Analyzing Genomes via Intelligent Algorithms & Architectures

Dr. Mohammed Alser
@mealser

ETH Zurich
21 May 2022
What is a Genome?

An organism’s complete set of genetic instructions
CCTCCTCAGTGCCACCCAGCCCACTGGCAGCTCCCAAACA
GGCTCTTTATTTAAAAACACCCCTGTGTTCCCTGCCCCCTTGAGGATG
AGGTGTCAAGGACCTAAACTTTAAATTTTAAGTAAATTCTTTGAAA
AAAACACTAATTCTAAGCCTTCTTCATGTCAAGGACCTAAATTTGA
TGCTAAACAGCACTTTTTTTTGACCATTATTTTGAGATCTGAAA
GAAATCAAGAATAAAATGAAAGGACCTTGAGATACATTTGGAAGA
GGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAGAAA
AAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGA
AAAAAACTAATTCTAAGCCTTCTTCATGTCAAGGACCTAAATTTGA
TTGACCATTATAT
GAAATCAAGAATAAAATGAAAGGACCTTGAGATACATTTGGAAGA
GGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAGAAA
AAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGA
AAAAAACTAATTCTAAGCCTTCTTCATGTCAAGGACCTAAATTTGA
TTGACCATTATAT
C
How Large is a Genome?

Prime Tower, Zurich

~3.2 billion genomic bases
How About Other Species?

- **Phi X174 virus**: 5.386 Kilo 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
Lack of Specialized Compute Capability

48 Human whole genomes in about 2 days

1 genome 32 CPU hours on a 48-core processor

Specialized Machine for Sequencing

General-Purpose Machine for Analysis

FAST

SLOW
Need for Speed
1 in 17 people in the world have a rare disease
That’s 350 Million

80% of rare diseases are genetic in origin

30% of sick children never reach 5th birthday

1 in 17 people

582,450

30 Million

http://www.raredisease.org.uk/what-is-a-rare-disease
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.

---

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

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

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

---

**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

---

Safari, Farnaes+, “Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization”, NPJ Genomic Medicine, 2018
Scalable Genome Analysis

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

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

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

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]

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

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

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

*Boroumand et al., “Google Workloads for Consumer Devices: Mitigating Data Movement Bottlenecks,” ASPLOS 2018
* Kestor et al., “Quantifying the Energy Cost of Data Movement in Scientific Applications,” IISWC 2013
* Pandiyan and Wu, “Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms,” IISWC 2014
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
- GenASpM [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]
Our Contributions

**Near-memory/In-memory Pre-alignment Filtering**

- **GRIM-Filter** [BMC Genomics'18]
- **GenASM** [MICRO 2020]
- **SeGraM** [ISCA 2022]
- **SneakySnake** [IEEE Micro'21]

**Specialized Pre-alignment Filtering Accelerators (GPU, FPGA)**

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

[Image of FPGA board(s)]

[Links]
emea.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html
NVIDIA Clara Parabricks (2020)

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

GPU board(s)

PERFORMANCE COMPARISON
Germline End-to-End Secondary Analysis

<table>
<thead>
<tr>
<th></th>
<th>CPU/GATK</th>
<th>8X T4</th>
<th>8X V100</th>
<th>8X A100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (minutes)</td>
<td>1,200</td>
<td>52</td>
<td>35</td>
<td>23</td>
</tr>
</tbody>
</table>
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

For Research Use Only.
Not for use in diagnostic procedures.
NVIDIA is claiming a **7x improvement** in dynamic programming algorithm (**DPX instructions**) performance on a single H100 versus naïve execution on an A100.

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


Onur Mutlu, **Computer Architecture Lecture 2b**, Fall 2019, ETH Zurich
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.
Key Takeaway

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

Think BIG, Aim HIGH!

https://safari.ethz.ch
Onur Mutlu’s SAFARI Research Group

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

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

Think BIG, Aim HIGH!

https://safari.ethz.ch
Dear SAFARI friends,

2019 and the first three months of 2020 have been very positive eventful times for SAFARI.
Dear SAFARI friends,

Happy New Year! We are excited to share our group highlights with you in this second edition of the SAFARI newsletter (You can find the first edition from April 2020 here). 2020 has
SAFARI Newsletter December 2021 Edition

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

Think Big, Aim High

ETH zurich

View in your browser
December 2021
Learning Materials

https://www.youtube.com/c/OnurMutluLectures/search?query=Mohammed%20alser
Intelligent Genome Analysis

https://arxiv.org/abs/2205.07957

Quantitative Biology > Genomics

[Submitted on 16 May 2022]

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

Mohammed Alser, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez–Luna, Onur Mutlu
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
Feedback From Our Community!

