P&S Genomics

Lecture 2: Intelligent Genomic Analyses

Dr. Mohammed Alser

ETH Zürich
Spring 2023
9 March 2023

Mohammed Alser

- Lecturer and Senior Researcher, SAFARI Research Group, ETH Zürich, since Sept. 2018.
- PhD from Bilkent University (Turkey) 2018, worked at UCLA, TU Dresden, and PETRONAS.
- Received the IEEE Turkey Doctoral Dissertation Award and a number of international prestigious awards.
- https://twitter.com/mealser
- My main research is in bioinformatics, computational genomics, metagenomics, and computer architecture.
- I am especially excited about **building** new data structures, algorithms, and architectures that **make intelligent genome analysis a reality**.

Agenda for Today

- What is Genome Analysis?
- What is Intelligent Genome Analysis?

- How we Analyze Genome?
- What are the Barriers to Enabling Intelligent Analyses?

- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

- Where is Genomic Analyses Going Next?

Agenda for Today

- **What is Genome Analysis?**
- What is Intelligent Genome Analysis?

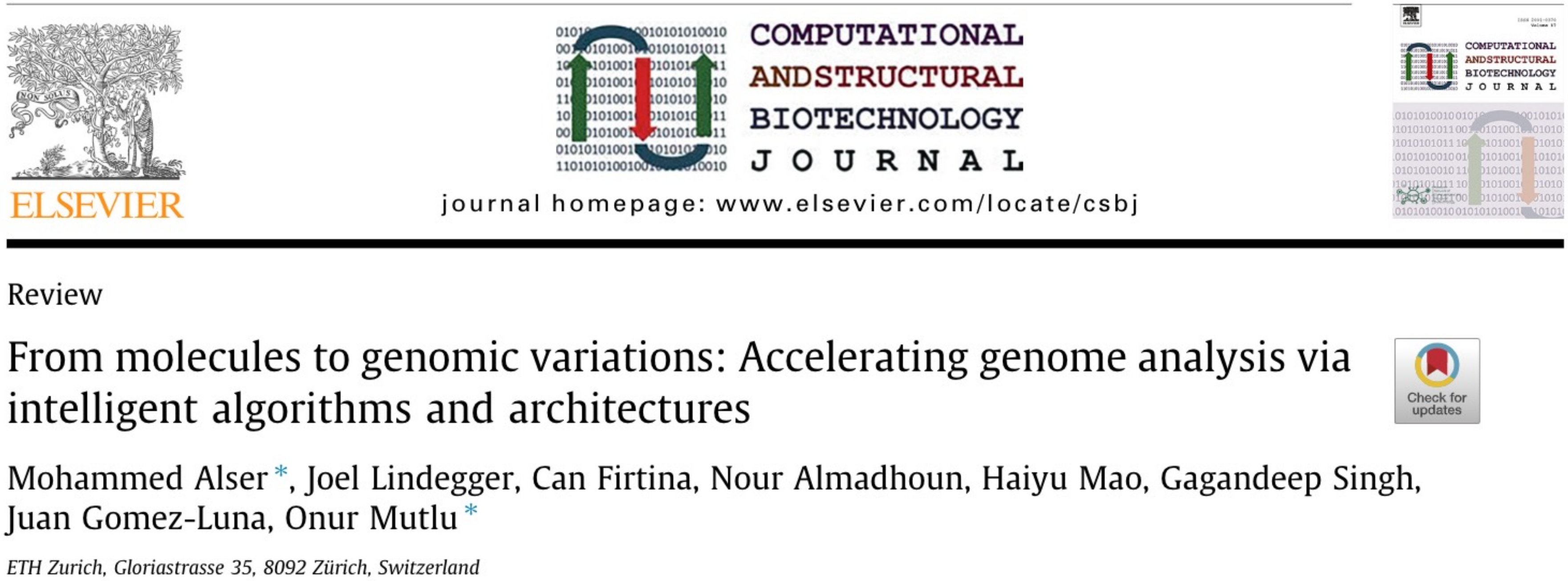
- How we Analyze Genome?
- What are the Barriers to Enabling Intelligent Analyses?

- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

- Where is Genomic Analyses Going Next?

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

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Review

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

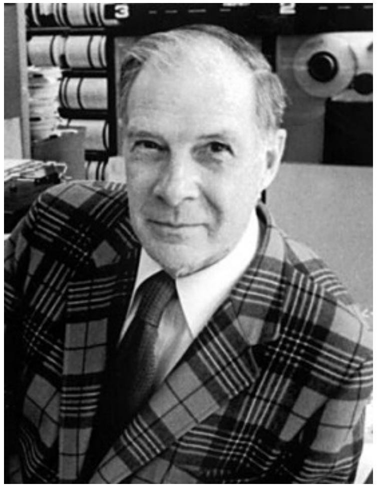
Mohammed Alser *, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu *

ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland

What is Data Analysis?

"The purpose of computing is [to gain] insight, not numbers"

Richard Hamming



We need to gain insights and observations much more efficiently than ever before

Major Generators of Big Data

Big data is everywhere …

Astronomy
25 zetta-bytes/year

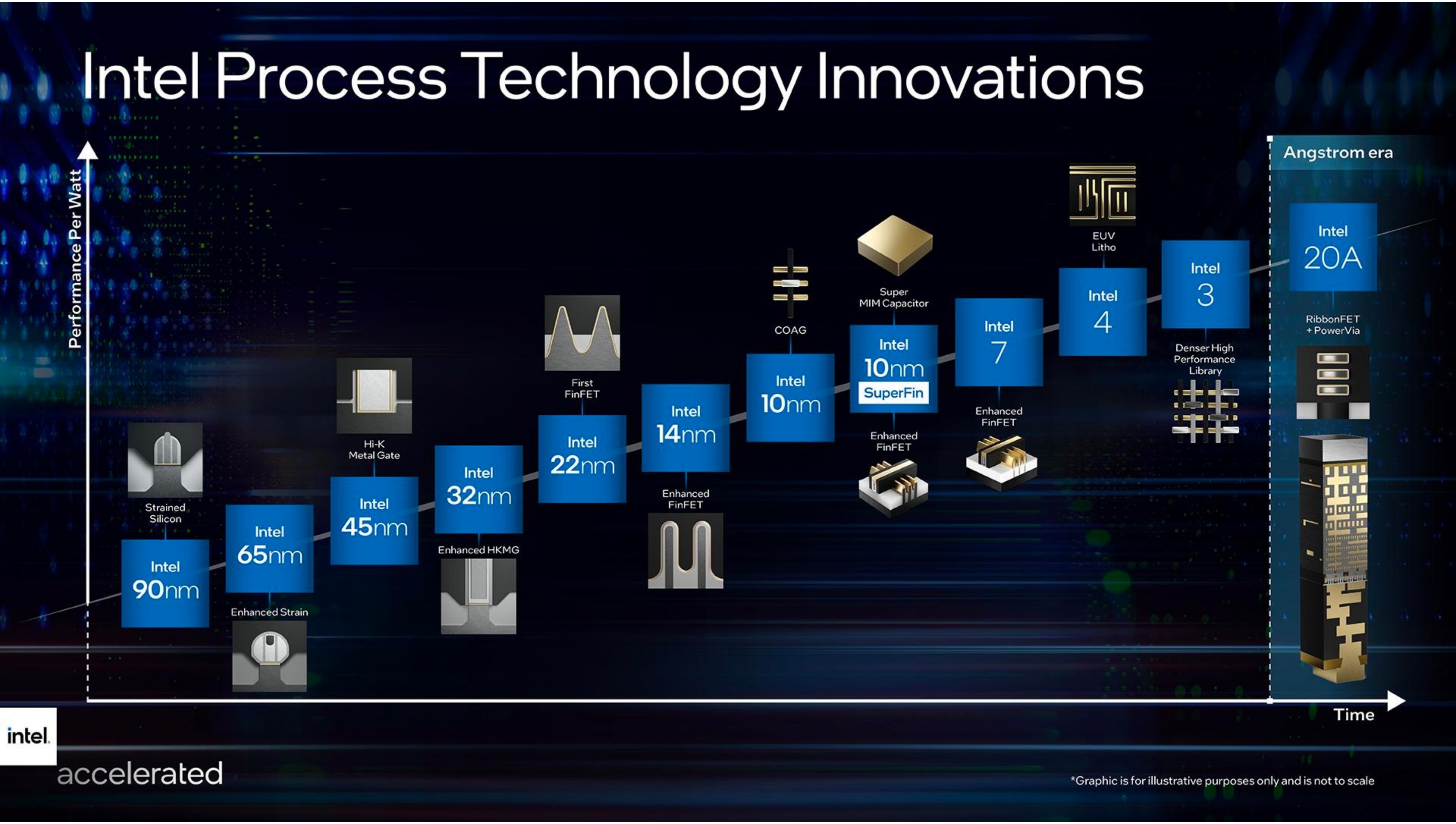
Twitter
0.5-15 billion tweets/year

YouTube
500-900 million hours/year

Genomics
1 zetta-bases/year

"Big data: astronomical or genomical?", PLoS biology, 2015.

Angstrom (10^{-10}m) Era of Semiconductors



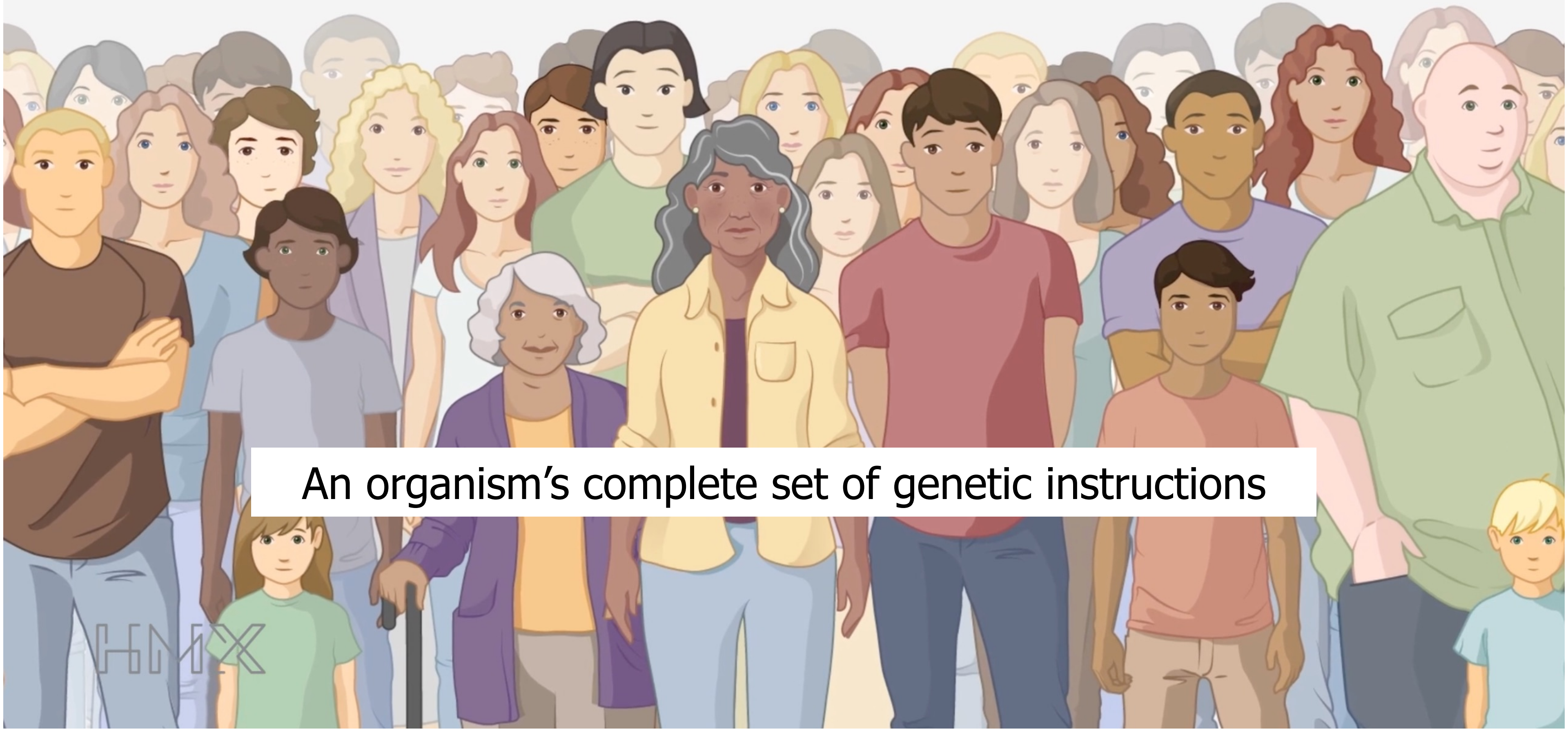
https://siliconangle.com/2021/07/26/angstrom-era-intel-unveils-ambitious-semiconductor-roadmap-goes-beyond-1nm-chips/

What is Intelligent Data Analysis?

- The **science and art** of revealing previously unknown and potentially valuable **information or knowledge** from **data** while meeting functional, performance, energy consumption, cost, and other specific goals

Adapted from the definition of data mining

What is a Genome?



An organism’s complete set of genetic instructions

https://onlinelearning.hms.harvard.edu/hmx/courses/genetic-testing/

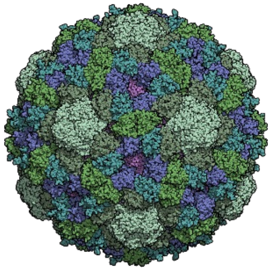
CCTCCTCAGTGCCACCCAGCCCACTGGCAGCTCCCAAACA
GGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTGGAGTG
AGGTGTCAAGGACCTAAACTAAAAAAAAAAAAAGAAAA
AGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAA
AAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATG
TGCTAAACAGCACTTTTT**TTGACCATTAT**TTTGGATCTGAAA
GAAATCAAGAATAAATGAAGGACTTGATACATTGGAAGA
GGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAGAAA
AAGAAAAGAAAAAGA**A**TTTAAAATTTAAGTAATTCTTTGA
AAAAAACTAATTTCTAAGCTTCTT**C**ATGTCAAGGACCTAAT
GTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAAGA
AAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTA
ATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAGGCC
GGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTGGAGTG

How Large is a Genome?

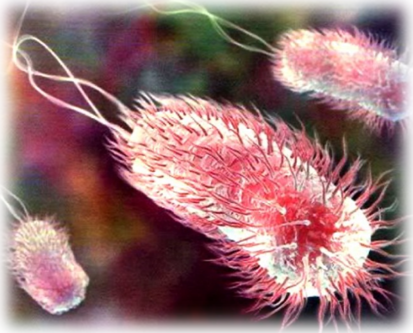
Prime Tower, Zurich

~3.2 billion genomic bases

How About Other Species?



Phi X174 virus
5.386 Killo bp



E. coli O157:H7
5.44 Million bp

Homo Sapiens
3.2 Billion bp

Onion, Allium Cepa
16 Billion bp

Paris Japonica
149 Billion bp

DNA Testing

Health + Ancestry Service

$199

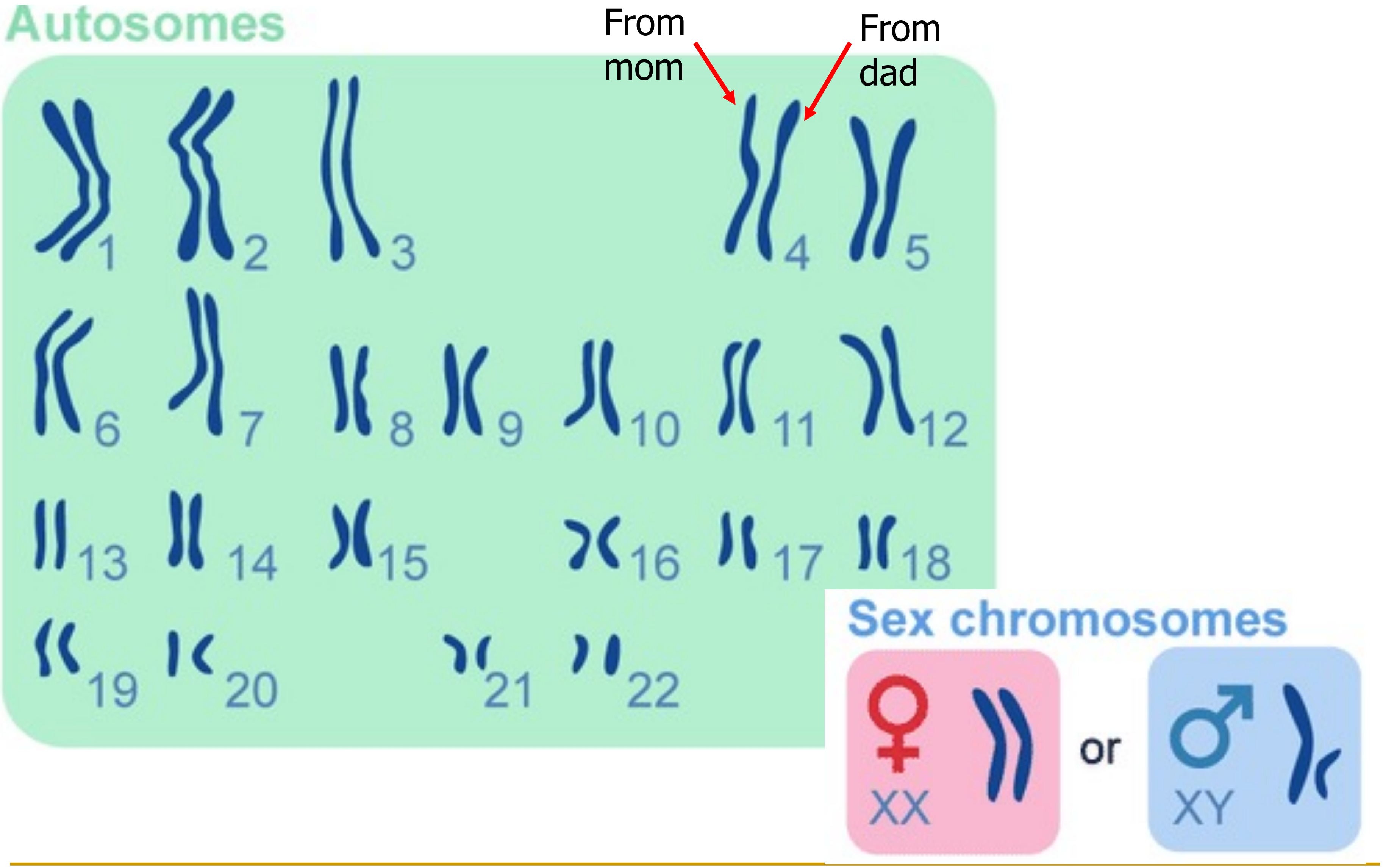
- Includes everything in Ancestry + Traits Service

PLUS

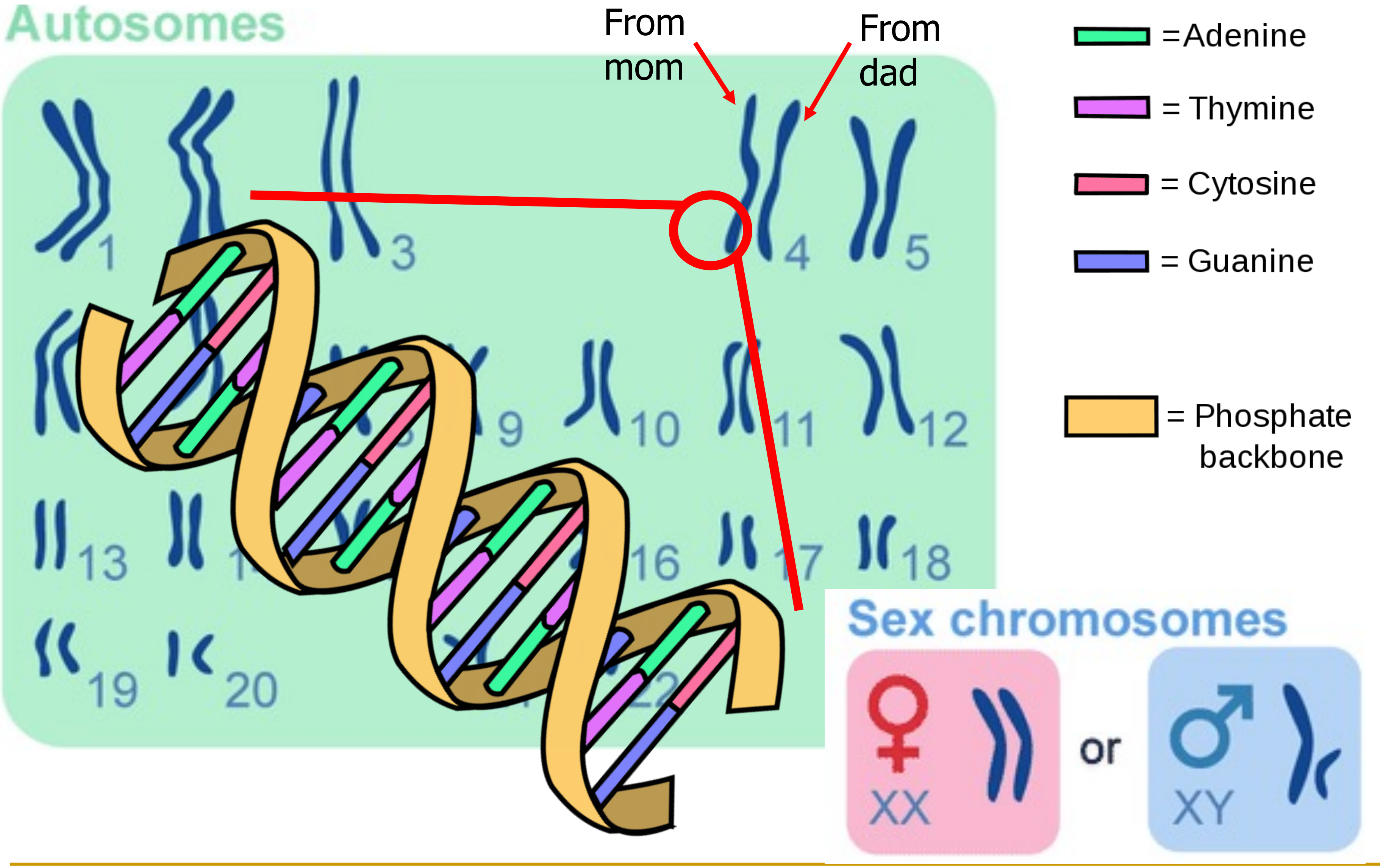
- 10+ Health Predisposition reports*
- 5+ Wellness reports
- 40+ Carrier Status reports*

https://www.myheritage.ch/dna https://www.23andme.com/

Human Chromosomes (23 Pairs)



Human Chromosomes (23 Pairs)



From mom

From dad

🟩 =Adenine

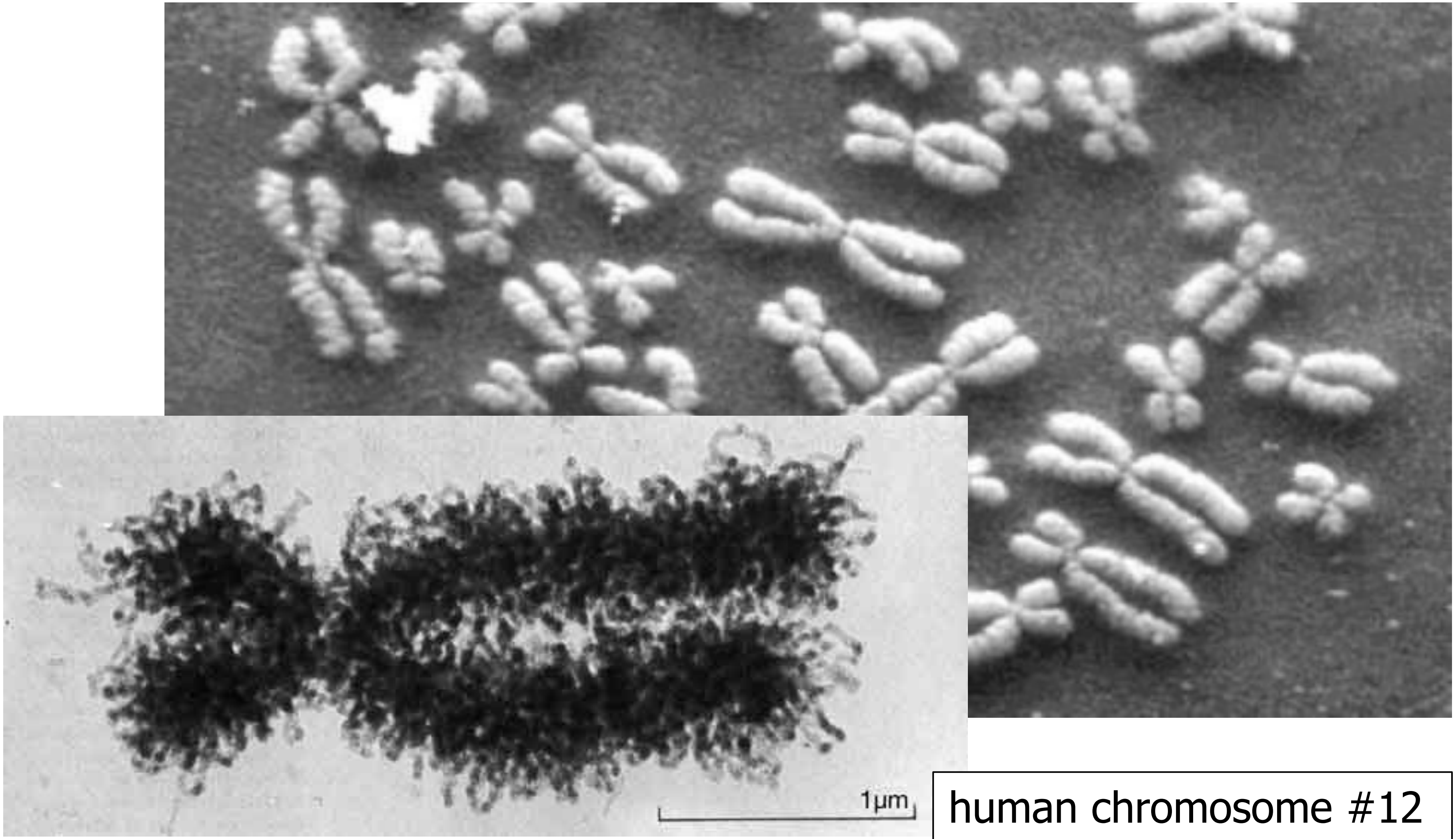
🟪 = Thymine

🟥 = Cytosine

🟦 = Guanine

🟨 = Phosphate backbone

DNA Under Electron Microscope



human chromosome #12 from HeLa's cell

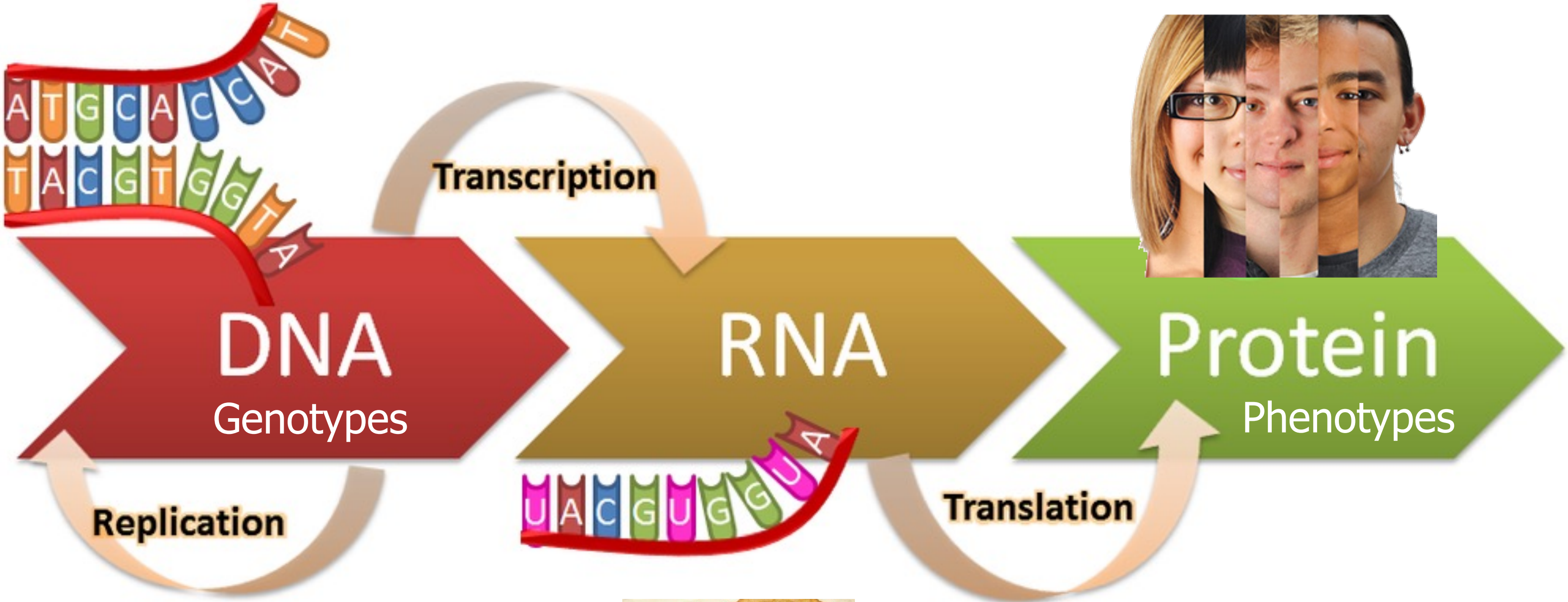
DNA Under Electron Microscope

human chromosome #12
from HeLa’s cell

DNA Under Electron Microscope

human chromosome #12 from HeLa's cell

The Central Dogma of Molecular Biology



Transcription

DNA

Genotypes

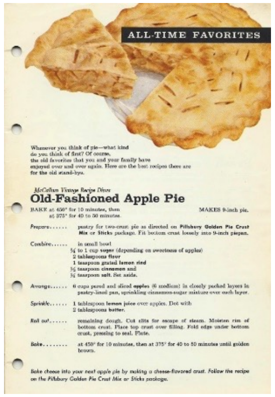
Replication

RNA

Translation

Protein

Phenotypes



ALL-TIME FAVORITES

Whenever you think of pie—what kind do you think of first? Of course, the old favorites that you and your family have enjoyed over and over again. Here are the best recipes there are for the old stand-bys.

McCallum Vintage Recipe Diver

Old-Fashioned Apple Pie

BAKE at 450° for 10 minutes, then at 375° for 40 to 50 minutes. MAKES 9-inch pie.

Prepare...... pastry for two-crust pie as directed on **Pillsbury Golden Pie Crust Mix** or **Sticks** package. Fit bottom crust loosely into 9-inch piepan.

Combine...... in small bowl
¾ to 1 cup **sugar** (depending on sweetness of apples)
2 tablespoons **flour**
1 teaspoon grated **lemon rind**
½ teaspoon **cinnamon** and
¼ teaspoon **salt**. Set aside.

Arrange...... 6 cups pared and sliced **apples** (6 medium) in closely packed layers in pastry-lined pan, sprinkling cinnamon-sugar mixture over each layer.

Sprinkle...... 1 tablespoon **lemon juice** over apples. Dot with 2 tablespoons **butter**.

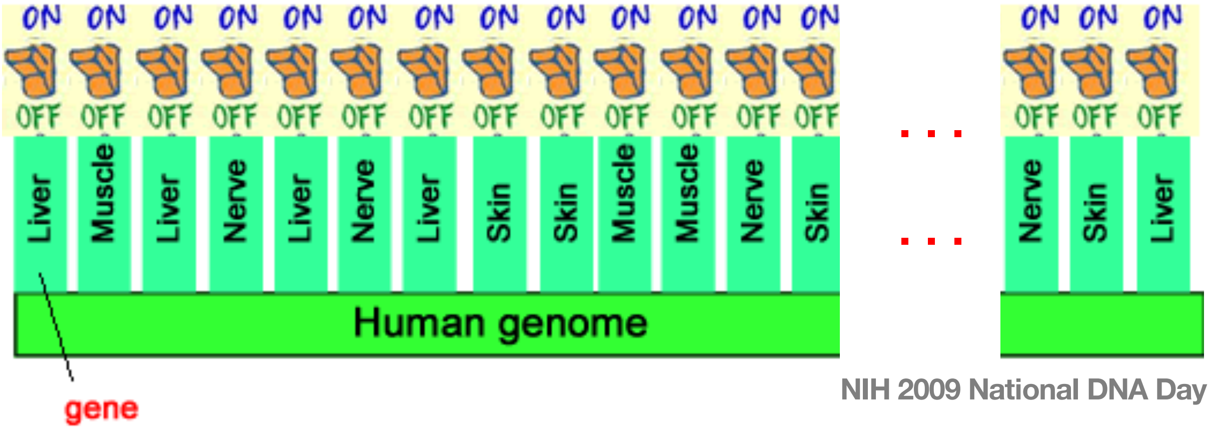
Roll out...... remaining dough. Cut slits for escape of steam. Moisten rim of bottom crust. Place top crust over filling. Fold edge under bottom crust, pressing to seal. Flute.

Bake........ at 450° for 10 minutes, then at 375° for 40 to 50 minutes until golden brown.

Bake cheese into your next apple pie by making a cheese-flavored crust. Follow the recipe on the Pillsbury Golden Pie Crust Mix or Sticks package.

Cells of Different Organs and Tissues

- All the cells in a person's body have the same DNA and the same genes.
  - Expression of the genes differs between cells.
  - But not all genes are used or expressed by those cells.



NIH 2009 National DNA Day

20,000-25,000 human genes

Finding SNPs Associated with Complex Trait

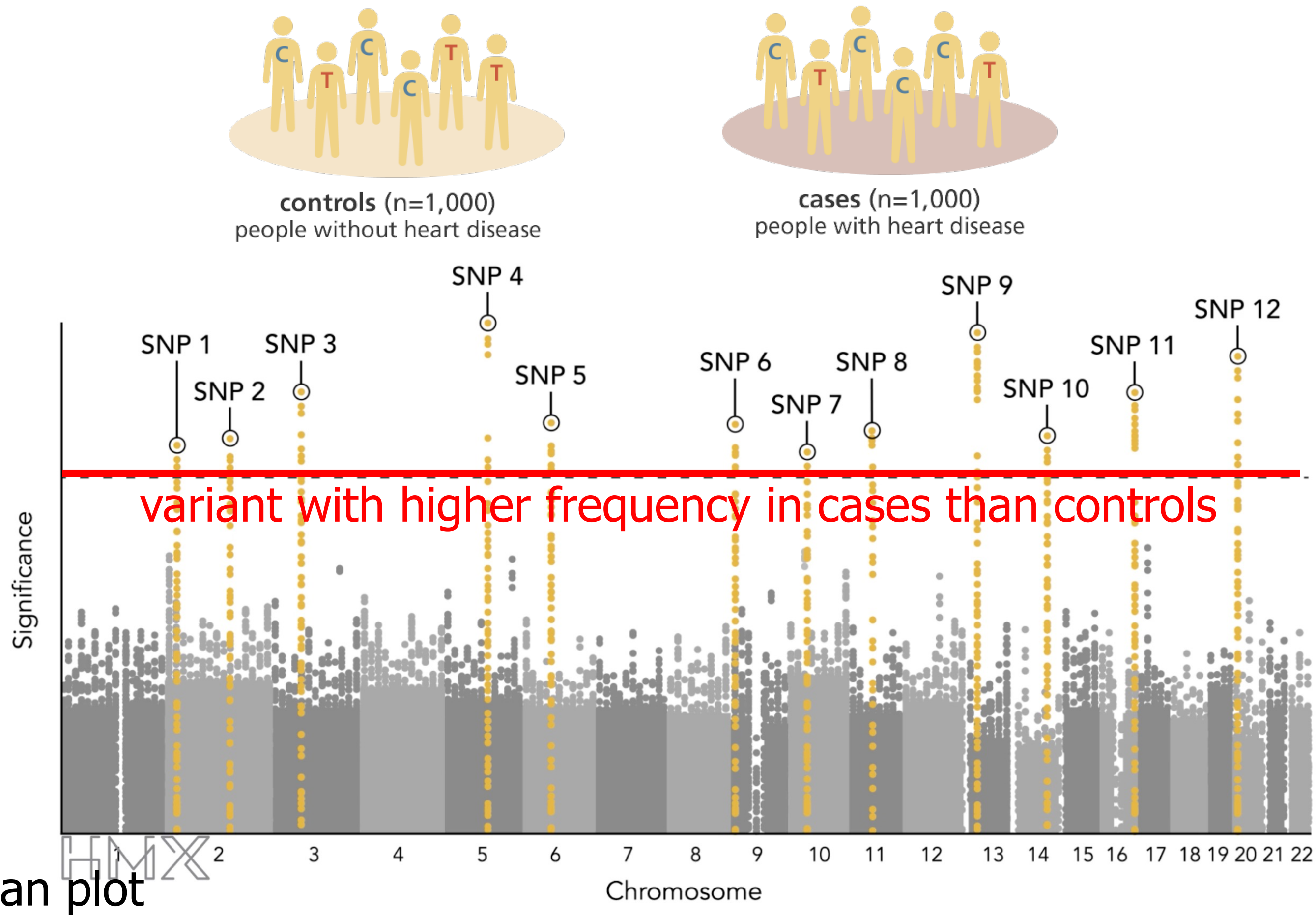
	SNP1 SNP2	Blood Pressure
Individual #1	...ACATGCCGACATTTCATAGGCC...	180
Individual #2	...ACATGCCGACATTTCATAAGCC...	175
Individual #3	...ACATGCCGACATTTCATAGGCC...	170
Individual #4	...ACATGCCGACATTTCATAAGCC...	165
Individual #5	...ACATGCCGACATTTCATAGGCC...	160
Individual #6	...ACATGCCGACATTTCATAGGCC...	145
Individual #7	...ACATGCCGACATTTCATAAGCC...	140
Individual #8	...ACATGCCGACATTTCATAAGCC...	130
Individual #9	...ACATGTCGACATTTCATAGGCC...	120
Individual #10	...ACATGTCGACATTTCATAAGCC...	120
Individual #11	...ACATGTCGACATTTCATAGGCC...	115
Individual #12	...ACATGTCGACATTTCATAAGCC...	110
Individual #13	...ACATGTCGACATTTCATAGGCC...	110
Individual #14	...ACATGTCGACATTTCATAAGCC...	110
Individual #15	...ACATGTCGACATTTCATAGGCC...	105
Individual #16	...ACATGTCGACATTTCATAAGCC...	100

SNP: single nucleotide polymorphism

Eleazar Eskin: Discovering the Causal Variants Involved in GWAS Studies, CGSI 2018, UCLA

Genome-Wide Association Study (GWAS)

- Detecting genetic variants associated with phenotypes using two groups of people.

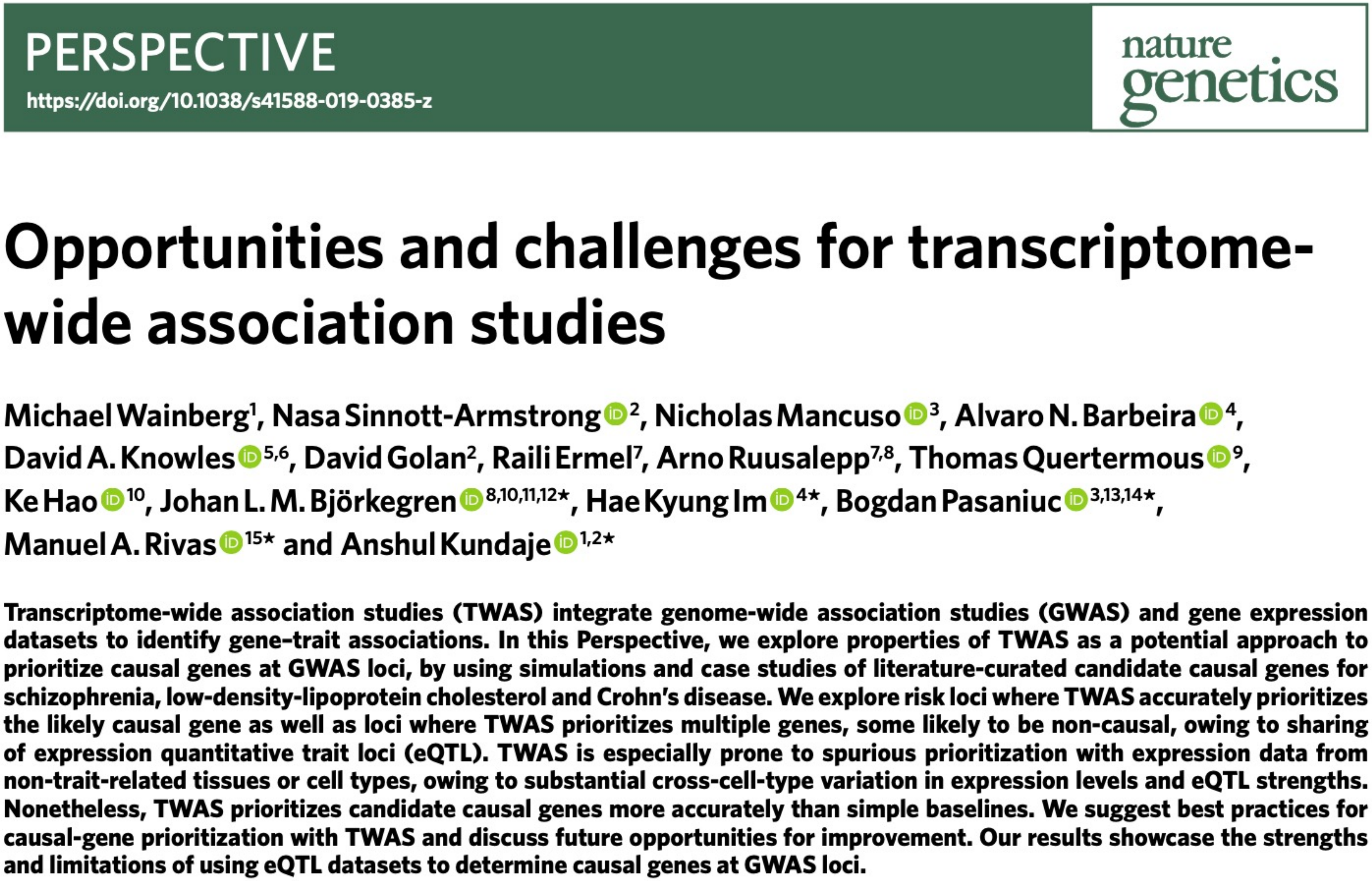


variant with higher frequency in cases than controls

Manhattan plot

https://onlinelearning.hms.harvard.edu/hmx/courses/genetic-testing/

Similar Association Studies



PERSPECTIVE
https://doi.org/10.1038/s41588-019-0385-z

nature genetics

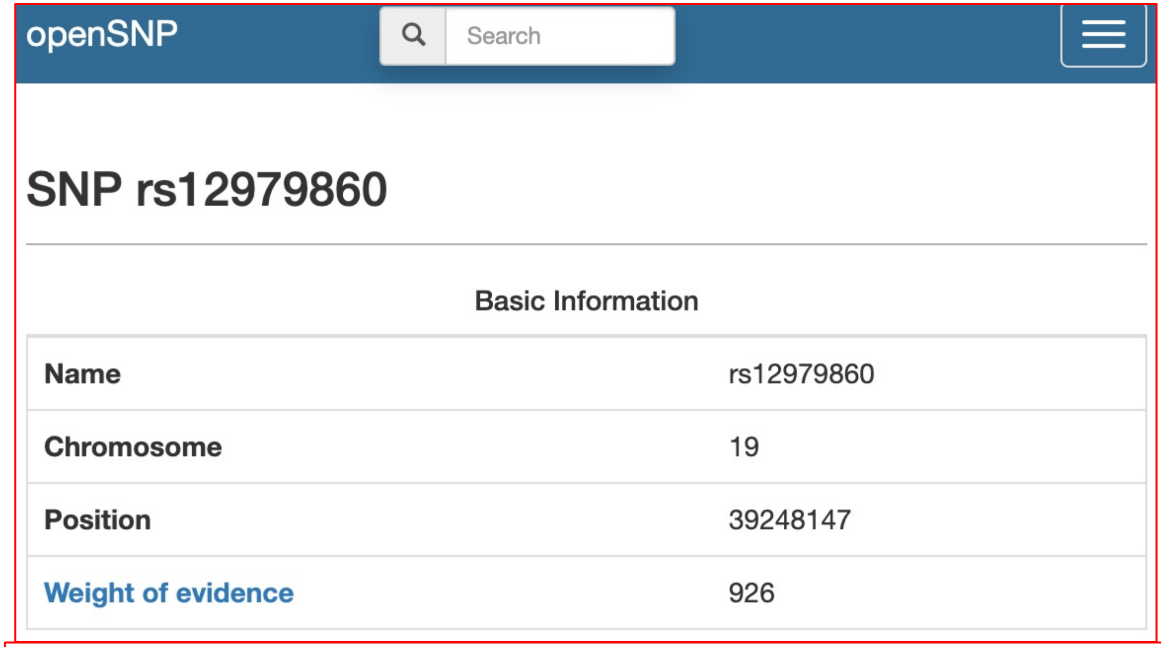
Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg[1], Nasa Sinnott-Armstrong[2], Nicholas Mancuso[3], Alvaro N. Barbeira[4], David A. Knowles[5,6], David Golan[2], Raili Ermel[7], Arno Ruusalepp[7,8], Thomas Quertermous[9], Ke Hao[10], Johan L. M. Björkegren[8,10,11,12]*, Hae Kyung Im[4]*, Bogdan Pasaniuc[3,13,14]*, Manuel A. Rivas[15]* and Anshul Kundaje[1,2]*

Transcriptome-wide association studies (TWAS) integrate genome-wide association studies (GWAS) and gene expression datasets to identify gene-trait associations. In this Perspective, we explore properties of TWAS as a potential approach to prioritize causal genes at GWAS loci, by using simulations and case studies of literature-curated candidate causal genes for schizophrenia, low-density-lipoprotein cholesterol and Crohn's disease. We explore risk loci where TWAS accurately prioritizes the likely causal gene as well as loci where TWAS prioritizes multiple genes, some likely to be non-causal, owing to sharing of expression quantitative trait loci (eQTL). TWAS is especially prone to spurious prioritization with expression data from non-trait-related tissues or cell types, owing to substantial cross-cell-type variation in expression levels and eQTL strengths. Nonetheless, TWAS prioritizes candidate causal genes more accurately than simple baselines. We suggest best practices for causal-gene prioritization with TWAS and discuss future opportunities for improvement. Our results showcase the strengths and limitations of using eQTL datasets to determine causal genes at GWAS loci.

Wainberg+, "Opportunities and challenges for transcriptome-wide association studies", *Nature genetics,* 2019.

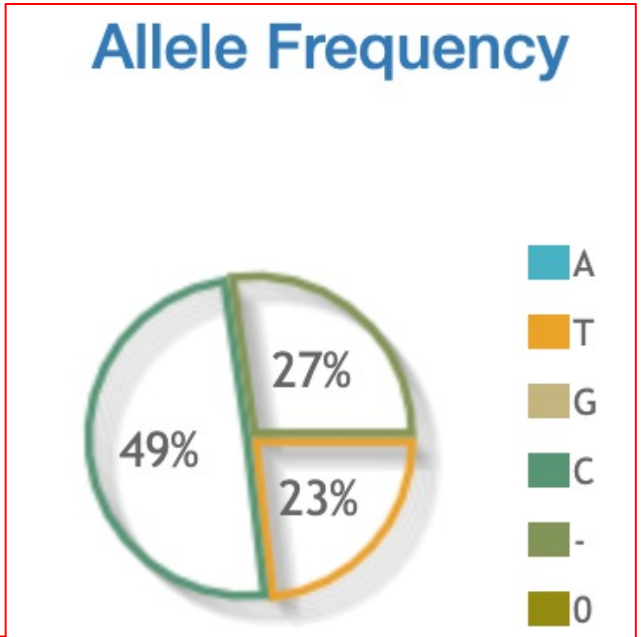
SNPs and Personalized Medicine

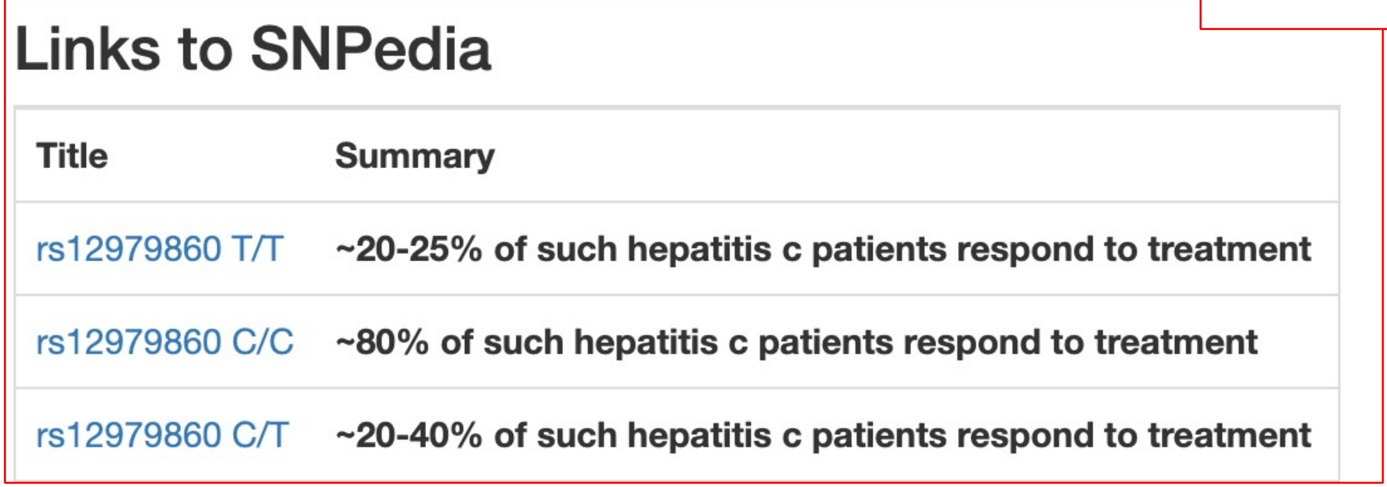


SNP rs12979860

Basic Information

Name	rs12979860
Chromosome	19
Position	39248147
Weight of evidence	926





Links to SNPedia

Title	Summary
rs12979860 T/T	~20-25% of such hepatitis c patients respond to treatment
rs12979860 C/C	~80% of such hepatitis c patients respond to treatment
rs12979860 C/T	~20-40% of such hepatitis c patients respond to treatment

https://opensnp.org/snps/rs12979860

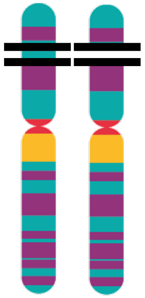
Much Larger Structural Variations!

AUTISM
Weiss, *N Eng J Med* 2008
Deletion of 593 kb

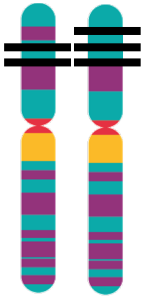
SCHIZOPHRENIA
McCarthy, *Nat Genet* 2009
Duplication of 593 kb

OBESITY
Walters, *Nature* 2010
Deletion of 593 kb

UNDERWEIGHT
Jacquemont, *Nature* 2011
Duplication of 593 kb



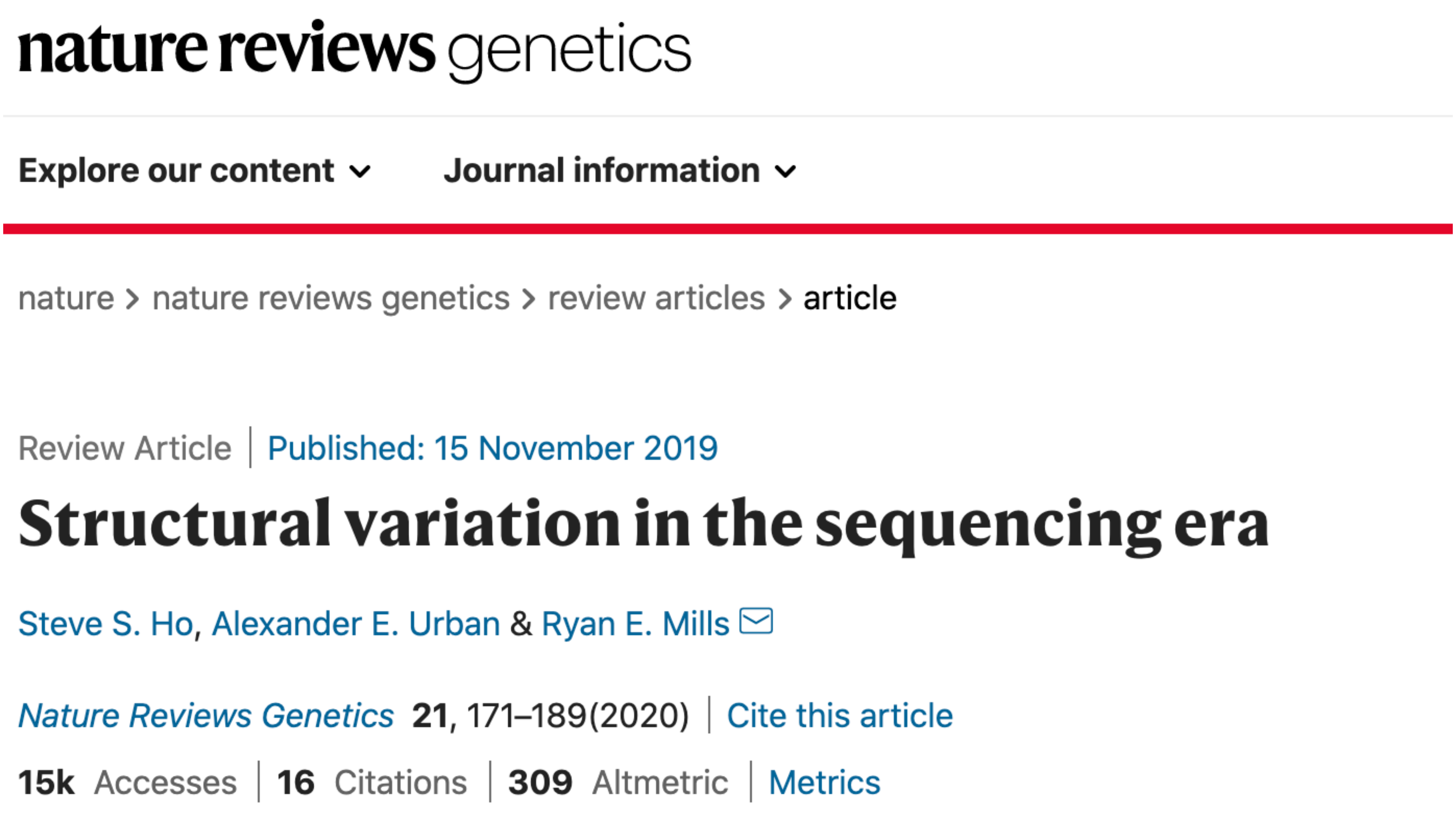
Deletion in the short arm of chromosome 16 (16p11.2)



Duplication in the short arm of chromosome 16 (16p11.2)

CNV: copy number variation

Recommended Reading



nature reviews genetics

Explore our content ⌄ Journal information ⌄

nature › nature reviews genetics › review articles › article

Review Article | Published: 15 November 2019

Structural variation in the sequencing era

Steve S. Ho, Alexander E. Urban & Ryan E. Mills ✉

Nature Reviews Genetics **21**, 171–189(2020) | Cite this article

15k Accesses | **16** Citations | **309** Altmetric | Metrics

Ho+, "Structural variation in the sequencing era", Nature Reviews Genetics, 2020

Agenda for Today

- What is Genome Analysis?
- **What is Intelligent Genome Analysis?**

- How we Analyze Genome?
- What are the Barriers to Enabling Intelligent Analyses?

- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

- Where is Genomic Analyses Going Next?

What is Intelligent Genome Analysis?

- Fast genome analysis
  - *Real-time analysis?*

Bandwidth

- Population-scale genome analysis
  - *Number of analyses per day!*

Scalability

- Using intelligent architectures
  - *Small specialized HW with less data movement*

Energy-efficiency & Portability

- DNA is a valuable asset
  - *Controlled-access analysis*

Privacy

- Avoiding erroneous analysis
  - *E.g., your father is not your father*

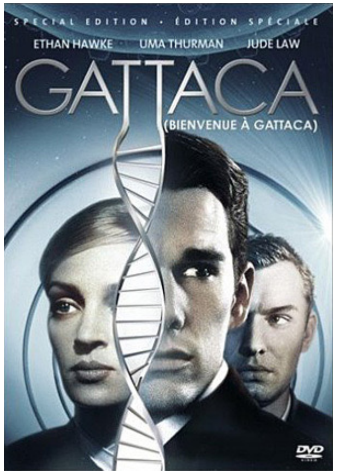
Accuracy

Does intelligent genome analysis really matter?

Fast Genome Analysis?

- **Fast** genome analysis in mere seconds using limited computational resources (i.e., personal computer or small hardware).

1997

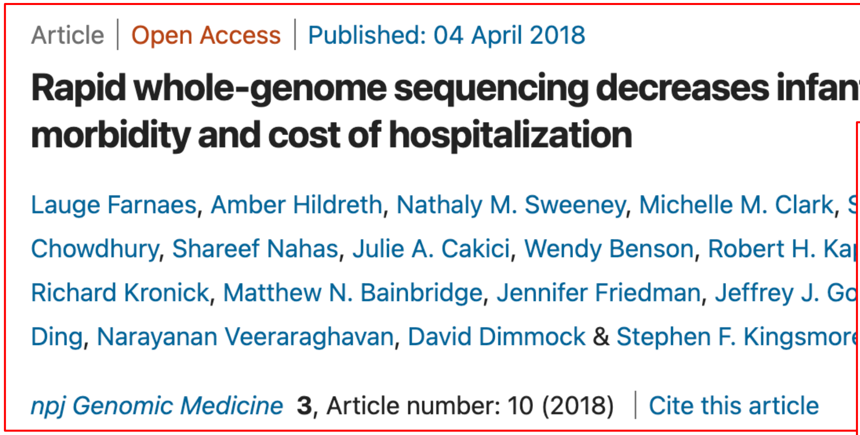


2015

FROM THE DIRECTOR OF THE INCREDIBLES & MISSION: IMPOSSIBLE - GHOST PROTOCOL
GEORGE CLOONEY
TOMORROWLAND
REMEMBER TOMORROW MAY 22
EXPERIENCE IT IN IMAX

Personalized Medicine for Critically Ill Infants

- rWGS can be performed in 2-day (costly) or 5-day time to interpretation.
- Diagnostic rWGS for infants
  - Avoids morbidity
  - Reduces hospital stay length by 6%-69%
  - Reduces inpatient cost by $800,000-$2,000,000.

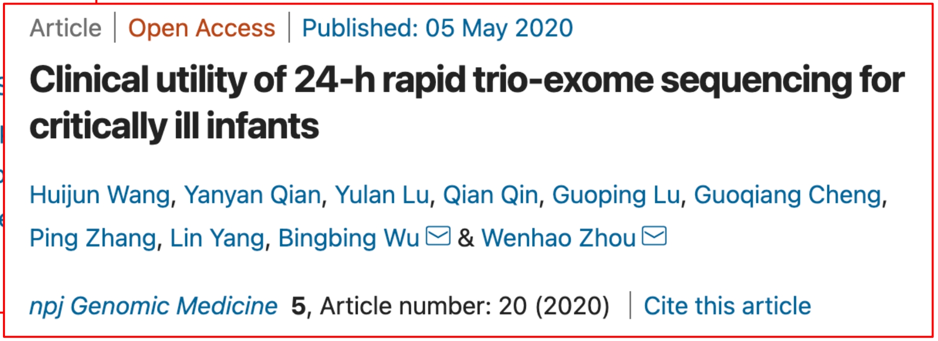


Article | Open Access | Published: 04 April 2018

Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization

Lauge Farnaes, Amber Hildreth, Nathaly M. Sweeney, Michelle M. Clark, S
Chowdhury, Shareef Nahas, Julie A. Cakici, Wendy Benson, Robert H. Ka
Richard Kronick, Matthew N. Bainbridge, Jennifer Friedman, Jeffrey J. Go
Ding, Narayanan Veeraraghavan, David Dimmock & Stephen F. Kingsmore

npj Genomic Medicine **3**, Article number: 10 (2018) | Cite this article



Article | Open Access | Published: 05 May 2020

Clinical utility of 24-h rapid trio-exome sequencing for critically ill infants

Huijun Wang, Yanyan Qian, Yulan Lu, Qian Qin, Guoping Lu, Guoqiang Cheng, Ping Zhang, Lin Yang, Bingbing Wu ✉ & Wenhao Zhou ✉

npj Genomic Medicine **5**, Article number: 20 (2020) | Cite this article

Farnaes+, "Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization", NPJ Genomic Medicine, 2018

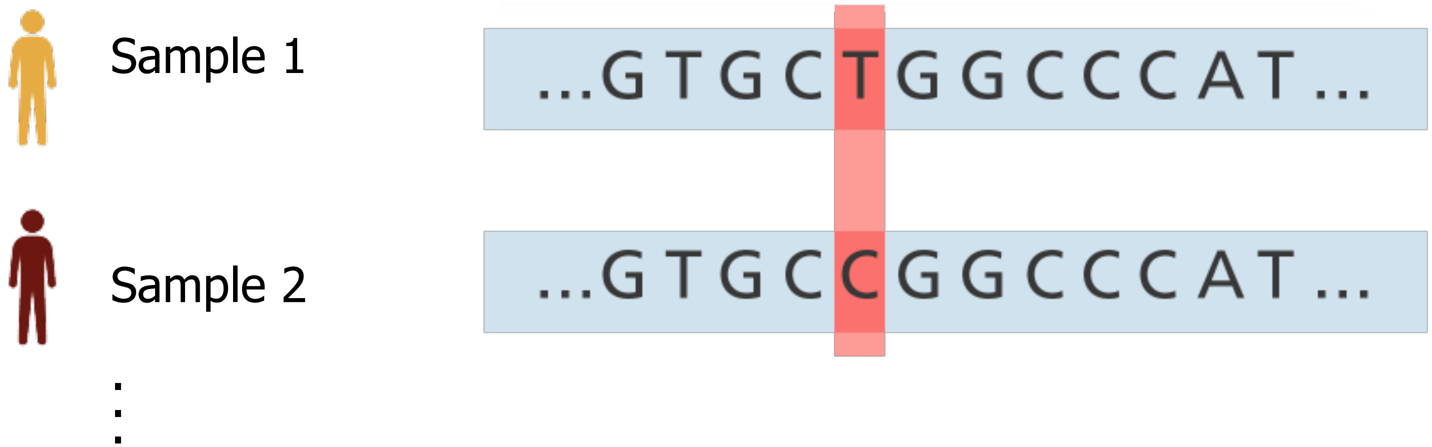
Personalized Medicine in UK

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

NHS
National Institute for Health Research

Population-Scale Genomics

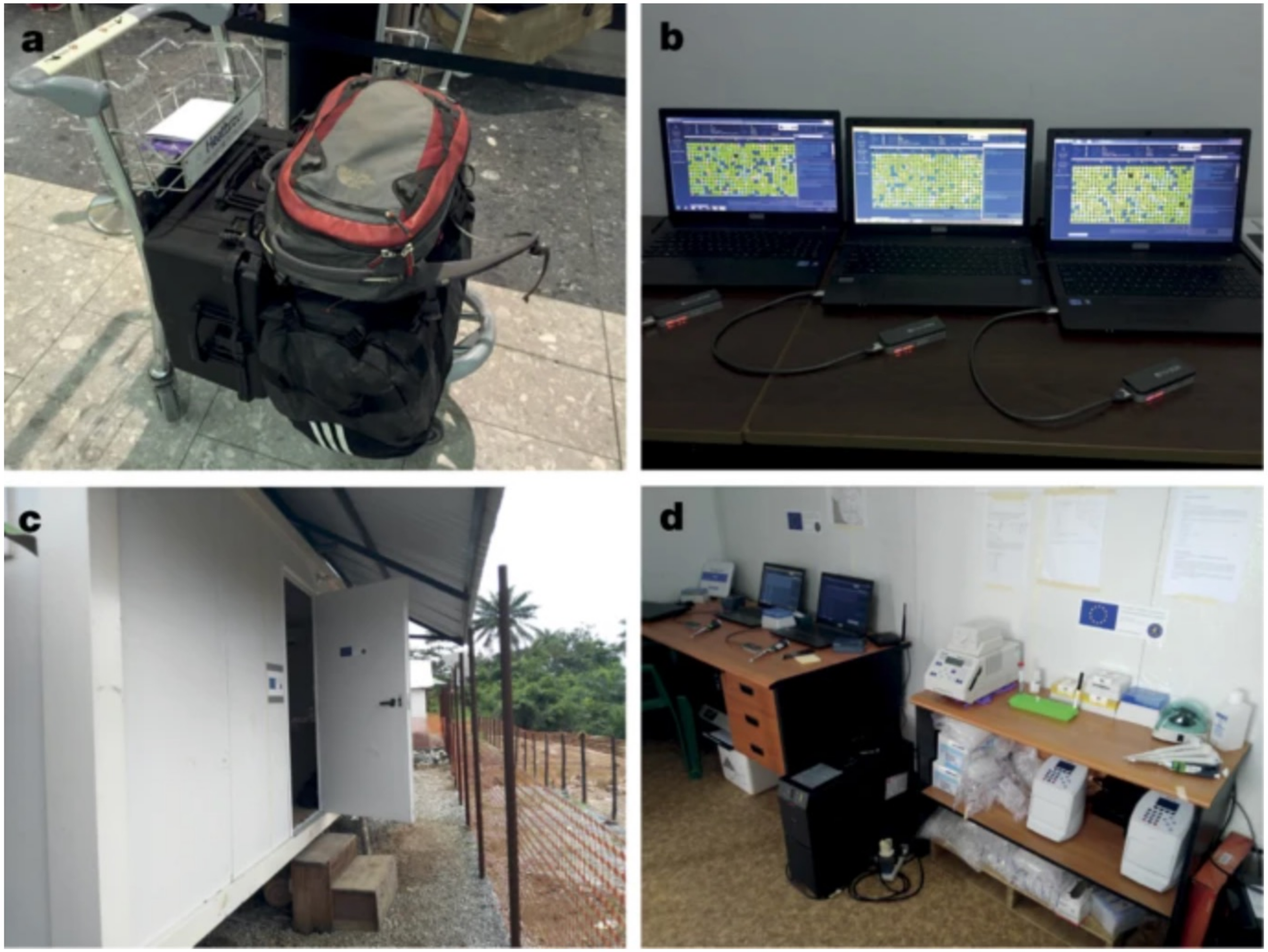
- Characterizing genomic variations of 49,962 Icelanders took **4.15 million CPU hours** or 83 CPU hours per sample on average


Sample 1
...G T G C T G G C C C A T ...
Sample 2
...G T G C C G G C C C A T ...
⋮

"GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs", Nature Communications, 2019

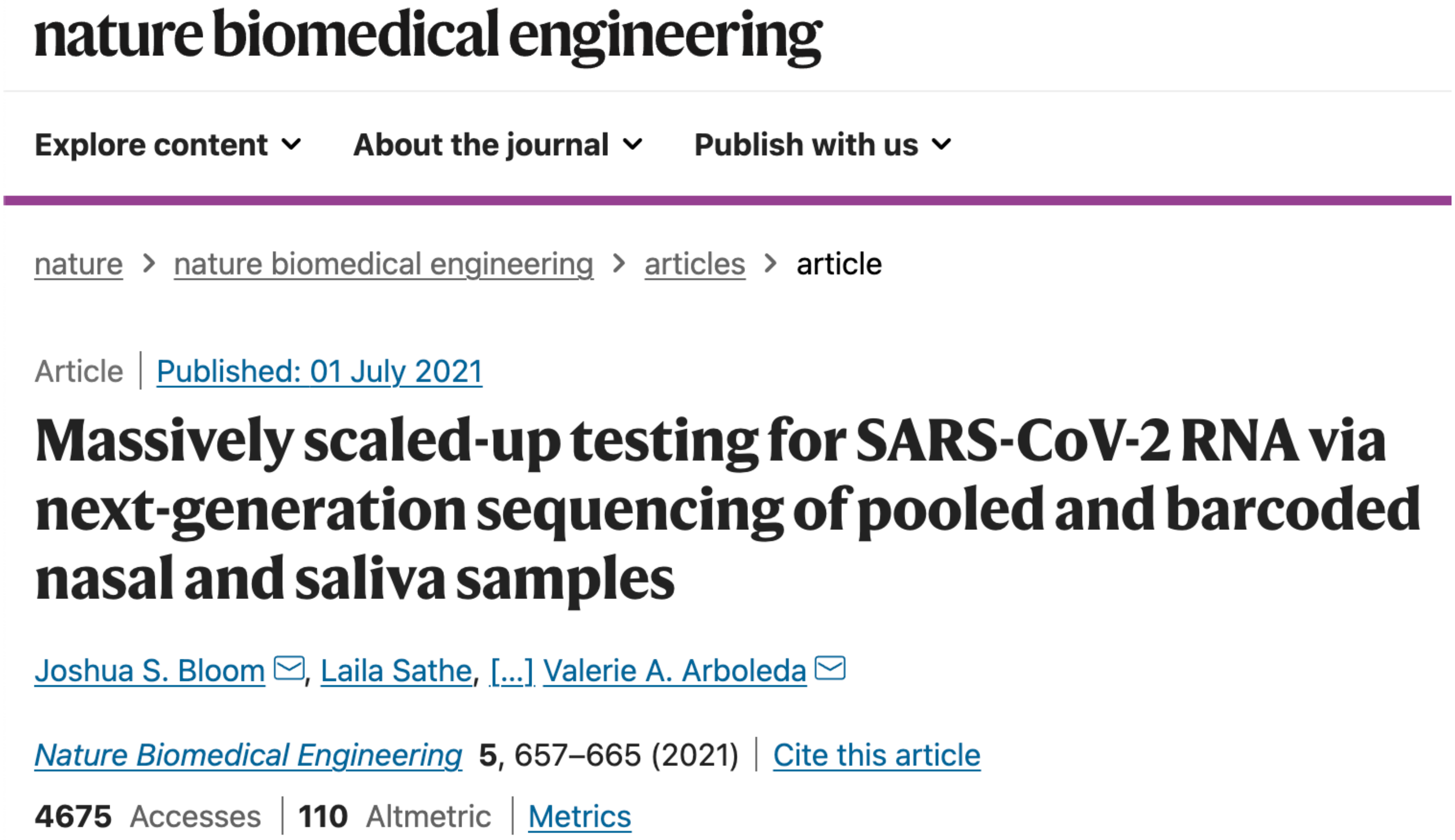
Rapid Surveillance of Disease Outbreaks?

Figure 1: Deployment of the portable genome surveillance system in Guinea.



Quick+, "Real-time, portable genome sequencing for Ebola surveillance", *Nature*, 2016

Scalable SARS-CoV-2 Testing



nature biomedical engineering

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > nature biomedical engineering > articles > article

Article | Published: 01 July 2021

Massively scaled-up testing for SARS-CoV-2 RNA via next-generation sequencing of pooled and barcoded nasal and saliva samples

Joshua S. Bloom ✉, Laila Sathe, [...] Valerie A. Arboleda ✉

Nature Biomedical Engineering **5**, 657–665 (2021) | Cite this article

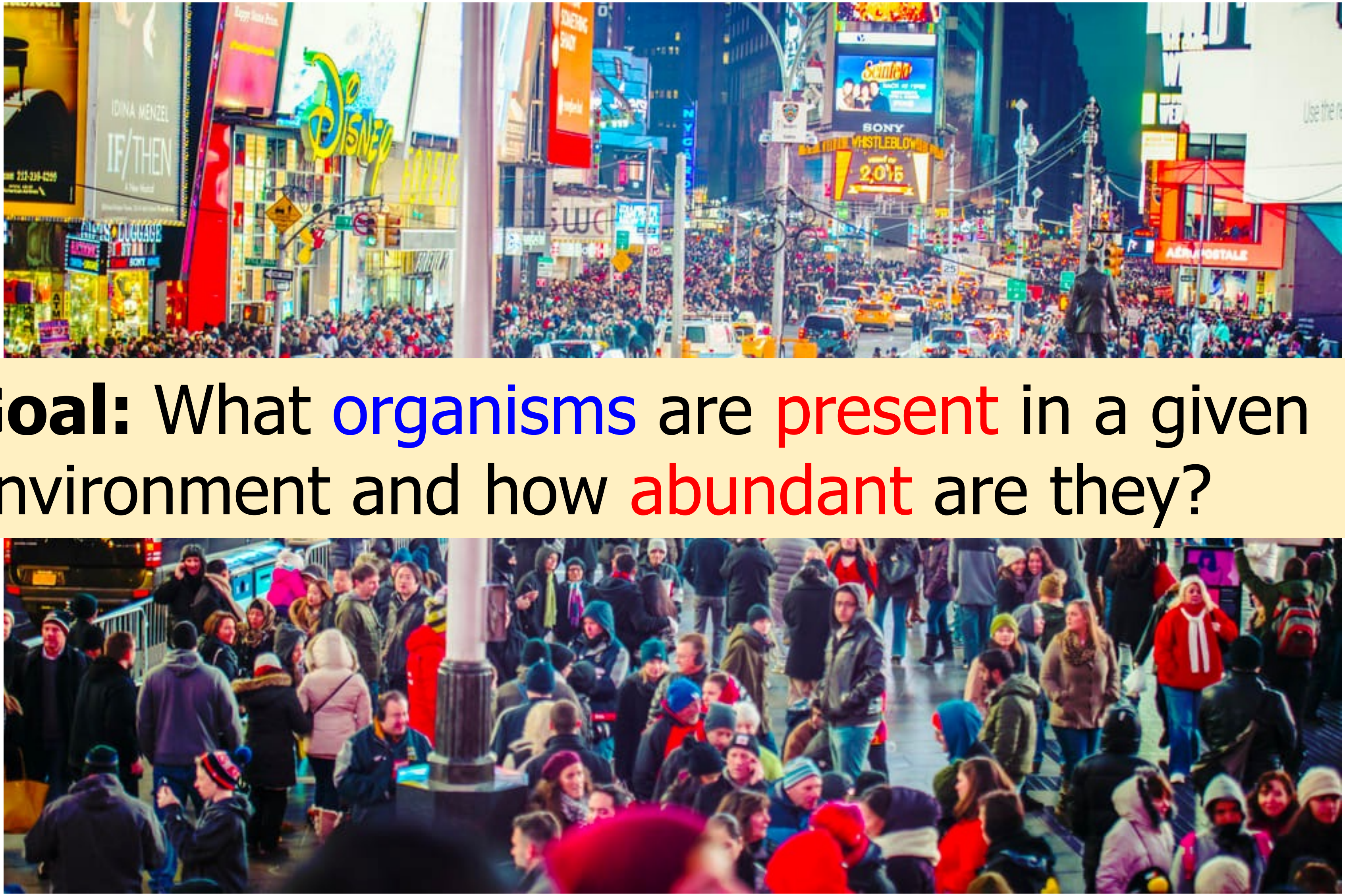
4675 Accesses | **110** Altmetric | Metrics

Bloom+, "Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing", *Nature Biomedical Engineering*, 2021

Population-Scale Microbiome Profiling

https://blog.wego.com/7-crowded-places-and-events-that-you-will-love/

Population-Scale Microbiome Profiling

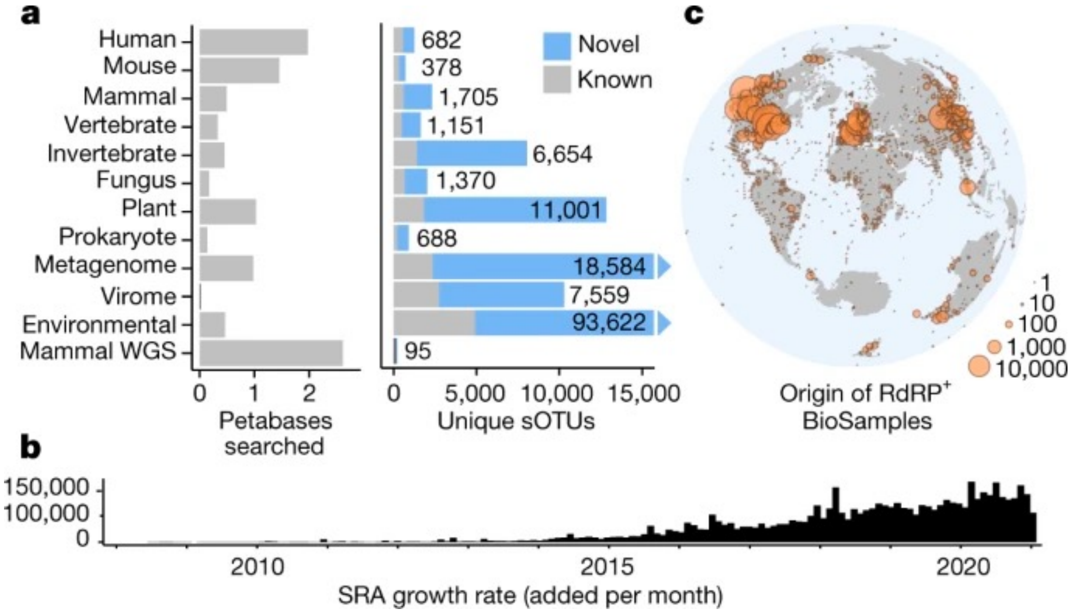


Goal: What organisms are present in a given environment and how abundant are they?

