RECOMB 2021
PADOVA (ITALY) AUG 29 - SEP 1

August 29 – September 1
Virtual

Accelerating Genome Analysis
A Primer on an Ongoing Journey

Monday, August 30 at 4:50 PM ET

Mohammed Alser

Senior Researcher and Lecturer
ETH Zurich

Accelerating Genome Analysis

A Primer on an Ongoing Journey

Mohammed Alser

ETH Zurich

@mealser

RECOMB 2021 - Highlights

30 August 2021

RECOMB 2021
PADOVA (ITALY) AUG 29 - SEP 1

SAFARI

ETH zürich

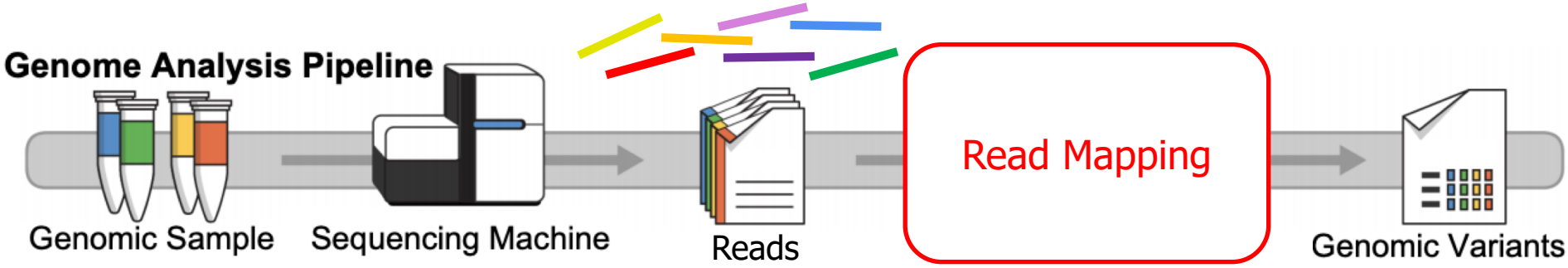
How to Analyze a Genome?

NO

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

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

Genome Analysis in Real Life



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

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

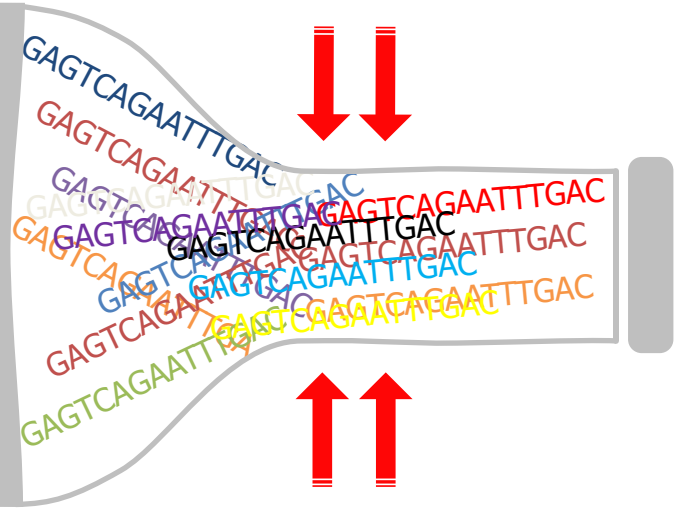
Analysis is Bottlenecked in Read Mapping!!

48 Human whole genomes

at 30× coverage

in about 2 days

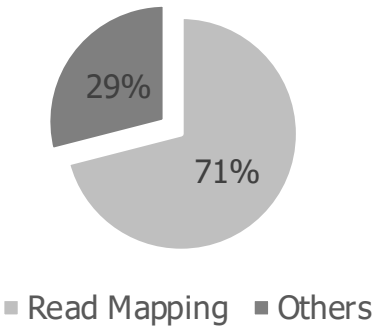
Illumina NovaSeq 6000



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.

