

P&S Mobile Genomics

Introduction & Project Proposals

Dr. Mohammed Alser



ETH Zurich

Spring 2022

8 March 2022

The Role of This Course

Projects & Seminars: Mobile Genomics

- We will cover the **basics** of **genome analysis** to understand the **speed-accuracy tradeoff** in using computationally-lightweight heuristics versus accurate computationally-expensive algorithms.
- Students will **experimentally** evaluate different heuristic **algorithms** and observe their effect on **the end results**.
- This evaluation will give the students the chance to carry out a **hands-on project** to implement one or more of these heuristic algorithms in **their smartphones** and **help the society by enabling on-site analysis of genomic data**.

Key Objectives

- Multiple components that are aimed at improving students'
 - Basic knowledge in genome analysis (**dry lab**)
 - Technical skills in genome analysis and computer architecture
 - Critical thinking and analysis
 - Familiarity with key research directions
 - Technical presentation of your project

Key Goal

(Learn how to)

efficiently implement

one of the key steps in genome

analysis on portable devices

Prerequisites of the Course

- No prior knowledge in bioinformatics or genome analysis is required.
- A good knowledge in C programming language and programming is required.
- Interest in making things efficient and solving problems

Course Info: Who Are We?



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**.

Course Info: Who Are We? (I)



Juan Gómez Luna
Senior Researcher and
Lecturer

Processing-In-Memory |
Heterogeneous computing |
Memory Systems | Bioinformatics |
Medical imaging



Nour Almadhoun Alserr
Senior Researcher

Data privacy | Bioinformatics |
Computational Genomics



Can Firtina
PhD Student

Genome Assembly |
Sequence Analysis &
Alignment | Biologically-
Inspired Computing
Paradigms | Brain-
Computer Interfaces |
Phase-change memory

Course Info: Who Are We? (II)



Jeremie Kim

PhD Student

DRAM
power/reliability/performance | Genome Sequence Analysis & Alignment | Hardware/Software Cooperation | Processing-in-Memory | Core Microarchitecture



Joël Lindegger

PhD Student

Acceleration of the bioinformatics pipeline | Current and future computer architectures | All kinds of algorithms and data structures

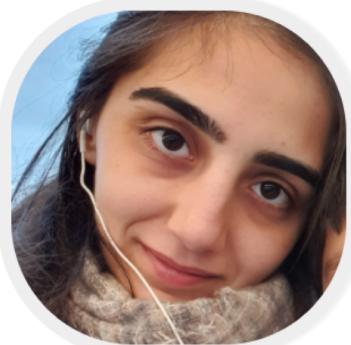
Course Info: Who Are We? (III)



Max Rumpf

Masters Student

Bioinformatics | Computational
Genomics | Sequence Analysis &
Alignment | Machine Learning



Banu Cavlak

Masters Student



Arvid Gollwitzer

Masters Student

Bioinformatics | Computational
Genomics | Sequence Analysis &
Alignment | Medical Applications |
Clinical Metagenomics



Course Requirements and Expectations

- Attendance required for all meetings
- Study the learning materials
- Each student will carry out a hands-on project
 - Build, implement, code, and design with close engagement from the supervisors
- Participation
 - Ask questions, contribute thoughts/ideas
 - Read relevant papers
- Presentation & GitHub repository

We will help the projects with good progress to get published in good venues!

Your Responsibilities

- 2 Lectures every week
 - Tuesday 10-11 AM
 - Friday 9-10 AM
- Attendance is mandatory
- Working on your project for ~6 hours per week
- Meeting your mentors weekly is required

Course Website

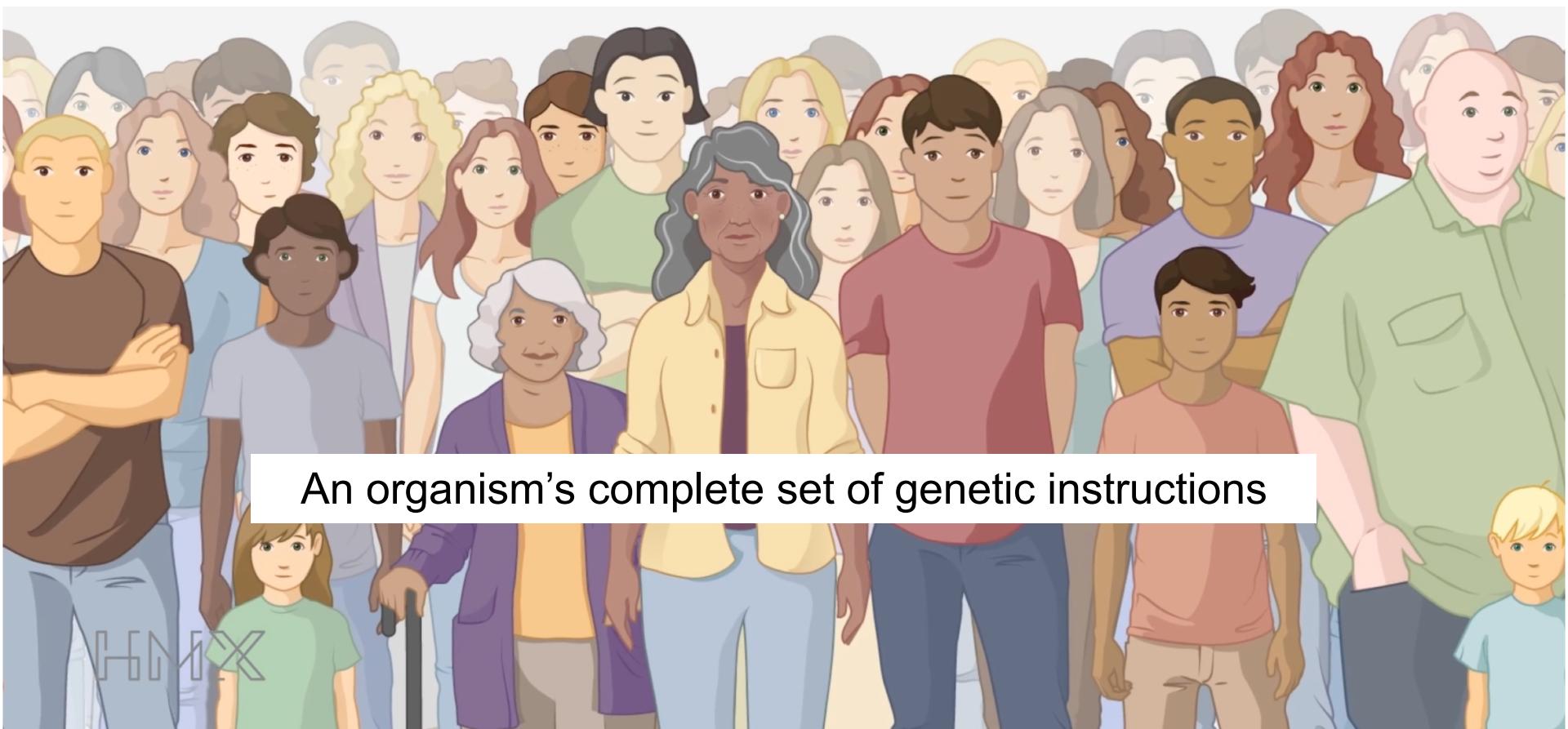
- https://safari.ethz.ch/projects_and_seminars/spring2022/oku.php?id=genome_seq_mobile
- Useful information for the course
- Check your email and Moodle frequently for announcements
- We will also have Moodle for Q&A, announcements, ..

