

SPONSORED BY THE



Federal Ministry
of Education
and Research



ARAB-GERMAN
YOUNG ACADEMY
OF SCIENCES AND
HUMANITIES

International Workshop of the
Arab-German Young Academy of Sciences and Humanities (AGYA)

Introduction of Bioinformatics to Arab Education Systems

21 – 22 November 2022

Tunis, Tunisia



agya.info

Preparing Students To Rethink Genomic Analyses

Mohammed Alser

ETH Zurich

 @meals

Institut Pasteur de Tunis

21 November 2022



What is a **Genome**?



An organism's complete set of genetic instructions

How to Analyze a Genome?



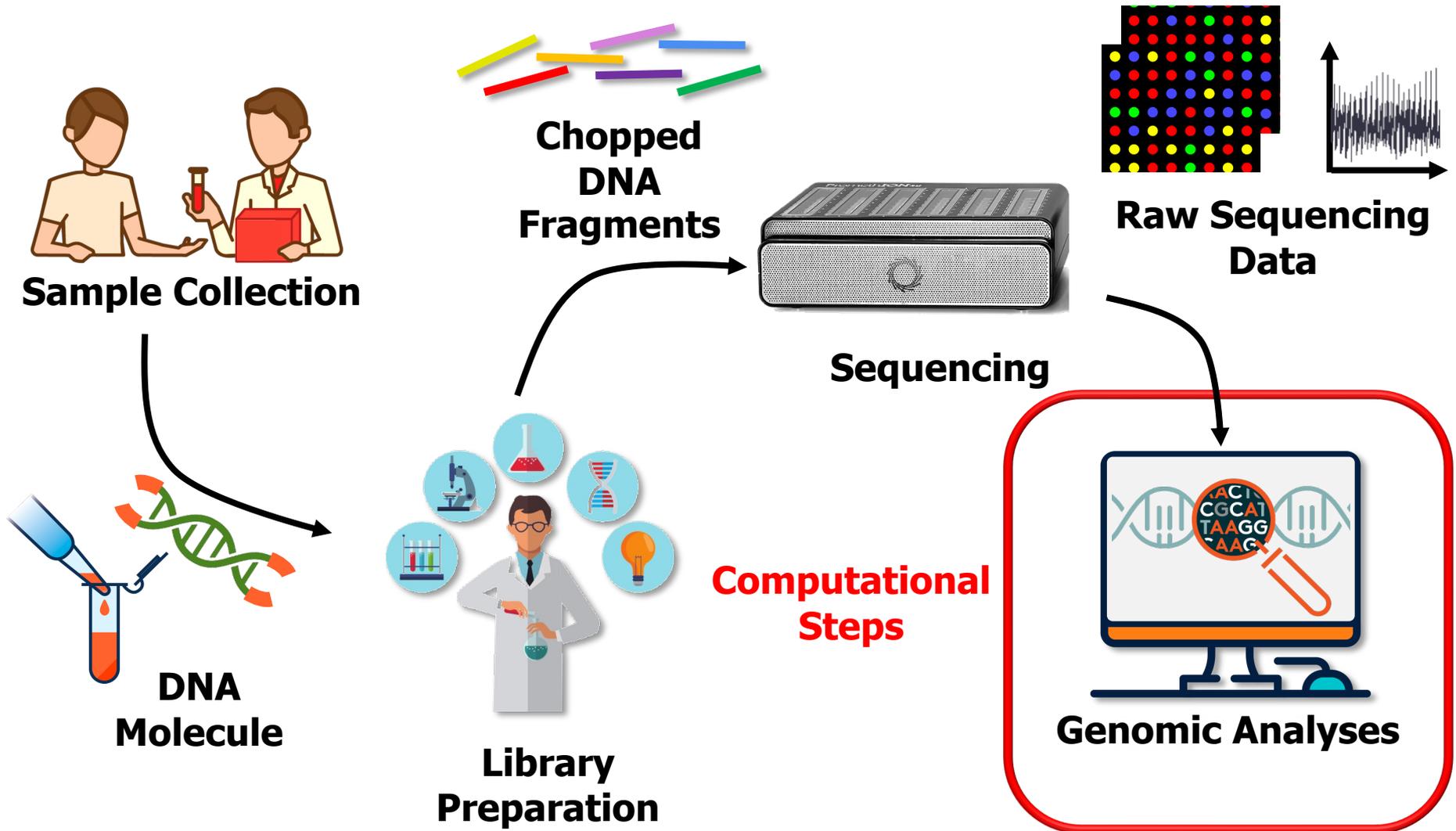
NO

machine gives the **complete sequence** of genome as output



```
>CCTCCTCAGTGCCACCCAGCCCAGCTGGCAGCTCCCAAACAGGCTCTTATTAACACCCCTGTTCCCTGCCCTTGGAGTGAGGTGTCAAG  
GACCTAAACTAAAAAAAAAAAAAAAAAGAAAAAGAAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTCTT  
CATGTCAAGGACCTAATGTGCTAACAGCACTTTTTTGACCATTATTTTGGATCTGAAAGAAATCAAGAATAAATGAAGGACTTGATACATTG  
GAAGAGGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAAAAGAAAAAGAAAAAGAAAAAGAATTAAAATTTAAGTAATTCTTTGAAAAAA  
ACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAGAAAAAGAATTTAAAATTT  
AAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAAGGCCAAGAGTTGCAAAAAAAAAAAAAAAAAAGAAAA  
GAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTAGCCAGAATGG  
TTGTGGGATGGGAGCCTCTGTGGACCGACCAGGTAGCTCTCTTTCCACACTGTAGTCTCAAAGCTTCTTCATGTGGTTTTCTCTGAGTGAAA  
AAAAAAAAAAGAAAAAGAAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTTTCATGTCAAGGACC  
TAATGTAGCTATACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCCTCCAGAAAGCTTCTTCA.....
```

Genome Analysis in Real Life



Solving the Puzzle



Reference genome



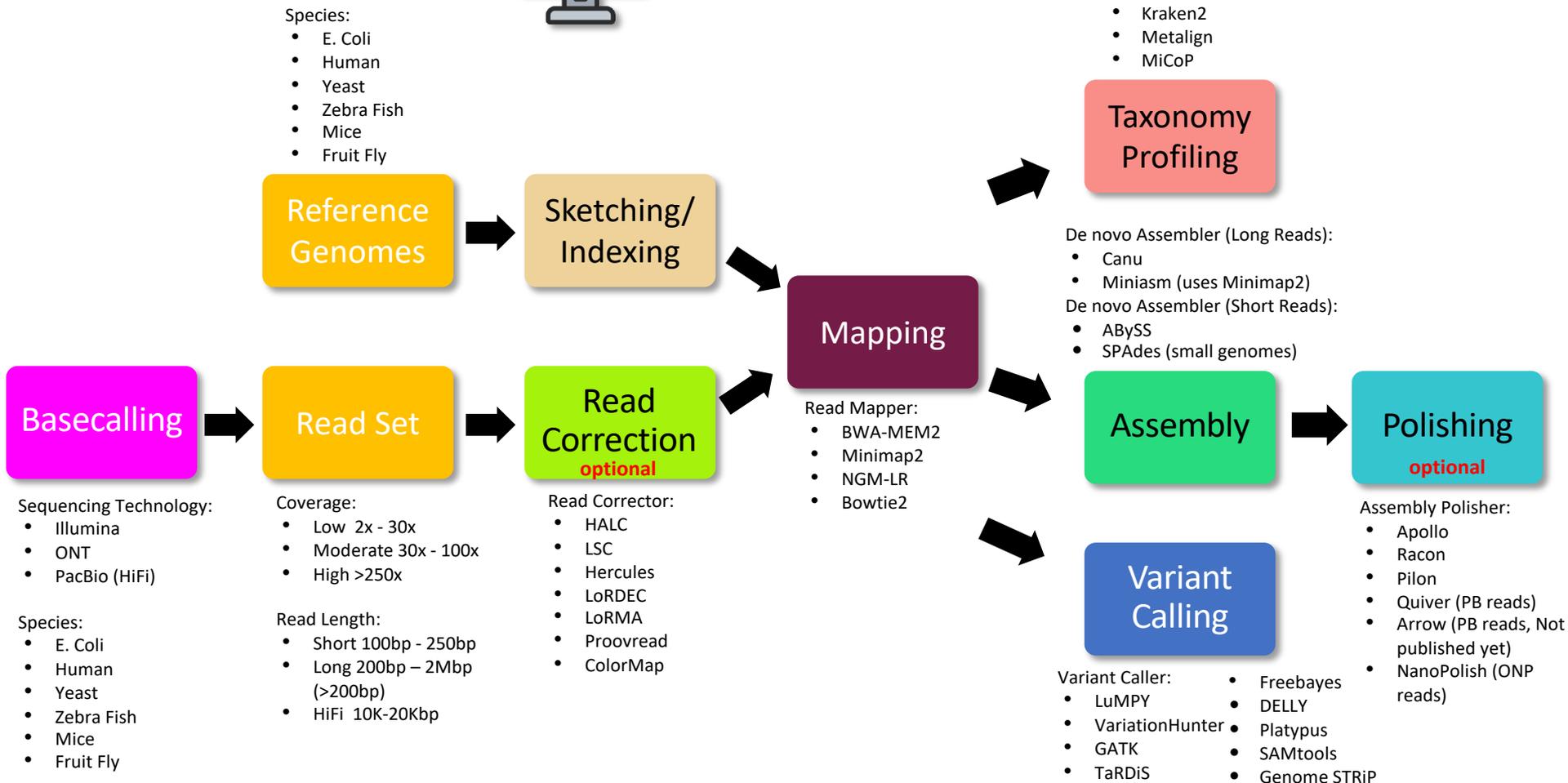
Reads

<https://www.pacb.com/smrt-science/smrt-sequencing/hifi-reads-for-highly-accurate-long-read-sequencing/>

Several Genome Analysis Pipelines



Genome Analysis



How Large is a Genome?



Hôtel El Mouradi Africa, Tunis



~3.2 billion genomic bases

Sequencing Technologies



... and more! All produce data with different properties.

Sequencing in Action



Chemistry type:

R10.4.1



Pack size:

Select ...



