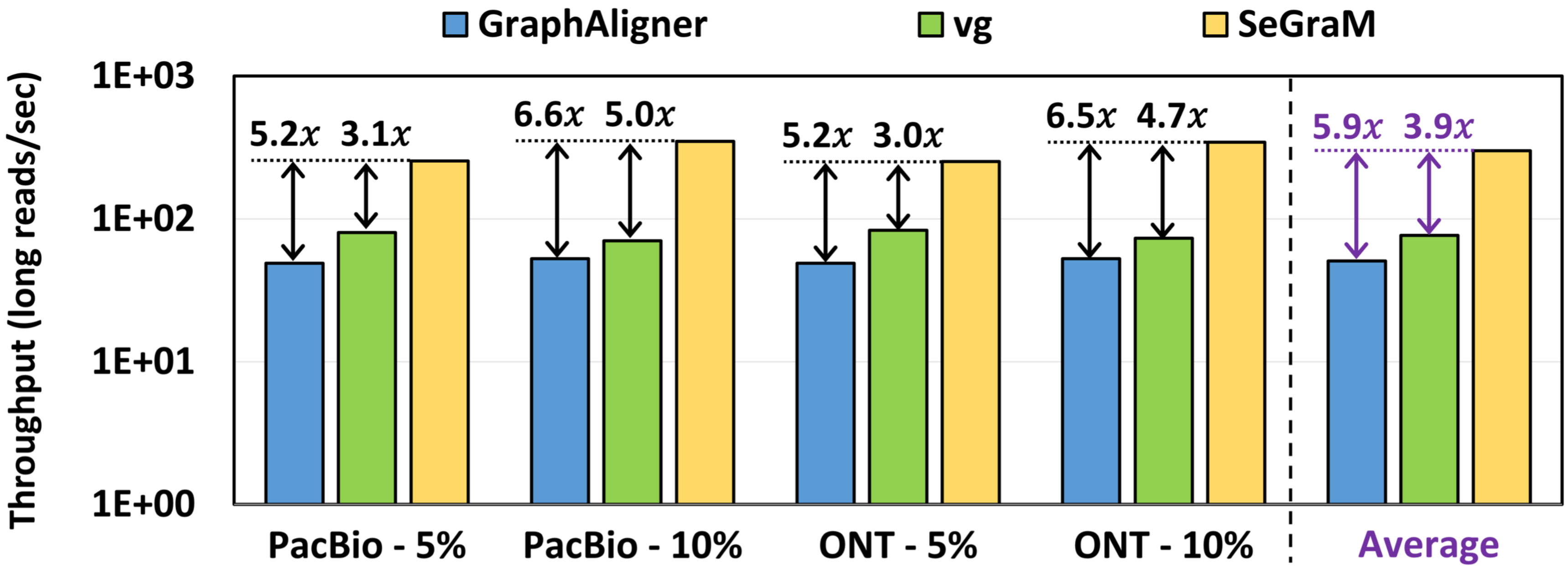
Good candidate mapping locations (subgraphs)

3 Alignment

Optimal alignment between read & subgraph

Seed-and-Extend Steps (Online)

Key Results – SeGraM with Long Reads



SeGraM provides **5.9× and 3.9× throughput improvement** over GraphAligner and vg,
while **reducing the power consumption by 4.1× and 4.4×**