Need for Speed

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

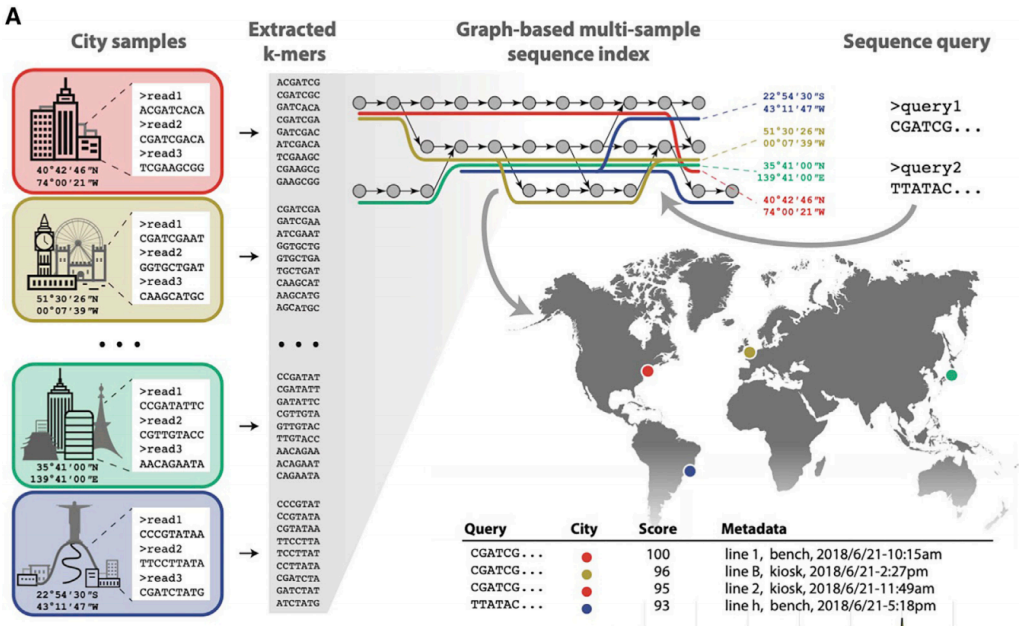
Farnaes+, "Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization", NPJ Genomic Medicine, 2018

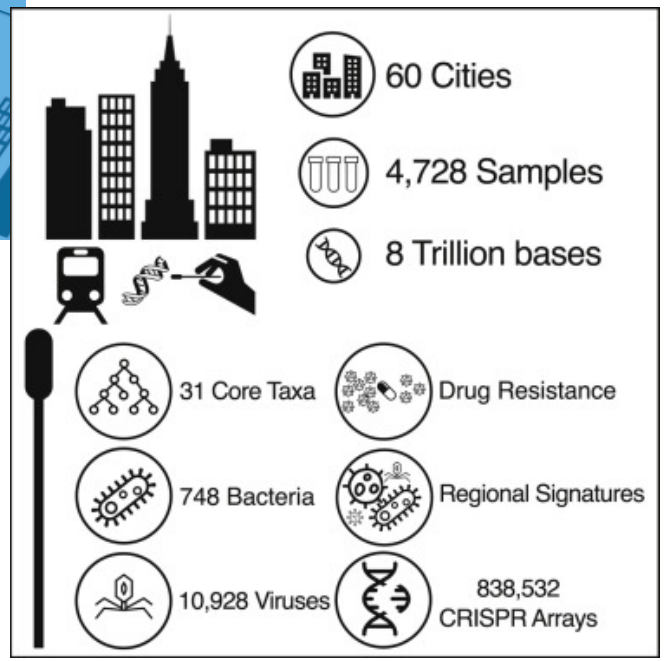
Scalable Genome Analysis

"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 Microbiome Profiling

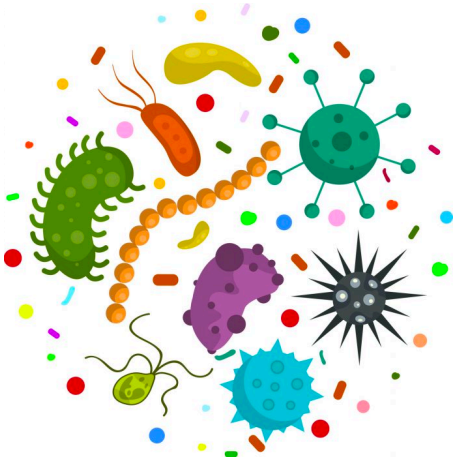




Danko+, "A global metagenomic map of urban microbiomes and antimicrobial resistance", Cell, 2021