1 Flow cell

\$900.00

\$900.00 each

12 Flow cells

\$9,480.00

\$790.00 each

MinION

Portable DNA/RNA sequencing for anyone



Technology Dictates Algorithm Complexity

Short Reads (Illumina)

1 Sequencing

Library preparation: 6.5 hours
Sequencing: 68.2 Gb/hour

2 Basecalling

104.4 Gb/hour

3 Quality Control

1339.2 Gb/hour

4 Read Mapping

0.2 Gb/hour

5 Variant Calling

1.2 Gb/hour

Ultra-long Reads (ONT)

1 Sequencing

Library preparation: 24 hours
Sequencing: 4.1 Gb/hour

2 Basecalling

0.833 Gb/hour

3 Quality Control

3420 Gb/hour

4 Read Mapping

1.7 Gb/hour

5 Variant Calling

0.044 Gb/hour

Accurate Long Reads (PacBio)

1 Sequencing

Library preparation: 24 hours
Sequencing: 5.3 Gb/hour

2 Basecalling

8.3 Gb/hour

3 Quality Control

1081 Gb/hour

4 Read Mapping

1.4 Gb/hour

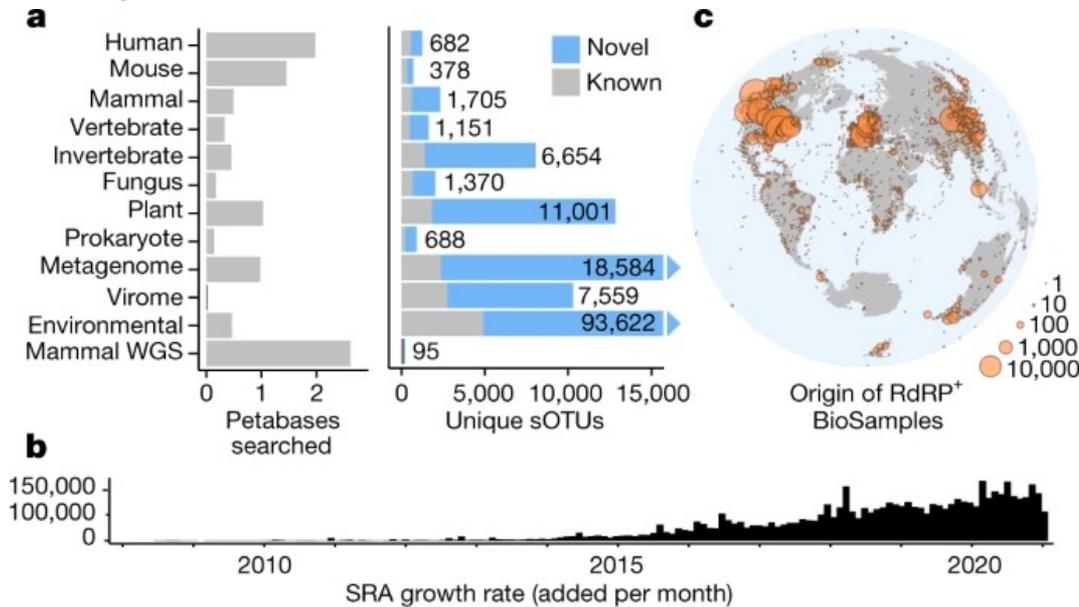
5 Variant Calling

1.1 Gb/hour

Alser+, [Going From Molecules to Genomic Variations to Scientific Discovery: Intelligent Algorithms and Architectures for Intelligent Genome Analysis](#), arXiv 2022

Petabase-scale Viral Discovery

- Building and Profiling 3,500 genomic assemblies needs **28,000 virtual AWS CPUs.**



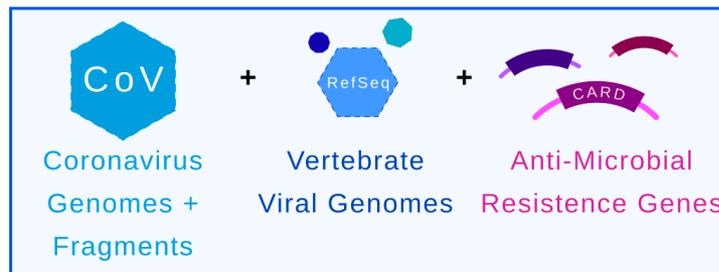
<https://serratus.io/>

Nucleotide



3.8m

ATGCATCAGGAATAGAC...
bowtie2



Edgar+, "[Petabase-scale sequence alignment catalyses viral discovery](#)", Nature 2022

Population-Scale Microbiome Profiling



CAMI Consortium

F. Meyer, A. Fritz, Z.L. Deng, D. Koslicki, A. Gurevich, G. Robertson, **Mohammed Alser**, and others

[“Critical Assessment of Metagenome Interpretation - the second round of challenges”](#), **Nature Methods**, 2022

[\[Source Code\]](#)

nature | **methods**

ANALYSIS

<https://doi.org/10.1038/s41592-022-01431-4>

Analysis | [Open Access](#) | [Published: 08 April 2022](#)

Critical Assessment of Metagenome Interpretation: the second round of challenges

[Fernando Meyer](#), [Adrian Fritz](#), ... [Alice Carolyn McHardy](#) 

+ Show authors

[Nature Methods](#) **19**, 429–440 (2022) | [Cite this article](#)

7302 Accesses | **79** Altmetric | [Metrics](#)

City-Scale Microbiome Profiling

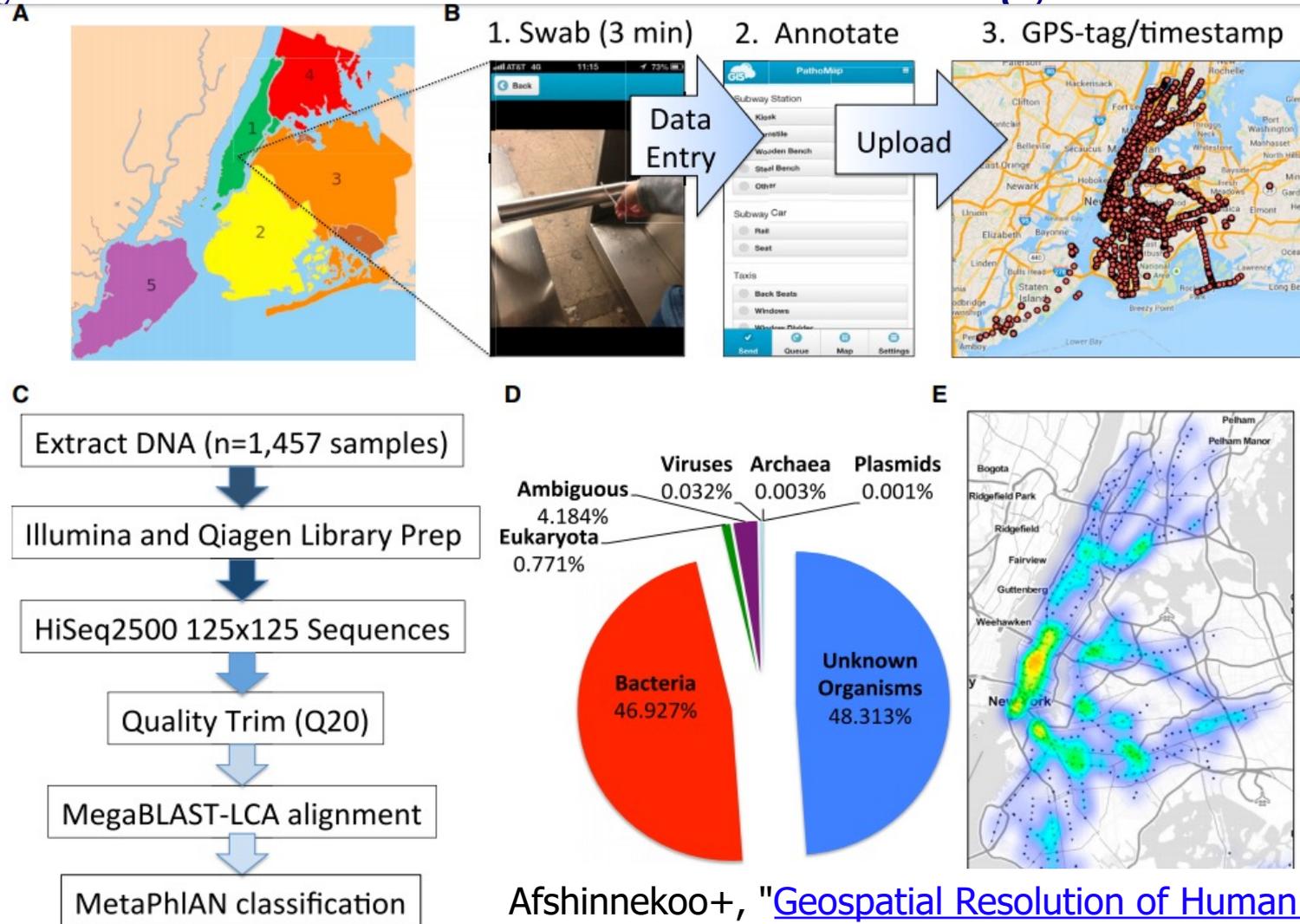


Figure 1. The Metagenome of New York City

(A) The five boroughs of NYC include (1) Manhattan (green)

(B) The collection from the 466 subway stations of NYC across the 24 subway lines involved three main steps: (1) collection with Copan Elution swabs, (2) data entry into the database, and (3) uploading of the data. An image is shown of the current collection database, taken from <http://pathomap.giscloud.com>.

(C) Workflow for sample DNA extraction, library preparation, sequencing, quality trimming of the FASTQ files, and alignment with MegaBLAST and MetaPhlan to discern taxa present

Afshinnekoo+, "[Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics](#)", Cell Systems, 2015

Population-Scale Microbiome Profiling

Cell

Log in Register Su

ARTICLE | ONLINE NOW

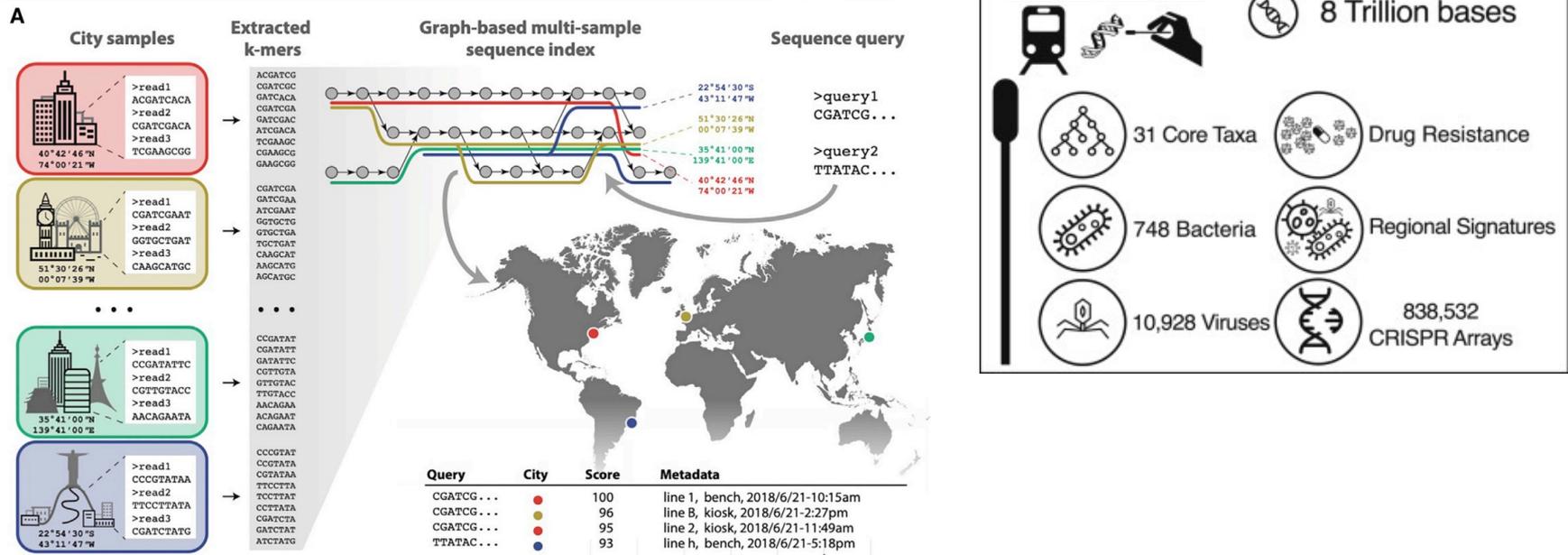
PDF [9 MB] Figures Save

A global metagenomic map of urban microbiomes and antimicrobial resistance

David Danko ⁶⁸ • Daniela Bezdán ⁶⁸ • Evan E. Afshin • ... Sibó Zhu • Christopher E. Mason ⁶⁹  

The International MetaSUB Consortium • [Show all authors](#) • [Show footnotes](#)

Open Access • Published: May 26, 2021 • DOI: <https://doi.org/10.1016/j.cell.2021.05.002>



Danko+, "A global metagenomic map of urban microbiomes and antimicrobial resistance", Cell, 2021

Personalized Medicine in UK

“From 2019, **all seriously ill children** in UK will be offered **whole genome sequencing** as part of their care”



Challenging Environment in Outer Space



We need intelligent algorithms
and intelligent architectures
that handle data well