Our Contributions

Near-memory Sequence Alignment

- **GenASM [MICRO 2020]**
- **SeGraM [ISCA 2022]**

Near-memory/In-memory Pre-alignment Filtering

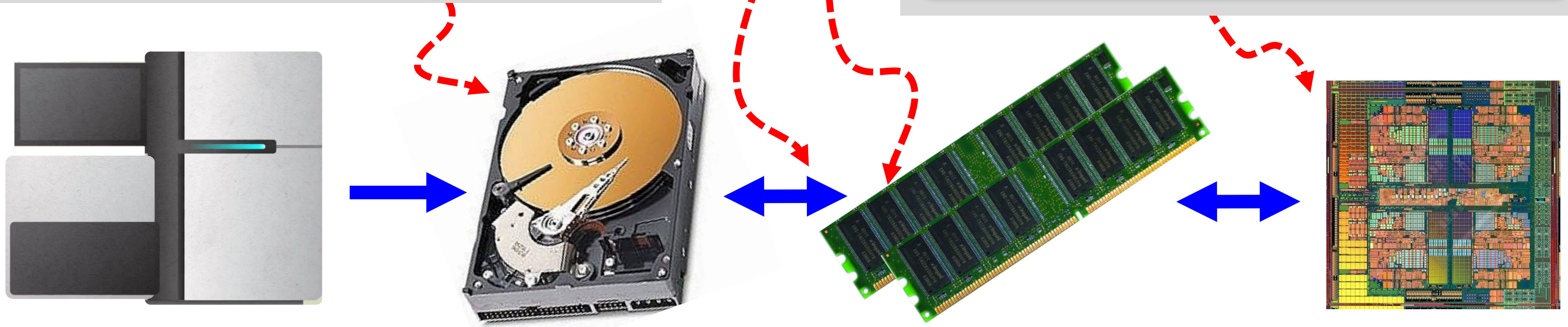
- **GRIM-Filter [BMC Genomics’18]**
- **SneakySnake [IEEE Micro’21]**
- **GenASM [MICRO 2020]**

Specialized Pre-alignment Filtering Accelerators (GPU, FPGA)

- **GateKeeper [Bioinformatics’17]**
- **MAGNET [AACBB’18]**
- **Shouji [Bioinformatics’19]**
- **GateKeeper-GPU [arXiv’21]**
- **SneakySnake [Bioinformatics’20]**

In-storage Sequence Alignment

- **GenStore [ASPLOS 2022]**



Our Contributions

Near-memory Sequence Alignment

GenASM [MICRO 2020]

SeGraM [ISCA 2022]

Near-memory/In-memory Pre-alignment Filtering

GRIM-Filter [BMC Genomics'18]

S

G

Specialized Pre-alignment Filtering Accelerators (GPU, FPGA)

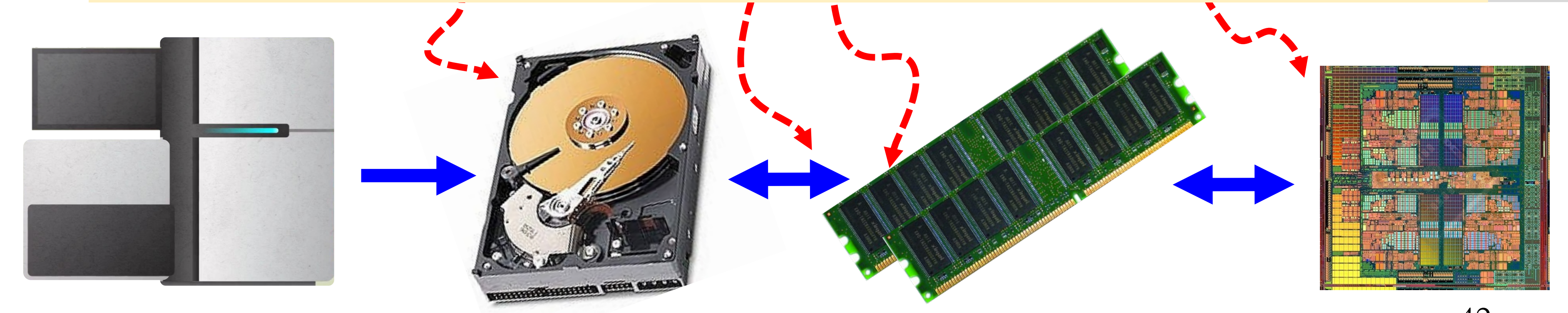
]

20]