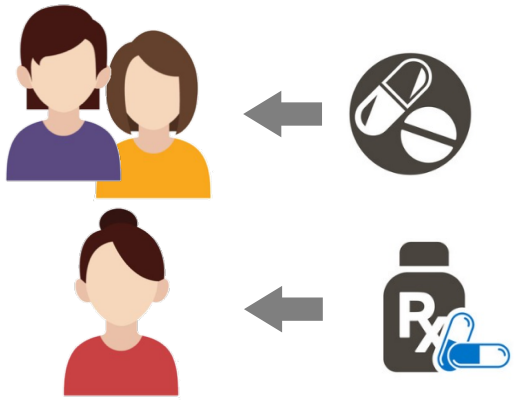
We Need Faster & Scalable Genome Analysis

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 …

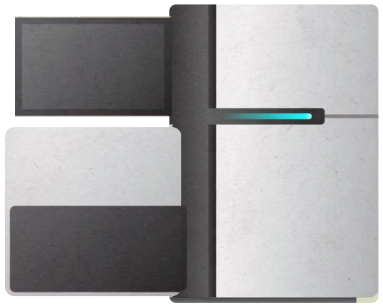
What makes read mapping a **bottleneck**?

A Tsunami of Sequencing Data

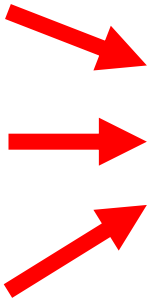
A Tera-scale increase in sequencing production in the past 25 years		
Genes & Operons	1990	**Kilo** = 1,000
Bacterial genomes	1995	**Mega** = 1,000,000
Human genome	2000	**Giga** = 1,000,000,000
Human microbiome	2005	**Tera** = 1,000,000,000,000
50K Microbiomes	2015	**Peta** = 1,000,000,000,000,000
what 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