Fostering Omics Research

National
Strategy

Genome Map

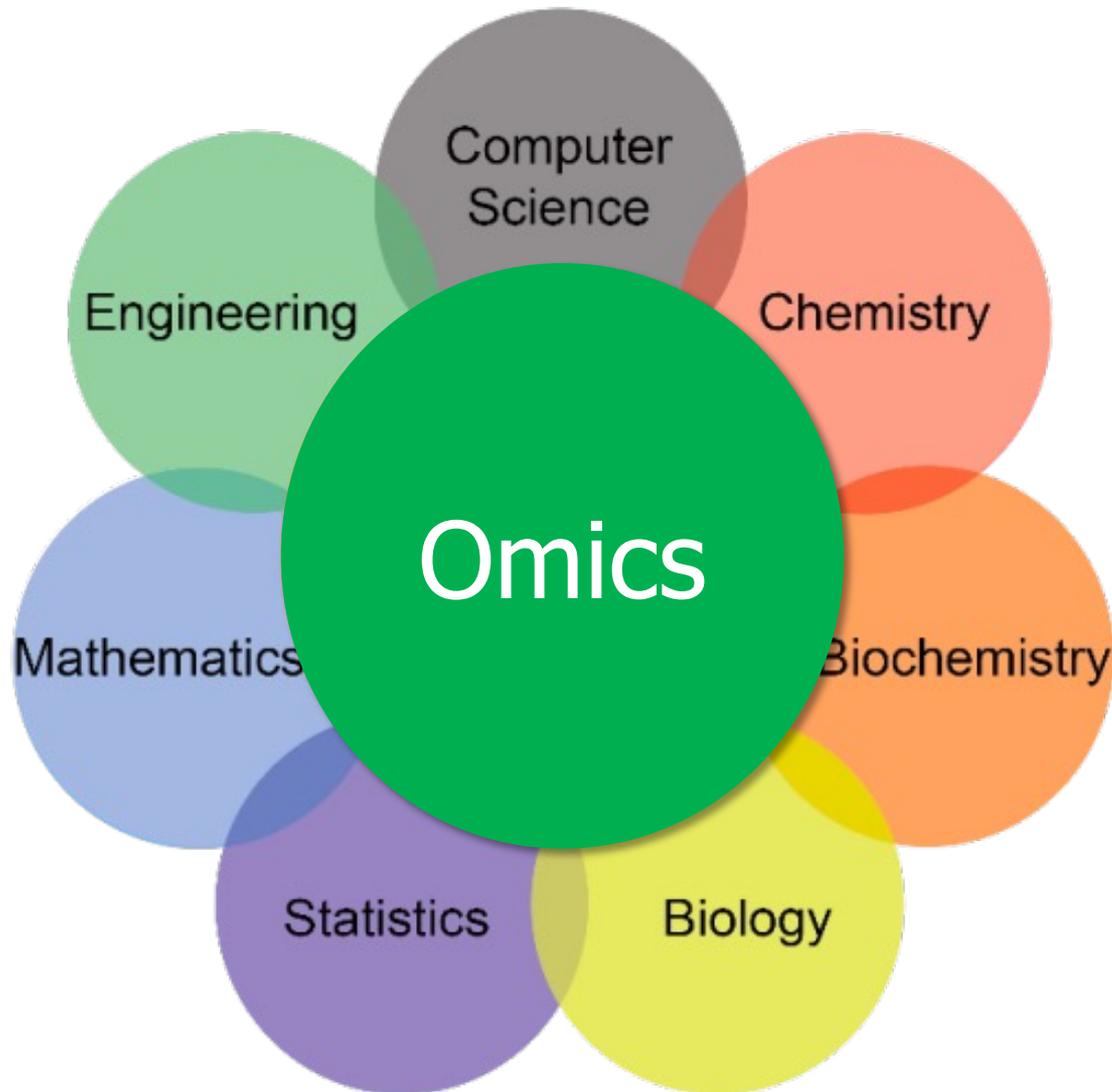
National
Genomics
Centers

Genome
Banks

Genomics
Startups &
Companies

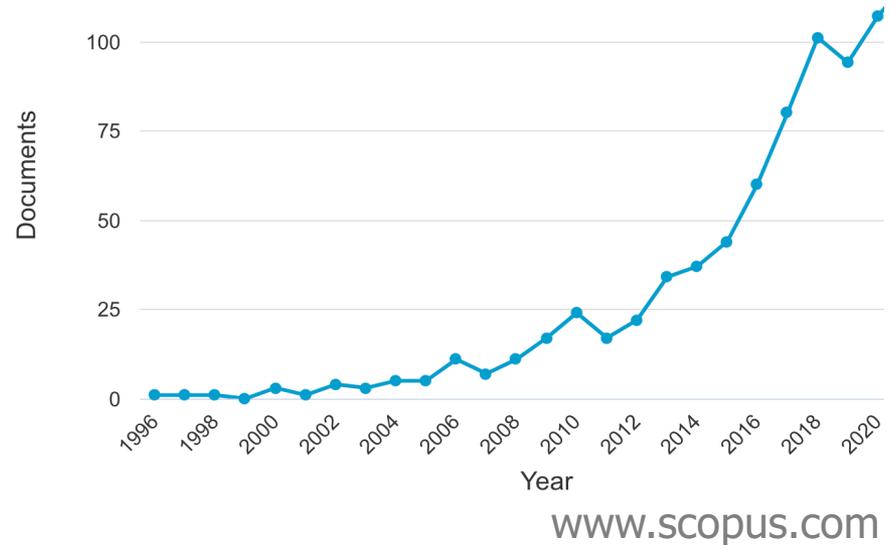
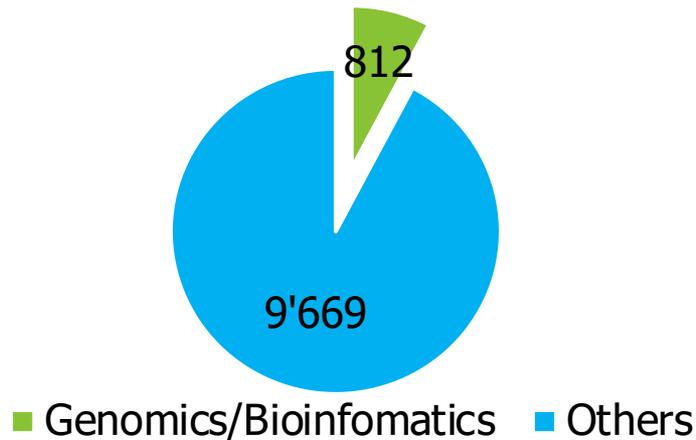
Genomics
Education &
Experts

Interdisciplinary Education



Genomics in Palestinian Universities

- Only 812 (**7%**) out of 10,481 Scopus-indexed research papers are **genomics/bioinformatics related**.

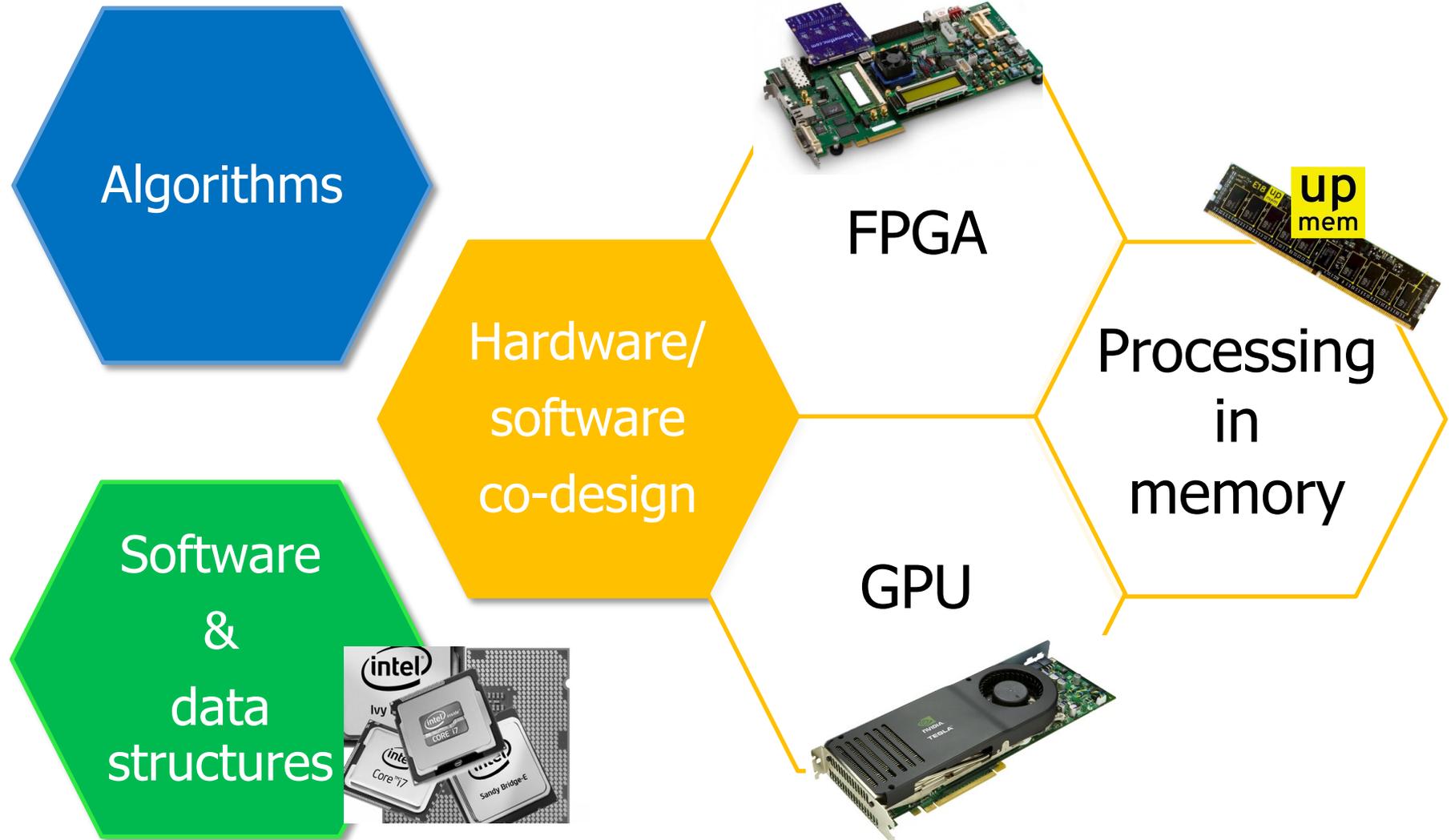


Omics is NOT Biology

- أحياء مسار بيوتكنولوجي في جامعة القدس
- أحياء مسار بيوتكنولوجي في جامعة بيرزيت
- التكنولوجيا الحيوية في الجامعة الإسلامية

- الأحياء التطبيقية في جامعة البوليتكنيك
- الأحياء والتكنولوجيا الحيوية في الجامعة الأمريكية
- الأحياء والبيوتكنولوجيا في جامعة النجاح

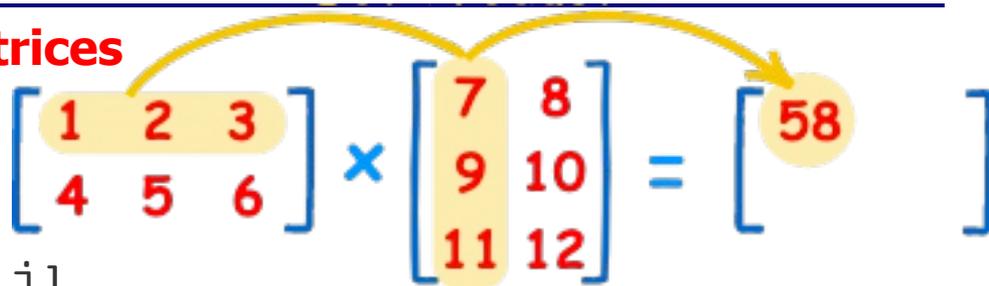
Toolkits



Software & Hardware Optimizations

Multiplying Two 4096-by-4096 Matrices

```
for i in xrange(4096):  
    for j in xrange(4096):  
        for k in xrange(4096):  
            C[i][j] += A[i][k] * B[k][j]
```