In

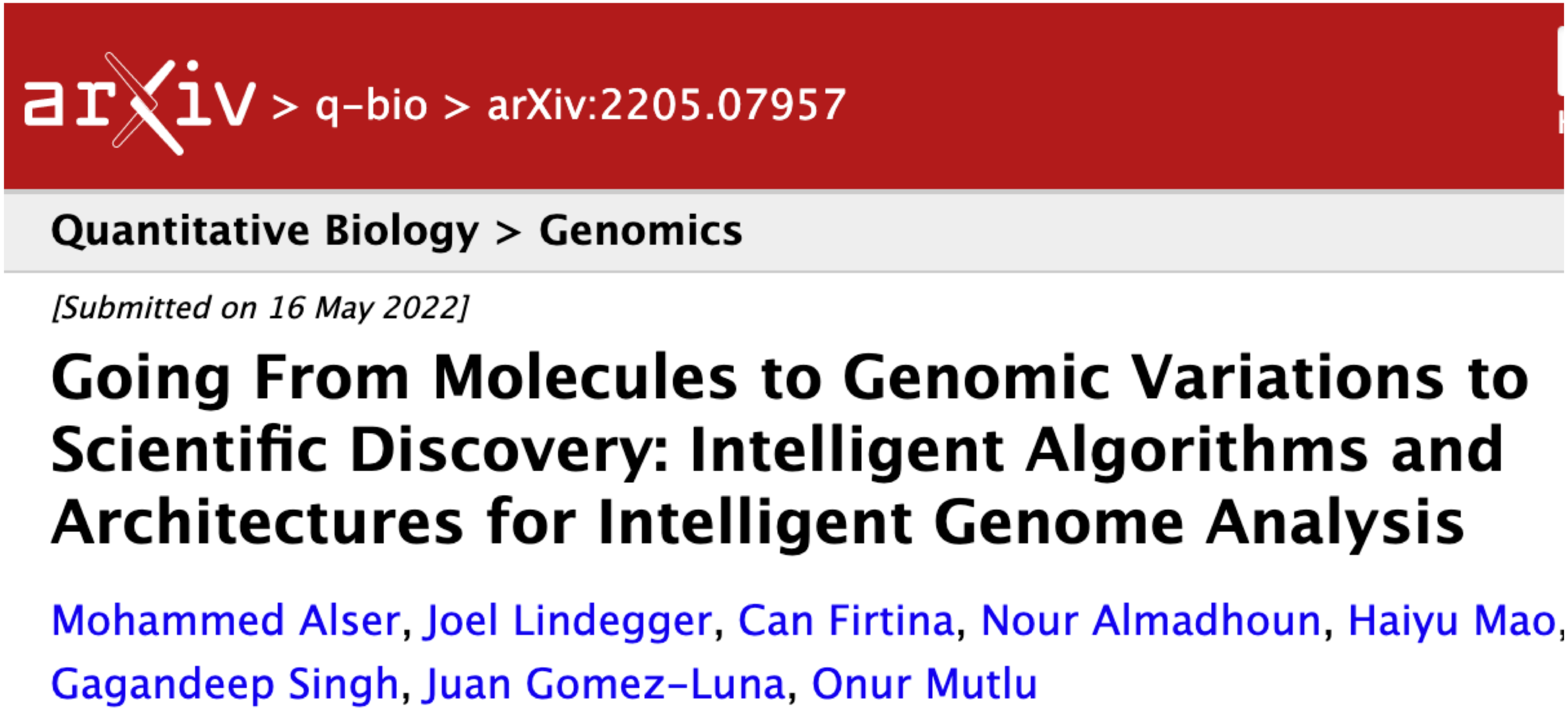
Ge

Improving **performance** and **energy efficiency** by 1-3 orders of magnitude



Intelligent Genome Analysis

https://arxiv.org/abs/2205.07957



arXiv > q-bio > arXiv:2205.07957

Quantitative Biology > Genomics

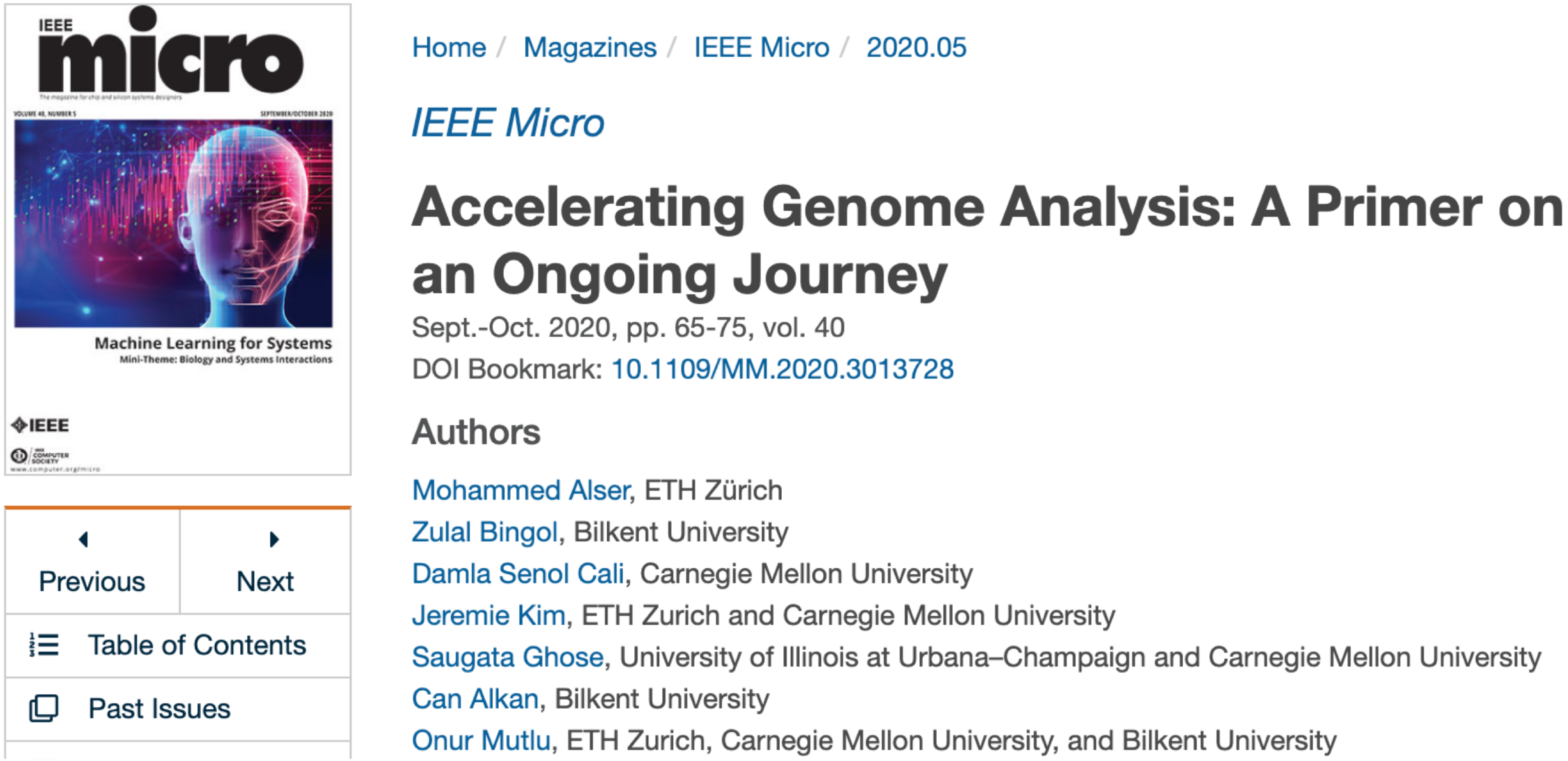
[Submitted on 16 May 2022]

Going From Molecules to Genomic Variations to Scientific Discovery: Intelligent Algorithms and Architectures for Intelligent Genome Analysis

Mohammed Alser, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu

Accelerating Genome Analysis

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu
"Accelerating Genome Analysis: A Primer on an Ongoing Journey"
IEEE Micro, August 2020.


IEEE
micro
The magazine for chip and silicon systems designers
VOLUME 40, NUMBER 5
SEPTEMBER/OCTOBER 2020
Machine Learning for Systems
Mini-Theme: Biology and Systems Interactions
IEEE
IEEE COMPUTER SOCIETY
www.computer.org/micro
Previous
Next
Table of Contents
Past Issues
Home / Magazines / IEEE Micro / 2020.05