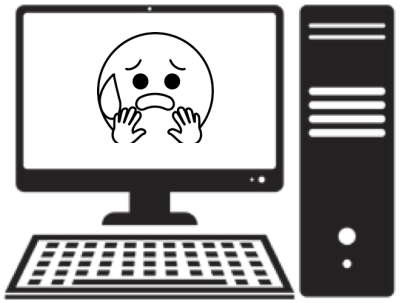
Source: @kyrpides

Lack of Specialized Compute Capability



Specialized Machine
for Sequencing





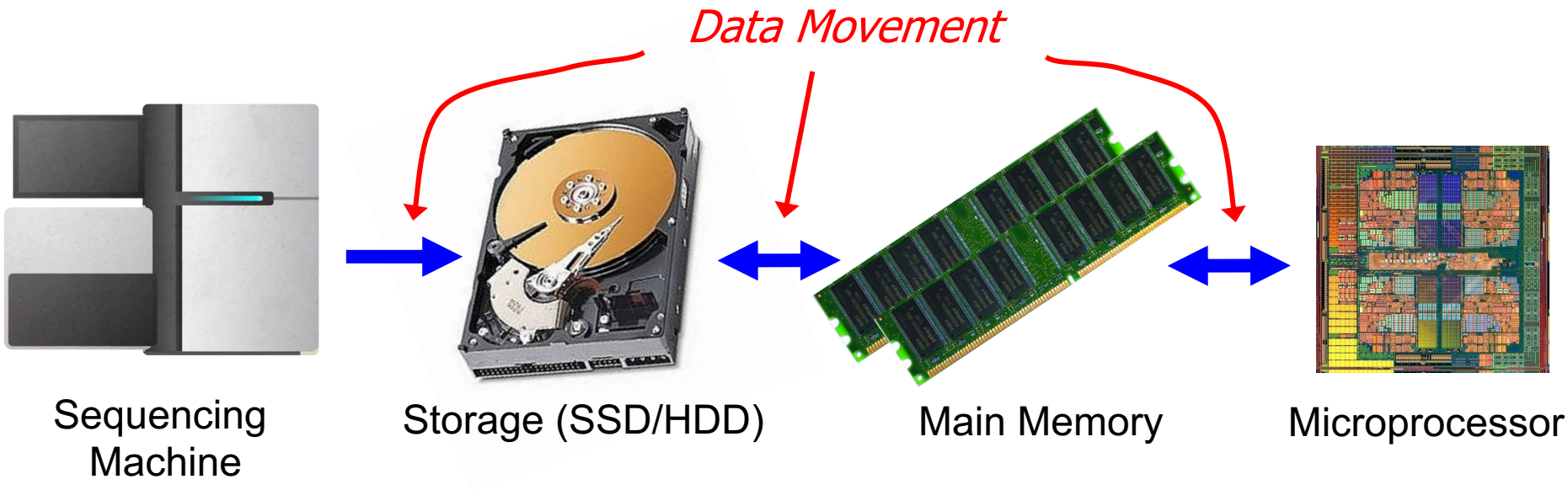
General-Purpose Machine
for Analysis

FAST

SLOW

Data Movement Dominates Performance

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



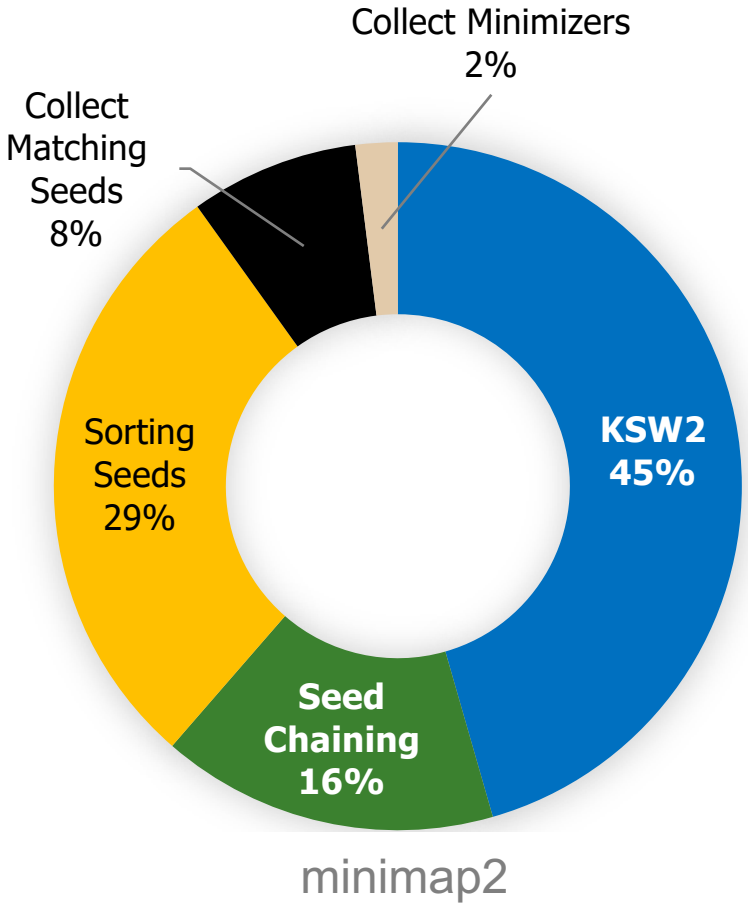
Single memory request consumes >160x-800x more energy compared to performing an addition operation

