P&S Mobile Genomics

Project Proposals

Dr. Mohammed Alser

Prof. Onur Mutlu

ETH Zürich

Spring 2021

9 March 2021

SAFARI

ETH zürich

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?

- **Onur Mutlu**
  - Full Professor @ ETH Zurich ITET (INFK), since September 2015
  - Strecker Professor @ Carnegie Mellon University ECE/CS, 2009-2016, 2016-…
  - PhD from UT-Austin, worked at Google, VMware, Microsoft Research, Intel, AMD
  - https://people.inf.ethz.ch/omutlu/
  - omutlu@gmail.com (Best way to reach me)
  - https://people.inf.ethz.ch/omutlu/projects.htm
- Research and Teaching in:
  - Computer architecture, computer systems, hardware security, bioinformatics
  - Memory and storage systems
  - Hardware security, safety, predictability
  - Fault tolerance
  - Hardware/software cooperation
  - Architectures for bioinformatics, health, medicine
  - …

Course Info: Who Are We?

- Senior Researcher and Lecturer, SAFARI Research Group, ETH Zürich, since Sept. 2018.
- PhD from Bilkent University (Turkey) 2018, worked at UCLA, TU Dresden, and PETRONAS.
- PhD these in accelerating genome analysis, advisors: Can Alkan and Onur Mutlu, awarded:
  - IEEE Turkey Doctoral Dissertation Award
  - TÜBITAK fellowship
  - The Best Palestinian PhD Student in Turkey
  - HiPEAC Collaboration Grant
- ALSERM@ethz.ch, https://mealser.github.io/, 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?

Juan Gómez Luna

Senior Researcher and Lecturer

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

Damla Senol Cali

PhD Student (at CMU)

Hardware acceleration for bioinformatics tools | Genome sequence analysis tools | Hardware/Software Cooperation | Processing-in-Memory | Memory systems

Jeremie Kim

PhD Student

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

Can Firtina

PhD Student

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

- Get to know them and their research: https://safari.ethz.ch/safari-group/

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!

Course Website

- https://safari.ethz.ch/projects_and_seminars/spring2021/doku.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 start the projects **next week**.

WHAT IS GENOME ANALYSIS?

What is Genome Analysis?

Genomic analysis

Atom RSS Feed

Genomic analysis is the identification, measurement or comparison of genomic features such as DNA sequence, structural variation, gene expression, or regulatory and functional element annotation at a genomic scale. Methods for genomic analysis typically require high-throughput sequencing or microarray hybridization and bioinformatics.

https://onlinelearning.hms.harvard.edu/hmx/courses/genetic-testing/
https://www.nature.com/subjects/genomic-analysis

Genome Analysis

NO machine can read the ***entire*** content of a genome

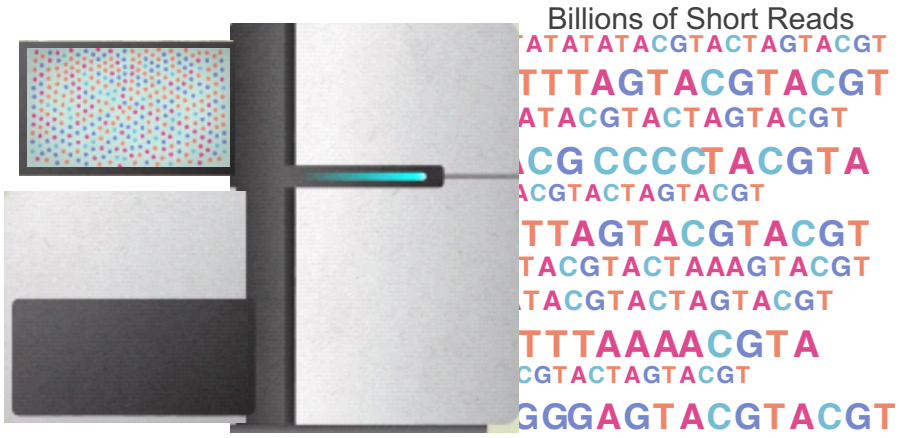
>CCTCCTCAGTGCCACCCAGCCCACTGGCAGCTCCCAAACAGGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTGGAGTGAGGTGTCAAG
GACCTAAACTAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTT
CATGTCAAGGACCTAATGTGCTAAACAGCACTTTTTTGACCATTATTTTGGATCTGAAAGAAATCAAGAATAAATGAAGGACTTGATACATTG
GAAGAGGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAA
ACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAAGAAAAAGAATTTAAAATTT
AAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAGGCCAAGAGTTGCAAAAAAAAAAAAAGAAAAA
GAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTAGCCAGAATGG
TTGTGGGATGGGAGCCTCTGTGGACCGACCAGGTAGCTCTCTTTTCCACACTGTAGTCTCAAAGCTTCTTCATGTGGTTTCTCTGAGTGAAA
AAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTTTTCATGTCAAGGACC
TAATGTAGCTATACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCACTTCCCAGAAAGCTTCTTCA.....