Implementation	Running time (s)	Absolute speedup
Python	25,552.48	1x
Java	2,372.68	11x
C	542.67	47x
Parallel loops	69.80	366x
Parallel divide and conquer	3.80	6,727x
plus vectorization	1.10	23,224x
plus AVX intrinsics	0.41	62,806x

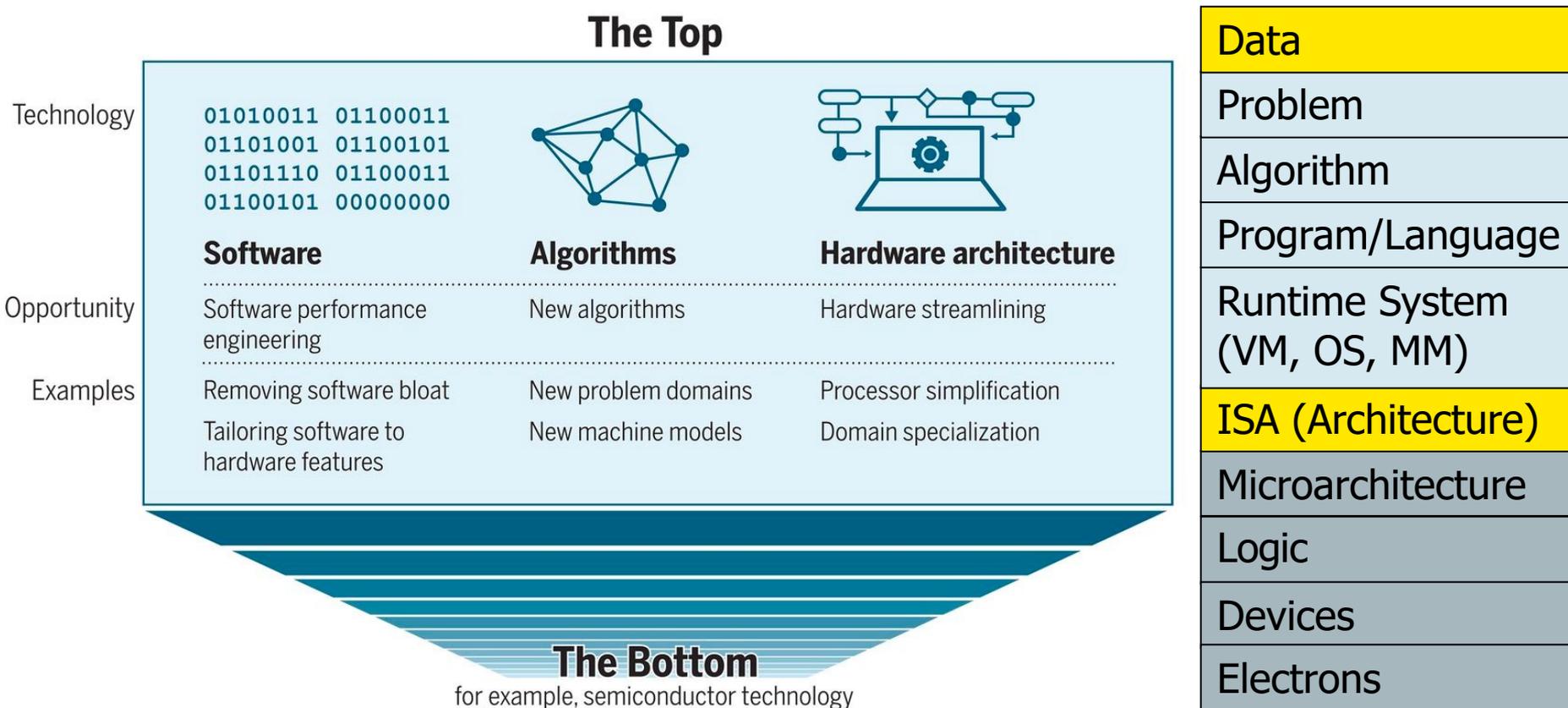
Leiserson+, "[There's plenty of room at the Top: What will drive computer performance after Moore's law?](#)", Science, 2020

FASTQ Parsing

Program	Language	t _{gzip} (s)	t _{plain} (s)	Comments
fqcnt_rs2_needletail.rs	Rust	9.3	0.8	needletail ; fasta/4-line fastq
fqcnt_c1_kseq.c	C	9.7	1.4	multi-line fasta/fastq
fqcnt_cr1_klib.cr	Crystal	9.7	1.5	kseq.h port
fqcnt_nim1_klib.nim	Nim	10.5	2.3	kseq.h port
fqcnt_jl1_klib.jl	Julia	11.2	2.9	kseq.h port
fqcnt_js1_k8.js	Javascript	17.5	9.4	kseq.h port
fqcnt_go1.go	Go	19.1	2.8	4-line only
fqcnt_lua1_klib.lua	LuaJIT	28.6	27.2	partial kseq.h port
fqcnt_py2_rfq.py	PyPy	28.9	14.6	partial kseq.h port
fqcnt_py2_rfq.py	Python	42.7	19.1	partial kseq.h port

Spanning The Full Computing Stack

Leiserson+, "[There's plenty of room at the Top: What will drive computer performance after Moore's law?](#)", Science, 2020

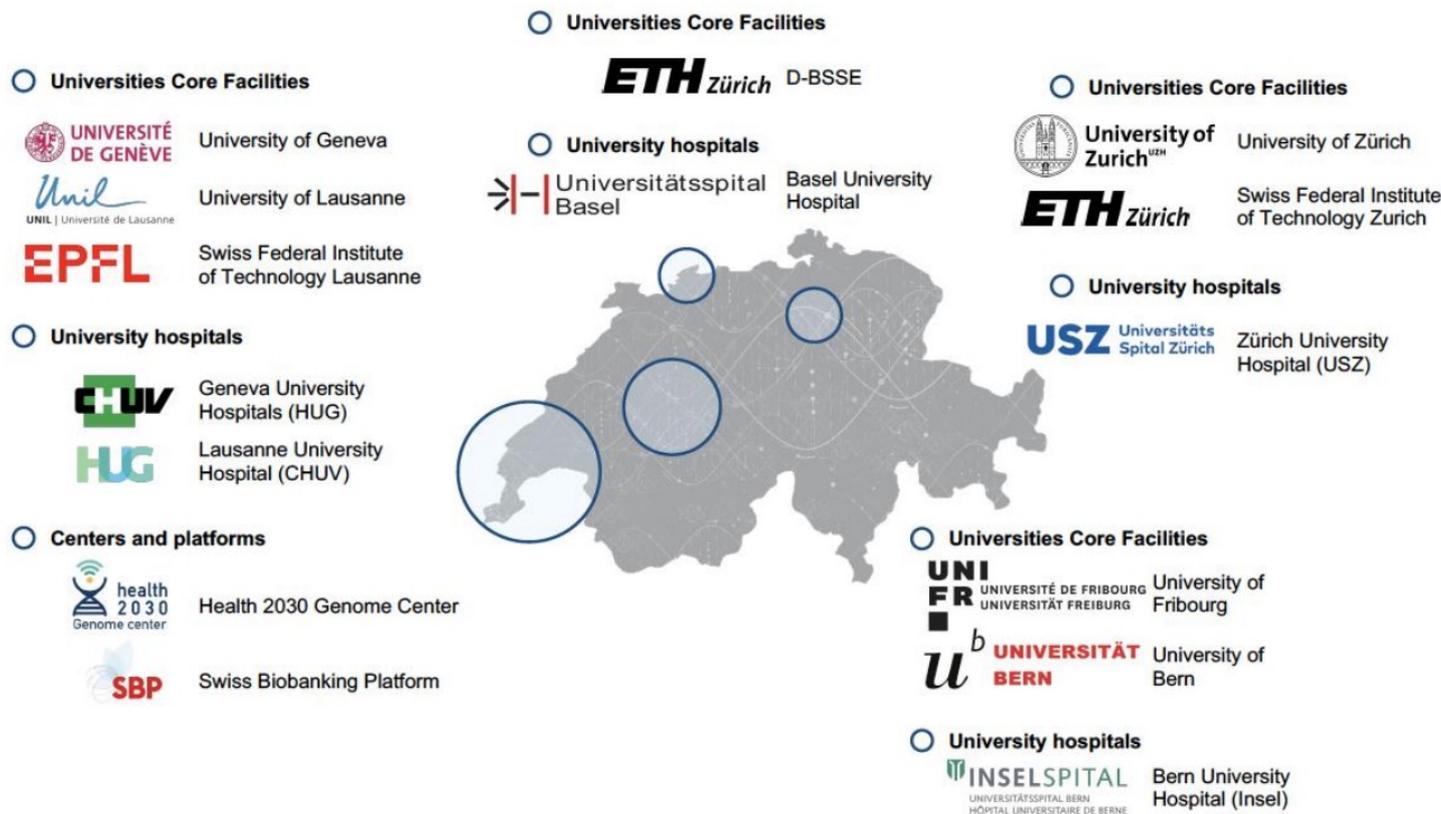


Richard Feynman, "[There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics](#)", a lecture given at Caltech, 1959.

Moving Forward ...

Switzerland (2019)...It is Never Too Late

The Swiss genomics landscape



Saudi Genome (2013)



الجينوم السعودي
SAUDI GENOME



مدينة الملك عبدالعزيز
للعلوم والتقنية KACST

أبرز إنجازات برنامج الجينوم السعودي

توثيق
7,500

متغير مسبب للأمراض الوراثية

والجينية بالمملكة , منها **3000** متغير
جيني مسبب لأكثر من **1230** مرضاً
وراثياً نادراً في المجتمع السعودي

فحص أكثر من
50,000
عينة



مشاركة **600**
باحث وباحثة



Qatar Genome Program (2013)

26000

جينوم كامل



عضو في مؤسسة قطر
Member of Qatar Foundation



The Emirati Genome Program (2021)



الإمارات اليوم

50
عام القمصين
1971

«برنامج الجينوم» يستهدف
جمع مليون عيّنة «DNA» لرسم
خريطة جينية للمواطنين



الجينوم الإماراتي
لمستقبل أجيالنا

EMIRATI GENOME
FUTURE OF OUR GENERATIONS

برنامج الجينوم الإماراتي

انطلاقاً من حرص قيادتنا الرشيدة على الارتقاء بالرعاية الصحية المقدمة لمواطني دولة الإمارات وإرساء مكانة الدولة كمركز للبحث والابتكار في مجال الجينوم، تأتي فكرة برنامج الجينوم الإماراتي لتعزيز هذه الأهداف الاستراتيجية الفريدة من نوعها في دولة الإمارات العربية المتحدة.