Next Meetings

- We will give you a chance to select a project,
- Then, we will have **1-1 meetings** to match your interests, skills, and background with a suitable project.
- It is important that you **study the learning materials** before our next meeting!
- We will assign the projects **next week**.

What is Genome Analysis?

What is a Genome?



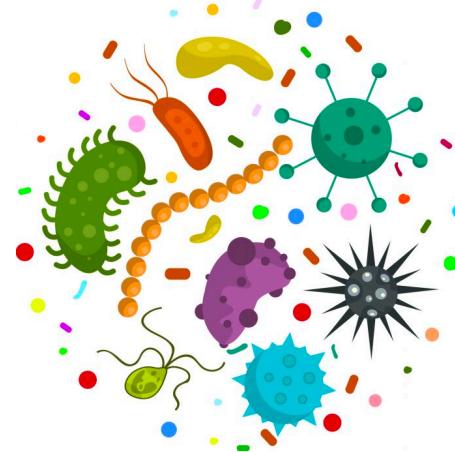
An organism's complete set of genetic instructions

CCTCCTCAGTGCCACCCAGCCCCTGGCAGCTCCAAACA
GGCTCTTATTAAAAACACCCTGTTCCCTGCCCTGGAGTG
AGGTGTCAAGGACCTAAACTAAAAAAAAAAAAGAAAAA
AGAAAAAGAAAAAGAATTAAAAATTAAAGTAATTCTTGAA
AAAAAACTAATTCTAAGCTTCTCATGTCAAGGACCTAATG
TGCTAACACAGCACTTT**TTGACCAATTAT**TTTGGATCTGAAA
GAAATCAAGAATAATGAAGGACTTGATACATTGGAAGA
GGAGAGTCAAGGACCTACAGAAAAAAAAGAAAAAAGAAA
AAGAAAAGAAAAAGA**A**TTTAAAATTAAAGTAATTCTTGAA
AAAAAACTAATTCTAAGCTTCTT**C**ATGTCAAGGACCTAAT
GTCTGTGTTGCAGGTCTTCTTGCAATTCCCTGTCAAAAGA
AAAAGAATTAAAATTAAAGTAATTCTTGAAAAAAACTA
ATTCTAAGCTTCTCATGTCAAGGACCTAATGTCAGGCC
GGCTCTTATTAAAAACACCCTGTTCCCTGCCCTGGAGTG

Applications of Genome Analysis



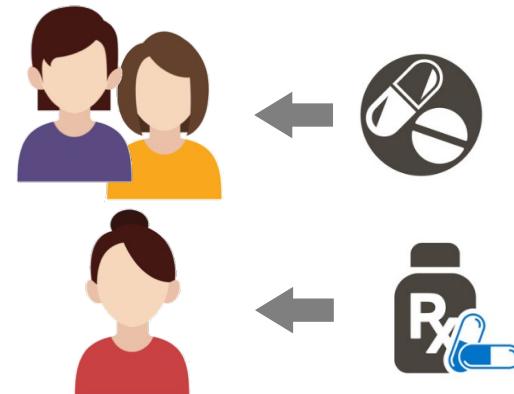
Understanding **genetic variations**



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



Rapid surveillance of **disease outbreaks**



Developing **personalized medicine**

And many other applications ...

How to Analyze a Genome?



NO

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



```
>CCTCCTCAGTGCCACCCAGCCCCTGGCAGCTCCAAACAGGGCTTTATTAAAACACCCGTCCCTGCCCCGGAGTGAGGTGTCAAG  
GACCTAAACTAAAAAAAAAAAAAGAAAAAGAAAAGAAAAGAATTAAAATTAAGTAATTCTTGAAAAAAACTAATTCTAAGCTTCTT  
CATGTCAAGGACCTAATGTGCTAACAGCACTTTTGACCATTATTTGGATCTGAAAGAAAATCAAGAATAATGAAGGACTGATACATTG  
GAAGAGGAGAGTCAGGACCTACAGAAAAAAAAAAAAAGAAAAAGAAAAGAAAAGAATTAAAATTAAGTAATTCTTGAAAAAA  
ACTAATTCTAAGCTTCTTCATGTCAAGGACCTAATGTCTGTGCAGGTCTTGCATTCCCTGTCAAAAGAAAAGAATTAAAATT  
AAGTAATTCTTGAAAAAAACTAATTCTAAGCTTCTTCAAGGTCAAGGACCTAATGTCAGGCCAAGAGTTGCAAAAAAAAAGAAAAA  
GAAAAGAAAAAGAATTAAAATTAAGTAATTCTTGAAAAAAACTAATTCTAAGCTTCTTCAAGGTCAAGGACCTAATGTAGCCAGAATGG  
TTGTGGGATGGGAGCCTCTGTGGACCGACCAGGTAGCTCTTTCCACACTGTAGTCTCAAAGCTTCTTCAAGGTCAAGGACCTAATGTAGTCAAGGACC  
TAATGTAGCTACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTAAGGTCAACACACCACCTCCAGAAAGCTTCTCA.....
```

DNA Testing



Fall DNA special
Just 55 CHF ~~89 CHF~~

Order now

The promotion ends today in 12 more hours!



<https://www.myheritage.ch/dna>

High-Throughput Sequencers



Illumina MiSeq



Illumina NovaSeq 6000



Pacific Biosciences RS II

Oxford
Nanopore
PromethION



Pacific
Biosciences
Sequel II



Oxford Nanopore MinION

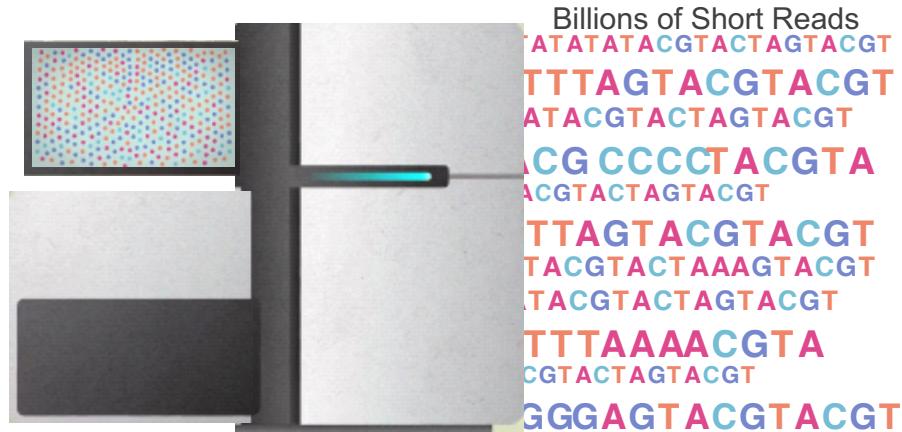


Oxford
Nanopore
SmidgION

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

Genome Sequencer is a Chopper

Regardless the sequencing machine,
reads still lack information about their order and location
(which part of genome they are originated from)