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)



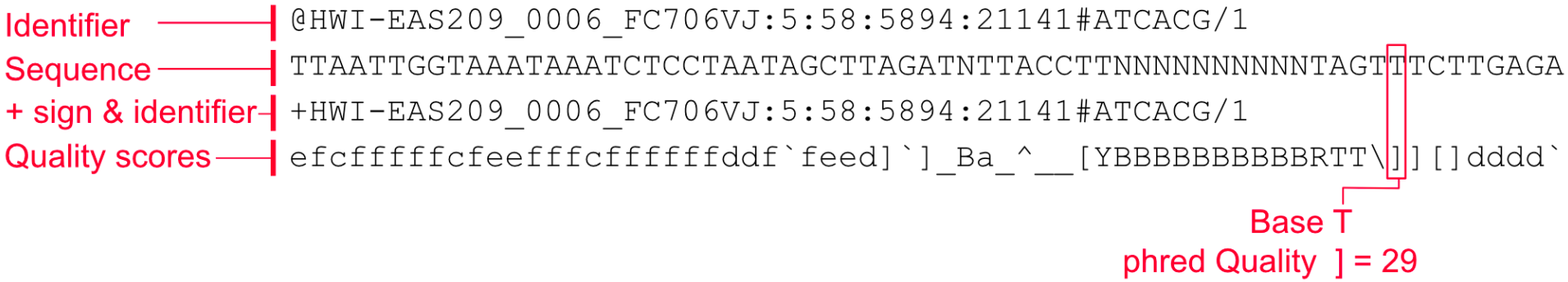
Reference Genome

.FASTA file:

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

Genomic Reads

.FASTQ file:

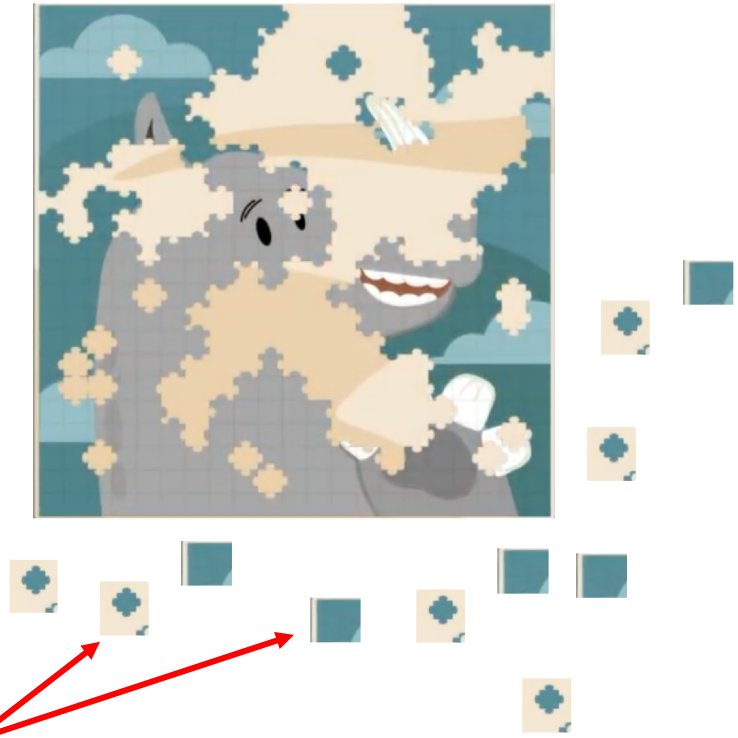


Solving the Puzzle

.FASTA file

Reference genome

.FASTQ file



Reads

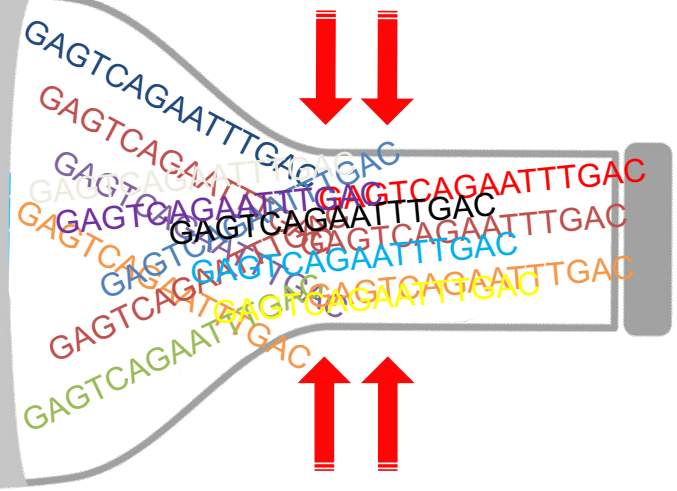
https://www.pacb.com/smrt-science/smrt-sequencing/hifi-reads-for-highly-accurate-long-read-sequencing/

Bottlenecked in Read Mapping!!

48 Human whole genomes

at 30× coverage

in about 2 days

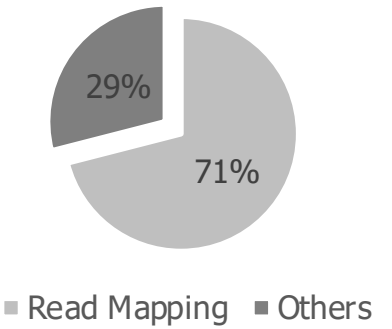


Illumina NovaSeq 6000

1 Human genome

32 CPU hours

on a 48-core processor

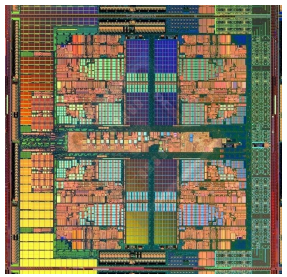


Goyal+, "Ultra-fast next generation human genome sequencing data processing using DRAGENTM bio-IT processor for precision medicine", *Open Journal of Genetics,* 2017.

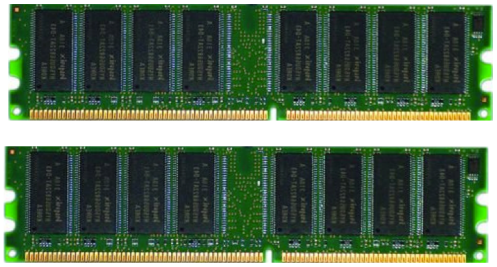
What is Intelligent Genome Analysis?

- 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*

Pushing Towards New Architectures



Microprocessor



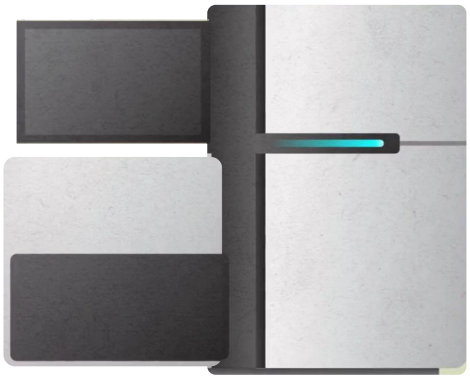
Main Memory

Storage (SSD/HDD)