البرنامج يساهم في التعرف إلى الطفرات الجينية للمواطنين. من المصدر

التاريخ: 28 يونيو 2021

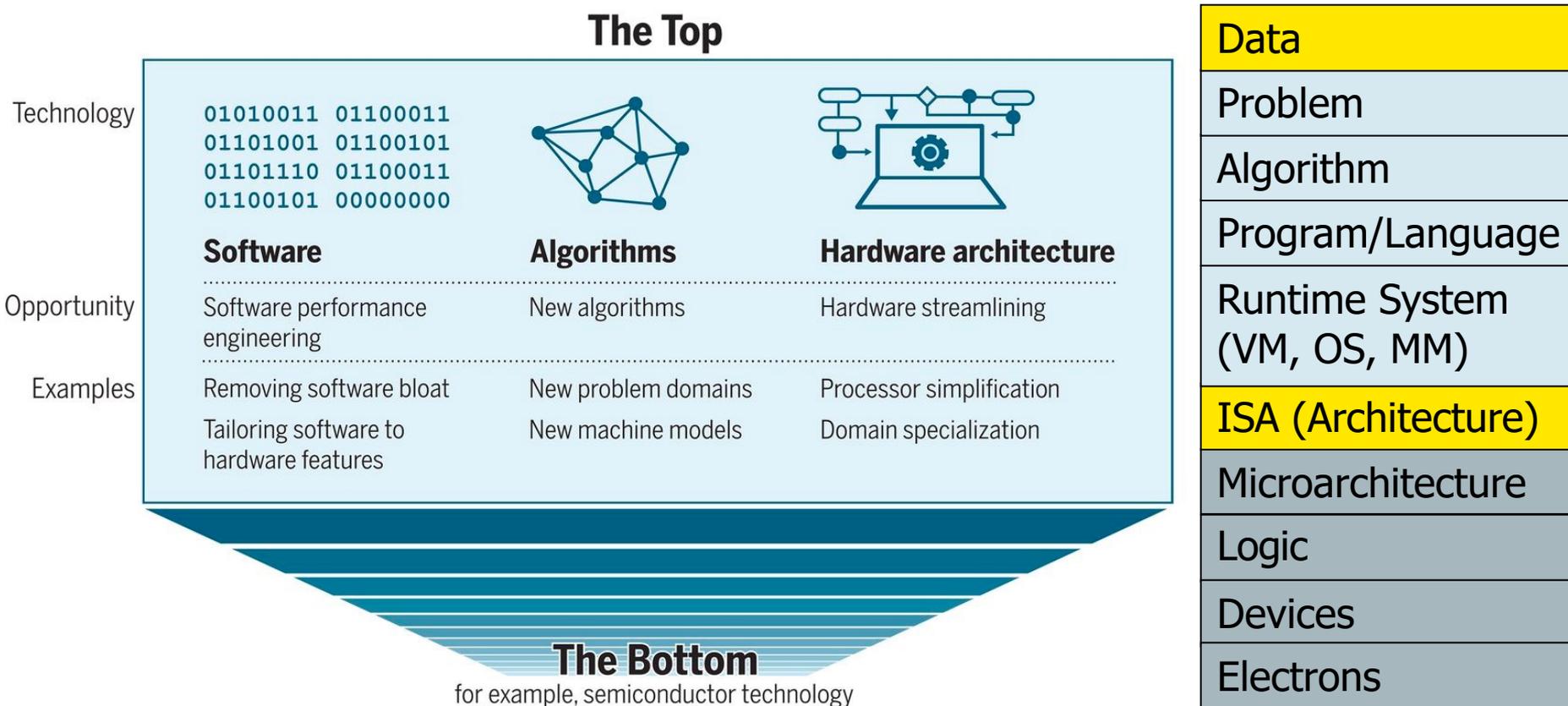
المصدر: عمرو بيومي ■ أبوظبي

Tunisia Can Do It Too ...



Revisiting The Full Computing Stack

Leiserson+, "[There's plenty of room at the Top: What will drive computer performance after Moore's law?](#)", Science, 2020



Richard Feynman, "[There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics](#)", a lecture given at Caltech, 1959.

Plenty of Room at the Top

Data Representation

Sparsified Genomics

<https://arxiv.org/abs/2211.08157>

arXiv > cs > arXiv:2211.08157

Search...

Help | Adv

Computer Science > Data Structures and Algorithms

[Submitted on 15 Nov 2022]

Taming Large-Scale Genomic Analyses via Sparsified Genomics

Mohammed Alser, Julien Eudine, Onur Mutlu

Searching for similar genomic sequences is an essential and fundamental step in biomedical research and an overwhelming majority of genomic analyses. State-of-the-art computational methods performing such comparisons fail to cope with the exponential growth of genomic sequencing data. We introduce the concept of sparsified genomics where we systematically exclude a large number of bases from genomic sequences and enable much faster and more memory-efficient processing of the sparsified, shorter genomic sequences, while providing similar or even higher accuracy compared to processing non-sparsified sequences. Sparsified genomics provides significant benefits to many genomic analyses and has broad applicability. We show that sparsifying genomic sequences greatly accelerates the state-of-the-art read mapper (minimap2) by 1.54–8.8x using real Illumina, HiFi, and ONT reads, while providing a higher number of mapped reads and more detected small and structural variations. Sparsifying genomic sequences makes containment search through very large genomes and very large databases 72.7–75.88x faster and 723.3x more storage-efficient than searching through non-sparsified genomic sequences (with CMash and KMC3). Sparsifying genomic sequences enables robust microbiome discovery by providing 54.15–61.88x faster and 720x more storage-efficient taxonomic profiling of metagenomic samples over the state-of-the-art tool (Metalign). We design and open-source a framework called Genome-on-Diet as an example tool for sparsified genomics, which can be freely downloaded from [this https URL](https://github.com/malser/genome-on-diet).

Minimizing Workload

ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA

Exact Match

ACCCTAACCCCTAACCCCTAACCCCTAACCCCTAA

Minimizing Workload

ACCCTAACCCCTAACCCCTAACCCCTAAC

A_C_T_A_C_T_A_C_T_A_C_T_A_C_T_A

Still Exact Match

ACCCTAACCCCTAACCCCTAACCCCTAAC

A_C_T_A_C_T_A_C_T_A_C_T_A_C_T_A

Minimizing Workload Is Challenging

ACCCTAACCCCTAACCCCTAACCCCTAA

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 _

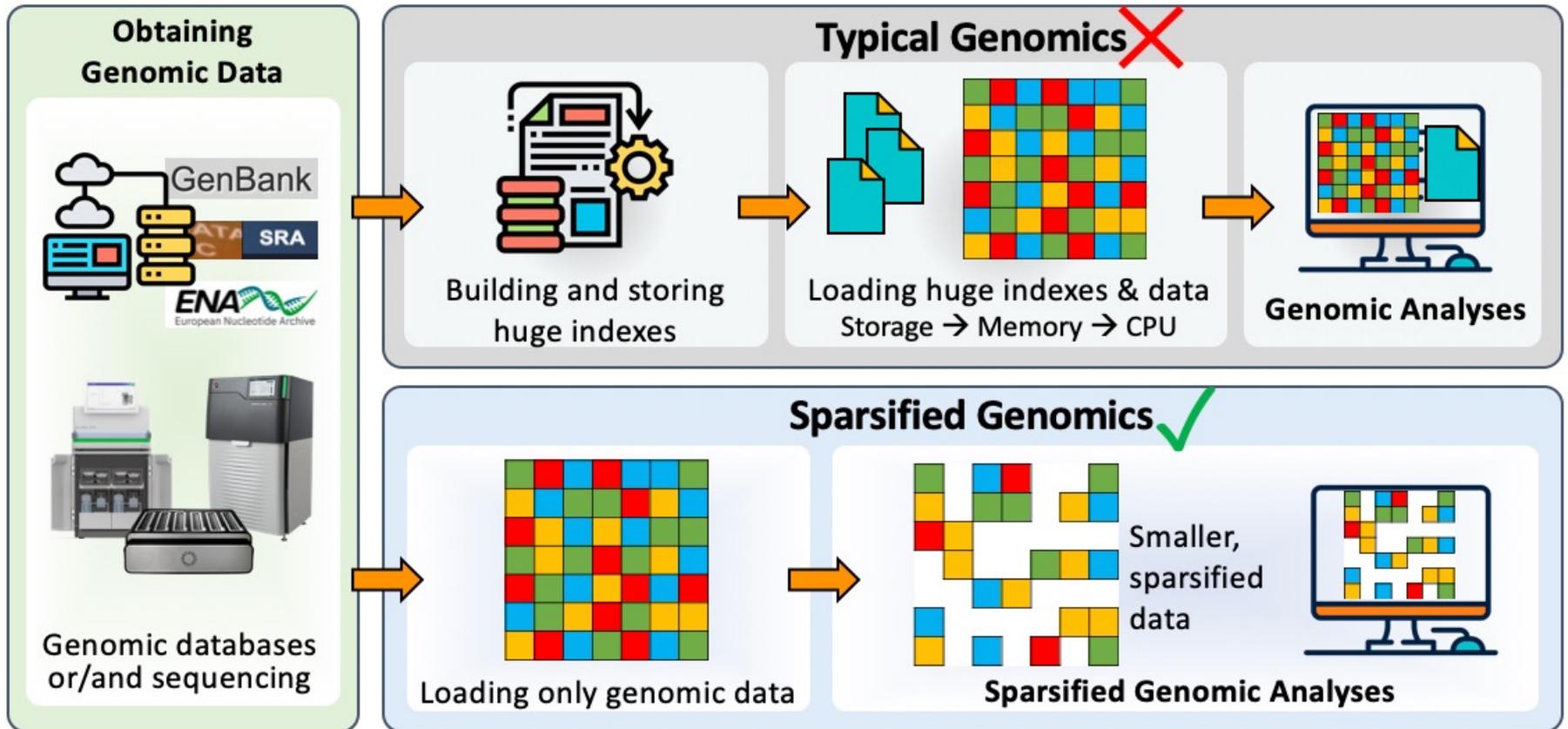
No Match 😞

_ CCCTAACCCCTAACCCCTAACCCCTAA

_ 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 _

Great Benefits by Manipulating Data

Taming Large-Scale Genomic Analyses via Sparsified Genomics



Great Benefits by Manipulating Data

- **Genome-on-Diet** is 1.54-8.8x faster and 2x more peak-memory-efficient compared to minimap2 for performing **read mapping**.
- **Genome-on-Diet** is 72.7-75.9x faster and 723.3x more storage-efficient than KMC3 combined with CMash, for performing **containment search**.
- **Genome-on-Diet** is 54.2-61.9x faster and 720x more storage-efficient than Metalign, for performing **taxonomic profiling of metagenomic samples**.

Plenty of Room at the Top

Efficient Algorithms

Efficient Algorithm for Pre-alignment Filtering

Mohammed Alser, Taha Shahroodi, Juan-Gomez Luna, Can Alkan, and Onur Mutlu,
"SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs"

Bioinformatics, 2020.

[[Source Code](#)]

[[Online link at Bioinformatics Journal](#)]

Bioinformatics



SneakySnake: a fast and accurate universal genome pre-alignment filter for CPUs, GPUs and FPGAs

Mohammed Alser ✉, Taha Shahroodi, Juan Gómez-Luna, Can Alkan ✉, Onur Mutlu ✉

Bioinformatics, btaa1015, <https://doi.org/10.1093/bioinformatics/btaa1015>

Published: 26 December 2020 **Article history** ▼

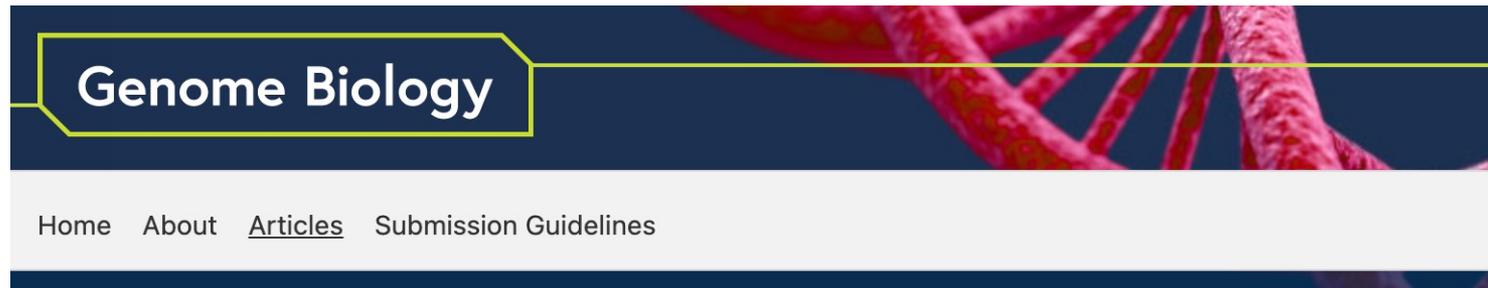
Metalign

Nathan LaPierre, **Mohammed Alser**, Eleazar Eskin, David Koslicki, Serghei Mangul
“[Metalign: efficient alignment-based metagenomic profiling via containment min hash](#)”

Genome Biology, September 2020.