Solving the Puzzle

.FASTA file



Reference genome

.FASTQ file



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

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)
ACCCCT[REDACTED]TCATTGACATTAAACTCTGGGGCAGG[REDACTED]GAACGC GGCTGT CAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACC CGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCC[REDACTED]CCCCGGCCGGCTCGGGGCCCGCGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTGCTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGGTG[REDACTED]CAAAAGTAGCA[REDACTED]CTCCTA[REDACTED]TCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTGCATCCAGACCTCCTCTGCATCGCAGTT[REDACTED]CGCTTGGGAAAG
TCCGTACCCGCGCCT[REDACTED]AAAGACACCCCTGCCGCGGGTCGGCGAGGTGCAGCAGAAGTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTCTCAGAAAGACGC
```

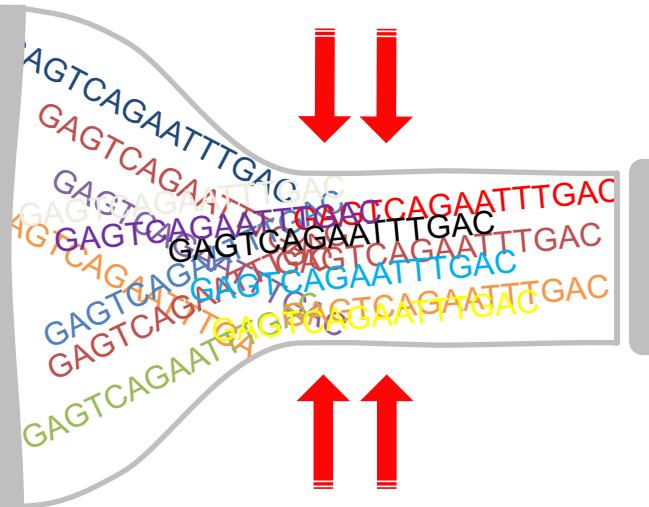
.FASTQ file:

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
T[REDACTED]AATAAAATCT[REDACTED]TTAGATN[REDACTED]NNNNNNNNNTAG
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeffffcfffffffdf`feed]`]_Ba_`__[YBBBBBBBBBBRTT
```

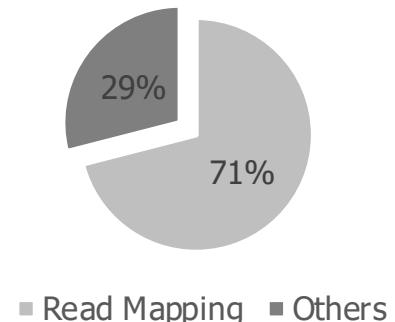
Analysis is Bottlenecked in Read Mapping!!

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

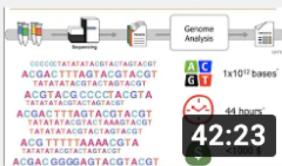


What is Intelligent Genome Analysis?

- Fast genome analysis Bandwidth
 - *Real-time analysis*
- Using intelligent architectures Energy-efficiency & Latency
 - *Specialized HW with less data movement*
- DNA is a valuable asset Privacy
 - *Controlled-access analysis*
- Population-scale genome analysis Scalability
 - *Sequence anywhere at large scale!*
- Avoiding erroneous analysis Accuracy
 - *E.g., your father is not your father*

Topics To Be Covered (I)

3



Mobile Genomics Course - Meeting 2: Introduction to Sequencing (Fall 2021)

Onur Mutlu Lectures

4

Mapping a read is similar to querying the yellow p

56:46

Mobile Genomics Course - Meeting 3: Read Mapping (Fall 2021)

Onur Mutlu Lectures

5

P&S Mobile Genomics
Lecture 4: GateKeeper

Dr. Mohammed Alser

@mealsr

ETH Zürich

Fall 2021

9 November 2021

1:05:51

Mobile Genomics Course - Meeting 4: GateKeeper (Fall 2021)

Onur Mutlu Lectures

6

P&S Mobile Genomics
Lecture 5: MAGNET & Shouji

Dr. Mohammed Alser

@mealsr

ETH Zürich

Fall 2021

10 November 2021

1:04:15

Mobile Genomics Course - Lecture 5: MAGNET & Shouji (Fall 2021)

Onur Mutlu Lectures

:

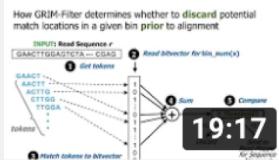
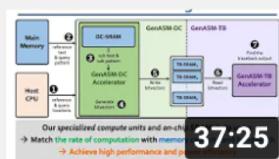
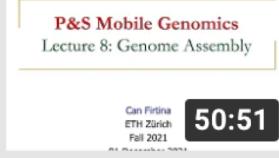
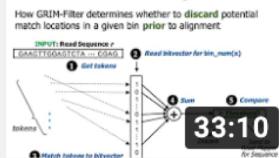
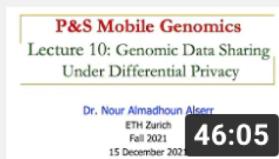
7



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

Onur Mutlu Lectures

Topics To Be Covered (II)

- 8 GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping w/ Processing-in-Memory - Jeremie Kim
Onur Mutlu Lectures
- 
- 19:17
- 9 Comp. Architecture - Lecture 9a: GenASM: Approx. String Matching Accelerator (ETH Zürich, Fall 2020)
Onur Mutlu Lectures
- 
- 37:25
- 10 Mobile Genomics Course - Lecture 8: Genome Assembly (Fall 2021)
Onur Mutlu Lectures
- 
- 50:51
- 11 Mobile Genomics Course - Lecture 9: GRIM-Filter (Fall 2021)
Onur Mutlu Lectures
- 
- 33:10
- 12 Mobile Genomics Course - Lecture 10: Genomic Data Sharing Under Differential Privacy (Fall 2021)
Onur Mutlu Lectures
- 
- 46:05

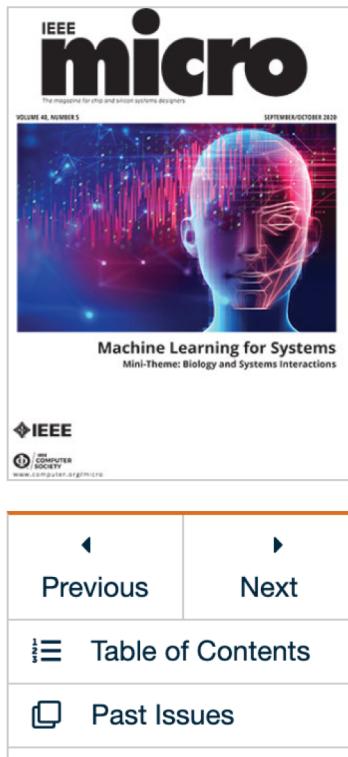
⋮

Near-memory Pre-alignment Filtering

Mohammed Alser, Zülal Bingöl, Damla Senol Cali, Jeremie Kim, Saugata Ghose, Can Alkan, Onur Mutlu

[“Accelerating Genome Analysis: A Primer on an Ongoing Journey”](#)

IEEE Micro, August 2020.