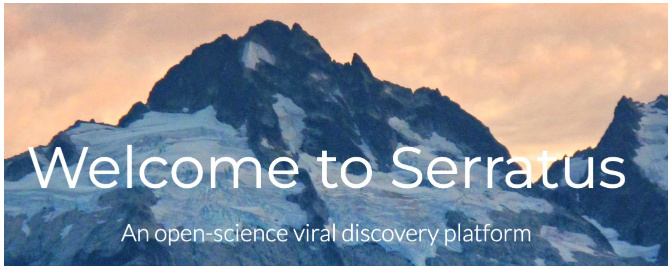
https://blog.wego.com/7-crowded-places-and-events-that-you-will-love/

Petabase-scale Viral Discovery

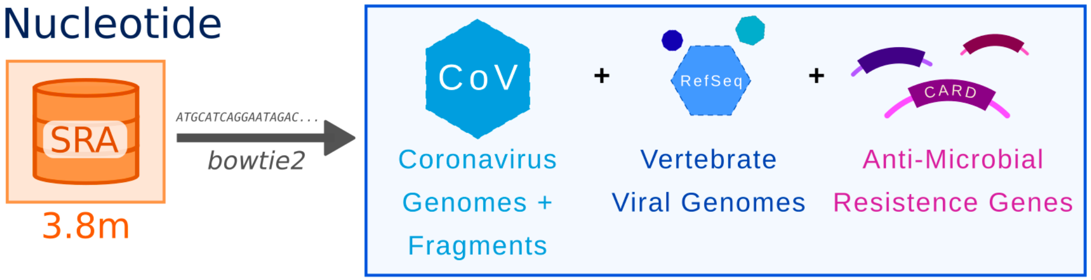
- Building and Profiling 3,500 genomic assemblies needs **28,000 virtual AWS CPUs**.



aws

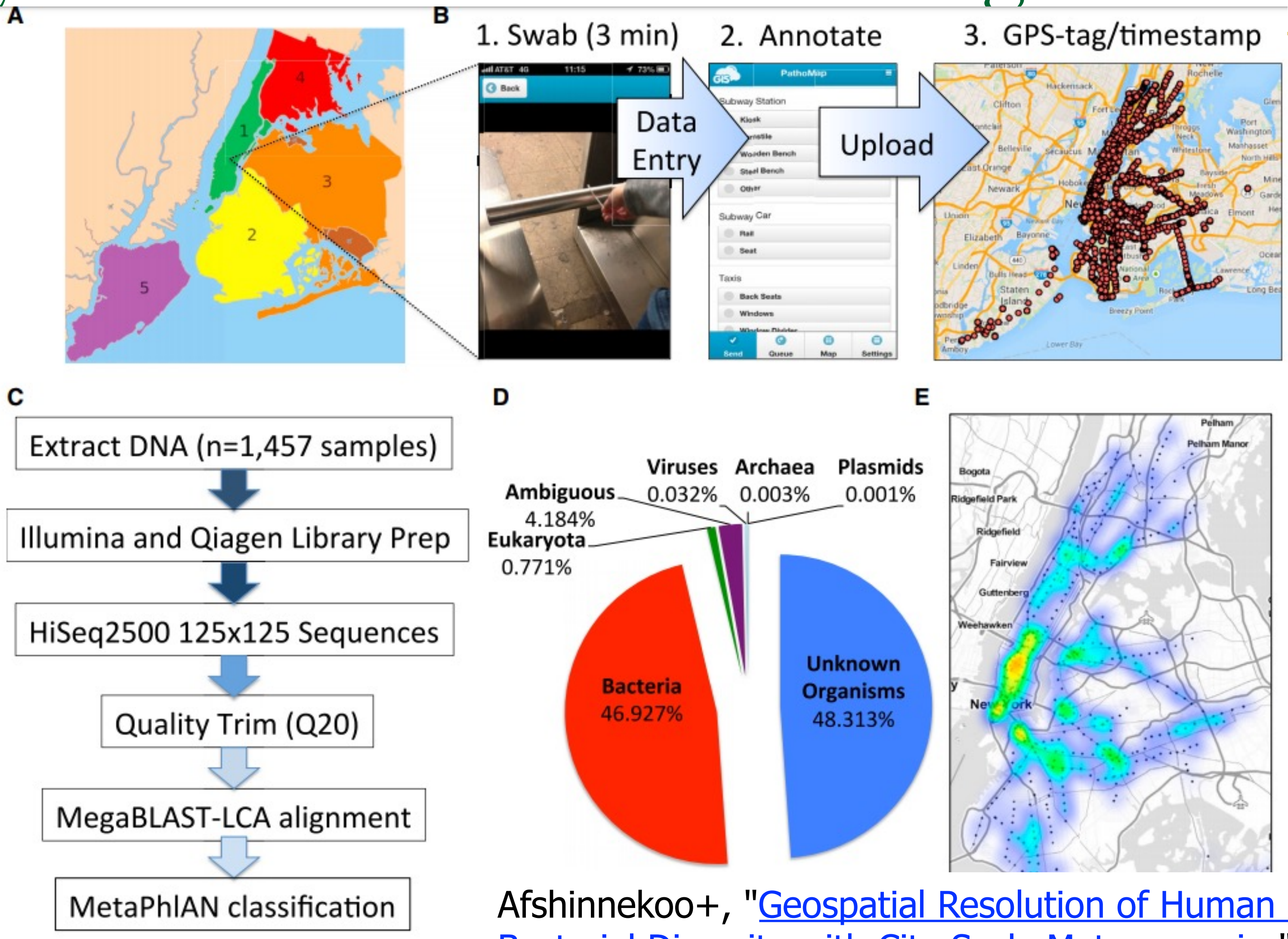


https://serratus.io/



Edgar+, "Petabase-scale sequence alignment catalyses viral discovery", Nature 2022

City-Scale Microbiome Profiling



Afshinnekoo+, "Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics", Cell Systems, 2015

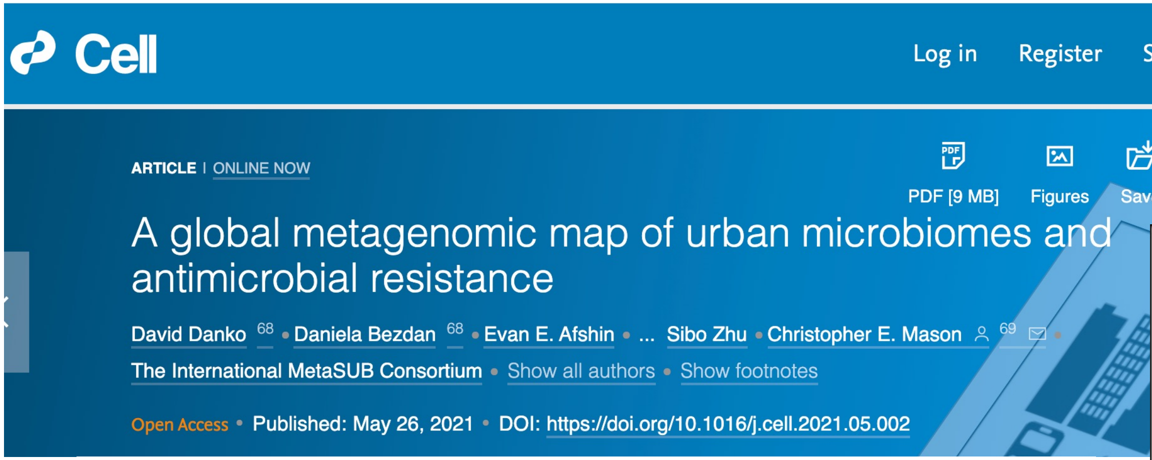
Figure 1. The Metagenome of New York City

(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

Population-Scale Microbiome Profiling

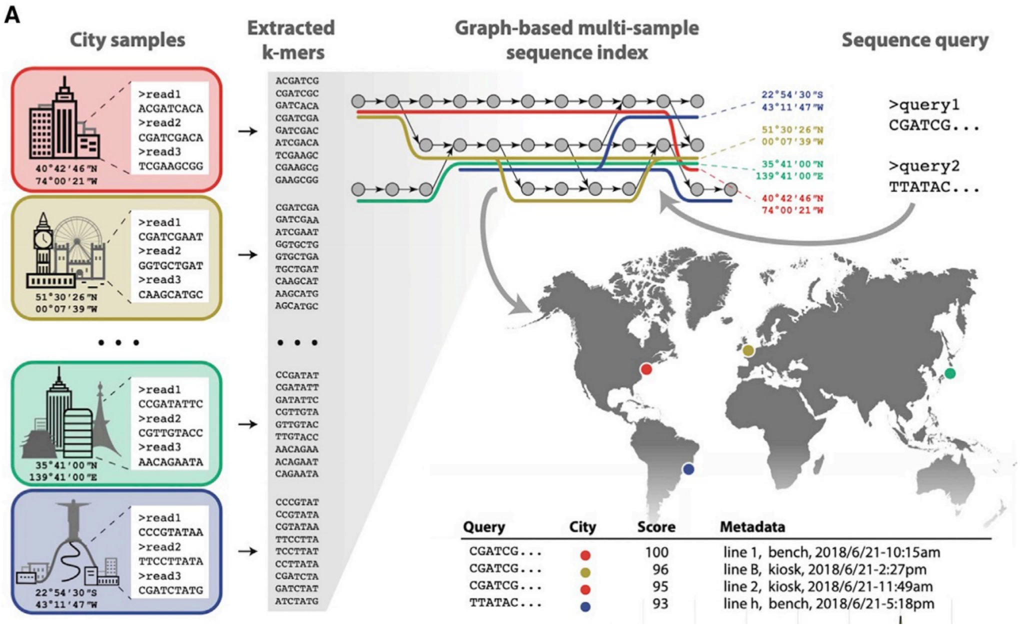


ARTICLE | ONLINE NOW

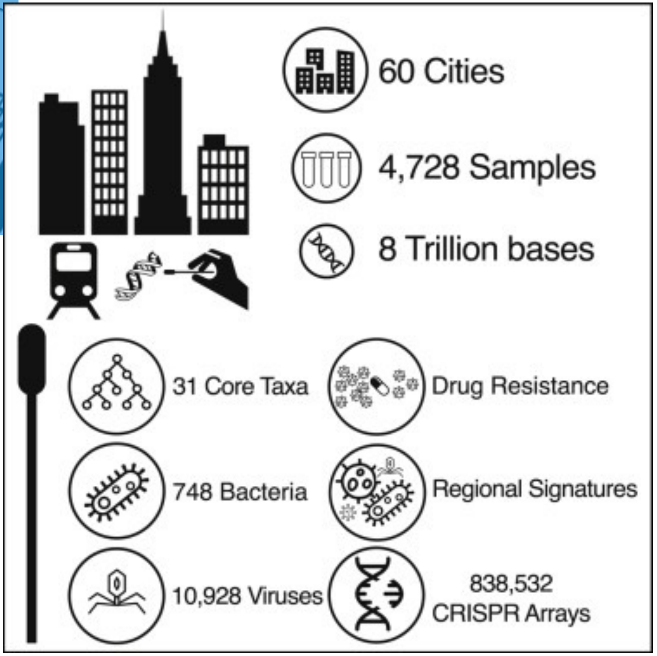
A global metagenomic map of urban microbiomes and antimicrobial resistance

David Danko [68] • Daniela Bezdan [68] • Evan E. Afshin • ... Sibo 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

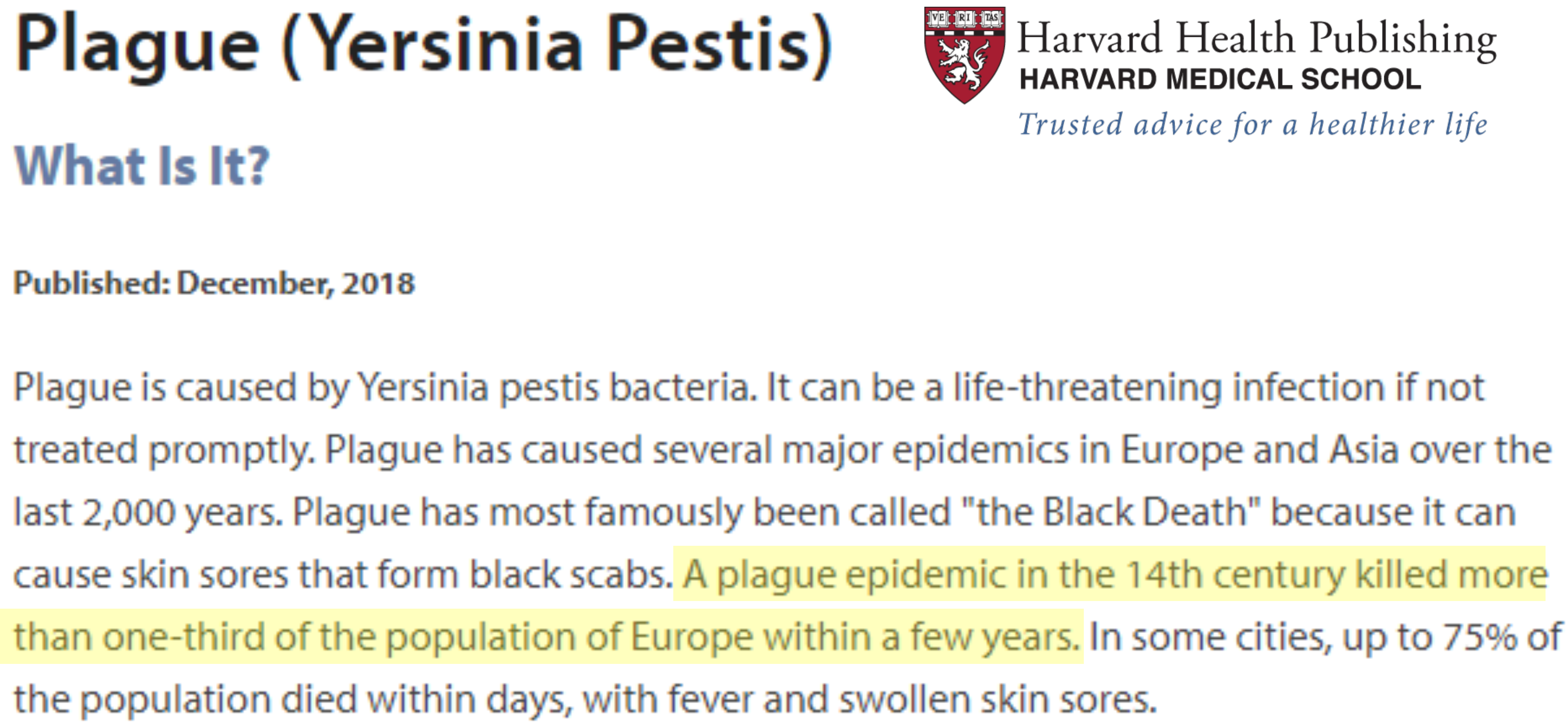


Query	City	Score	Metadata
CGATCG...	●	100	line 1, bench, 2018/6/21-10:15am
CGATCG...	●	96	line B, kiosk, 2018/6/21-2:27pm
CGATCG...	●	95	line 2, kiosk, 2018/6/21-11:49am
TTATAC...	●	93	line h, bench, 2018/6/21-5:18pm



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

Plague in New York Subway System?



Harvard Health Publishing
HARVARD MEDICAL SCHOOL
Trusted advice for a healthier life

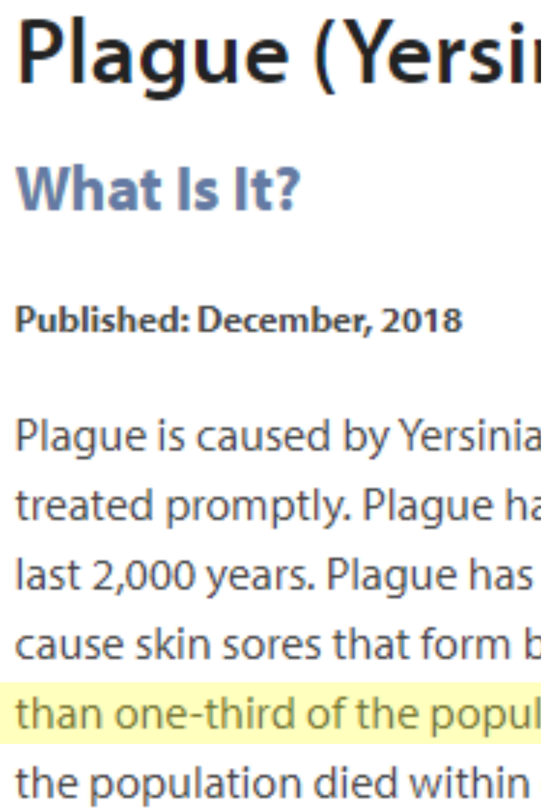
Plague (Yersinia Pestis)

What Is It?

Published: December, 2018

Plague is caused by Yersinia pestis bacteria. It can be a life-threatening infection if not treated promptly. Plague has caused several major epidemics in Europe and Asia over the last 2,000 years. Plague has most famously been called "the Black Death" because it can cause skin sores that form black scabs. A plague epidemic in the 14th century killed more than one-third of the population of Europe within a few years. In some cities, up to 75% of the population died within days, with fever and swollen skin sores.

Plague in New York Subway System?



Plague (Yersi

What Is It?

Published: December, 2018

Plague is caused by Yersinia

treated promptly. Plague ha he

last 2,000 years. Plague has ı

cause skin sores that form b re

than one-third of the popul s of

the population died within

The New York Times

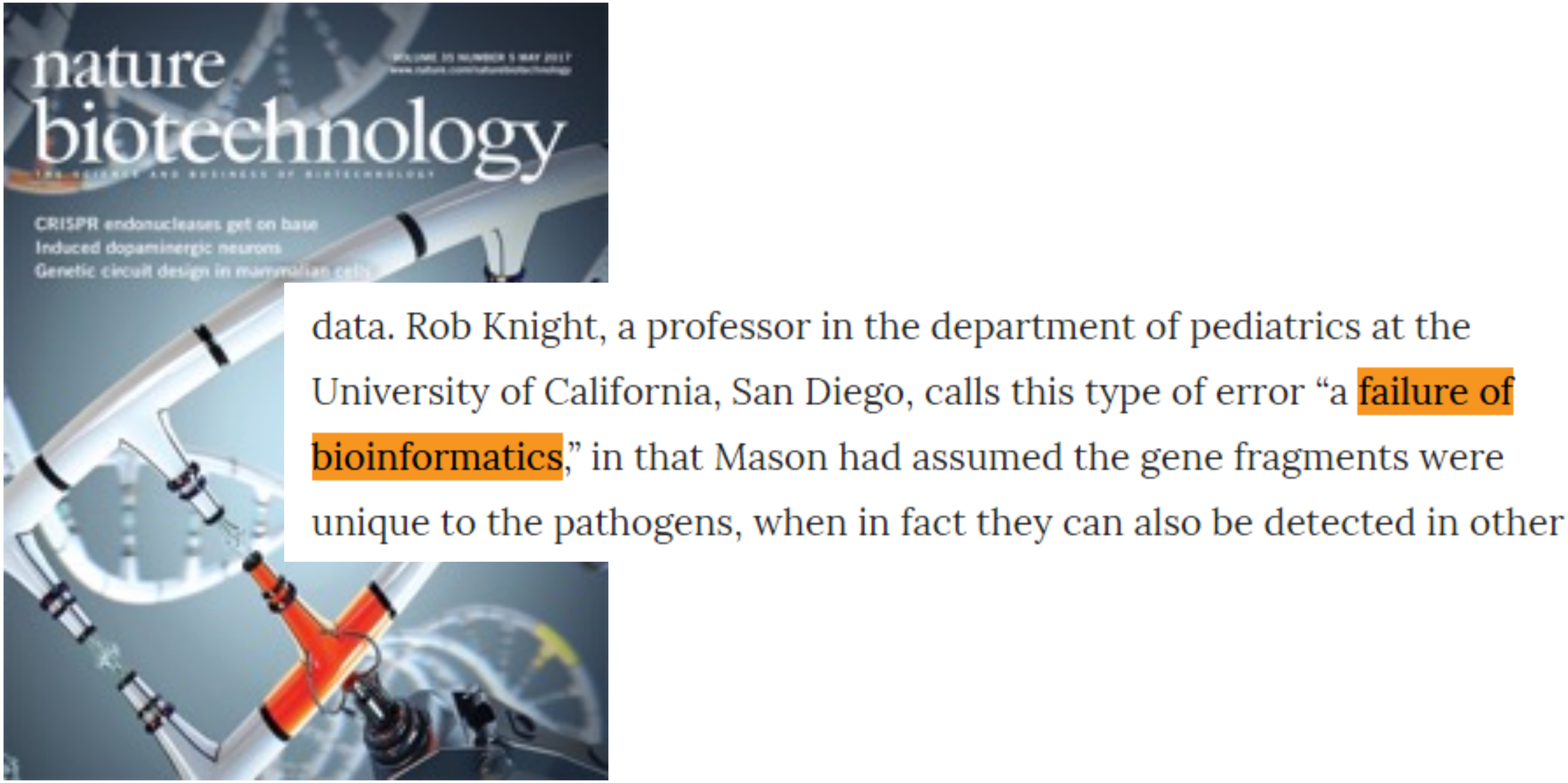
Bubonic Plague in the Subway System? Don't Worry About It

In October, riders were not deterred after reports that an Ebola-infected man had ridden the subway just before he fell ill. Robert Stolarik for The New York Times

https://www.nytimes.com/2015/02/07/nyregion/bubonic-plague-in-the-subway-system-dont-worry-about-it.html

The findings of Yersinia Pestis in the subway received wide coverage in the lay press, causing some alarm among New York residents

Failure of Bioinformatics


nature biotechnology
VOLUME 35 NUMBER 5 MAY 2017
www.nature.com/naturebiotechnology
THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY
CRISPR endonucleases get on base
Induced dopaminergic neurons
Genetic circuit design in mammalian cells

data. Rob Knight, a professor in the department of pediatrics at the University of California, San Diego, calls this type of error “a failure of bioinformatics,” in that Mason had assumed the gene fragments were unique to the pathogens, when in fact they can also be detected in other

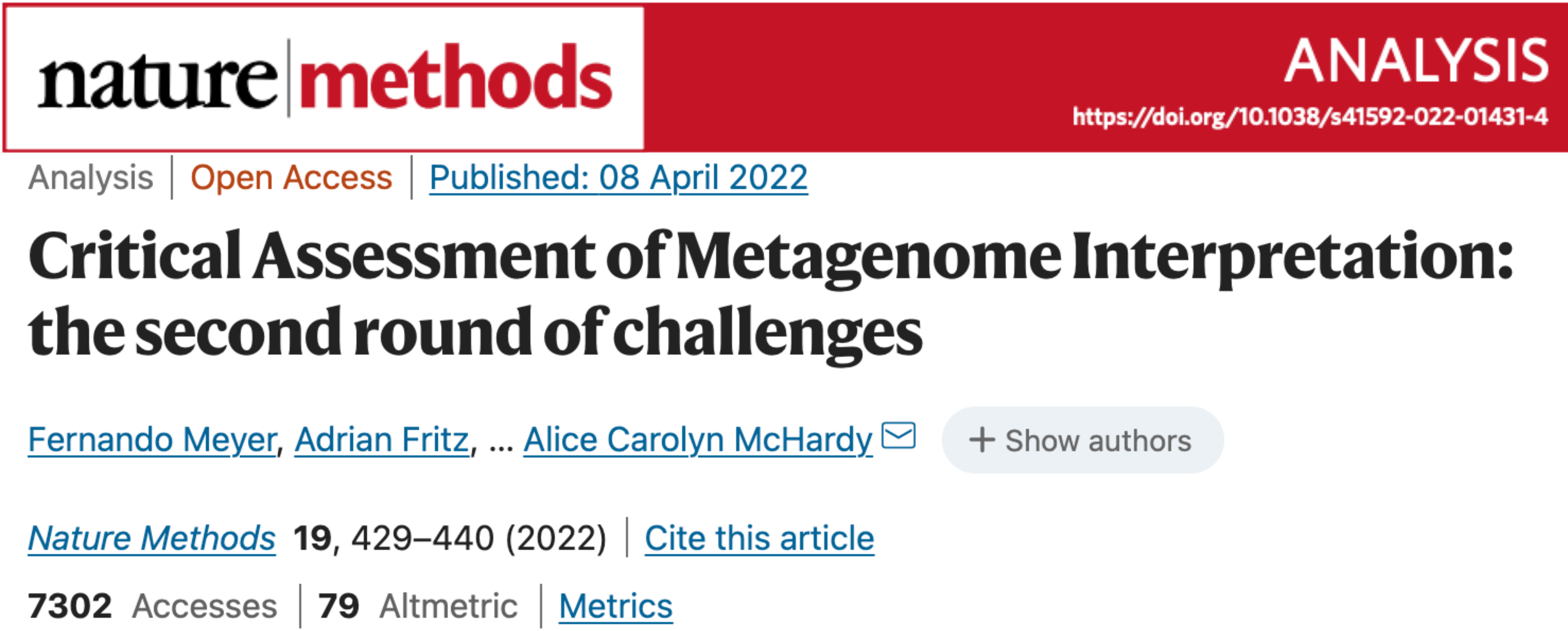
Living in a microbial world
Charles Schmidt
Nature Biotechnology, **volume 35**, pages401–403 (2017)
https://www.nature.com/articles/nbt.3868

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

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

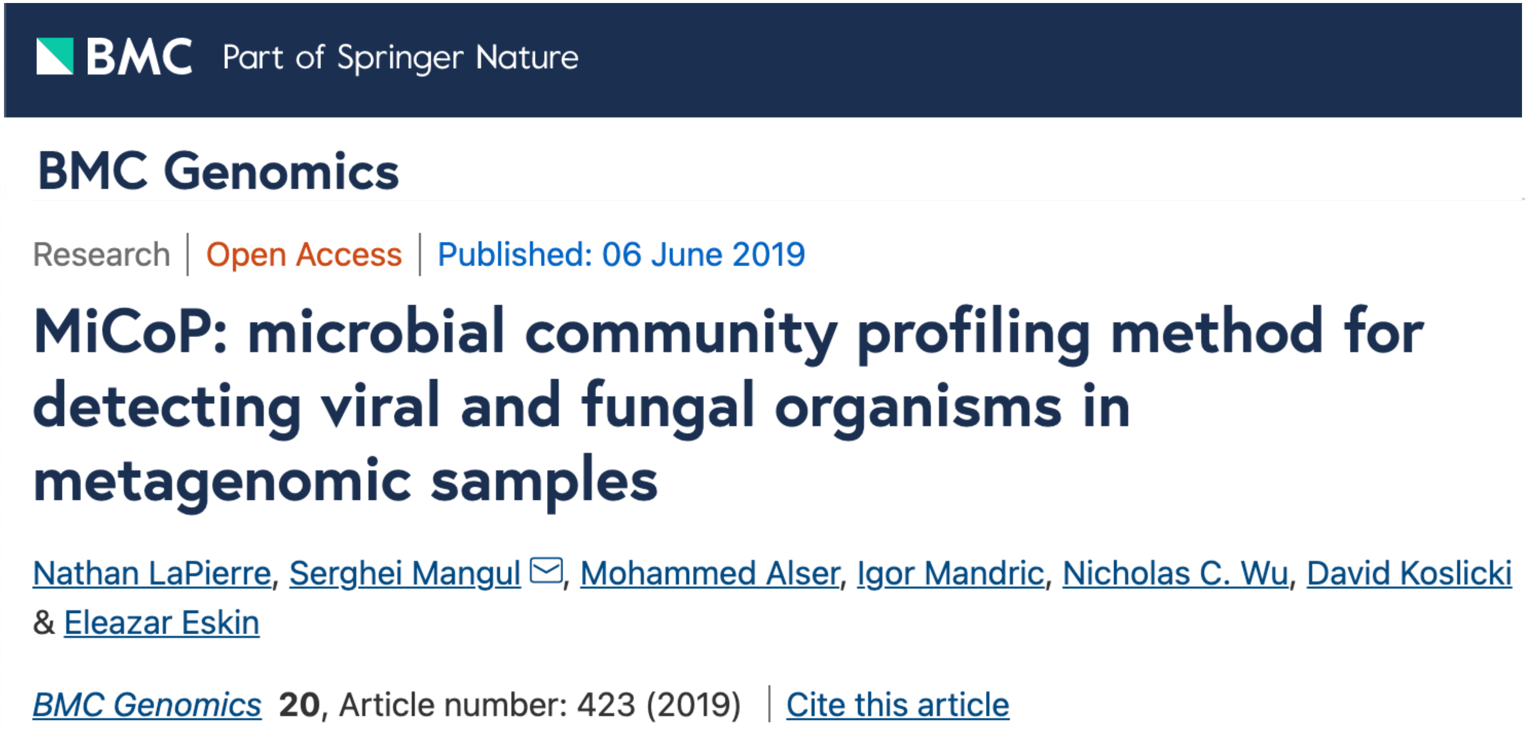
Metalign: efficient alignment-based metagenomic profiling via containment min hash

Nathan LaPierre ✉, Mohammed Alser, Eleazar Eskin, David Koslicki ✉ & Serghei Mangul ✉

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

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

How About Reliability?

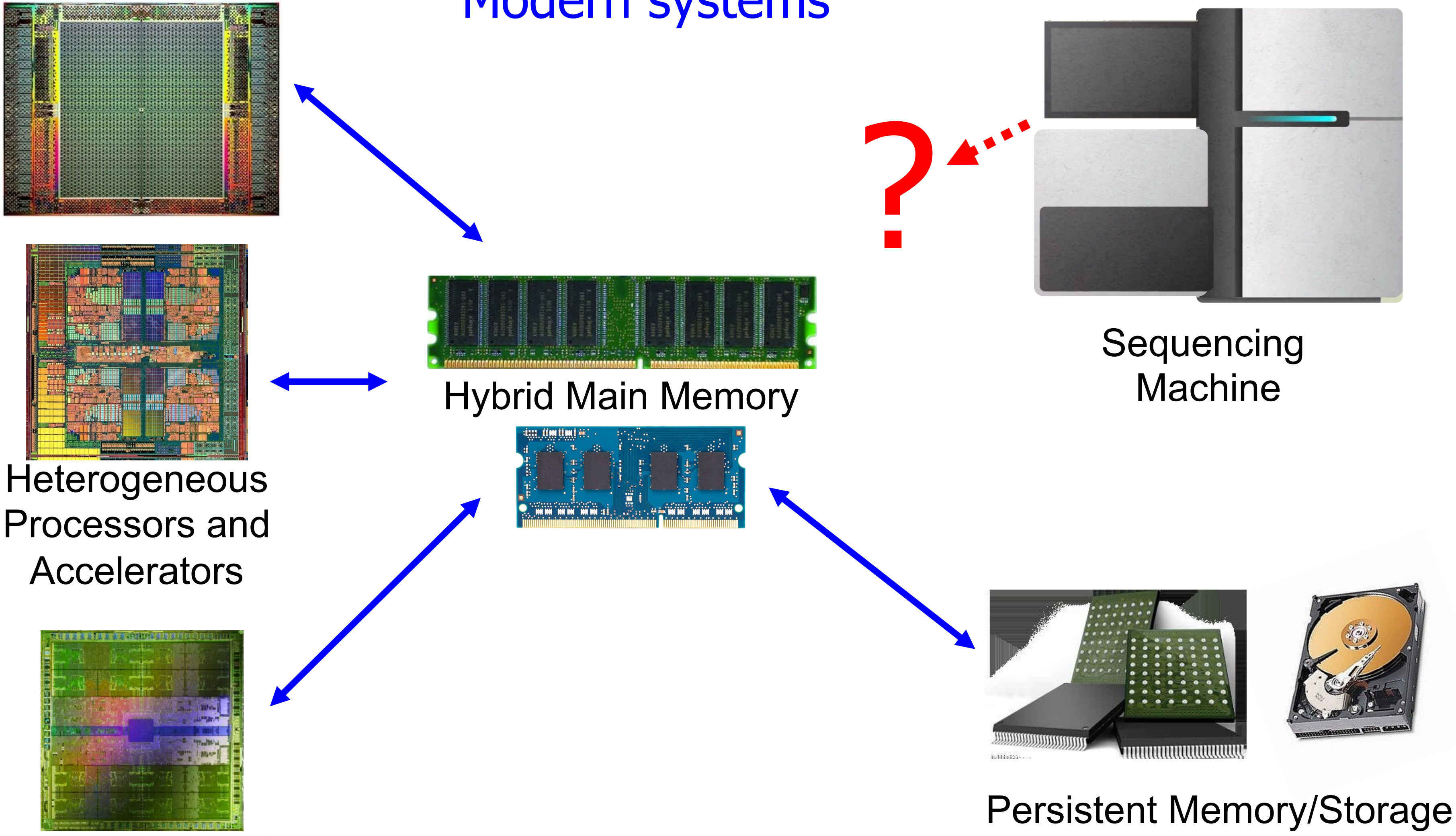
https://www.bbc.com/future/article/20221011-how-space-weather-causes-computer-errors

Challenging Environment in Outer Space

https://spaceref.com/space-stations/nasa-space-station-on-orbit-status-6-august-2020-working-in-the-kibo-laboratory/

Intelligent Architecture?

Modern systems



FPGAs

Heterogeneous Processors and Accelerators

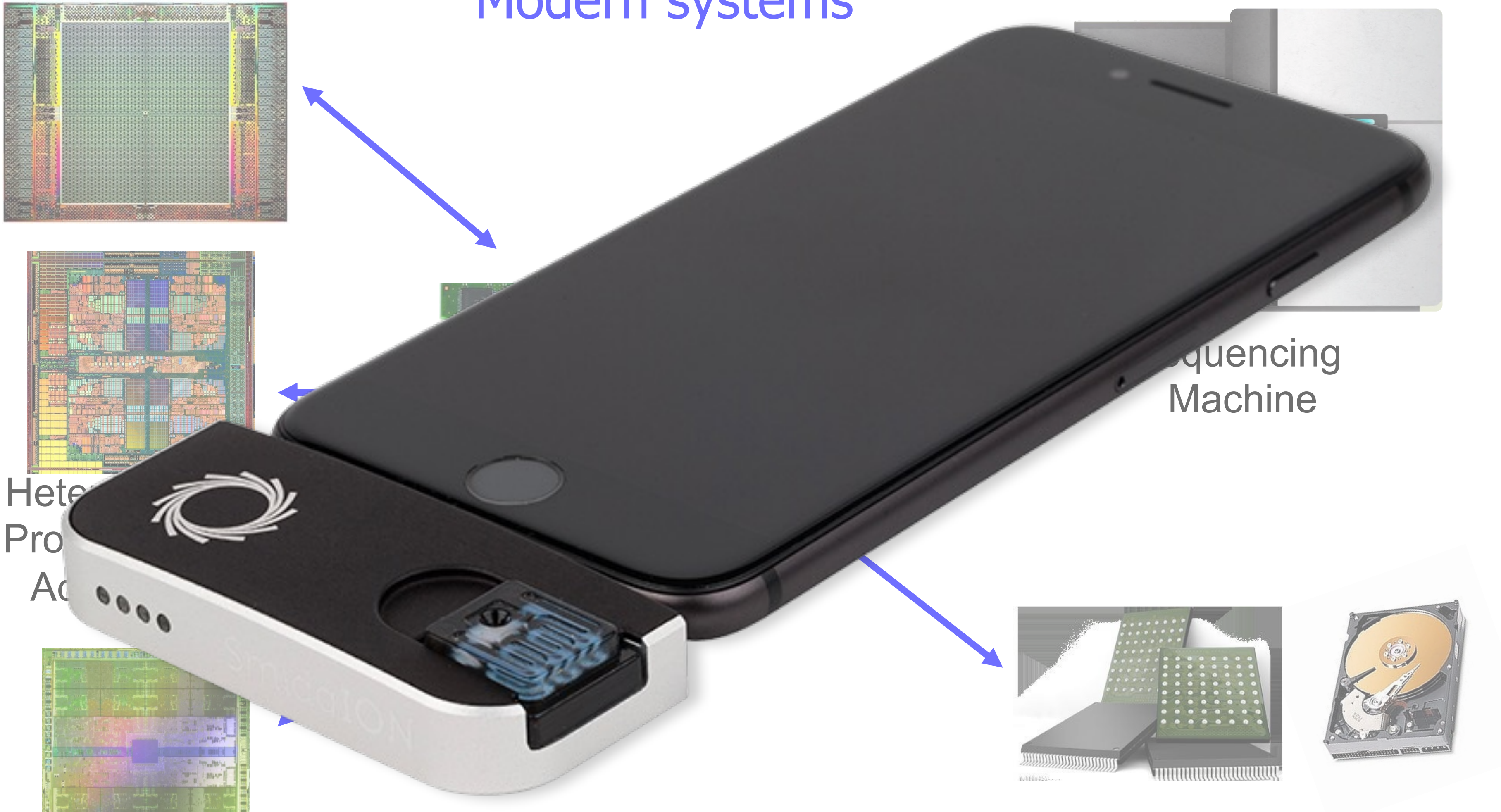
(General Purpose) GPUs

Hybrid Main Memory

Sequencing Machine

Persistent Memory/Storage

Intelligent Architecture?



FPGAs

Modern systems

quencing
Machine

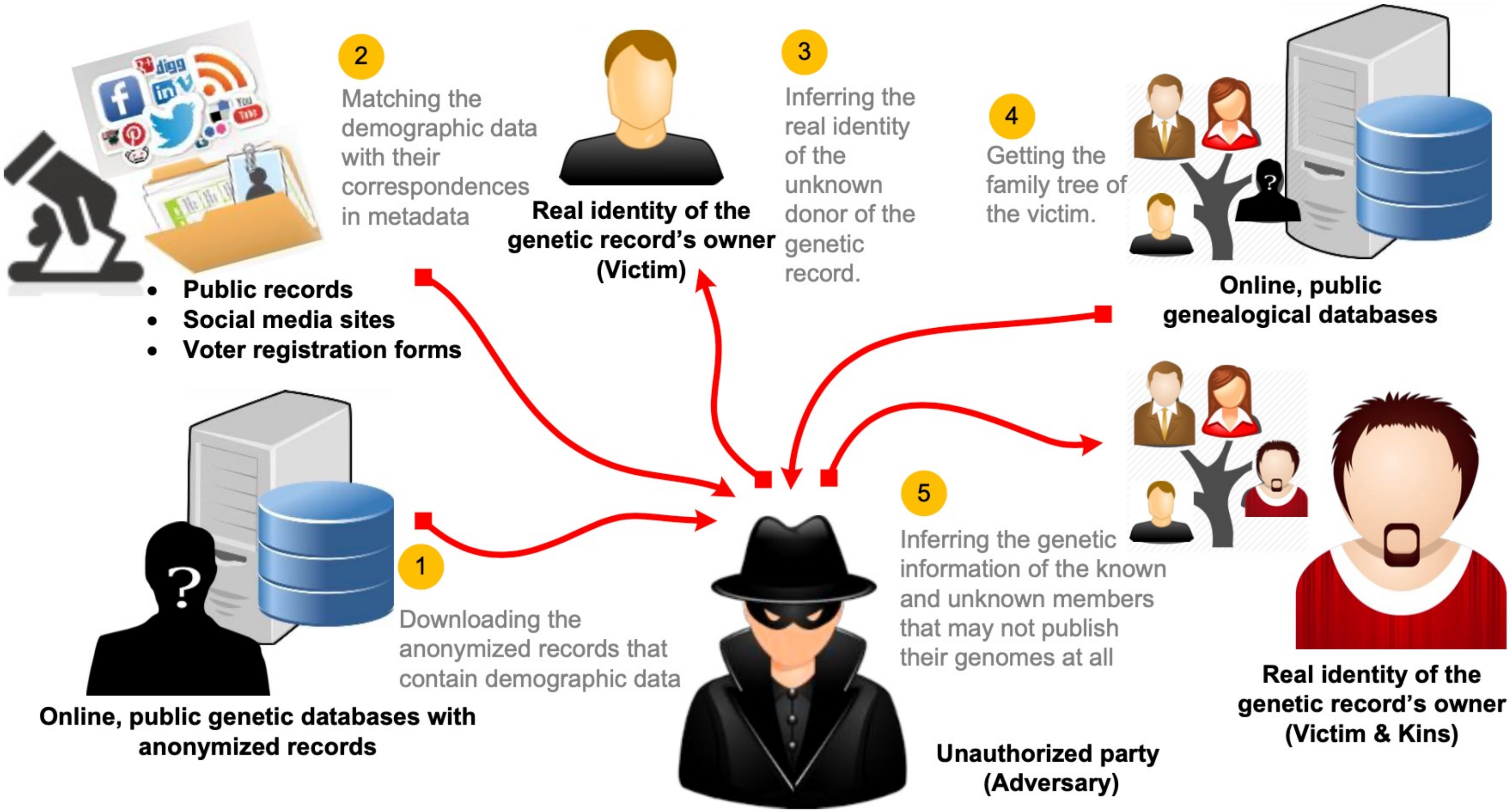
Hete
Pro
Ac

Persistent Memory/Storage

(General Purpose) GPUs

https://nanoporetech.com/products/smidgion

Privacy-Preserving Genome Analysis?

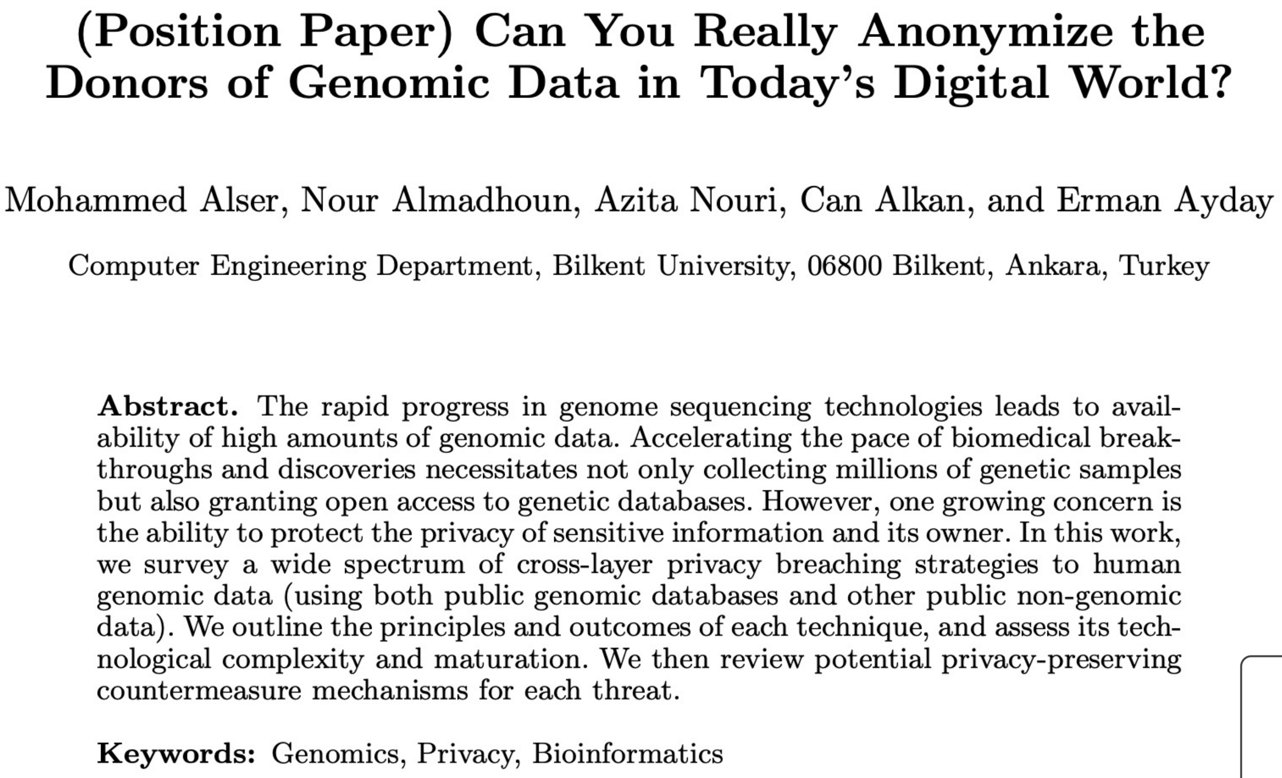


2
Matching the demographic data with their correspondences in metadata
Real identity of the genetic record's owner (Victim)
3
Inferring the real identity of the unknown donor of the genetic record.
4
Getting the family tree of the victim.
Public records
Social media sites
Voter registration forms
Online, public genealogical databases
1
Downloading the anonymized records that contain demographic data
Online, public genetic databases with anonymized records
5
Inferring the genetic information of the known and unknown members that may not publish their genomes at all
Unauthorized party (Adversary)
Real identity of the genetic record's owner (Victim & Kins)

Fig. 5. A completion attack.

Alser+, "**Can you really anonymize the donors of genomic data in today's digital world?**" *10th International Workshop on Data Privacy Management (DPM)*, 2015.

Can you Really Anonymize the Donors?



(Position Paper) Can You Really Anonymize the Donors of Genomic Data in Today's Digital World?

Mohammed Alser, Nour Almadhoun, Azita Nouri, Can Alkan, and Erman Ayday

Computer Engineering Department, Bilkent University, 06800 Bilkent, Ankara, Turkey

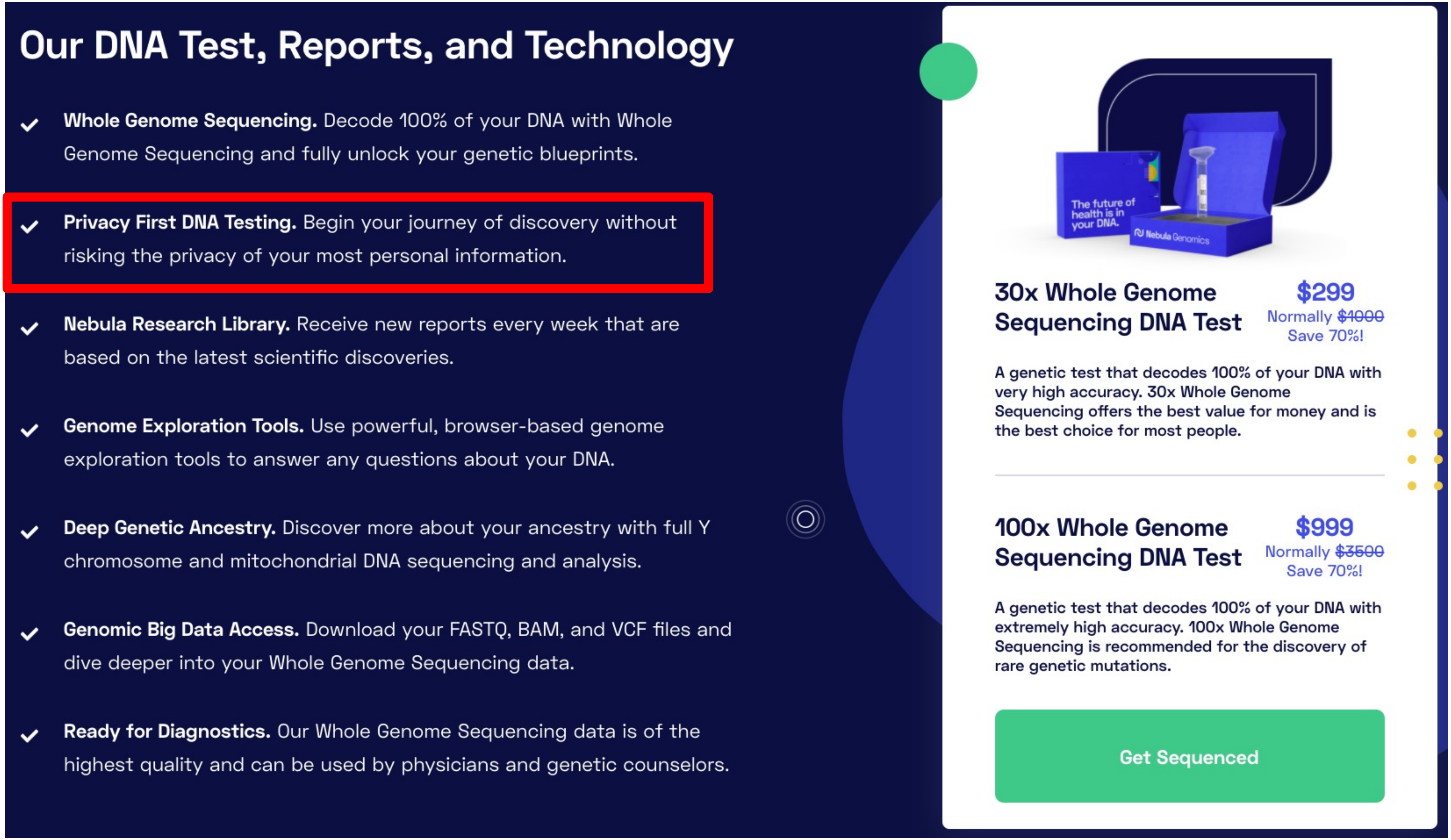
Abstract. The rapid progress in genome sequencing technologies leads to availability of high amounts of genomic data. Accelerating the pace of biomedical breakthroughs and discoveries necessitates not only collecting millions of genetic samples but also granting open access to genetic databases. However, one growing concern is the ability to protect the privacy of sensitive information and its owner. In this work, we survey a wide spectrum of cross-layer privacy breaching strategies to human genomic data (using both public genomic databases and other public non-genomic data). We outline the principles and outcomes of each technique, and assess its technological complexity and maturation. We then review potential privacy-preserving countermeasure mechanisms for each threat.

Keywords: Genomics, Privacy, Bioinformatics

DPM 2015
Vienna, Austria
September 21-22, 2015

Alser+, "Can you really anonymize the donors of genomic data in today's digital world?" *10th International Workshop on Data Privacy Management (DPM)*, 2015.

Privacy-Preserving DNA Test



Our DNA Test, Reports, and Technology

- ✓ **Whole Genome Sequencing.** Decode 100% of your DNA with Whole Genome Sequencing and fully unlock your genetic blueprints.
- ✓ **Privacy First DNA Testing.** Begin your journey of discovery without risking the privacy of your most personal information.
- ✓ **Nebula Research Library.** Receive new reports every week that are based on the latest scientific discoveries.
- ✓ **Genome Exploration Tools.** Use powerful, browser-based genome exploration tools to answer any questions about your DNA.
- ✓ **Deep Genetic Ancestry.** Discover more about your ancestry with full Y chromosome and mitochondrial DNA sequencing and analysis.
- ✓ **Genomic Big Data Access.** Download your FASTQ, BAM, and VCF files and dive deeper into your Whole Genome Sequencing data.
- ✓ **Ready for Diagnostics.** Our Whole Genome Sequencing data is of the highest quality and can be used by physicians and genetic counselors.

30x Whole Genome Sequencing DNA Test — **$299**
Normally ~~$4000~~
Save 70%!

A genetic test that decodes 100% of your DNA with very high accuracy. 30x Whole Genome Sequencing offers the best value for money and is the best choice for most people.

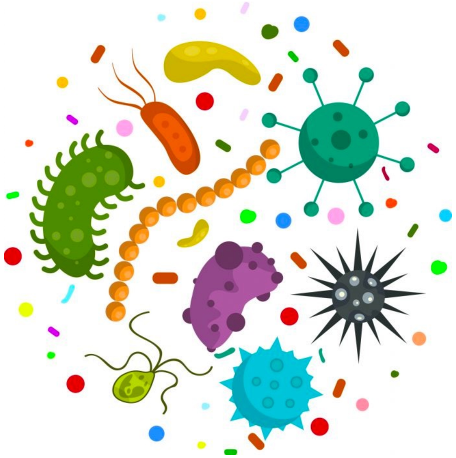
100x Whole Genome Sequencing DNA Test — **$999**
Normally ~~$3500~~
Save 70%!

A genetic test that decodes 100% of your DNA with extremely high accuracy. 100x Whole Genome Sequencing is recommended for the discovery of rare genetic mutations.

https://nebula.org/whole-genome-sequencing/

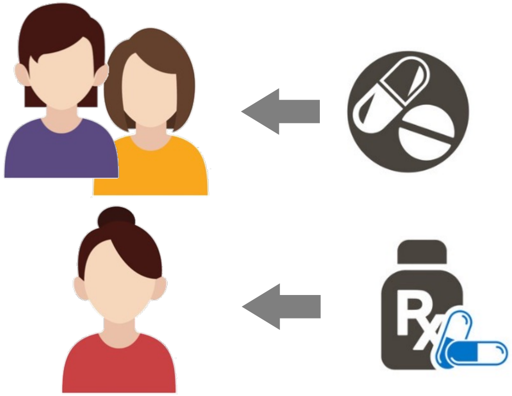
We Need Faster & Scalable Genome Analysis

Understanding **genetic variations**



Predicting the presence and relative abundances of **microbes** in a sample

Rapid surveillance of **disease outbreaks**



Developing **personalized medicine**

And many other applications …

Applications are only limited by our imagination

Fundamentally New Storage Architectures

215,000 terabytes of data stored
in a single gram of DNA

"A DNA-of-things storage architecture to create materials with embedded memory", *Nature Biotechnology,* 2020

New Personalized Shopping Paradigm

https://www.dnanudge.com/

Achieving Intelligent Genome Analysis?

How and where to enable

fast, accurate, cheap,

privacy-preserving, and exabyte scale

analysis of genomic data?

Agenda for Today

- What is Genome Analysis?
- What is Intelligent Genome Analysis?

- **How we Analyze Genome?**
- What are the Barriers to Enabling Intelligent Analyses?

- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

- Where is Genomic Analyses Going Next?

How to Analyze a Genome?

NO

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

>CCTCCTCAGTGCCACCCAGCCCACTGGCAGCTCCCAAACAGGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTGGAGTGAGGTGTCAAG
GACCTAAACTAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTT
CATGTCAAGGACCTAATGTGCTAAACAGCACTTTTTTGACCATTATTTTGGATCTGAAAGAAATCAAGAATAAATGAAGGACTTGATACATTG
GAAGAGGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGA**A**TTTAAAATTTAAGTAATTCTTTGAAAAAA
ACTAATTTCTAAGCTTCTT**C**ATGTCAAGGACCTAATGTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAAGAAAAAGAATTTAAAATTT
AAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAGGCCAAGAGTTGCAAAAAAAAAAAAAAGAAAAA
GAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTAGCCAGAATGG
TTGTGGGATGGGAGCCTCTGTGGACCGACCAGGTAGCTCTCTTTTCCACACTGTAGTCTCAAAGCTTCTTCATGTGGTTTCTCTGAGTGAAA
AAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTT**T**TTCATGTCAAGGACC
TAATGTAGCTATACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCACTTCCCAGAAAGCTTCTTCA……

How to Analyze a Genome?

NO

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

Why?!

>CCT CAAG

GACC TCTT

CATGT CATTG

GAAG AAAA

ACTA ATTT

AAGT AAAA

GAAA ATGG

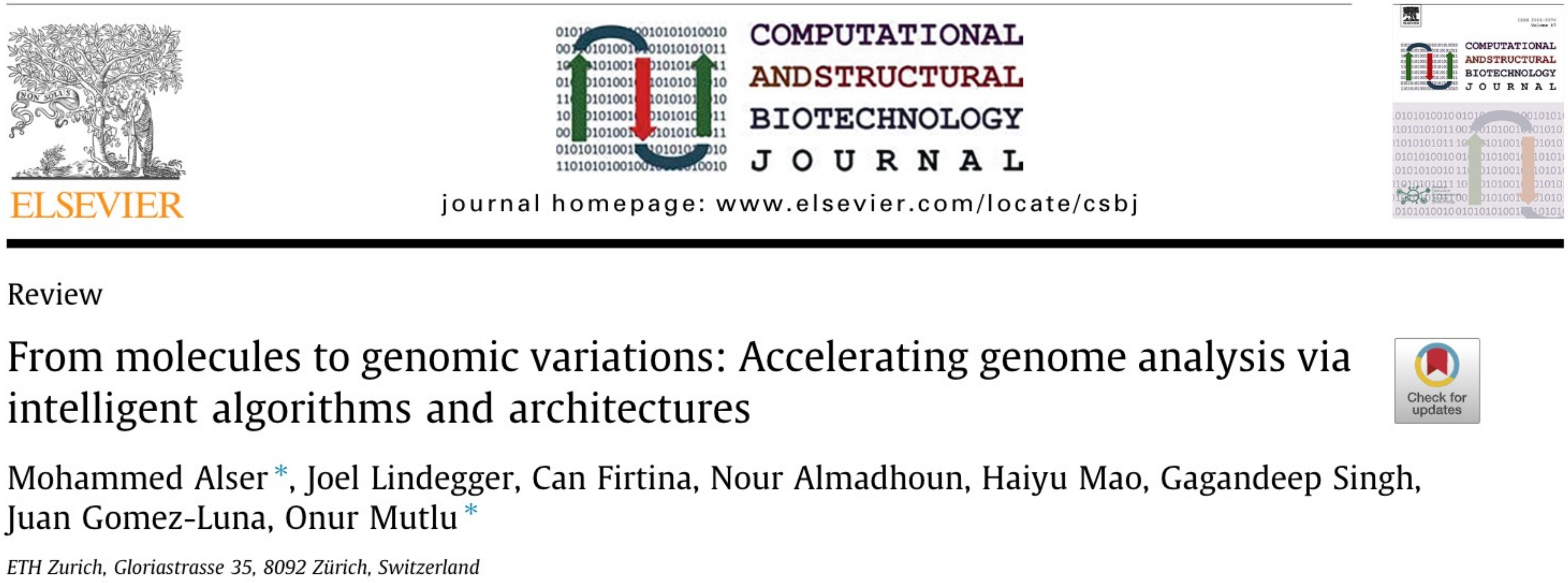
TTGTC GAAA

AAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTTTTTCATGTCAAGGACC

TAATGTAGCTATACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCACTTCCCAGAAAGCTTCTTCA……

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

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Review

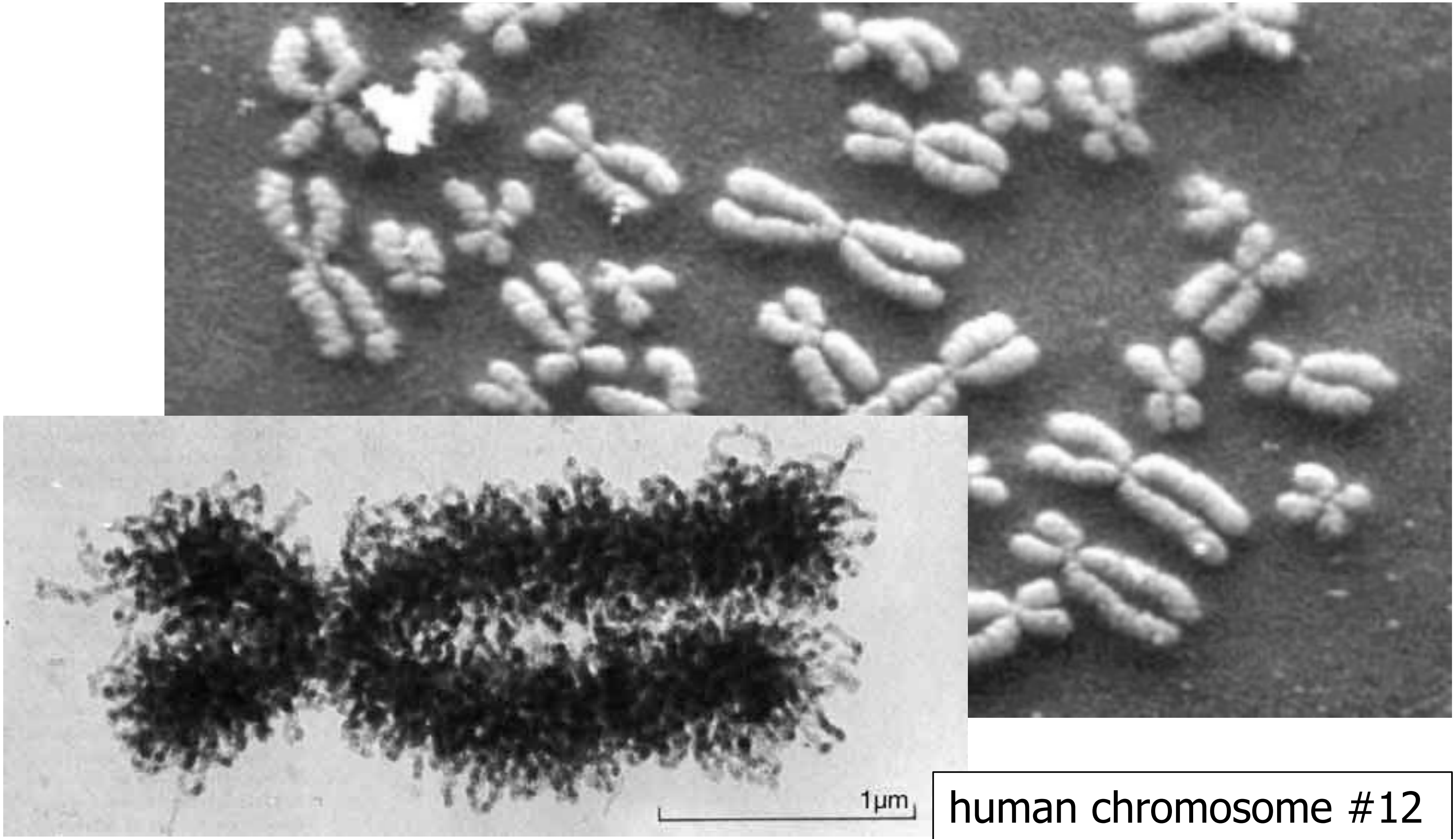
From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures

Check for updates

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

ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland

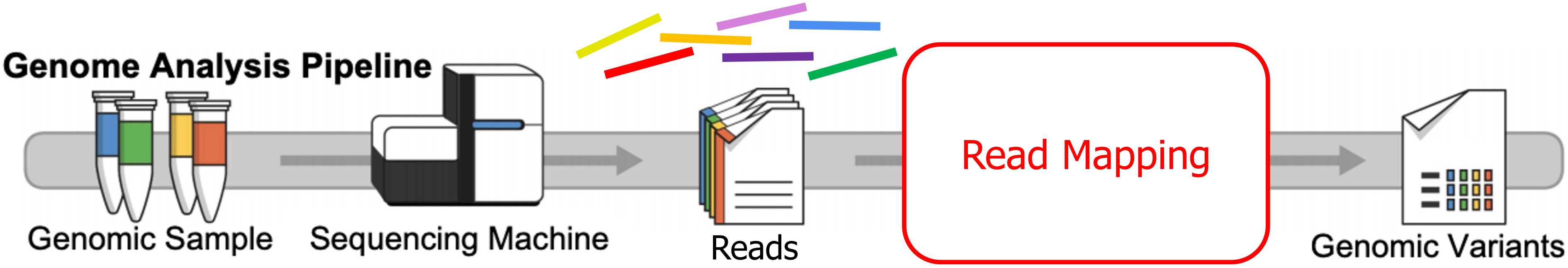
DNA Under Electron Microscope



human chromosome #12 from HeLa’s cell

Untangling Yarn Balls & DNA Sequencing

Genome Sequencer is a Chopper



CCCCCCTATATATACGTACTAGTACGT

ACGACTTTAGTACGTACGT

TATATATACGTACTAGTACGT

ACGTACG CCCCTACGTA

TATATATACGTACTAGTACGT

ACGACTTTAGTACGTACGT

TATATATACGTACTAAAGTACGT

TATATATACGTACTAGTACGT

ACG TTTTTAAAACGTA

TATATATACGTACTAGTACGT

ACGAC GGGGAGTACGTACGT

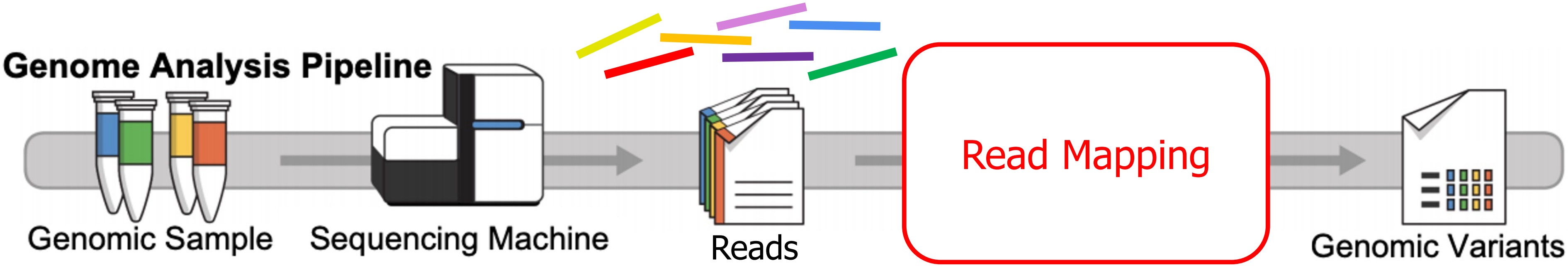
1×10^{12} bases*

44 hours*

<1000 $

* NovaSeq 6000

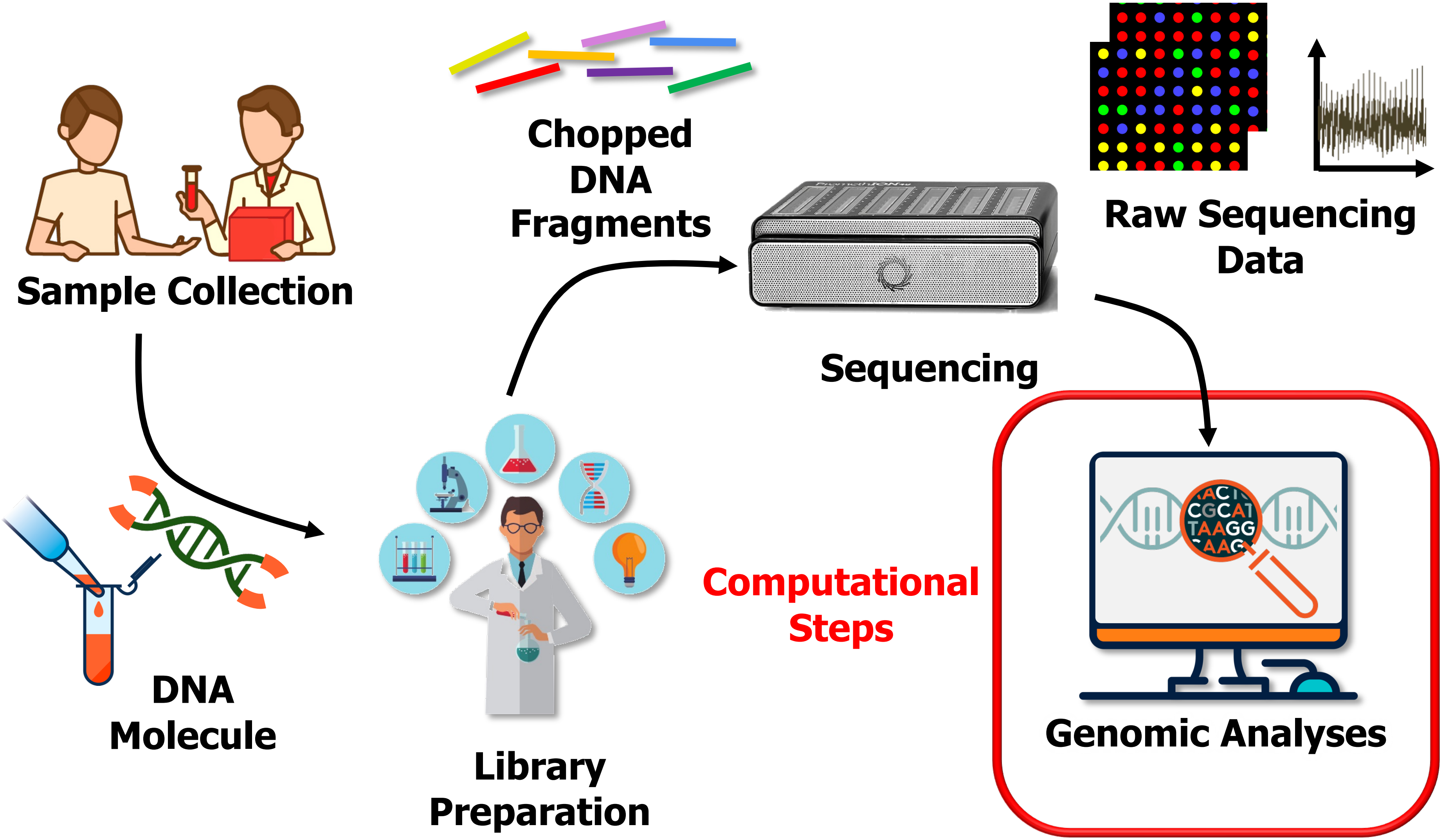
Genome Sequencer is a Chopper



Current sequencing machine provides **small randomized fragments** of the original DNA sequence

Alser+, "Technology dictates algorithms: Recent developments in read alignment", Genome Biology, 2021

Genome Analysis in Real Life



Sample Collection

DNA Molecule

Library Preparation

Chopped DNA Fragments

Sequencing

Raw Sequencing Data

Computational Steps

Genomic Analyses

Sequencing Technologies

Oxford Nanopore (ONT)

PacBio (HiFi, CLR)

Illumina

… and more! All produce data with different properties.

Oxford Nanopore Sequencers

MinION Mk1B

MinION Mk1C

GridION Mk1

PromethION 24/48

	MinION Mk1B	MinION Mk1C	GridION Mk1	PromethION 24	PromethION 48
Read length	> 2Mb	> 2Mb	> 2Mb	> 2Mb	> 2Mb
Yield per flow cell	50 Gb	50 Gb	50 Gb	220 Gb	220 Gb
Number of flow cells per device	1	1	5	24	48
Yield per device	<50 Gb	<50 Gb	<250 Gb	<5.2 Tb	<10.5 Tb
Starting price	$1,000	$4,990	$49,995	$195,455	$327,455

https://nanoporetech.com/products/comparison

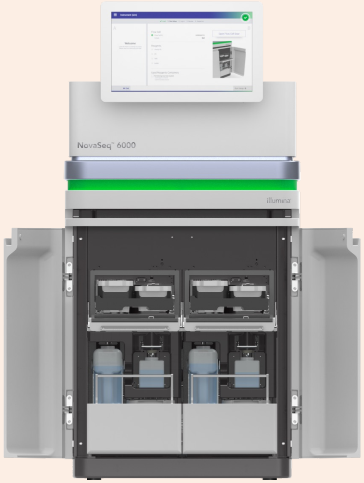
Illumina Sequencers

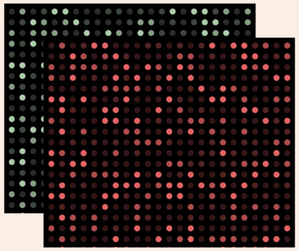
	iSeq 100	MiniSeq	MiSeq	NextSeq 550	NextSeq 2000	NovaSeq 6000
Run time	9.5–19 hrs	4–24 hrs	4–55 hrs	12–30 hrs	24-48 hrs	13-44 hrs
Max. reads per run	4 million	25 million	25 million	400 million	1 billion	20 billion
Max. read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 x 250
Max. output	1.2 Gb	7.5 Gb	15 Gb	120 Gb	300 Gb	6000 Gb
Estimated price	$19,900	$49,500	$128,000	$275,000	$335,000	$985,000

https://www.illumina.com/systems/sequencing-platforms.html

Different Raw Sequencing Data

Illumina

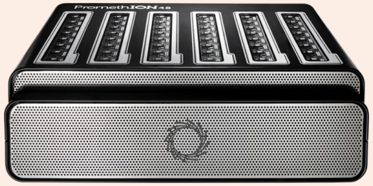


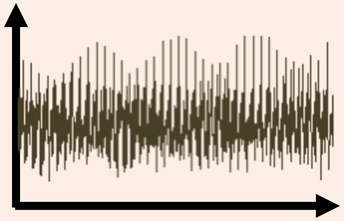


Multiple images

.BCL/.CBCL

ONT

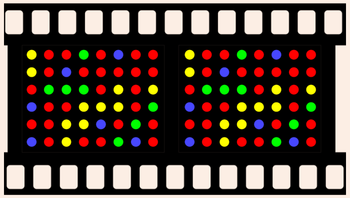




Squiggle

.FAST5

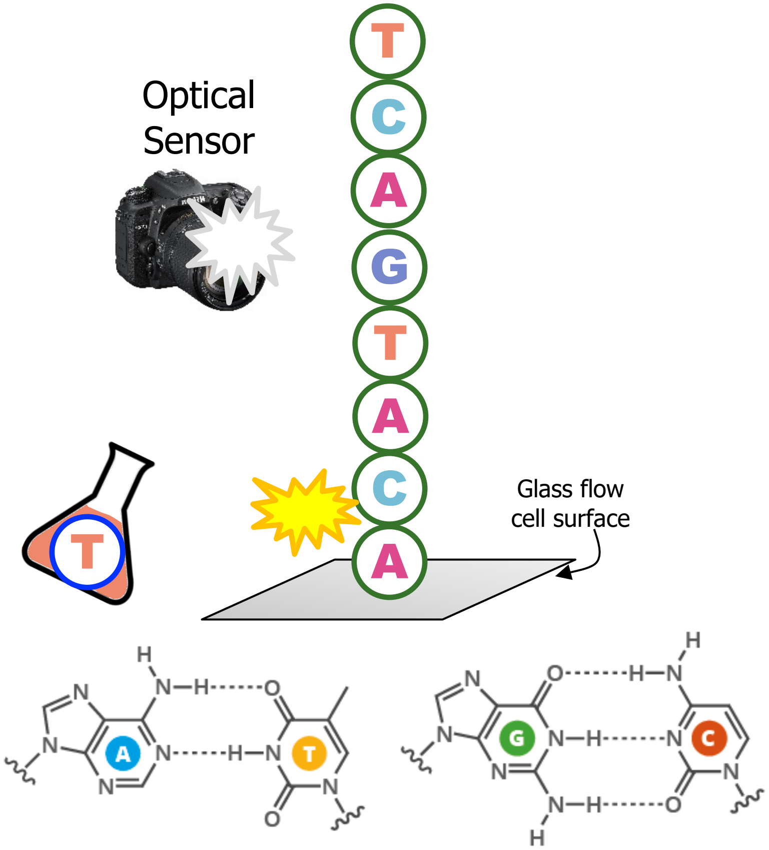
PacBio

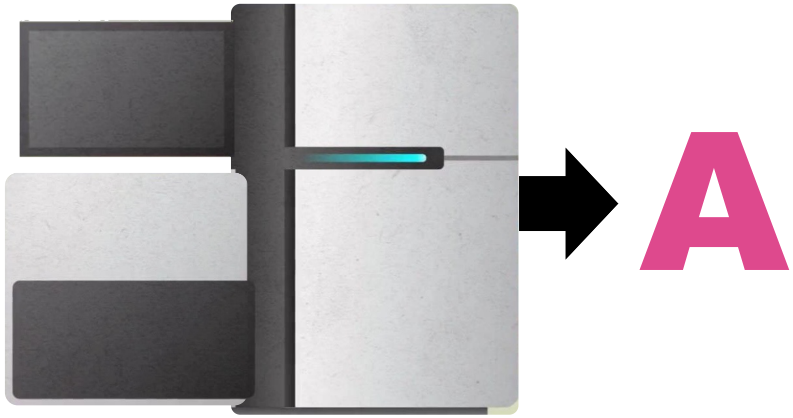


30-hour movie

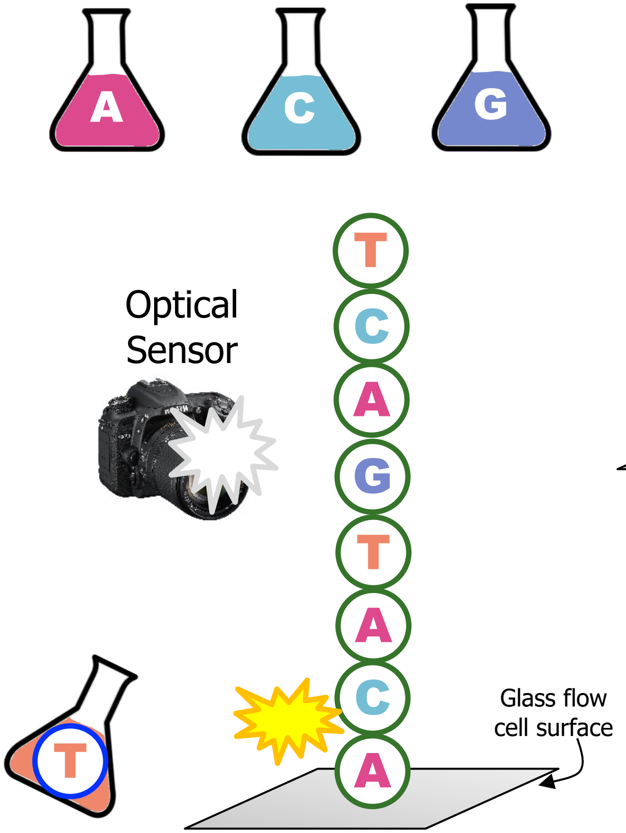
.BAM

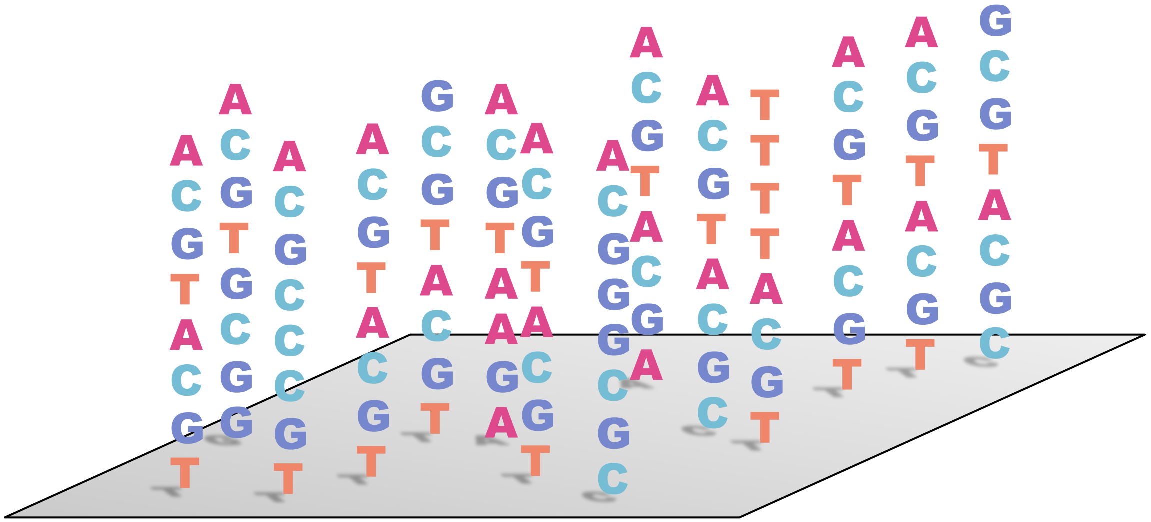
How Does Illumina Machine Work?

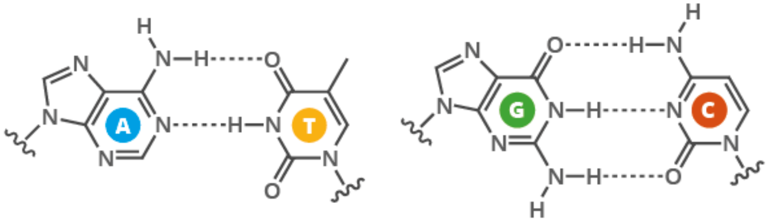


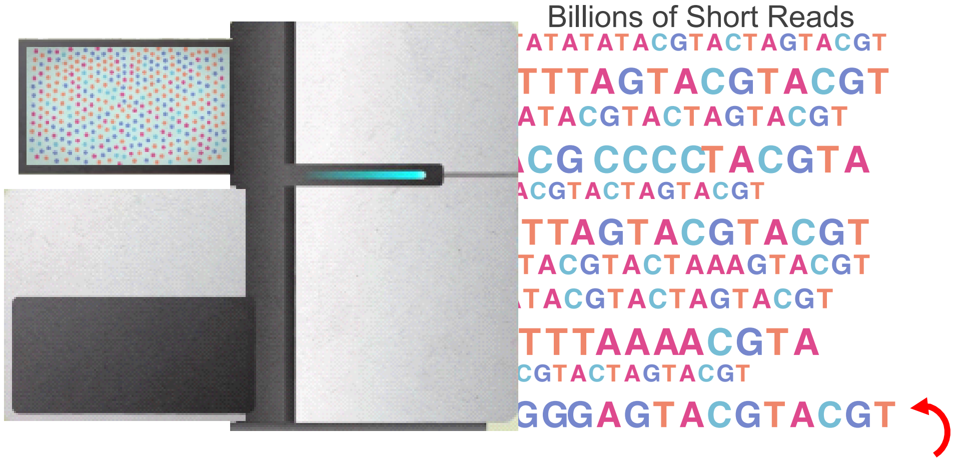


How Does Illumina Machine Work?