[[Talk Video](#) (7 minutes) at ISMB 2020]

[[Source code](#)]



Software | [Open Access](#) | [Published: 10 September 2020](#)

Metalign: efficient alignment-based metagenomic profiling via containment min hash

[Nathan LaPierre](#) ✉, [Mohammed Alser](#), [Eleazar Eskin](#), [David Koslicki](#) ✉ & [Serghei Mangul](#) ✉

Genome Biology **21**, Article number: 242 (2020) | [Cite this article](#)

CAMI Consortium

F. Meyer, A. Fritz, Z.L. Deng, D. Koslicki, A. Gurevich, G. Robertson, **Mohammed Alser**, and others

[“Critical Assessment of Metagenome Interpretation - the second round of challenges”](#), **Nature Methods**, 2022

[\[Source Code\]](#)

nature | **methods**

ANALYSIS

<https://doi.org/10.1038/s41592-022-01431-4>

Analysis | [Open Access](#) | [Published: 08 April 2022](#)

Critical Assessment of Metagenome Interpretation: the second round of challenges

[Fernando Meyer](#), [Adrian Fritz](#), ... [Alice Carolyn McHardy](#) 

[+ Show authors](#)

[Nature Methods](#) **19**, 429–440 (2022) | [Cite this article](#)

7302 Accesses | **79** Altmetric | [Metrics](#)

MiCoP

Nathan LaPierre, Serghei Mangul, **Mohammed Alser**, Igor Mandric, Nicholas C. Wu, David Koslicki & Eleazar Eskin

[“MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples”](#)

BMC Genomics, June 2019.

[\[Source code\]](#)

 **BMC** Part of Springer Nature

BMC Genomics

Research | [Open Access](#) | [Published: 06 June 2019](#)

MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples

[Nathan LaPierre](#), [Serghei Mangul](#) , [Mohammed Alser](#), [Igor Mandric](#), [Nicholas C. Wu](#), [David Koslicki](#) & [Eleazar Eskin](#)

[BMC Genomics](#) **20**, Article number: 423 (2019) | [Cite this article](#)

AirLift

Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali,
Mohammed Alser, Nastaran Hajinazar, Can Alkan, Onur Mutlu
“[AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes](#)”
arXiv 2022
GitHub: <https://github.com/CMU-SAFARI/AirLift>

arXiv > q-bio > arXiv:1912.08735

Search...

Help | Advanced

Quantitative Biology > Genomics

[Submitted on 18 Dec 2019 (v1), last revised 12 Aug 2022 (this version, v3)]

AirLift: A Fast and Comprehensive Technique for Remapping Alignments between Reference Genomes

Jeremie S. Kim, Can Firtina, Meryem Banu Cavlak, Damla Senol Cali, Mohammed Alser,
Nastaran Hajinazar, Can Alkan, Onur Mutlu

Read Mapping in 111 pages!

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

Mohammed Alser, Jeremy Rotman, Dhriti Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyung 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†}, Dhriti Deshpande⁵, Kodi Taraszka⁴, Huwenbo Shi^{6,7}, Pelin Icer Baykal⁸, Harry Taegyung Yang^{4,9}, Victor Xue⁴, Sergey Knyazev⁸, Benjamin D. Singer^{10,11,12}, Brunilda Balliu¹³, David Koslicki^{14,15,16}, Pavel Skums⁸, Alex Zelikovsky^{8,17}, Can Alkan^{2,18}, Onur Mutlu^{1,2,3†} and Serghei Mangul^{5*†} 

Intelligent Genome Analysis

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

["From Molecules to Genomic Variations: Intelligent Algorithms and Architectures for Intelligent Genome Analysis"](#)

Computational and Structural Biotechnology Journal, 2022

[[Source code](#)]



ELSEVIER

01010100100101010010
0010101001010101011
1010101001010101011
010101001010101010
110101001010101010
1010101001010101011
0010101001010101011
010101001010101010
11010101001010101010

COMPUTATIONAL
AND STRUCTURAL
BIOTECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj



Review

From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures



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

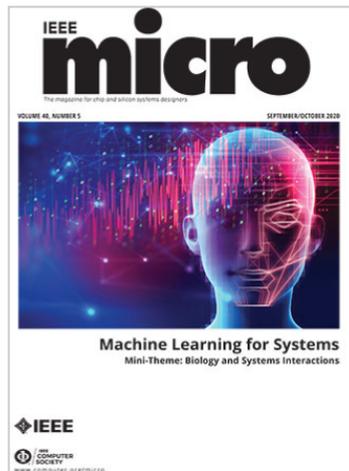
ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland

Accelerating Read Mapping

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.



[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](https://doi.org/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

◀	▶
Previous	Next
☰	Table of Contents
📄	Past Issues

Improving Omics Usability & Reproducibility

Mohammed Alser, Sharon Waymost, Ram Ayyala, Brendan Lawlor, Richard J. Abdill, Neha Rajkumar, Nathan LaPierre, Jaqueline Brito, Andre M. Ribeiro-dos-Santos, Can Firtina, Nour Almadhoun, Varuni Sarwal, Eleazar Eskin, Qiyang Hu, Derek Strong, Byoung-Do (BD)Kim, Malak S. Abedalthagafi, Onur Mutlu, Serghei Mangul

["Packaging, containerization, and virtualization of computational omics methods: Advances, challenges, and opportunities"](#)

arXiv 2022

**Packaging, containerization, and virtualization of computational omics methods:
Advances, challenges, and opportunities**

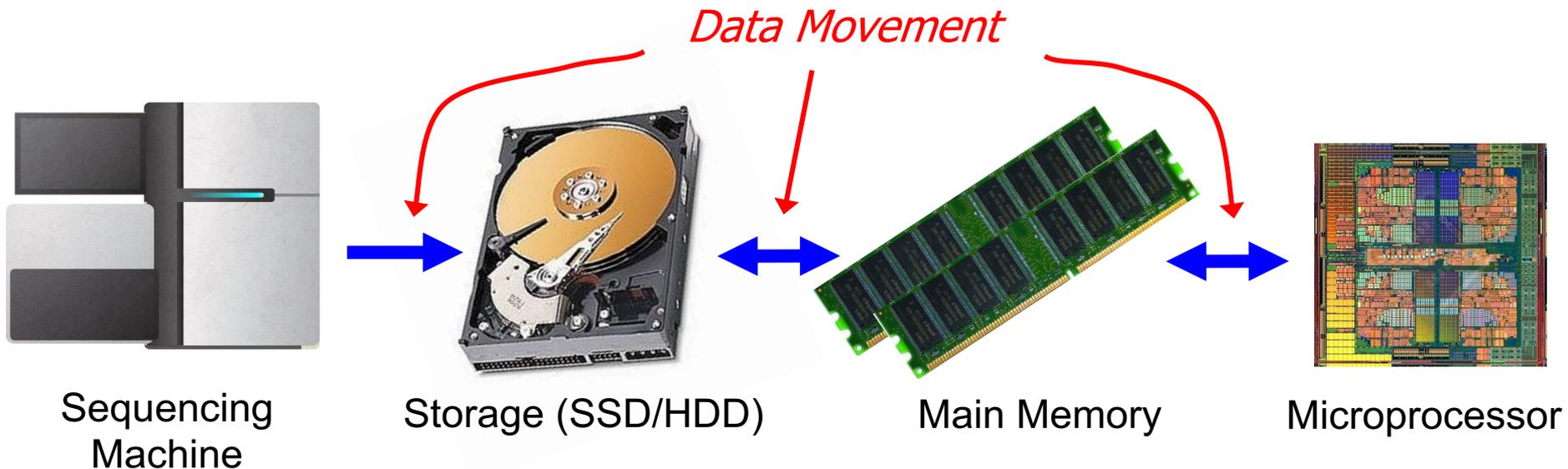
Mohammed Alser¹, Sharon Waymost², Ram Ayyala^{3,4}, Brendan Lawlor⁵, Richard J. Abdill⁶, Neha Rajkumar⁷, Nathan LaPierre², Jaqueline Brito⁴, André M. Ribeiro-dos-Santos⁸, Can Firtina¹, Nour Almadhoun¹, Varuni Sarwal², Eleazar Eskin^{2,9,10}, Qiyang Hu¹¹, Derek Strong¹², Byoung-Do (BD) Kim¹², Malak S. Abedalthagafi^{13,14,15*}, Onur Mutlu^{1,*}, Serghei Mangul^{4,*}

Plenty of Room at the Bottom

Efficient HW/SW Co-design

Data Movement Dominates Performance

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



* 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

Overview

Near-memory/In-memory Pre-alignment Filtering

GRIM-Filter [BMC Genomics'18]

SneakySnake [IEEE Micro'21]

GenASM [MICRO 2020]

In-storage Sequence Alignment

GenStore [ASPLOS 2022]

Near-memory Sequence Alignment

GenASM [MICRO 2020]

SeGraM [ISCA 2022]

Specialized Pre-alignment Filtering Accelerators (GPU, FPGA)

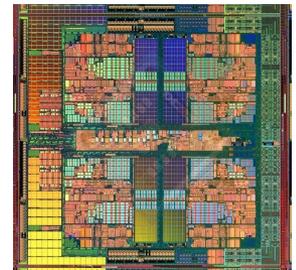
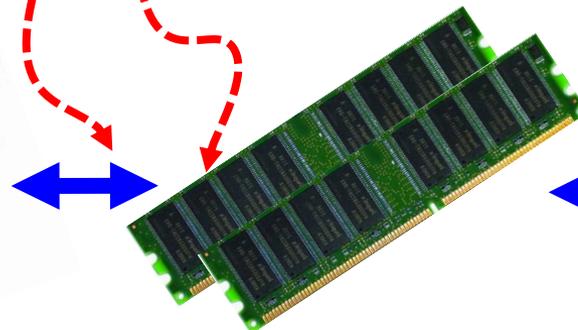
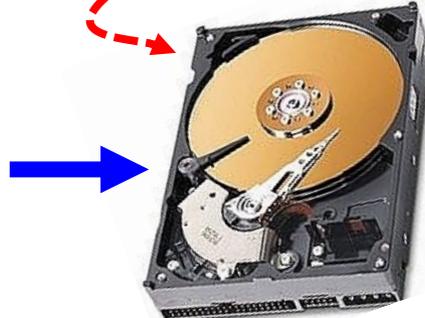
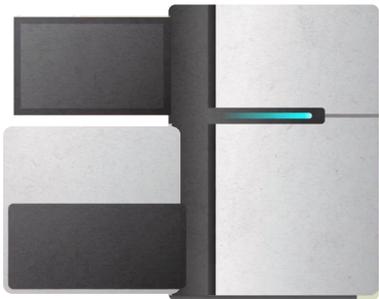
GateKeeper [Bioinformatics'17]

MAGNET [AACBB'18]

Shouji [Bioinformatics'19]

GateKeeper-GPU [arXiv'21]

SneakySnake [Bioinformatics'20]



Sequencing Machine

Storage (SSD/HDD)