[Home](#) / [Magazines](#) / [IEEE Micro](#) / [2020.05](#)

IEEE Micro

Accelerating Genome Analysis: A Primer on an Ongoing Journey

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

DOI Bookmark: [10.1109/MM.2020.3013728](https://doi.org/10.1109/MM.2020.3013728)

Authors

[Mohammed Alser](#), ETH Zürich

[Zulal Bingol](#), Bilkent University

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

[Jeremie Kim](#), ETH Zurich and Carnegie Mellon University

[Saugata Ghose](#), University of Illinois at Urbana–Champaign and Carnegie Mellon University

[Can Alkan](#), Bilkent University

[Onur Mutlu](#), ETH Zurich, Carnegie Mellon University, and Bilkent University

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

Feedback From Our Community!



James Ferguson

@Psy_Fer_

This is awesome! I've got my evening reading sorted.



Stéphane Le Crom

@sleclrom

Very complete article on the evolution of read alignment algorithms. #NGS #genomics



Svetlana Gorokhova

@SGorokhova

An impressive overview of read alignment methods over the last three decades



BContrerasMoreira @BrunoContrerasM · Sep 10

Replies to [@mealser](#) [@GenomeBiology](#) and 3 others

Buen hilo de repaso sobre la evolución de los algoritmos de alineamiento de secuencias a medida que ha mejorado la tecnología de secuenciación

...

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]

Premieres in 23 hours
October 5, 4:30 PM

Storage Set reminder

Main Memory

Microprocessor

Specialized Pre-alignment Filtering Accelerators (GPU, FPGA)

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

(⌚) Premieres in 23 hours
October 5, 4:30 PM

Set reminder

SAFARI

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

19

0

SHARE

SAVE

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 \(at IEEE Micro, 2020\)\]](#)

The screenshot shows a video player interface with a presentation slide on the left and a video feed of the speaker on the right.

Our Solution: GateKeeper

1st FPGA-based Alignment Filter.

x10¹² mappings

x10³ mappings

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

High throughput DNA sequencing (HTS) technologies

Read Pre-Alignment Filtering

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)

35

More on Fast Genome Analysis ...

- Onur Mutlu,

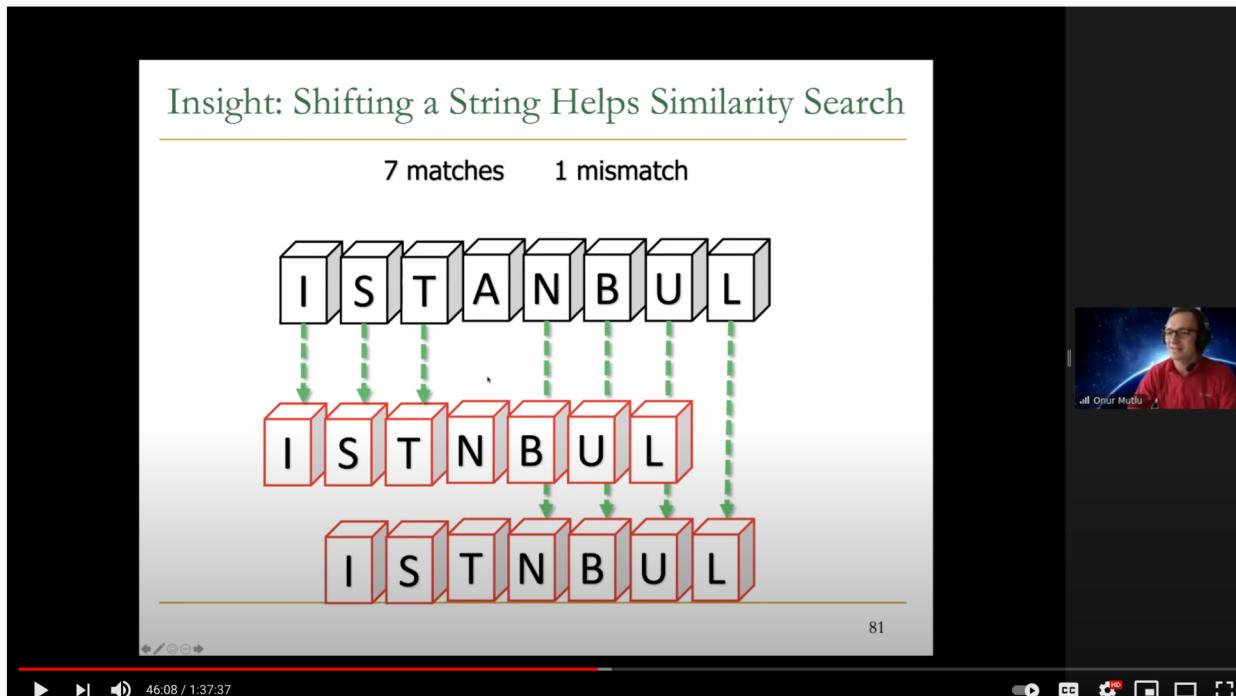
[**"Accelerating Genome Analysis: A Primer on an Ongoing Journey"**](#)

Invited Lecture at [Technion](#), Virtual, 26 January 2021.

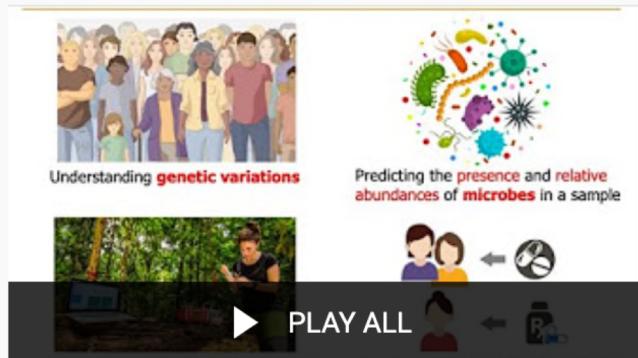
[[Slides \(pptx\)](#) ([pdf](#))]

[[Talk Video](#) (1 hour 37 minutes, including Q&A)]

[[Related Invited Paper \(at IEEE Micro, 2020\)](#)]



Two P&S Genomics Courses



Livestream - P&S Genome Sequencing on Mobile Devices (Fall 2021)

9 videos • 75 views • Updated 5 days ago



Onur Mutlu
Lectures

SUBSCRIBED



- 1
Mobile Genomics Course - Meeting 1: Course...
Onur Mutlu Lectures
- 2
Intelligent Genome Analysis Course - Meeting 1: Cours...
Onur Mutlu Lectures
- 3
Mobile Genomics Course - Meeting 2: Introduction to...
Onur Mutlu Lectures
- 4
Mobile Genomics Course - Meeting 3: Read Mapping...
Onur Mutlu Lectures
- 5
Mobile Genomics Course - Meeting 4: GateKeeper (Fall...
Onur Mutlu Lectures