Single memory request consumes

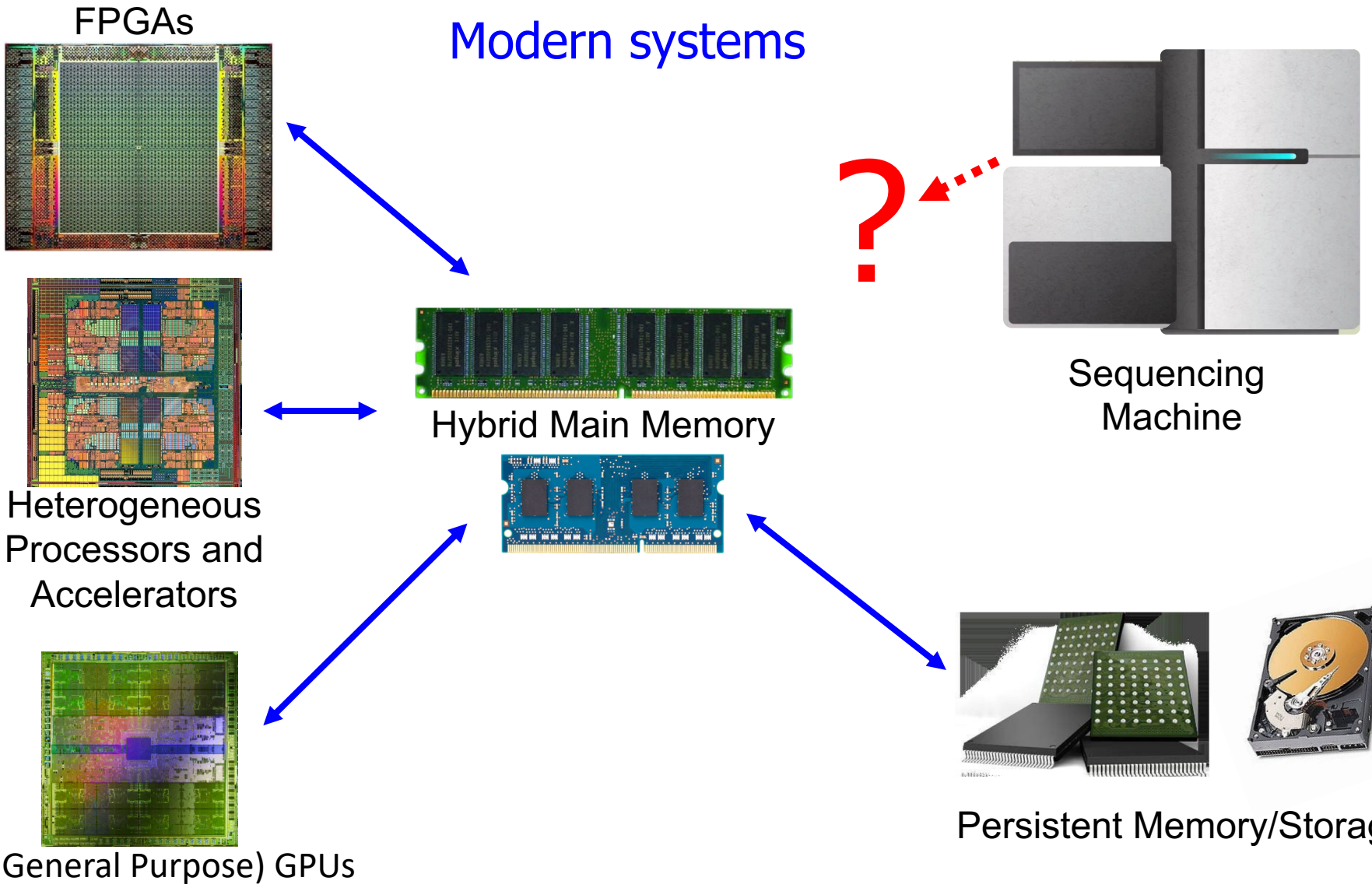
>160x-800x more energy compared to

performing a complex add operation



Sequencing Machine

Processing Genomic Data Where it Makes Sense



Achieving Intelligent Genome Analysis?