James Ferguson
@Psy_Fer_

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

Stéphane Le Crom
@slecrom

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
Replying 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

https://twitter.com/mealser/status/1435223377644503040
Genomics (Fall 2021)

- **Fall 2021 Edition:**

- **Youtube Livestream:**
  - [https://www.youtube.com/watch?v=MnogTeMjY8k&list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J](https://www.youtube.com/watch?v=MnogTeMjY8k&list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J)

- **Project course**
  - Taken by Bachelor’s/Master’s students
  - Genomics lectures
  - Hands-on research exploration
  - Many research readings

---

### Fall 2021 Meetings/Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Livestream</th>
<th>Meeting</th>
<th>Learning Materials</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>5.10 Tue.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M1: P&amp;S Accelerating Genomics Course Introduction &amp; Project Proposals</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W2</td>
<td>20.10 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M2: Introduction to Sequencing</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W3</td>
<td>27.10 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M3: Read Mapping</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W4</td>
<td>3.11 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M4: GateKeeper</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W5</td>
<td>10.11 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M5: MAGNET &amp; Shouji</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W6</td>
<td>17.11 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M6.1: SneakySnake</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M6.2: GRIM-Filter</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W7</td>
<td>24.11 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M7: GenASM</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W8</td>
<td>01.12 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M8: Genome Assembly</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W9</td>
<td>13.12 Mon.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M9: GRIM-Filter</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
<tr>
<td>W10</td>
<td>15.12 Wed.</td>
<td><a href="https://www.youtube.com/watch?v=MnogTeMjY8k&amp;list=PL5Q2soXYZi8sgH-TrNznDhDkPq55J9J">YouTube Live</a></td>
<td>M10: Genomic Data Sharing Under Differential Privacy</td>
<td><a href="#">PDF</a></td>
<td><a href="#">PPT</a></td>
</tr>
</tbody>
</table>
# Spring 2022 Meetings/Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Livestream</th>
<th>Meeting</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>8.3</td>
<td>Live</td>
<td><strong>M1: P&amp;S Mobile Genomics Course</strong> Introduction &amp; Project Proposals</td>
</tr>
<tr>
<td></td>
<td>Tue.</td>
<td></td>
<td><a href="#">PDF</a> <a href="#">PPT</a></td>
</tr>
<tr>
<td>W2</td>
<td>15.3</td>
<td>Live</td>
<td><strong>M2: Introduction to Sequencing</strong></td>
</tr>
<tr>
<td></td>
<td>Tue.</td>
<td></td>
<td><a href="#">PDF</a> <a href="#">PPT</a></td>
</tr>
<tr>
<td>W3</td>
<td>22.3</td>
<td>Live</td>
<td><strong>M3: Read Mapping</strong></td>
</tr>
<tr>
<td></td>
<td>Tue.</td>
<td></td>
<td><a href="#">PDF</a> <a href="#">PPT</a></td>
</tr>
<tr>
<td>W4</td>
<td>29.3</td>
<td>Live</td>
<td><strong>M4: GateKeeper</strong></td>
</tr>
<tr>
<td></td>
<td>Tue.</td>
<td></td>
<td><a href="#">PDF</a> <a href="#">PPT</a></td>
</tr>
</tbody>
</table>

Livestream - **P&S Genome Sequencing on Mobile**

Onur Mutlu Lectures - 1 / 3

[Live Stream](#)

---

[https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=genome_seq_mobile](https://safari.ethz.ch/projects_and_seminars/spring2022/doku.php?id=genome_seq_mobile)
Analyzing Genomes via Intelligent Algorithms & Architectures

Dr. Mohammed Alser

@mealser

ETH Zurich

21 May 2022
Analyzing Genomes via Intelligent Algorithms & Architectures
Saturday, May 21 at 8:00 AM CET

Mohammed Alser
ETH Zurich, Switzerland