Main Memory

Microprocessor

Overview

Near-memory Sequence Alignment

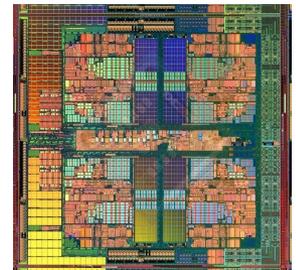
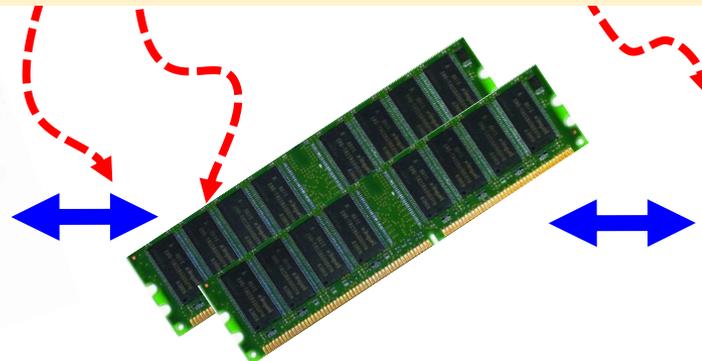
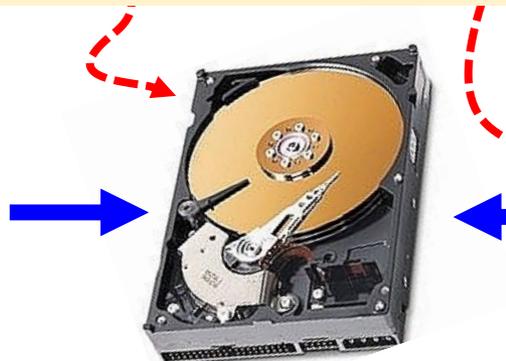
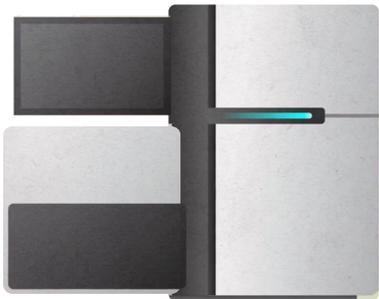
GenASM [MICRO 2020]

SeGraM [ISCA 2022]

Near-memory/In-memory
Pre-alignment Filtering

Specialized Pre-alignment Filtering
Accelerators (GPU, FPGA)

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



Sequencing Machine

Storage (SSD/HDD)

Main Memory

Microprocessor

RUBICON (2022)

Gagandeep Singh, **Mohammed Alser**, Alireza Khodamoradi, Kristof Denolf, Can Firtina, Meryem Banu Cavlak, Henk Corporaal, Onur Mutlu,
“[A Framework for Designing Efficient Deep Learning-Based Genomic Basecallers](#)”,
arXiv 2022

A Framework for Designing Efficient Deep Learning-Based Genomic Basecallers

Gagandeep Singh^a Mohammed Alser^{*a} Alireza Khodamoradi^{*b}
Kristof Denolf^b Can Firtina^a Meryem Banu Cavlak^a
Henk Corporaal^c Onur Mutlu^a
^aETH Zürich ^bAMD ^cEindhoven University of Technology

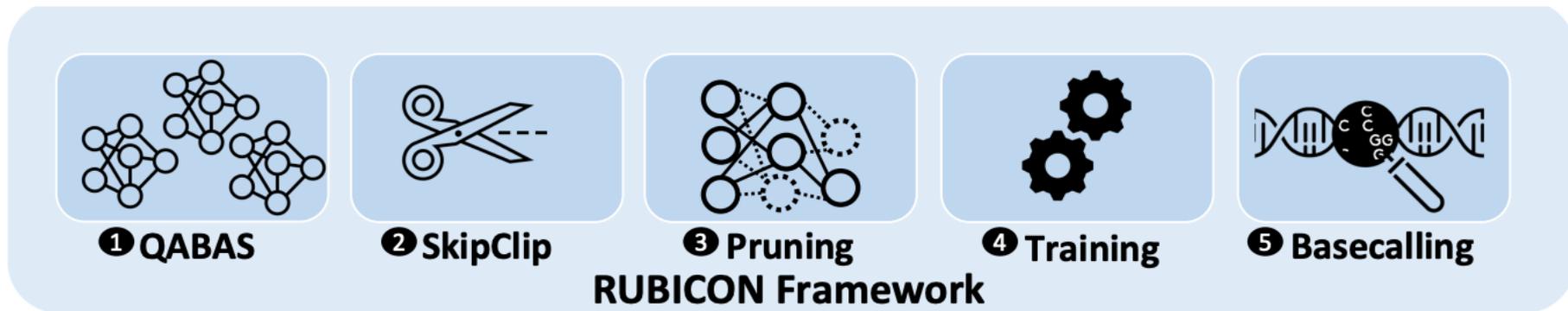


Figure 1: Overview of RUBICON framework.

GenStore (ASPLOS 2022)

Nika Mansouri Ghiasi, Jisung Park, Harun Mustafa, Jeremie Kim, Ataberk Olgun, Arvid Gollwitzer, Damla Senol Cali, Can Firtina, Haiyu Mao, Nour Almadhoun Alserr, Rachata Ausavarungnirun, Nandita Vijaykumar, **Mohammed Alser**, Onur Mutlu

["GenStore: A High-Performance and Energy-Efficient In-Storage Computing System for Genome Sequence Analysis"](#),

ASPLOS 2022

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

Nika Mansouri Ghiasi
ETH Zürich
Switzerland

Jisung Park
ETH Zürich
Switzerland

Harun Mustafa
ETH Zürich
Switzerland

Jeremie Kim
ETH Zürich
Switzerland

Ataberk Olgun
ETH Zürich
Switzerland

Arvid Gollwitzer
ETH Zürich
Switzerland

Damla Senol Cali
Bionano Genomics
USA

Can Firtina
ETH Zürich
Switzerland

Haiyu Mao
ETH Zürich
Switzerland

Nour Almadhoun
Alserr
ETH Zürich
Switzerland

Rachata
Ausavarungnirun
KMUTNB
Thailand

Nandita Vijaykumar
University of Toronto
Canada

Mohammed Alser
ETH Zürich
Switzerland

Onur Mutlu
ETH Zürich
Switzerland

GenPIP (MICRO 2022)

Haiyu Mao, **Mohammed Alser**, Mohammad Sadrosadati, Can Firtina, Akanksha Baranwal, Damla Senol Cali, Aditya Manglik, Nour Almadhoun Alserr, Onur Mutlu

[“GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping”](#)

Proceedings of the [55rd International Symposium on Microarchitecture \(MICRO\)](#), 2022.

GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping

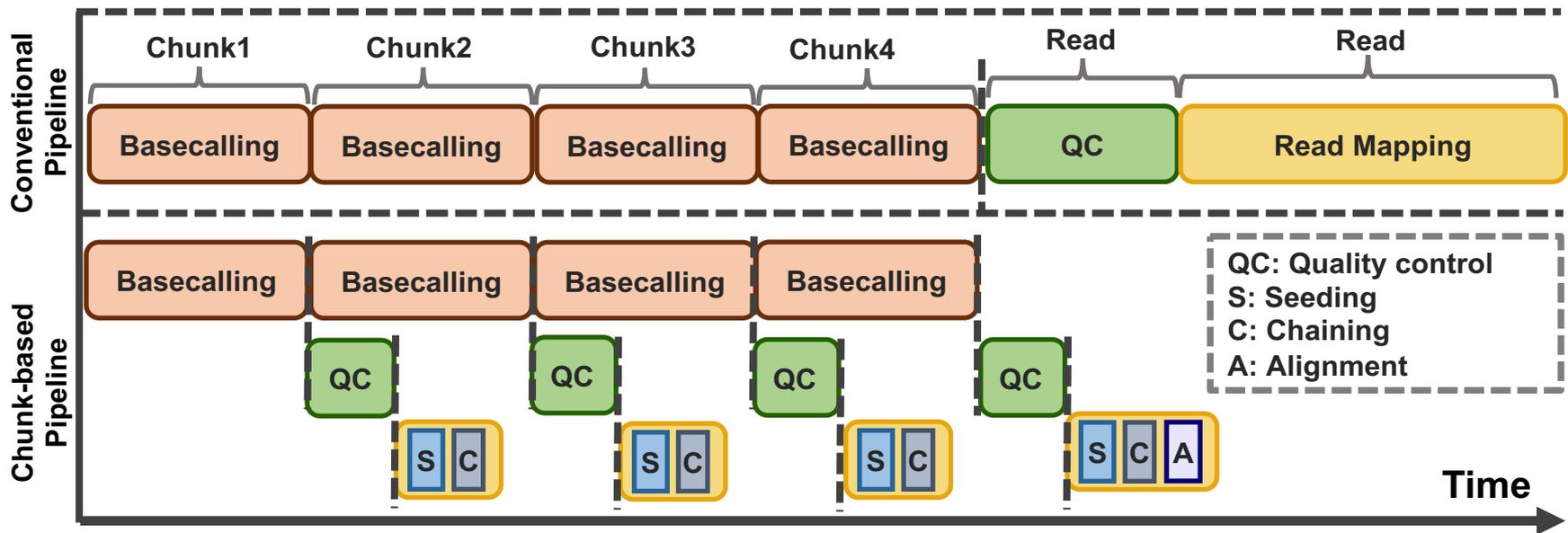
Haiyu Mao¹ Mohammed Alser¹ Mohammad Sadrosadati¹ Can Firtina¹ Akanksha Baranwal¹
Damla Senol Cali² Aditya Manglik¹ Nour Almadhoun Alserr¹ Onur Mutlu¹

¹*ETH Zürich*

²*Bionano Genomics*

Innovations Require Change

- CP processes reads at the granularity of a chunk instead of the complete read sequence, increasing parallelism and resource utilization by overlapping the execution of different steps.



GenPIP provides 41.6x and 8.4x speedup and 32.8x and 20.8x energy reduction compared to CPU and GPU state-of-the-art solutions.

SeGraM (ISCA 2022)

Damla Senol Cali, Konstantinos Kanellopoulos, Joel Lindegger, Zülal Bingöl, Gurpreet S. Kalsi, Ziyi Zuo, Can Firtina, Meryem Banu Cavlak, Jeremie Kim, Nika Mansouri Ghiasi, Gagandeep Singh, Juan Gómez-Luna, Nour Almadhoun Alserr, **Mohammed Alser**, Sreenivas Subramoney, Can Alkan, Saugata Ghose, Onur Mutlu
["SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping"](#),
Proceedings of the International Symposium on Computer Architecture (ISCA 2022).

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

Damla Senol Cali¹ Konstantinos Kanellopoulos² Joël Lindegger² Zülal Bingöl³
Gurpreet S. Kalsi⁴ Ziyi Zuo⁵ Can Firtina² Meryem Banu Cavlak² Jeremie Kim²
Nika Mansouri Ghiasi² Gagandeep Singh² Juan Gómez-Luna² Nour Almadhoun Alserr²
Mohammed Alser² Sreenivas Subramoney⁴ Can Alkan³ Saugata Ghose⁶ Onur Mutlu²

¹Bionano Genomics ²ETH Zürich ³Bilkent University ⁴Intel Labs

⁵Carnegie Mellon University ⁶University of Illinois Urbana-Champaign

Specialized Hardware for Pre-alignment Filtering

Mohammed Alser, Taha Shahroodi, Juan-Gomez Luna, Can Alkan, and Onur Mutlu,
"SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs"

Bioinformatics, 2020.