https://www.youtube.com/playlist?list=PL5Q2soXY2Zi_U2F8yrrNPD9CjcM6CFQXv

Course Materials

2021 Meetings/Schedule

Week	Date	Livestream	Meeting	Learning Materials	Assignments
W1	5.10 Tue.	YouTube Live	M1: P&S Accelerating Genomics Course Introduction & Project Proposals PDF (PDF) PPT (PPT) YouTube Video	Required Materials Recommended Materials	
W2	20.10 Wed.	YouTube Live	M2: Introduction to Sequencing PDF (PDF) PPT (PPT)		
W3	27.10 Wed.	YouTube Live	M3: Read Mapping PDF (PDF) PPT (PPT)		
W4	3.11 Wed.	YouTube Live	M4: GateKeeper PDF (PDF) PPT (PPT)		
W5	10.11 Wed.	YouTube Live	M5: MAGNET & Shouji PDF (PDF) PPT (PPT)		
W6	17.11 Wed.		M6.1: SneakySnake PDF (PDF) PPT (PPT) Video		
			M6.2: GRIM-Filter PDF (PDF) PPT (PPT) YouTube Video		
W7	24.11 Wed.		M7: GenASM PDF (PDF) PPT (PPT) YouTube Video		

https://safari.ethz.ch/projects_and_seminars/fall2021/doku.php?id=bioinformatics

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 + "["SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs."](#)" to appear in *Bioinformatics*, 2020.
 - Senol Cali+, "["GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis"](#)", *MICRO* 2020.
 - Alser+, "["Technology dictates algorithms: Recent developments in read alignment"](#)", to appear in *Genome Biology*, 2021.
 - 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 Mobile Genomics

Introduction & Project Proposals

Dr. Mohammed Alser



ETH Zurich

Spring 2022

8 March 2022

BACKUP SLIDES

Reference Genome

.FASTA file:

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTTTTCTTATCATTGACATTAAACTCTGGGGCAGGTCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGCGGTGAGAAGTGTGGAACCGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAAACGCCCGGGCTCCGGCCCCGGCTCGGGGCCCGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCGCCCAAGTGGCCCCGGGCTTGATTTGCTTTAAAAG
GAGGCATAAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTGCAAAAGTAGCAAAATGTTCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGAGTAGGGGGCGGGAGTCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTGCATCCAGACCTCCTGCATCGCAGTTCACGACATCCACGCTGGAAAG
TCCGTACCCGCGCCTGGAGCGCTAAAGACACCCCTGCCCGGGTCGGCGAGGTGCAGCAGAAGTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGAACAAAGTCTAGAGATGGGTTCGTTCTCAGAAAGACGC
```

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_00001.11 Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
....
```

Genomic Reads

.FASTQ file:

Identifier ————— @HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
Sequence ————— TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTNNNNNNNNNTAGTTCTTGAGA
'+' sign ————— +
Quality scores ————— efcfffffffcfеffffcfffffdff`feed] `] _Ba_ ^ __ [YBBBBBBBBBBRTT\]] [] dddd`

Base T
phred Quality] = 29

Obtaining .FASTQ Files

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

The screenshot shows the NCBI SRA search interface. At the top, there's a navigation bar with the NCBI logo, 'Resources' (with a dropdown arrow), and 'How To'. Below the bar, the text 'SRA' is displayed, followed by a dropdown menu set to 'SRA' and an empty search input field. A link to 'Advanced' search is also present. A prominent orange banner at the top of the main content area contains a large exclamation mark icon and the text 'COVID-19 is an emerging, rapidly evolving situation.' Below the banner, there are links to 'Public health information (CDC)', 'Research information (NIH)', 'SARS-CoV-2 data (NCBI)', and 'Prevention and treatment information (WHO)'. The main content area displays study details for 'ERX215261: Whole Genome Sequencing of human TSI NA20754'. It includes a summary, design information, submission details, study information, sample details, library information, and run statistics.

Full ▾

Send to: ▾

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

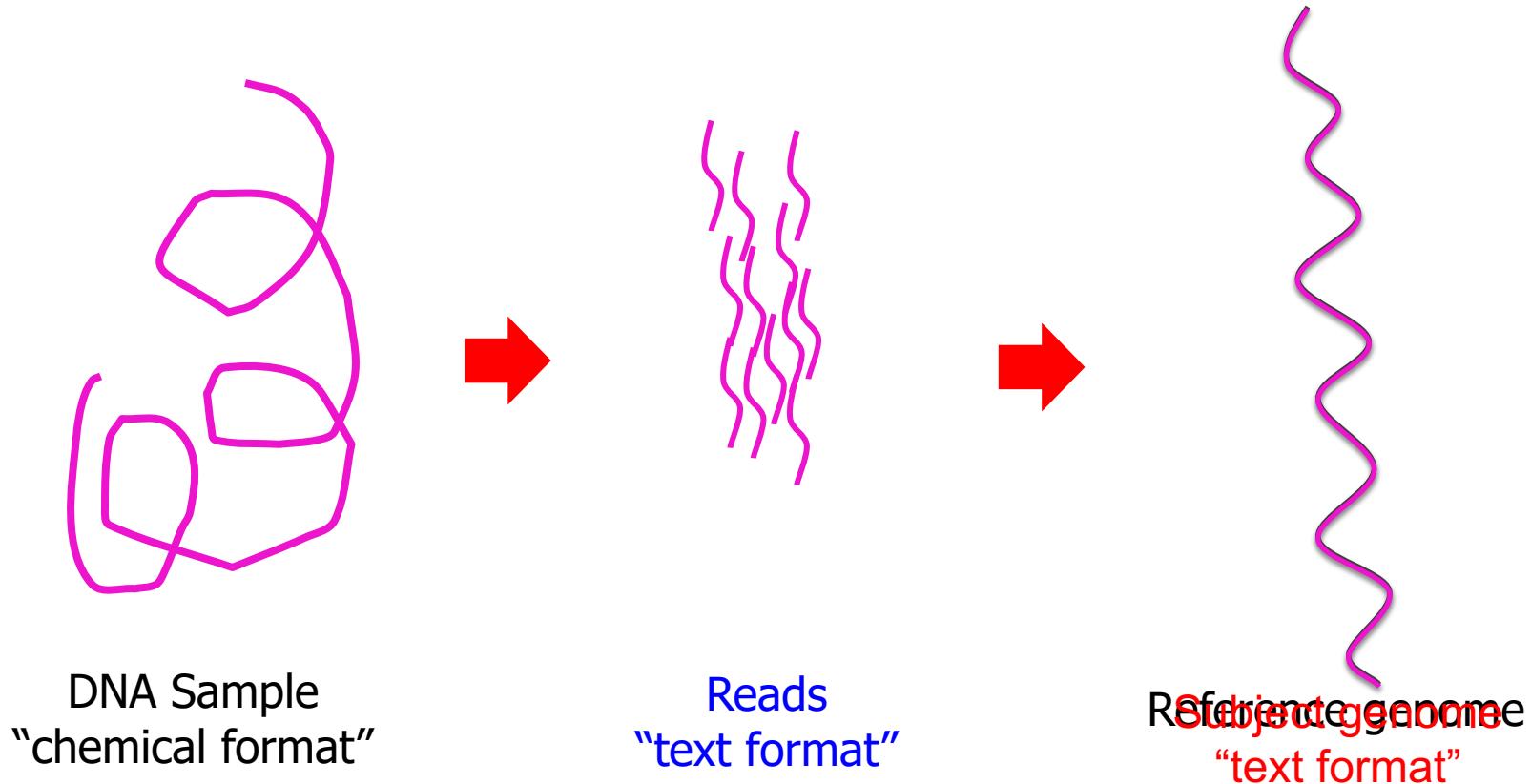
Construction protocol: Standard

Runs: 1 run, 4.1M spots, 818.7M bases, [387.2Mb](#)