IEEE Micro

Accelerating Genome Analysis: A Primer on an Ongoing Journey

Sept.-Oct. 2020, pp. 65-75, vol. 40

DOI Bookmark: 10.1109/MM.2020.3013728

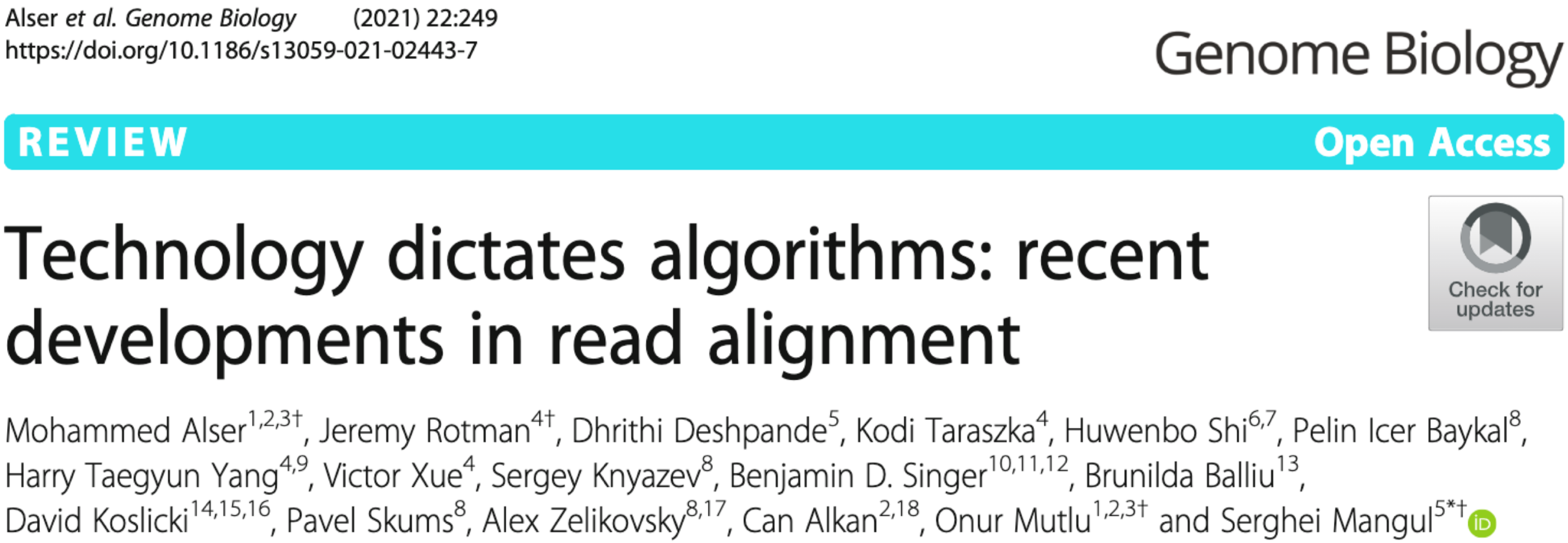
Authors

Mohammed Alser, ETH Zürich
Zulal Bingol, Bilkent University
Damla Senol Cali, Carnegie Mellon University
Jeremie Kim, ETH Zurich and Carnegie Mellon University
Saugata Ghose, University of Illinois at Urbana–Champaign and Carnegie Mellon University
Can Alkan, Bilkent University
Onur Mutlu, ETH Zurich, Carnegie Mellon University, and Bilkent University

Read Mapping in 111 pages!

In-depth analysis of 107 read mappers (1988-2020)

Mohammed Alser, Jeremy Rotman, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D. Singer, Brunilda Balliu, David Koslicki, Pavel Skums, Alex Zelikovsky, Can Alkan, Onur Mutlu, Serghei Mangul
"Technology dictates algorithms: Recent developments in read alignment"
Genome Biology, 2021
[Source code]



Alser *et al. Genome Biology* (2021) 22:249
https://doi.org/10.1186/s13059-021-02443-7

Genome Biology

REVIEW Open Access

Technology dictates algorithms: recent developments in read alignment

Mohammed Alser[1,2,3†], Jeremy Rotman[4†], Dhrithi Deshpande[5], Kodi Taraszka[4], Huwenbo Shi[6,7], Pelin Icer Baykal[8], Harry Taegyun Yang[4,9], Victor Xue[4], Sergey Knyazev[8], Benjamin D. Singer[10,11,12], Brunilda Balliu[13], David Koslicki[14,15,16], Pavel Skums[8], Alex Zelikovsky[8,17], Can Alkan[2,18], Onur Mutlu[1,2,3†] and Serghei Mangul[5*†]

Key Takeaway

Most speedup comes from

parallelism enabled by

novel architectures and algorithms

SAFARI Research Group

SAFARI

SAFARI Research Group

safari.ethz.ch

Think BIG, Aim HIGH!