* 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 Execution Time

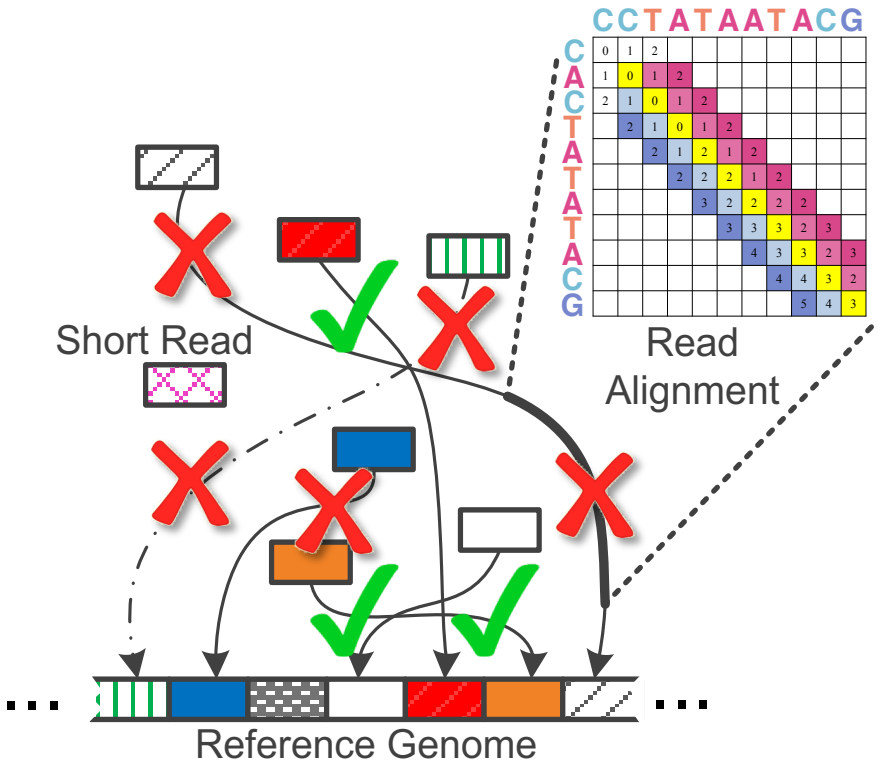
>60%

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



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

Large Search Space for Mapping Location



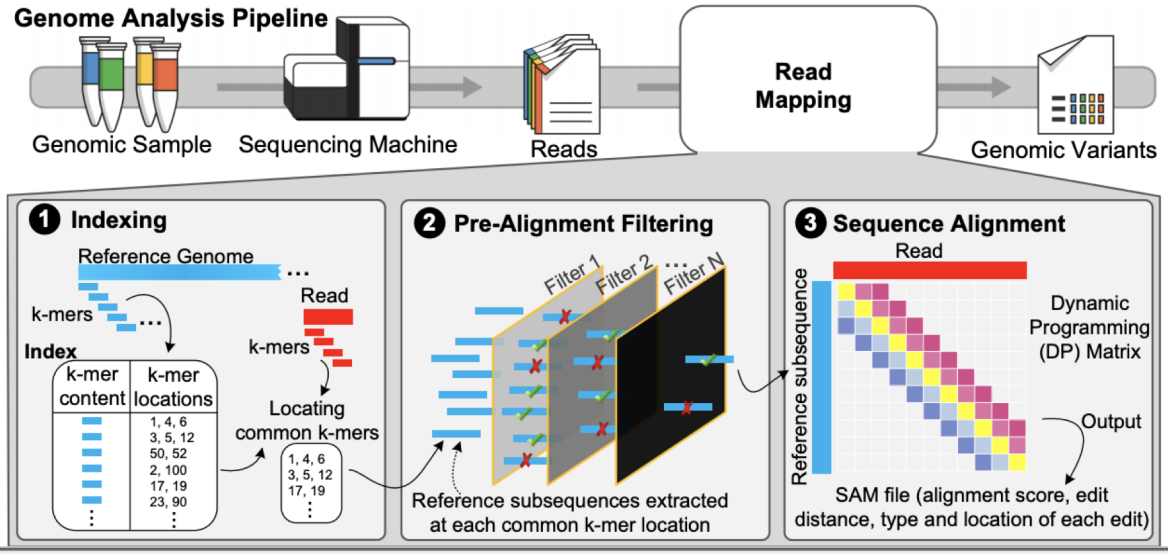
98%

of candidate locations have high dissimilarity with a given read

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

We need intelligent algorithms and intelligent architectures that handle data well

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

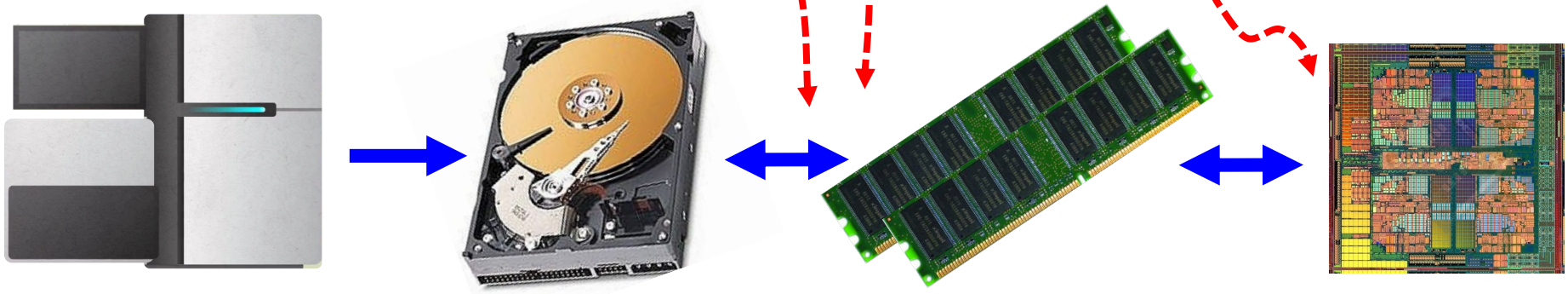
- **GRIM-Filter [BMC Genomics'18]**
- **SneakySnake [IEEE Micro'21]**
- **GenASM [MICRO 2020]**