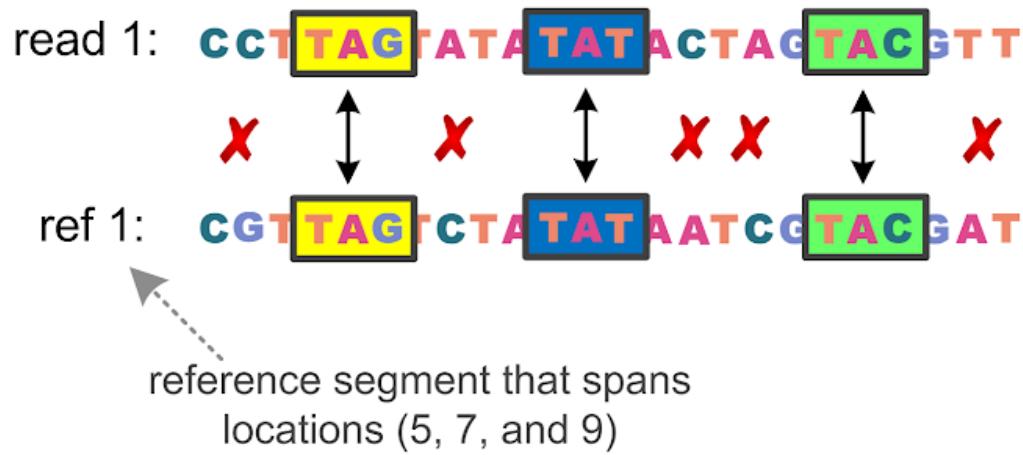
Run	# of Spots	# of Bases	Size	Published
ERR240727	4,093,747	818.7M	387.2Mb	2013-03-22

Read Mapping

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



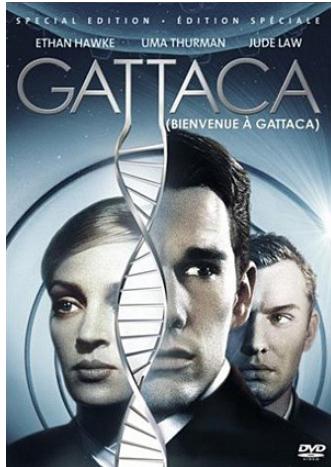
Base-by-Base Comparison



Fast Genome Analysis?