How and where to enable

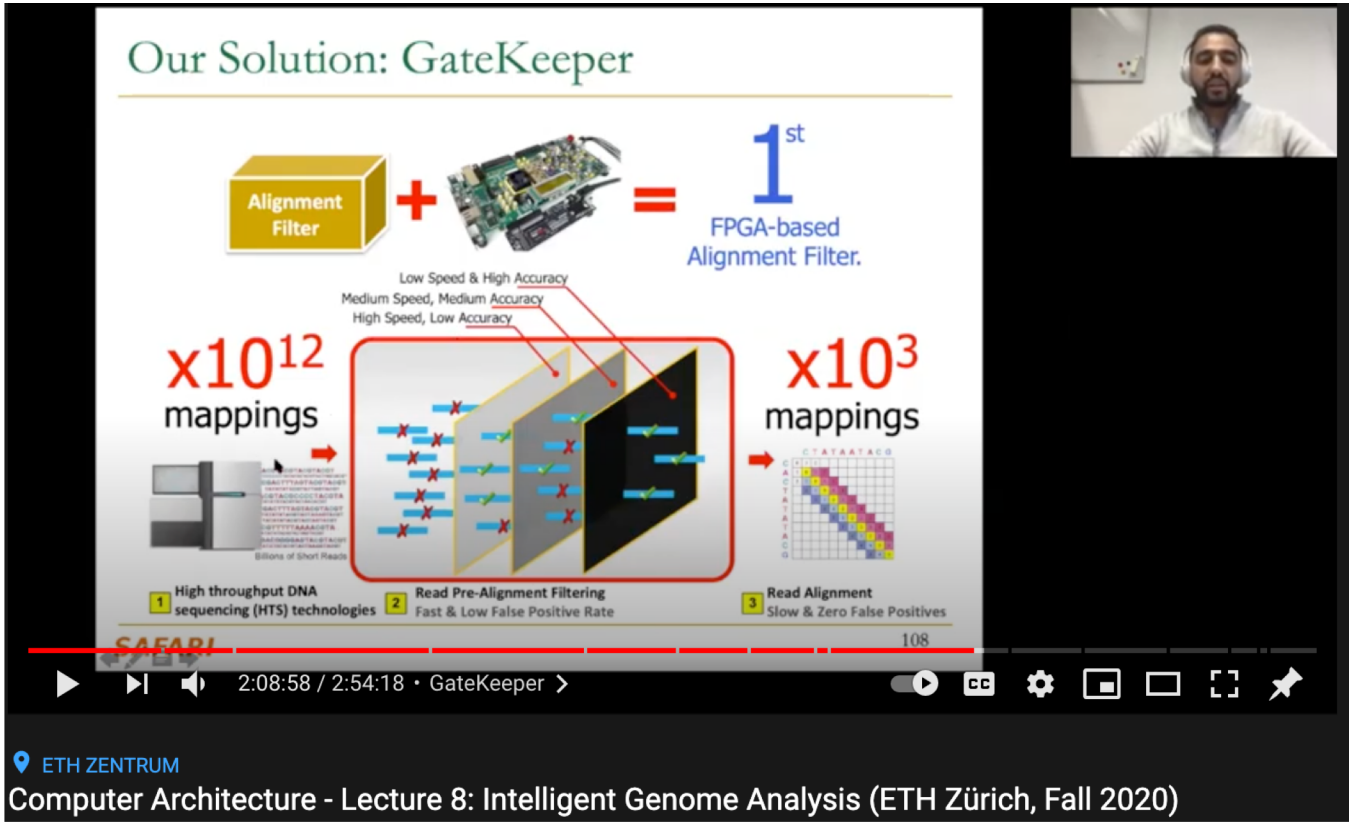
fast, accurate, cheap,

privacy-preserving, and exabyte scale

analysis of genomic data?

Most speedup comes from parallelism enabled by novel architectures and algorithms

More on This Topic!



https://www.youtube.com/watch?v=ygmQpdDTL7o

Prior Research on Genome Analysis (1/2)

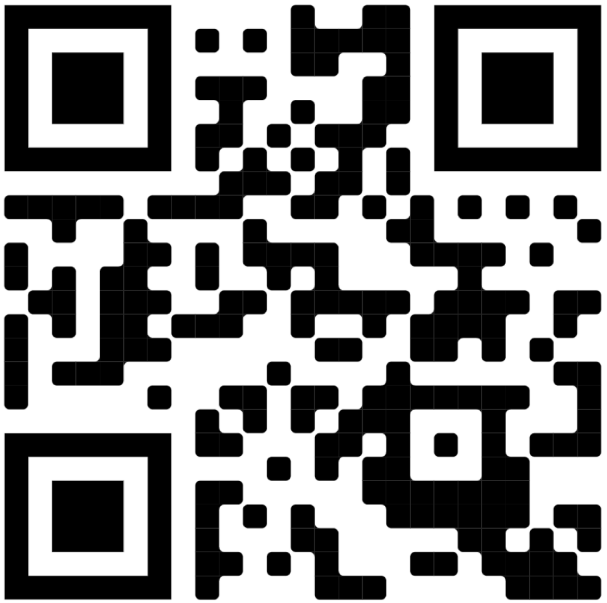
- Alser + "SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs.", *Bioinformatics,* 2020.
- Senol Cali+, "GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis", *MICRO* 2020.
- Alser+, "Technology dictates algorithms: Recent developments in read alignment", *arXiv*, 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.

Openings @ SAFARI

- We are **hiring** enthusiastic and motivated students and researchers at all levels.
- Join us now: safari.ethz.ch/apply



P&S Mobile Genomics

Project Proposals

Dr. Mohammed Alser

Prof. Onur Mutlu

ETH Zürich

Spring 2021

9 March 2021

SAFARI

ETH zürich