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]**



Sequencing Machine | Storage (SSD/HDD) | Main Memory | Microprocessor

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

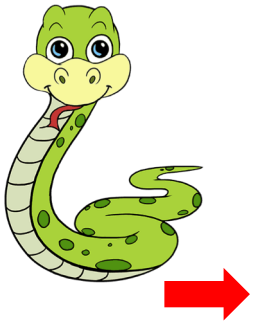
Mohammed Alser ✉, Taha Shahroodi, Juan Gómez-Luna, Can Alkan ✉, Onur Mutlu ✉

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

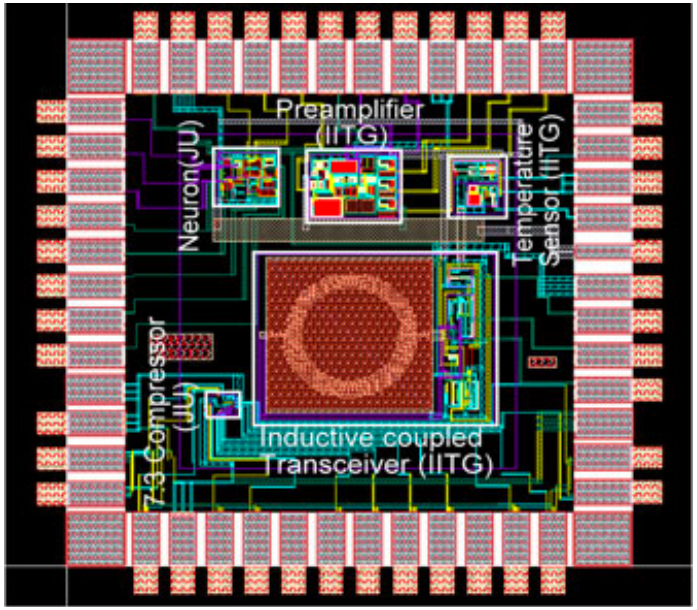
Published: 26 December 2020 **Article history ▾**