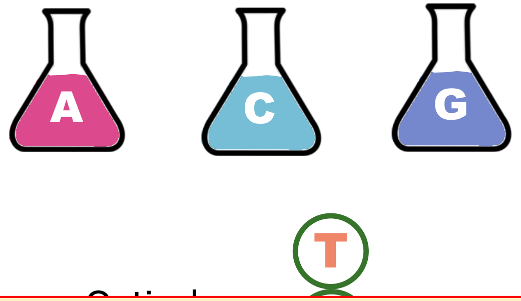






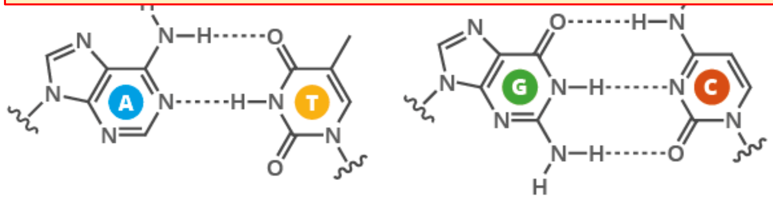
DNA fragment = Read

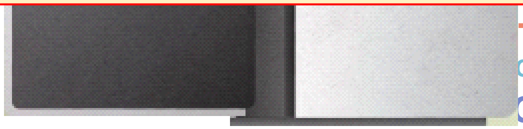
How Does Illumina Machine Work?



Check Illumina virtual tour:

https://emea.illumina.com/systems/sequencing-platforms/iseq/tour.html

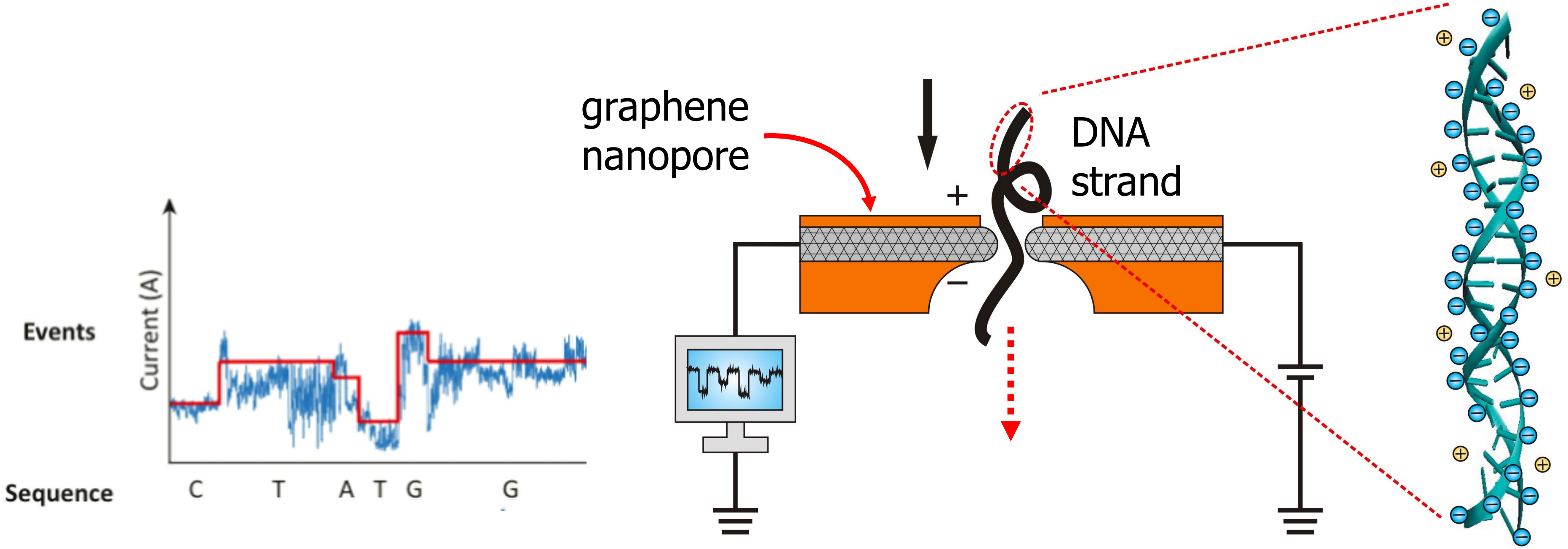




TTTAAAACGTA
CGTACTAGTACGT
GGGAGTACGTACGT

DNA fragment = Read

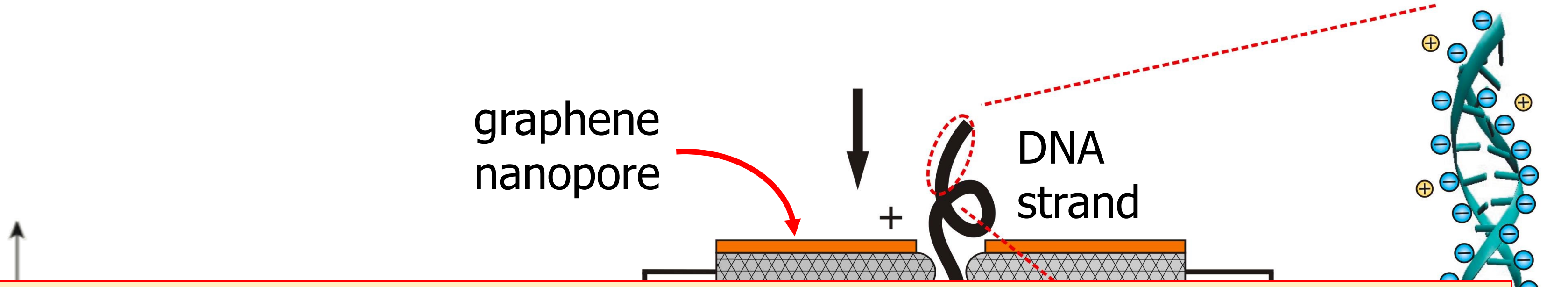
How Does Nanopore Machine Work?



- **Nanopore** is a nano-scale hole (<20nm).
- In nanopore sequencers, an **ionic current** passes through the nanopores
- When the DNA strand passes through the nanopore, the sequencer measures the the **change in current**
- This change is used to identify the bases in the strand with the help of **different electrochemical structures** of the different bases

Figure is adapted from: https://phys.org/news/2013-12-gene-sequencing-future.html

How Does Nanopore Machine Work?



Check Nanopore virtual tour:

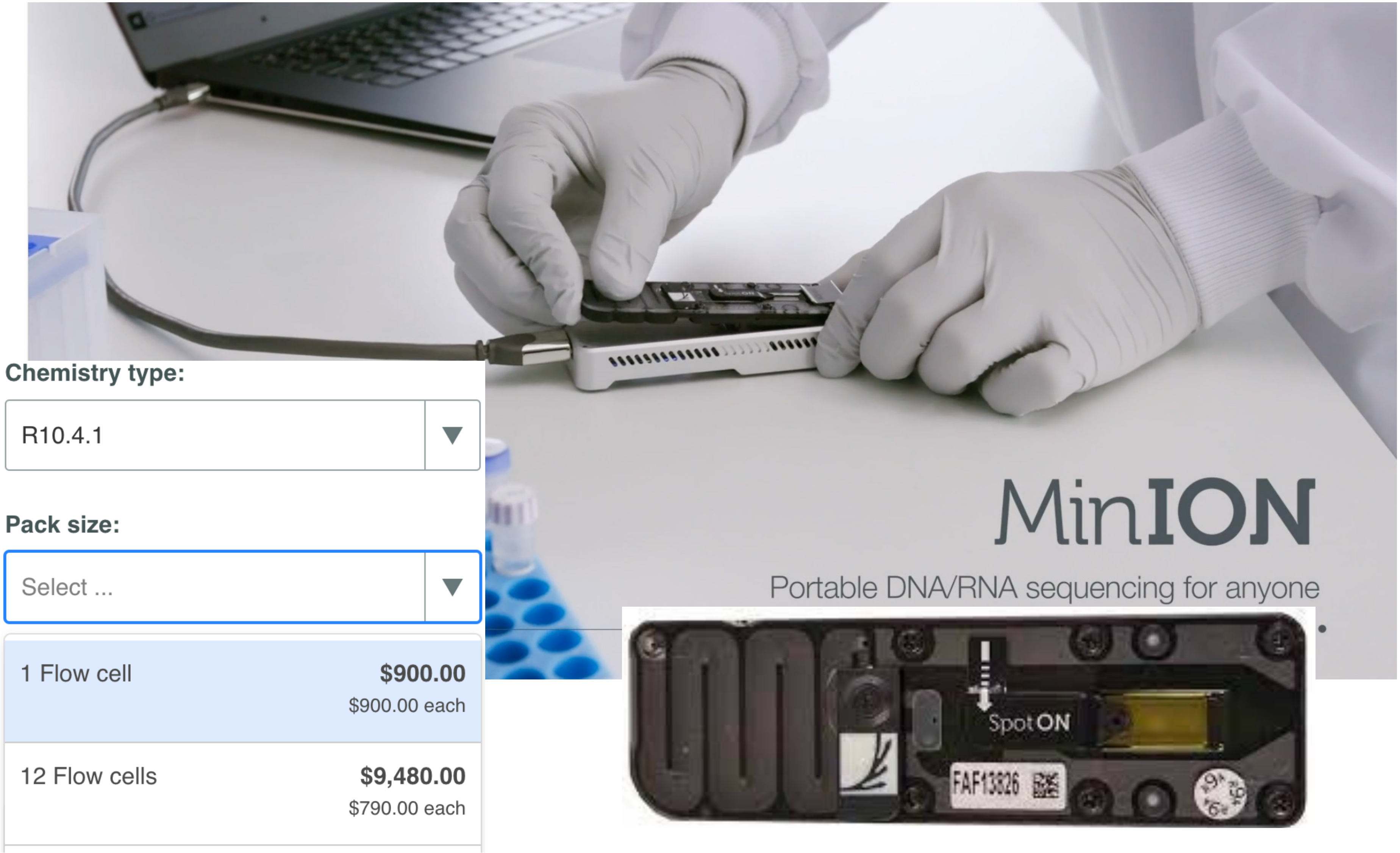
https://nanoporetech.com/resource-centre/minion-video

measures the the **change in current**

- This change is used to identify the bases in the strand with the help of **different electrochemical structures** of the different bases

Figure is adapted from: https://phys.org/news/2013-12-gene-sequencing-future.html

Sequencing in Action



Chemistry type:

R10.4.1 ▼

Pack size:

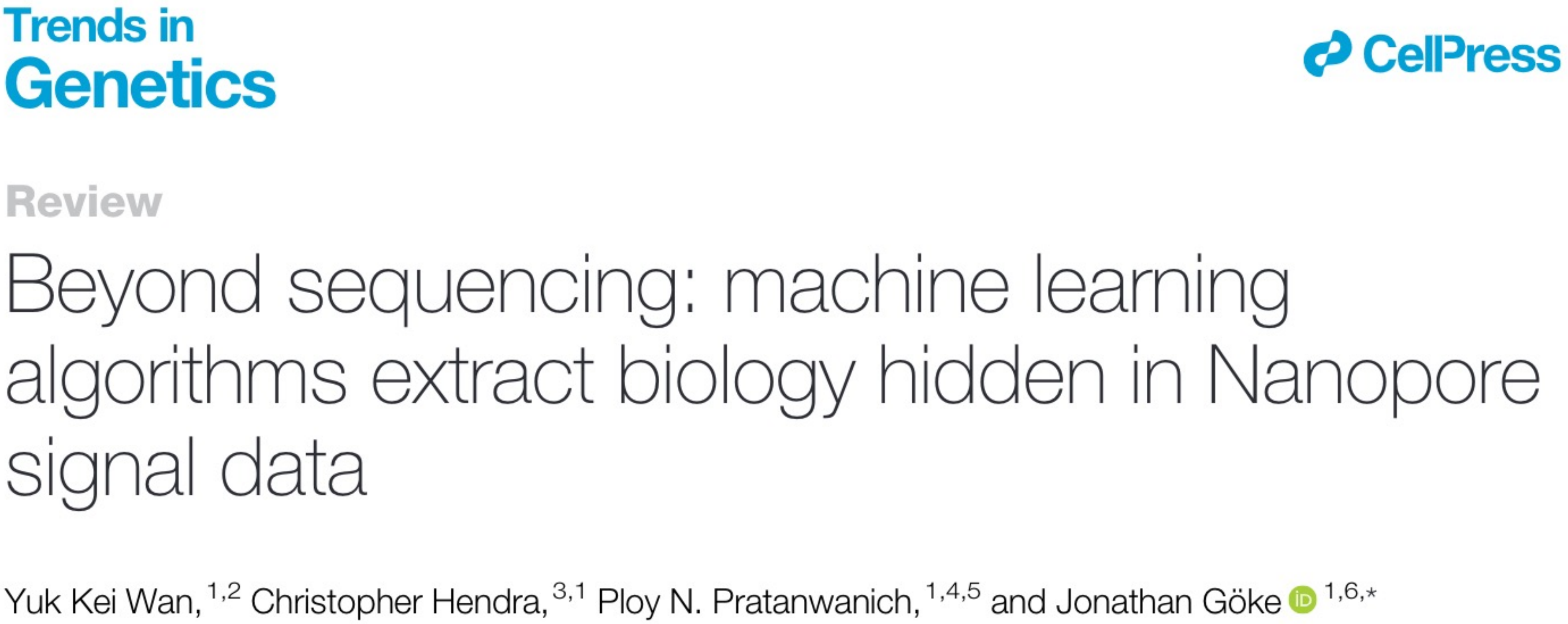
Select ... ▼

1 Flow cell	**$900.00** $900.00 each
12 Flow cells	**$9,480.00** $790.00 each

https://store.nanoporetech.com/flow-cell-r9-4-1.html

Machine Learning for Nanopore Machine

Wan+
“Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data”
Trends in Genetics, October 25, 2021



Trends in
Genetics

CellPress

Review

Beyond sequencing: machine learning algorithms extract biology hidden in Nanopore signal data

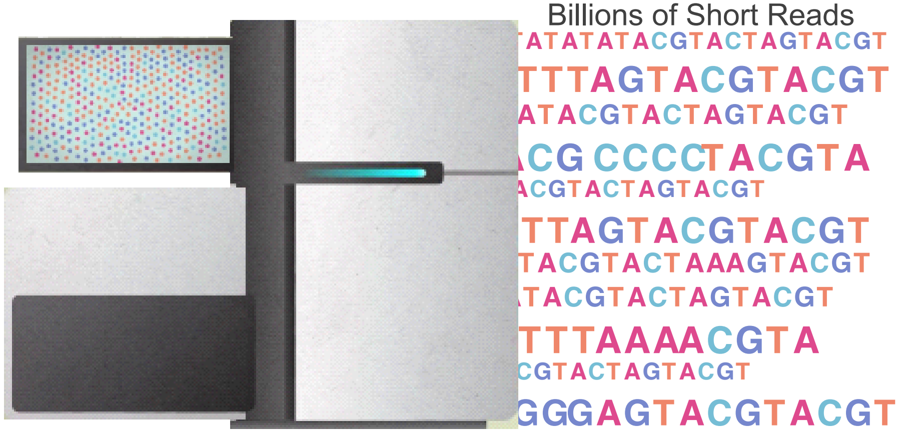
Yuk Kei Wan,[1,2] Christopher Hendra,[3,1] Ploy N. Pratanwanich,[1,4,5] and Jonathan Göke[1,6,*]

Common Disadvantages!

Regardless the sequencing machine,

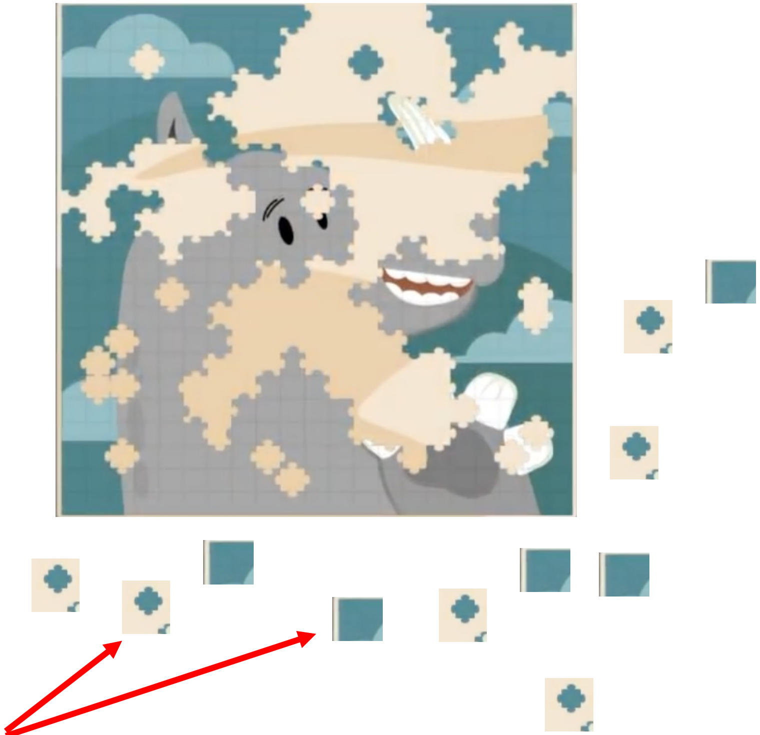
reads still lack information about their order and location

(which part of genome they are originated from)



Solving the Puzzle

Reference genome



Reads

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

HTS Sequencing Output

Small pieces of a puzzle
short reads (Illumina)

Large pieces of a puzzle
long reads (ONT & PacBio)

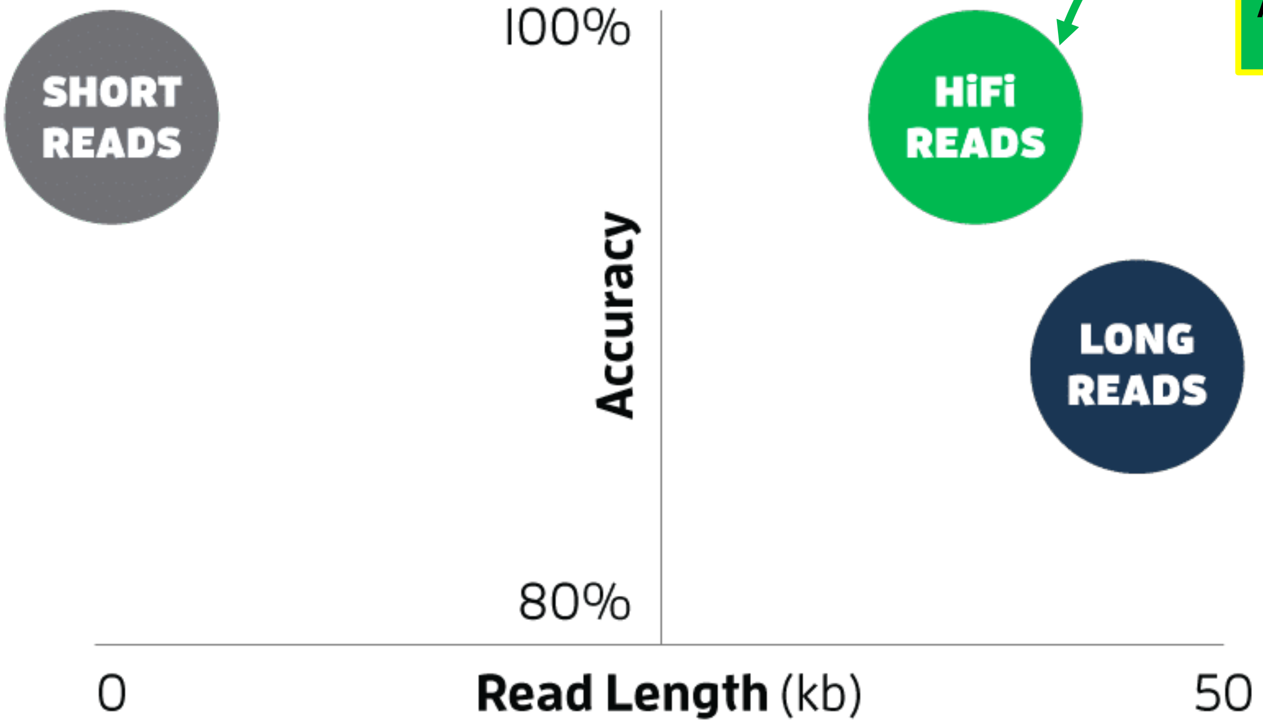
Which sequencing technology is the best?

- ☐ 100-300 bp
- ☐ low error rate (~0.1%)

- ☐ 500-2M bp
- ☐ high error rate (~15%)

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

HiFi Reads (PacBio)



Long: 10-20 kb
Accurate: 99.8%

But still very expensive!

Wenger+, "[Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome]()", *Nature Biotechnology*, 2019

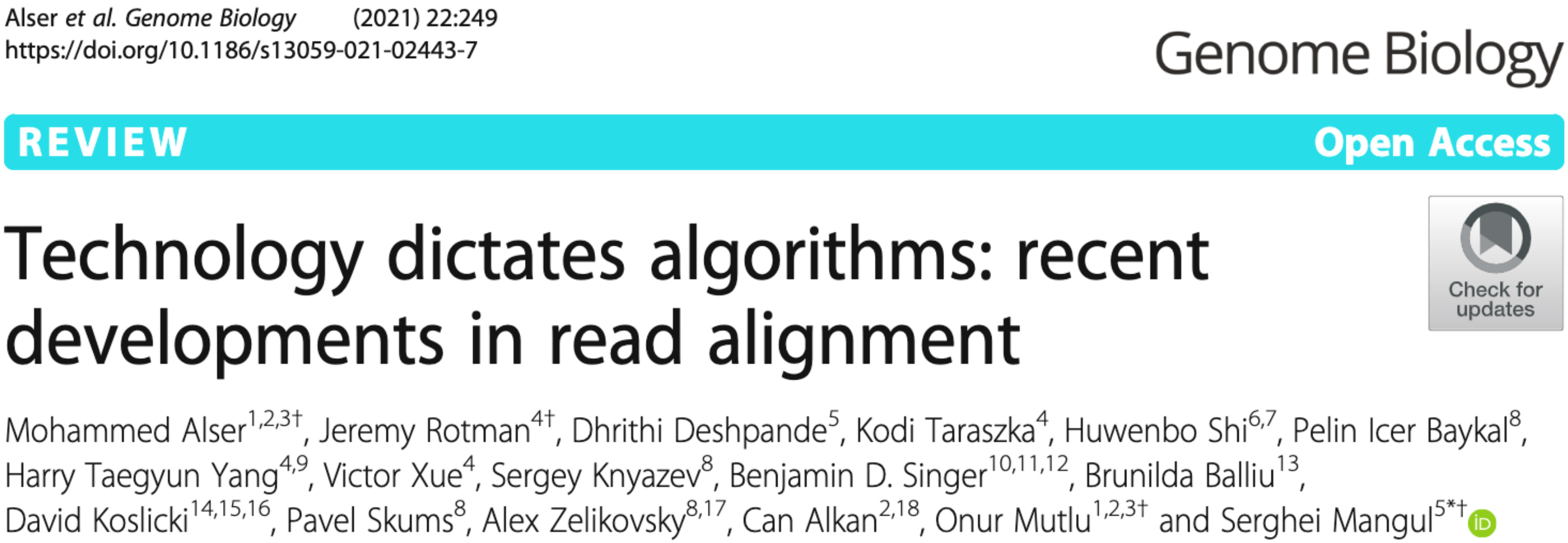
https://labs.wsu.edu/genomicscore/illumina-sequencing/
https://pacbio.gs.washington.edu/

Changes in sequencing technologies can render some read mapping algorithms irrelevant

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

Check for updates

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*†]

Looking forward,

Will we be able to read

the entire genome sequence?

Agenda for Today

- What is Genome Analysis?
- What is Intelligent Genome Analysis?

- How we Analyze Genome?
- **What are the Barriers to Enabling Intelligent Analyses?**

- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

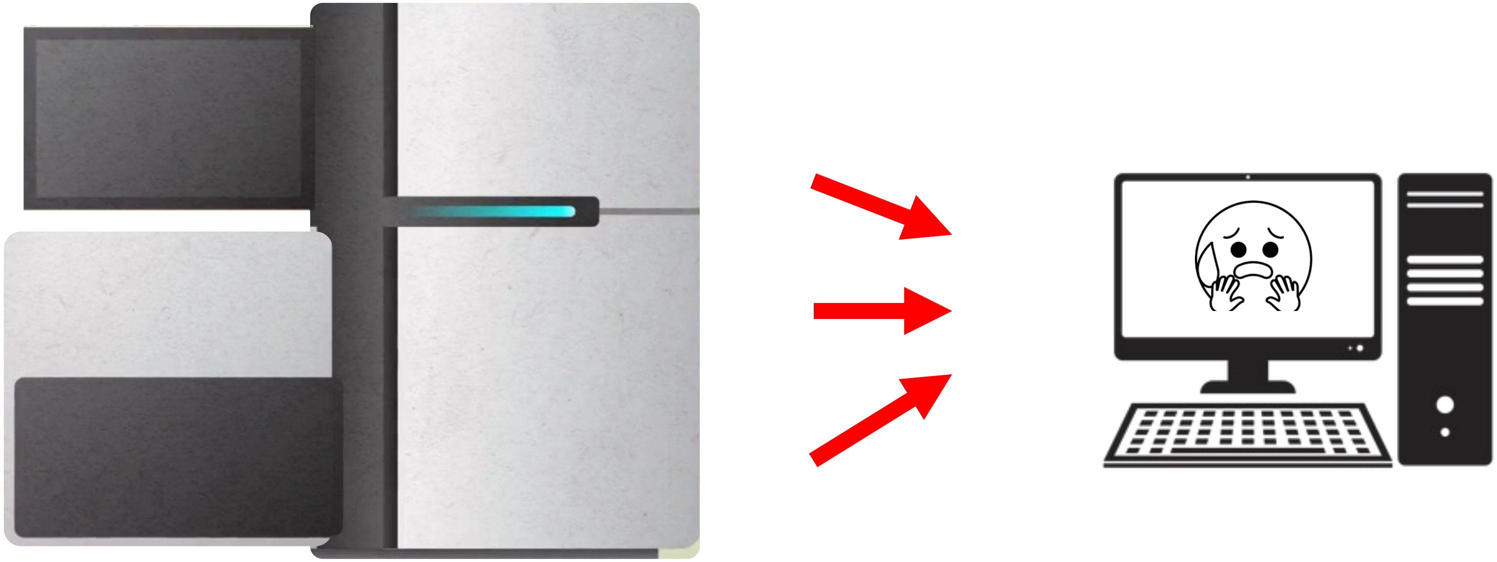
- Where is Genomic Analyses Going Next?

Significant barriers
to intelligent analyses

Significant Barriers to Intelligent Analyses

1. Performance gap between data generation and data processing

Lack of Specialized Compute Capability



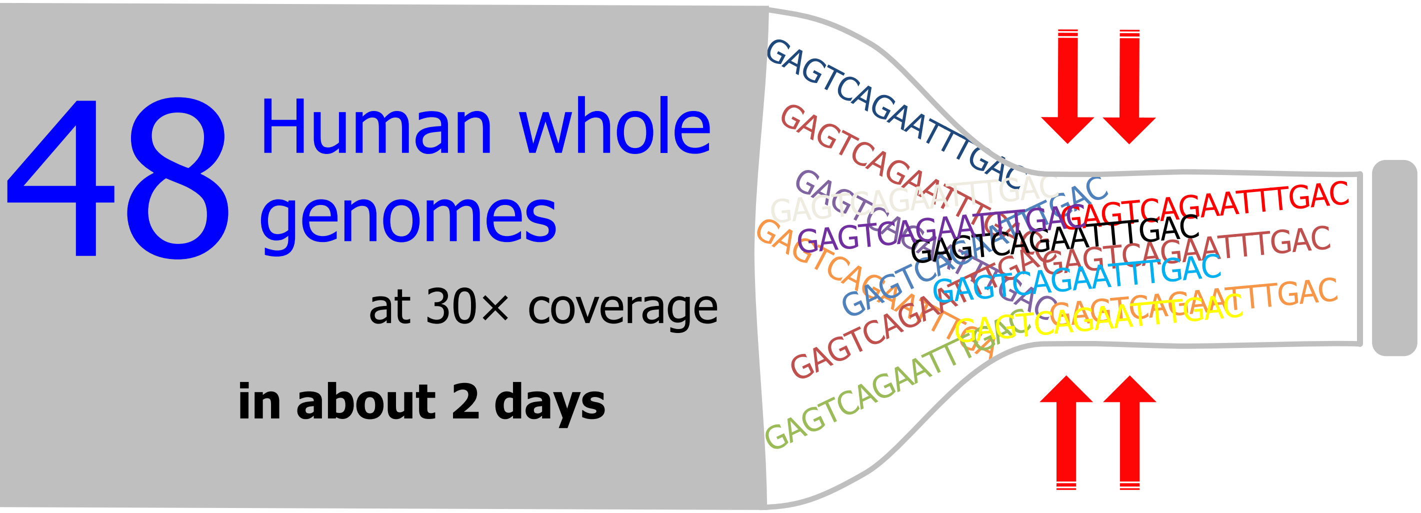
Specialized Machine
for Sequencing

General-Purpose Machine
for Analysis

FAST

SLOW

Analysis is Bottlenecked in Read Mapping!!



48 Human whole genomes

at 30× coverage

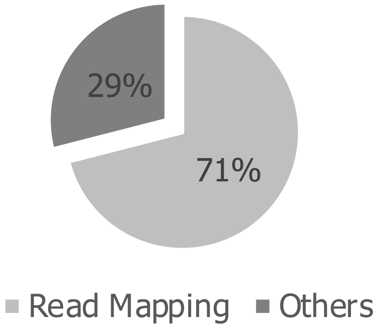
in about 2 days

Illumina NovaSeq 6000

1 Human genome

32 CPU hours

on a 48-core processor



Goyal+, "Ultra-fast next generation human genome sequencing data processing using DRAGENTM bio-IT processor for precision medicine", *Open Journal of Genetics,* 2017.

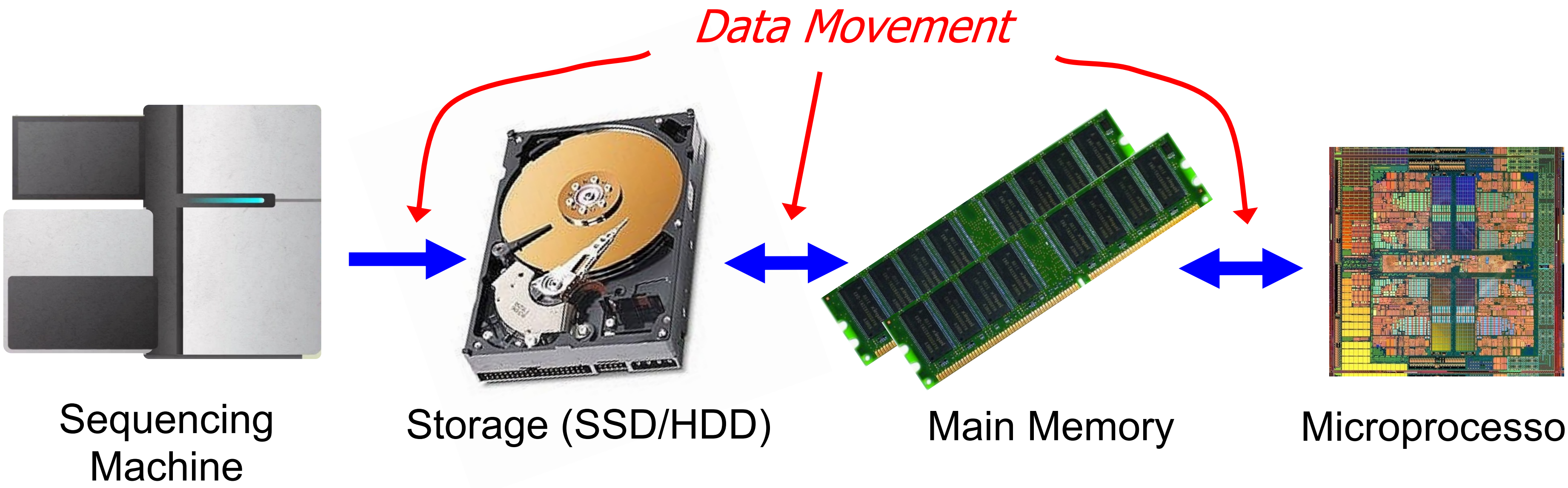
Significant Barriers to Intelligent Analyses

1. Performance gap between data generation and data processing

2. Expensive data movements

Data Movement Dominates Performance

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



Single memory request consumes >160x-800x more energy compared to performing an addition operation

✲ 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

Data analysis

is performed

far away from the data

Significant Barriers to Intelligent Analyses

1. Performance gap between data generation and data processing

2. Expensive data movements

3. Neglecting metadata
   1. Types of sequencing data
   2. Properties of intermediate data
   3. Quality of data
   4. Genome structure

Significant Barriers to Intelligent Analyses

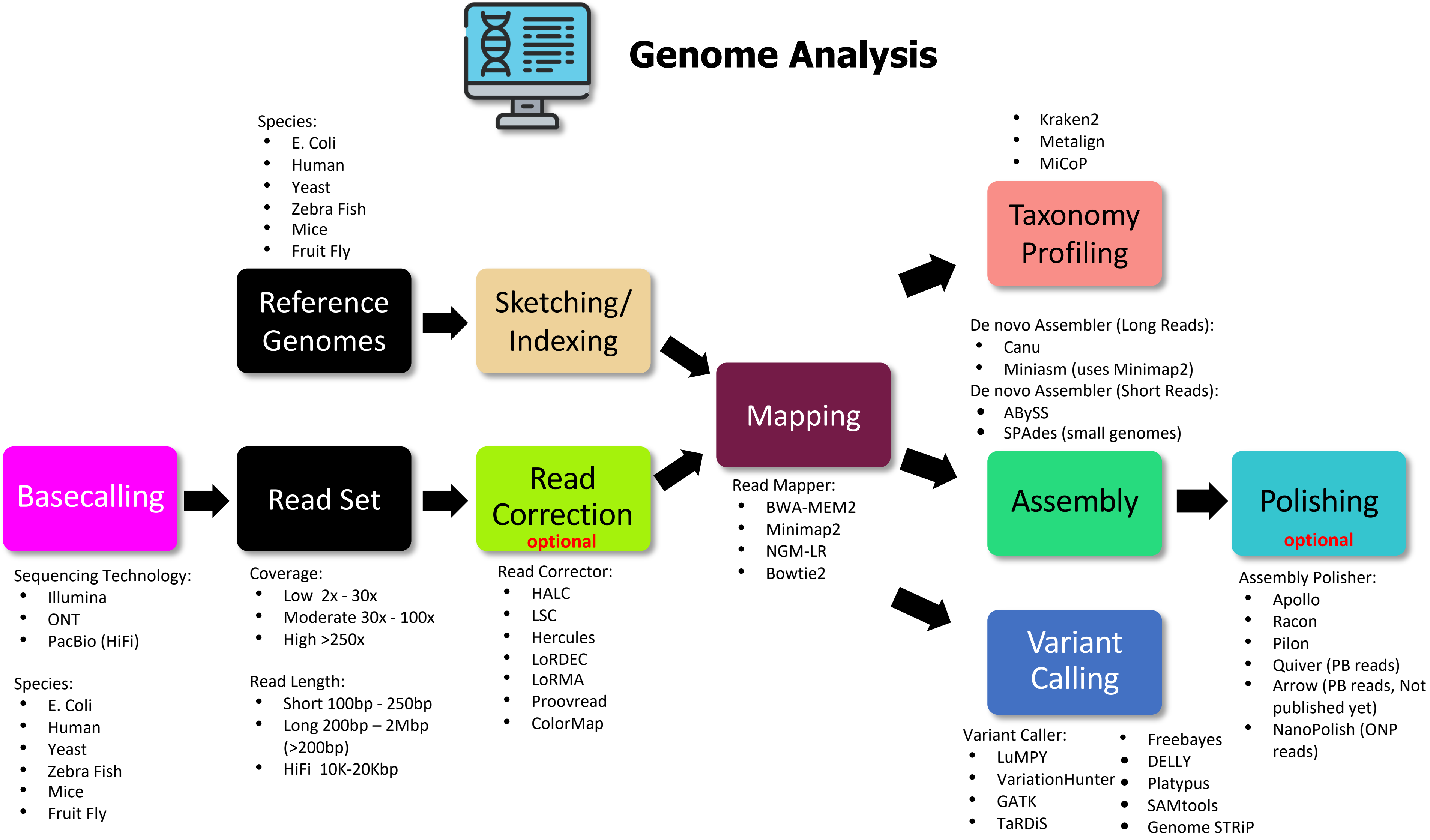
1. Performance gap between data generation and data processing

2. Expensive data movements

3. Neglecting metadata

4. And many more barriers specific to each computational step …

Several Genome Analysis Pipelines



Genome Analysis

Species:
- E. Coli
- Human
- Yeast
- Zebra Fish
- Mice
- Fruit Fly

Reference Genomes

Sketching/ Indexing

Basecalling

Sequencing Technology:
- Illumina
- ONT
- PacBio (HiFi)

Species:
- E. Coli
- Human
- Yeast
- Zebra Fish
- Mice
- Fruit Fly

Read Set

Coverage:
- Low 2x - 30x
- Moderate 30x - 100x
- High >250x

Read Length:
- Short 100bp - 250bp
- Long 200bp – 2Mbp (>200bp)
- HiFi 10K-20Kbp

Read Correction
optional

Read Corrector:
- HALC
- LSC
- Hercules
- LoRDEC
- LoRMA
- Proovread
- ColorMap

Mapping

Read Mapper:
- BWA-MEM2
- Minimap2
- NGM-LR
- Bowtie2

- Kraken2
- Metalign
- MiCoP

Taxonomy Profiling

De novo Assembler (Long Reads):
- Canu
- Miniasm (uses Minimap2)

De novo Assembler (Short Reads):
- ABySS
- SPAdes (small genomes)

Assembly

Polishing
optional

Assembly Polisher:
- Apollo
- Racon
- Pilon
- Quiver (PB reads)
- Arrow (PB reads, Not published yet)
- NanoPolish (ONP reads)

Variant Calling

Variant Caller:
- LuMPY
- VariationHunter
- GATK
- TaRDiS
- Freebayes
- DELLY
- Platypus
- SAMtools
- Genome STRiP

Challenges in Genome Analysis

- ❑ Basecalling: Each sequencing technology provides different types of raw sequencing data.

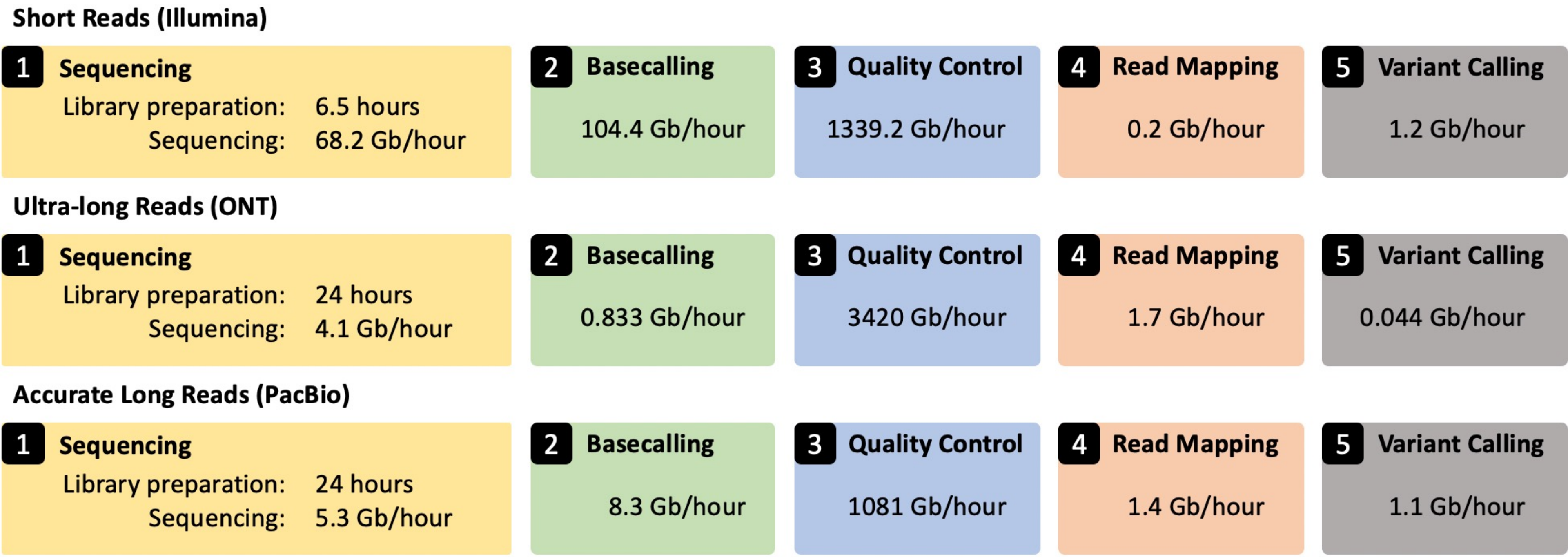
- ❑ Error correction & quality control: Sequencing error rates vary from 0.1%-15%

- ❑ Read mapping: Regardless the sequencing machine, reads are still small randomized fragments of the original DNA sequence with unknown order and location.

- ❑ Variant calling: Small & complex genomic differences need to be maintained.

- ❑ Metagenomic profiling: The sample donor is unknown.

Technology Dictates Algorithm Complexity

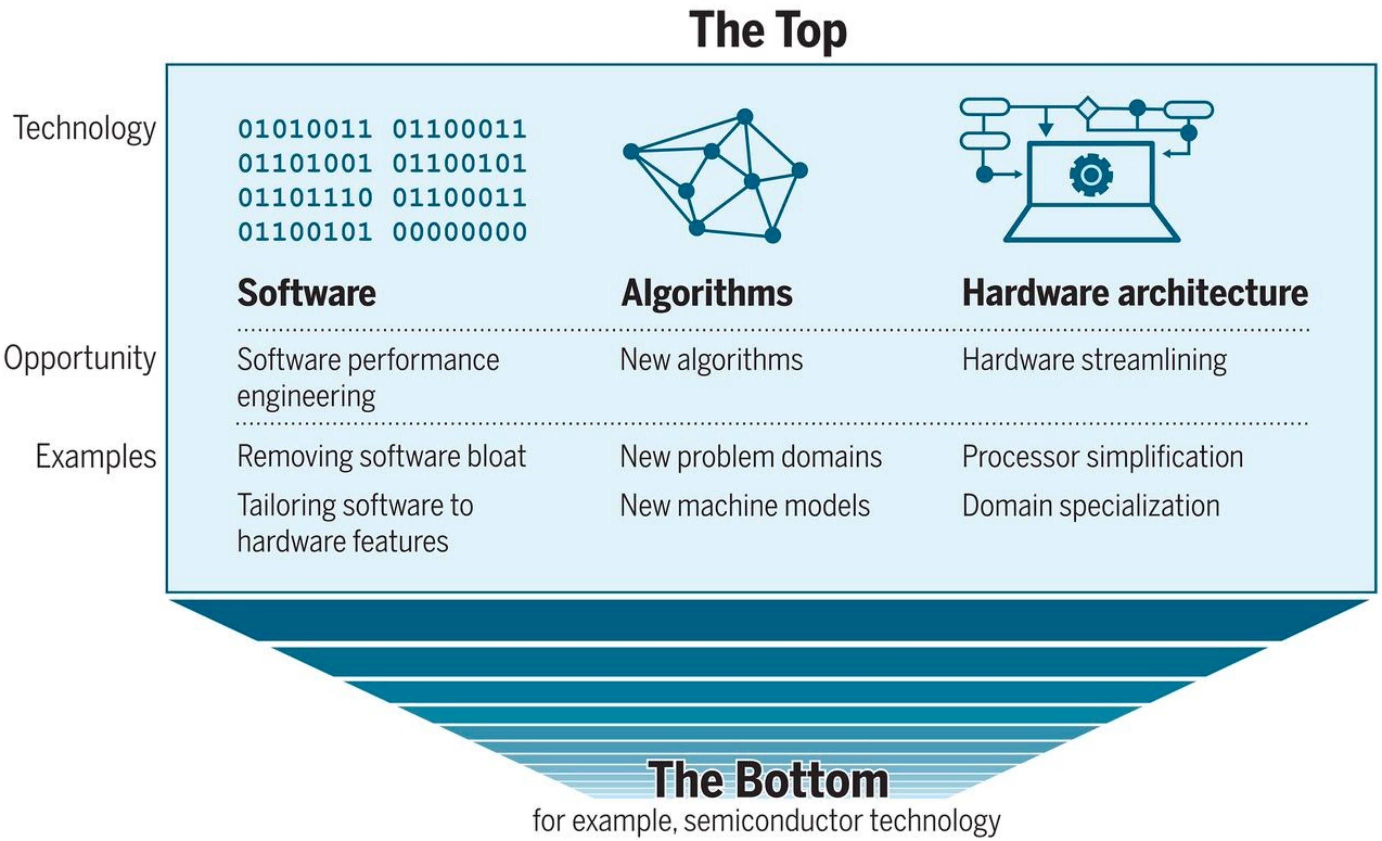


	1 Sequencing	2 Basecalling	3 Quality Control	4 Read Mapping	5 Variant Calling
Short Reads (Illumina)	Library preparation: 6.5 hours Sequencing: 68.2 Gb/hour	104.4 Gb/hour	1339.2 Gb/hour	0.2 Gb/hour	1.2 Gb/hour
Ultra-long Reads (ONT)	Library preparation: 24 hours Sequencing: 4.1 Gb/hour	0.833 Gb/hour	3420 Gb/hour	1.7 Gb/hour	0.044 Gb/hour
Accurate Long Reads (PacBio)	Library preparation: 24 hours Sequencing: 5.3 Gb/hour	8.3 Gb/hour	1081 Gb/hour	1.4 Gb/hour	1.1 Gb/hour

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

Computing System

Leiserson+, "There’s plenty of room at the Top: What will drive computer performance after Moore’s law?", Science, 2020



The Top

	Software	Algorithms	Hardware architecture
Technology	01010011 01100011 01101001 01100101 01101110 01100011 01100101 00000000		
Opportunity	Software performance engineering	New algorithms	Hardware streamlining
Examples	Removing software bloat Tailoring software to hardware features	New problem domains New machine models	Processor simplification Domain specialization

The Bottom

for example, semiconductor technology

- Data
- Problem
- Algorithm
- Program/Language
- Runtime System (VM, OS, MM)
- ISA (Architecture)
- Microarchitecture
- Logic
- Devices
- Electrons

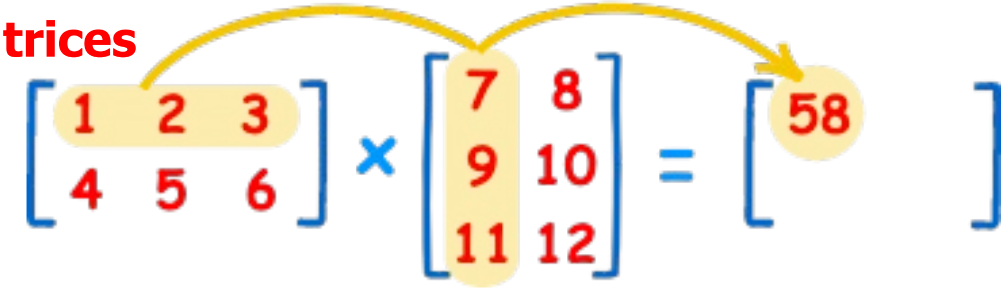
Richard Feynman, "There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics”, a lecture given at Caltech, 1959.

Image source: https://science.sciencemag.org/content/368/6495/eaam9744

Software & Hardware Optimizations

Multiplying Two 4096-by-4096 Matrices

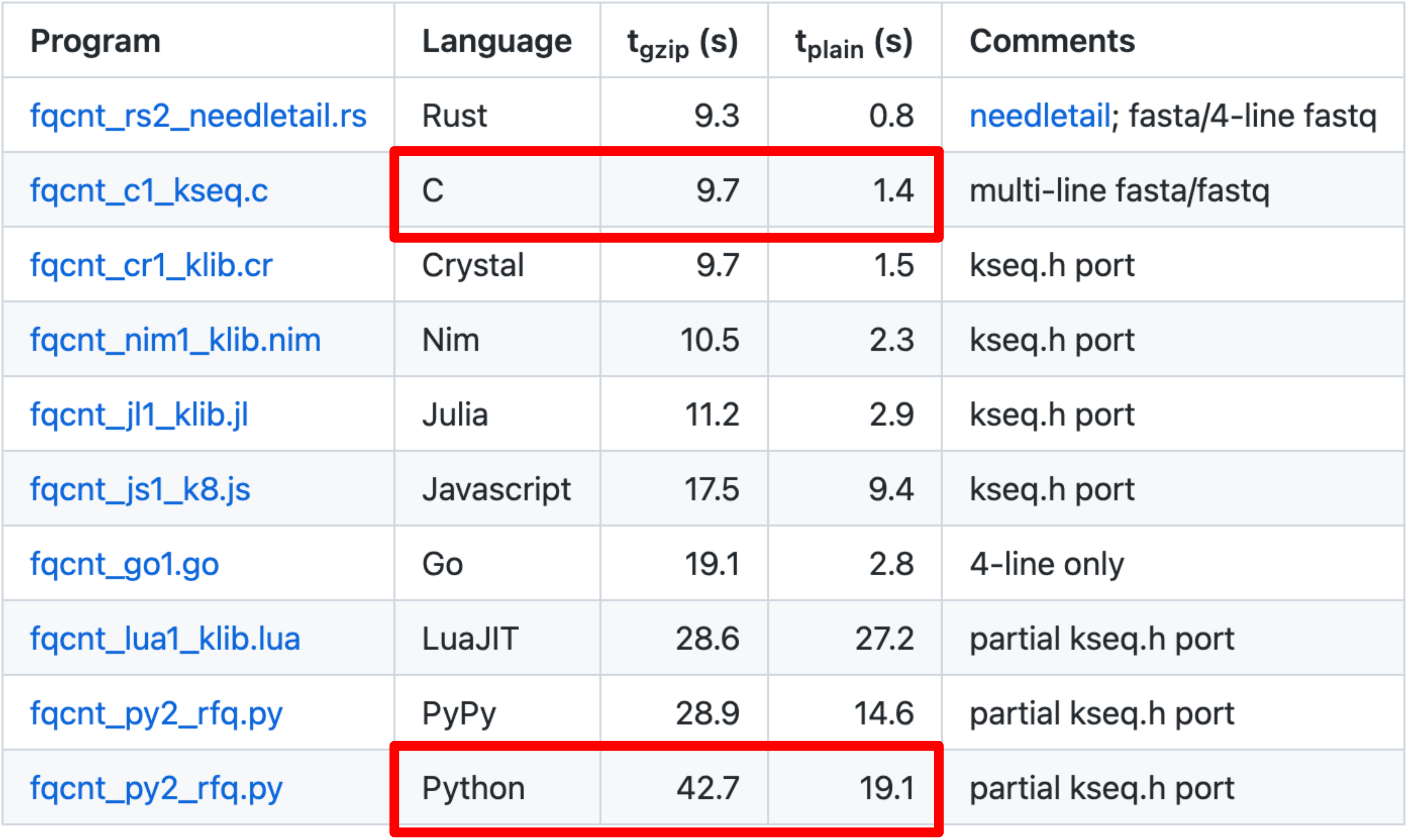
```
for i in xrange(4096):
  for j in xrange(4096):
    for k in xrange(4096):
      C[i][j] += A[i][k] *
B[k][j]
```



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

https://github.com/lh3/biofast

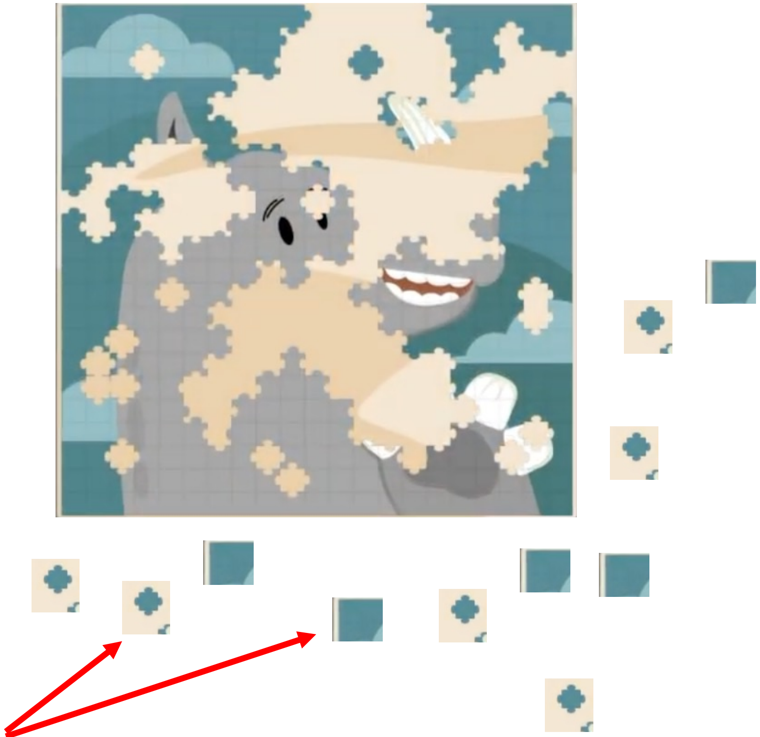
We need intelligent algorithms and intelligent architectures that handle data well

Solving the Puzzle

.FASTA file

Reference genome

.FASTQ file



Reads

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

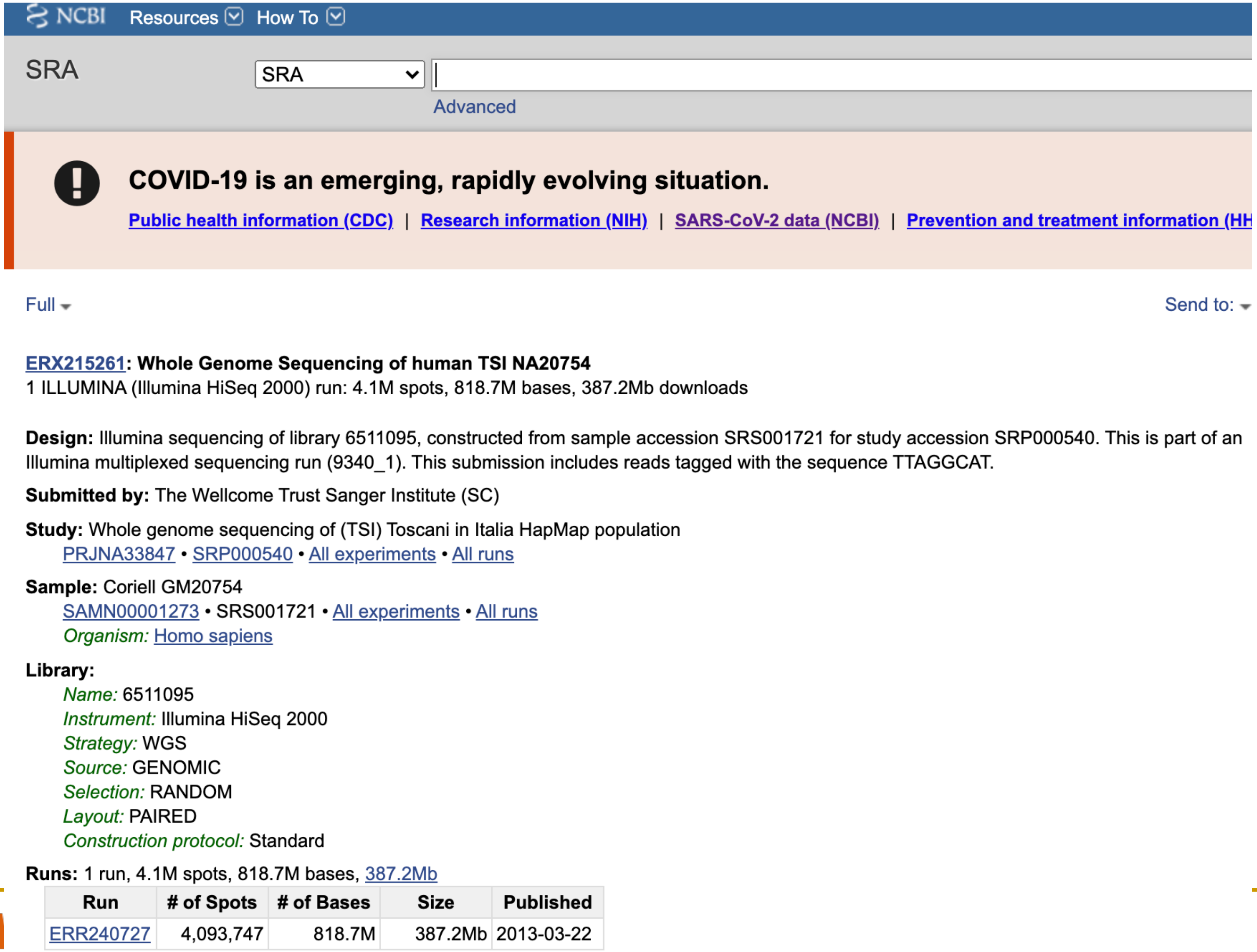
Obtaining the Human Reference Genome

- **GRCh38.p13**
- Description: Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13)
- Organism name: Homo sapiens (human)
- Date: 2019/02/28
- 3,099,706,404 bases
- Compressed .fna file (964.9 MB)
- https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.39

```
>NC_000001.11 Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
….
```

Obtaining .FASTQ Files

- https://www.ncbi.nlm.nih.gov/sra/ERR240727



COVID-19 is an emerging, rapidly evolving situation.

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HH

ERX215261: Whole Genome Sequencing of human TSI NA20754
1 ILLUMINA (Illumina HiSeq 2000) run: 4.1M spots, 818.7M bases, 387.2Mb downloads

Design: Illumina sequencing of library 6511095, constructed from sample accession SRS001721 for study accession SRP000540. This is part of an Illumina multiplexed sequencing run (9340_1). This submission includes reads tagged with the sequence TTAGGCAT.

Submitted by: The Wellcome Trust Sanger Institute (SC)

Study: Whole genome sequencing of (TSI) Toscani in Italia HapMap population
PRJNA33847 • SRP000540 • All experiments • All runs

Sample: Coriell GM20754
SAMN00001273 • SRS001721 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: 6511095
Instrument: Illumina HiSeq 2000
Strategy: WGS
Source: GENOMIC
Selection: RANDOM
Layout: PAIRED
Construction protocol: Standard

Runs: 1 run, 4.1M spots, 818.7M bases, 387.2Mb

Run	# of Spots	# of Bases	Size	Published
ERR240727	4,093,747	818.7M	387.2Mb	2013-03-22

Let’s learn
how to map a read

Read Mapping: A Brute Force Algorithm

Reference
Read

Very expensive!

$O(m^2kn)$

m: read length

k: no. of reads

n: reference genome length

Matching Each Read with Reference Genome

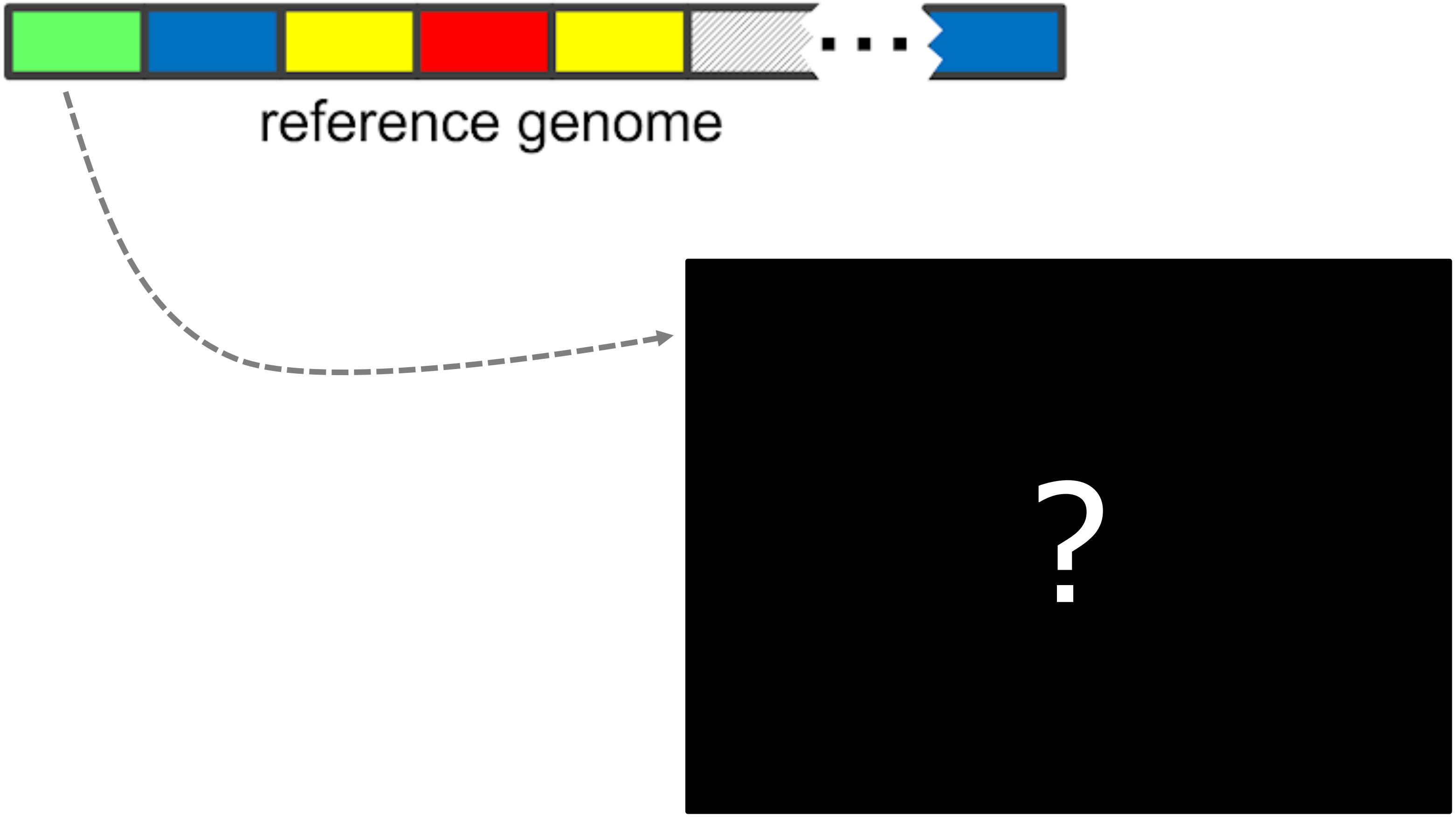
.FASTA file:

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCT🟩TCATTGACATTTAAACTCTGGGGCAGGT🟦GAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCC🟨CCCCGGCCCGGCTCGGGGCCCGCGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGT🟦CAAAAGTAGCAA🟩CTCCTAA🟥TCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCA🟨ACGCTTGGAAAG
TCCGTACCCGCGCCT🟥AAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

.FASTQ file:

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TT🟦AATAAATCTC🟩TTAGATN🟥NNNNNNNNNTAG
+
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT
```

Step 1: Indexing the Reference Genome

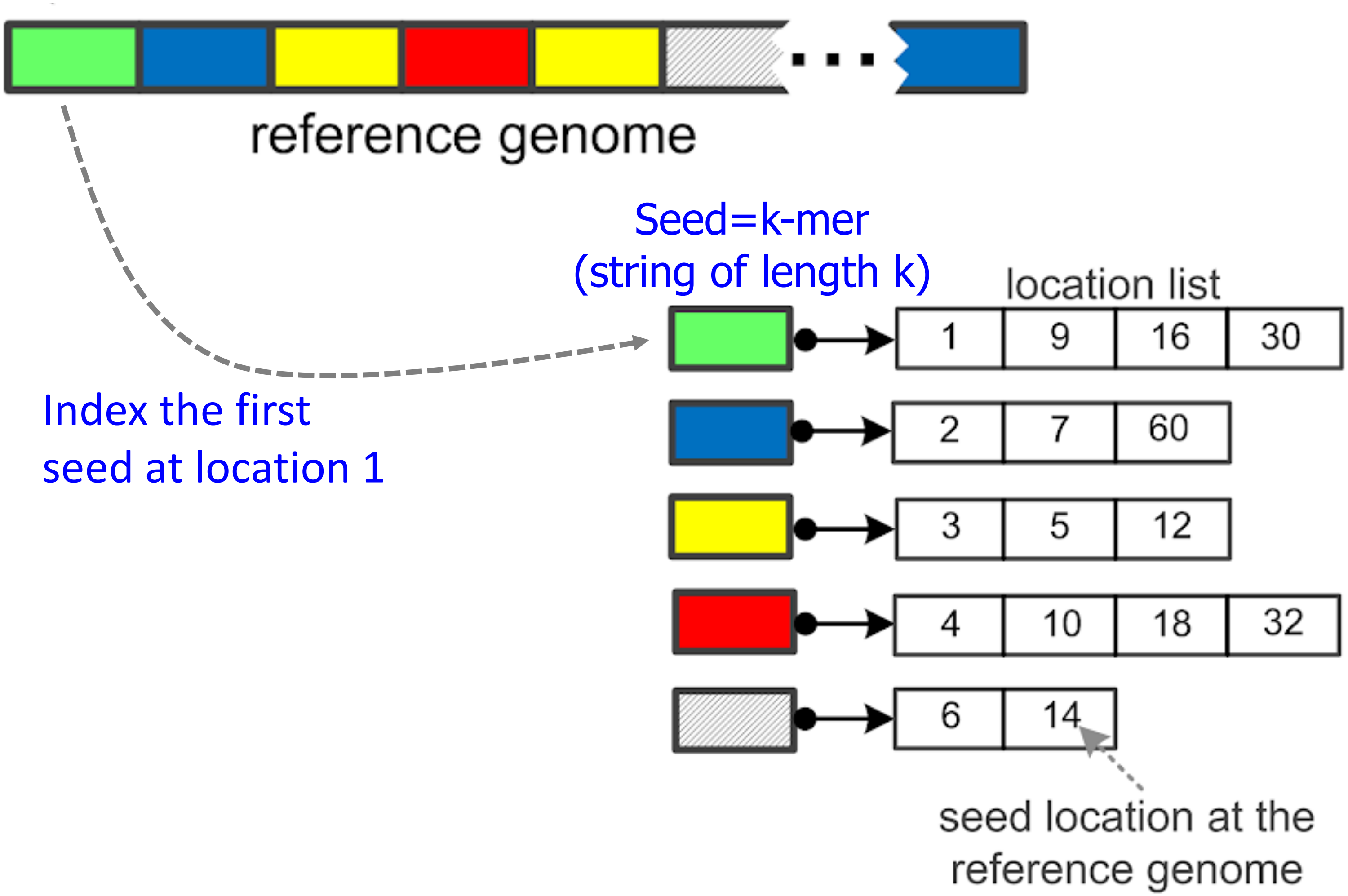


Popular Indexing Technique

Hashing is the most popular indexing technique for read mapping since 1988

Alser+, "Technology dictates algorithms: Recent developments in read alignment", Genome Biology, 2021

Step 1: Indexing the Reference Genome



Seed=k-mer
(string of length k)

Index the first
seed at location 1

Genome Index Properties

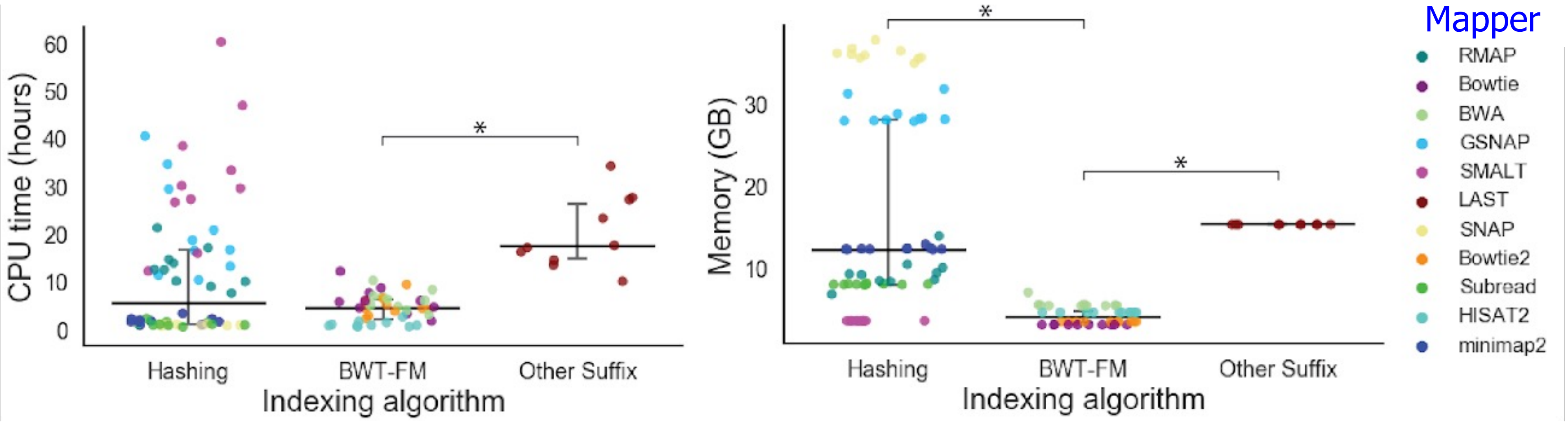
- The index is built only once for each reference.

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

Tool	Version	Index Size*	Indexing Time
mrFAST	2.2.5	16.5 GB	20.00 min
minimap2	0.12.7	7.2 GB	3.33 min
BWA-MEM	0.7.17	4.7 GB	49.96 min

*Human genome = 3.2 GB

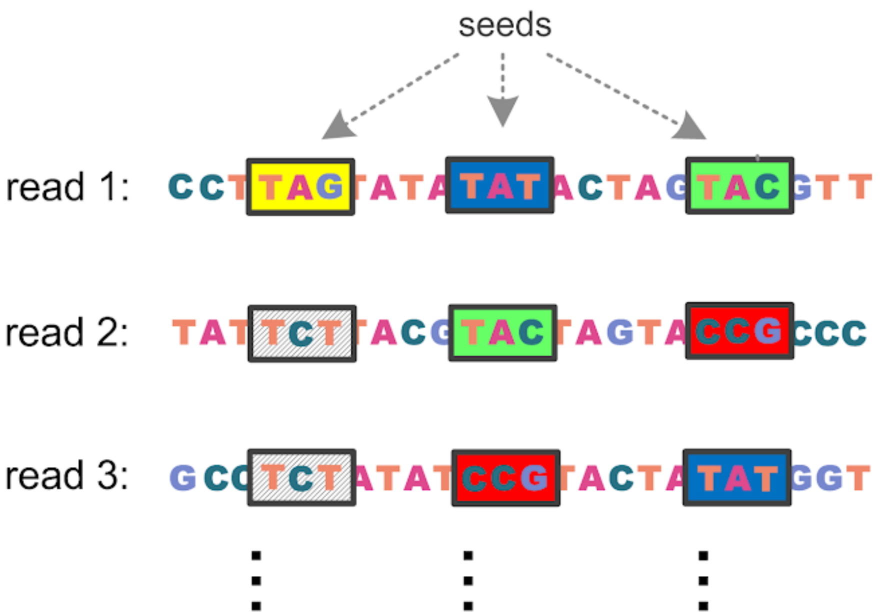
Performance of Human Genome Indexing



Alser+, "Technology dictates algorithms: Recent developments in read alignment", Genome Biology, 2021

SAFARI

Step 2: Query the Index Using Read Seeds



seeds

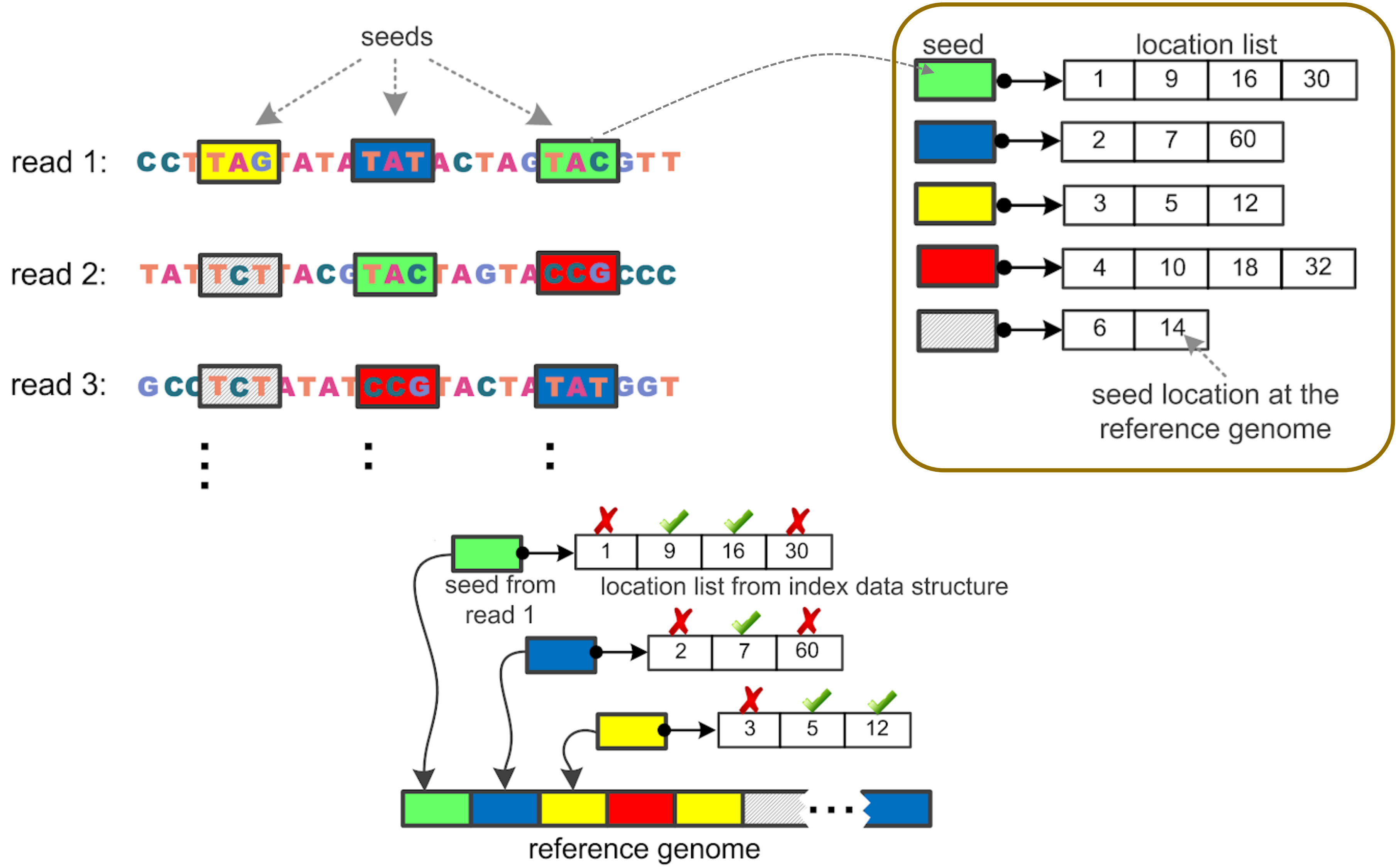
read 1: CCTTAGTATATATACTAGTACGTT

read 2: TATTCTTACGTACTAGTACCGCCC

read 3: GCCTCTATATCCGTACTATATGGT

⋮ ⋮ ⋮

Step 2: Query the Index Using Read Seeds

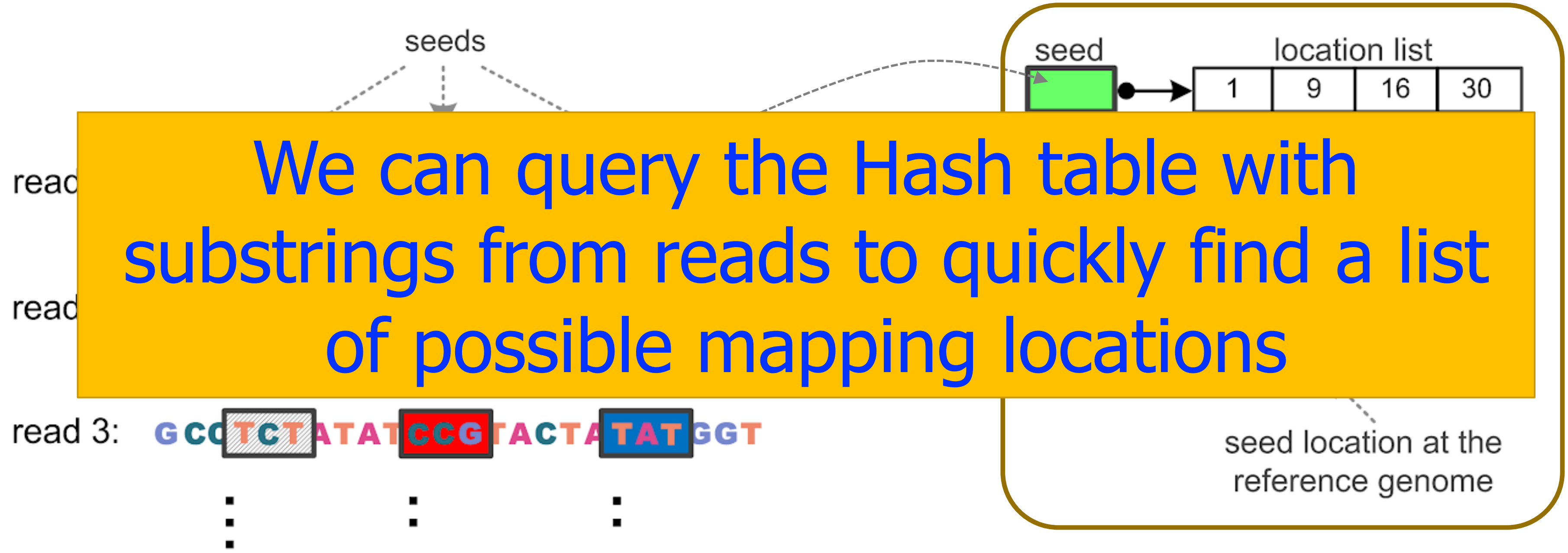

seeds
read 1: CCTTAGTATATATACTAGTACGTT
read 2: TATTCTTACGTACTAGTACCGCCC
read 3: GCCTCTATATCCGTACTATATGGT
seed
location list
seed location at the reference genome
seed from read 1
location list from index data structure
reference genome

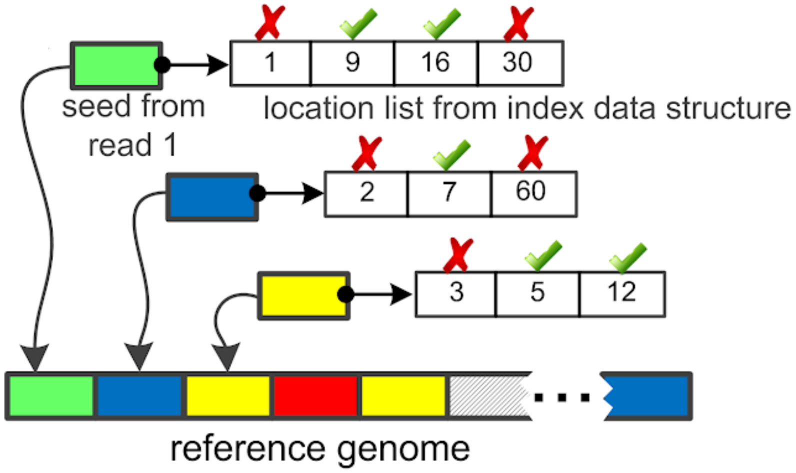
seed	location list			
🟩	1	9	16	30
🟦	2	7	60	
🟨	3	5	12	
🟥	4	10	18	32
▨	6	14		

seed from read 1	location list from index data structure			
🟩	1 ✗	9 ✓	16 ✓	30 ✗
🟦	2 ✗	7 ✓	60 ✗	
🟨	3 ✗	5 ✓	12 ✓	

Step 2: Query the Index Using Read Seeds

We can query the Hash table with substrings from reads to quickly find a list of possible mapping locations



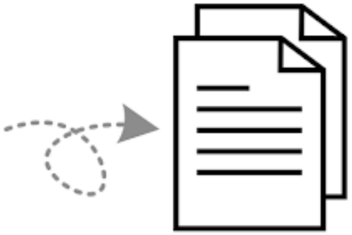


Step 3: Sequence Alignment (Verification)

		C	G	T	T	A	G	T	C	T	A
	0	0	0	0	0	0	0	0	0	0	0
C	0	2	2	2	2	2	2	2	2	2	2
C	0	2	3	3	3	3	3	3	4	4	4
T	0	2	3	5	5	5	5	5	5	6	6
T	0	2	3	5	7	7	7	7	7	7	7
A	0	3	3	5	7	9	9	9	9	9	9
G	0	2	4	5	7	9	11	11	11	11	11
T	0	2	4	6	7	9	11	13	13	13	13
A	0	2	4	6	7	9	11	13	14	14	15
T	0	2	4	6	8	9	11	13	14	16	16

…

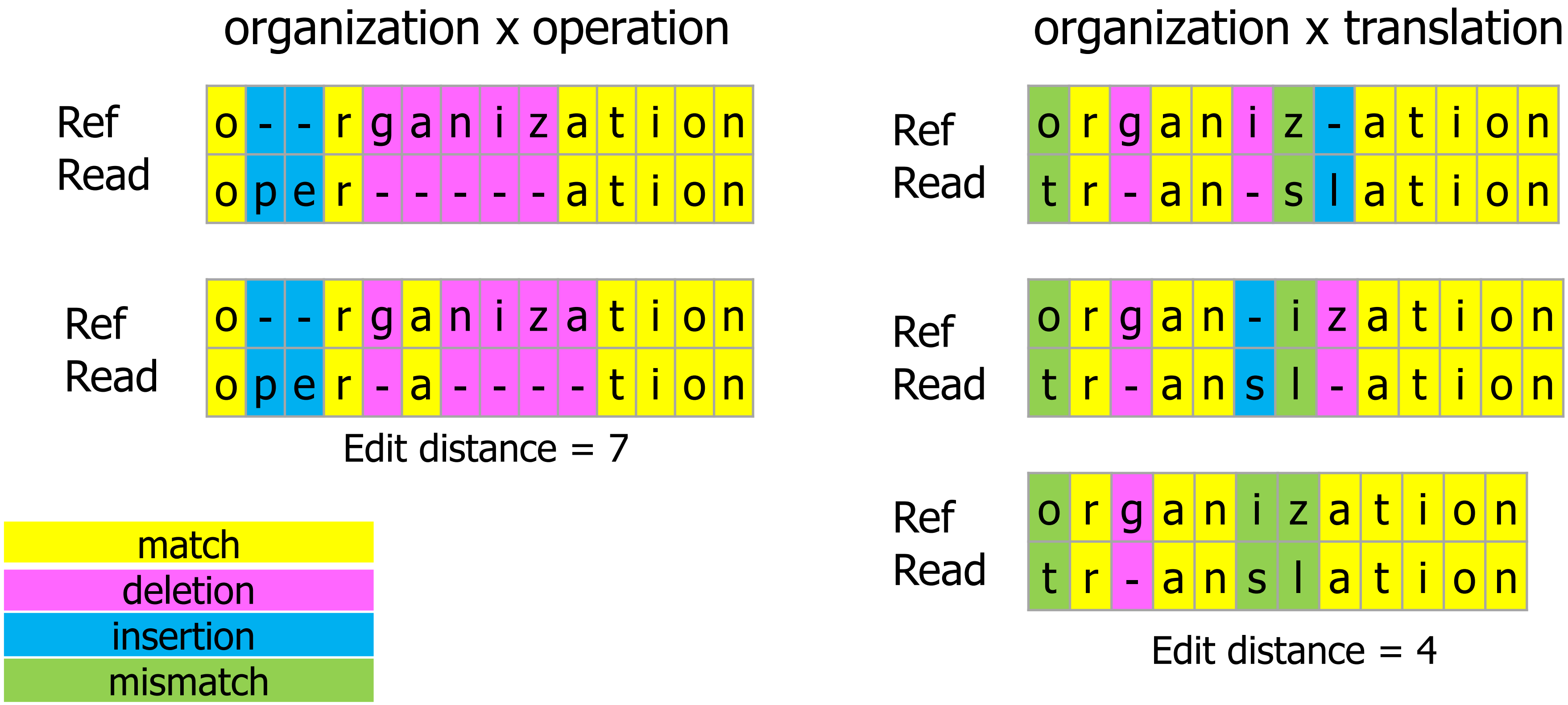
⋮



.bam/.sam file contains necessary alignment information (e.g., type, location, and number of each edit)

Step 3: Sequence Alignment (Verification)

- **Edit distance** is defined as the minimum number of edits (i.e. insertions, deletions, or substitutions) needed to make the read exactly match the reference segment.



organization x operation

Ref	o	-	-	r	g	a	n	i	z	a	t	i	o	n
Read	o	p	e	r	-	-	-	-	-	a	t	i	o	n

Ref	o	-	-	r	g	a	n	i	z	a	t	i	o	n
Read	o	p	e	r	-	a	-	-	-	-	t	i	o	n

Edit distance = 7

organization x translation

Ref	o	r	g	a	n	i	z	-	a	t	i	o	n
Read	t	r	-	a	n	-	s	l	a	t	i	o	n

Ref	o	r	g	a	n	-	i	z	a	t	i	o	n
Read	t	r	-	a	n	s	l	-	a	t	i	o	n

Ref	o	r	g	a	n	i	z	a	t	i	o	n
Read	t	r	-	a	n	s	l	a	t	i	o	n

Edit distance = 4

match
deletion
insertion
mismatch

Popular Algorithms for Sequence Alignment

Smith-Waterman remains

the most popular algorithm

since 1988

Hamming distance is

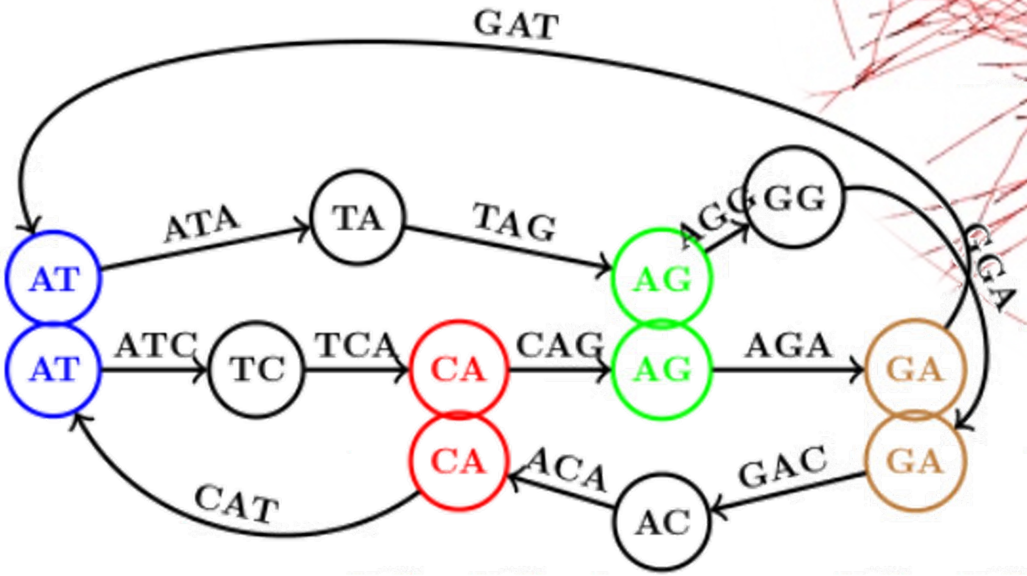
the second most popular technique

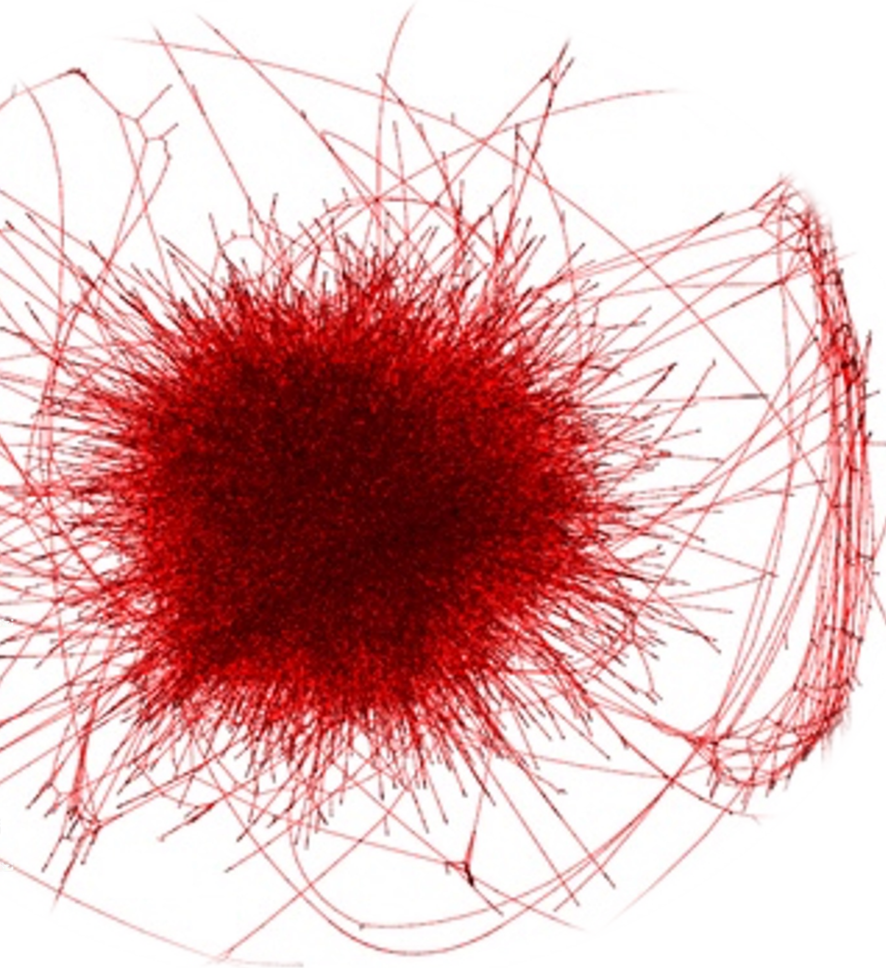
since 2008

Alser+, "Technology dictates algorithms: Recent developments in read alignment", Genome Biology, 2021

De Novo Genome Assembly

Reference-free



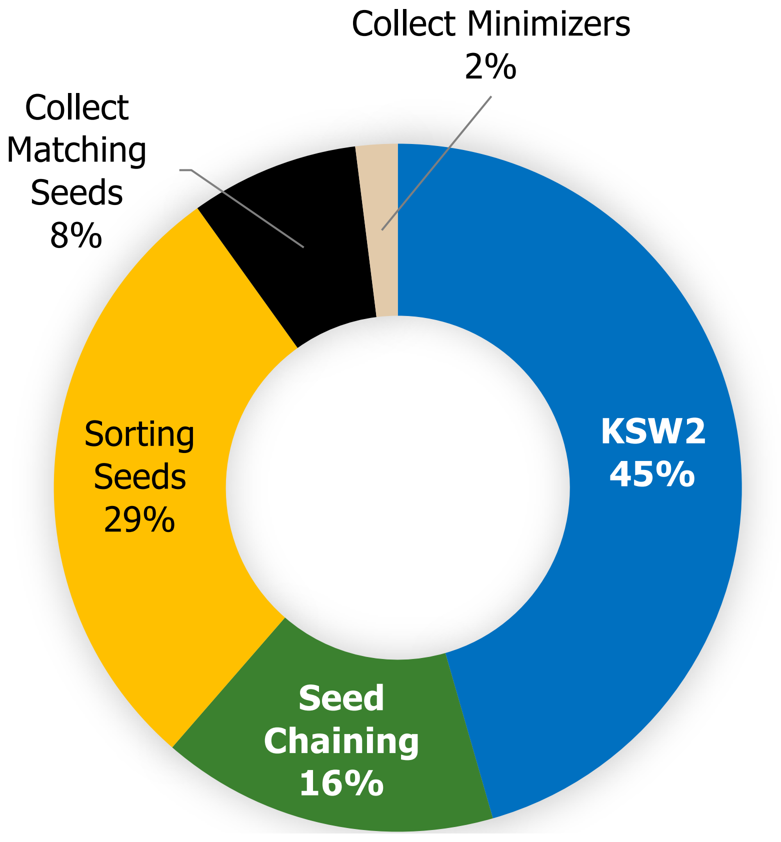


computationalgenomics.bioinformatics.ucla.edu/portfolio/david-koslicki-the-cami-project-assessment-of-computational-techniques-in-metagenomics/

Read Mapping Execution Time

>60%

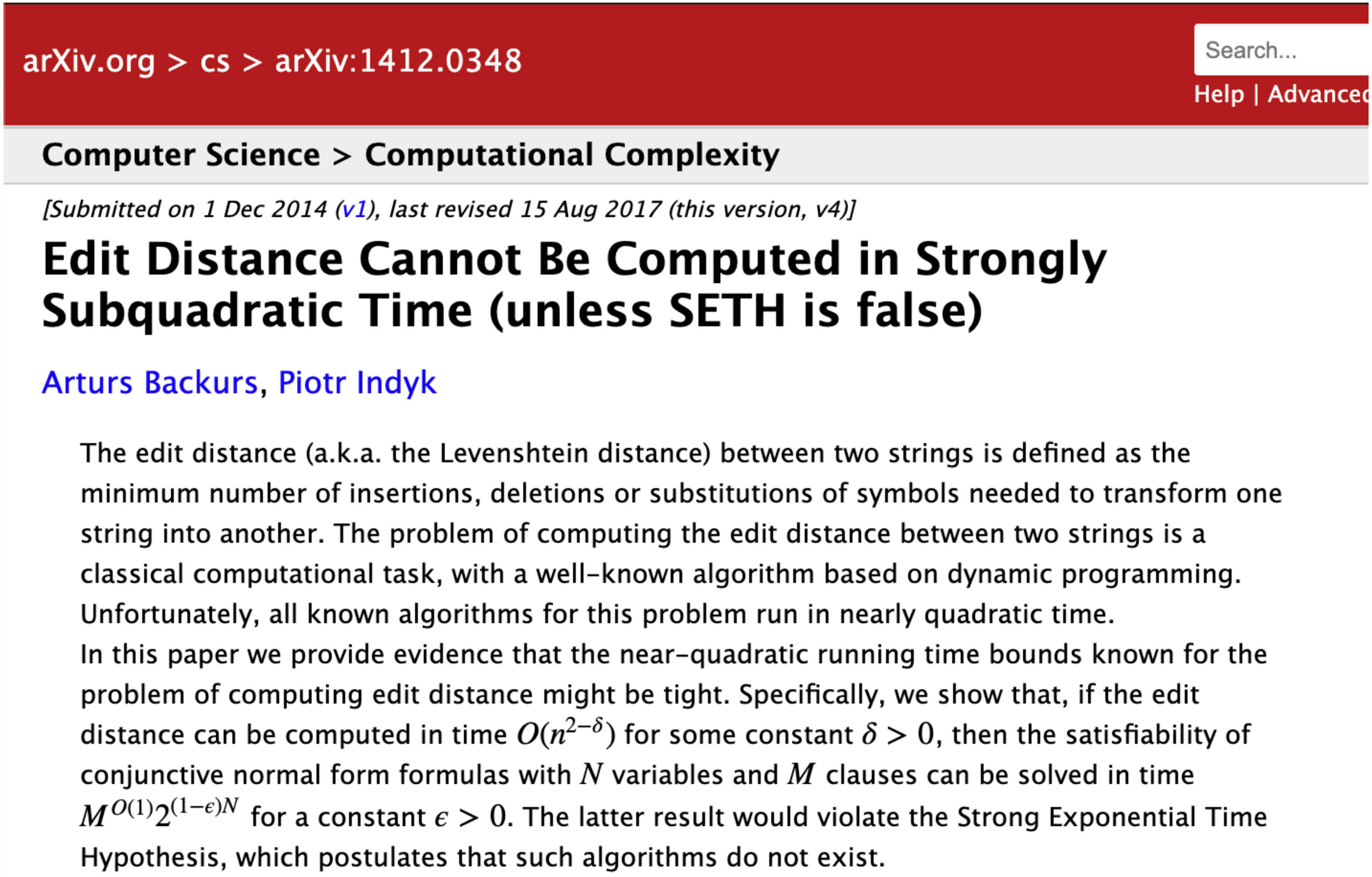
of the read mapper's execution time is spent in sequence alignment



minimap2

ONT FASTQ size: 103MB (151 reads), Mean length: 356,403 bp, std: 173,168 bp, longest length: 817,917 bp

Computational Cost is Mathematically Proven



arXiv.org > cs > arXiv:1412.0348

Computer Science > Computational Complexity

[Submitted on 1 Dec 2014 (v1), last revised 15 Aug 2017 (this version, v4)]

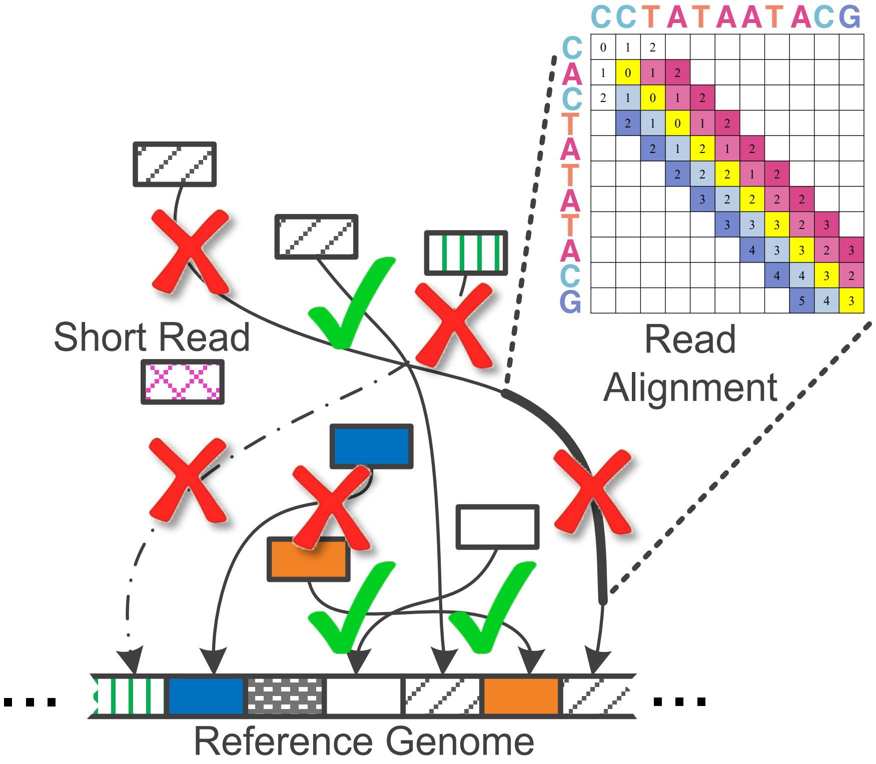
Edit Distance Cannot Be Computed in Strongly Subquadratic Time (unless SETH is false)

Arturs Backurs, Piotr Indyk

The edit distance (a.k.a. the Levenshtein distance) between two strings is defined as the minimum number of insertions, deletions or substitutions of symbols needed to transform one string into another. The problem of computing the edit distance between two strings is a classical computational task, with a well-known algorithm based on dynamic programming. Unfortunately, all known algorithms for this problem run in nearly quadratic time.
In this paper we provide evidence that the near-quadratic running time bounds known for the problem of computing edit distance might be tight. Specifically, we show that, if the edit distance can be computed in time $O(n^{2-\delta})$ for some constant $\delta > 0$, then the satisfiability of conjunctive normal form formulas with N variables and M clauses can be solved in time $M^{O(1)} 2^{(1-\epsilon)N}$ for a constant $\epsilon > 0$. The latter result would violate the Strong Exponential Time Hypothesis, which postulates that such algorithms do not exist.

https://arxiv.org/abs/1412.0348

Large Search Space for Mapping Location



	C	C	T	A	T	A	A	T	A	C	G
C	0	1	2								
A	1	0	1	2							
C	2	1	0	1	2						
T		2	1	0	1	2					
A			2	1	2	1	2				
T				2	2	2	1	2			
A					3	2	2	2	2		
T						3	3	3	2	3	
A							4	3	3	2	3
C								4	4	3	2
G									5	4	3

98%

of candidate locations have high dissimilarity with a given read

Cheng *et al*, *BMC bioinformatics* (2015)
Xin *et al*, *BMC genomics* (2013)

Agenda for Today

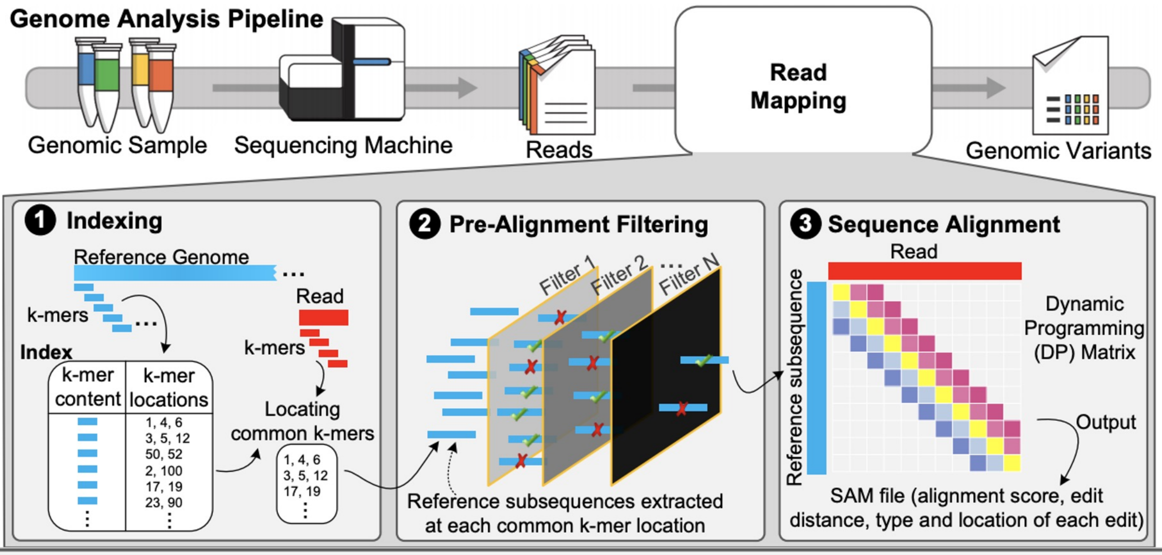
- What is Genome Analysis?
- What is Intelligent Genome Analysis?

- How we Analyze Genome?
- What are the Barriers to Enabling Intelligent Analyses?

- **Algorithmic & Hardware Acceleration**
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

- Where is Genomic Analyses Going Next?

Accelerating Read Mapping



Accelerating Indexing

Reducing the number of seeds

Reducing data movement during indexing

Accelerating Pre-Alignment Filtering

q-gram filtering

Pigeonhole principle

Base counting

Sparse DP

Accelerating Alignment

Accurate alignment accelerators

Heuristic-based alignment accelerators

Alser+, "[Accelerating Genome Analysis: A Primer on an Ongoing Journey]", IEEE Micro, 2020.

Our Contributions

Near-memory/In-memory Pre-alignment Filtering

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

In-storage Sequence Alignment

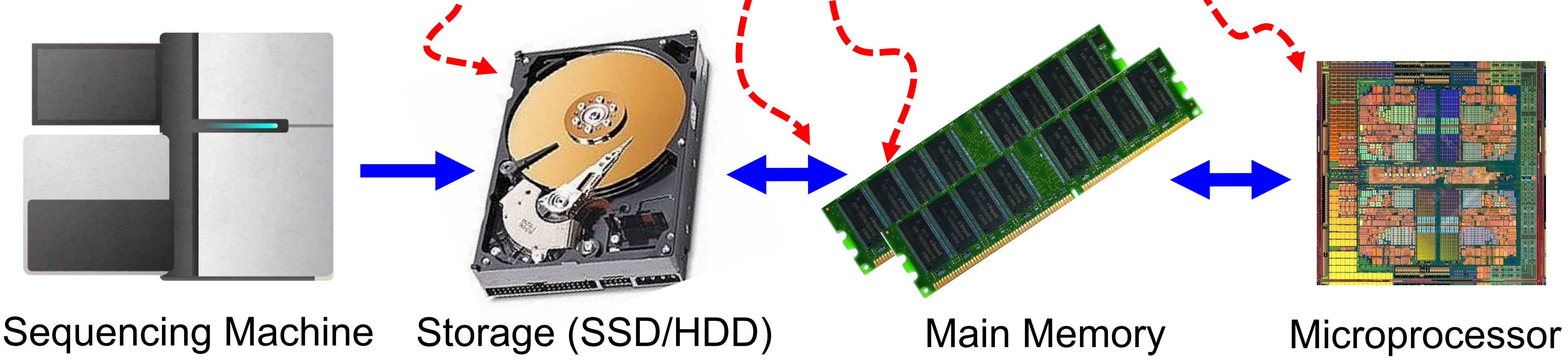
- **GenStore** [ASPLOS 2022]

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]



Ongoing Directions

- **Seed Filtering Technique:**
  - Goal: Reducing the number of seed (k-mer) locations.
    - Heuristic (limits the number of mapping locations for each seed).
    - Supports exact matches only.

- **Pre-alignment Filtering Technique:**
  - Goal: Reducing the number of *invalid mappings (>E)*.
    - Supports both exact and inexact matches.
    - Provides some falsely-accepted mappings.

- **Read Alignment Acceleration:**
  - Goal: Performing read alignment at scale.
    - Limits the numeric range of each cell in the DP table and hence supports limited scoring function.
    - May not support backtracking step due to random memory accesses.

Ongoing Directions

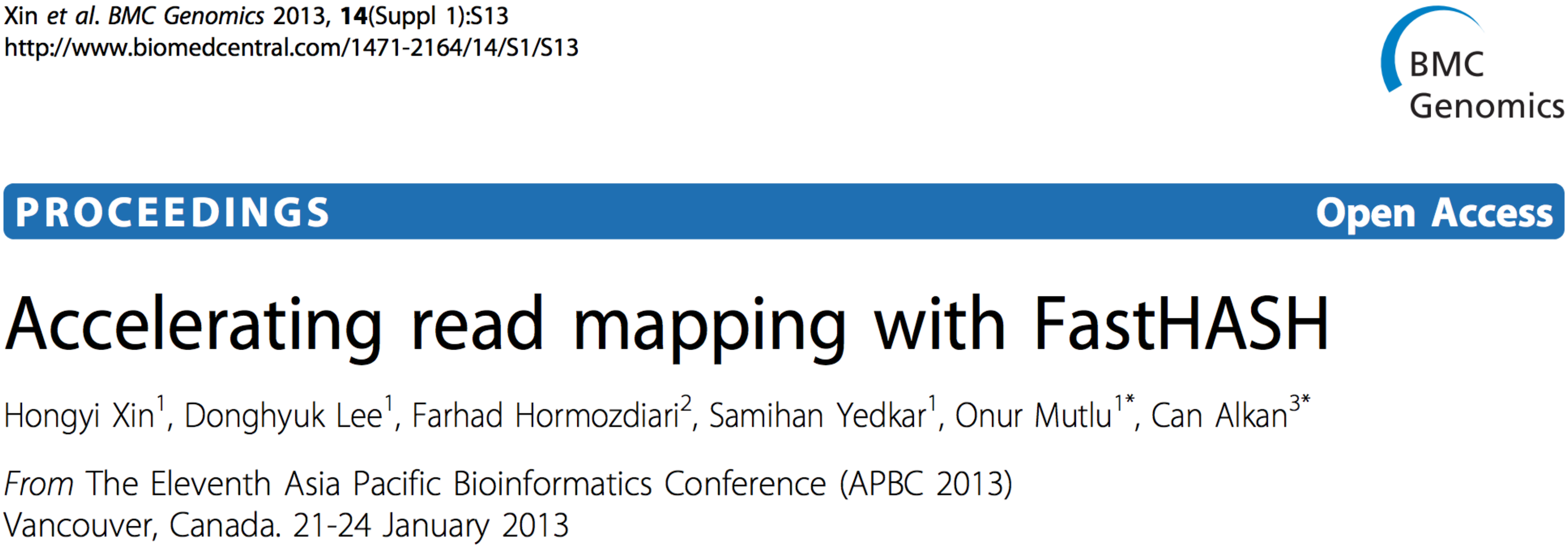
- **Seed Filtering Technique:**
  - Goal: Reducing the number of seed (k-mer) locations.
    - Heuristic (limits the number of mapping locations for each seed).
    - Supports exact matches only.

- **Pre-alignment Filtering Technique:**
  - Goal: Reducing the number of *invalid mappings (>E)*.
    - Supports both exact and inexact matches.
    - Provides some falsely-accepted mappings.

- **Read Alignment Acceleration:**
  - Goal: Performing read alignment at scale.
    - Limits the numeric range of each cell in the DP table and hence supports limited scoring function.
    - May not support backtracking step due to random memory accesses.

FastHASH

- Goal: Reducing the number of seed (k-mer) locations.
  - Heuristic (limits the number of mapping locations for each seed).
  - Supports exact matches only.



Xin *et al. BMC Genomics* 2013, **14**(Suppl 1):S13
http://www.biomedcentral.com/1471-2164/14/S1/S13

BMC Genomics

PROCEEDINGS — Open Access

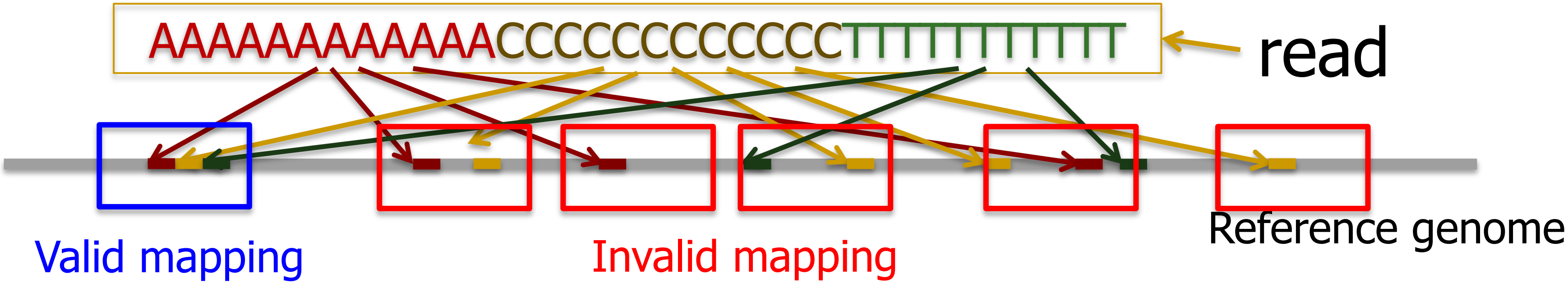
Accelerating read mapping with FastHASH

Hongyi Xin[1], Donghyuk Lee[1], Farhad Hormozdiari[2], Samihan Yedkar[1], Onur Mutlu[1*], Can Alkan[3*]

From The Eleventh Asia Pacific Bioinformatics Conference (APBC 2013)
Vancouver, Canada. 21-24 January 2013

Key Observations

- **Observation 1 (Adjacent k-mers)**
  - **Key insight:** Adjacent k-mers in the read should also be adjacent in the reference genome
  - **Key idea:** 1) sort the location list based on their number of locations and 2) search for adjacent locations in the k-mers' location lists



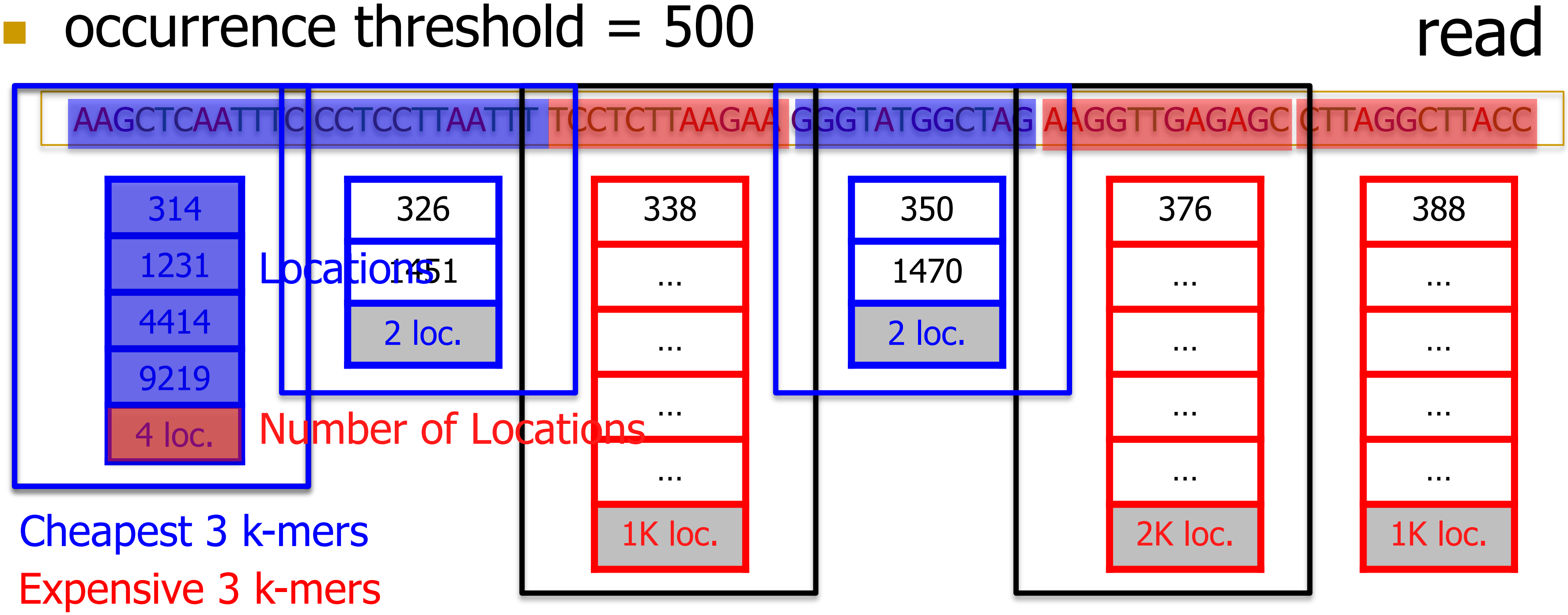
Key Observations

- **Observation 1 (Adjacent k-mers)**
  - **Key insight:** Adjacent k-mers in the read should also be adjacent in the reference genome
  - **Key idea:** 1) sort the location list based on their number of locations and 2) search for adjacent locations in the k-mers' location lists

- **Observation 2 (Cheap k-mers)**
  - **Key insight:** Some k-mers are cheaper to verify than others because they have shorter location lists (they occur less frequently in the reference genome)
  - **Key Idea:** Read mapper can choose the cheapest k-mers and verify their locations

Cheap K-mer Selection

- occurrence threshold = 500



Previous work needs to verify:

3004 locations

FastHASH verifies only:

8 locations

FastHASH Conclusion

- Problem: Existing read mappers perform poorly in mapping billions of short reads to the reference genome, in the presence of errors

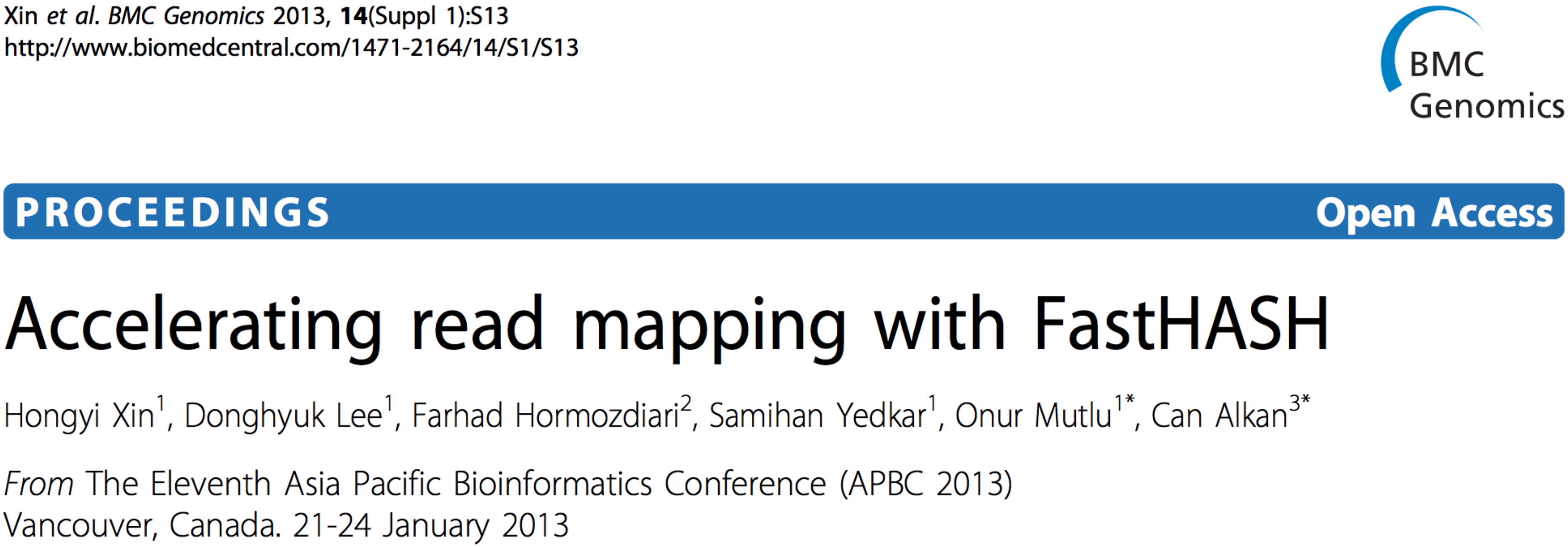
- Observation: Most of the verification calculations are unnecessary → filter them out

- Key Idea: To reduce the cost of unnecessary verification
  - Select Cheap and Adjacent k-mers.

- Key Result: FastHASH obtains up to 19x speedup over the state-of-the-art mapper without losing valid mappings

More on FastHASH

- Download source code and try for yourself
  - Download link to FastHASH



Xin *et al. BMC Genomics* 2013, **14**(Suppl 1):S13
http://www.biomedcentral.com/1471-2164/14/S1/S13

BMC Genomics

PROCEEDINGS **Open Access**

Accelerating read mapping with FastHASH

Hongyi Xin[1], Donghyuk Lee[1], Farhad Hormozdiari[2], Samihan Yedkar[1], Onur Mutlu[1*], Can Alkan[3*]

From The Eleventh Asia Pacific Bioinformatics Conference (APBC 2013)
Vancouver, Canada. 21-24 January 2013

Ongoing Directions

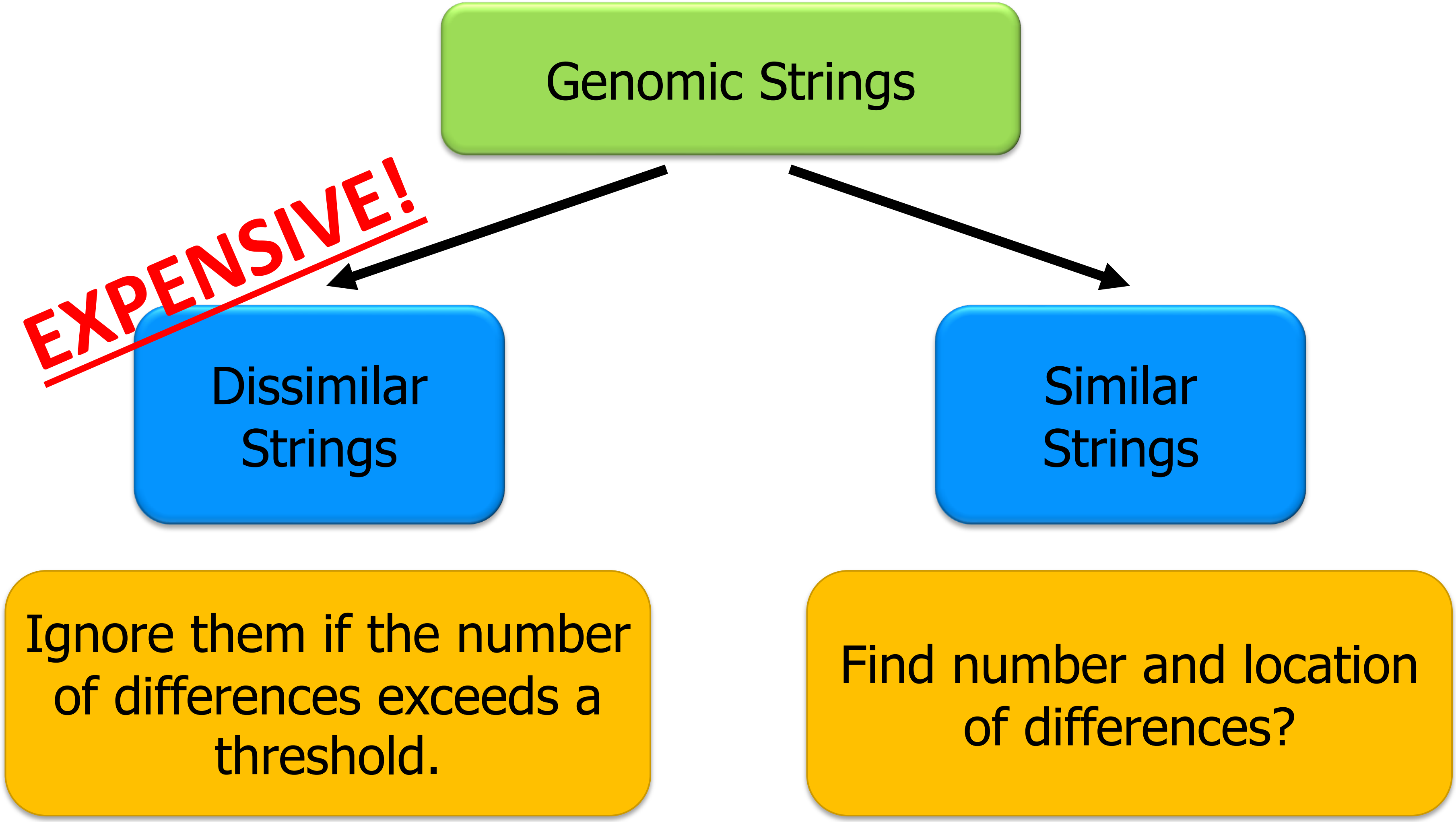
- **Seed Filtering Technique:**
  - Goal: Reducing the number of seed (k-mer) locations.
    - Heuristic (limits the number of mapping locations for each seed).
    - Supports exact matches only.
- **Pre-alignment Filtering Technique:**
  - Goal: Reducing the number of *invalid mappings (>E)*.
    - Supports both exact and inexact matches.
    - Provides some falsely-accepted mappings.
- **Read Alignment Acceleration:**
  - Goal: Performing read alignment at scale.
    - Limits the numeric range of each cell in the DP table and hence supports limited scoring function.
    - May not support backtracking step due to random memory accesses.

Pre-alignment Filtering Technique

Sequence Alignment is expensive

Our goal is to reduce the need for dynamic programming algorithms

Key Idea



Genomic Strings

EXPENSIVE!

Dissimilar Strings

Similar Strings

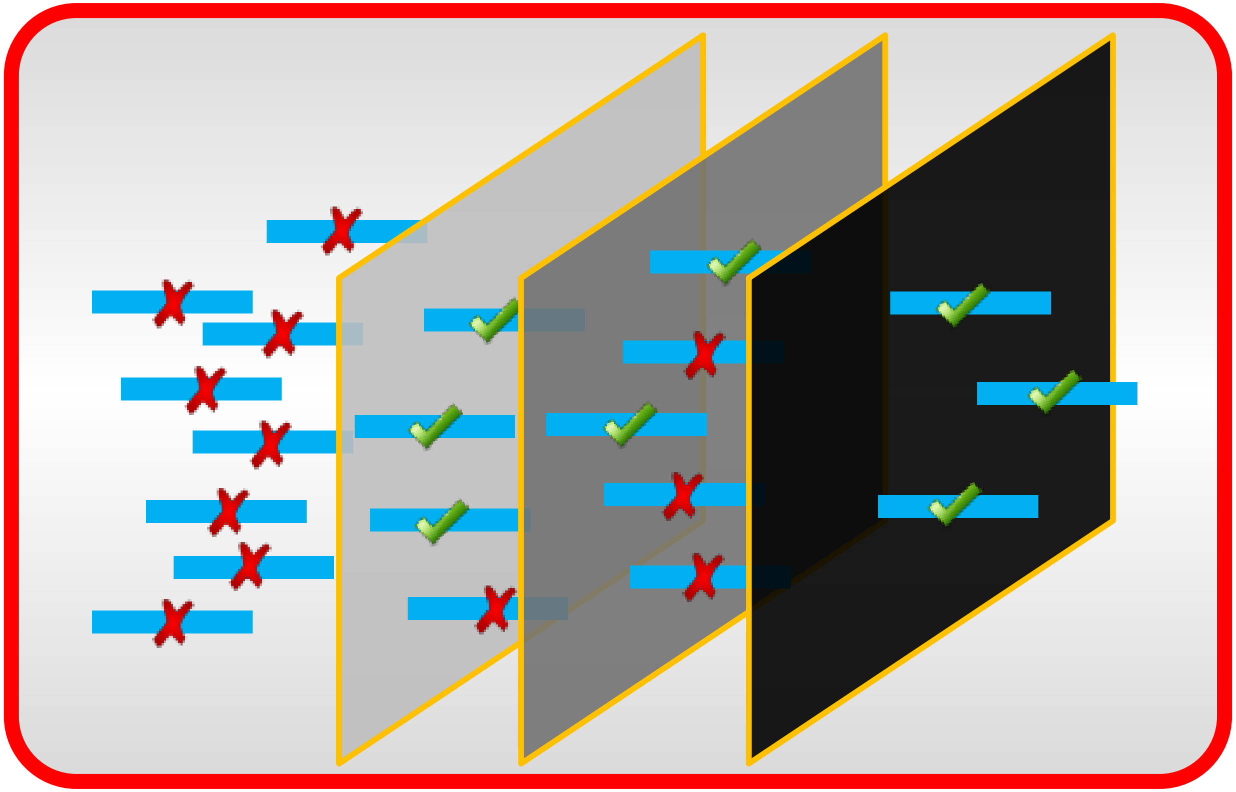
Ignore them if the number of differences exceeds a threshold.

Find number and location of differences?

Ideal Filtering Algorithm

Step 2

Query the Index

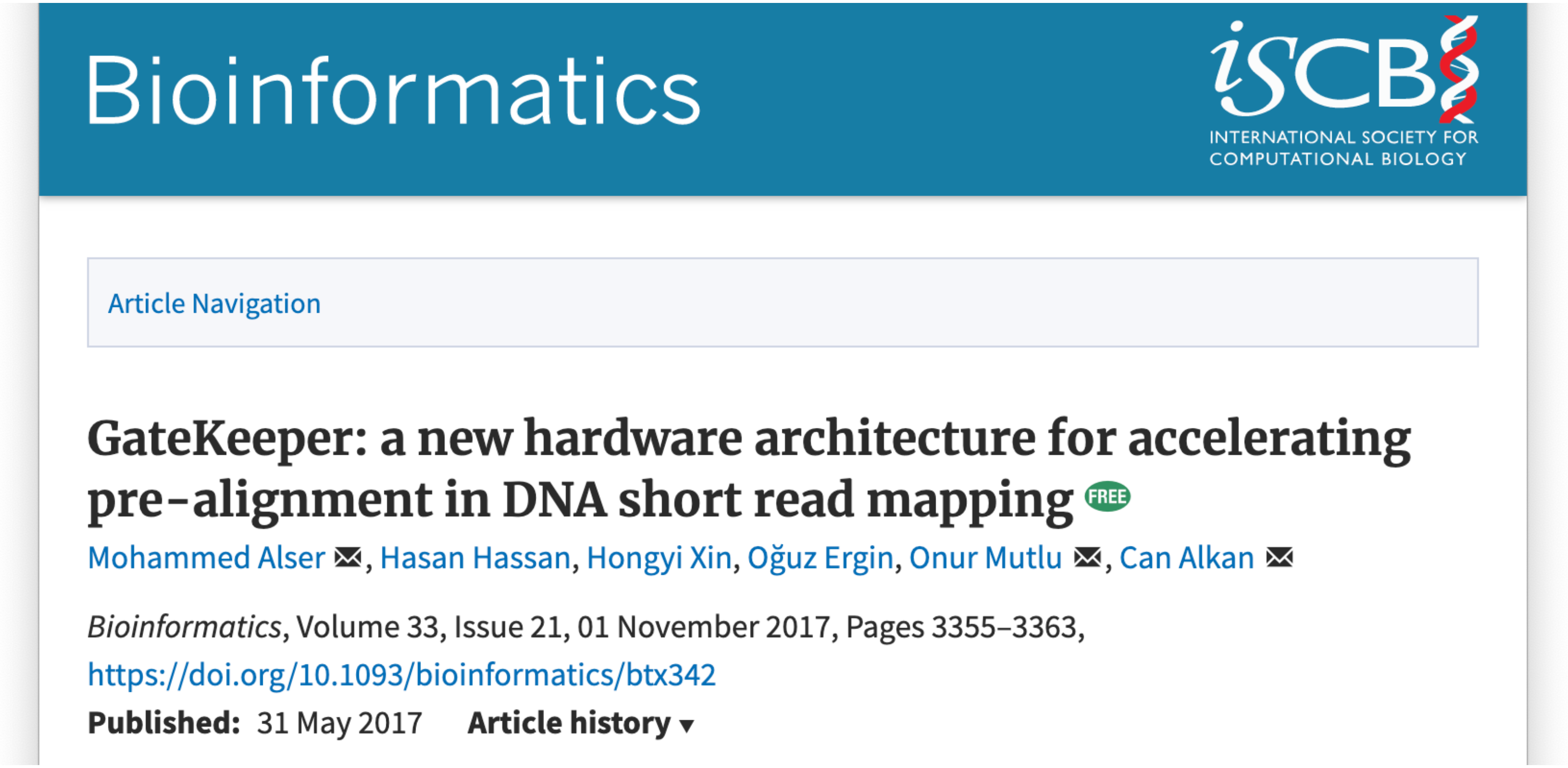


Step 3

Read Alignment

1. Filter out most of incorrect mappings.
2. Preserve all correct mappings.
3. Do it quickly.

GateKeeper



Bioinformatics

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** ▾

Alser+, "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping", Bioinformatics, 2017.

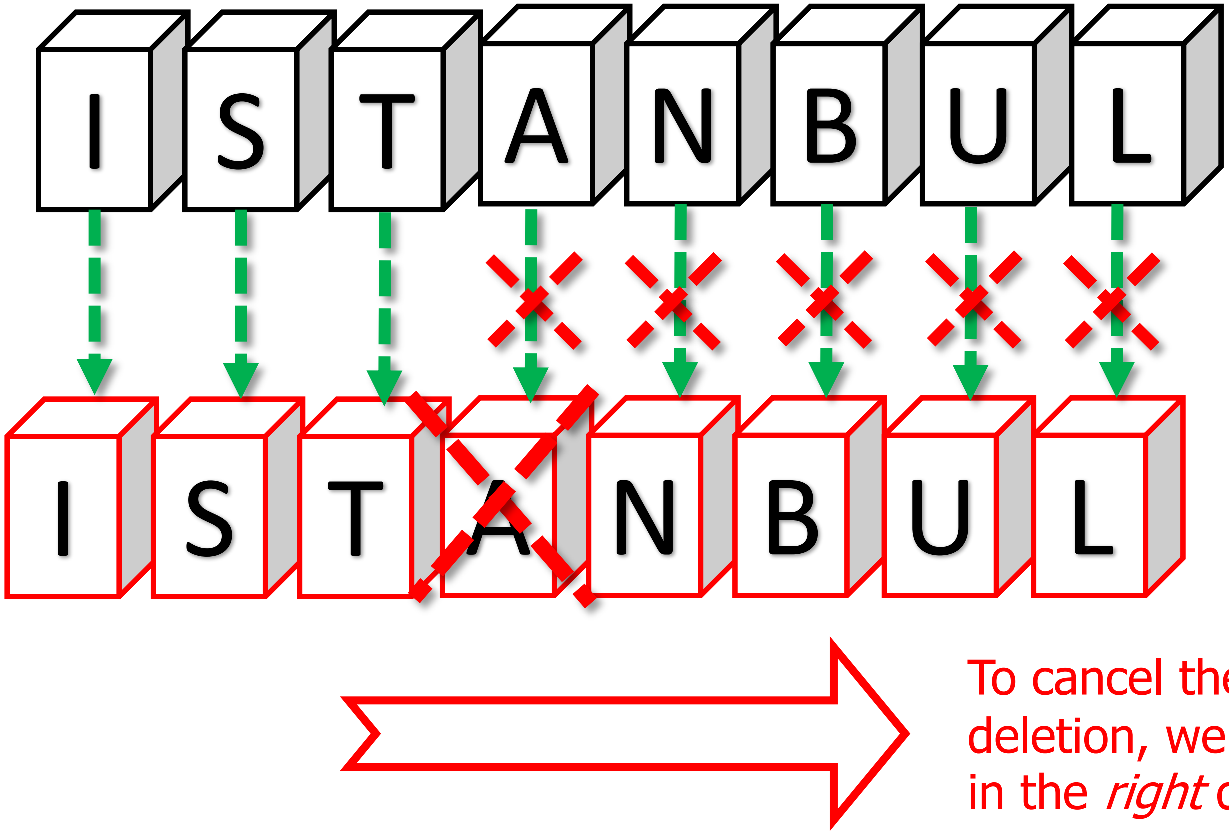
GateKeeper

- **Key observation:**
  - If two strings differ by *E* edits, then every bp match can be aligned in at most 2*E* shifts.
- **Key idea:**
  - Compute "Shifted Hamming Distance": AND of 2*E*+1 Hamming vectors of two strings, to identify invalid mappings
    - Uses *bit-parallel operations* that nicely map to FPGA architectures
- **Key result:**
  - GateKeeper is 90x-130x faster than SHD (Xin et al., 2015) and the Adjacency Filter (Xin et al., 2013), with only a 7% false positive rate
  - The addition of GateKeeper to the mrFAST mapper (Alkan et al., 2009) results in 10x end-to-end speedup in read mapping

Hamming Distance (∑⊕)

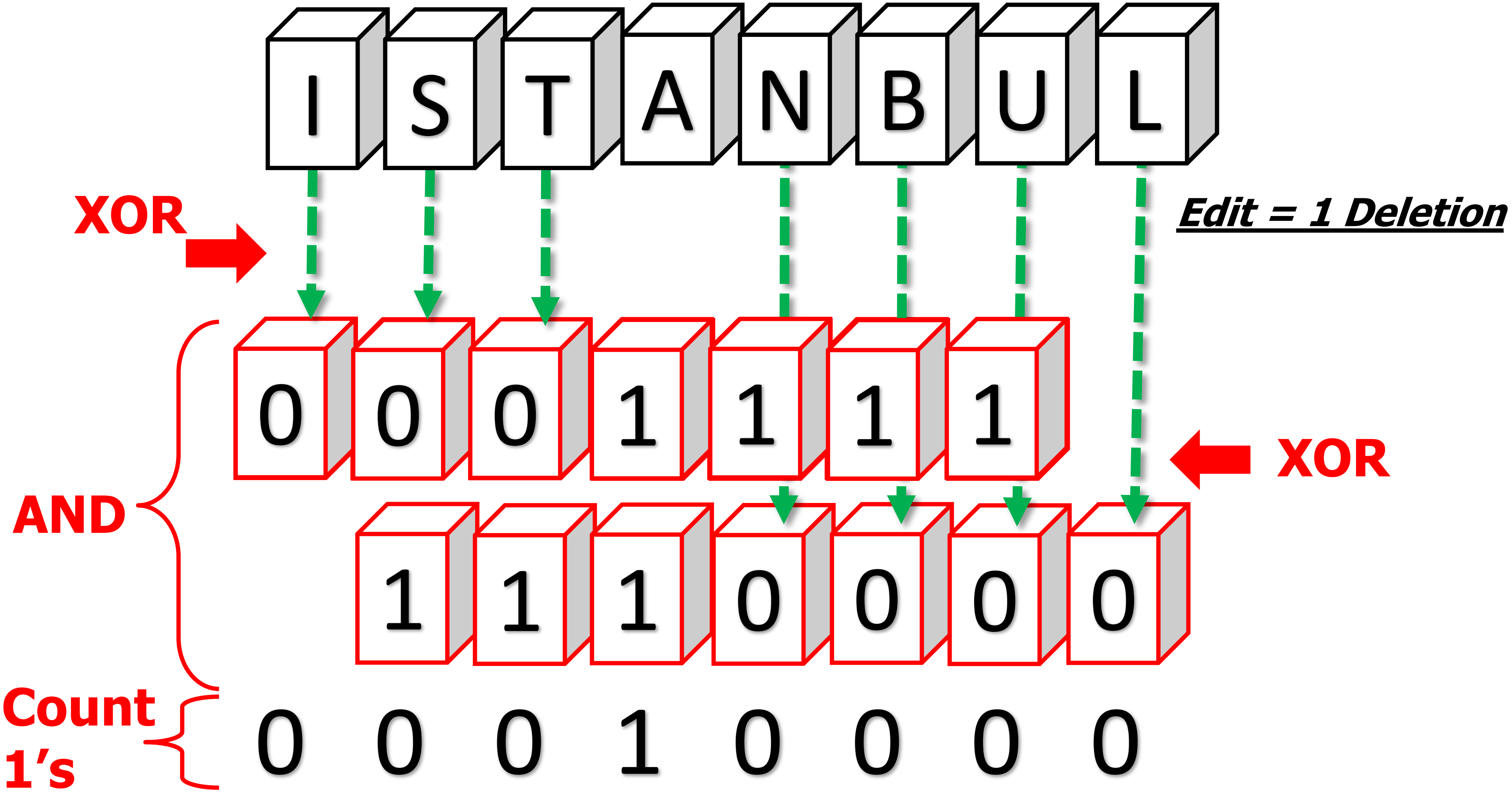
3 matches 5 mismatches

Edit = 1 Deletion



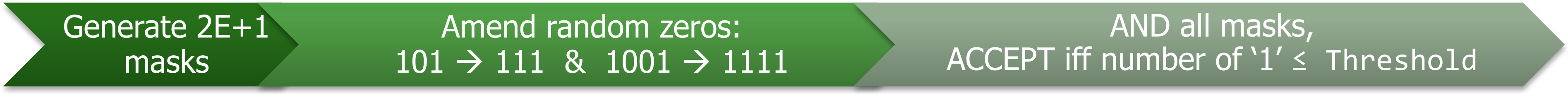
To cancel the effect of a deletion, we need to shift in the *right* direction

Shifted Hamming Distance (Xin+ 2015)



7 matches 1 mismatches

GateKeeper Walkthrough



Generate 2E+1 masks

Amend random zeros: 101 → 111 & 1001 → 1111

AND all masks, ACCEPT iff number of '1' ≤ Threshold

```
            Query :GAGAGAGATATTTAGTGTTGCAGCACTACAACACAAAAGAGGACCAACTTACGTGTCTAAAAGGGGAACATTGTTGGGCCGGA
        Reference :GAGAGAGATAGTTAGTGTTGCAGCCACTACAACACAAAAGAGGACCAACTTACGTGTCTAAAAGGGAGACATTGTTGGGCCGG

     Hamming Mask :000000000010000000000000111111101111000111011010110111111111000100000111101101001010 1
  1-Deletion Mask :111111111110011111011111000000000000000000000000000000000000000000011000000000000000 0
  2-Deletion Mask :000000001011011100111111111111101111000111011010110111111111000100010011101101001010
  3-Deletion Mask :111111111110111011001101110111011000100100111111111111110010110011010110111011101111
 1-Insertion Mask :111111111110111110111111011101100010010011111111111110010110011000101001110111011110
 2-Insertion Mask :000000100111110011111111100100011010101001101011111111111111011100111111100011101100
 3-Insertion Mask :111111110111011001100011111111101011011111100110010111011111111011101111010111001000

         AND Mask :000000000010000000000000100000000000000000000000000000000000000000000000000000000000
```

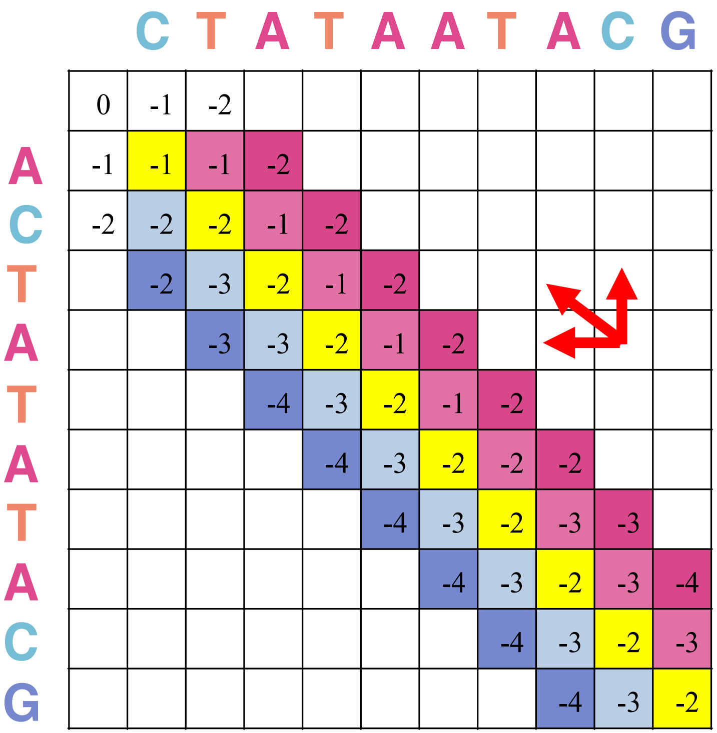
```
1-D                                      .111
2-D                                      0000
3-D                                      .110
1-In                                     .111
2-In                                     .110
3-In                                     .100
                                         .000
                                         0000
```

Our goal to track the diagonally consecutive matches in the neighborhood map.