[[Source Code](#)]

[[Online link at Bioinformatics Journal](#)]

Bioinformatics



SneakySnake: a fast and accurate universal genome pre-alignment filter for CPUs, GPUs and FPGAs

Mohammed Alser ✉, Taha Shahroodi, Juan Gómez-Luna, Can Alkan ✉, Onur Mutlu ✉

Bioinformatics, btaa1015, <https://doi.org/10.1093/bioinformatics/btaa1015>

Published: 26 December 2020 **Article history** ▼

GenStore (ASPLOS 2022)

Nika Mansouri Ghiasi, Jisung Park, Harun Mustafa, Jeremie Kim, Ataberk Olgun, Arvid Gollwitzer, Damla Senol Cali, Can Firtina, Haiyu Mao, Nour Almadhoun Alserr, Rachata Ausavarungnirun, Nandita Vijaykumar, **Mohammed Alser**, Onur Mutlu

["GenStore: A High-Performance and Energy-Efficient In-Storage Computing System for Genome Sequence Analysis"](#),

ASPLOS 2022

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

Nika Mansouri Ghiasi
ETH Zürich
Switzerland

Jisung Park
ETH Zürich
Switzerland

Harun Mustafa
ETH Zürich
Switzerland

Jeremie Kim
ETH Zürich
Switzerland

Ataberk Olgun
ETH Zürich
Switzerland

Arvid Gollwitzer
ETH Zürich
Switzerland

Damla Senol Cali
Bionano Genomics
USA

Can Firtina
ETH Zürich
Switzerland

Haiyu Mao
ETH Zürich
Switzerland

Nour Almadhoun
Alserr
ETH Zürich
Switzerland

Rachata
Ausavarungnirun
KMUTNB
Thailand

Nandita Vijaykumar
University of Toronto
Canada

Mohammed Alser
ETH Zürich
Switzerland

Onur Mutlu
ETH Zürich
Switzerland

GateKeeper [Alser+, Bioinformatics 2017]

Mohammed Alser, Hasan Hassan, Hongyi Xin, Oguz Ergin, Onur Mutlu, and Can Alkan
"GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping"

Bioinformatics, [published online, May 31], 2017.

[\[Source Code\]](#)

[\[Online link at Bioinformatics Journal\]](#)

Bioinformatics

iSCB
INTERNATIONAL SOCIETY FOR
COMPUTATIONAL BIOLOGY

Article Navigation

GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping FREE

Mohammed Alser ✉, Hasan Hassan, Hongyi Xin, Oğuz Ergin, Onur Mutlu ✉, Can Alkan ✉

Bioinformatics, Volume 33, Issue 21, 01 November 2017, Pages 3355–3363,

<https://doi.org/10.1093/bioinformatics/btx342>

Published: 31 May 2017 **Article history** ▼

MAGNET

Mohammed Alser, Onur Mutlu, and Can Alkan.

["MAGNET: understanding and improving the accuracy of genome pre-alignment filtering"](#)

IPSI Transaction (2017).

[[Source code](#)]

MAGNET: Understanding and Improving the Accuracy of Genome Pre-Alignment Filtering

Alser, Mohammed; Mutlu, Onur; and Alkan, Can

Shouji (障子) [Alser+, Bioinformatics 2019]

Mohammed Alser, Hasan Hassan, Akash Kumar, Onur Mutlu, and Can Alkan,
"Shouji: A Fast and Efficient Pre-Alignment Filter for Sequence Alignment"
Bioinformatics, [published online, March 28], 2019.

[\[Source Code\]](#)

[\[Online link at Bioinformatics Journal\]](#)

Bioinformatics, 2019, 1–9

doi: 10.1093/bioinformatics/btz234

Advance Access Publication Date: 28 March 2019

Original Paper



Sequence alignment

Shouji: a fast and efficient pre-alignment filter for sequence alignment

**Mohammed Alser^{1,2,3,*}, Hasan Hassan¹, Akash Kumar², Onur Mutlu^{1,3,*}
and Can Alkan^{3,*}**

¹Computer Science Department, ETH Zürich, Zürich 8092, Switzerland, ²Chair for Processor Design, Center For Advancing Electronics Dresden, Institute of Computer Engineering, Technische Universität Dresden, 01062 Dresden, Germany and ³Computer Engineering Department, Bilkent University, 06800 Ankara, Turkey

In-Memory Sequence Analysis GRIM-Filter

- Jeremie S. Kim, Damla Senol Cali, Hongyi Xin, Donghyuk Lee, Saugata Ghose, **Mohammed Alser**, Hasan Hassan, Oguz Ergin, Can Alkan, and Onur Mutlu, "[GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies](#)"
to appear in [BMC Genomics](#), 2018.
*Proceedings of the [16th Asia Pacific Bioinformatics Conference \(APBC\)](#),
Yokohama, Japan, January 2018.
[arxiv.org Version \(pdf\)](#)*

BMC Genomics

Research | [Open Access](#) | [Published: 09 May 2018](#)

GRIM-Filter: Fast seed location filtering in DNA read mapping using processing-in-memory technologies

[Jeremie S. Kim](#) ✉, [Damla Senol Cali](#), [Hongyi Xin](#), [Donghyuk Lee](#), [Saugata Ghose](#), [Mohammed Alser](#), [Hasan Hassan](#), [Oguz Ergin](#), [Can Alkan](#) ✉ & [Onur Mutlu](#) ✉

[BMC Genomics](#) **19**, Article number: 89 (2018) | [Cite this article](#)

4340 Accesses | **39** Citations | **9** Altmetric | [Metrics](#)

Near-memory Pre-alignment Filtering

Gagandeep Singh, **Mohammed Alser**, Damla Senol Cali, Dionysios Diamantopoulos, Juan Gomez-Luna, Henk Corporaal, Onur Mutlu,

[“FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications”](#)

IEEE Micro, 2021.

[\[Source Code\]](#)



[Home](#) / [Magazines](#) / [IEEE Micro](#) / [2021.04](#)

IEEE Micro

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications

July-Aug. 2021, pp. 39-48, vol. 41

DOI Bookmark: [10.1109/MM.2021.3088396](https://doi.org/10.1109/MM.2021.3088396)

Authors

[Gagandeep Singh](#), ETH Zürich, Zürich, Switzerland

[Mohammed Alser](#), ETH Zürich, Zürich, Switzerland

[Damla Senol Cali](#), Carnegie Mellon University, Pittsburgh, PA, USA

[Dionysios Diamantopoulos](#), Zürich Lab, IBM Research Europe, Rüschlikon, Switzerland

[Juan Gomez-Luna](#), ETH Zürich, Zürich, Switzerland

[Henk Corporaal](#), Eindhoven University of Technology, Eindhoven, The Netherlands

[Onur Mutlu](#), ETH Zürich, Zürich, Switzerland

◀	▶
Previous	Next
☰ Table of Contents	
📄 Past Issues	

GenASM Framework [MICRO 2020]

- Damla Senol Cali, Gurpreet S. Kalsi, Zülal Bingöl, Can Firtina, Lavanya Subramanian, Jeremie S. Kim, Rachata Ausavarungnirun, **Mohammed Alser**, Juan Gomez-Luna, Amirali Boroumand, Anant Nori, Allison Scibisz, Sreenivas Subramoney, Can Alkan, Saugata Ghose, and Onur Mutlu, "[GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis](#)"
Proceedings of the [53rd International Symposium on Microarchitecture \(MICRO\)](#), Virtual, October 2020.
[[Lightning Talk Video](#) (1.5 minutes)]
[[Lightning Talk Slides \(pptx\)](#) ([pdf](#))]
[[Talk Video](#) (18 minutes)]
[[Slides \(pptx\)](#) ([pdf](#))]

GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

Damla Senol Cali[†][✕] Gurpreet S. Kalsi[✕] Zülal Bingöl[∇] Can Firtina[◇] Lavanya Subramanian[‡] Jeremie S. Kim[◇][†]
Rachata Ausavarungnirun[○] Mohammed Alser[◇] Juan Gomez-Luna[◇] Amirali Boroumand[†] Anant Nori[✕]
Allison Scibisz[†] Sreenivas Subramoney[✕] Can Alkan[∇] Saugata Ghose^{*†} Onur Mutlu[◇][∇]
[†]Carnegie Mellon University [✕]Processor Architecture Research Lab, Intel Labs [∇]Bilkent University [◇]ETH Zürich
[‡]Facebook [○]King Mongkut's University of Technology North Bangkok ^{*}University of Illinois at Urbana-Champaign

Demeter (HD Food Microbiome Profiling)

Taha Shahroodi, Mahdi Zahedi, Can Firtina, **Mohammed Alser**, Stephan Wong, Onur Mutlu, Said Hamdioui

[“Demeter: A Fast and Energy-Efficient Food Profiler using Hyperdimensional Computing in Memory”](#)

IEEE Access, 2022

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

 **RESEARCH ARTICLE**

Demeter: A Fast and Energy-Efficient Food Profiler Using Hyperdimensional Computing in Memory

**TAHA SHAHROODI^{ID1}, MAHDI ZAHEDI^{ID1}, CAN FIRTINA², MOHAMMED ALSER^{ID2},
STEPHAN WONG¹, (Senior Member, IEEE), ONUR MUTLU^{ID2}, (Fellow, IEEE),
AND SAID HAMDIOUI^{ID1}, (Senior Member, IEEE)**

¹Q&CE Department, EEMCS Faculty, Delft University of Technology (TU Delft), 2628 CD Delft, The Netherlands

²SAFARI Research Group, D-ITET, ETH Zürich, 8092 Zürich, Switzerland

AIM (PIM Sequence Alignment Framework)

Safaa Diab, Amir Nassereldine, **Mohammed Alser**, Juan Gómez-Luna,
Onur Mutlu, Izzat El Hajj

[“A Framework for High-throughput Sequence Alignment using Real Processing-in-Memory Systems”](#)

arXiv, 2022

[\[Source code\]](#)

A Framework for High-throughput Sequence Alignment using Real Processing-in-Memory Systems

Safaa Diab¹, Amir Nassereldine¹, Mohammed Alser², Juan Gómez Luna², Onur Mutlu², Izzat El Hajj¹

¹*American University of Beirut, Lebanon* ²*ETH Zürich, Switzerland*

Fostering Omics Research

National
Strategy

Genome Map

National
Genomics
Centers

Genome
Banks

Genomics
Startups &
Companies

Genomics
Education &
Experts

Intelligent Genome Analysis

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

["From Molecules to Genomic Variations: Intelligent Algorithms and Architectures for Intelligent Genome Analysis"](#)

Computational and Structural Biotechnology Journal, 2022

[[Source code](#)]

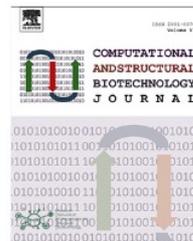


ELSEVIER

010100100101010010
0010101001010101011
1010101001010101011
010101001010101010
110101001010101010
1010101001010101011
0010101001010101011
010101001010101010
11010101001010101010

COMPUTATIONAL
AND STRUCTURAL
BIOTECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj



Review

From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures



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

ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland

Accelerating Genome Analysis

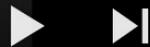
How Large is a Genome?



Prime Tower, Zurich



~3.2 billion genomic bases



7:02 / 1:44:35

SAFARI



Livestream - Seminar in Computer Architecture - ETH Zürich (Spring 2022)

Seminar in Computer Arch. - Lecture 5: Accelerating Genome Analysis (Spring 2022)

More on Intelligent Genome Analysis ...

- Mohammed Alser,
"Computer Architecture - Lecture 8: Intelligent Genome Analysis"
ETH Zurich, Computer Architecture Course, Lecture 8, Virtual, 15 October 2021.
[\[Slides \(pptx\) \(pdf\)\]](#)
[\[Talk Video \(2 hour 54 minutes, including Q&A\)\]](#)
[\[Related Invited Paper \(at IEEE Micro, 2020\)\]](#)

Our Solution: GateKeeper

Alignment Filter + [FPGA board] = 1st FPGA-based Alignment Filter.

Low Speed & High Accuracy
Medium Speed, Medium Accuracy
High Speed, Low Accuracy

x10¹² mappings

x10³ mappings

1 High throughput DNA sequencing (HTS) technologies
2 Read Pre-Alignment Filtering Fast & Low False Positive Rate
3 Read Alignment Slow & Zero False Positives

Billions of Short Reads

108

2:08:58 / 2:54:18 • GateKeeper >

ETH ZENTRUM

Preparing Students To Rethink Genomic Analyses

Mohammed Alser

ETH Zurich

 @meals

Institut Pasteur de Tunis

21 November 2022