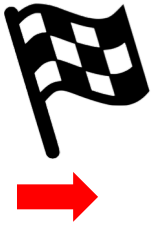
SneakySnake

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



VLSI chip layout





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

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

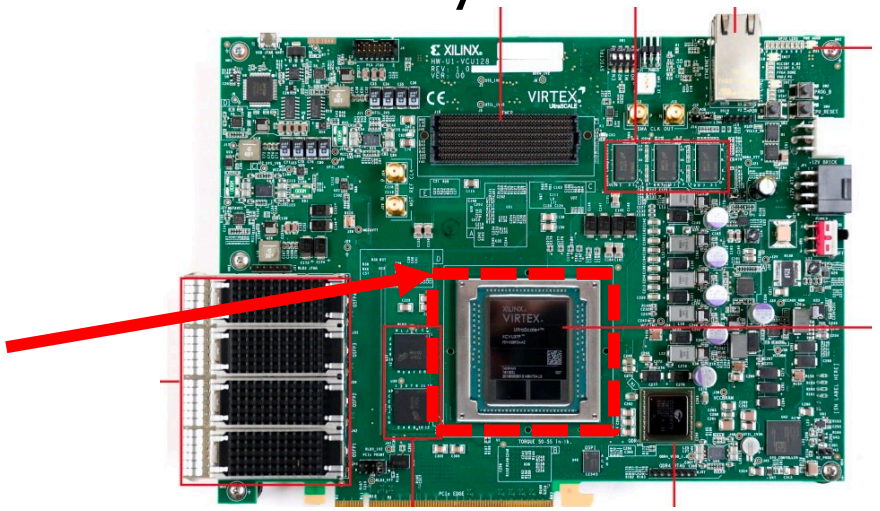
Authors

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

Near-memory SneakySnake

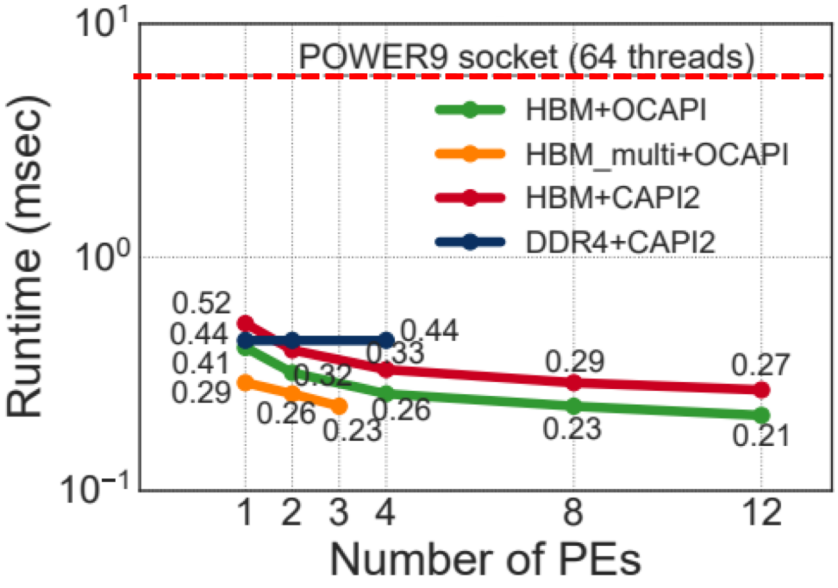
- Problem: Read Mapping is heavily bottlenecked by data movement from main memory
- Solution: Perform read mapping near where data resides (i.e., near-memory)
- We carefully redesigned the accelerator logic of SneakySnake to exploit near-memory computation capability on modern FPGA boards with high-bandwidth memory

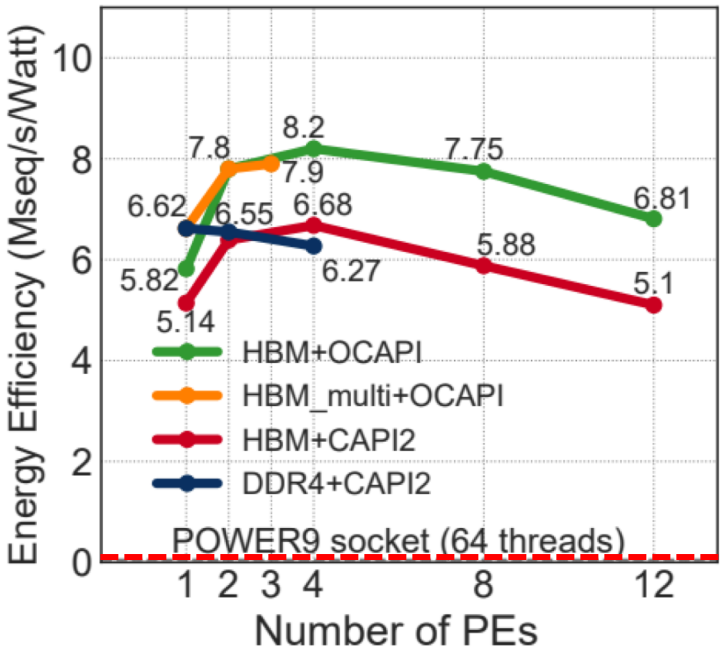
FPGA + high-bandwidth memory on the same package substrate



Xilinx Virtex UltraScale+ HBM VCU128 FPGA

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

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**)