```
Needleman-Wunsch   GAGAGAGATATTTAGTGTTGCAG-CACTACAACACAAAAGAGGACCAACTTACGTGTCTAAAAGGGGAACATTGTTGGGCCGG
       Alignment  :|||||||||| |||||||||||| |||||||||||||||||||||||||||||||||||||||||||::|||||||||||||||
                   GAGAGAGATAGTTAGTGTTGCAGCCACTACAACACAAAAGAGGACCAACTTACGTGTCTAAAAGGGAGACATTGTTGGGCCGG
```

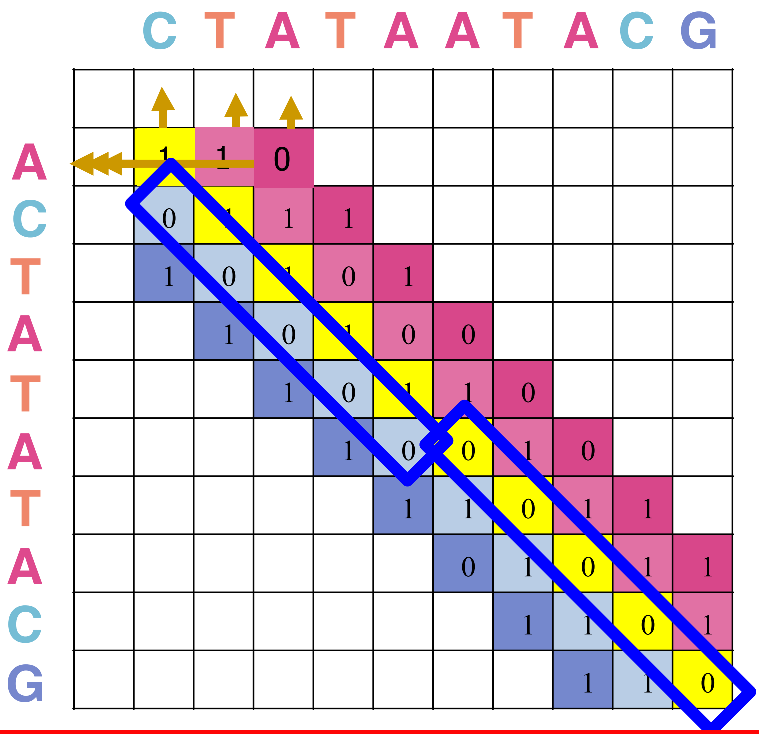
Alignment Matrix vs. Neighborhood Map

Needleman-Wunsch



		C	T	A	T	A	A	T	A	C	G
	0	-1	-2								
A	-1	-1	-1	-2							
C	-2	-2	-2	-1	-2						
T		-2	-3	-2	-1	-2					
A			-3	-3	-2	-1	-2				
T				-4	-3	-2	-1	-2			
A					-4	-3	-2	-2	-2		
T						-4	-3	-2	-3	-3	
A							-4	-3	-2	-3	-4
C								-4	-3	-2	-3
G									-4	-3	-2

Neighborhood Map

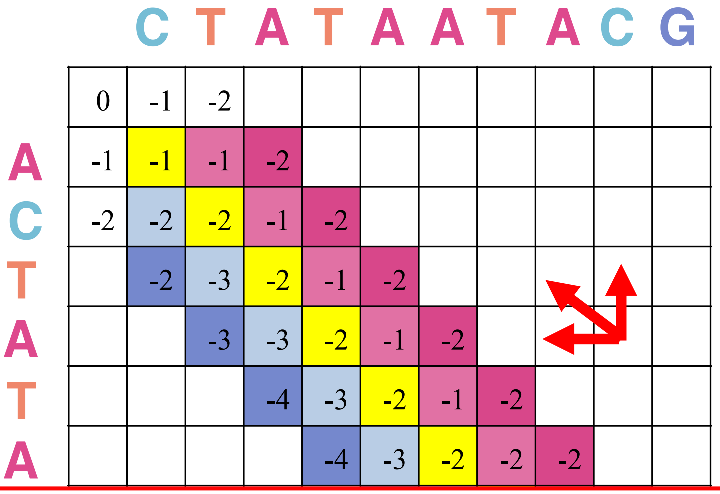


		C	T	A	T	A	A	T	A	C	G
A		1	1	0							
C		0	1	1	1						
T		1	0	1	0	1					
A			1	0	1	0	0				
T				1	0	1	1	0			
A					1	0	0	1	0		
T						1	1	0	1	1	
A							0	1	0	1	1
C								1	1	0	1
G									1	1	0

Our goal to track the diagonally consecutive matches in the neighborhood map.

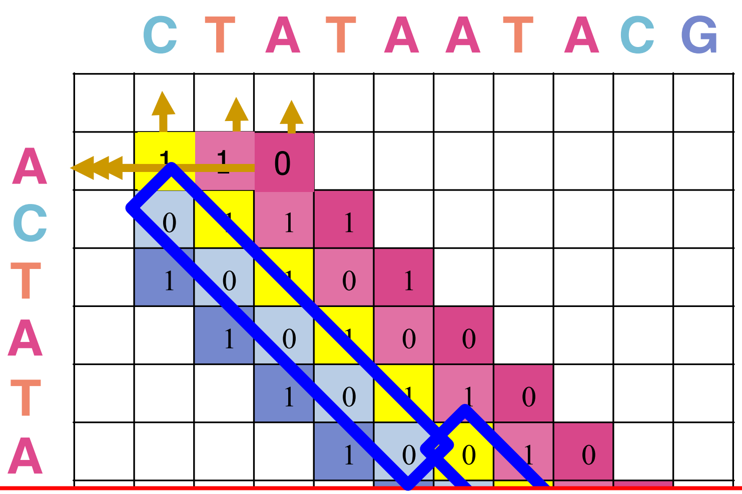
Alignment Matrix vs. Neighborhood Map

Needleman-Wunsch



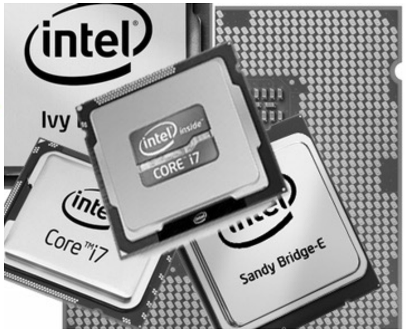
		C	T	A	T	A	A	T	A	C	G
	0	-1	-2								
A	-1	-1	-1	-2							
C	-2	-2	-2	-1	-2						
T		-2	-3	-2	-1	-2					
A			-3	-3	-2	-1	-2				
T				-4	-3	-2	-1	-2			
A					-4	-3	-2	-2	-2		

Neighborhood Map



		C	T	A	T	A	A	T	A	C	G
A		1	1	0							
C		0	1	1	1						
T		1	0	1	0	1					
A			1	0	1	0	0				
T				1	0	1	1	0			
A					1	0	0	1	0		

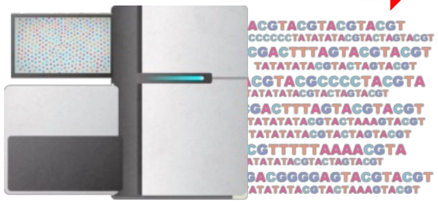
Independent vectors can be processed in parallel using hardware technologies



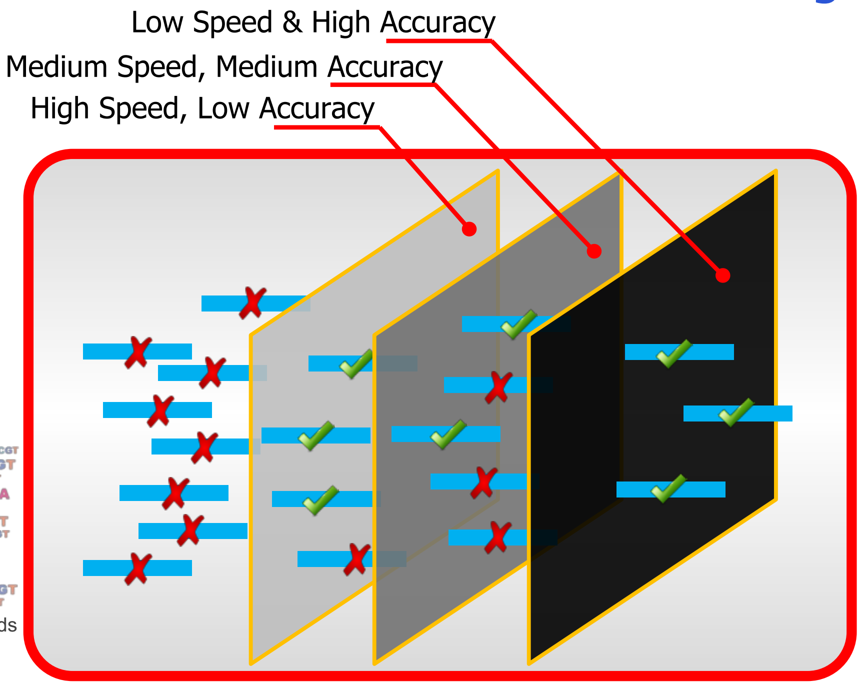
Our Solution: GateKeeper

Alignment Filter + = 1st FPGA-based Alignment Filter.

x10^{12} mappings



Billions of Short Reads



Low Speed & High Accuracy

Medium Speed, Medium Accuracy

High Speed, Low Accuracy

x10^{3} mappings

		C	T	A	T	A	A	T	A	C	G
C	0	1	2								
A	1	0	1	2							
C	2	1	0	1	2						
T		2	1	0	1	2					
A			2	1	2	1	2				
T				2	2	2	1	2			
A					3	2	2	2	2		
T						3	3	3	2	3	
A							4	3	3	2	3
C								4	4	3	2
G									5	4	3

1 **High throughput DNA sequencing (HTS) technologies**

2 **Read Pre-Alignment Filtering**
Fast & Low False Positive Rate

3 **Read Alignment**
Slow & Zero False Positives

GateKeeper Walkthrough (cont'd)

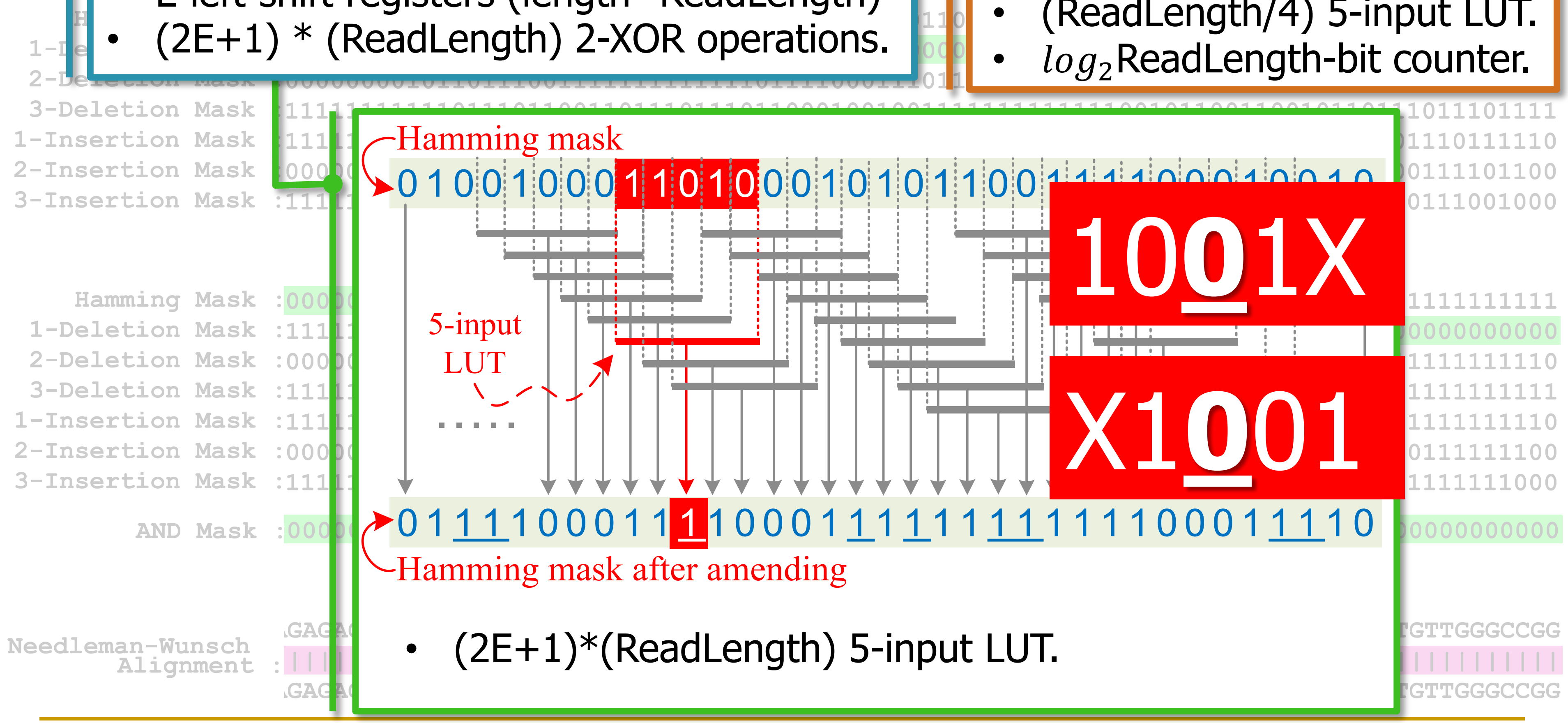
Generate 2E+1 masks

Amend random zeros: 101 → 111 & 1001 → 1111

AND all masks, ACCEPT iff number of '1' ≤ Threshold

- E right-shift registers (length=ReadLength)
- E left-shift registers (length=ReadLength)
- (2E+1) * (ReadLength) 2-XOR operations.

- (2E)*(ReadLength) 2-AND operations.
- (ReadLength/4) 5-input LUT.
- log_2ReadLength-bit counter.



- (2E+1)*(ReadLength) 5-input LUT.

Virtex-7 FPGA Layout

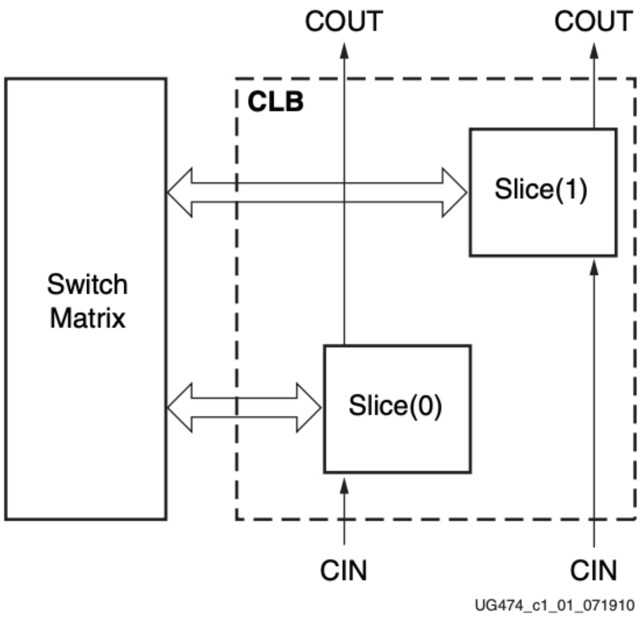
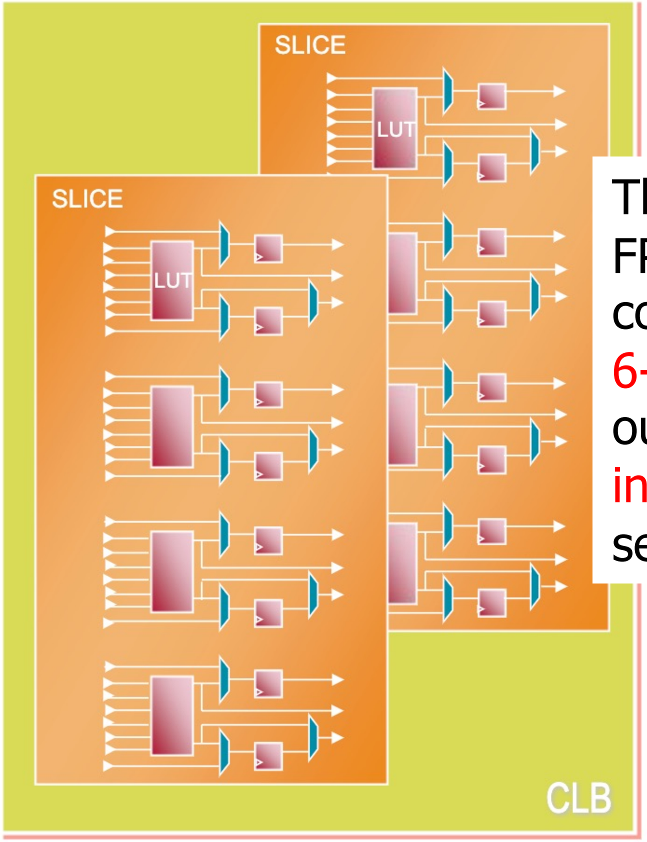
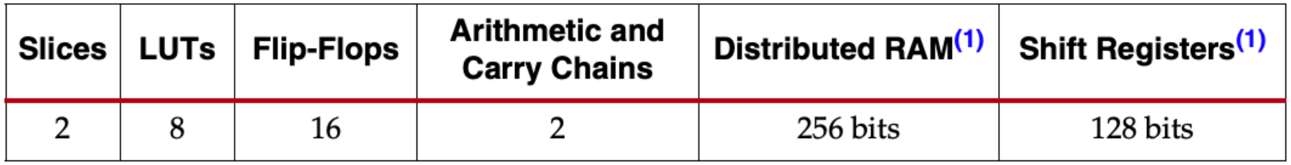


Figure 1-1: **Arrangement of Slices within the CLB**



The LUTs in 7 series FPGAs can be configured as either a 6-input LUT with one output, or as two 5-input LUTs with separate outputs

Table 2-1: **Logic Resources in One CLB**



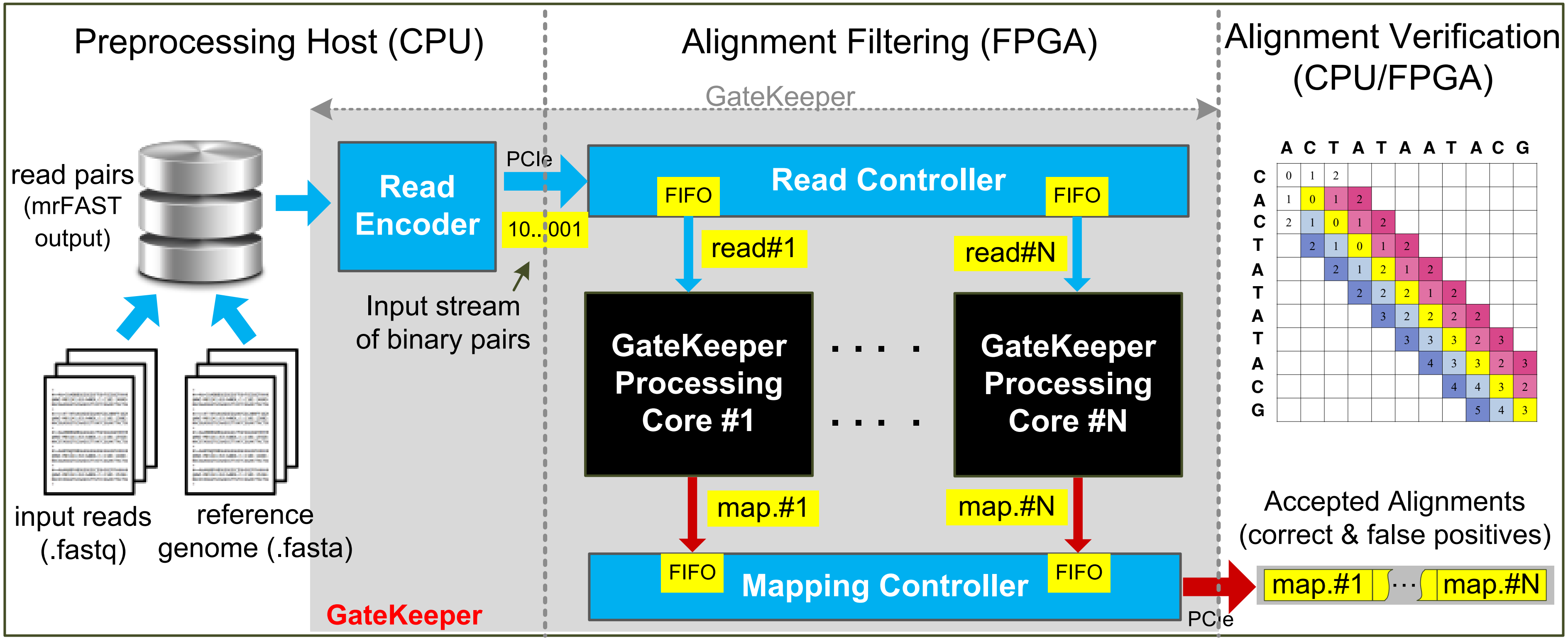
Slices	LUTs	Flip-Flops	Arithmetic and Carry Chains	Distributed RAM[(1)]	Shift Registers[(1)]
2	8	16	2	256 bits	128 bits

SAFARI

"7 Series FPGAs Configurable Logic Block", User Guide, Xilinx 2016

GateKeeper Accelerator Architecture

- **Maximum data throughput** =~13.3 billion bases/sec
  - Can examine **8 (300 bp) or 16 (100 bp) mappings concurrently** at 250 MHz
  - **Occupies 50%** (100 bp) to **91%** (300 bp) of the FPGA slice LUTs and registers



Preprocessing Host (CPU)

Alignment Filtering (FPGA)

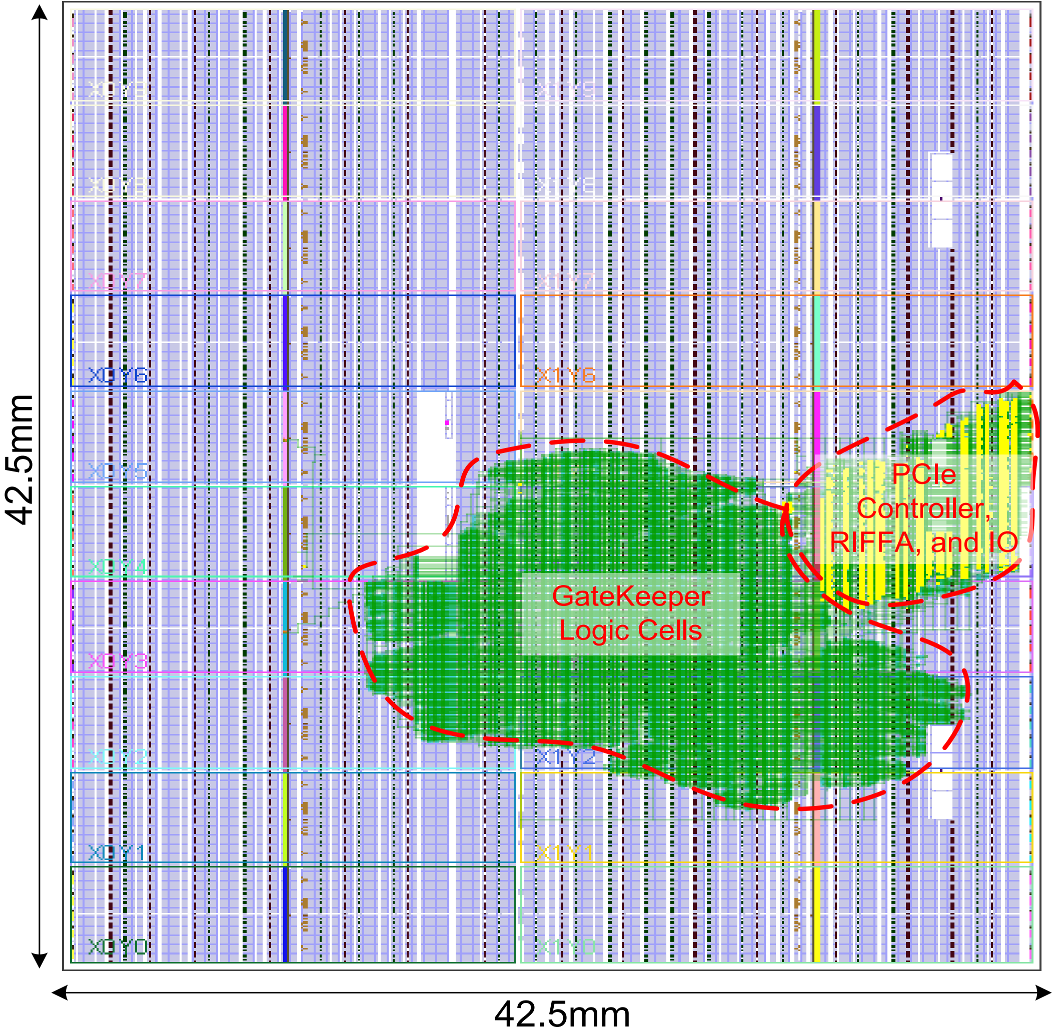
Alignment Verification (CPU/FPGA)

GateKeeper
read pairs (mrFAST output)
input reads (.fastq)
reference genome (.fasta)
Read Encoder
PCIe
10...001
Input stream of binary pairs
Read Controller
FIFO
FIFO
read#1
read#N
GateKeeper Processing Core #1
. . . .
GateKeeper Processing Core #N
map.#1
map.#N
FIFO
FIFO
Mapping Controller
GateKeeper
PCIe
Accepted Alignments (correct & false positives)
map.#1 … map.#N

	A	C	T	A	T	A	A	T	A	C	G
C	0	1	2								
A	1	0	1	2							
C	2	1	0	1	2						
T		2	1	0	1	2					
A			2	1	2	1	2				
T				2	2	2	1	2			
A					3	2	2	2	2		
T						3	3	3	2	3	
A							4	3	3	2	3
C								4	4	3	2
G									5	4	3

FPGA Chip Layout

GateKeeper: 17.6%, PCIe Controller, RIFFA, and IO: 5%



300 bp

E=15

GateKeeper: Speed & Accuracy Results

90x-130x faster filter

than SHD (Xin et al., 2015) and the Adjacency Filter (Xin et al., 2013)

4x lower false accept rate

than the Adjacency Filter (Xin et al., 2013)

10x speedup in read mapping

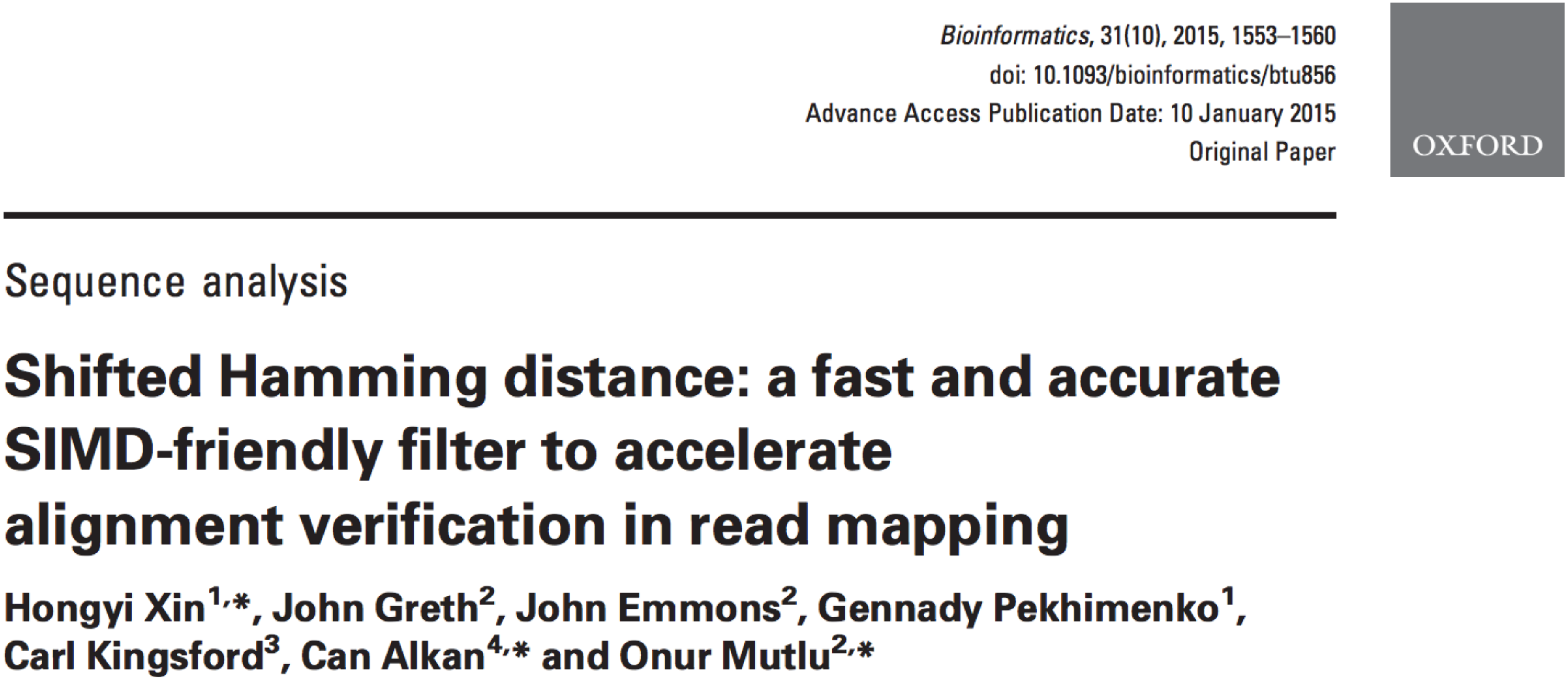
with the addition of GateKeeper to the mrFAST mapper (Alkan et al., 2009)

Freely available online

github.com/BilkentCompGen/GateKeeper

More on SHD (SIMD Implementation)

- Download and test for yourself
- https://github.com/CMU-SAFARI/Shifted-Hamming-Distance



Bioinformatics, 31(10), 2015, 1553–1560
doi: 10.1093/bioinformatics/btu856
Advance Access Publication Date: 10 January 2015
Original Paper

OXFORD

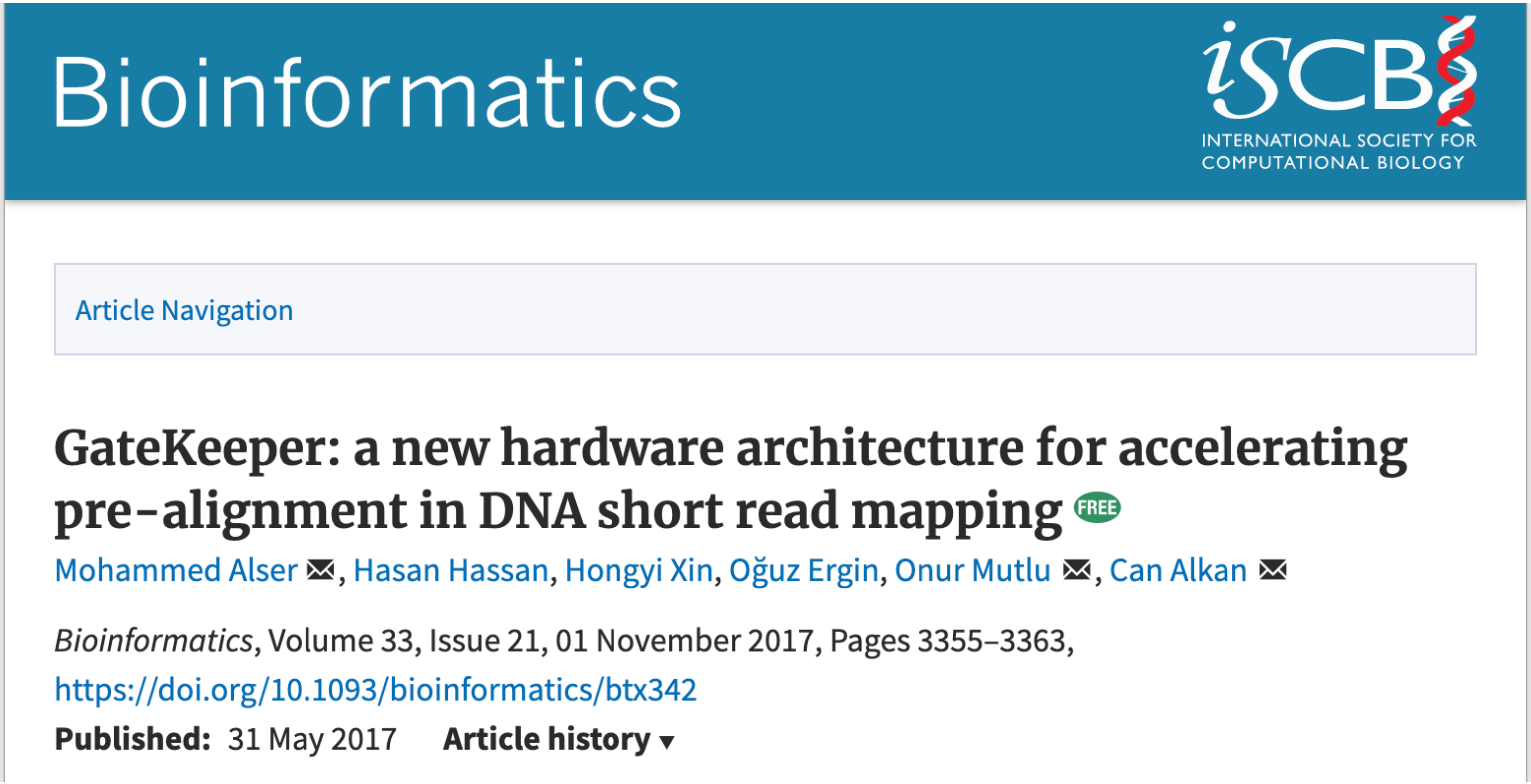
Sequence analysis

Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping

Hongyi Xin[1,*], John Greth[2], John Emmons[2], Gennady Pekhimenko[1], Carl Kingsford[3], Can Alkan[4,*] and Onur Mutlu[2,*]

More on GateKeeper

- Download and test for yourself
  https://github.com/BilkentCompGen/GateKeeper



Bioinformatics

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

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

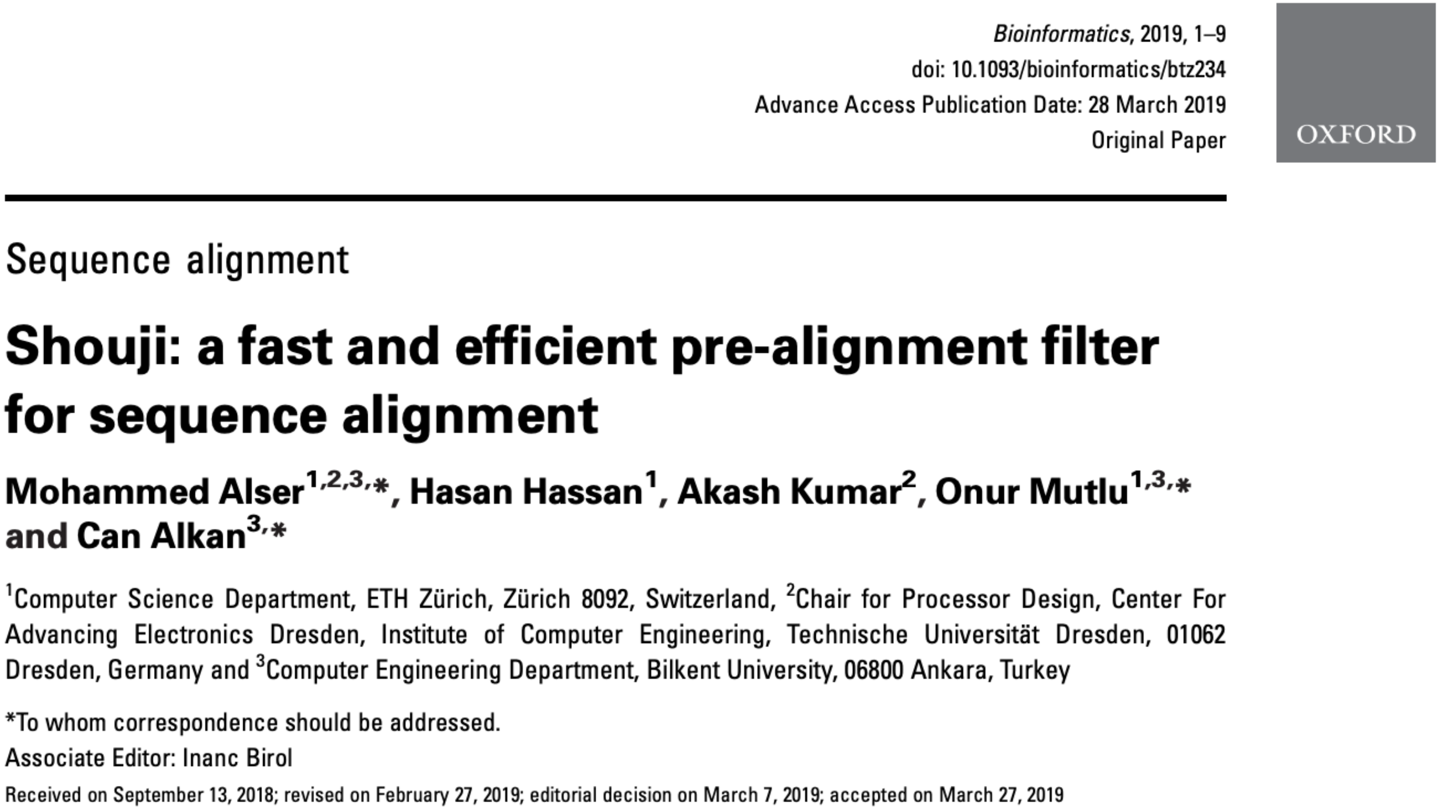
Bioinformatics, Volume 33, Issue 21, 01 November 2017, Pages 3355–3363, https://doi.org/10.1093/bioinformatics/btx342

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

Alser+, "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping", Bioinformatics, 2017.

Can we do better? Scalability?

Shouji (障子)



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 Hassan[1], Akash Kumar[2], Onur Mutlu[1,3,*] and Can Alkan[3,*]

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

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on September 13, 2018; revised on February 27, 2019; editorial decision on March 7, 2019; accepted on March 27, 2019

Alser+, "Shouji: a fast and efficient pre-alignment filter for sequence alignment", *Bioinformatics* 2019,
https://doi.org/10.1093/bioinformatics/btz234

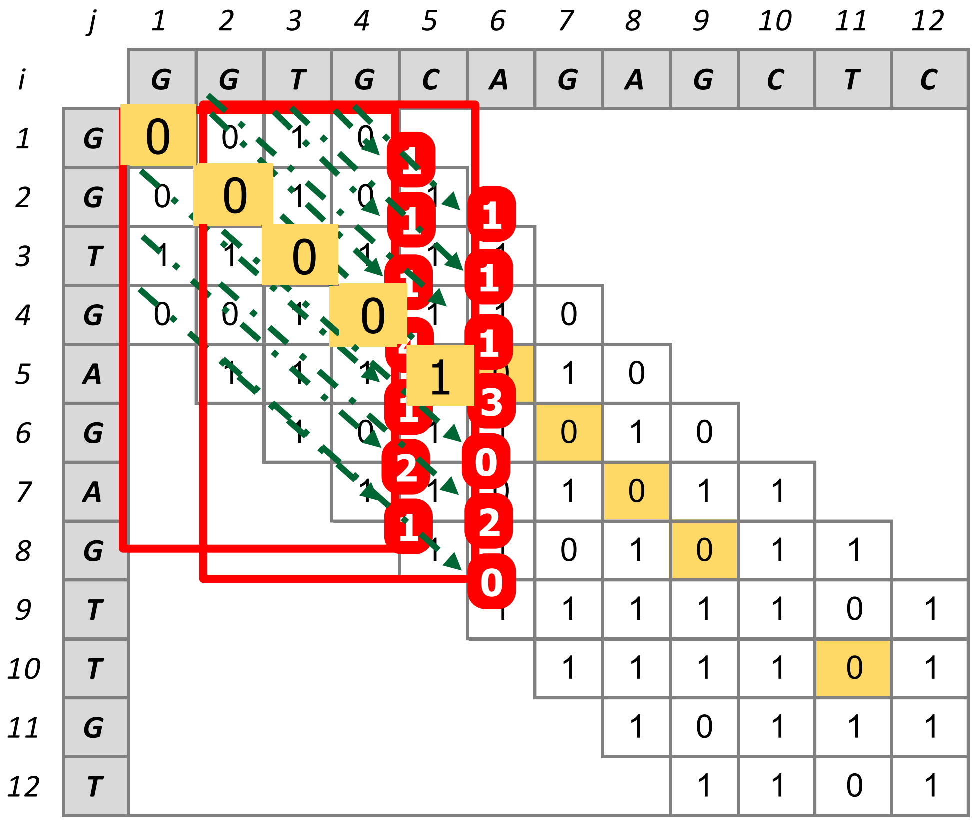
Shouji

- **Key observation:**
  - Correct alignment always includes long identical subsequences.
  - Processing the entire mapping at once is ineffective for hardware design.
- **Key idea:**
  - Use overlapping sliding window approach to quickly and accurately find all long segments of consecutive zeros.
- **Key result:**
  - Shouji on FPGA is up to three orders of magnitude faster than its CPU implementation.
  - Shouji accelerates best-performing CPU read aligner Edlib (Bioinformatics 2017) by up to 18.8x using 16 filtering units that work in parallel.
  - Shouji is 2.4x to 467x more accurate than GateKeeper (Bioinformatics 2017) and SHD (Bioinformatics 2015).

Shouji Walkthrough

Building the Neighborhood Map

Finding all common subsequences (diagonal segments of consecutive zeros) shared between two given sequences.



Storing it @ Shouji Bit-vector

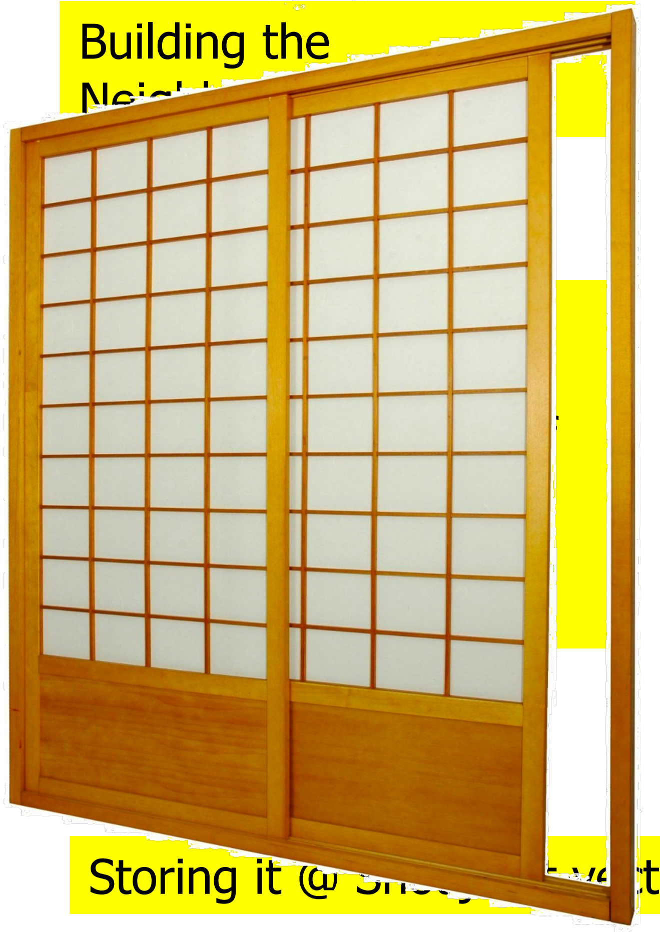
0	0	0	0	**1**	0	0	0	0	**1**	0	**1**

ACCEPT iff number of '1' ≤ Threshold

Shouji: a fast and efficient pre-alignment filter for sequence alignment, *Bioinformatics* 2019, https://doi.org/10.1093/bioinformatics/btz234

Shouji Walkthrough

Building the Neigh



Storing it @ Shouji bit-vector

j		1	2	3	4	5	6	7	8	9	10	11	12
i		G	G	T	G	C	A	G	A	G	C	T	C
1	G	0	0	1	0								
2	G	0	0	1	0	1							
3	T	1	1	0	1	1	1						
4	G	0	0	1	0	1	1	0					
5	A		1	1	1	1	0	1	0				
6	G			1	0	1	1	0	1	0			
7	A				1	1	0	1	0	1	1		
8	G					1	1	0	1	0	1	1	
9	T						1	1	1	1	1	0	1
10	T							1	1	1	1	0	1
11	G								1	0	1	1	1
12	T									1	1	0	1

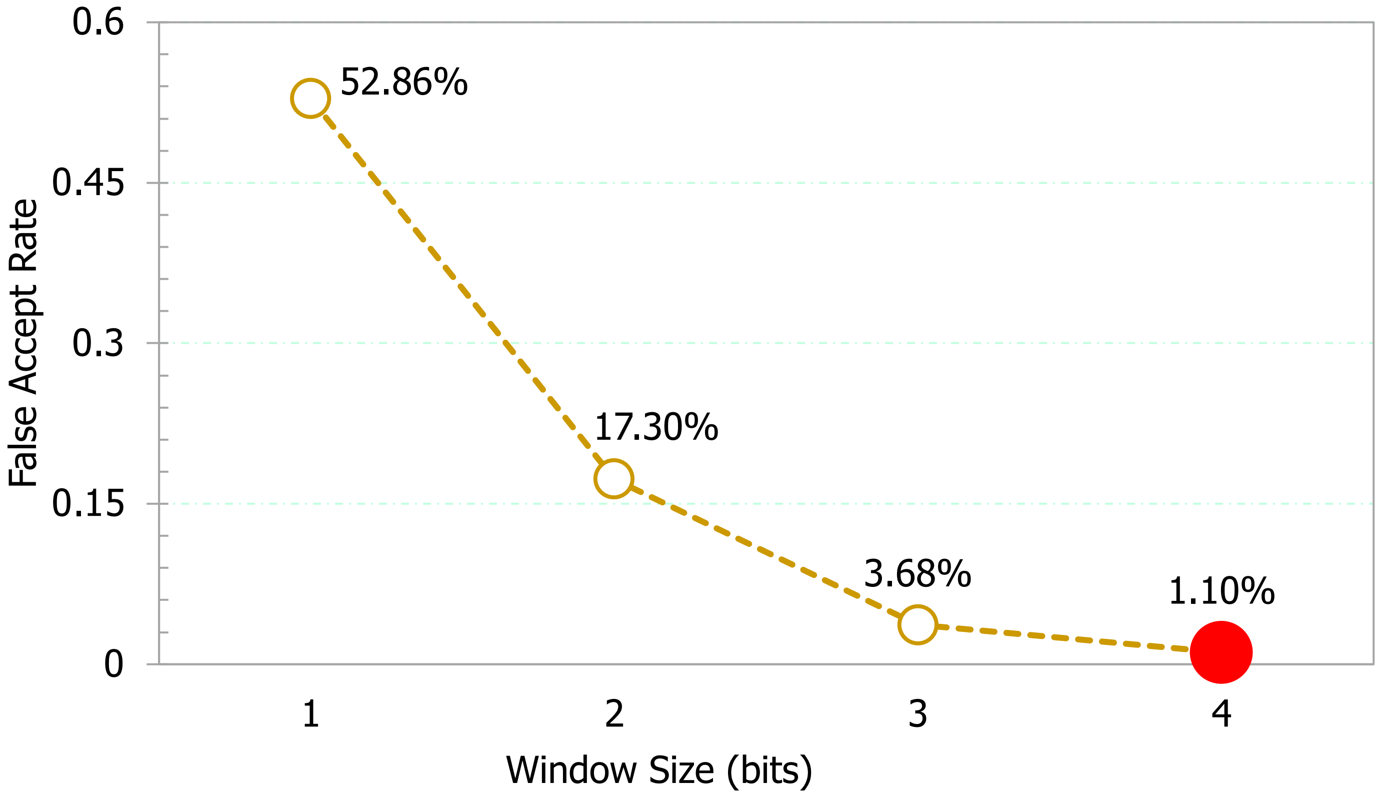
0	0	0	0	**1**	0	0	0	0	**1**	0	**1**

ACCEPT iff number of '1' ≤ Threshold

Shouji: a fast and efficient pre-alignment filter for sequence alignment, *Bioinformatics* 2019, https://doi.org/10.1093/bioinformatics/btz234

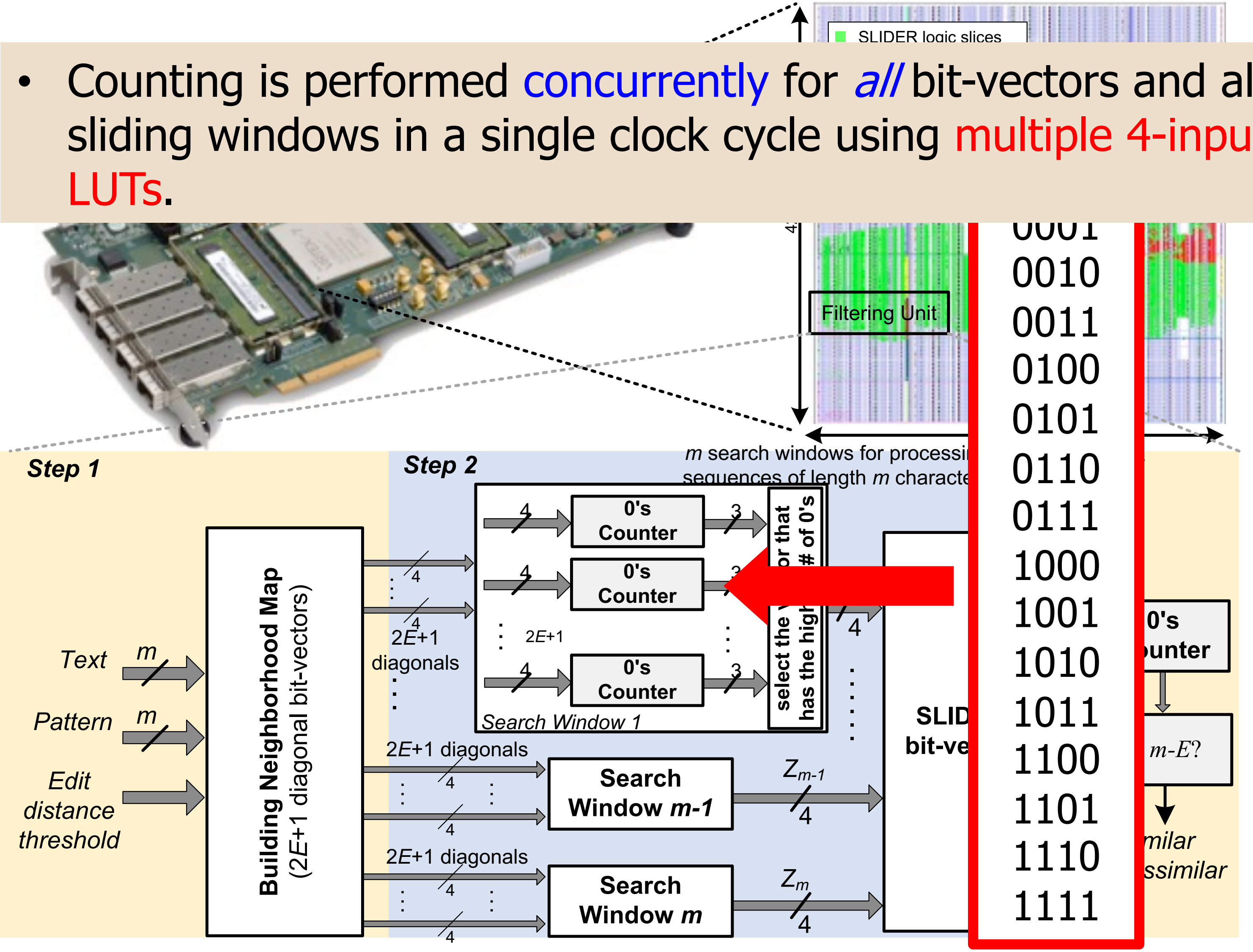
Sliding Window Size

- The reason behind the selection of the window size is due to the minimal possible length of the identical subsequence that is a single match (e.g., such as `101').



Hardware Implementation

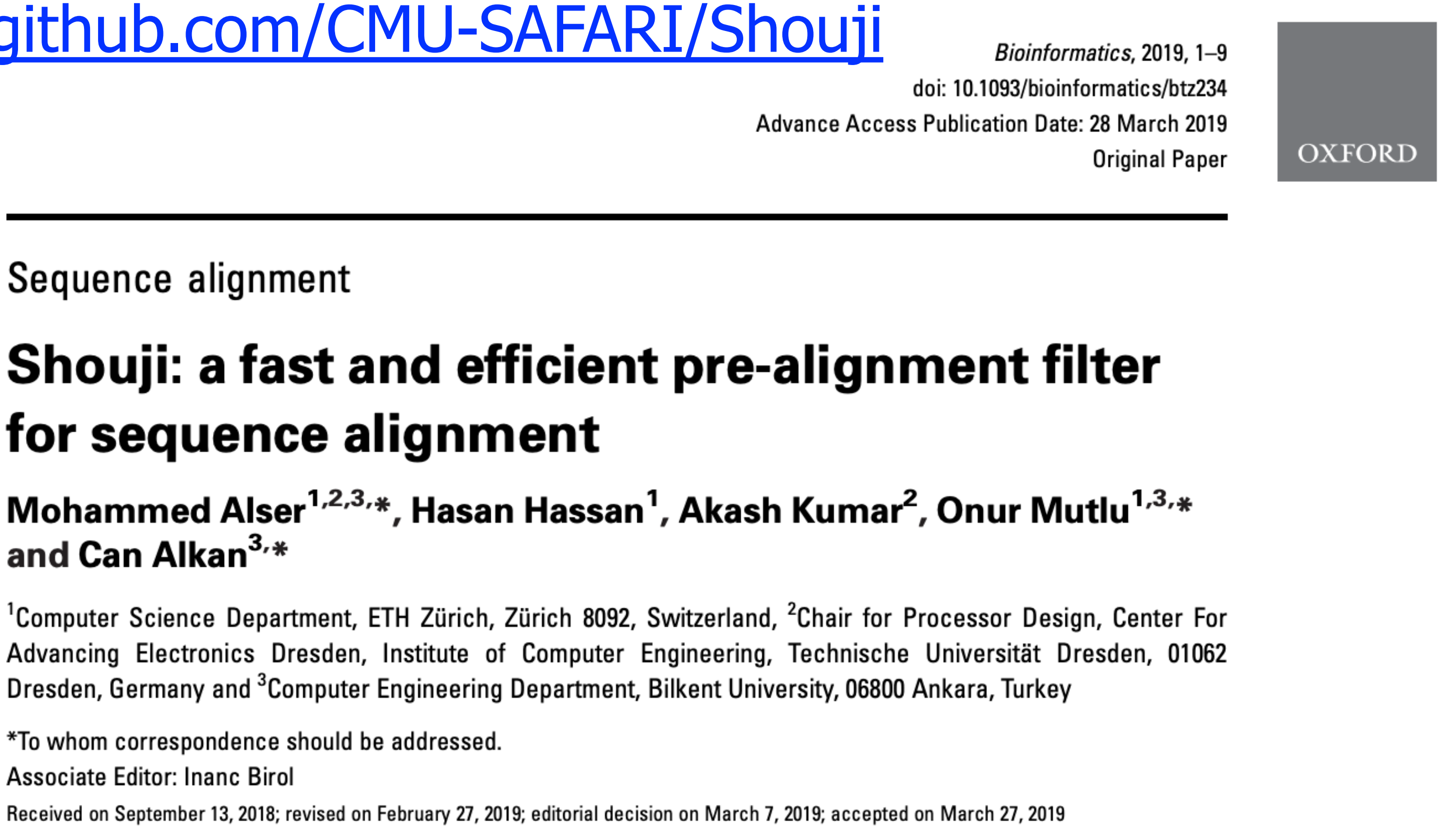
- Counting is performed concurrently for *all* bit-vectors and all sliding windows in a single clock cycle using multiple 4-input LUTs.



More on Shouji

Download and test for yourself

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



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 Hassan[1], Akash Kumar[2], Onur Mutlu[1,3,*] and Can Alkan[3,*]

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

*To whom correspondence should be addressed.

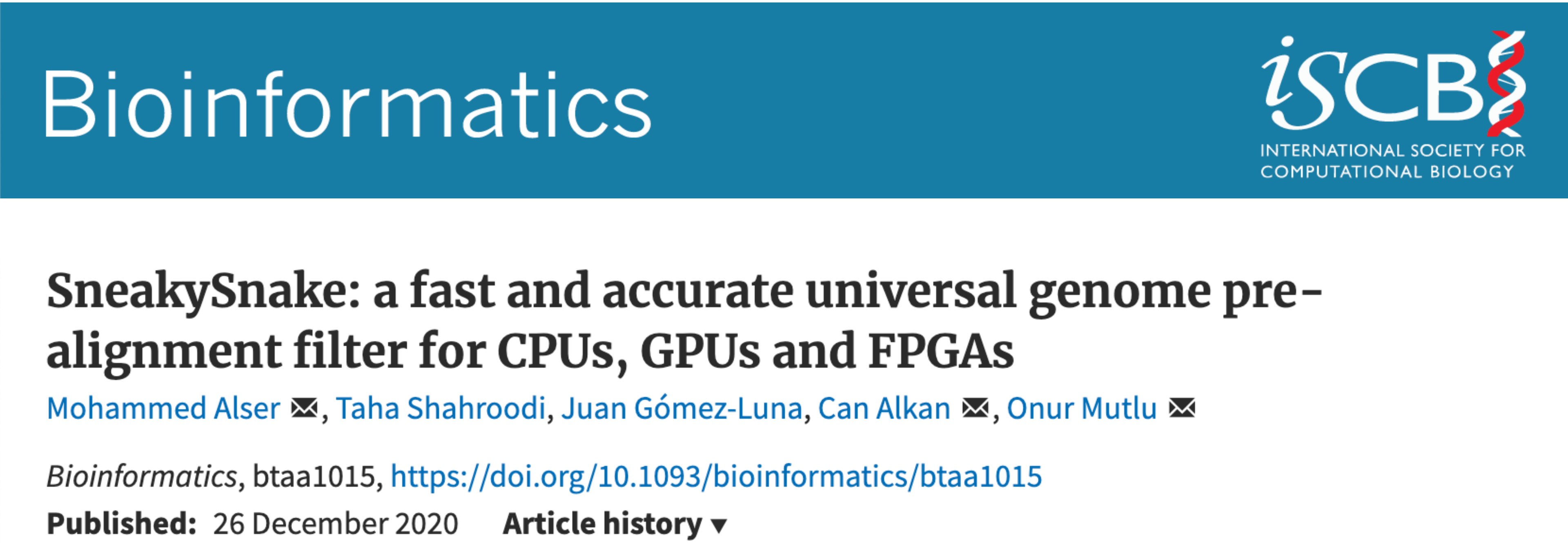
Associate Editor: Inanc Birol

Received on September 13, 2018; revised on February 27, 2019; editorial decision on March 7, 2019; accepted on March 27, 2019

Alser+, "Shouji: a fast and efficient pre-alignment filter for sequence alignment", *Bioinformatics* 2019, https://doi.org/10.1093/bioinformatics/btz234

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]



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

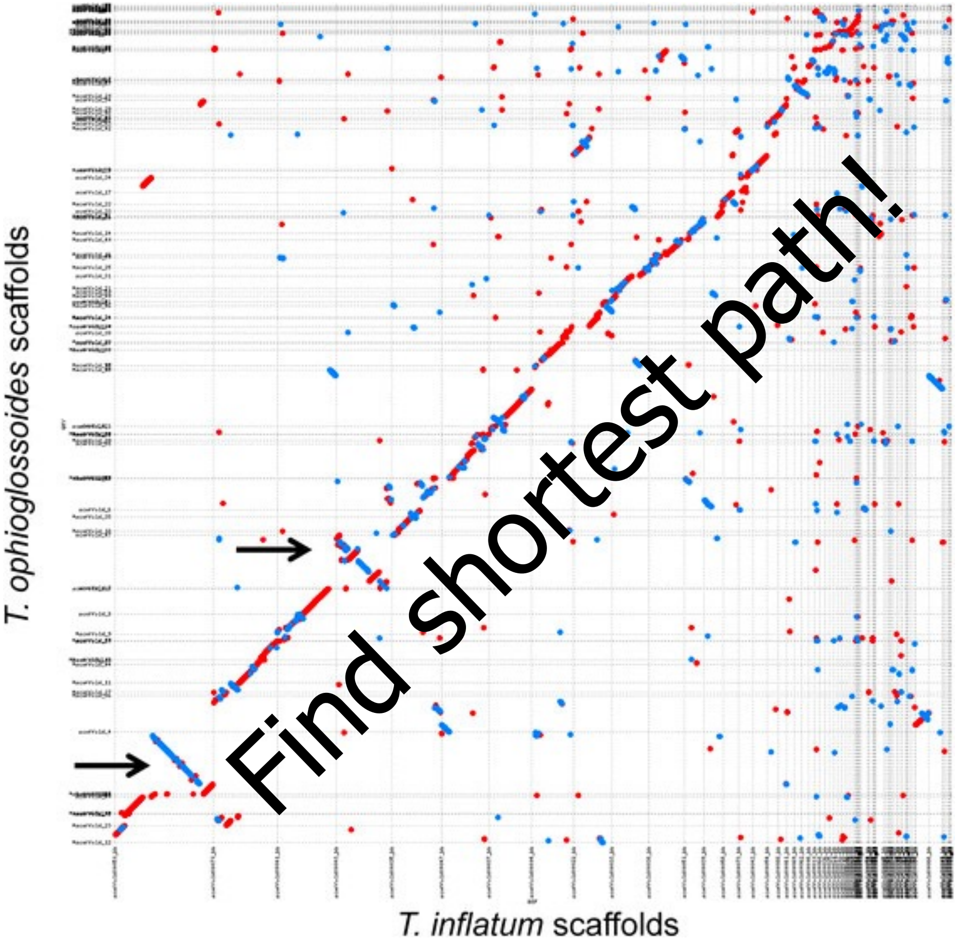
Mohammed Alser ✉, Taha Shahroodi, Juan Gómez-Luna, Can Alkan ✉, Onur Mutlu ✉

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

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

SneakySnake

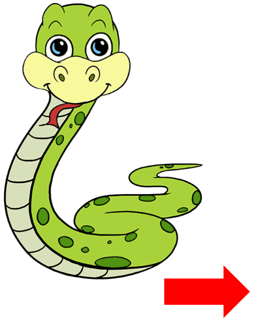
- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches.



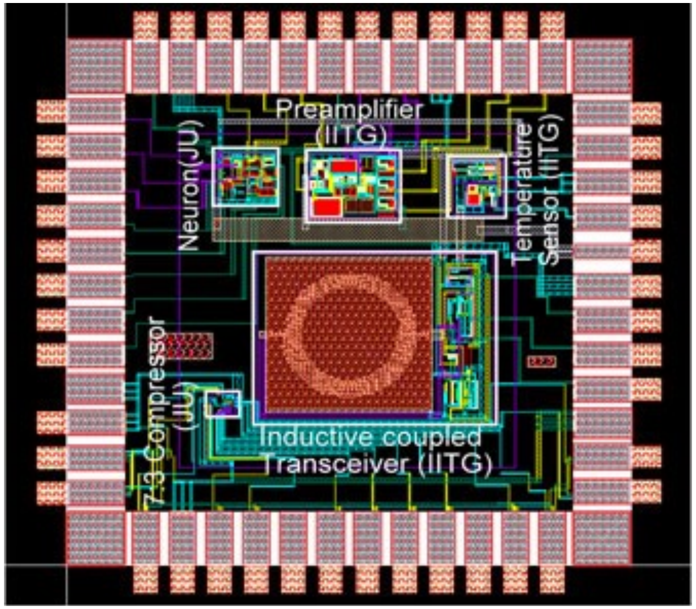
Dot plot, dot matrix
(Lipman and Pearson, 1985)

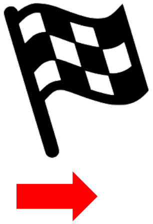
SneakySnake

- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches
- **Key idea:**
  - Approximate edit distance calculation is similar to Single Net Routing problem in VLSI chip



VLSI chip layout


Preamplifier (IITG)
Neuron(JU)
Temperature Sensor (IITG)
Compressor (JU)
Inductive coupled Transceiver (IITG)



SneakySnake Walkthrough

Building Neighborhood Map → Finding the Optimal Routing Path → Examining the Snake Survival

Given two genomic sequences, a reference sequence $R[1 \ldots m]$ and a query sequence $Q[1 \ldots m]$, and an edit distance threshold E, we calculate the entry $Z[i, j]$ of the chip maze, where $1 \leq i \leq (2E+1)$ and $1 \leq j \leq m$, as follows:

$$Z[i,j] = \begin{cases} 0, & if \;\; i = E+1, \; Q[j] = R[j], \\ 0, & if \;\; 1 \leq i \leq E, \; Q[j-i] = R[j], \\ 0, & if \;\; i > E+1, \; Q[j+i-E-1] = R[j], \\ 1, & otherwise \end{cases} \qquad (1)$$

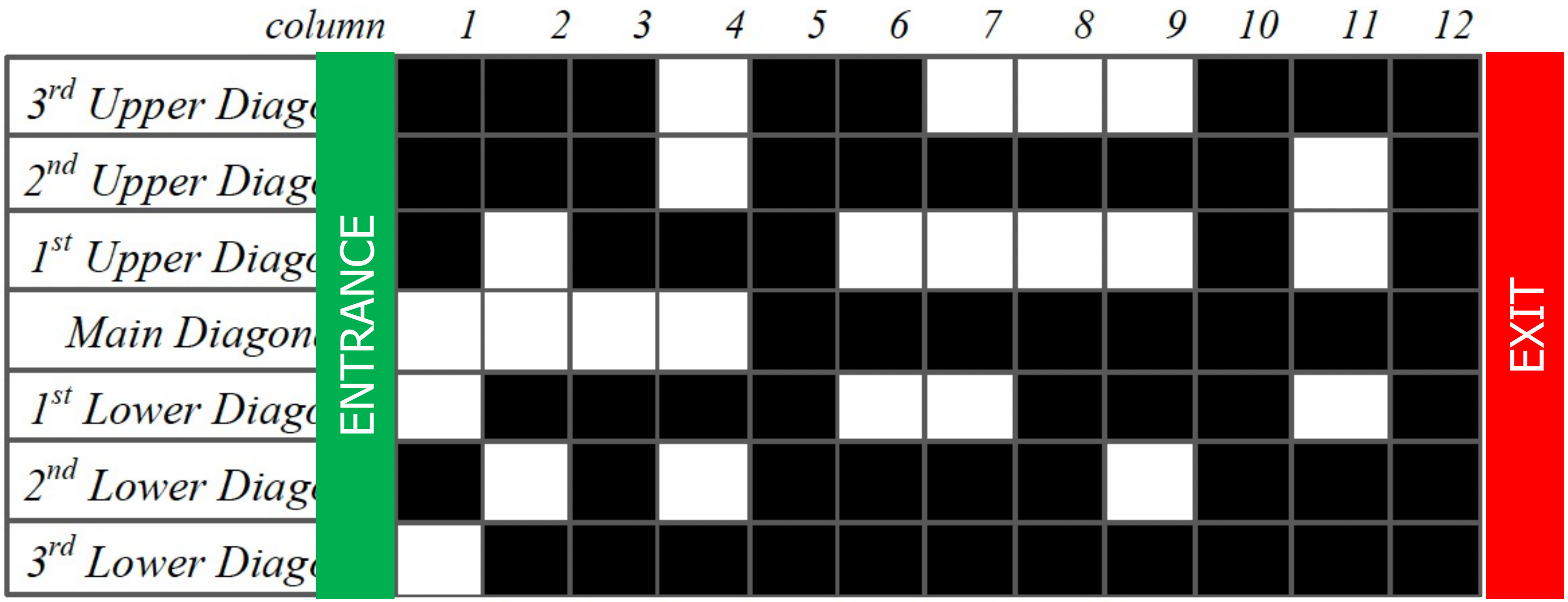
$E = 3$

column	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
3rd Upper Diagonal	1	1	1	0	1	1	0	0	0	1	1	1
2nd Upper Diagonal	1	1	1	0	1	1	1	1	1	1	0	1
1st Upper Diagonal	1	0	1	1	1	0	0	0	0	1	0	1
Main Diagonal	0	0	0	0	1	1	1	1	1	1	1	1
1st Lower Diagonal	0	1	1	1	1	0	0	1	1	1	0	1
2nd Lower Diagonal	1	0	1	0	1	1	1	1	0	1	1	1
3rd Lower Diagonal	0	1	1	1	1	1	1	1	1	1	1	1

SneakySnake Walkthrough

Building Neighborhood Map → Finding the Optimal Routing Path → Examining the Snake Survival

$E = 3$

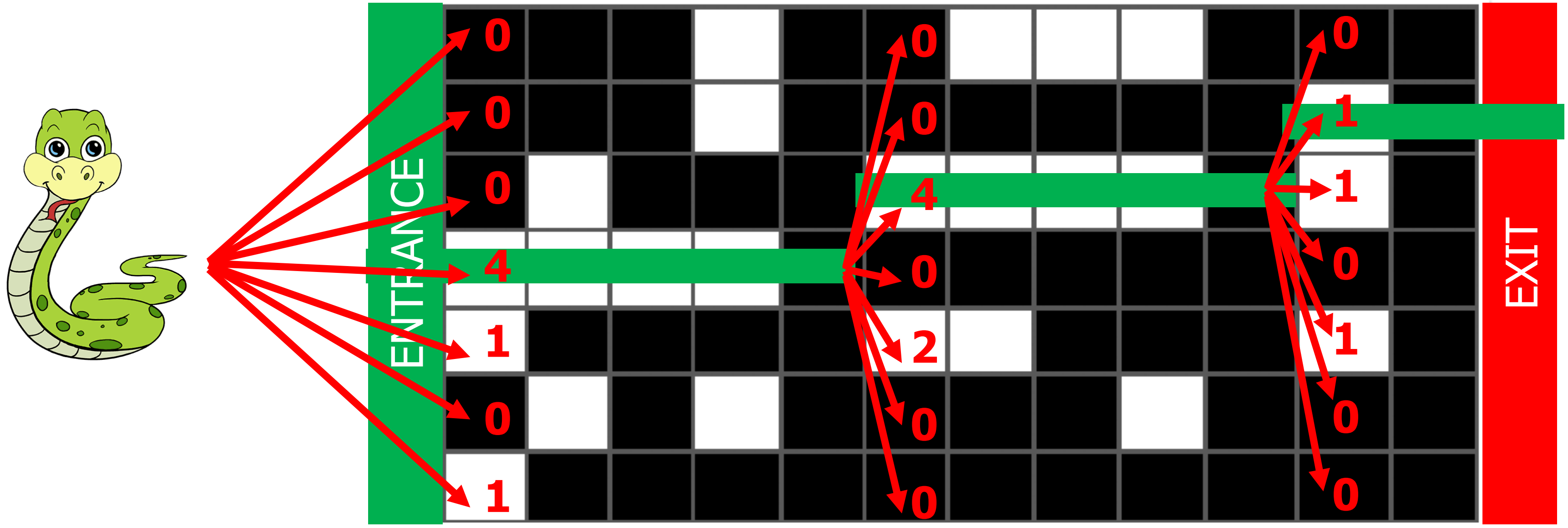


column	1	2	3	4	5	6	7	8	9	10	11	12
3rd Upper Diag	■	■	■	□	■	■	□	□	□	■	■	■
2nd Upper Diag	■	■	■	□	■	■	■	■	■	■	□	■
1st Upper Diag	■	□	■	■	■	□	□	□	□	■	□	■
Main Diagon	□	□	□	□	■	■	■	■	■	■	■	■
1st Lower Diag	□	■	■	■	■	□	□	■	■	■	□	■
2nd Lower Diag	■	□	■	□	■	■	■	■	□	■	■	■
3rd Lower Diag	□	■	■	■	■	■	■	■	■	■	■	■

SneakySnake Walkthrough

Building Neighborhood Map → Finding the Optimal Routing Path → Examining the Snake Survival

■ [illegible]

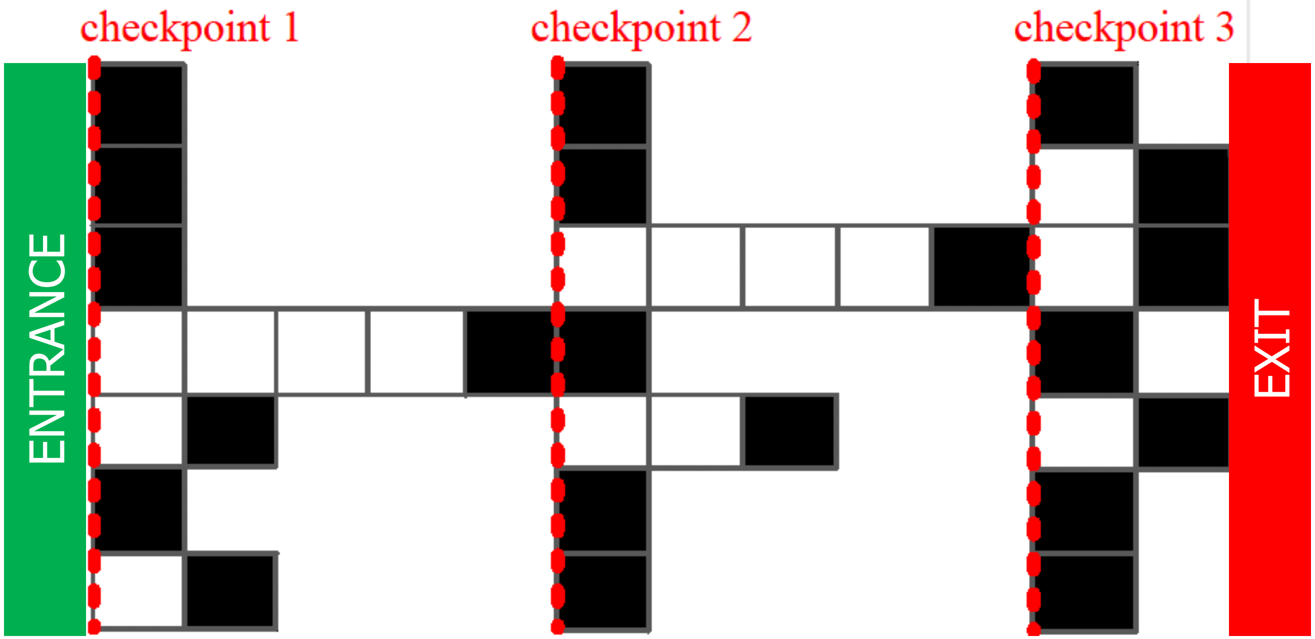


SneakySnake Walkthrough

Building Neighborhood Map | Finding the Routing Travel Path | Examining the Snake Survival

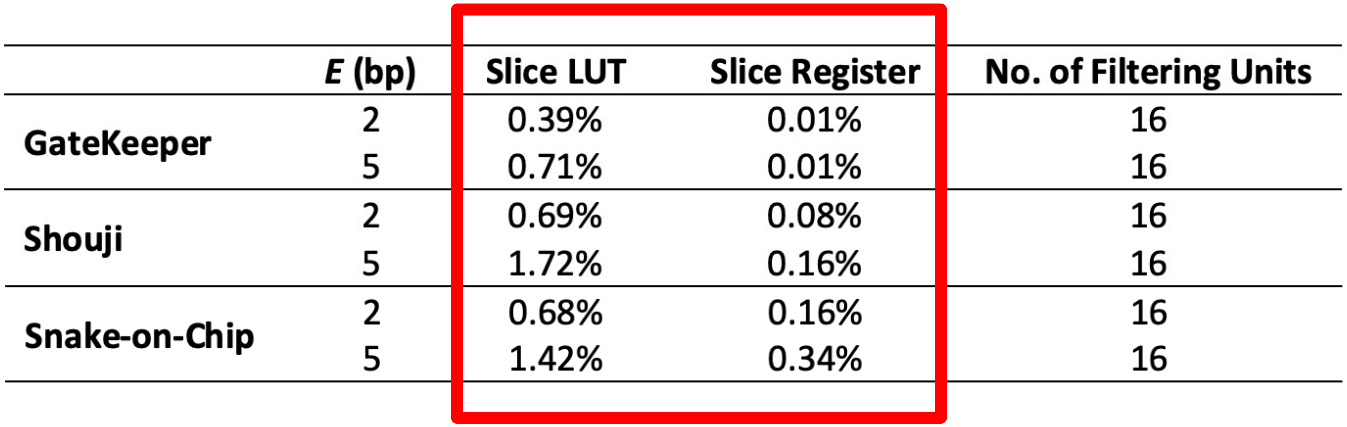
■ 3

This is what you actually need to build and it can be done on-the-fly!



FPGA Resource Analysis

- FPGA resource usage for a single filtering unit of GateKeeper, Shouji, and Snake-on-Chip for a sequence length of 100 and under different edit distance thresholds (E).



	E (bp)	Slice LUT	Slice Register	No. of Filtering Units
GateKeeper	2	0.39%	0.01%	16
	5	0.71%	0.01%	16
Shouji	2	0.69%	0.08%	16
	5	1.72%	0.16%	16
Snake-on-Chip	2	0.68%	0.16%	16
	5	1.42%	0.34%	16

Key Results of SneakySnake

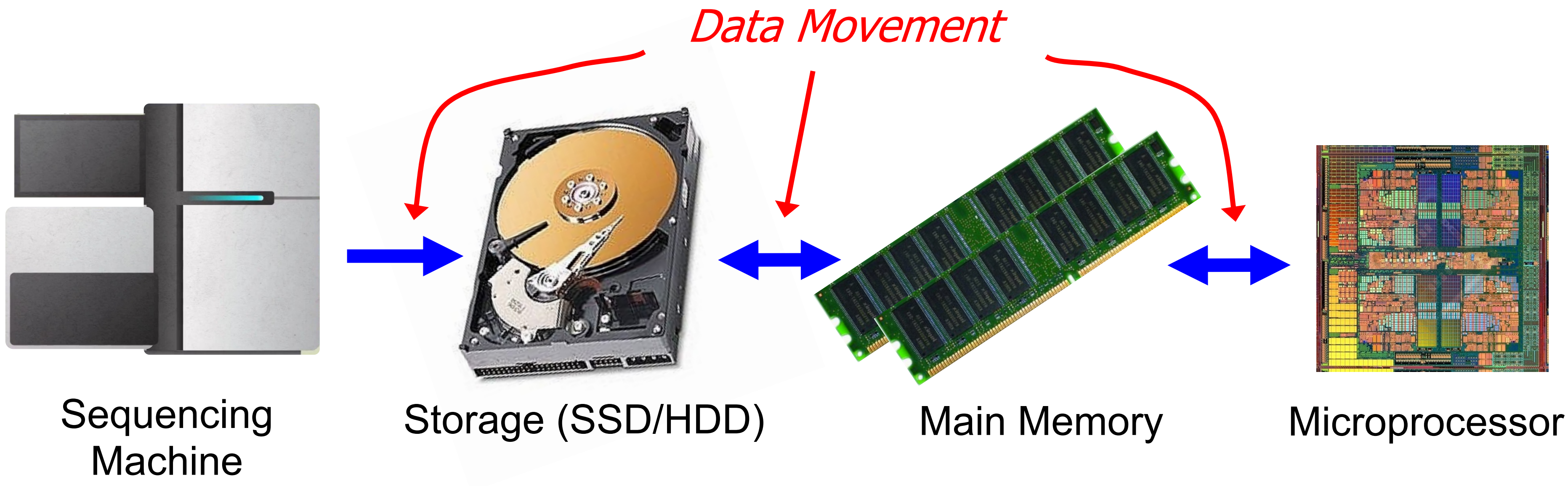
- SneakySnake is up to four orders of magnitude more accurate than Shouji (Bioinformatics'19) and GateKeeper (Bioinformatics'17)

- Using short reads, SneakySnake accelerates Edlib (Bioinformatics'17) and Parasail (BMC Bioinformatics'16) by
  - up to 37.7× and 43.9× (>12× on average), on CPUs
  - up to 413× and 689× (>400× on average) with *FPGA/GPU acceleration*

- Using long reads, SneakySnake accelerates Parasail and KSW2 by 140.1× and 17.1× on average, respectively, on CPUs

Data Movement Dominates Performance

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



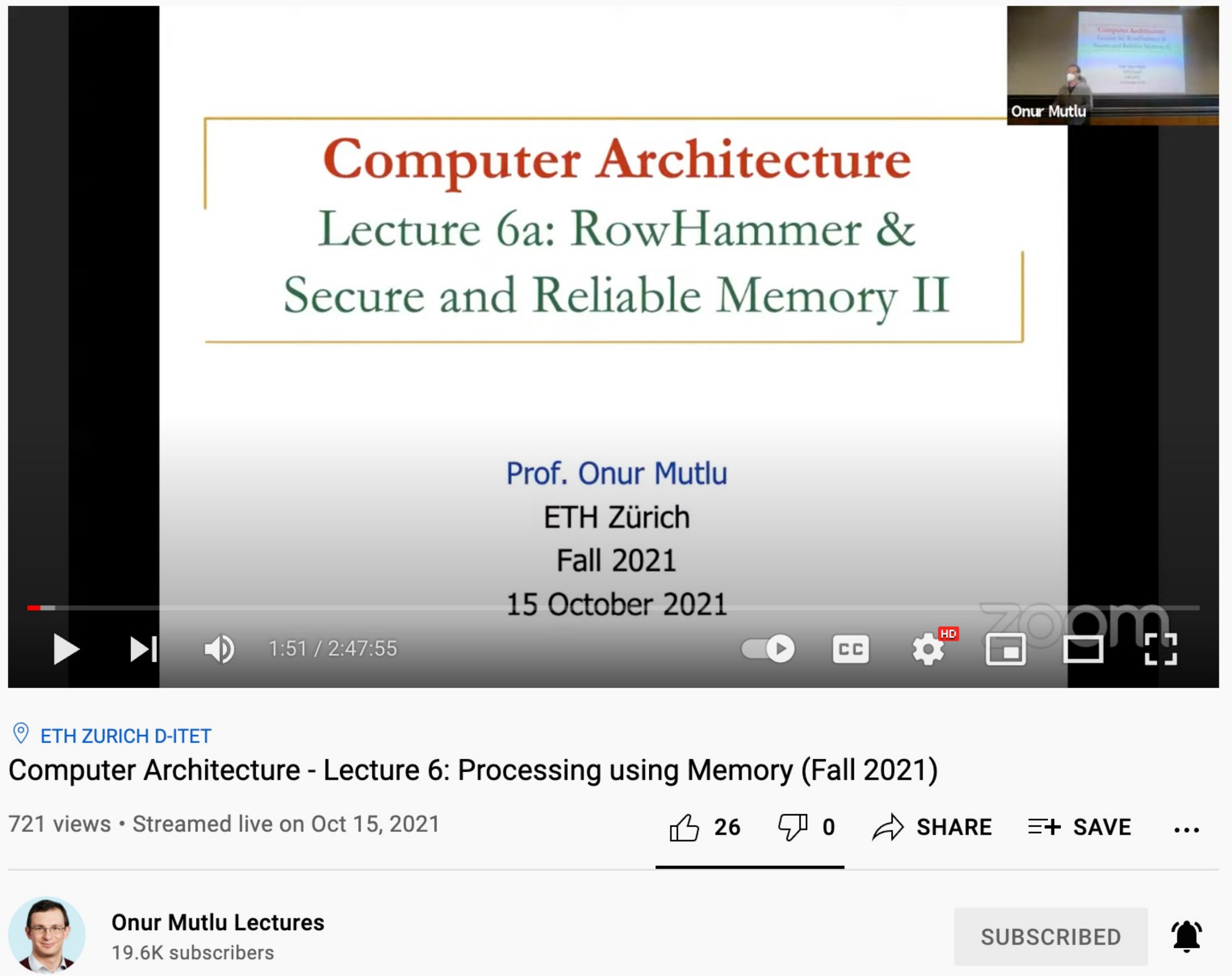
Single memory request consumes >160x-800x more energy compared to performing an addition operation

✻ 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

Read Mapping & Filtering in Memory