https://safari.ethz.ch

Contributors

Mohammed Alser

Julien Eudine

Onur Mutlu

Can Alkan

Damla Senol Cali

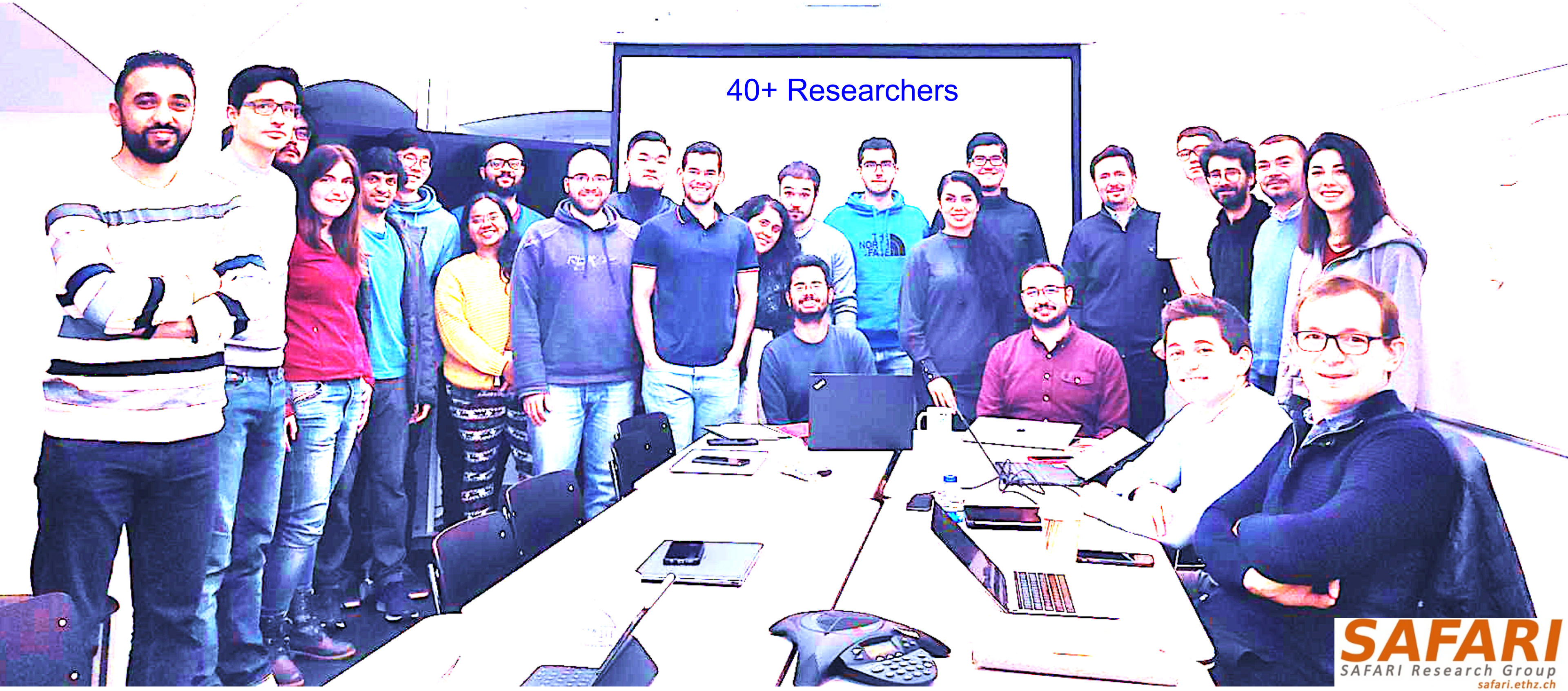
Nika Mansourighiasi

And many more …

SAFARI Research Group

Computer architecture, HW/SW, systems, bioinformatics, security, memory

https://safari.ethz.ch/safari-newsletter-december-2021/



Think BIG, Aim HIGH!

https://safari.ethz.ch

Intelligent Genome Analysis via Intelligent Algorithms and Architectures

Mohammed Alser, PhD

@mealser

IGGSy 2022

7 July 2022

IGGSY 2022
July 3-7

SAFARI
SAFARI Research Group

ETH zürich