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

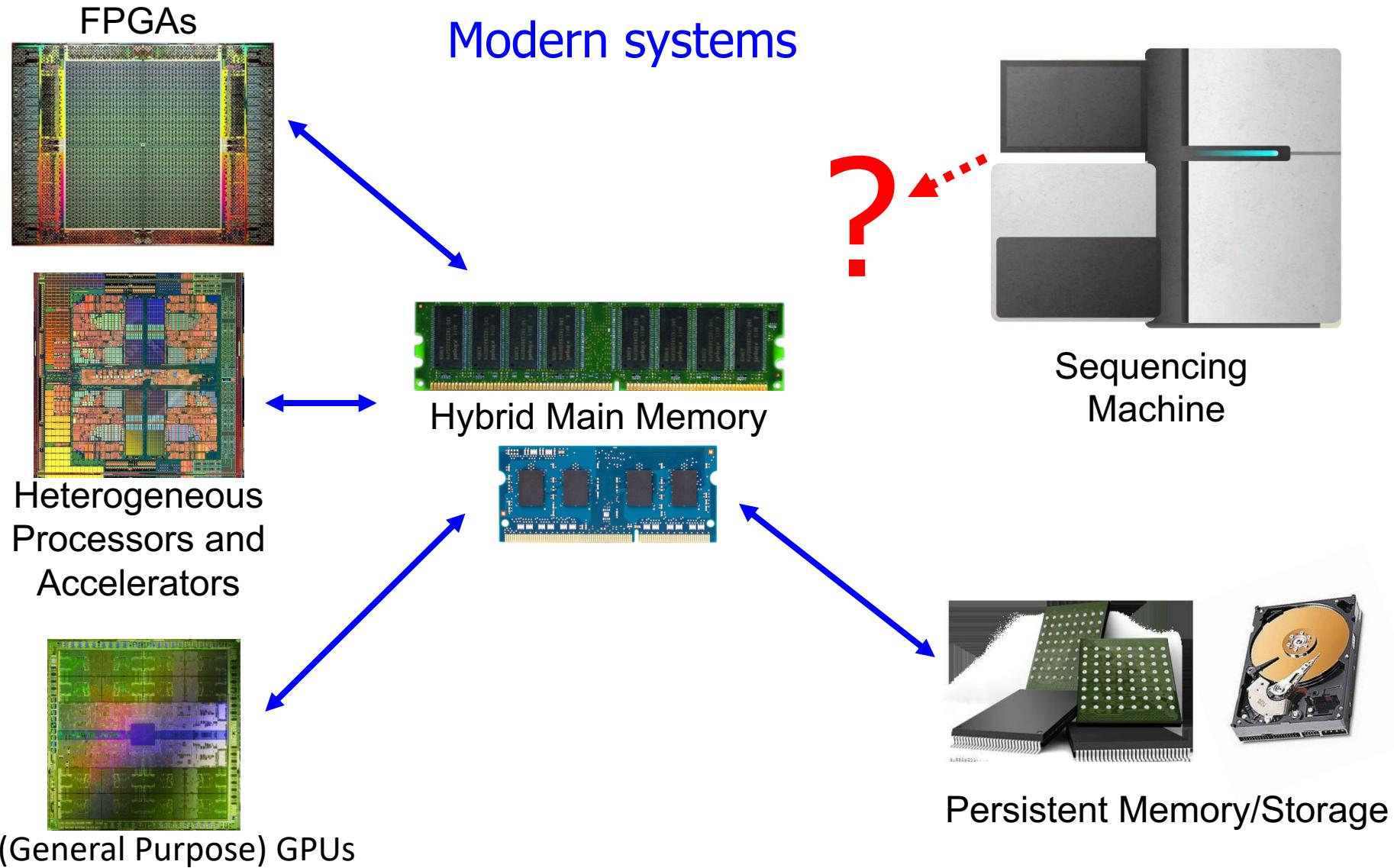
1997



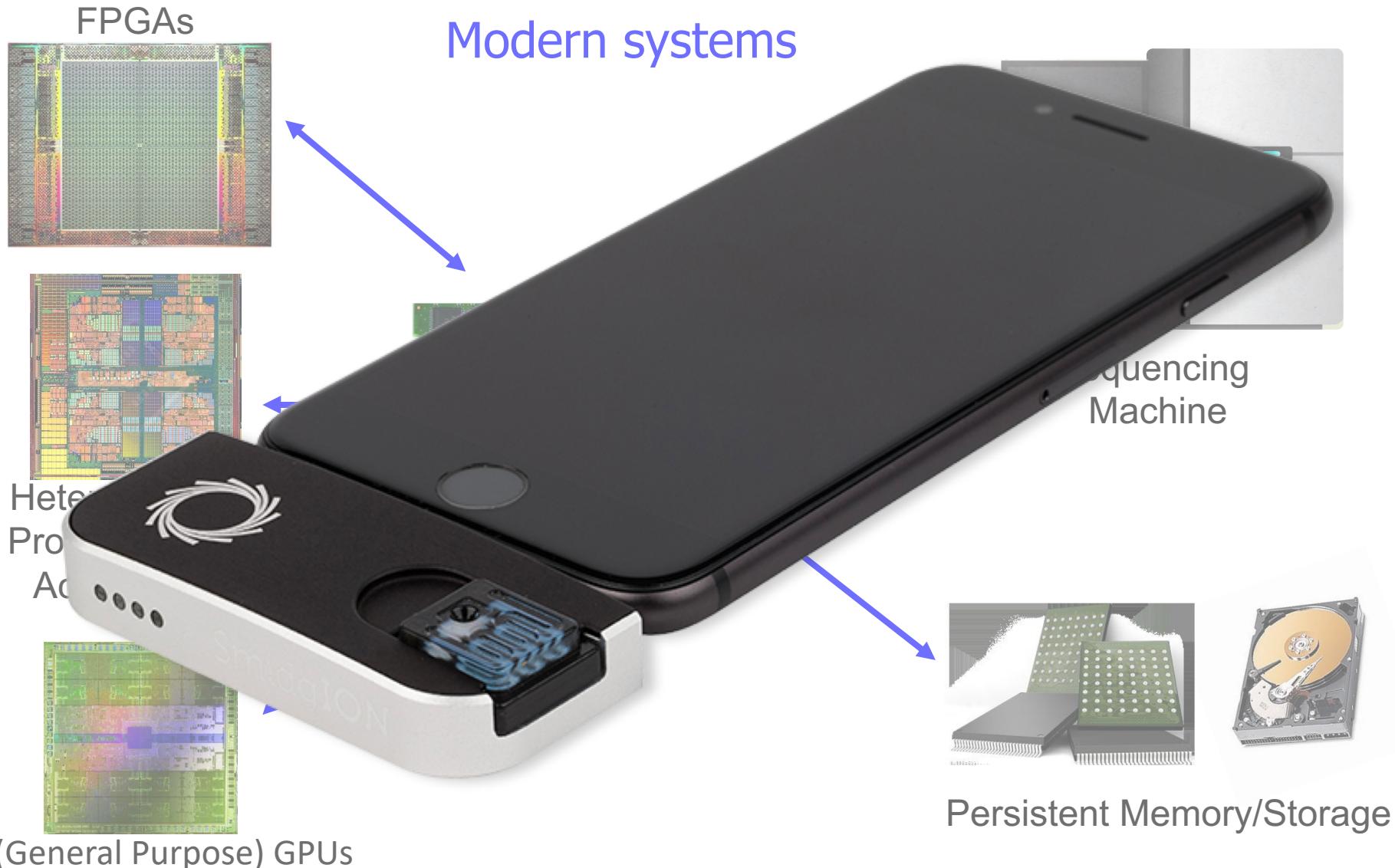
2015



Intelligent Architecture?



Intelligent Architecture?



Privacy-Preserving Genome Analysis?

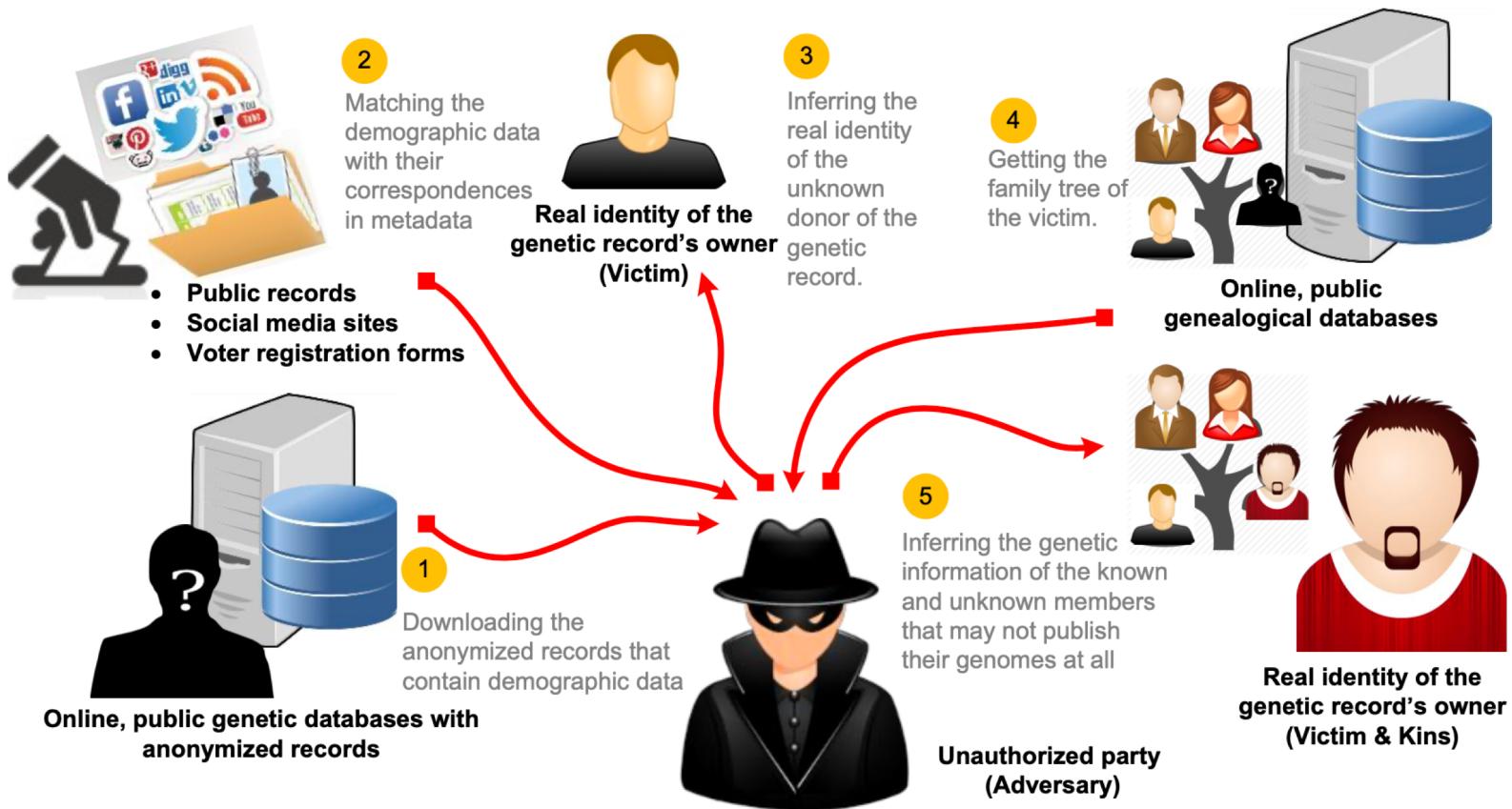


Fig. 5. A completion attack.

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.