We need to design
mapping & filtering algorithms
that fit processing-in-memory

Processing Using Memory



Computer Architecture

Lecture 6a: RowHammer & Secure and Reliable Memory II

Prof. Onur Mutlu

ETH Zürich

Fall 2021

15 October 2021

ETH ZURICH D-ITET

Computer Architecture - Lecture 6: Processing using Memory (Fall 2021)

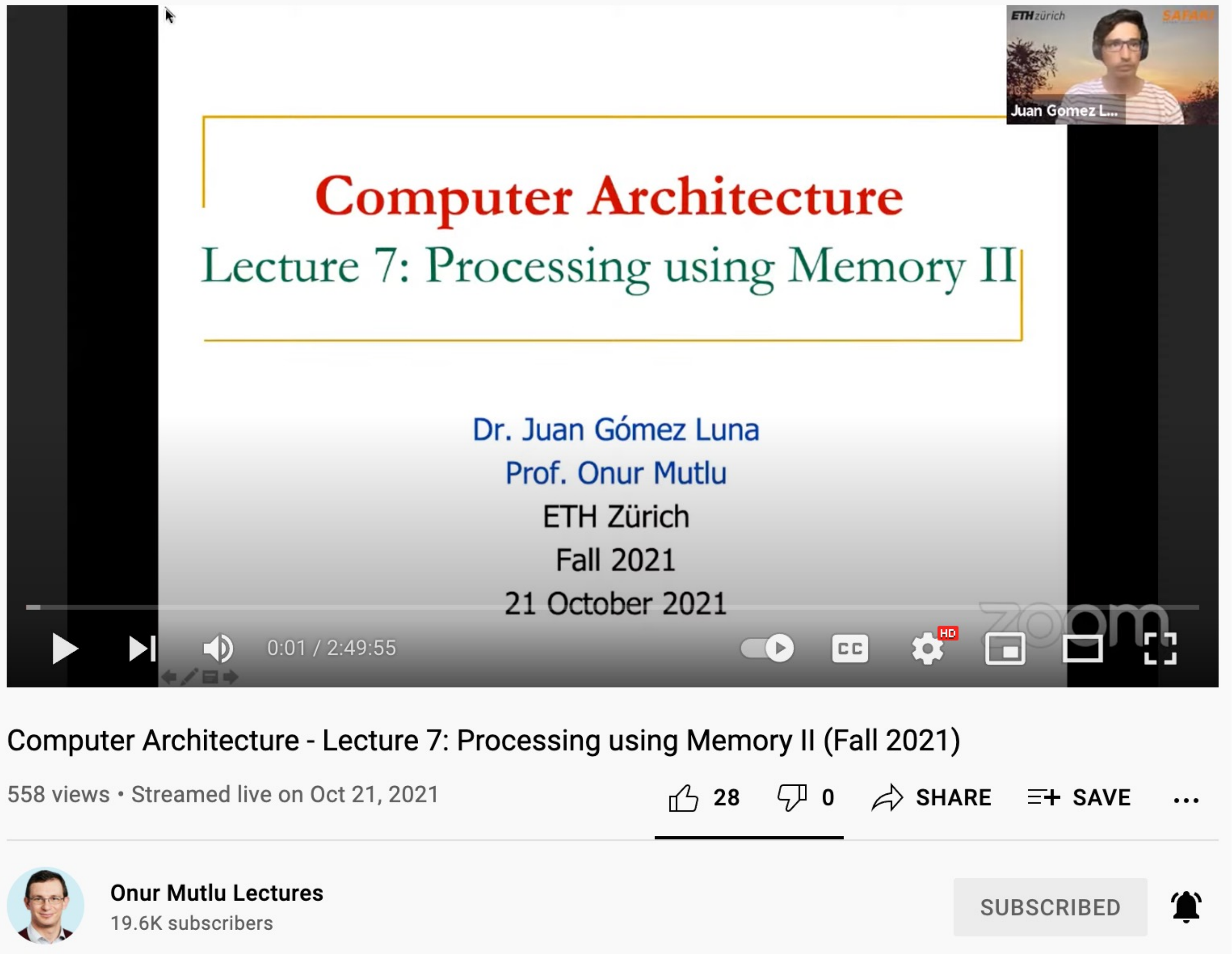
721 views • Streamed live on Oct 15, 2021

Onur Mutlu Lectures

19.6K subscribers

https://www.youtube.com/watch?v=HNd4skQrt6I

Processing Using Memory II



Computer Architecture
Lecture 7: Processing using Memory II

Dr. Juan Gómez Luna
Prof. Onur Mutlu
ETH Zürich
Fall 2021
21 October 2021

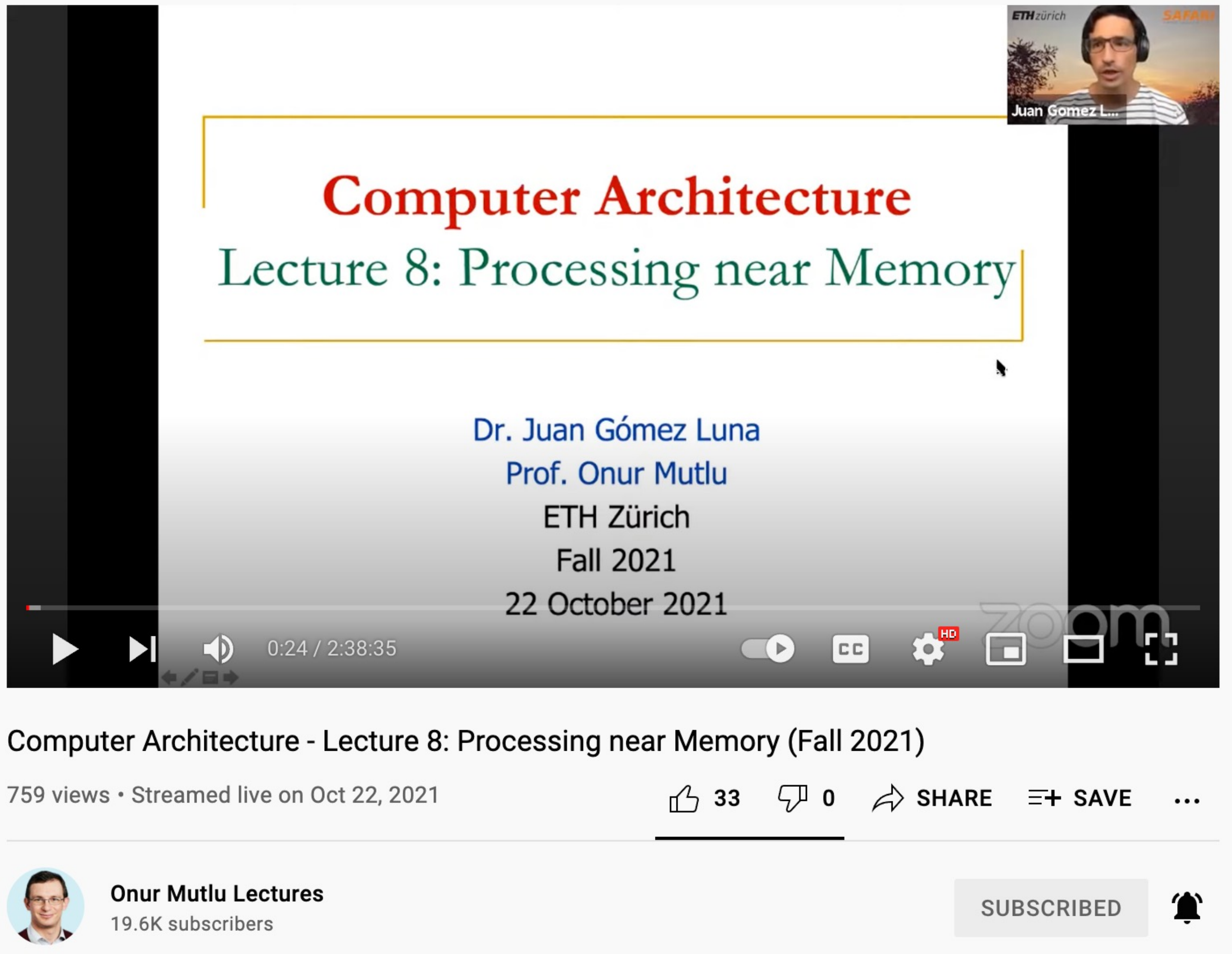
Computer Architecture - Lecture 7: Processing using Memory II (Fall 2021)

558 views • Streamed live on Oct 21, 2021

Onur Mutlu Lectures
19.6K subscribers

https://www.youtube.com/watch?v=k56x2qcaXWY

Processing Near Memory



Computer Architecture

Lecture 8: Processing near Memory

Dr. Juan Gómez Luna

Prof. Onur Mutlu

ETH Zürich

Fall 2021

22 October 2021

Computer Architecture - Lecture 8: Processing near Memory (Fall 2021)

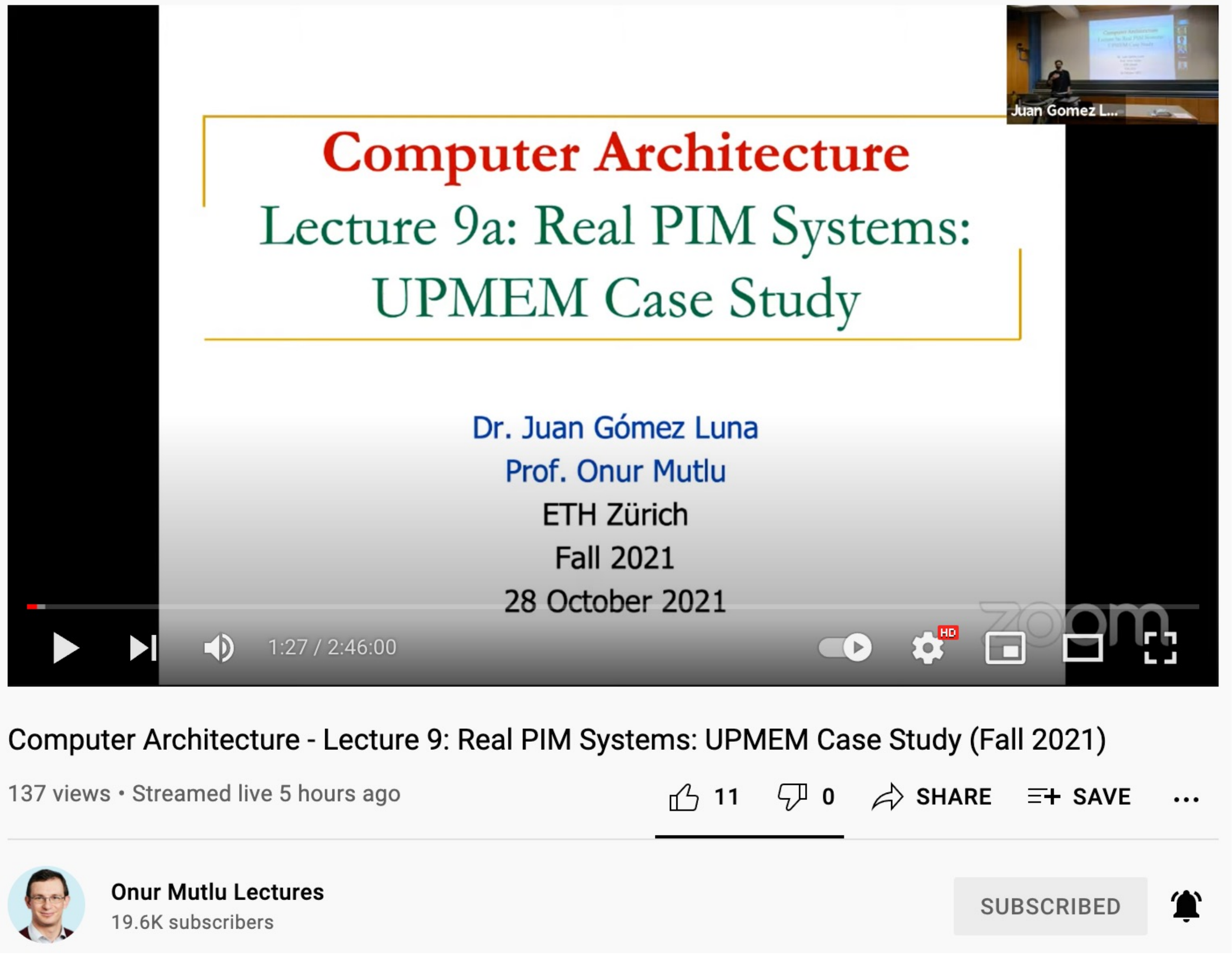
759 views • Streamed live on Oct 22, 2021

Onur Mutlu Lectures

19.6K subscribers

https://www.youtube.com/watch?v=kpgLmX9sdcI

Using Real PIM System



Computer Architecture

Lecture 9a: Real PIM Systems: UPMEM Case Study

Dr. Juan Gómez Luna

Prof. Onur Mutlu

ETH Zürich

Fall 2021

28 October 2021

Computer Architecture - Lecture 9: Real PIM Systems: UPMEM Case Study (Fall 2021)

137 views • Streamed live 5 hours ago

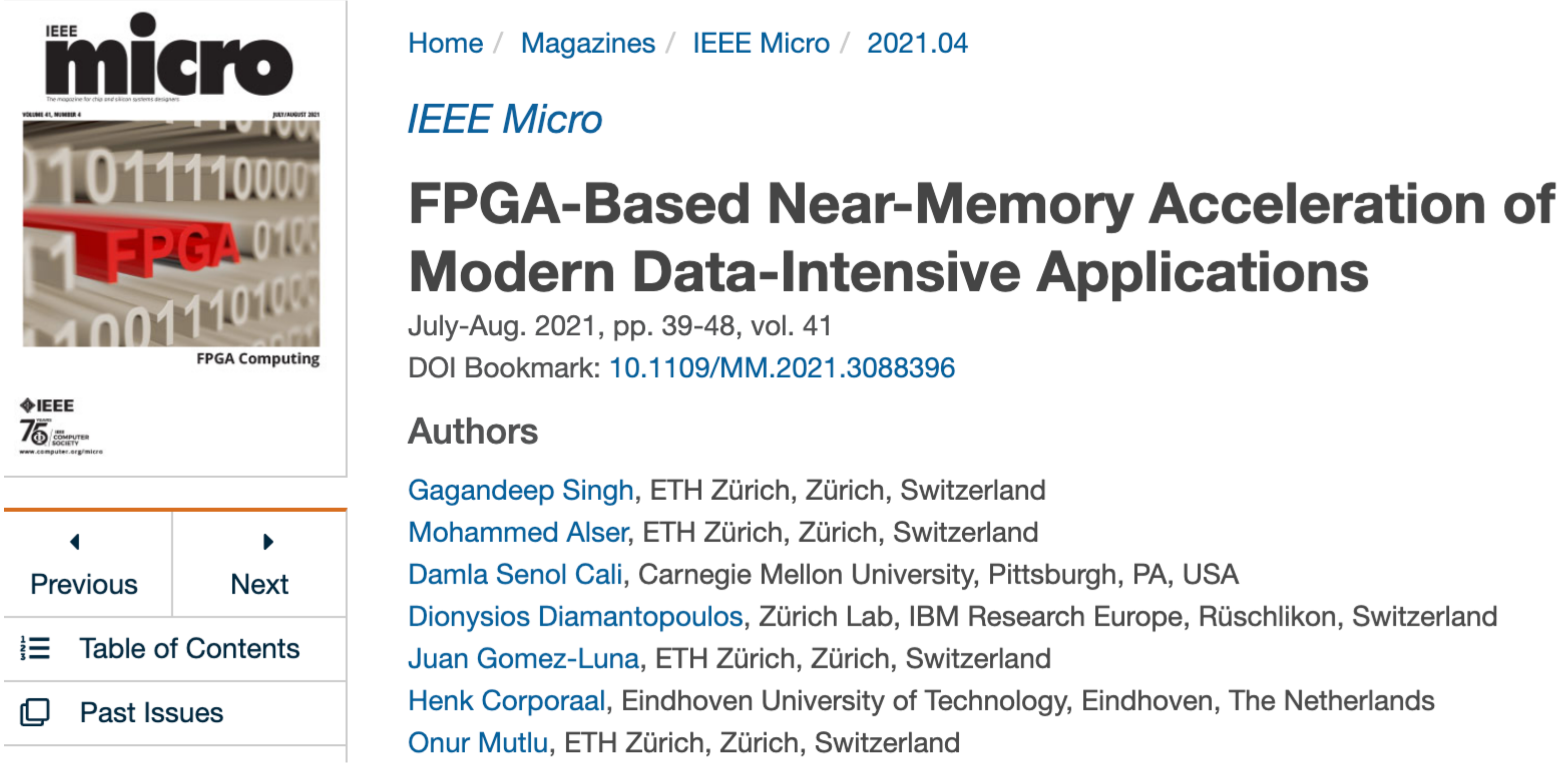
Onur Mutlu Lectures

19.6K subscribers

https://www.youtube.com/watch?v=TuVw_SKaTCo

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]



◂ Previous | ▸ Next

Table of Contents

Past Issues

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

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

Near-memory SneakySnake

- Problem: Read Mapping is heavily bottlenecked by data movement from main memory

- Solution: Perform read mapping near where data resides (i.e., near-memory)

- We carefully redesigned the accelerator logic of SneakySnake to exploit near-memory computation capability on modern FPGA boards with high-bandwidth memory

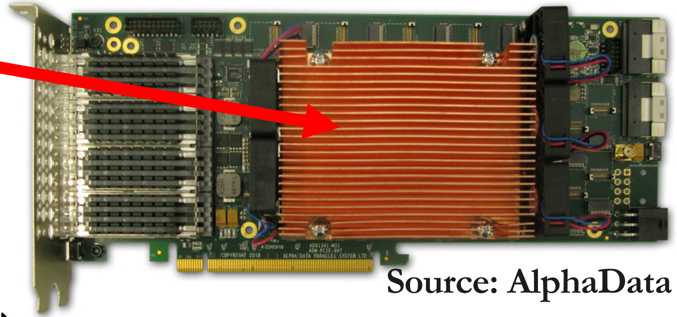
Heterogeneous System: CPU+FPGA

We evaluate two POWER9+FPGA systems:

1. **HBM-based AD9H7 board:** Xilinx Virtex Ultrascale+™ XCVU37P-2
2. **DDR4-based AD9V3 board:** Xilinx Virtex Ultrascale+™ XCVU3P-2

HBM-based AD9H7 board

FPGA + HBM on the same package substrate



Source: AlphaData

Source: IBM

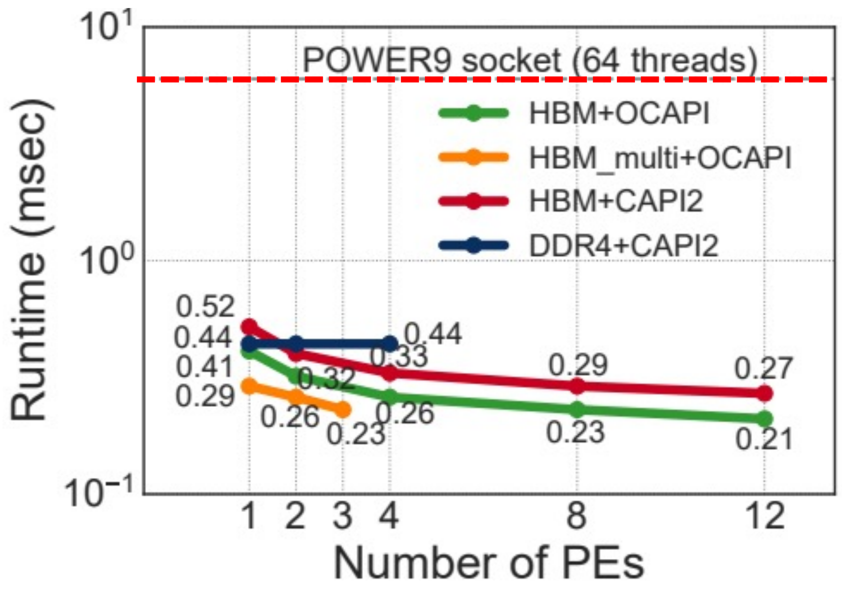
POWER9 AC922

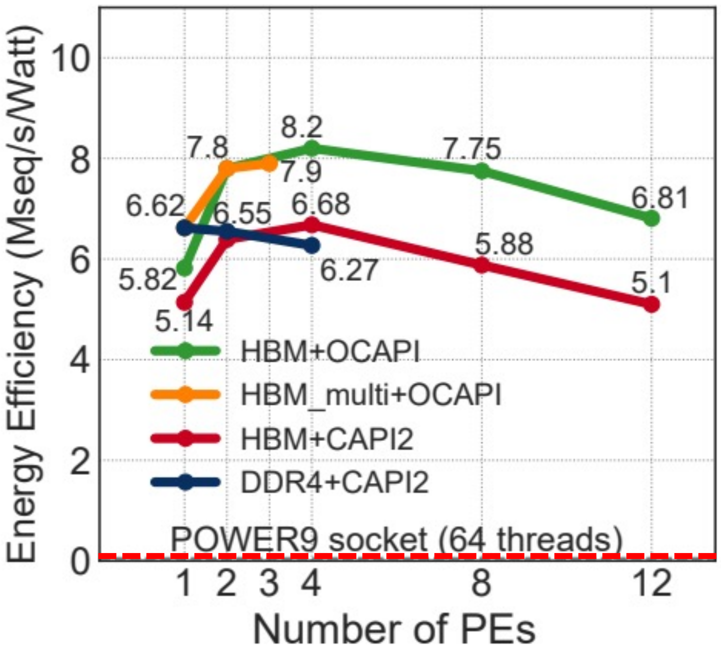
CAPI2

Source: AlphaData

DDR4-based AD9V3 board

Key Results of Near-memory SneakySnake

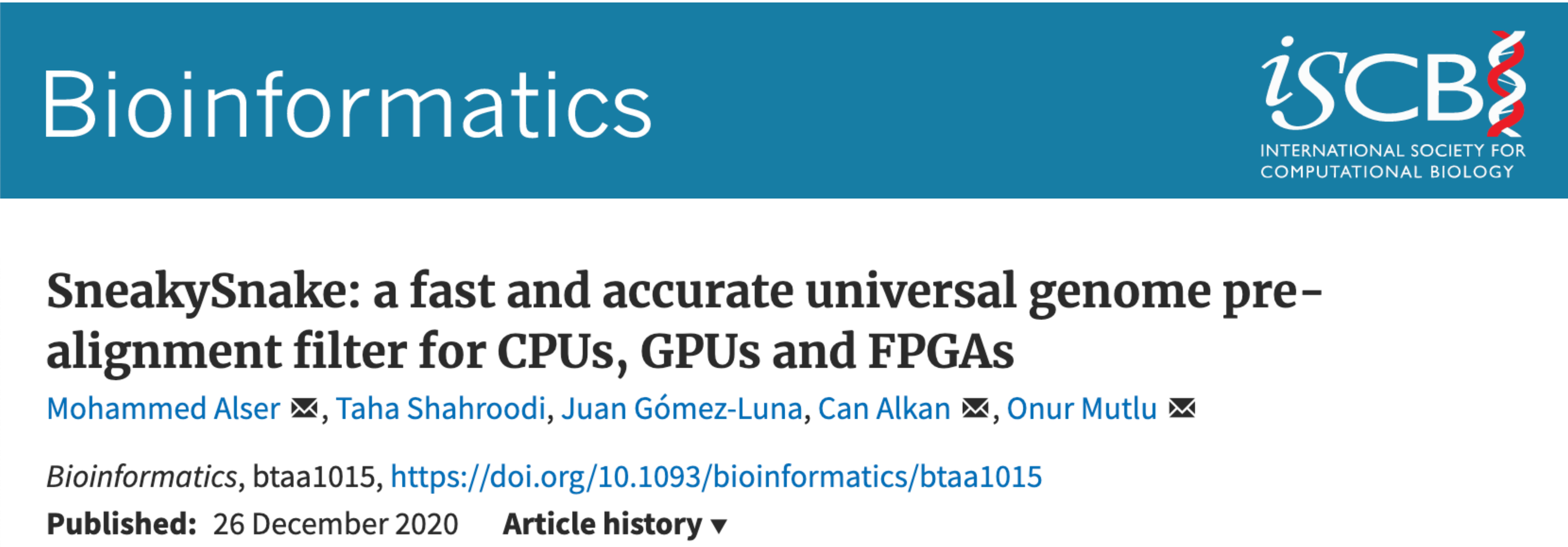




Near-memory pre-alignment filtering improves **performance** and **energy efficiency** by 27.4× and 133×, respectively, over a 16-core (64 hardware threads) IBM POWER9 CPU

More on SneakySnake [Bioinformatics 2020]

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]



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

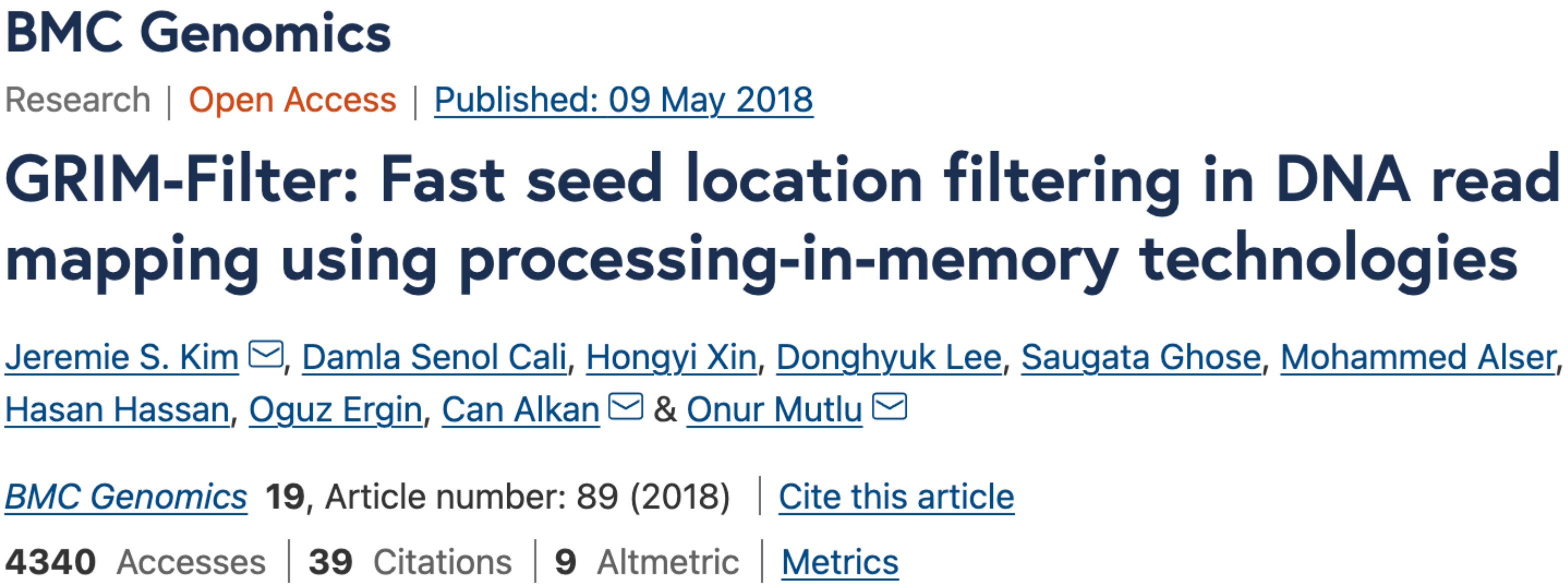
Mohammed Alser ✉, Taha Shahroodi, Juan Gómez-Luna, Can Alkan ✉, Onur Mutlu ✉

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

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

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

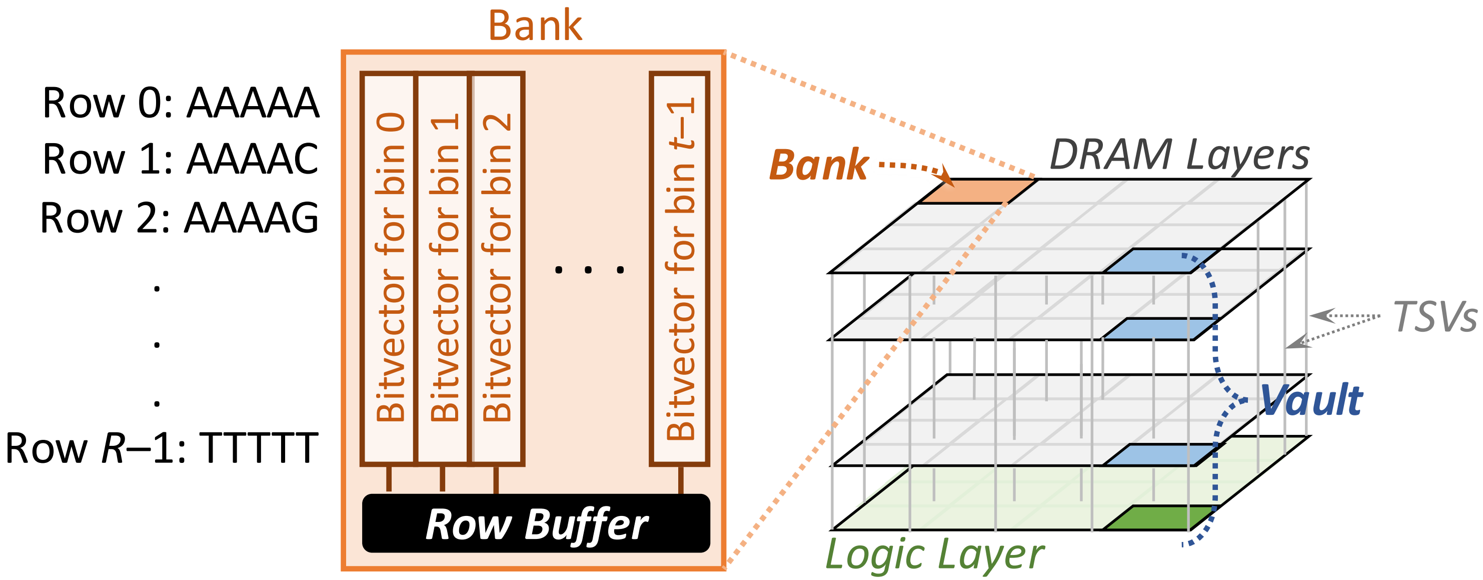
GRIM-Filter

- **Key observation:** FPGA and GPU accelerators are Heavily bottlenecked by Data Movement.

- **Key idea:** exploiting the high memory bandwidth and the logic layer of 3D-stacked memory to perform highly-parallel filtering in the DRAM chip itself.

- **Key results:**
  - We propose an algorithm called **GRIM-Filter**
  - GRIM-Filter with processing-in-memory is 1.8x-3.7x (2.1x on average) faster than FastHASH filter (BMC Genomics'13) across real data sets.
  - GRIM-Filter has 5.6x-6.4x (6.0x on average) lower falsely accepted pairs than FastHASH filter (BMC Genomics'13) across real data sets.

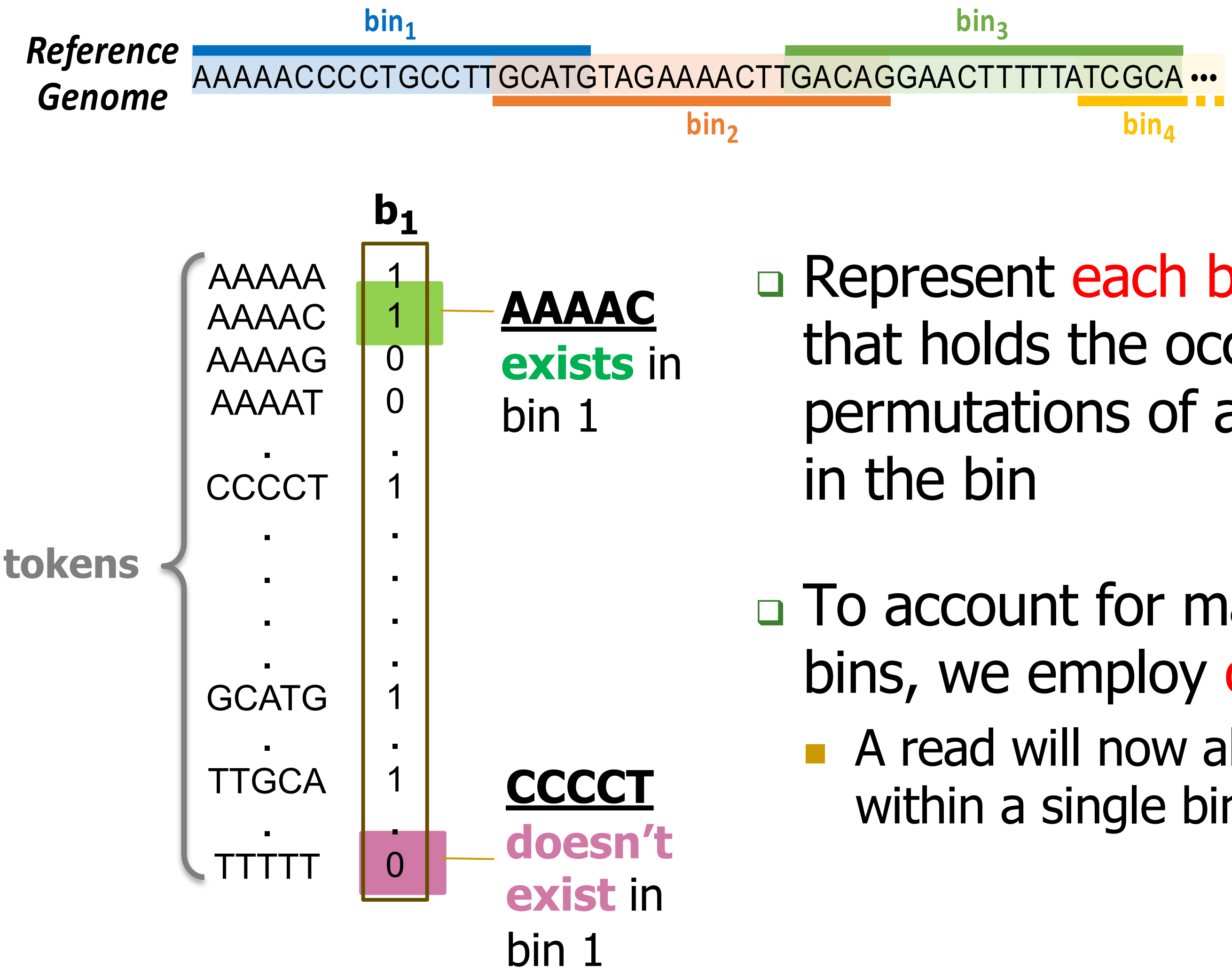
GRIM-Filter in 3D-Stacked DRAM



- Each DRAM layer is organized as an array of **banks**
  - A **bank** is an array of cells with a row buffer to transfer data

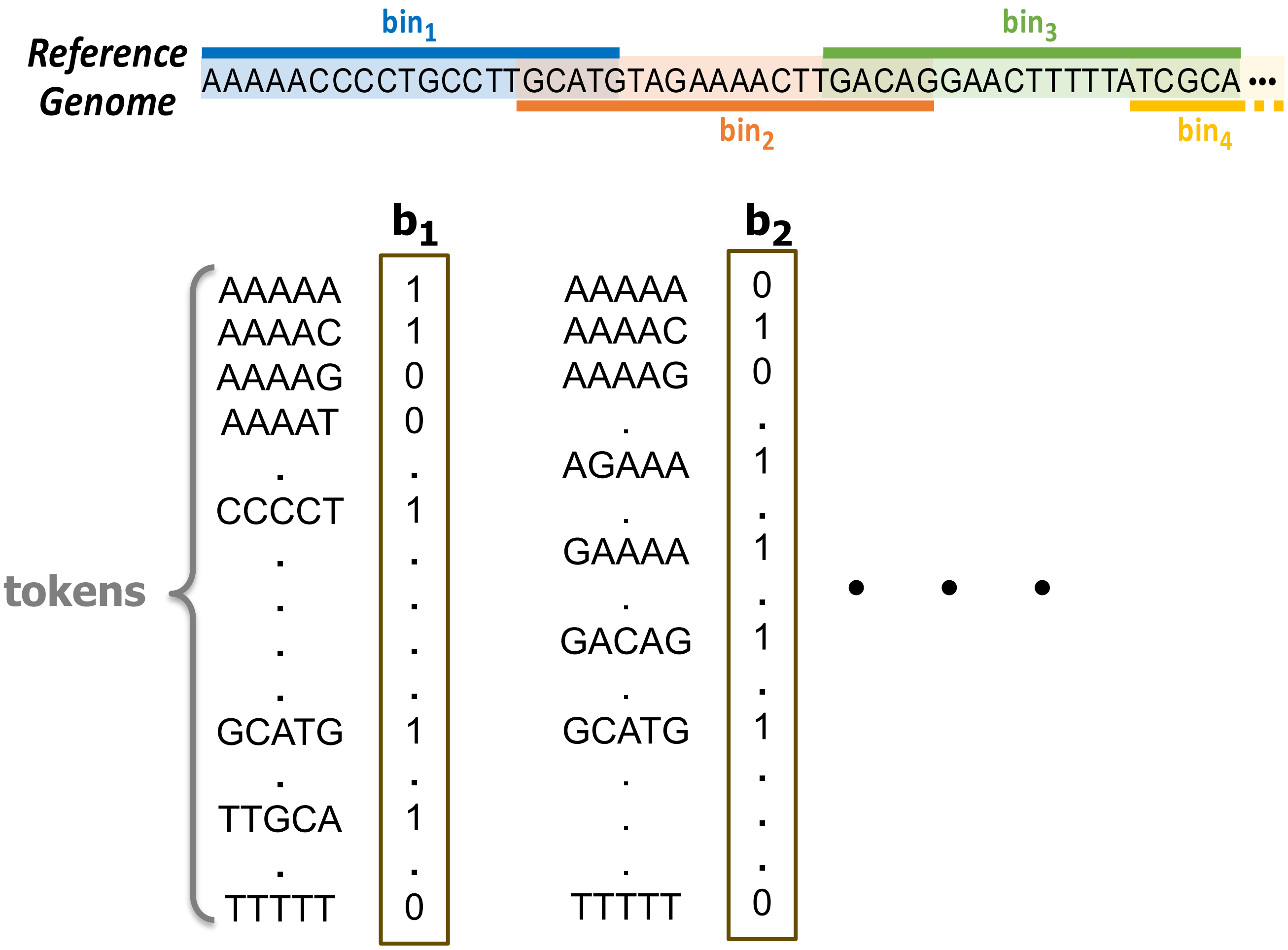
- The layout of bitvectors in a bank enables filtering many bins in parallel

GRIM-Filter: Bitvectors



- Represent each bin with a **bitvector** that holds the occurrence of all permutations of a small string (**token**) in the bin
- To account for matches that straddle bins, we employ overlapping bins
  - A read will now always completely fall within a single bin

GRIM-Filter: Bitvectors

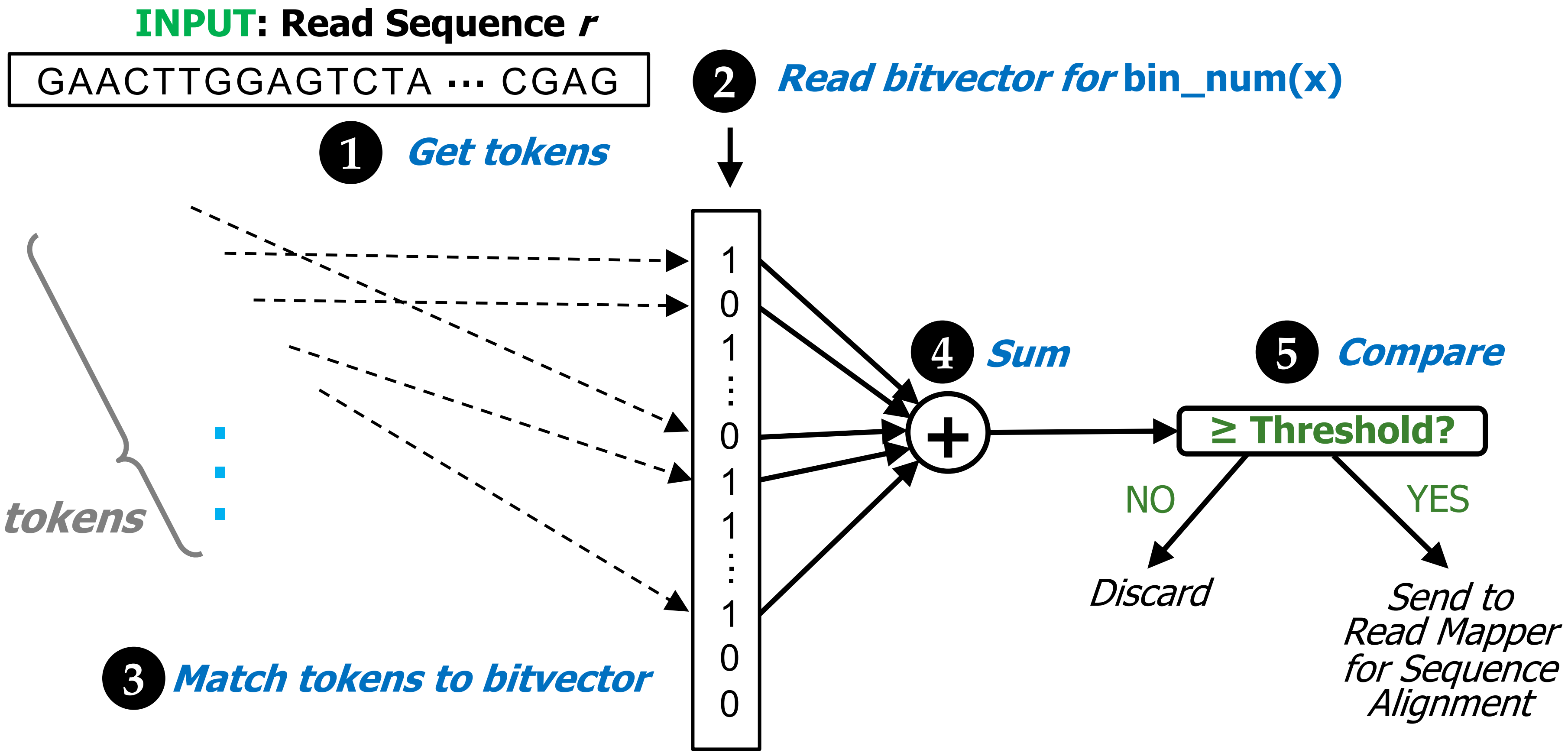


Storing all bitvectors requires $\underline{4^n * t}$ bits in memory, where
t = number of bins
&
n = token length.

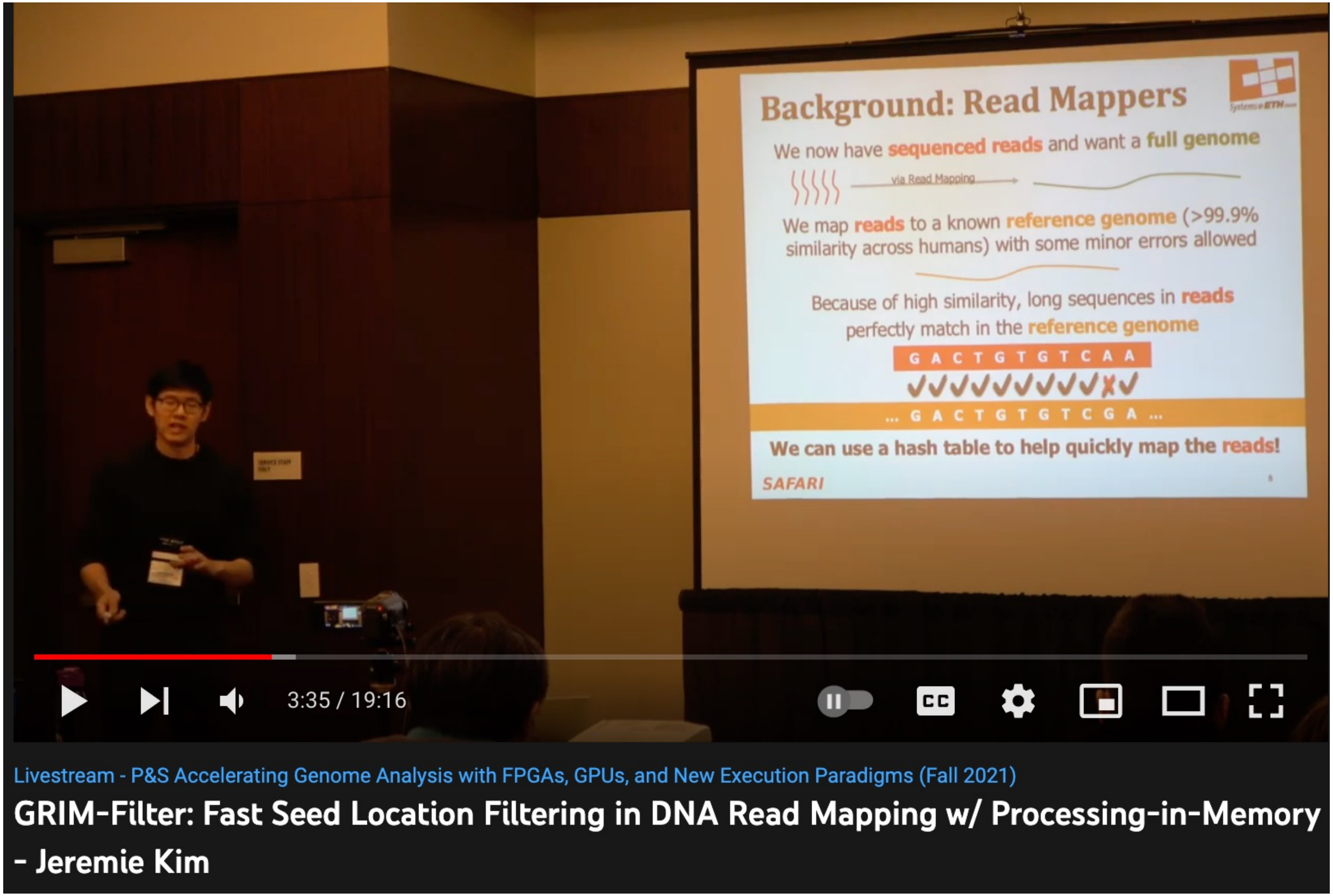
For **bin size** ~200, and **n** = 5, **memory footprint** ~3.8 GB

GRIM-Filter: Checking a Bin

How GRIM-Filter determines whether to **discard** potential match locations in a given bin **prior** to alignment



More on GRIM-Filter



Background: Read Mappers

We now have sequenced reads and want a full genome

We map reads to a known reference genome (>99.9% similarity across humans) with some minor errors allowed

Because of high similarity, long sequences in reads perfectly match in the reference genome

We can use a hash table to help quickly map the reads!

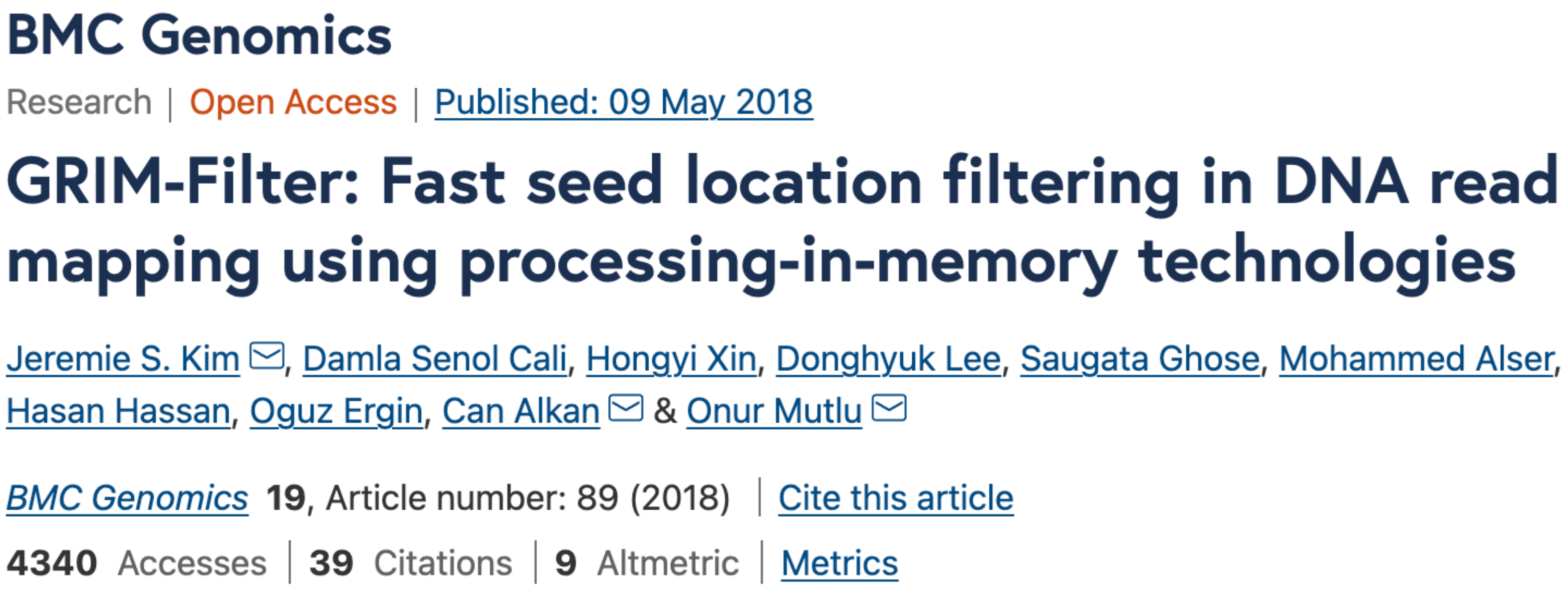
Livestream - P&S Accelerating Genome Analysis with FPGAs, GPUs, and New Execution Paradigms (Fall 2021)

GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping w/ Processing-in-Memory - Jeremie Kim

https://www.youtube.com/watch?v=j5-I84iNVd8

More on 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

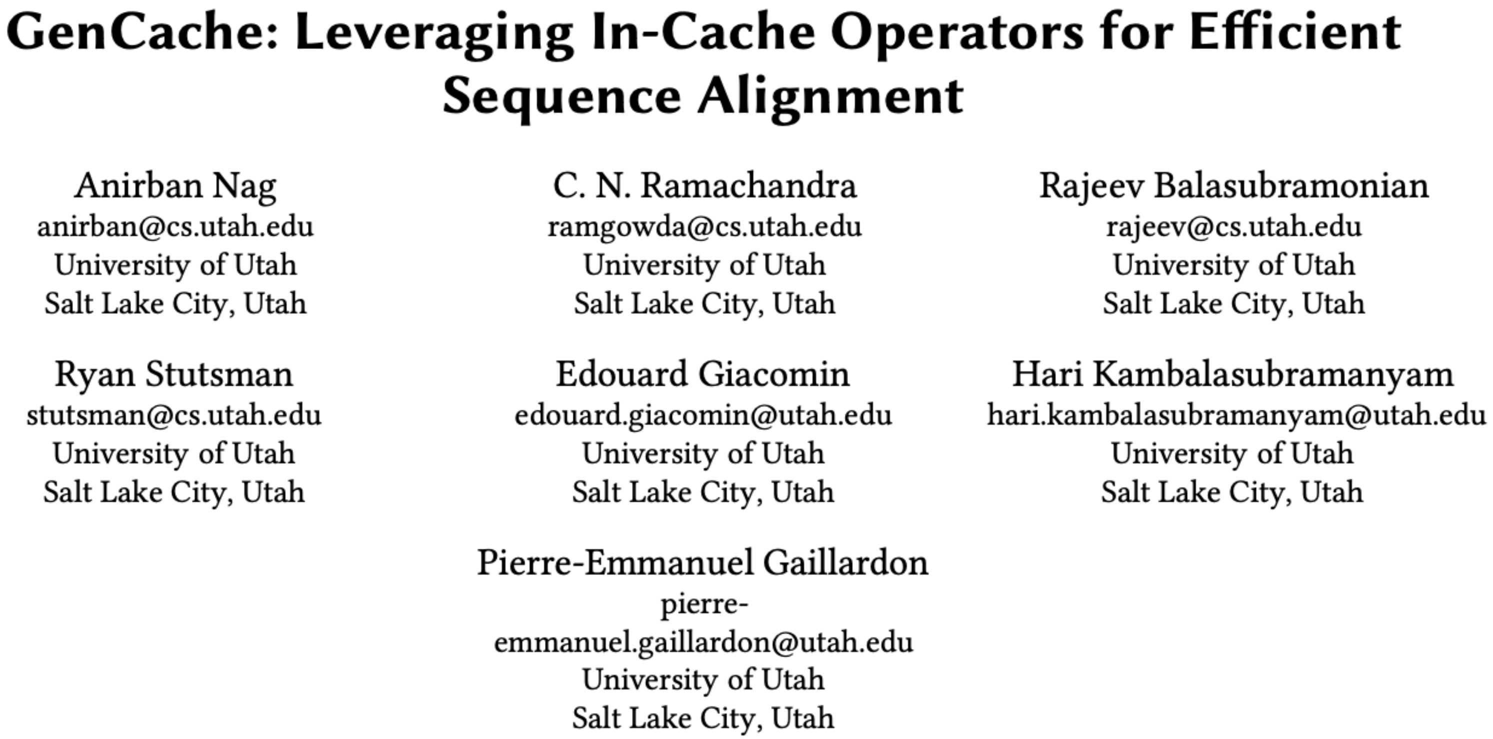
GRIM-Filter: Fast seed location filtering in DNA read mapping using processing-in-memory technologies

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

GenCache



GenCache: Leveraging In-Cache Operators for Efficient Sequence Alignment

Anirban Nag
anirban@cs.utah.edu
University of Utah
Salt Lake City, Utah

C. N. Ramachandra
ramgowda@cs.utah.edu
University of Utah
Salt Lake City, Utah

Rajeev Balasubramonian
rajeev@cs.utah.edu
University of Utah
Salt Lake City, Utah

Ryan Stutsman
stutsman@cs.utah.edu
University of Utah
Salt Lake City, Utah

Edouard Giacomin
edouard.giacomin@utah.edu
University of Utah
Salt Lake City, Utah

Hari Kambalasubramanyam
hari.kambalasubramanyam@utah.edu
University of Utah
Salt Lake City, Utah

Pierre-Emmanuel Gaillardon
pierre-emmanuel.gaillardon@utah.edu
University of Utah
Salt Lake City, Utah

Nag, Anirban, et al. **"GenCache: Leveraging In-Cache Operators for Efficient Sequence Alignment."** *Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture (**MICRO 52**)* , ACM, 2019.

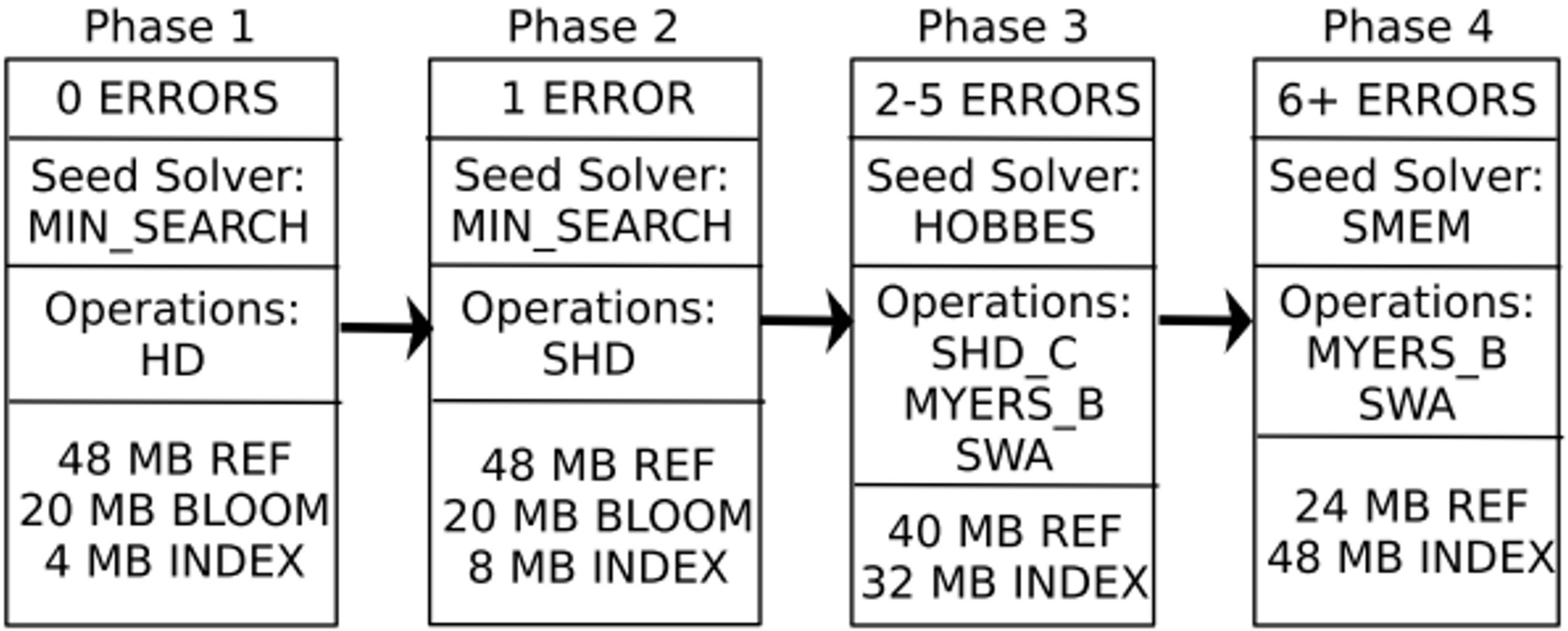
GenCache

- **Key observation:** State-of-the-art alignment accelerators are still bottlenecked by memory.

- **Key ideas:**
  - Performing in-cache alignment + pre-alignment filtering by enabling processing-in-cache using previous proposal, ComputeCache (HPCA'17).
  - Using different Pre-alignment filters depending on the selected edit distance threshold.

- Results:
  - GenCache on CPU is 1.36x faster than GenAx (ISCA 2018). GenCache in cache is 5.26x faster than GenAx.
  - GenCache chip has 16.4% higher area, 34.7% higher peak power, and 15% higher average power than GenAx.

GenCache's Four Phases



Phase 1	Phase 2	Phase 3	Phase 4
0 ERRORS	1 ERROR	2-5 ERRORS	6+ ERRORS
Seed Solver: MIN_SEARCH	Seed Solver: MIN_SEARCH	Seed Solver: HOBBES	Seed Solver: SMEM
Operations: HD	Operations: SHD	Operations: SHD_C MYERS_B SWA	Operations: MYERS_B SWA
48 MB REF 20 MB BLOOM 4 MB INDEX	48 MB REF 20 MB BLOOM 8 MB INDEX	40 MB REF 32 MB INDEX	24 MB REF 48 MB INDEX

Figure 7: Four phases in the new alignment algorithm that exploits in-cache operators.

Throughput Results

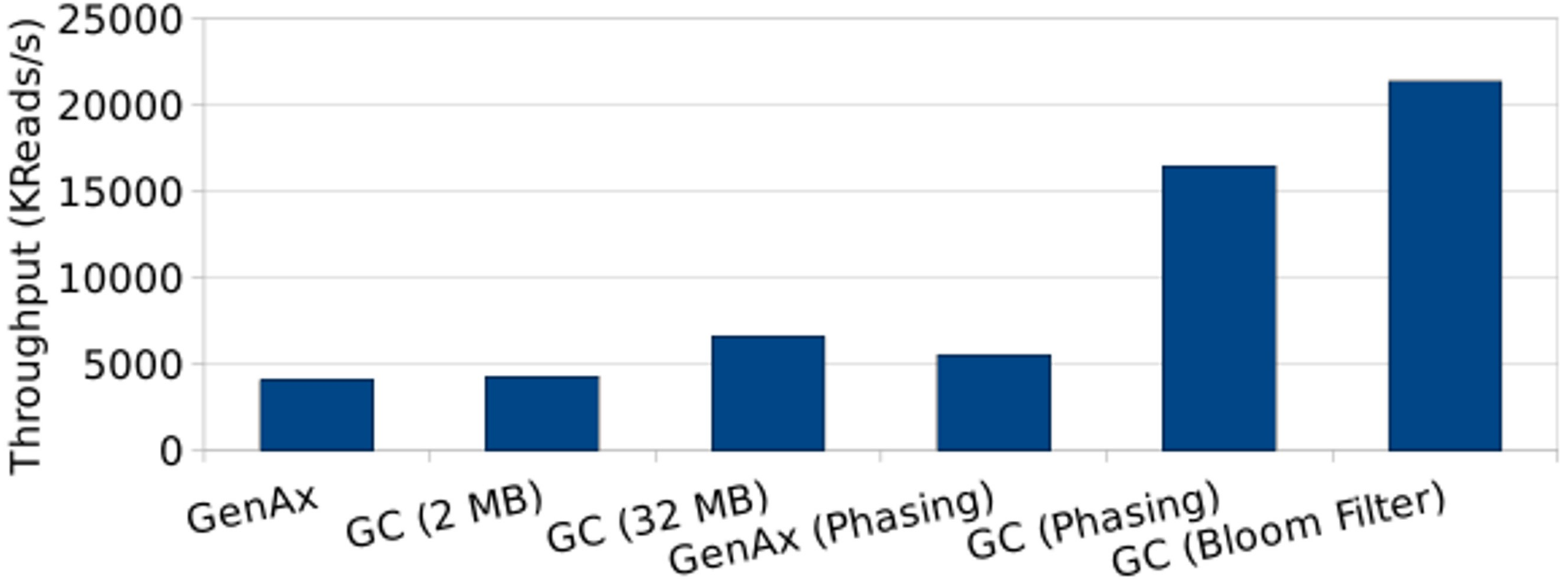


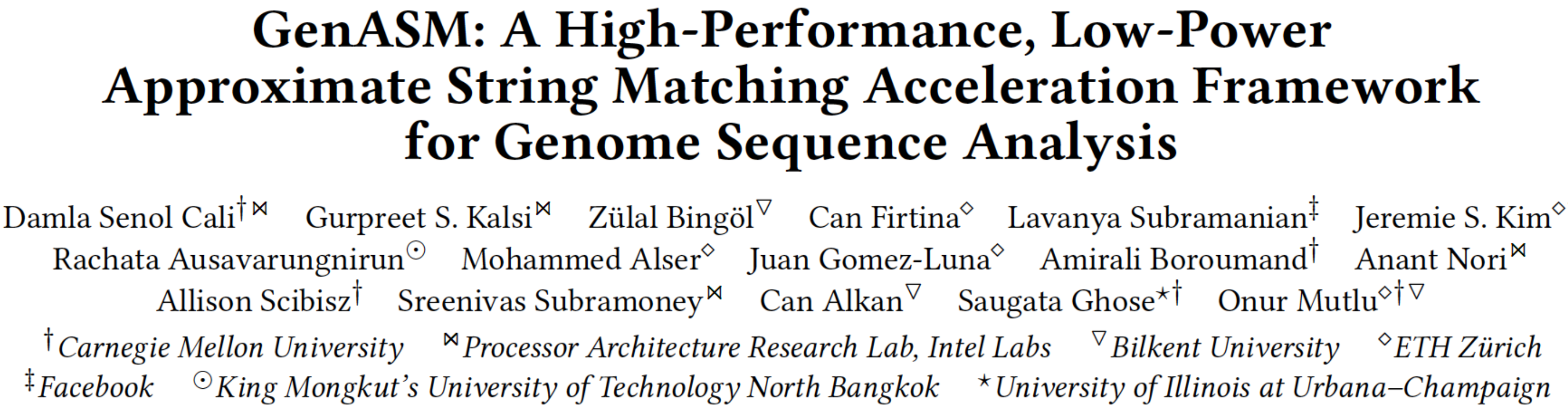
Figure 9: Throughput improvement of GenCache (Hardware & Software).

Ongoing Directions

- **Seed Filtering Technique:**
  - Goal: Reducing the number of seed (k-mer) locations.
    - Heuristic (limits the number of mapping locations for each seed).
    - Supports exact matches only.
- **Pre-alignment Filtering Technique:**
  - Goal: Reducing the number of *invalid mappings (>E)*.
    - Supports both exact and inexact matches.
    - Provides some falsely-accepted mappings.
- **Read Alignment Acceleration:**
  - Goal: Performing read alignment at scale.
    - Limits the numeric range of each cell in the DP table and hence supports limited scoring function.
    - May not support backtracking step due to random memory accesses.

GenASM Framework **[MICRO 2020]**

- Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, 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*

Near-memory GenASM Framework

- **Our goal:** Accelerate approximate string matching (ASM) by designing a fast and flexible framework, which can accelerate multiple steps of genome sequence analysis.

- **Key ideas:** Exploit the high memory bandwidth and the logic layer of 3D-stacked memory to perform highly-parallel ASM in the DRAM chip itself.
- Modify and extend Bitap[1,2], ASM algorithm with fast and simple bitwise operations, such that it now:
  - Supports long reads
  - Supports traceback
  - Is highly parallelizable
- Co-design of our modified scalable and memory-efficient algorithms with low-power and area-efficient hardware accelerators

[1] R. A. Baeza-Yates and G. H. Gonnet. "A New Approach to Text Searching." *CACM,* 1992.
[2] S. Wu and U. Manber. "Fast Text Searching: Allowing Errors." *CACM,* 1992.

Key Results of the GenASM Framework

(1) Read Alignment

- **116×** speedup, **37×** less power than **Minimap2** (state-of-the-art **SW**)
- **111×** speedup, **33×** less power than **BWA-MEM** (state-of-the-art **SW**)
- **3.9×** better throughput, **2.7×** less power than **Darwin** (state-of-the-art **HW**)
- **1.9×** better throughput, **82%** less logic power than **GenAx** (state-of-the-art **HW**)

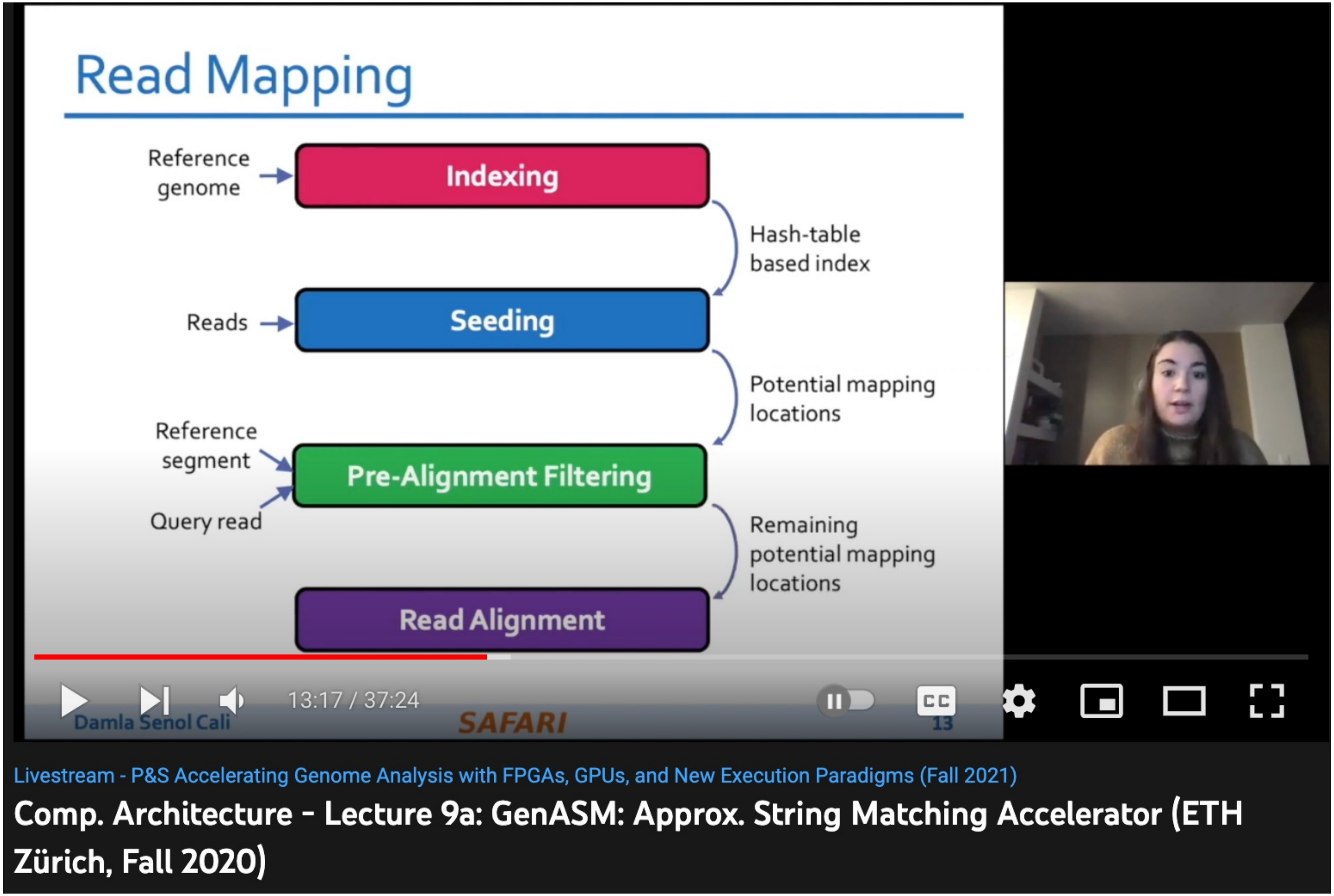
(2) Pre-Alignment Filtering

- **3.7×** speedup, **1.7×** less power than **Shouji** (state-of-the-art **HW**)

(3) Edit Distance Calculation

- **22–12501×** speedup, **548–582×** less power than **Edlib** (state-of-the-art **SW**)
- **9.3–400×** speedup, **67×** less power than **ASAP** (state-of-the-art **HW**)

More on GenASM



Read Mapping

Reference genome → Indexing

Hash-table based index

Reads → Seeding

Potential mapping locations

Reference segment → Pre-Alignment Filtering

Query read → Pre-Alignment Filtering

Remaining potential mapping locations

Read Alignment

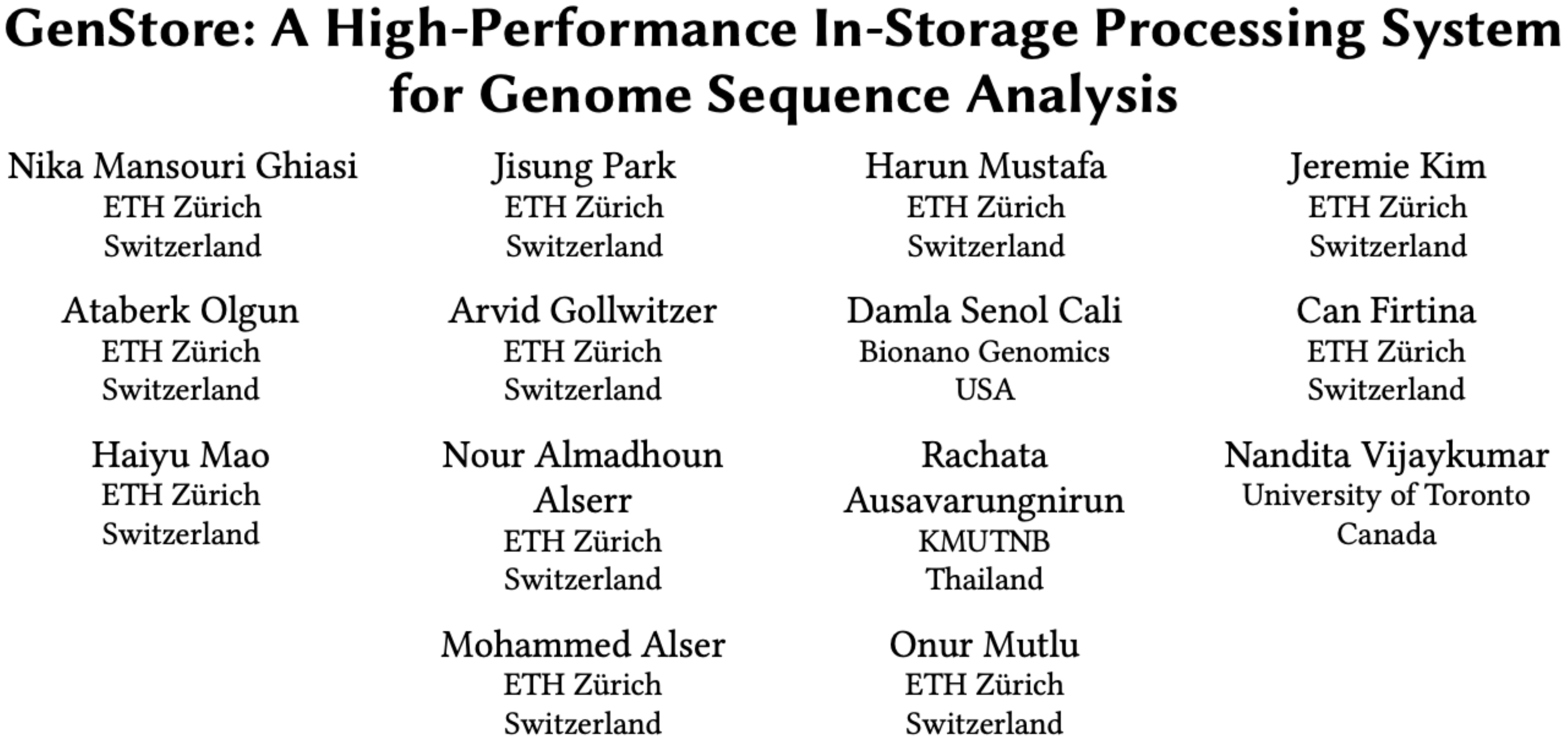
Livestream - P&S Accelerating Genome Analysis with FPGAs, GPUs, and New Execution Paradigms (Fall 2021)

Comp. Architecture - Lecture 9a: GenASM: Approx. String Matching Accelerator (ETH Zürich, Fall 2020)

https://www.youtube.com/watch?v=XoLpzmN-Pas

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

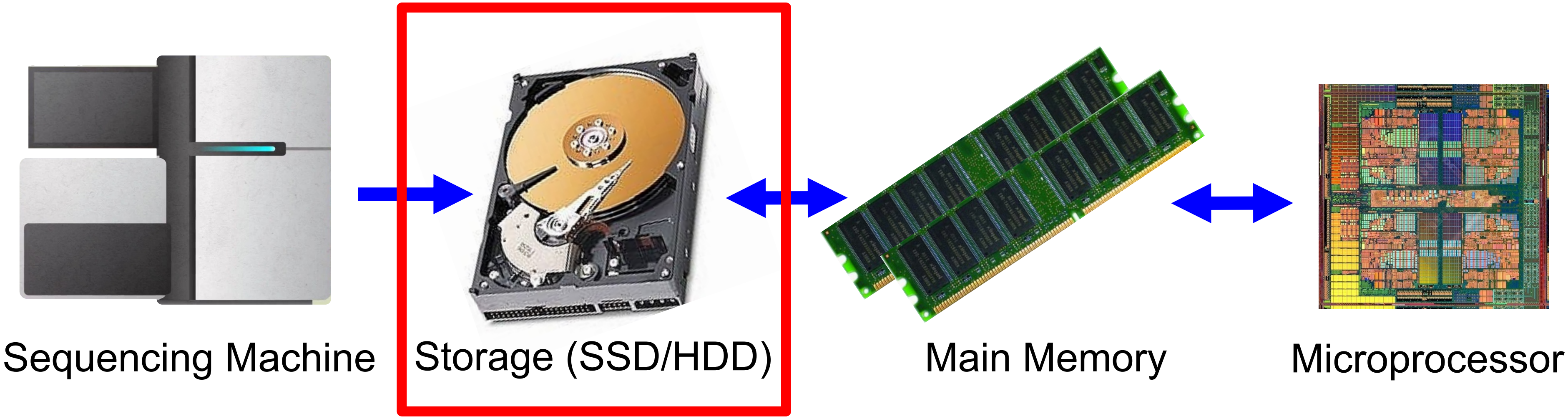
Mohammed Alser
ETH Zürich
Switzerland

Onur Mutlu
ETH Zürich
Switzerland

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

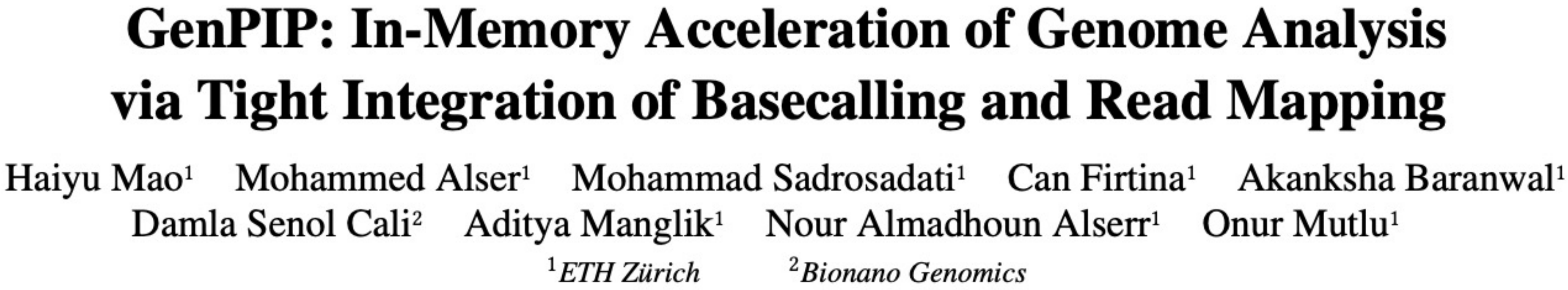
Microprocessor

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.

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.

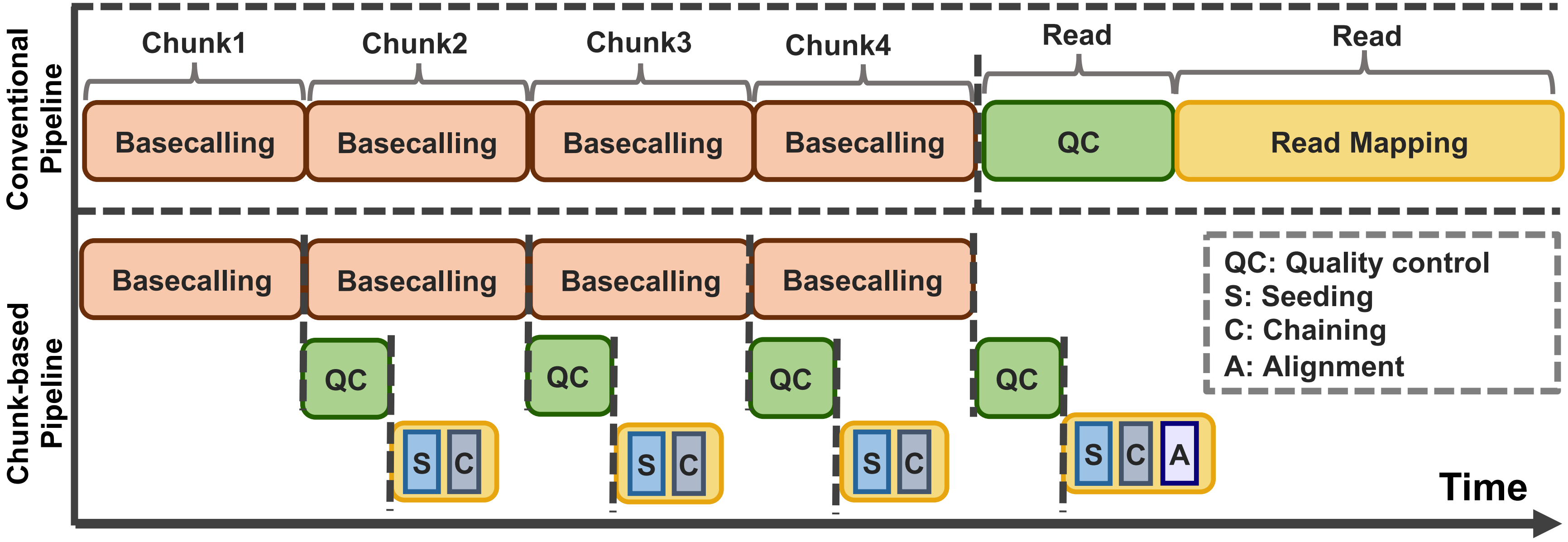


GenPIP: In-Memory Acceleration of Genome Analysis via Tight Integration of Basecalling and Read Mapping

Haiyu Mao[1] Mohammed Alser[1] Mohammad Sadrosadati[1] Can Firtina[1] Akanksha Baranwal[1]
Damla Senol Cali[2] Aditya Manglik[1] Nour Almadhoun Alserr[1] Onur Mutlu[1]
[1] *ETH Zürich* [2] *Bionano Genomics*

Innovations Require Change

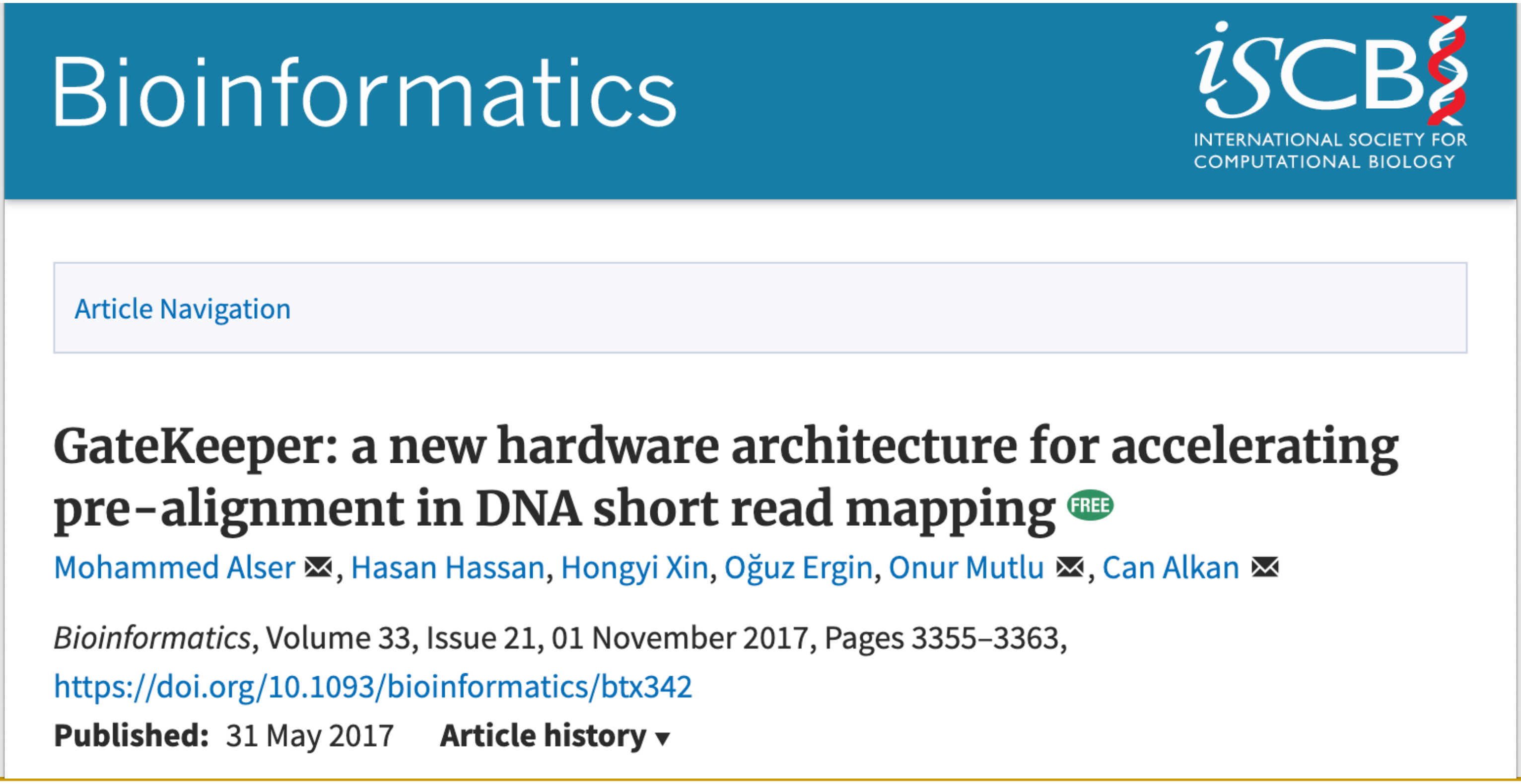
- CP processes reads at the granularity of a chunk instead of the complete read sequence, increasing parallelism and resource utilization by overlapping the execution of different steps.



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.

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

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

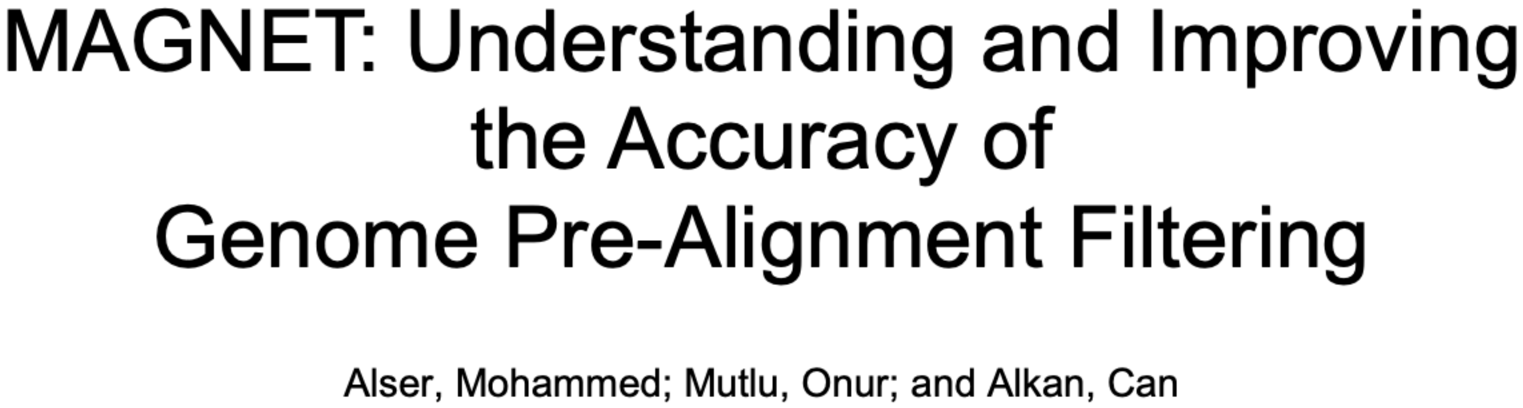
Mohammed Alser ✉, Hasan Hassan, Hongyi Xin, Oğuz Ergin, Onur Mutlu ✉, Can Alkan ✉

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]

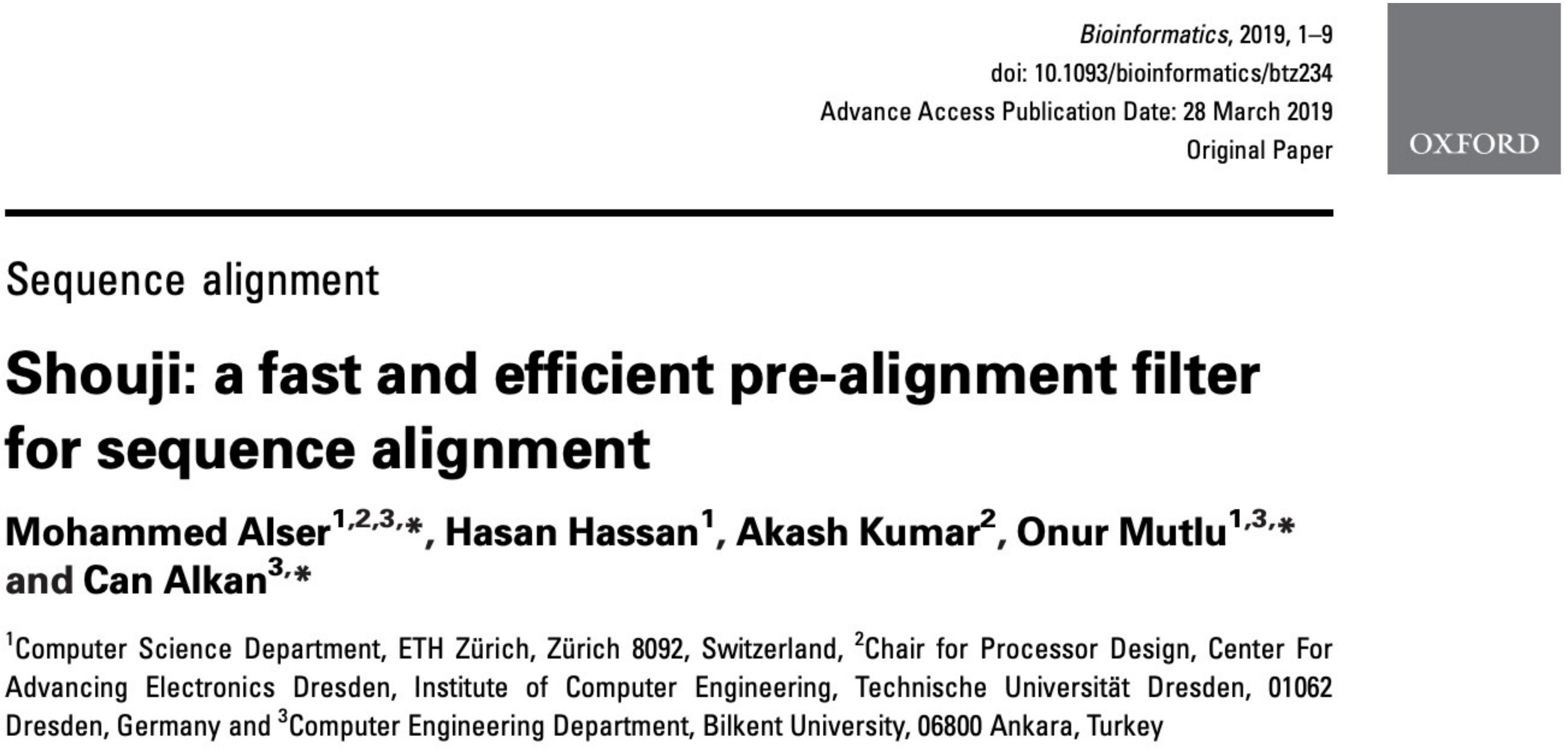


MAGNET: Understanding and Improving the Accuracy of Genome Pre-Alignment Filtering

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

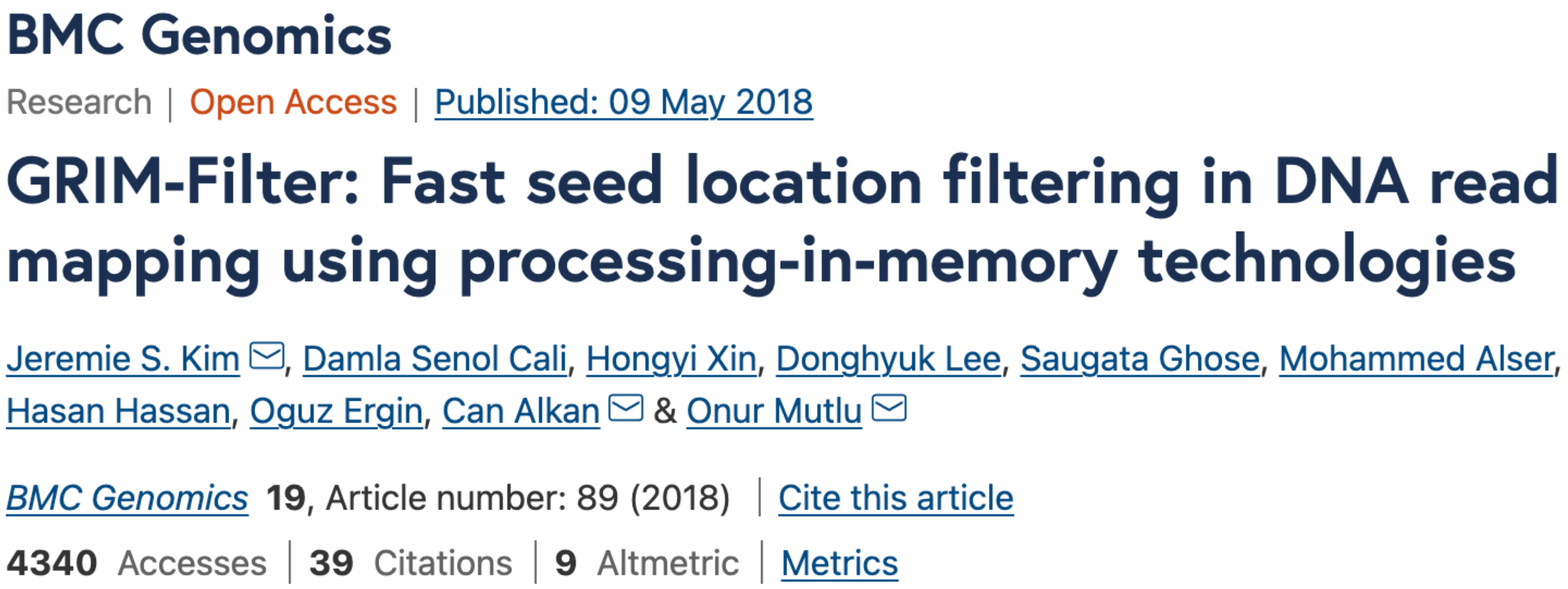
Shouji: a fast and efficient pre-alignment filter for sequence alignment

Mohammed Alser[1,2,3,*], Hasan Hassan[1], Akash Kumar[2], Onur Mutlu[1,3,*] and Can Alkan[3,*]

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

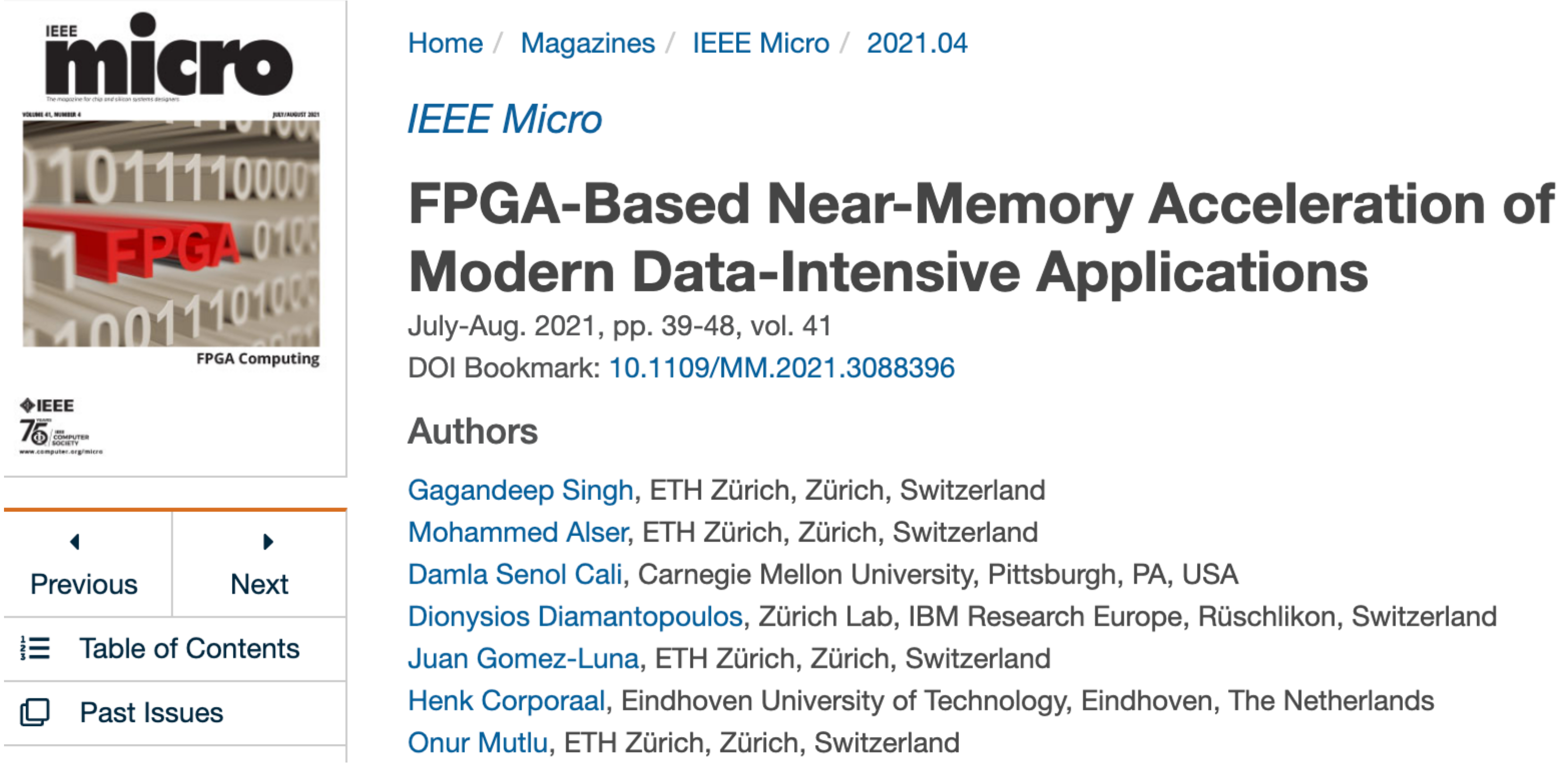
Jeremie S. Kim ✉, Damla Senol Cali, Hongyi Xin, Donghyuk Lee, Saugata Ghose, Mohammed Alser, Hasan Hassan, Oguz Ergin, Can Alkan ✉ & Onur Mutlu ✉

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

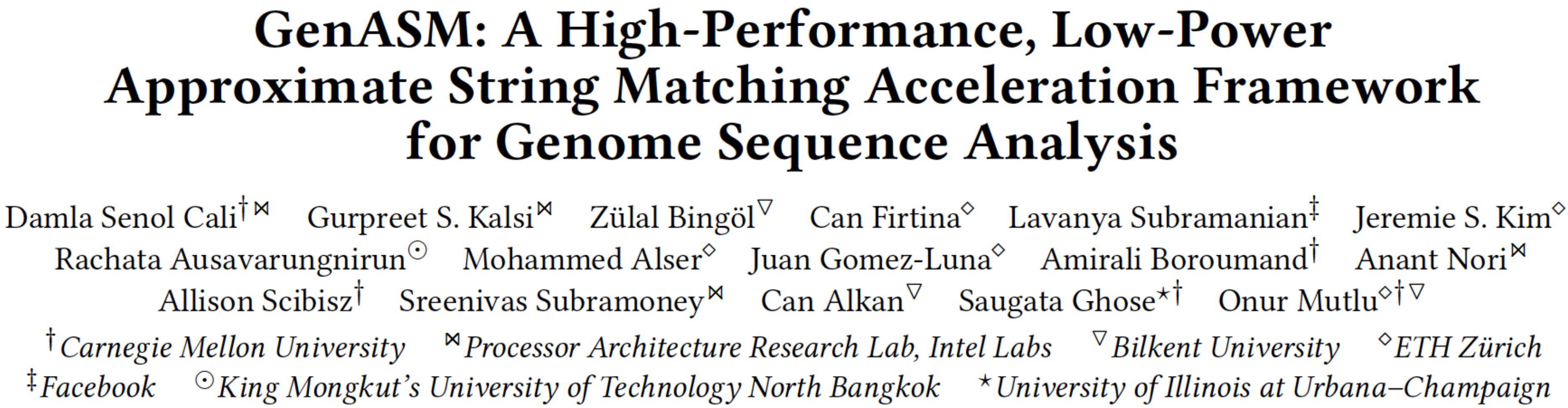
July-Aug. 2021, pp. 39-48, vol. 41
DOI Bookmark: 10.1109/MM.2021.3088396

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

GenASM Framework [MICRO 2020]

- Damla Senol Cali, Gurpreet S. Kalsi, Zulal Bingol, 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)]



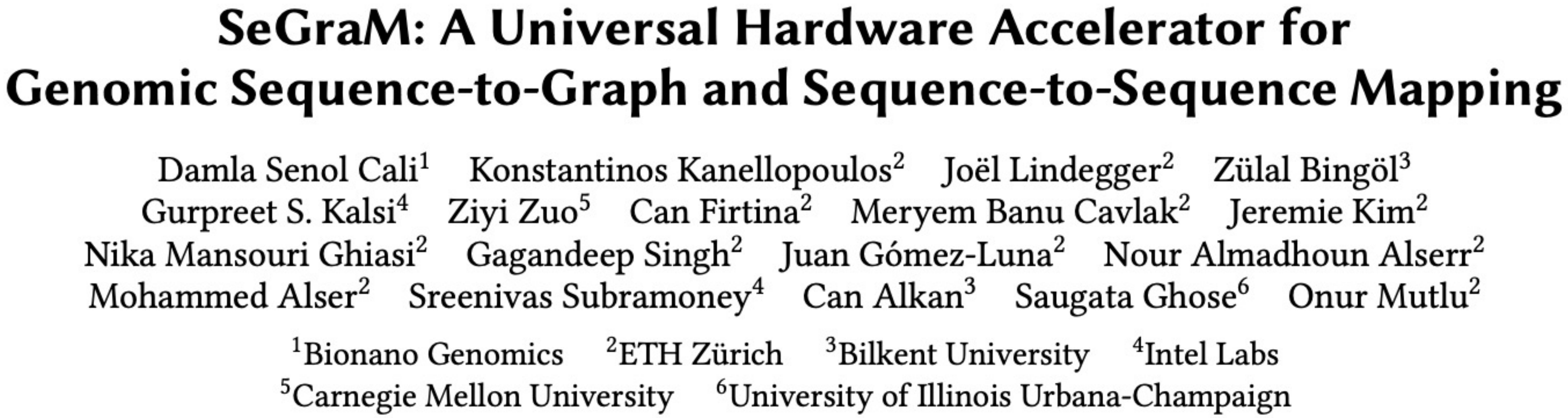
GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis

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*

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"
ISCA 2022



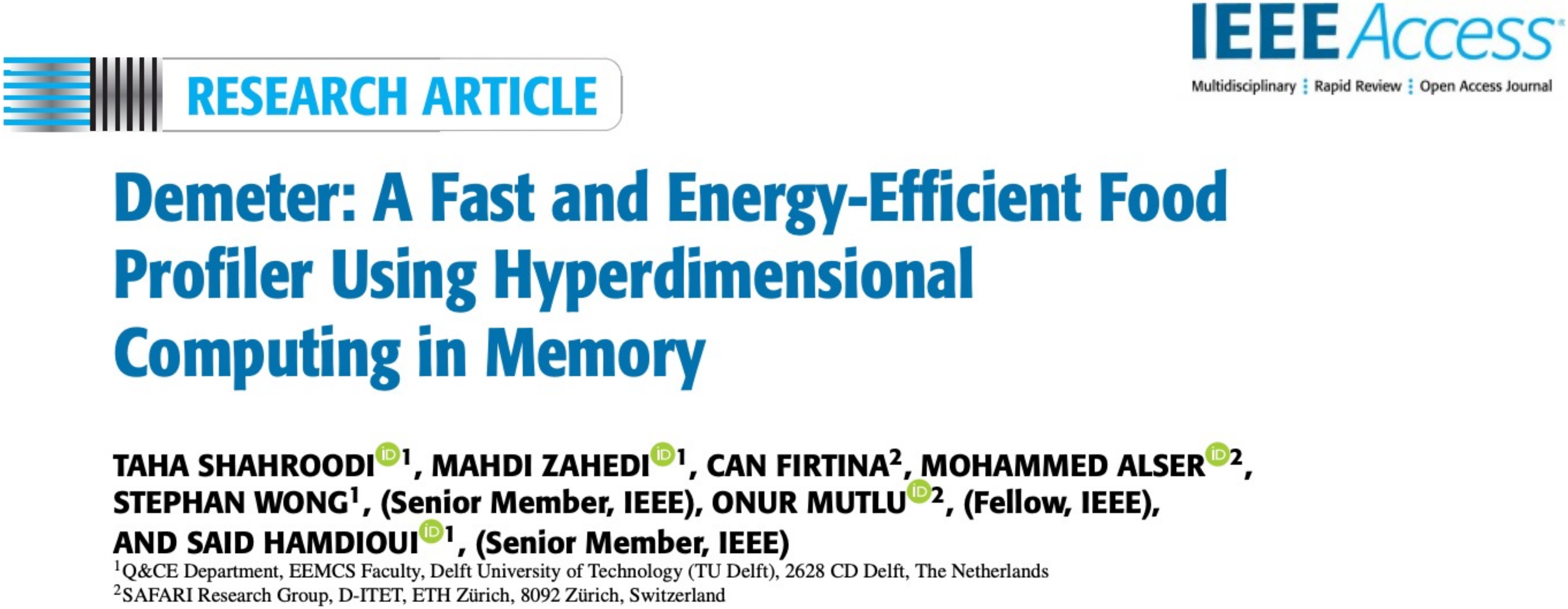
SeGraM: A Universal Hardware Accelerator for Genomic Sequence-to-Graph and Sequence-to-Sequence Mapping

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

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

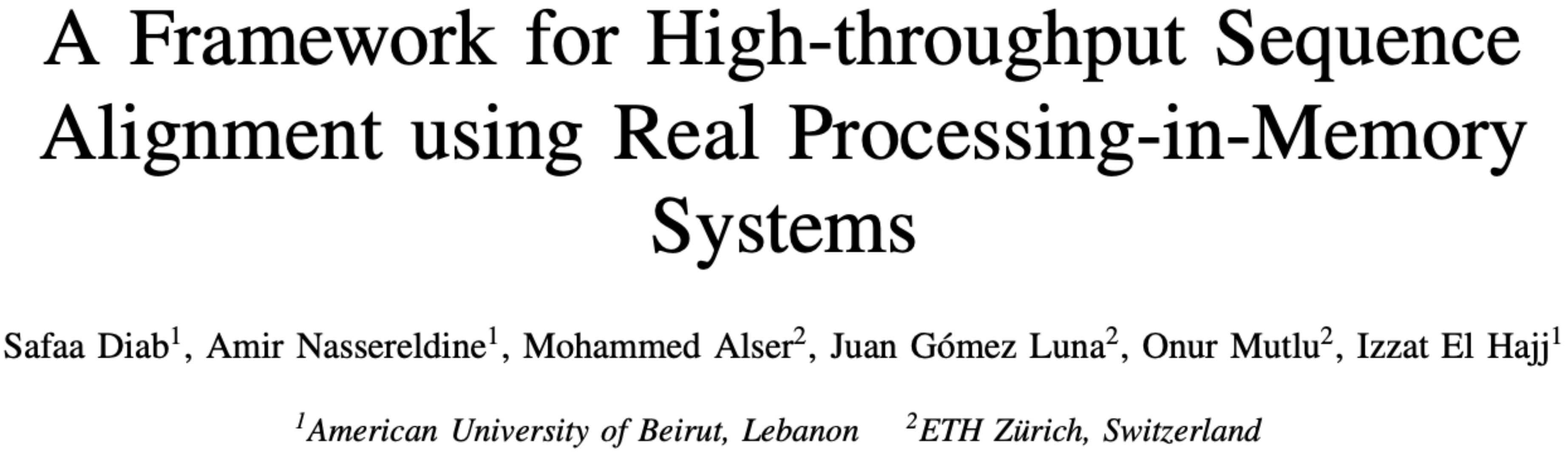
Demeter: A Fast and Energy-Efficient Food Profiler Using Hyperdimensional Computing in Memory

TAHA SHAHROODI[1], MAHDI ZAHEDI[1], CAN FIRTINA[2], MOHAMMED ALSER[2], STEPHAN WONG[1], (Senior Member, IEEE), ONUR MUTLU[2], (Fellow, IEEE), AND SAID HAMDIOUI[1], (Senior Member, IEEE)

[1]Q&CE Department, EEMCS Faculty, Delft University of Technology (TU Delft), 2628 CD Delft, The Netherlands
[2]SAFARI 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 Diab[1], Amir Nassereldine[1], Mohammed Alser[2], Juan Gómez Luna[2], Onur Mutlu[2], Izzat El Hajj[1]

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

Our Contributions

Near-memory/In-memory Pre-alignment Filtering

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

In-storage Sequence Alignment

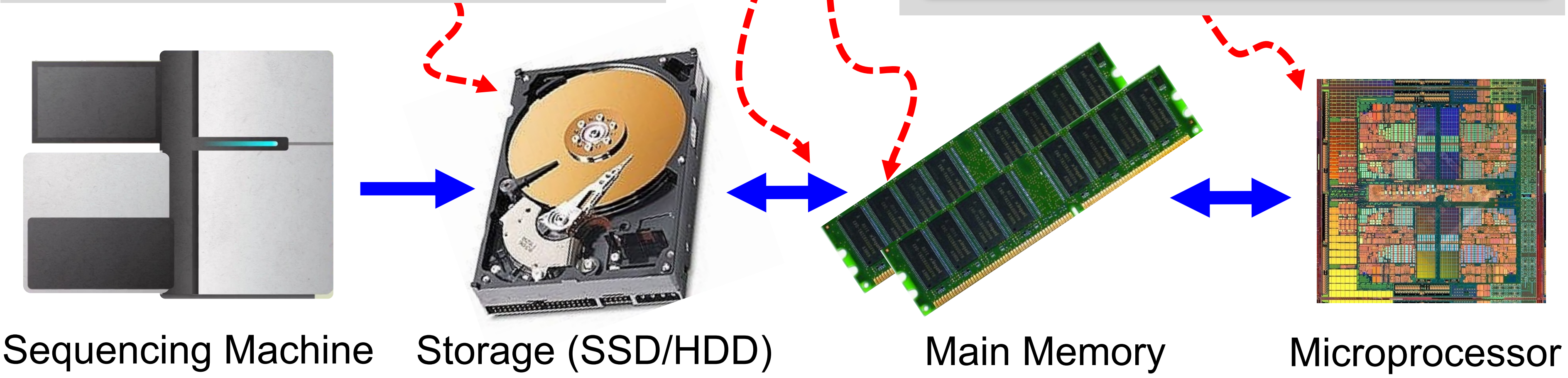
- **GenStore [ASPLOS 2022]**

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]**



Conclusion on Ongoing Directions

- Read alignment can be substantially accelerated using computationally inexpensive and accurate pre-alignment filtering algorithms designed for specialized hardware.

- All the three directions are used by mappers today, but filtering has replaced alignment as the bottleneck.

- Pre-alignment filtering does *not* sacrifice any of the aligner capabilities, as it does *not* modify or replace the alignment step.

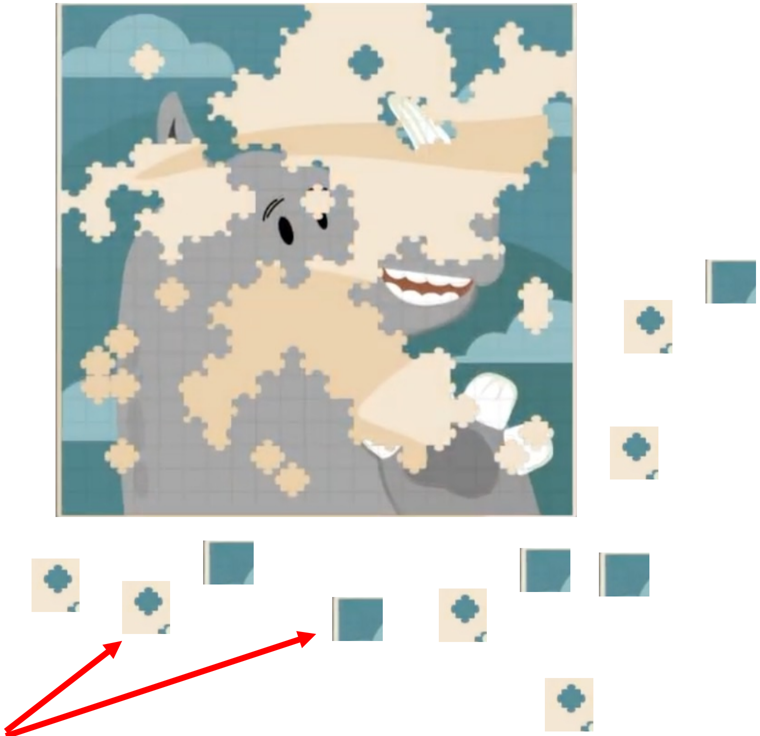
What else can be done?

What if we got a new version of the reference genome?

.FASTA file

Reference genome

.FASTQ file



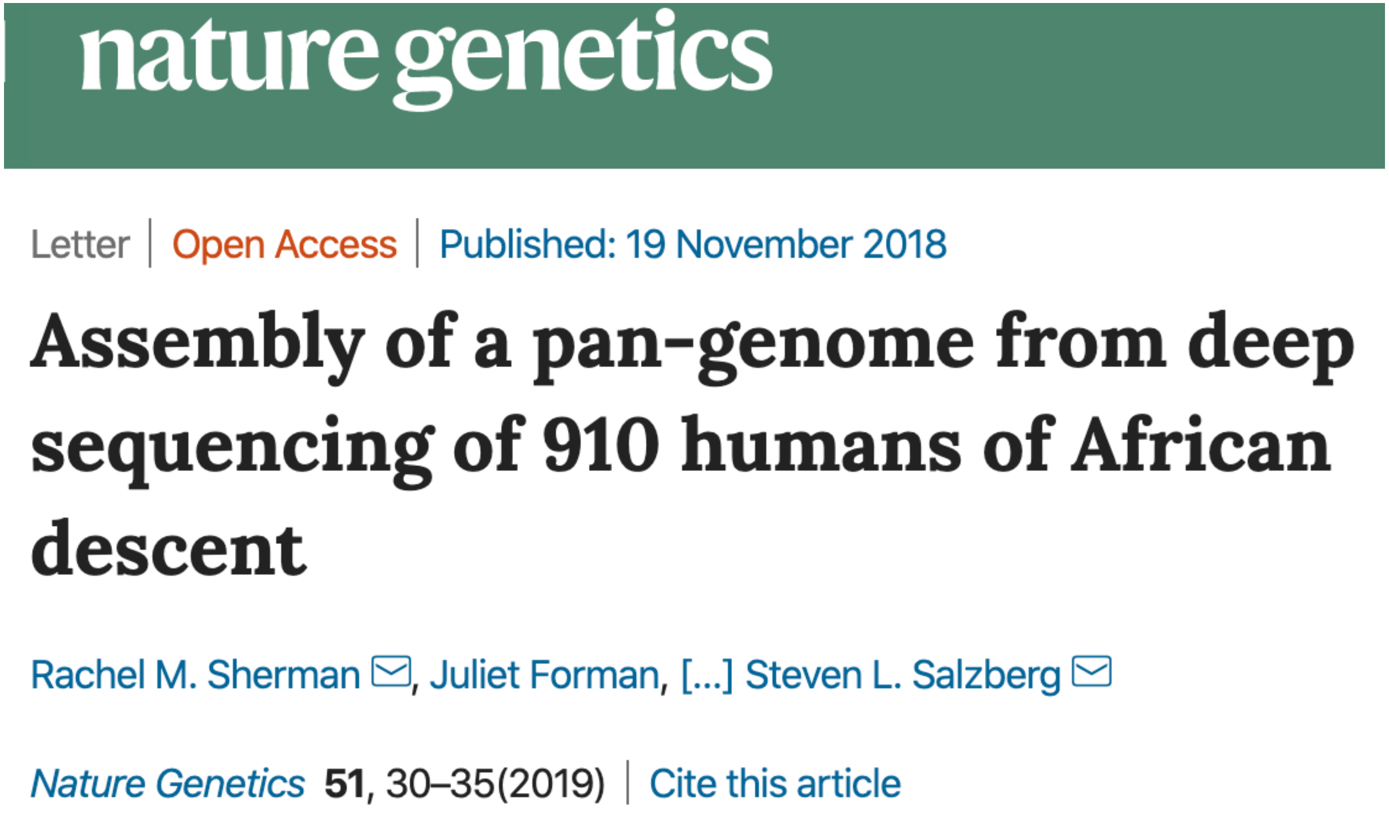
Reads

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

Revisiting the Puzzle

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

Reference Genome Bias



nature genetics

Letter | Open Access | Published: 19 November 2018

Assembly of a pan-genome from deep sequencing of 910 humans of African descent

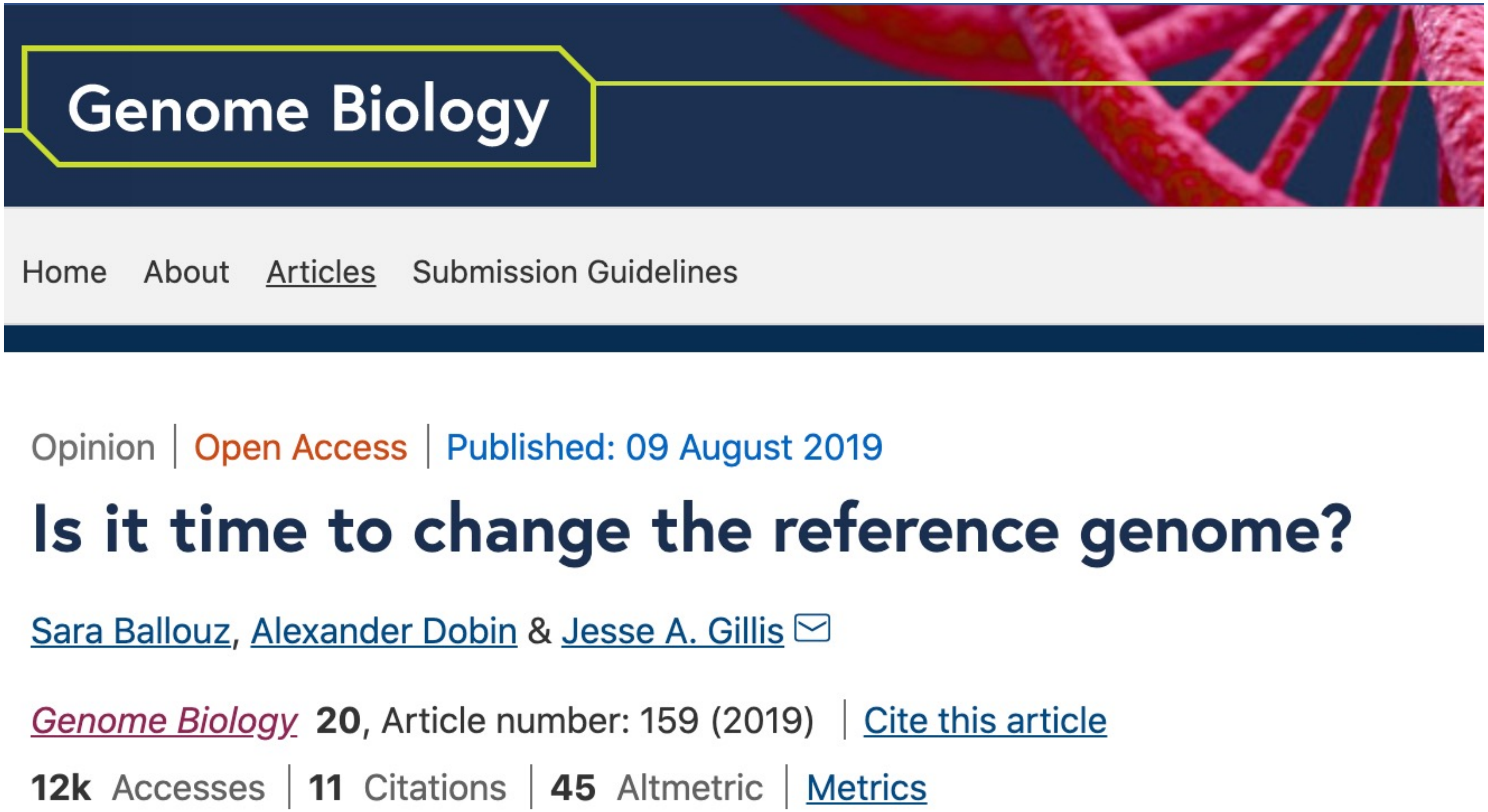
Rachel M. Sherman, Juliet Forman, [...] Steven L. Salzberg

Nature Genetics **51**, 30–35(2019) | Cite this article

"African pan-genome contains ~10% more DNA bases than the current human reference genome"

Sherman+, "Assembly of a pan-genome from deep sequencing of 910 humans of African descent" *Nature genetics*, 2019.

Time to Change the Reference Genome



Genome Biology

Home About Articles Submission Guidelines

Opinion | Open Access | Published: 09 August 2019

Is it time to change the reference genome?

Sara Ballouz, Alexander Dobin & Jesse A. Gillis

Genome Biology **20**, Article number: 159 (2019) | Cite this article

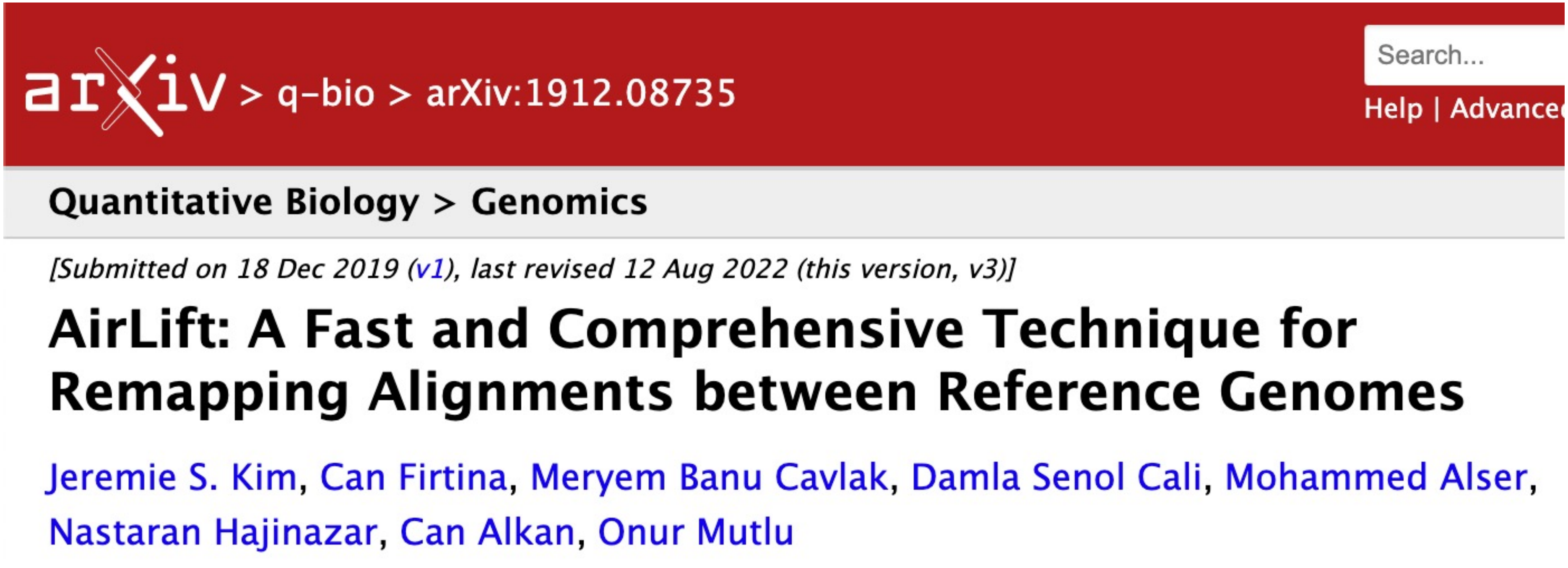
12k Accesses | **11** Citations | **45** Altmetric | Metrics

"Switching to a consensus reference would offer important advantages over the continued use of the current reference with few disadvantages"

Ballouz+, "Is it time to change the reference genome?", Genome Biology, 2019

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



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

AirLift

- **Key observation:** Reference genomes are updated frequently. Repeating *read mapping is a computationally expensive workload*.

- **Key idea:** Update the mapping results of only affected reads depending on how a region in the old reference relates to another region in the new reference.

- **Key results:**
  - reduces number of reads that needs to be re-mapped to new reference by up to 99%
  - reduces overall runtime to re-map reads by 6.94x, 208x, and 16.4x for large (human), medium (C. elegans), and small (yeast) reference genomes

Clustering the Reference Genome Regions

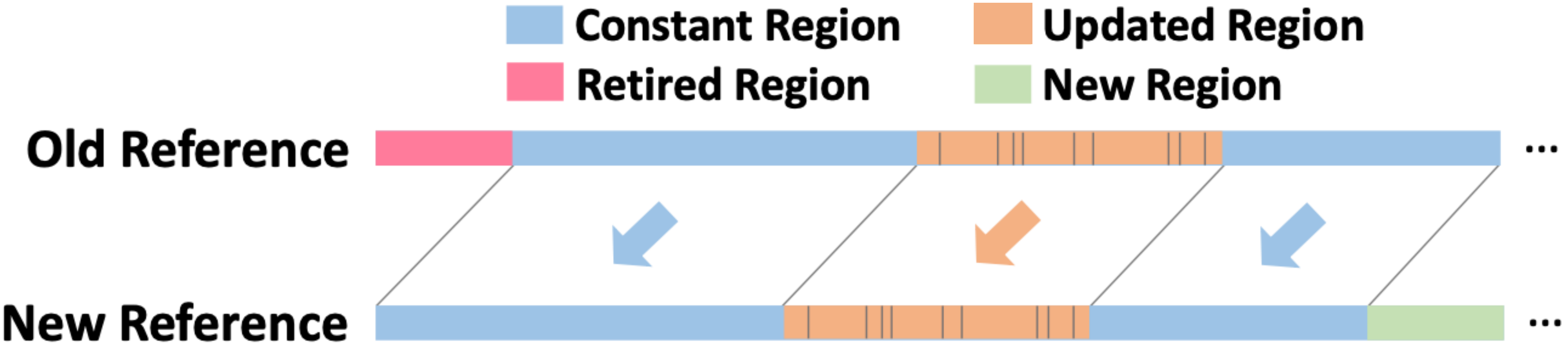
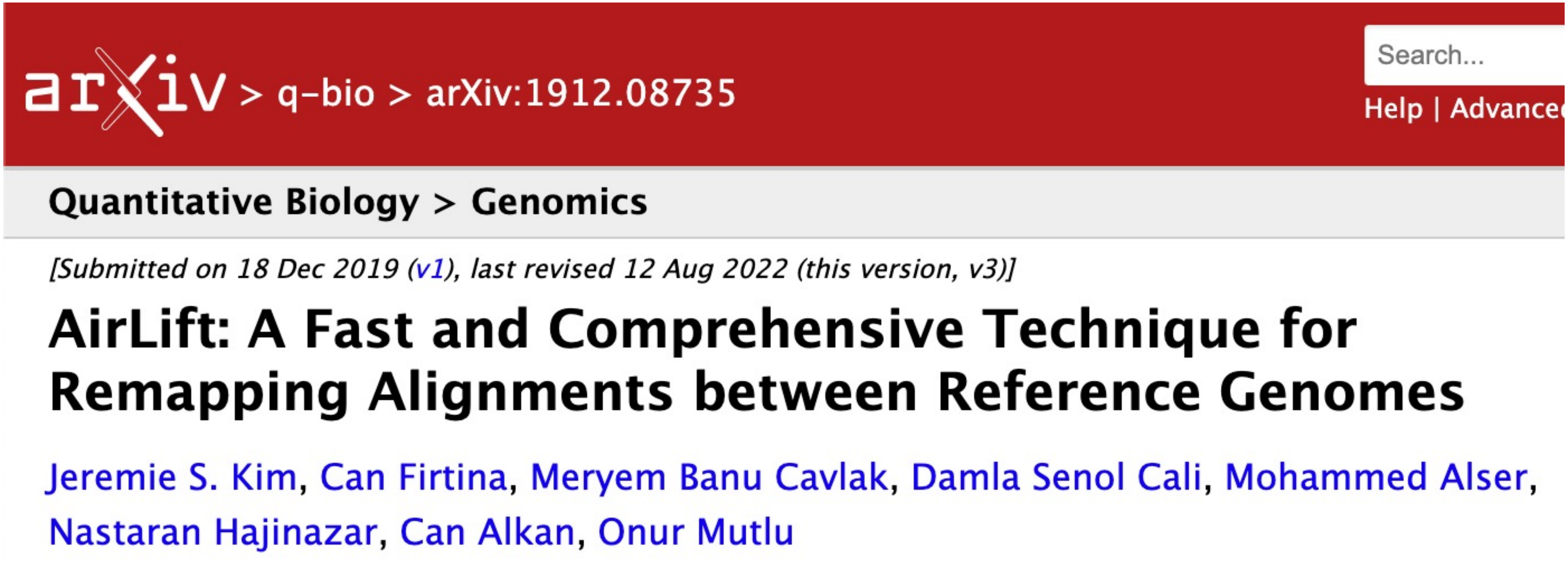


Fig. 2. Reference Genome Regions.

More Details on 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



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

Agenda for Today

- What is Genome Analysis?
- What is Intelligent Genome Analysis?

- How we Analyze Genome?
- What are the Barriers to Enabling Intelligent Analyses?

- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

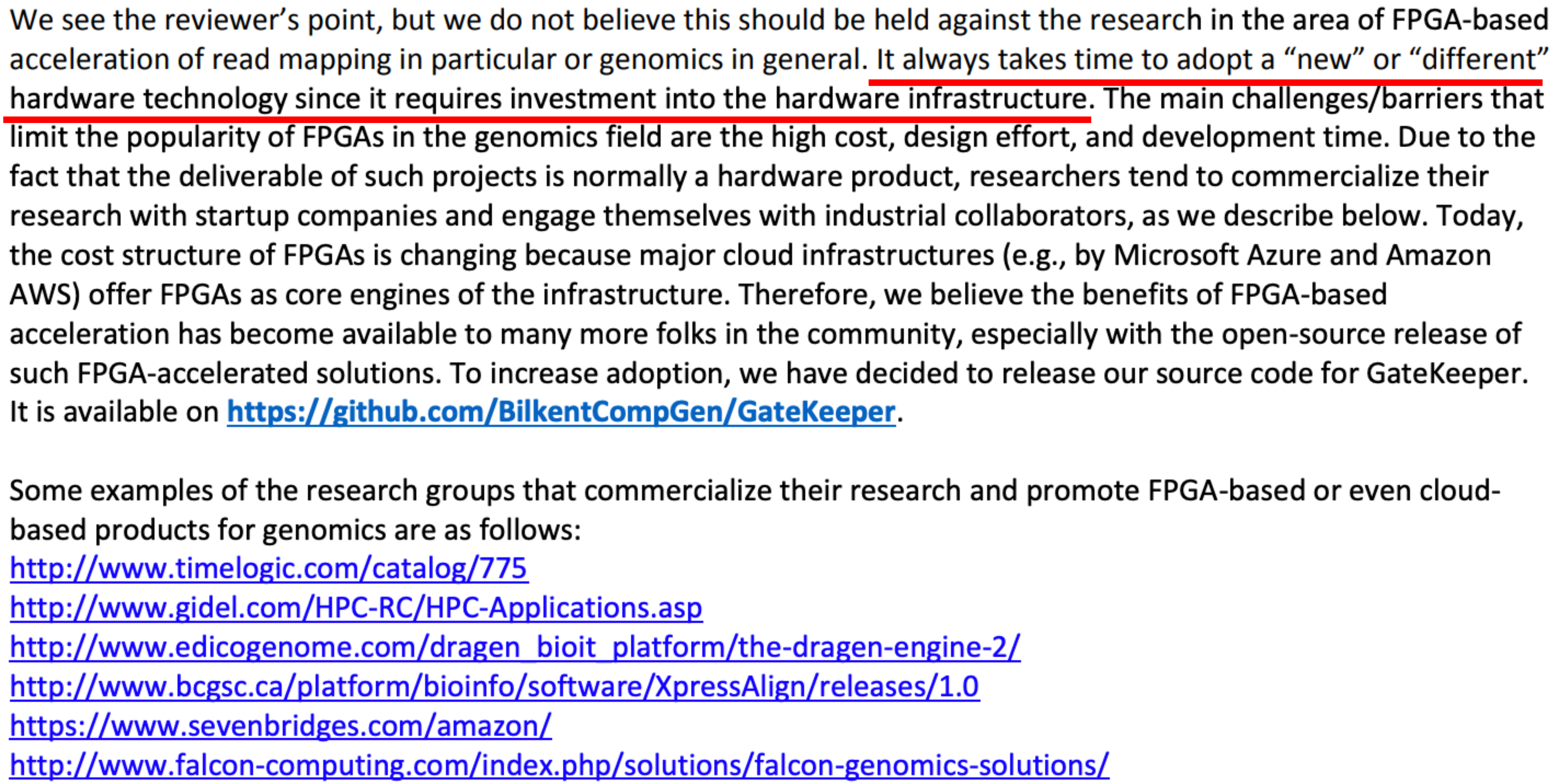
- **Where is Genomic Analyses Going Next?**

Adoption of hardware accelerators in genome analysis

Bioinformatics: **Reviewer #6** (Dec. 2016)

I have a major concern with the work that is actually not a problem with the manuscript at all. Specifically, I have the concern that there has been little to no adoption of previous specialized hardware solutions related to improving the speed of alignment. While there has been considerable work in this area (which the authors do an admirable job of citing), it does not seem that these hardware-based solutions have gained any type of real traction in the community, as the vast majority of alignment is still performed on "regular" CPUs, where the extent of hardware acceleration is the adoption of specific SIMD or vectorized instructions. While I don't think that this practical concern should preclude publication of the current work, it is something worth considering (what, if any, of the proposed improvements to the SHD filter could be "back-ported" to a software-only solution).

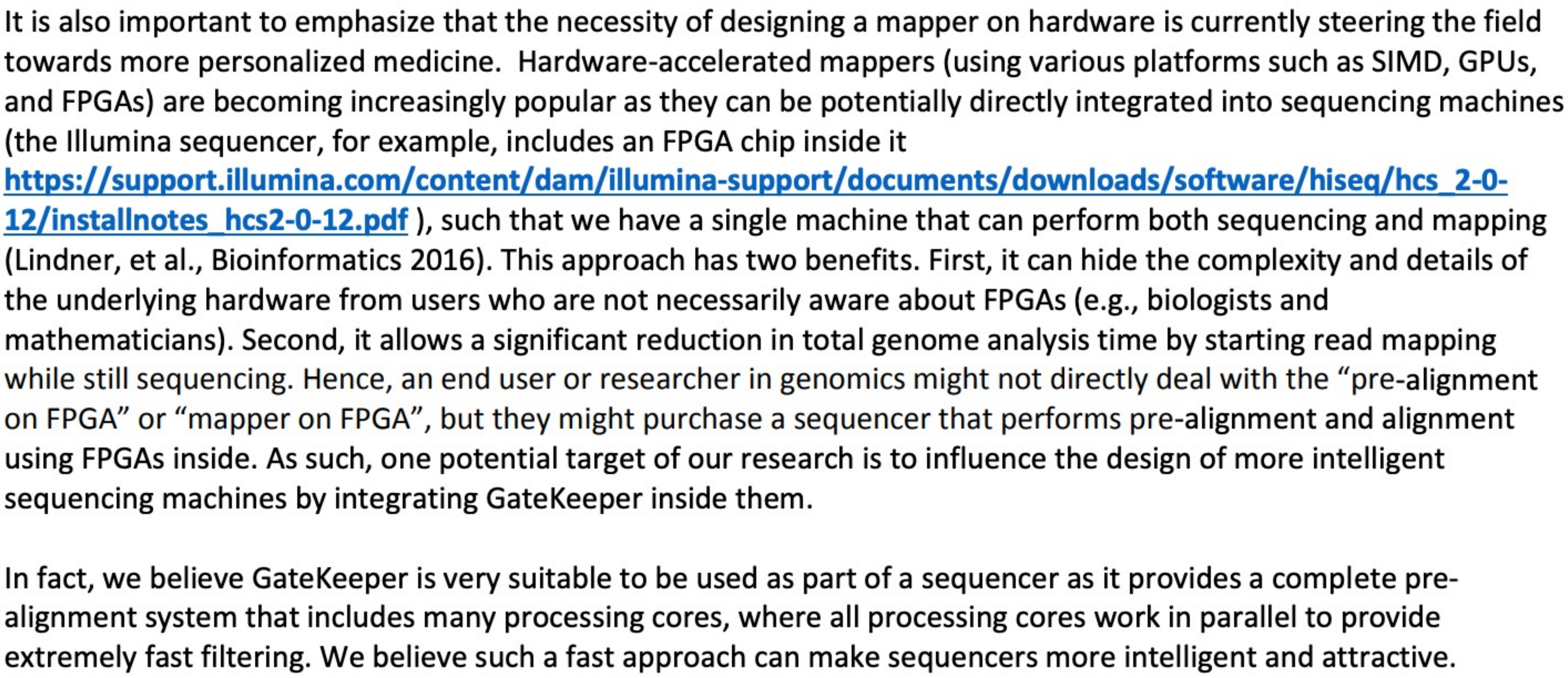
Our Response



We see the reviewer's point, but we do not believe this should be held against the research in the area of FPGA-based acceleration of read mapping in particular or genomics in general. It always takes time to adopt a "new" or "different" hardware technology since it requires investment into the hardware infrastructure. The main challenges/barriers that limit the popularity of FPGAs in the genomics field are the high cost, design effort, and development time. Due to the fact that the deliverable of such projects is normally a hardware product, researchers tend to commercialize their research with startup companies and engage themselves with industrial collaborators, as we describe below. Today, the cost structure of FPGAs is changing because major cloud infrastructures (e.g., by Microsoft Azure and Amazon AWS) offer FPGAs as core engines of the infrastructure. Therefore, we believe the benefits of FPGA-based acceleration has become available to many more folks in the community, especially with the open-source release of such FPGA-accelerated solutions. To increase adoption, we have decided to release our source code for GateKeeper. It is available on https://github.com/BilkentCompGen/GateKeeper.

Some examples of the research groups that commercialize their research and promote FPGA-based or even cloud-based products for genomics are as follows:
http://www.timelogic.com/catalog/775
http://www.gidel.com/HPC-RC/HPC-Applications.asp
http://www.edicogenome.com/dragen_bioit_platform/the-dragen-engine-2/
http://www.bcgsc.ca/platform/bioinfo/software/XpressAlign/releases/1.0
https://www.sevenbridges.com/amazon/
http://www.falcon-computing.com/index.php/solutions/falcon-genomics-solutions/

Our Response (cont'd)



It is also important to emphasize that the necessity of designing a mapper on hardware is currently steering the field towards more personalized medicine. Hardware-accelerated mappers (using various platforms such as SIMD, GPUs, and FPGAs) are becoming increasingly popular as they can be potentially directly integrated into sequencing machines (the Illumina sequencer, for example, includes an FPGA chip inside it https://support.illumina.com/content/dam/illumina-support/documents/downloads/software/hiseq/hcs_2-0-12/installnotes_hcs2-0-12.pdf), such that we have a single machine that can perform both sequencing and mapping (Lindner, et al., Bioinformatics 2016). This approach has two benefits. First, it can hide the complexity and details of the underlying hardware from users who are not necessarily aware about FPGAs (e.g., biologists and mathematicians). Second, it allows a significant reduction in total genome analysis time by starting read mapping while still sequencing. Hence, an end user or researcher in genomics might not directly deal with the "pre-alignment on FPGA" or "mapper on FPGA", but they might purchase a sequencer that performs pre-alignment and alignment using FPGAs inside. As such, one potential target of our research is to influence the design of more intelligent sequencing machines by integrating GateKeeper inside them.

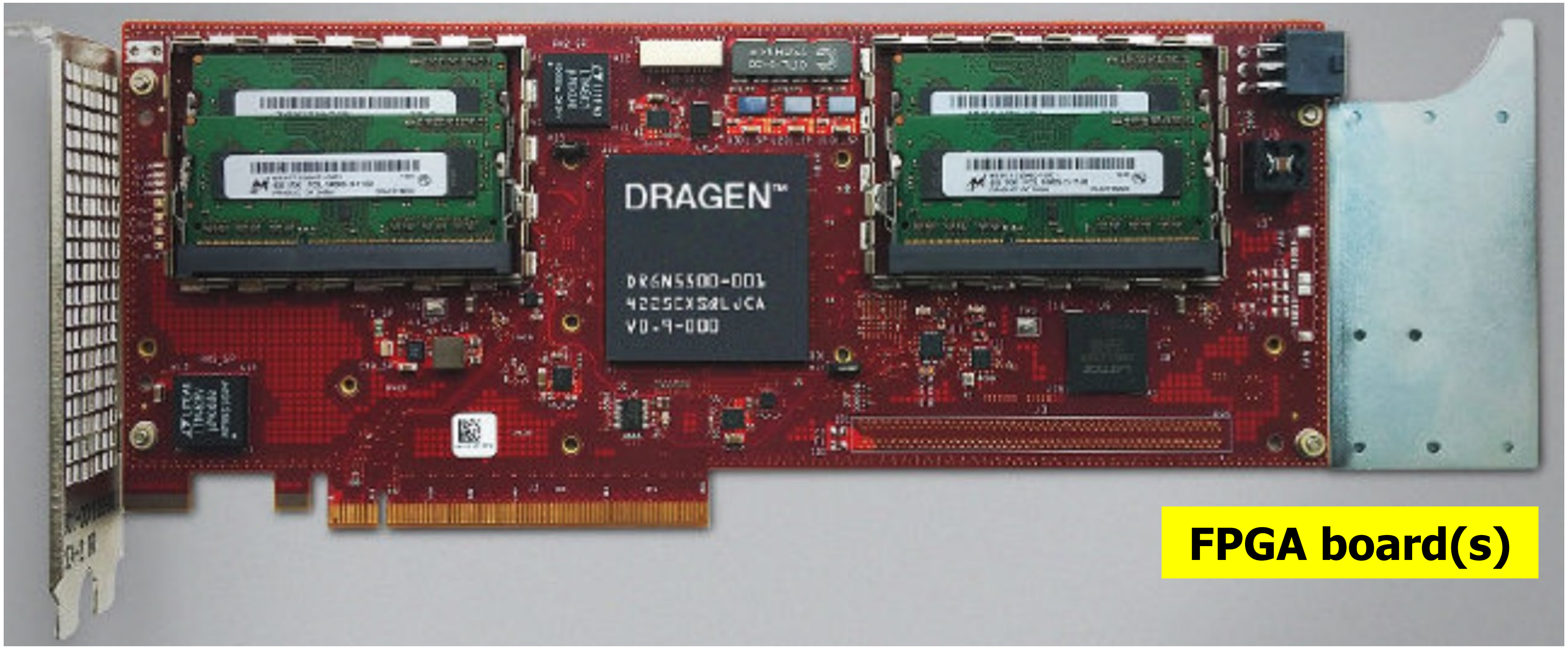
In fact, we believe GateKeeper is very suitable to be used as part of a sequencer as it provides a complete pre-alignment system that includes many processing cores, where all processing cores work in parallel to provide extremely fast filtering. We believe such a fast approach can make sequencers more intelligent and attractive.

Dream
and, they will come

Computing landscape is very different from 10-20 years ago

Illumina DRAGEN Bio-IT Platform (2018)

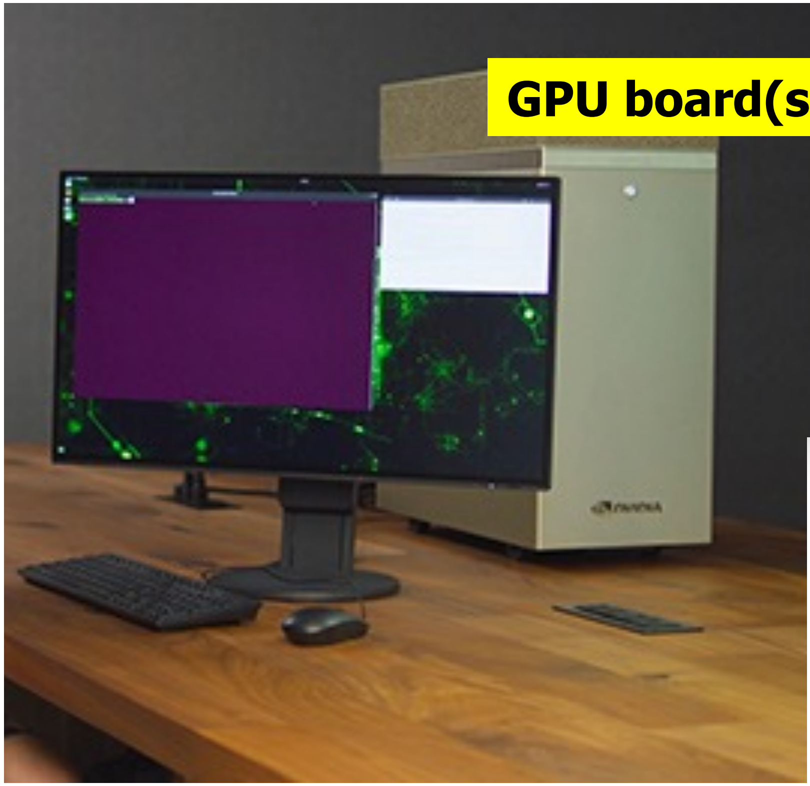
- Processes whole genome at 30x coverage in ~25 minutes with hardware support for data compression



FPGA board(s)

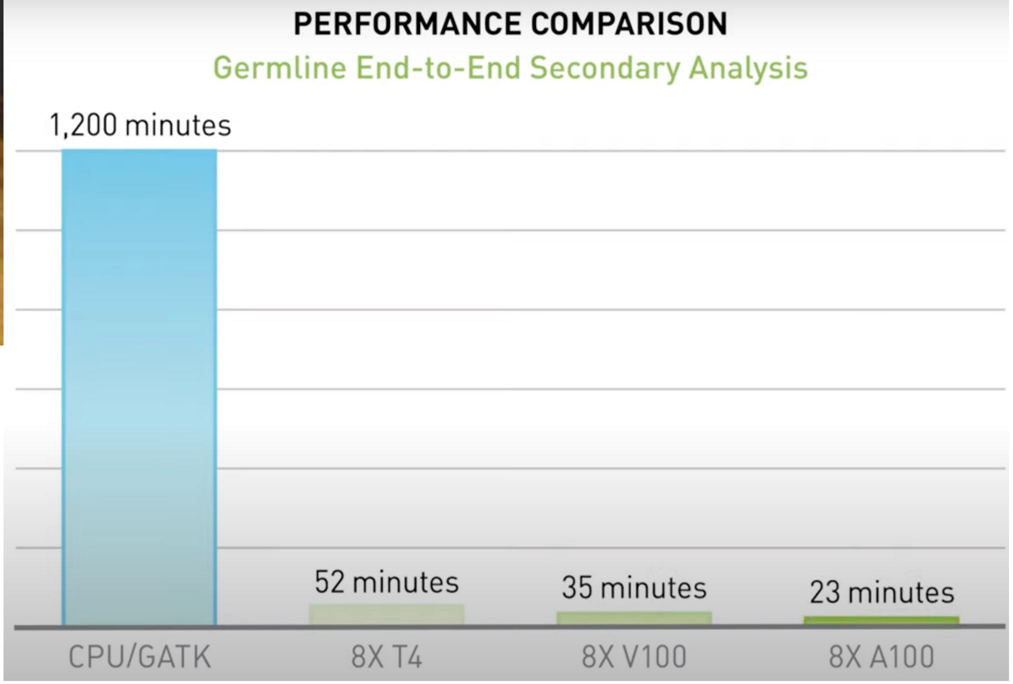
emea.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html
emea.illumina.com/company/news-center/press-releases/2018/2349147.html

NVIDIA Clara Parabricks (2020)



GPU board(s)

A University of Michigan's startup in 2018 and joined NVIDIA in 2020



https://developer.nvidia.com/clara-parabricks

Computing is Still Bottlenecked by Data Movement

Adoption Challenges of Hardware Accelerators

- Accelerate the entire read mapping process rather than its individual steps (Amdahl's law)

- Reduce the high amount of data movement
  - Working directly on compressed data
  - Filter out unlikely-reused data at the very first component of the compute system

- Develop flexible hardware architectures that do NOT conservatively **limit the range** of supported parameter values at design time

- Adapt existing genomic data formats for hardware accelerators or develop more efficient file formats

Adoption Challenges of Hardware Accelerators

- Maintaining the same (or better) accuracy/sensitivity of the output results of the software version
  - Using heuristic algorithms to gain speedup!

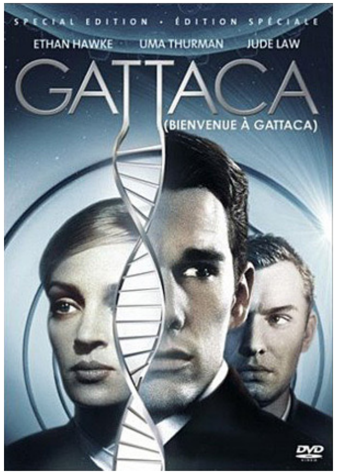
- High hardware cost

- Long development life-cycle for FPGA platforms

Did we Achieve Our Goal?

- **Fast** genome analysis in mere seconds using limited computational resources (i.e., personal computer or small hardware).

1997


SPECIAL EDITION · ÉDITION SPÉCIALE
ETHAN HAWKE UMA THURMAN JUDE LAW
GATTACA
(BIENVENUE À GATTACA)
DVD

2015

FROM THE DIRECTOR OF THE INCREDIBLES & MISSION: IMPOSSIBLE - GHOST PROTOCOL
GEORGE CLOONEY
TOMORROWLAND
REMEMBER TOMORROW MAY 22
EXPERIENCE IT IN IMAX

Open Questions

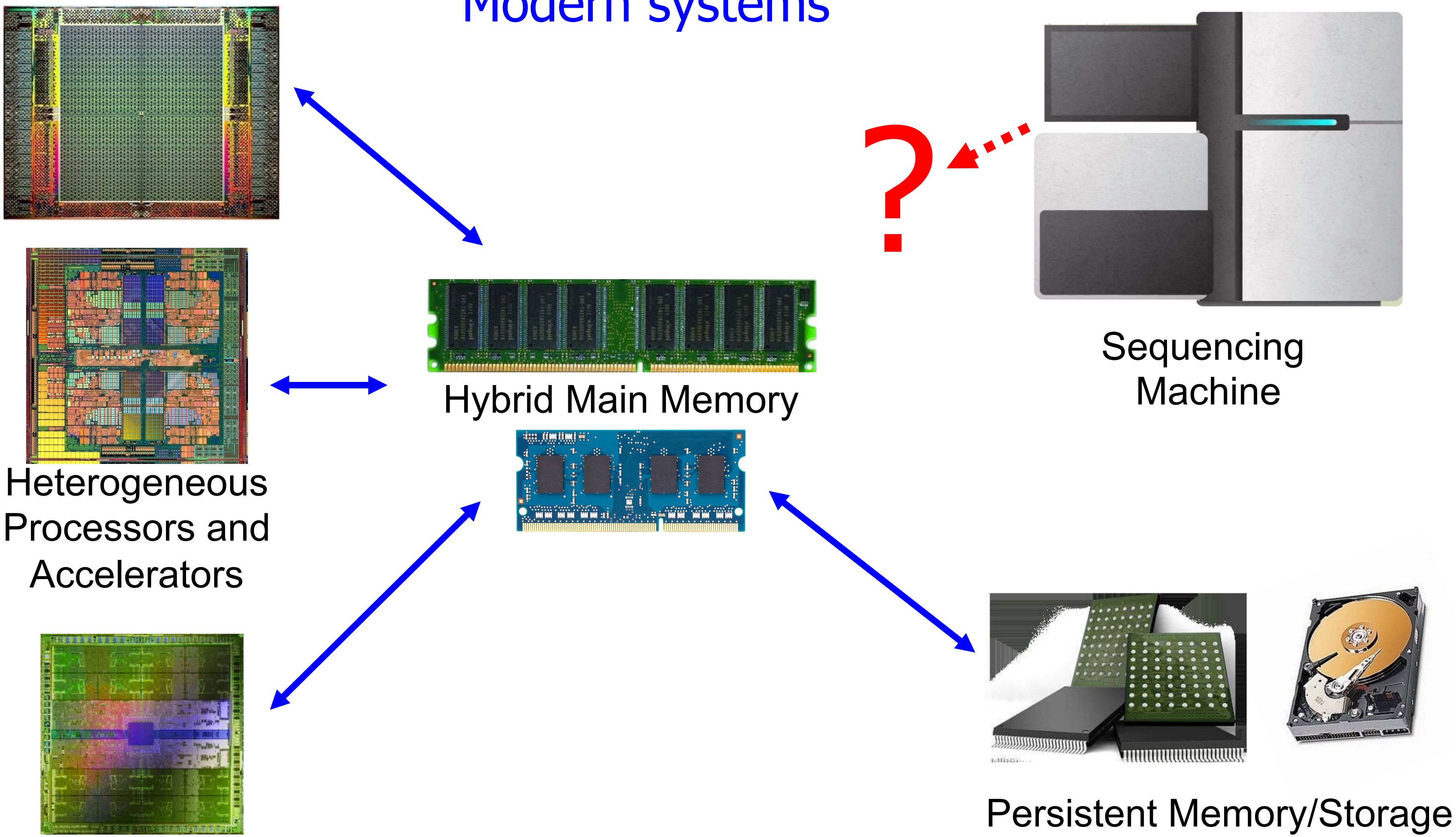
How and where to enable

fast, accurate, cheap,

privacy-preserving, and exabyte scale

analysis of genomic data?

Pushing Towards New Architectures



FPGAs

Modern systems

?

Sequencing Machine

Hybrid Main Memory

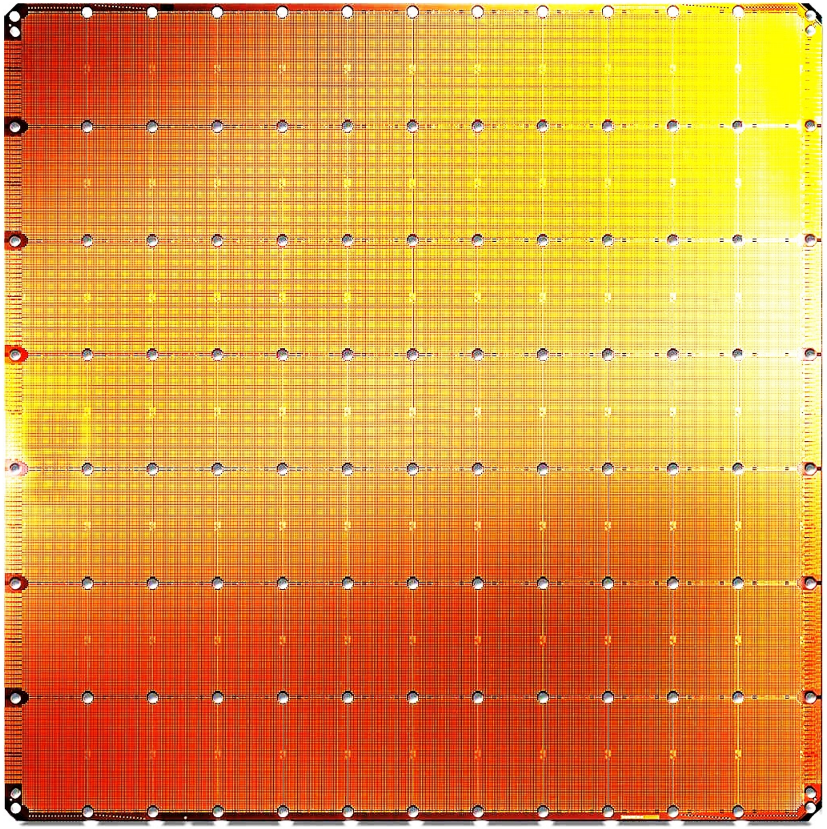
Heterogeneous Processors and Accelerators

Persistent Memory/Storage

(General Purpose) GPUs

Cerebras's Wafer Scale Engine (2019)

- The largest ML accelerator chip
- 400,000 cores



NVIDIA TITAN V

Cerebras WSE
1.2 Trillion transistors
46,225 mm^2

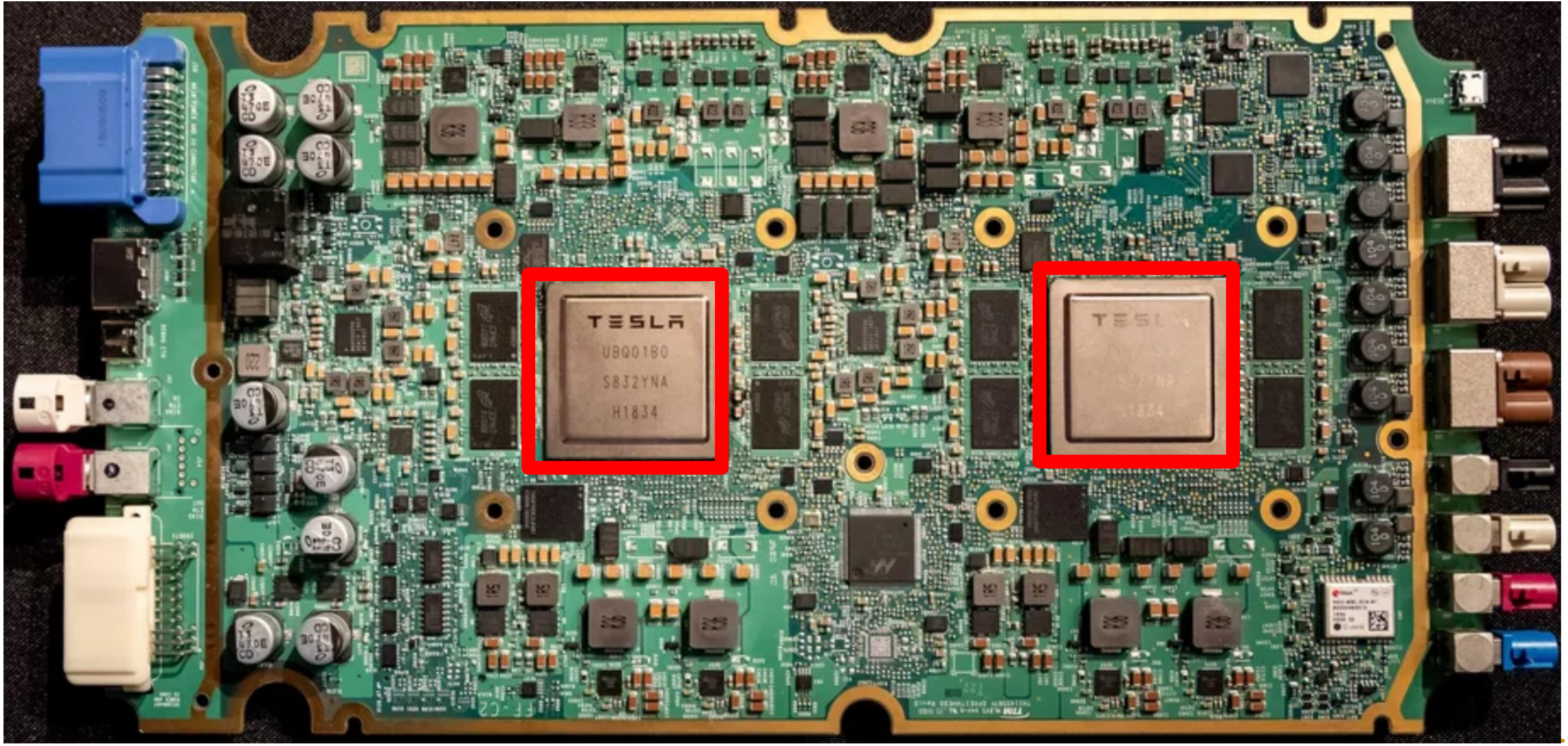
Largest GPU
21.1 Billion transistors
815 mm^2

https://www.cerebras.net/cerebras-wafer-scale-engine-why-we-need-big-chips-for-deep-learning/

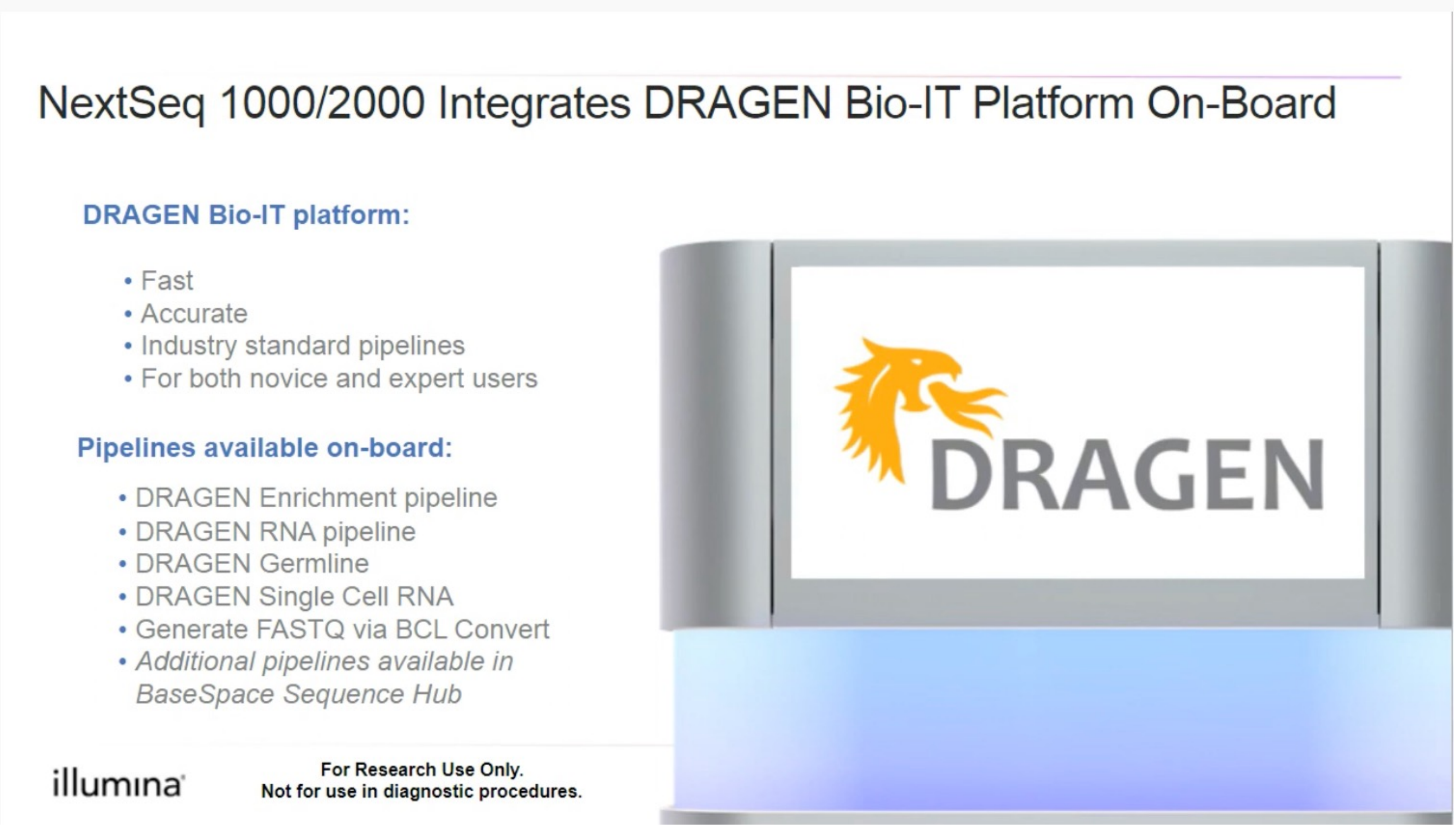
TESLA Full Self-Driving Computer (2019)

- ML accelerator: 260 mm^2, 6 billion transistors, 600 GFLOPS GPU, 12 ARM 2.2 GHz CPUs.
- Two redundant chips for better safety.

https://youtu.be/Ucp0TTmvqOE?t=4236



NextSeq 2000 with Analysis Capability



NextSeq 1000/2000 Integrates DRAGEN Bio-IT Platform On-Board

DRAGEN Bio-IT platform:

- Fast
- Accurate
- Industry standard pipelines
- For both novice and expert users

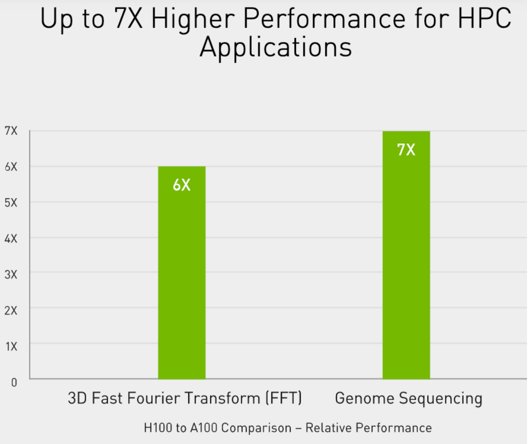
Pipelines available on-board:

- DRAGEN Enrichment pipeline
- DRAGEN RNA pipeline
- DRAGEN Germline
- DRAGEN Single Cell RNA
- Generate FASTQ via BCL Convert
- *Additional pipelines available in BaseSpace Sequence Hub*

illumina

For Research Use Only.
Not for use in diagnostic procedures.

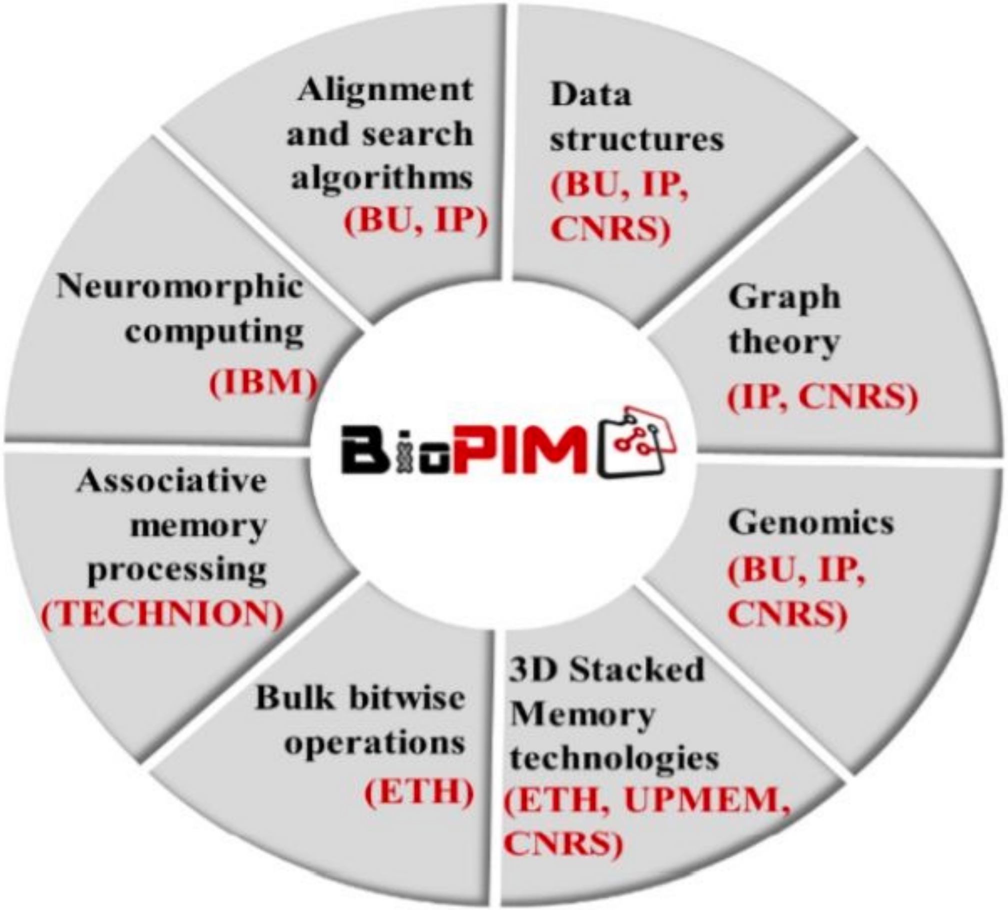
NVIDIA H100 (2022)


Up to 7X Higher Performance for HPC Applications
7X
6X
5X
4X
3X
2X
1X
0
6X
7X
3D Fast Fourier Transform (FFT)
Genome Sequencing
H100 to A100 Comparison – Relative Performance

NVIDIA is claiming a **7x improvement** in dynamic programming algorithm (**DPX instructions**) performance on a single H100 versus naïve execution on an A100.

https://www.nvidia.com/en-us/data-center/h100/

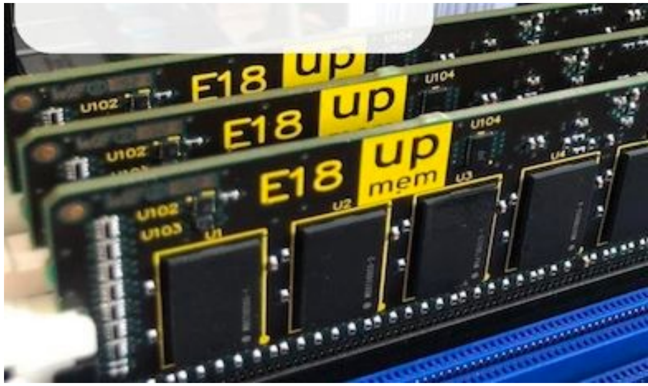
BioPIM (2022)


Alignment and search algorithms (BU, IP)
Data structures (BU, IP, CNRS)
Graph theory (IP, CNRS)