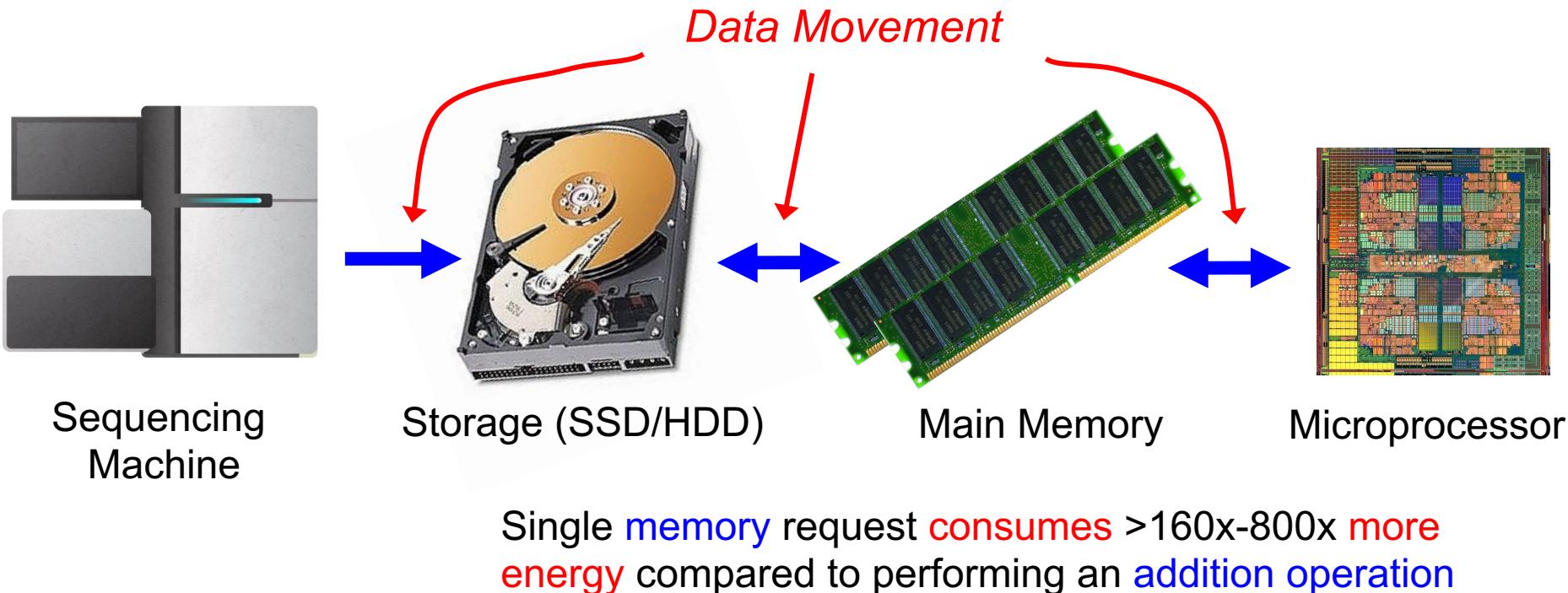
Keywords: Genomics, Privacy, Bioinformatics



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

Pushing Towards New Architectures

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

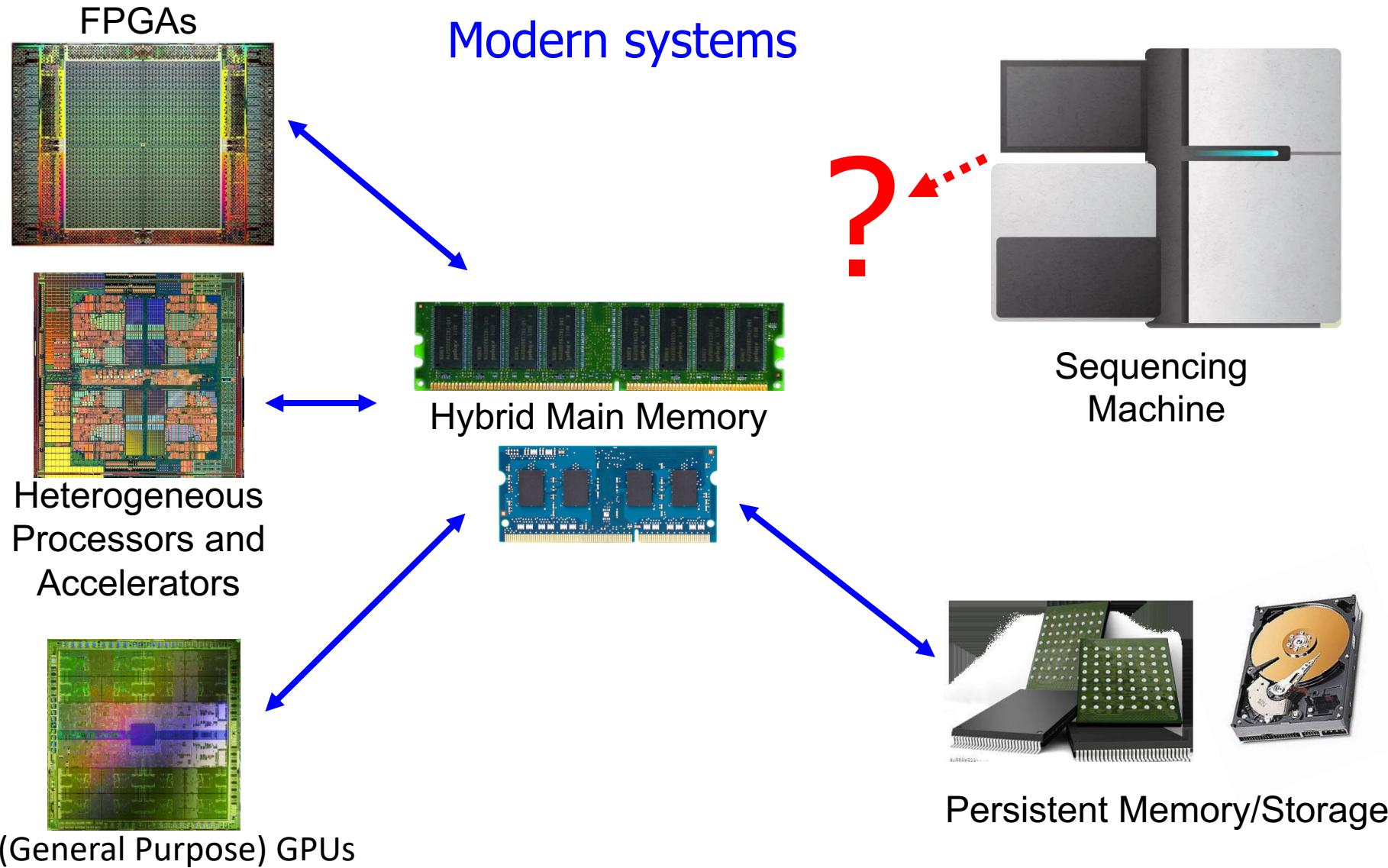


* Boroumand et al., “Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks,” ASPLOS 2018

★ Kestor et al., “Quantifying the Energy Cost of Data Movement in Scientific Applications,” IISWC 2013

★ Pandiyan and Wu, “Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms,” IISWC 2014

Processing Genomic Data Where it Makes Sense



Achieving Intelligent Genome Analysis?

How and where to enable
fast, accurate, cheap,
privacy-preserving, and exabyte scale
analysis of genomic data?

Most speedup comes from **parallelism** enabled
by novel architectures and **algorithms**