Genomics (BU, IP, CNRS)
3D Stacked Memory technologies (ETH, UPMEM, CNRS)
Bulk bitwise operations (ETH)
Associative memory processing (TECHNION)
Neuromorphic computing (IBM)
BioPIM

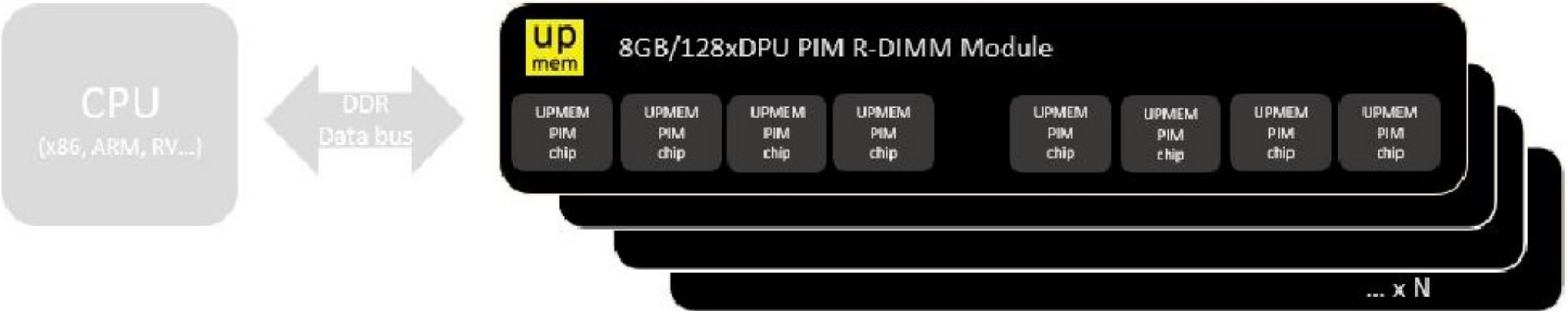
The vision of BioPIM is the realization of **cheap, ultra-fast and ultra-low energy mobile genomics** that eliminates the current dependence of sequence analysis on large and power-hungry computing clusters/data-centers.

UPMEM Processing-in-DRAM Engine (2019)

- Processing in DRAM Engine
- Includes **standard DIMM modules**, with a **large number of DPU processors** combined with DRAM chips.



- Replaces **standard** DIMMs
  - DDR4 R-DIMM modules
    - 8GB+128 DPUs (16 PIM chips)
    - Standard 2x-nm DRAM process
  - **Large amounts of** compute & memory bandwidth



https://www.anandtech.com/show/14750/hot-chips-31-analysis-inmemory-processing-by-upmem

https://www.upmem.com/video-upmem-presenting-its-true-processing-in-memory-solution-hot-chips-2019/

Where is Read Mapping Going Next?

Will 100% accurate genome-long reads alleviate/eliminate the need for read mapping?

Think about metagenomics, pan-genomics, ...

Lecture Conclusion

- System design for bioinformatics is a critical problem
  - It has large scientific, medical, societal, personal implications

- This lecture is about accelerating a key step in bioinformatics: genome sequence analysis
  - In particular, read mapping

- Many bottlenecks exist in accessing and manipulating huge amounts of genomic data during analysis

- We cover various recent ideas to accelerate read mapping
  - A journey since September 2006

Key Takeaways

- Population-scale analyses are not an easy task

- You need to consider **many** things in designing a new system + have good intuition/insight into ideas/tradeoffs

- But, it is fun and can be **very rewarding/impactful**

- And, enables a great future
  - It has large scientific, medical, societal, personal implications

- **Very hot topic for graduate studies and research!**

Key Conclusion

Most speedup comes from

parallelism enabled by

novel architectures and algorithms

Acknowledgments

Onur Mutlu, ETH Zurich

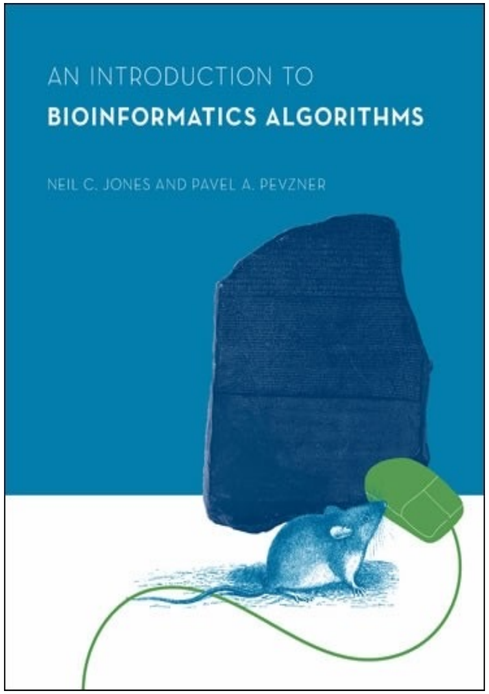
Can Alkan, Bilkent University

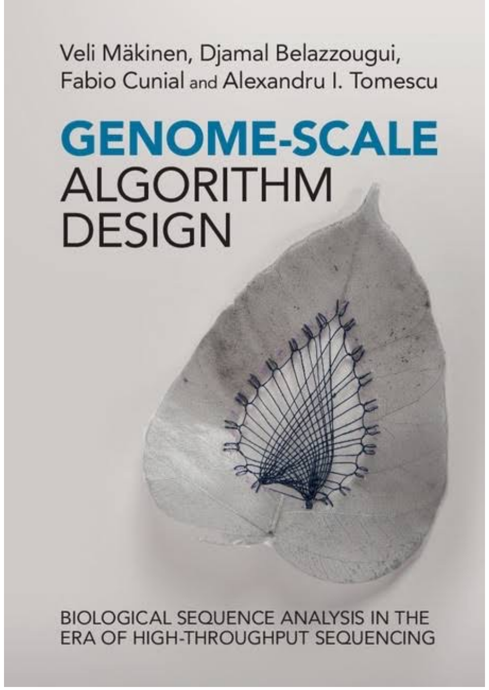
Serghei Mangul, USC

- Many colleagues and collaborators
  - Damla Senol Cali, Jeremie Kim, Hasan Hassan, Can Firtina, Juan Gómez Luna, Hongyi Xin, …
- Funders:
  - NIH and Industrial Partners (Alibaba, AMD, Google, Facebook, HP Labs, Huawei, IBM, Intel, Microsoft, Nvidia, Oracle, Qualcomm, Rambus, Samsung, Seagate, VMware)
- All papers, source code, and more are at:
  - https://people.inf.ethz.ch/omutlu/projects.htm

Recommended Readings

- Jones, Neil C. and Pavel Pevzner. "An introduction to bioinformatics algorithms," MIT press, 2004.
- Mäkinen, Veli, Djamal Belazzougui, Fabio Cunial, and Alexandru I. Tomescu. "Genome-scale algorithm design," Cambridge University Press, 2015.

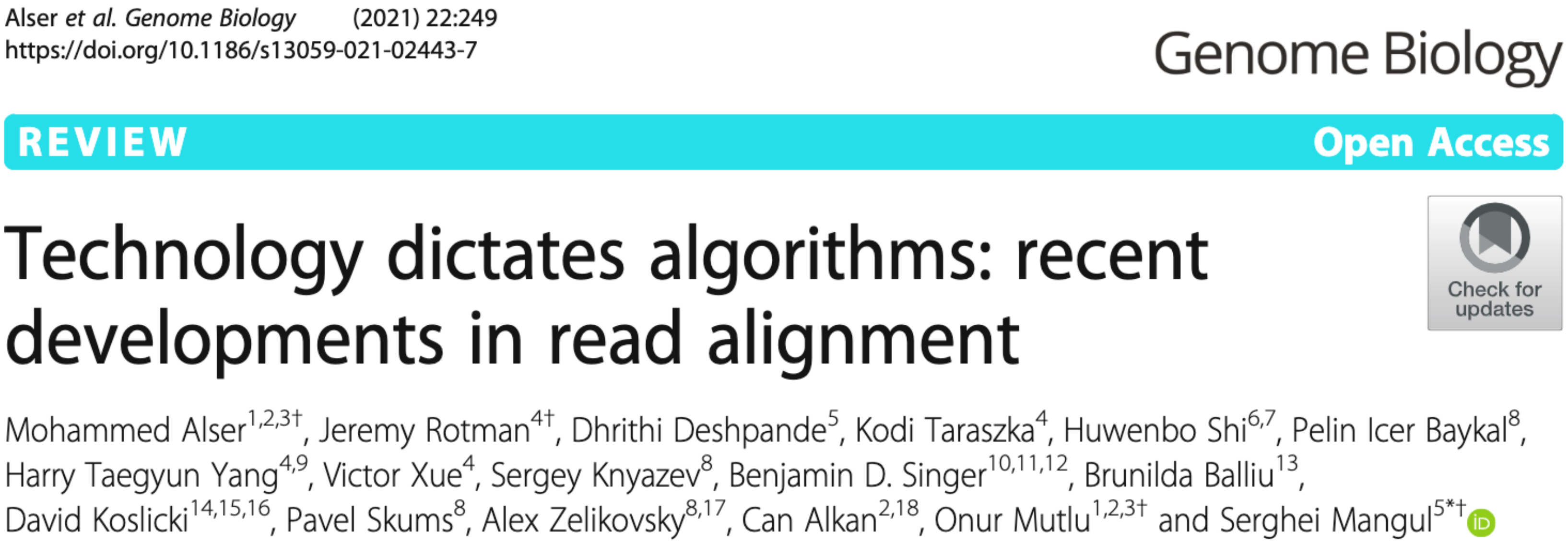



Veli Mäkinen, Djamal Belazzougui,
Fabio Cunial and Alexandru I. Tomescu
GENOME-SCALE
ALGORITHM
DESIGN
BIOLOGICAL SEQUENCE ANALYSIS IN THE
ERA OF HIGH-THROUGHPUT SEQUENCING

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

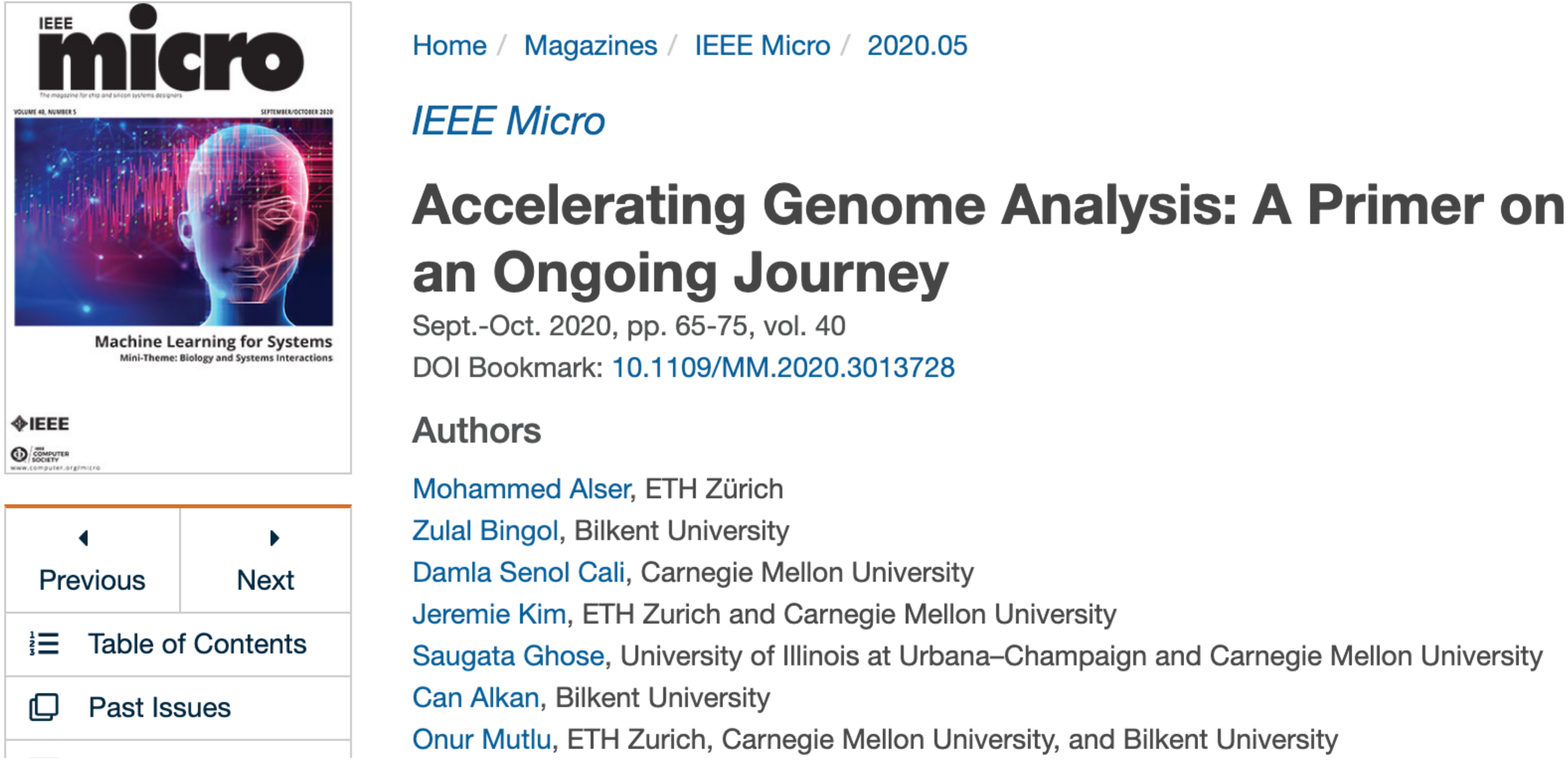
REVIEW Open Access

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*†]

Detailed Analysis of Tackling the Bottleneck

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

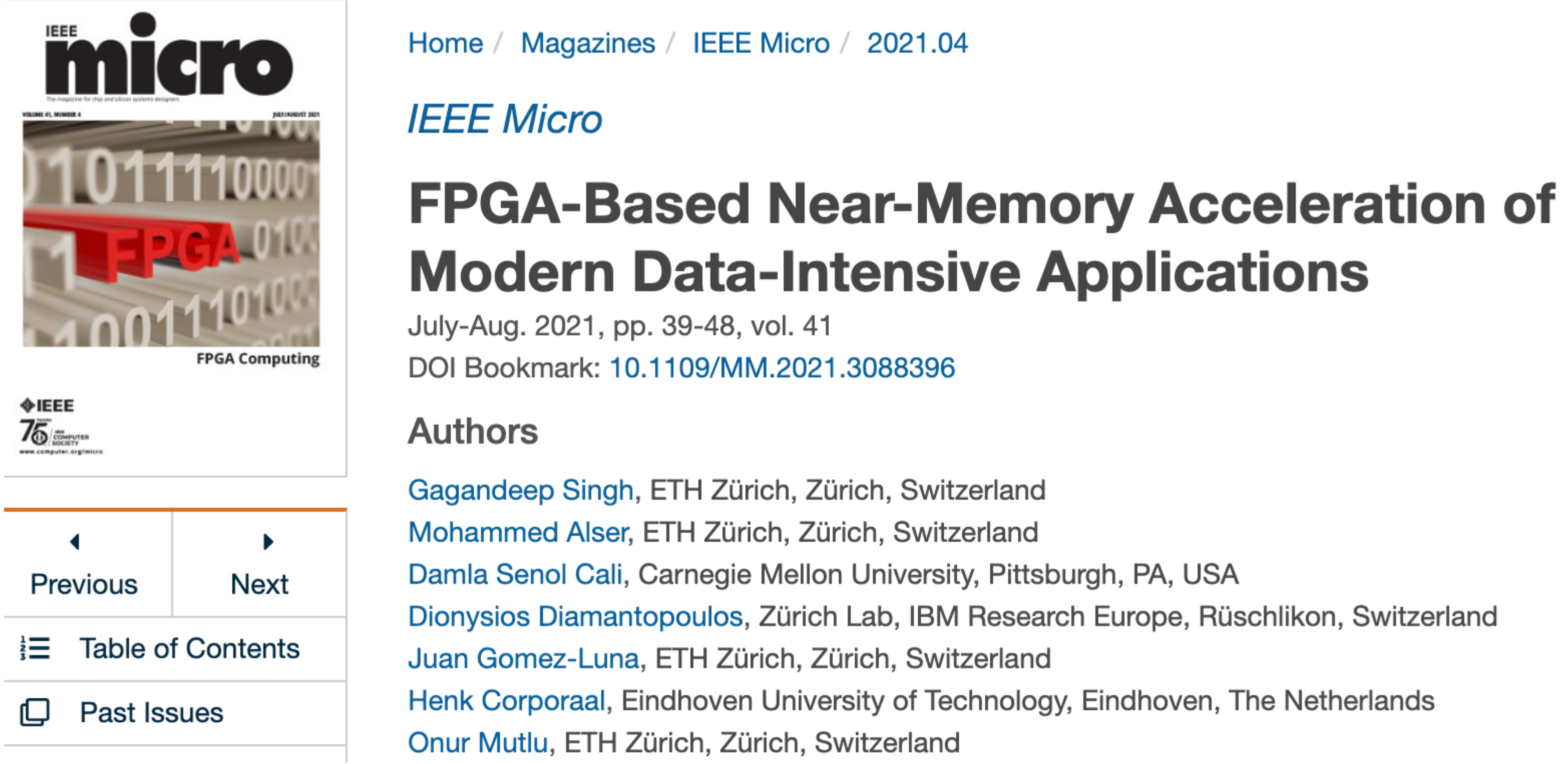
DOI Bookmark: 10.1109/MM.2020.3013728

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

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

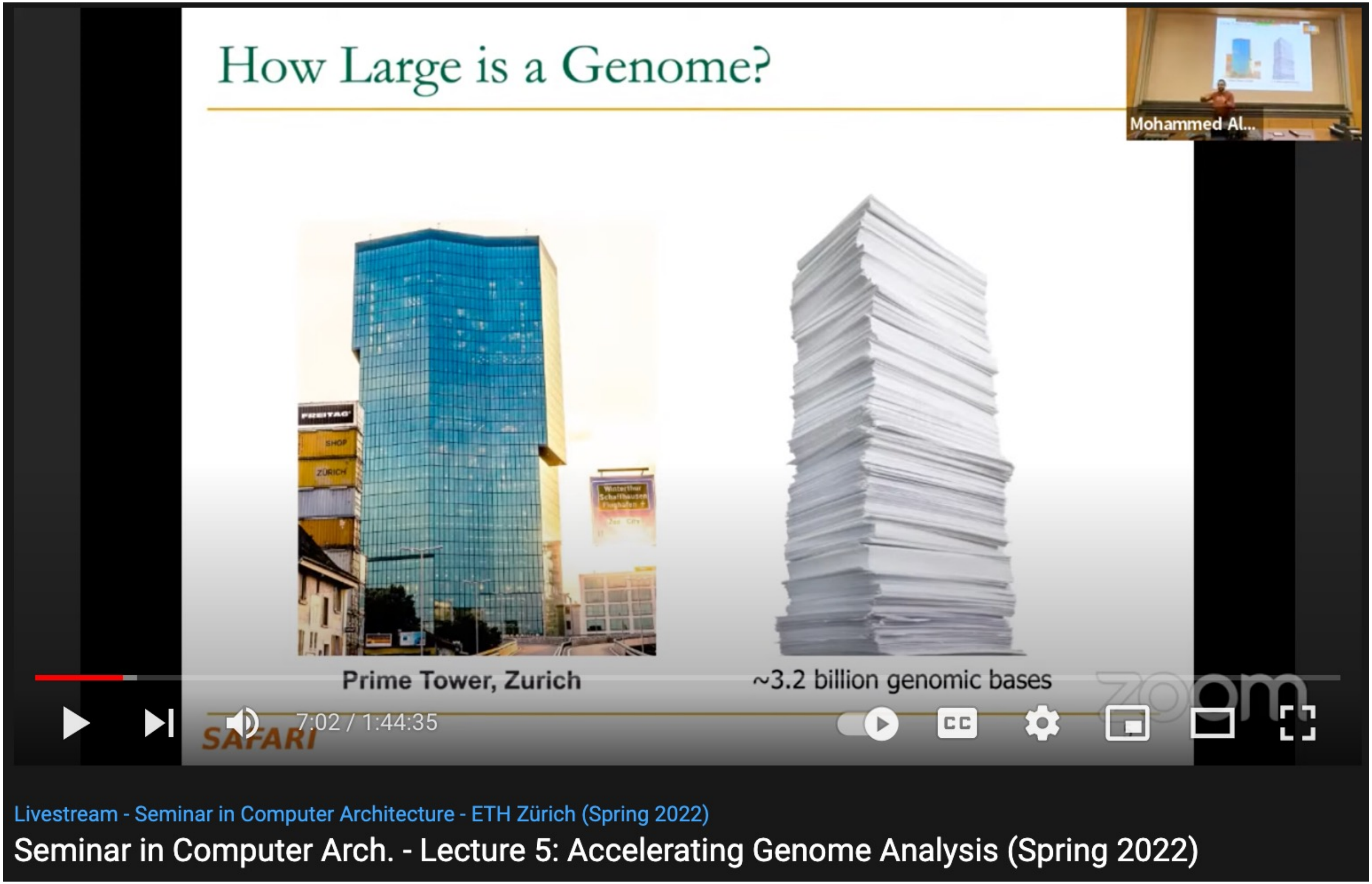
July-Aug. 2021, pp. 39-48, vol. 41

DOI Bookmark: 10.1109/MM.2021.3088396

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

Accelerating Genome Analysis



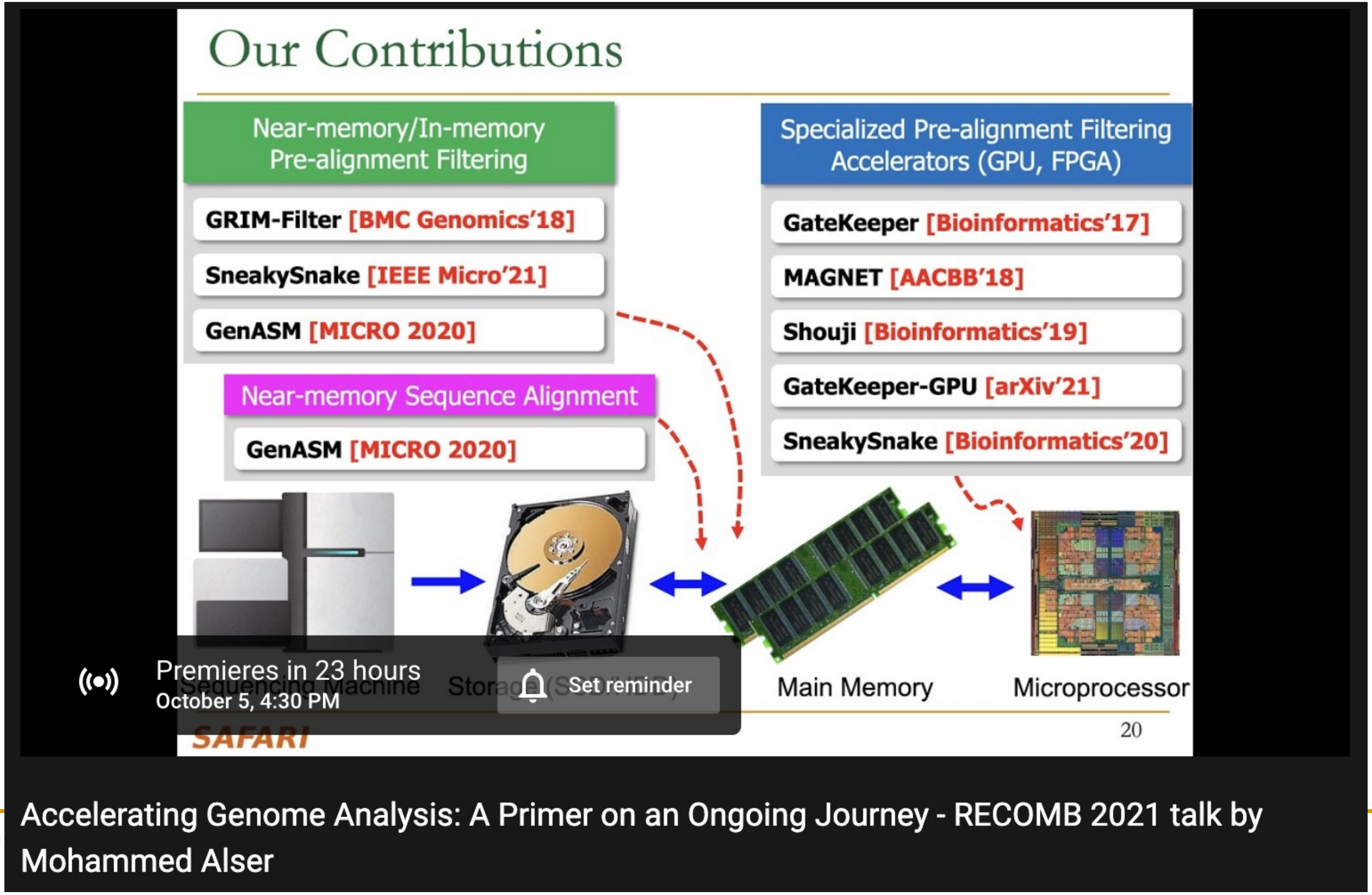
Livestream - Seminar in Computer Architecture - ETH Zürich (Spring 2022)

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

https://www.youtube.com/watch?v=qPIiiwUVFug

More on Accelerating Genome Analysis ...

- Mohammed Alser,
  "Accelerating Genome Analysis: A Primer on an Ongoing Journey"
  Talk at RECOMB 2021, Virtual, August 30, 2021.
  [Slides (pptx) (pdf)]
  [Talk Video (27 minutes)]
  [Related Invited Paper (at IEEE Micro, 2020)]



Our Contributions

Near-memory/In-memory Pre-alignment Filtering

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

Near-memory Sequence Alignment

- **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]

Accelerating Genome Analysis: A Primer on an Ongoing Journey - RECOMB 2021 talk by Mohammed Alser

More on Intelligent Genome Analysis …

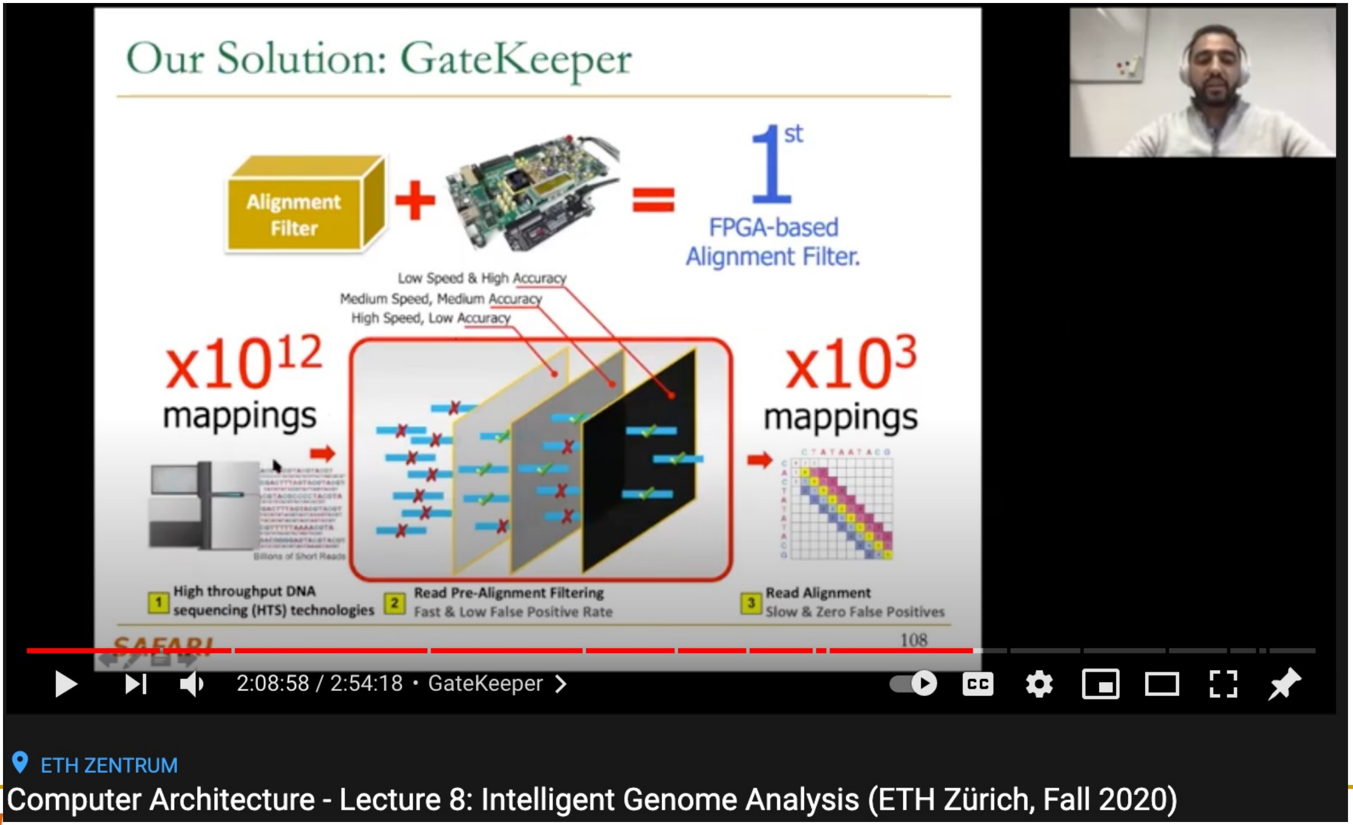
- Mohammed Alser,
  "Computer Architecture - Lecture 10: Intelligent Genome Analysis"
  ETH Zurich, Computer Architecture Course, Fall2021, Lecture 10, Virtual, 29 October 2021.
  [Slides (pptx) (pdf)]
  [Talk Video (3 hour 2 minutes, including Q&A)]
  [Related Invited Paper (at IEEE Micro, 2020)]

Computer Architecture - Lecture 10: Intelligent Genome Analysis (Fall 2021)

412 views • Streamed live on Oct 29, 2021

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

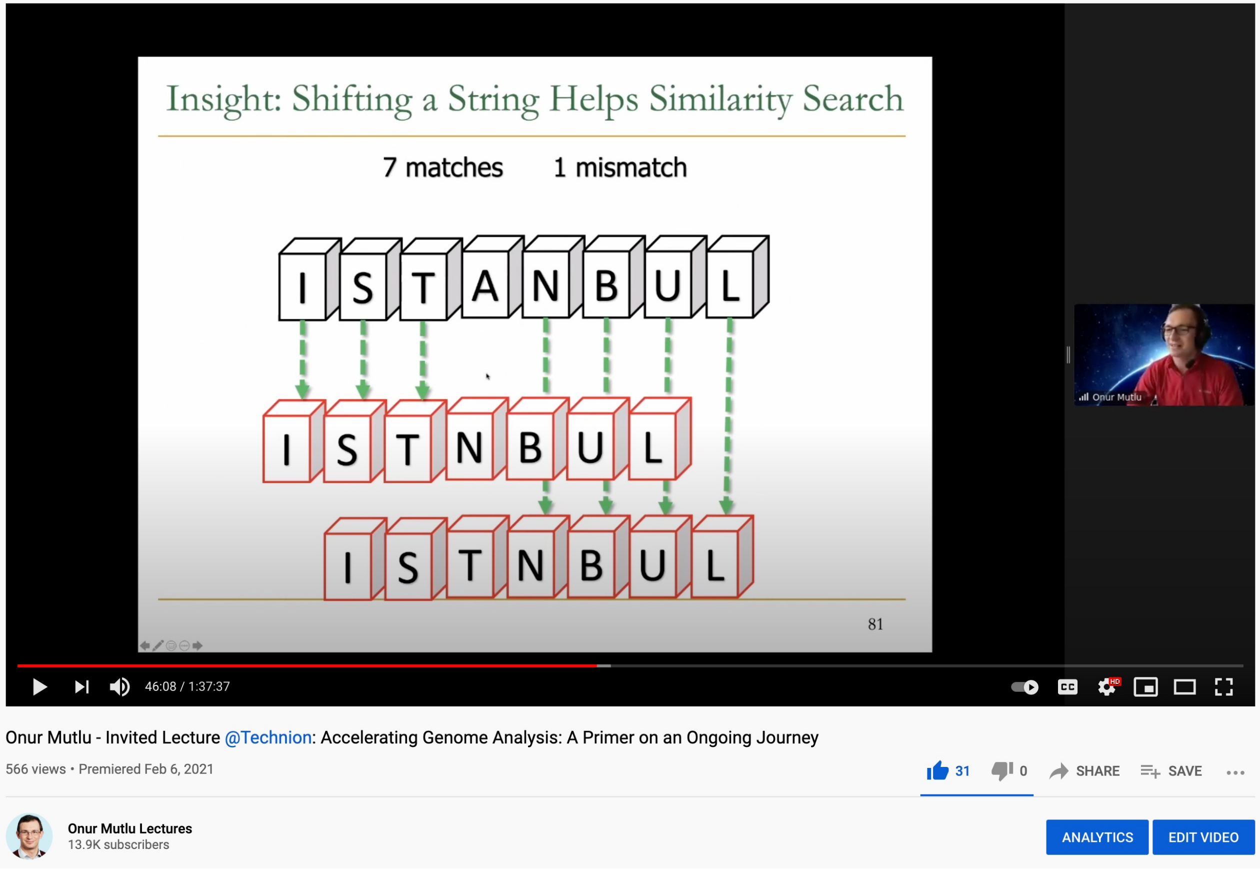
Alignment Filter
+
=
1st
FPGA-based Alignment Filter.
Low Speed & High Accuracy
Medium Speed, Medium Accuracy
High Speed, Low Accuracy
x10^12
mappings
x10^3
mappings
Billions of Short Reads
1 High throughput DNA sequencing (HTS) technologies
2 Read Pre-Alignment Filtering Fast & Low False Positive Rate
3 Read Alignment Slow & Zero False Positives
SAFARI
108
2:08:58 / 2:54:18 • GateKeeper

ETH ZENTRUM

Computer Architecture - Lecture 8: Intelligent Genome Analysis (ETH Zürich, Fall 2020)

More on Fast Genome Analysis …

- Onur Mutlu,
  "Accelerating Genome Analysis: A Primer on an Ongoing Journey"
  Invited Lecture at Technion, Virtual, 26 January 2021.
  [Slides (pptx) (pdf)]
  [Talk Video (1 hour 37 minutes, including Q&A)]
  [Related Invited Paper (at IEEE Micro, 2020)]



Insight: Shifting a String Helps Similarity Search

7 matches 1 mismatch

Onur Mutlu - Invited Lecture @Technion: Accelerating Genome Analysis: A Primer on an Ongoing Journey

566 views • Premiered Feb 6, 2021

Onur Mutlu Lectures
13.9K subscribers

Detailed Lectures on Genome Analysis

- Computer Architecture, Fall 2020, Lecture 3a
  - **Introduction to Genome Sequence Analysis** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=CrRb32v7SJc&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=5
- Computer Architecture, Fall 2020, Lecture 8
  - **Intelligent Genome Analysis** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=ygmQpdDTL7o&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=14
- Computer Architecture, Fall 2020, Lecture 9a
  - **GenASM: Approx. String Matching Accelerator** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=XoLpzmN-Pas&list=PL5Q2soXY2Zi9xidyIgBxUz7xRPS-wisBN&index=15
- Accelerating Genomics Project Course, Fall 2020, Lecture 1
  - **Accelerating Genomics** (ETH Zürich, Fall 2020)
  - https://www.youtube.com/watch?v=rgjl8ZyLsAg&list=PL5Q2soXY2Zi9E2bBVAgCqLgwiDRQDTyId

Prior Research on Genome Analysis (1/2)

- Alser+, "Technology dictates algorithms: Recent developments in read alignment", *Genome Biology*, 2021.

- Alser + "SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs.", *Bioinformatics,* 2020.

- Senol Cali+, "GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis", *MICRO* 2020.

- Kim+, "AirLift: A Fast and Comprehensive Technique for Translating Alignments between Reference Genomes", *arXiv*, 2020

- Alser+, "Accelerating Genome Analysis: A Primer on an Ongoing Journey", *IEEE Micro*, 2020.

Prior Research on Genome Analysis (2/2)

- Firtina+, “Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm”, *Bioinformatics*, 2019.

- Alser+, “Shouji: a fast and efficient pre-alignment filter for sequence alignment”, *Bioinformatics* 2019.

- Kim+, "GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies”, *BMC Genomics*, 2018.

- Alser+, "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping”, *Bioinformatics*, 2017.

- Alser+, "MAGNET: understanding and improving the accuracy of genome pre-alignment filtering”, *IPSI Transaction*, 2017.

P&S Genomics

Lecture 2: Intelligent Genomic Analyses

Dr. Mohammed Alser

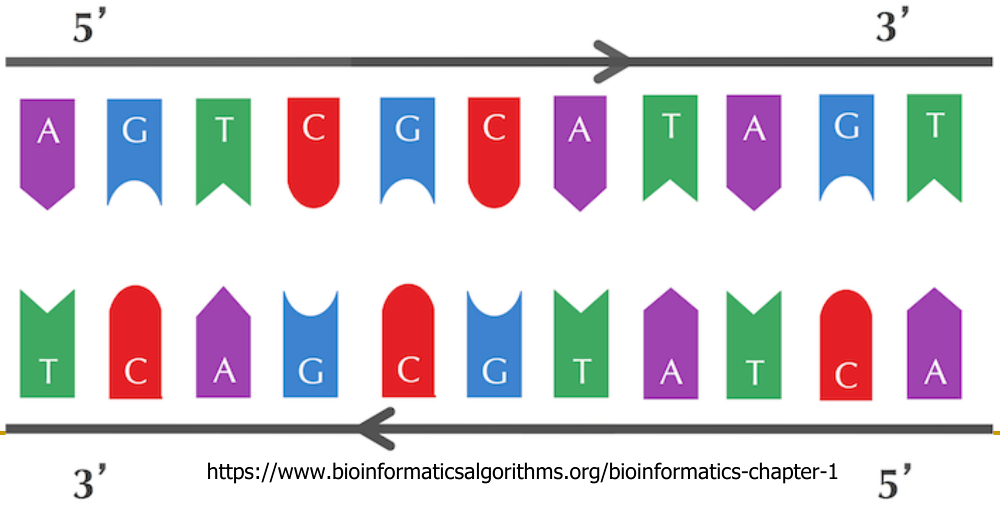
ETH Zürich

Spring 2023

9 March 2023

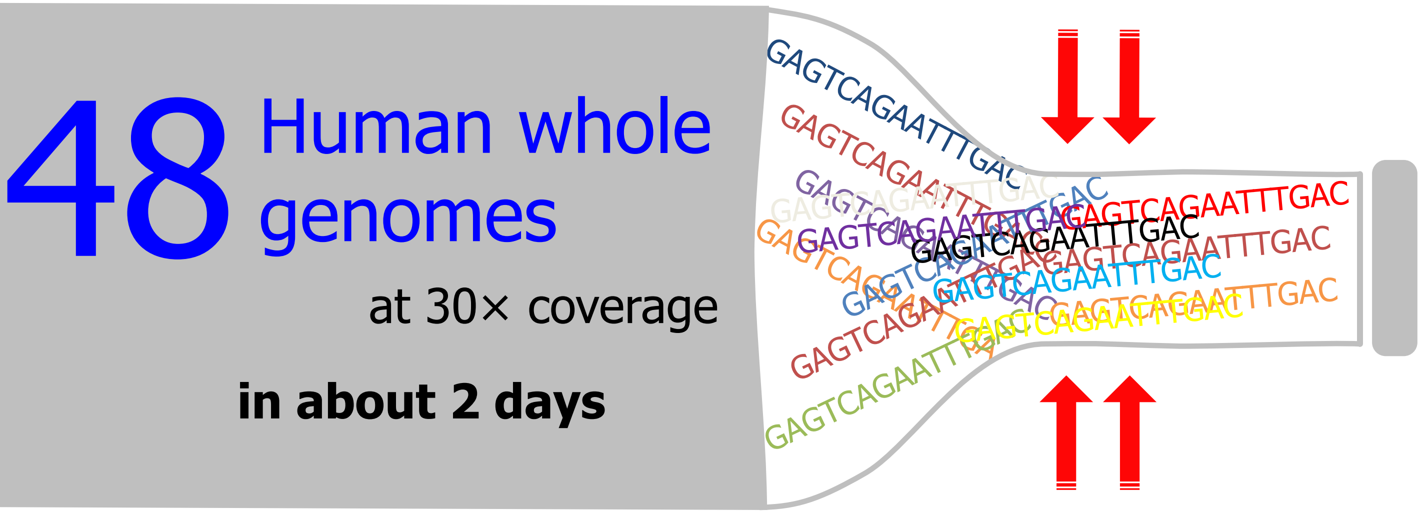
Challenges in Read Mapping

- Need to find many mappings of each read
- Need to tolerate variances/sequencing errors in each read
- Need to map each read very fast (i.e., performance is important, life critical in some cases)
- Need to map reads to both forward and reverse strands



https://www.bioinformaticsalgorithms.org/bioinformatics-chapter-1

Analysis is Bottlenecked in Read Mapping!!



48 Human whole genomes

at 30× coverage

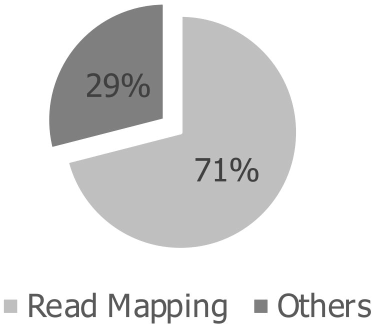
in about 2 days

Illumina NovaSeq 6000

1 Human genome

32 CPU hours

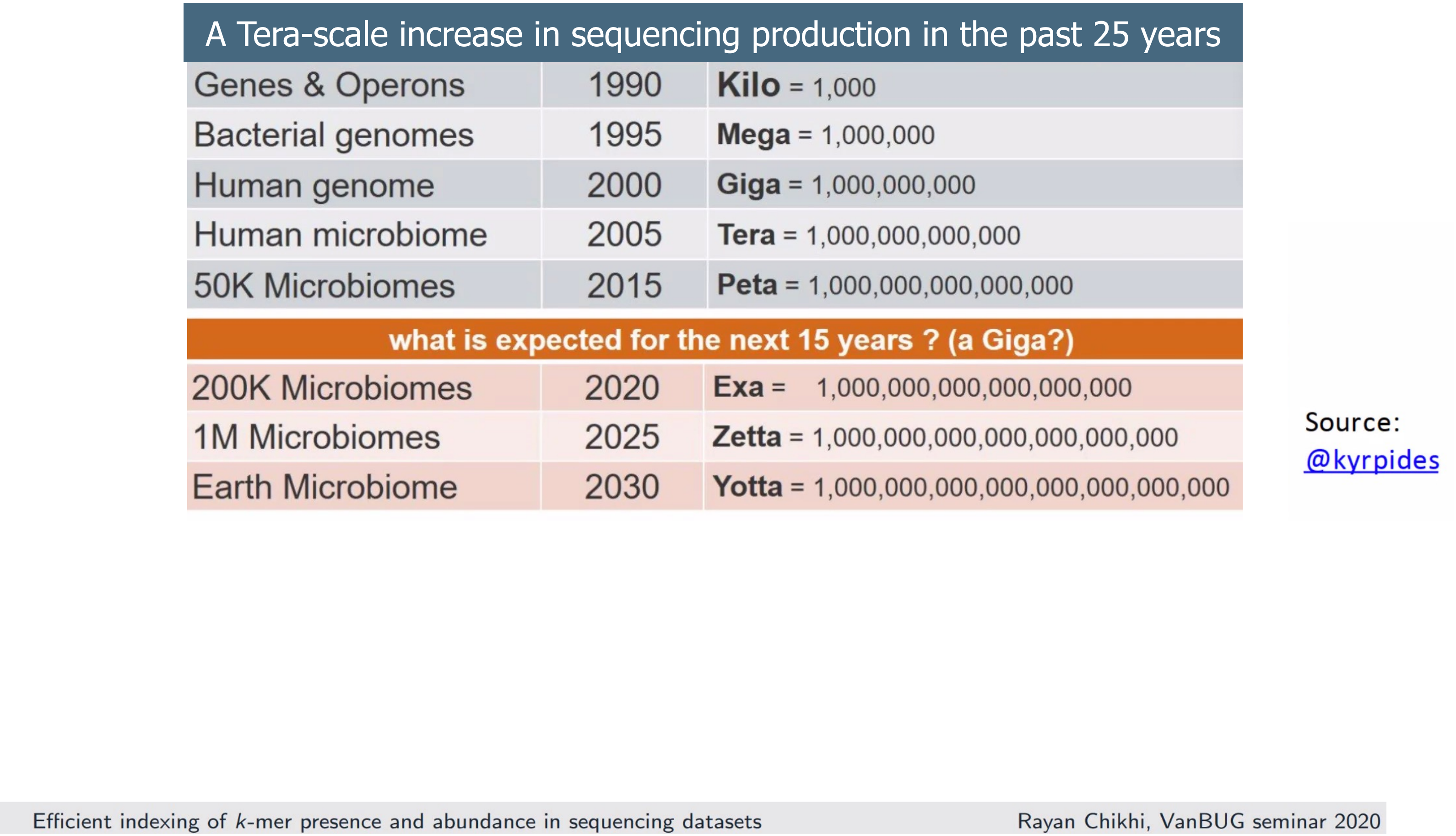
on a 48-core processor



SAFARI Goyal+, "Ultra-fast next generation human genome sequencing data processing using DRAGENTM bio-IT processor for precision medicine", *Open Journal of Genetics,* 2017.

What makes read mapping a **bottleneck**?

A Tsunami of Sequencing Data



A Tera-scale increase in sequencing production in the past 25 years

Genes & Operons	1990	**Kilo** = 1,000
Bacterial genomes	1995	**Mega** = 1,000,000
Human genome	2000	**Giga** = 1,000,000,000
Human microbiome	2005	**Tera** = 1,000,000,000,000
50K Microbiomes	2015	**Peta** = 1,000,000,000,000,000

what is expected for the next 15 years ? (a Giga?)

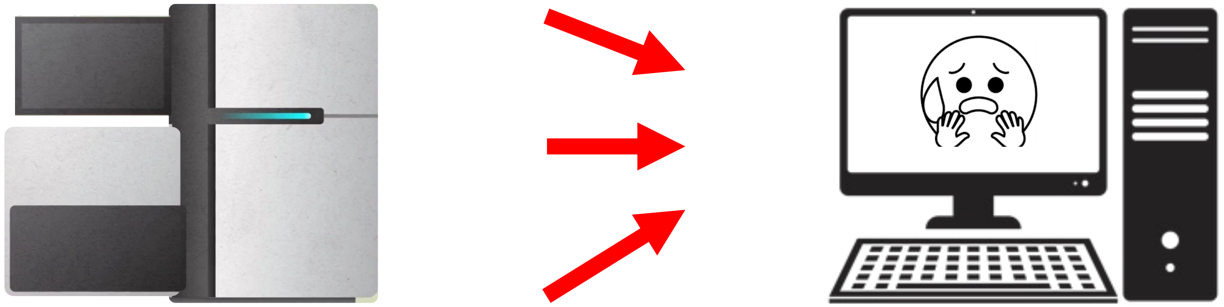
200K Microbiomes	2020	**Exa** = 1,000,000,000,000,000,000
1M Microbiomes	2025	**Zetta** = 1,000,000,000,000,000,000,000
Earth Microbiome	2030	**Yotta** = 1,000,000,000,000,000,000,000,000

Source: @kyrpides

Efficient indexing of *k*-mer presence and abundance in sequencing datasets

Rayan Chikhi, VanBUG seminar 2020

Lack of Specialized Compute Capability



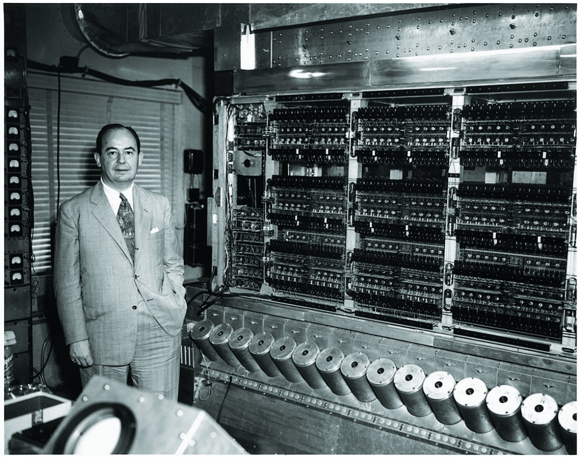
Specialized Machine
for Sequencing

General-Purpose Machine
for Analysis

FAST

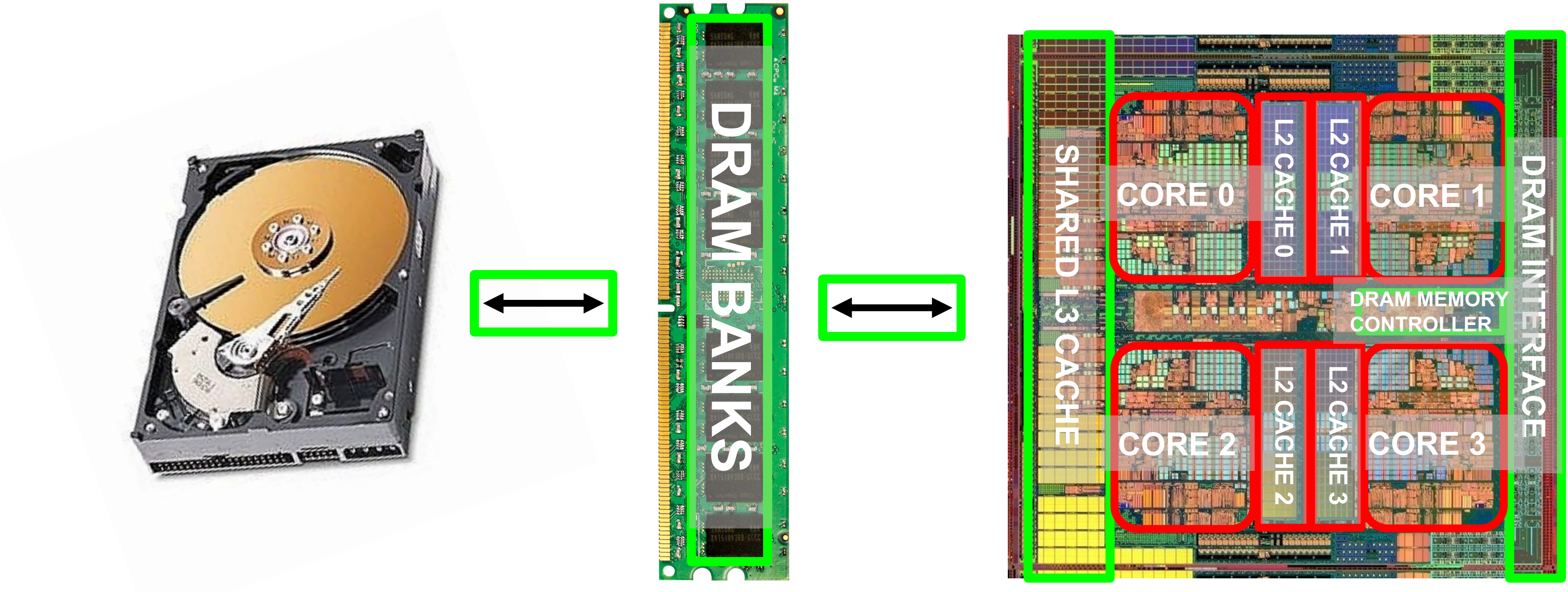
SLOW

Today's Computing Systems



von Neumann model, 1945

where the **CPU** can **access data** stored in an off-chip main memory only through **power-hungry bus**



*Die photo credit: AMD Barcelona

Burks, Goldstein, von Neumann, "Preliminary discussion of the logical design of an electronic computing instrument," 1946.

The Problem

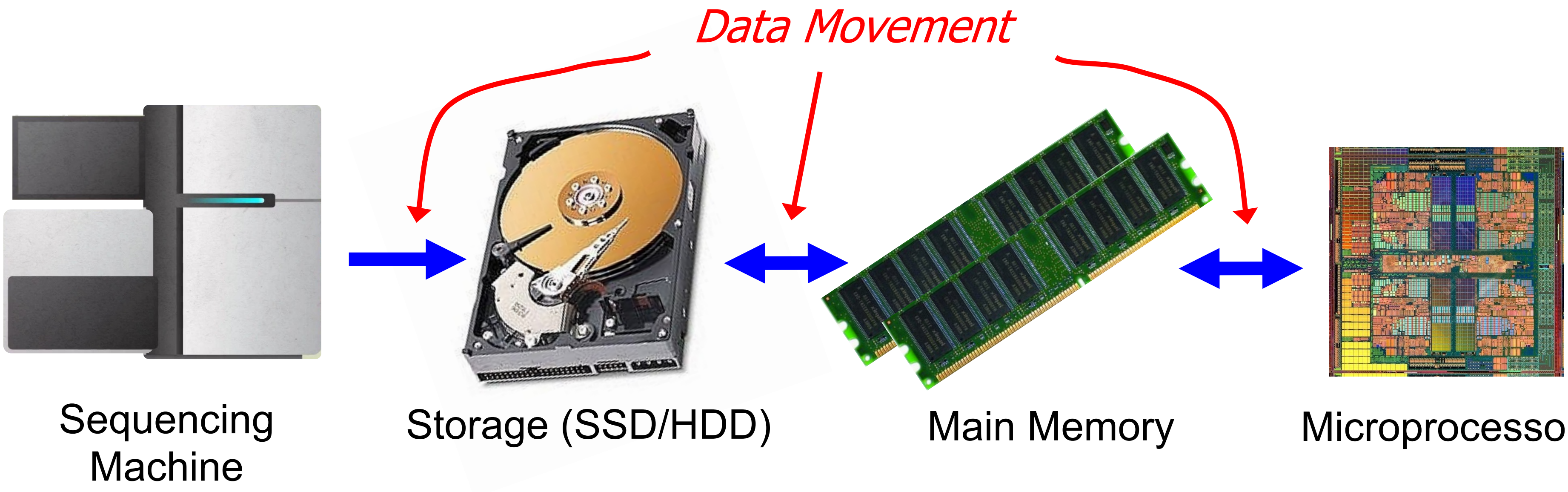
Data analysis

is performed

far away from the data

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

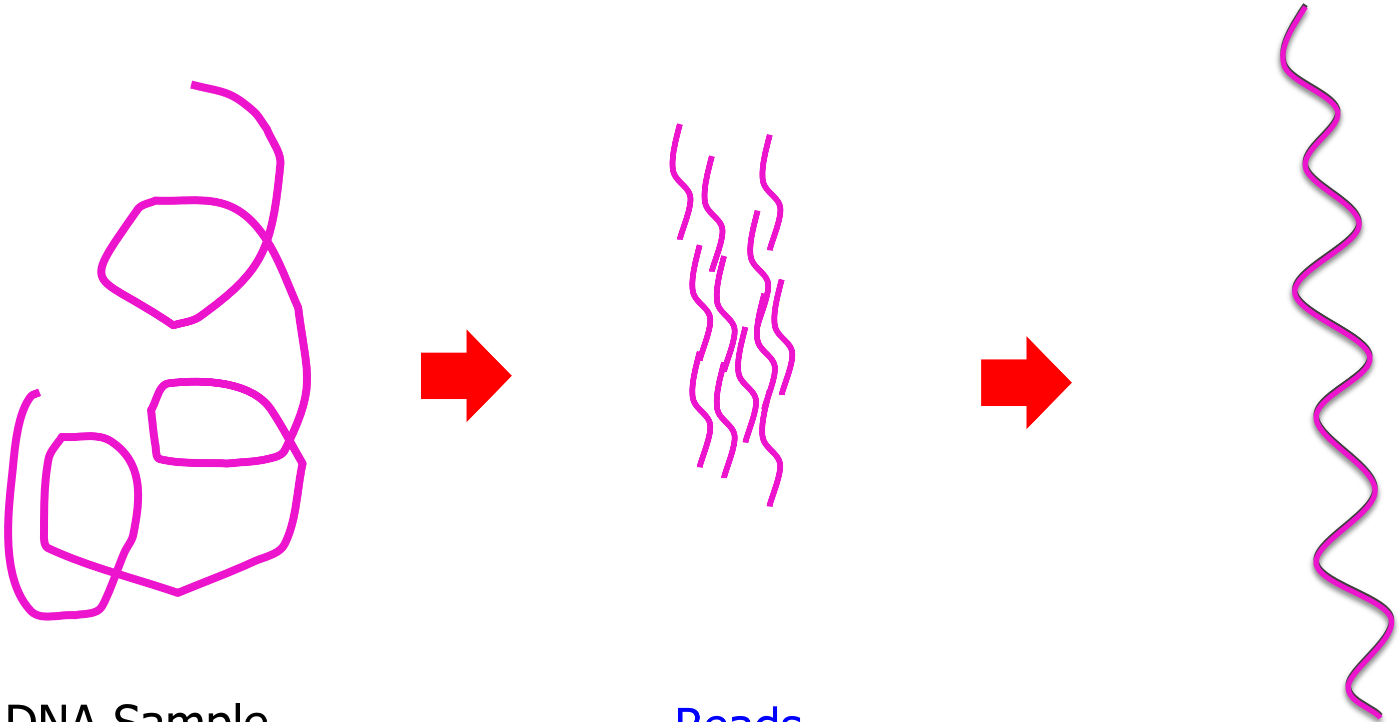
Microprocessor

Single memory request consumes >160x-800x more energy compared to performing an addition operation

✻ 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

Read Mapping

Map reads to a known reference genome with some minor differences allowed



DNA Sample
"chemical format"

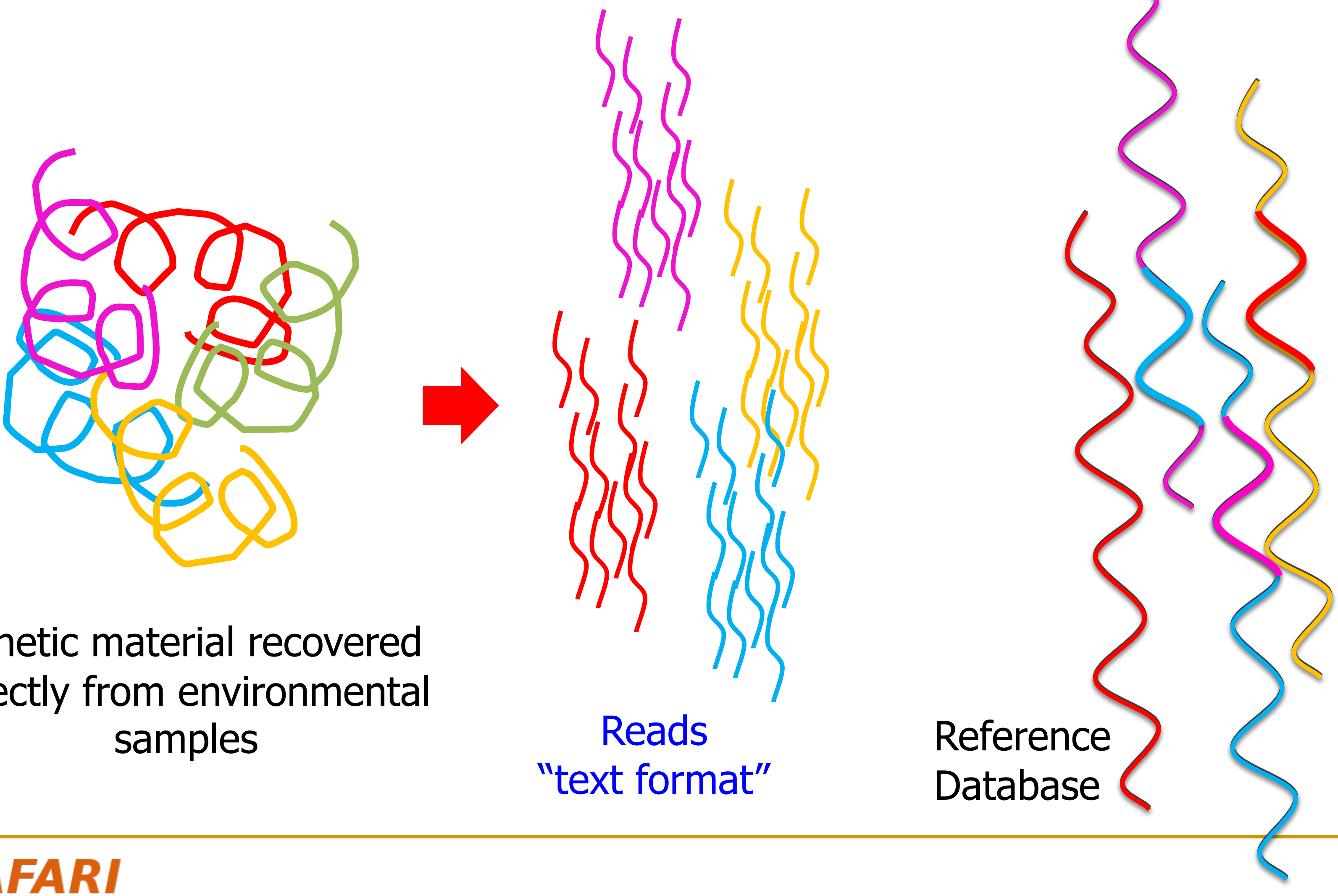
Reads
"text format"

Reference genome

Subject genome
"text format"

Metagenomics Analysis

Reads from different unknown donors at sequencing time are mapped to many known reference genomes

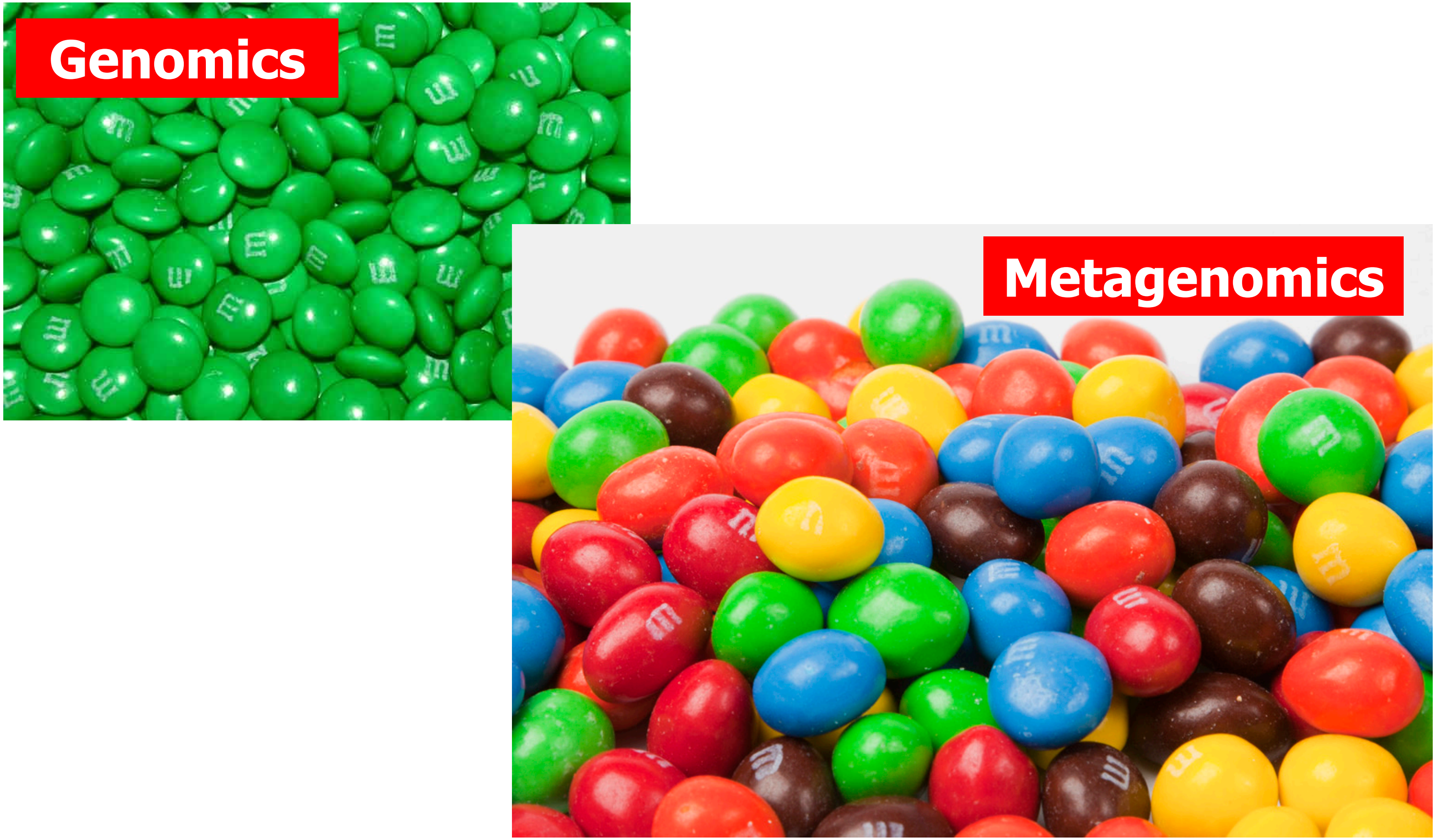


genetic material recovered directly from environmental samples

Reads
“text format”

Reference Database

Genomics vs. Metagenomics

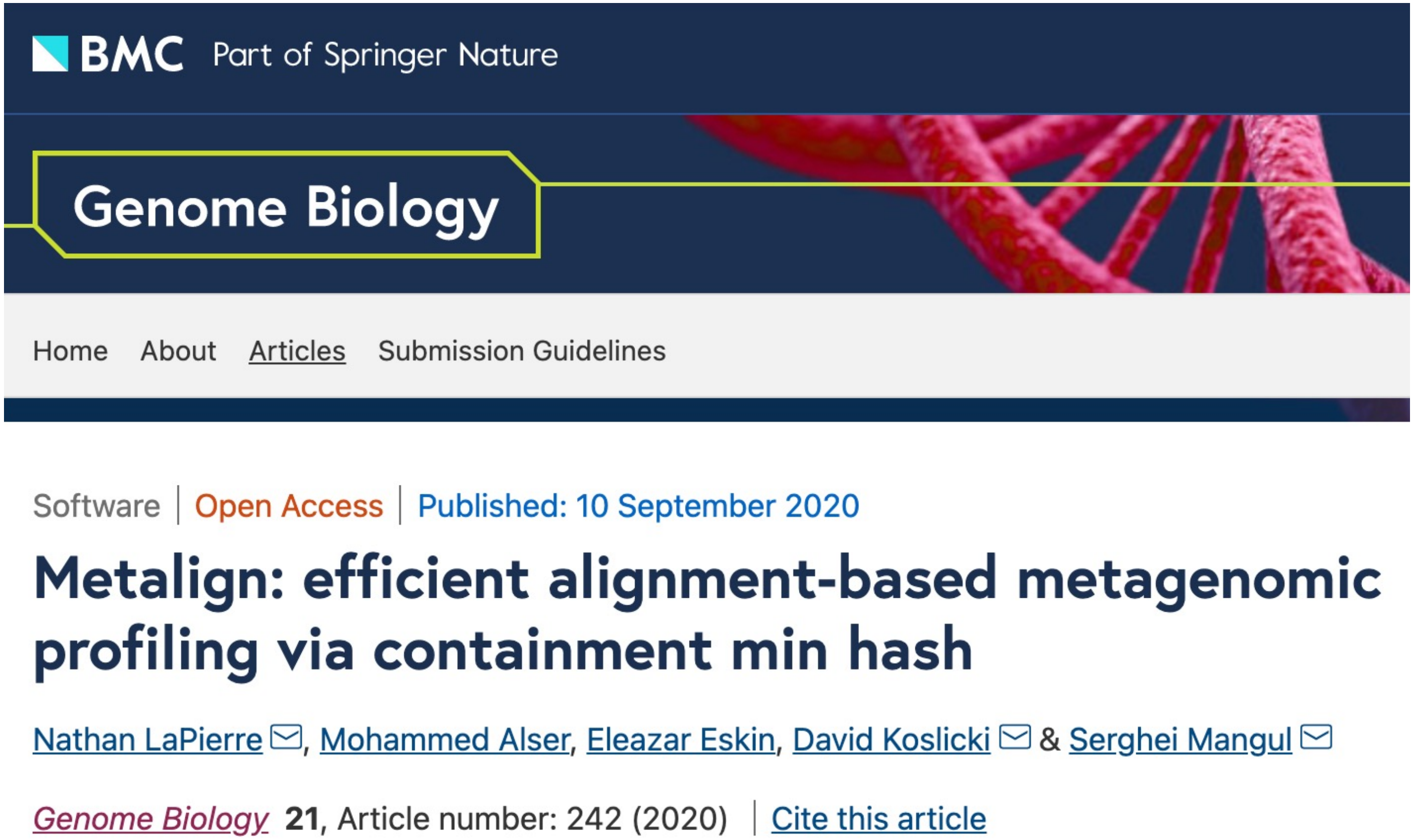


More on Metagenomic Profiling: Metalign

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]



BMC Part of Springer Nature

Genome Biology

Home About Articles Submission Guidelines

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

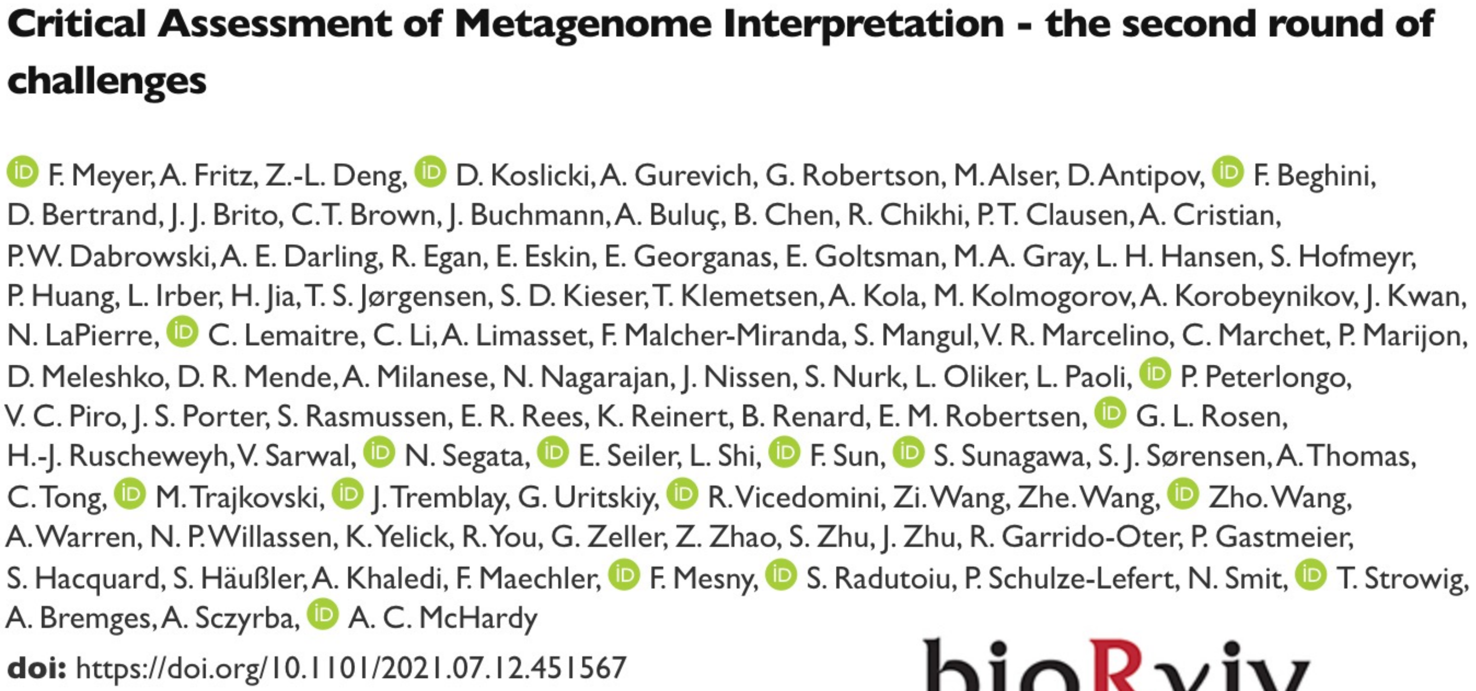
Check Also CAMI II Paper

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”

bioRxiv, 2021

[Source Code]



Critical Assessment of Metagenome Interpretation - the second round of challenges

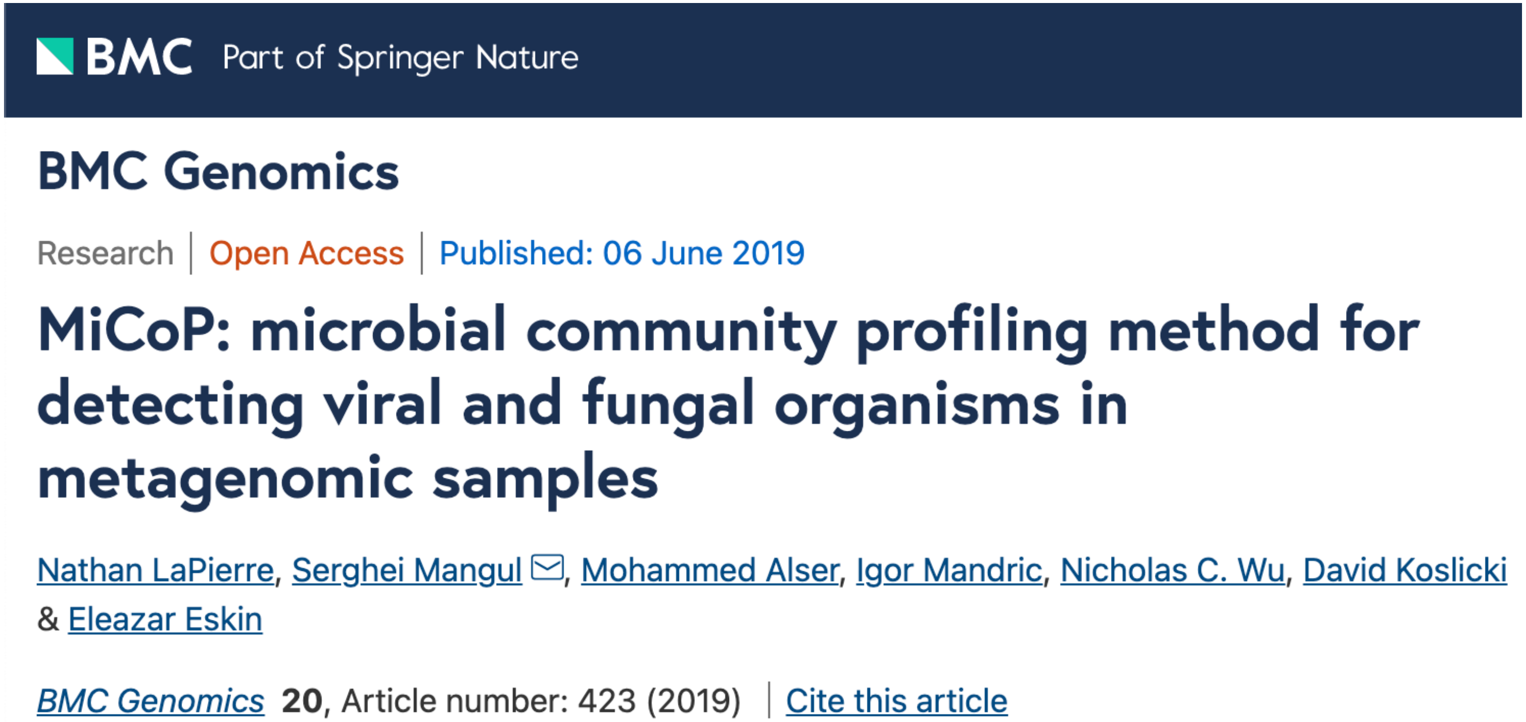
F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini, D. Bertrand, J. J. Brito, C.T. Brown, J. Buchmann, A. Buluç, B. Chen, R. Chikhi, P.T. Clausen, A. Cristian, P.W. Dabrowski, A. E. Darling, R. Egan, E. Eskin, E. Georganas, E. Goltsman, M.A. Gray, L. H. Hansen, S. Hofmeyr, P. Huang, L. Irber, H. Jia, T. S. Jørgensen, S. D. Kieser, T. Klemetsen, A. Kola, M. Kolmogorov, A. Korobeynikov, J. Kwan, N. LaPierre, C. Lemaitre, C. Li, A. Limasset, F. Malcher-Miranda, S. Mangul, V. R. Marcelino, C. Marchet, P. Marijon, D. Meleshko, D. R. Mende, A. Milanese, N. Nagarajan, J. Nissen, S. Nurk, L. Oliker, L. Paoli, P. Peterlongo, V. C. Piro, J. S. Porter, S. Rasmussen, E. R. Rees, K. Reinert, B. Renard, E. M. Robertsen, G. L. Rosen, H.-J. Ruscheweyh, V. Sarwal, N. Segata, E. Seiler, L. Shi, F. Sun, S. Sunagawa, S. J. Sørensen, A. Thomas, C. Tong, M. Trajkovski, J. Tremblay, G. Uritskiy, R. Vicedomini, Zi. Wang, Zhe. Wang, Zho. Wang, A. Warren, N. P. Willassen, K. Yelick, R. You, G. Zeller, Z. Zhao, S. Zhu, J. Zhu, R. Garrido-Oter, P. Gastmeier, S. Hacquard, S. Häußler, A. Khaledi, F. Maechler, F. Mesny, S. Radutoiu, P. Schulze-Lefert, N. Smit, T. Strowig, A. Bremges, A. Sczyrba, A. C. McHardy

doi: https://doi.org/10.1101/2021.07.12.451567

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Check Also 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