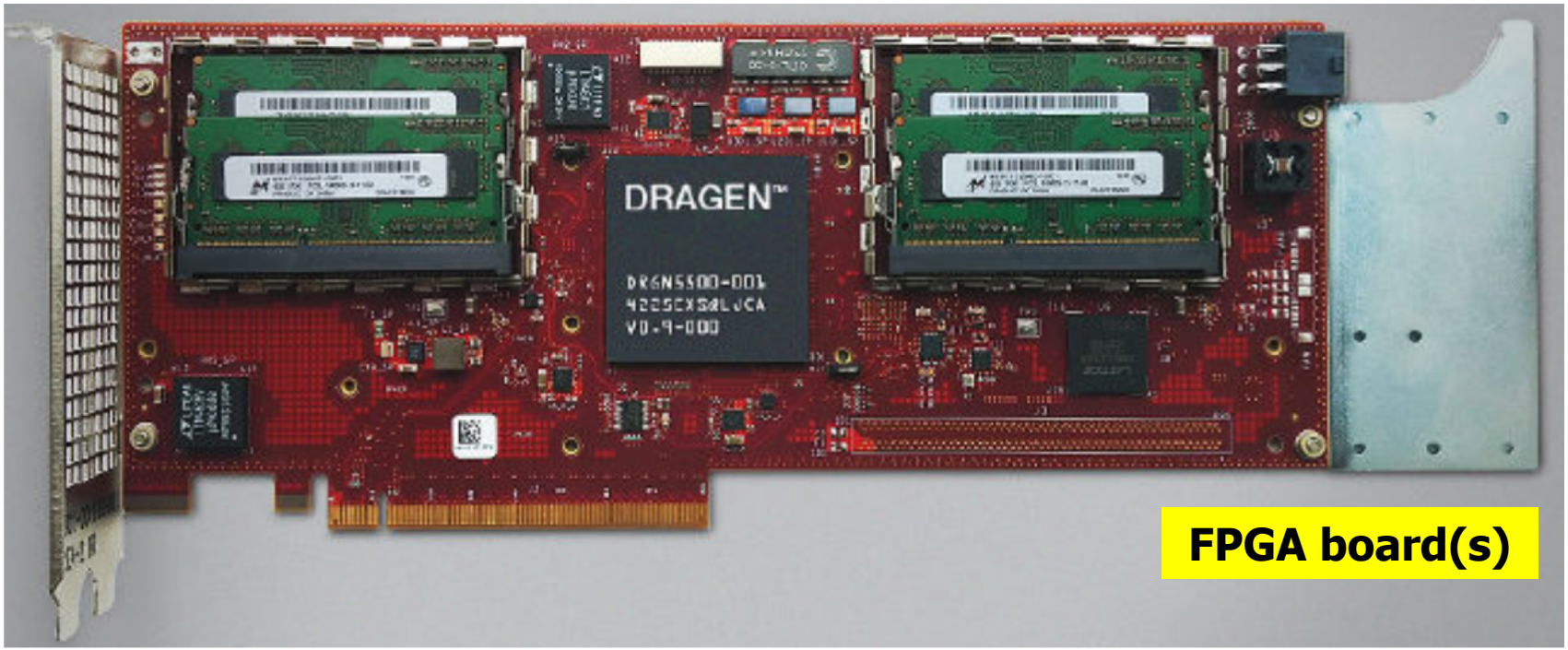
(3) Edit Distance Calculation

- **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**)

Adoption of hardware accelerators in genome analysis

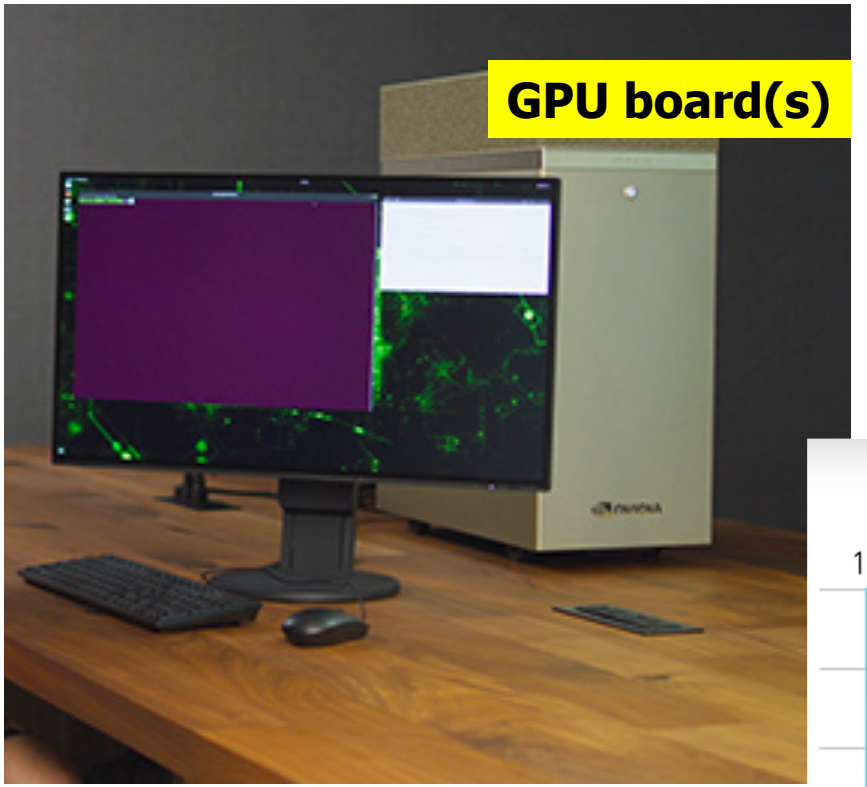
Illumina DRAGEN Bio-IT Platform

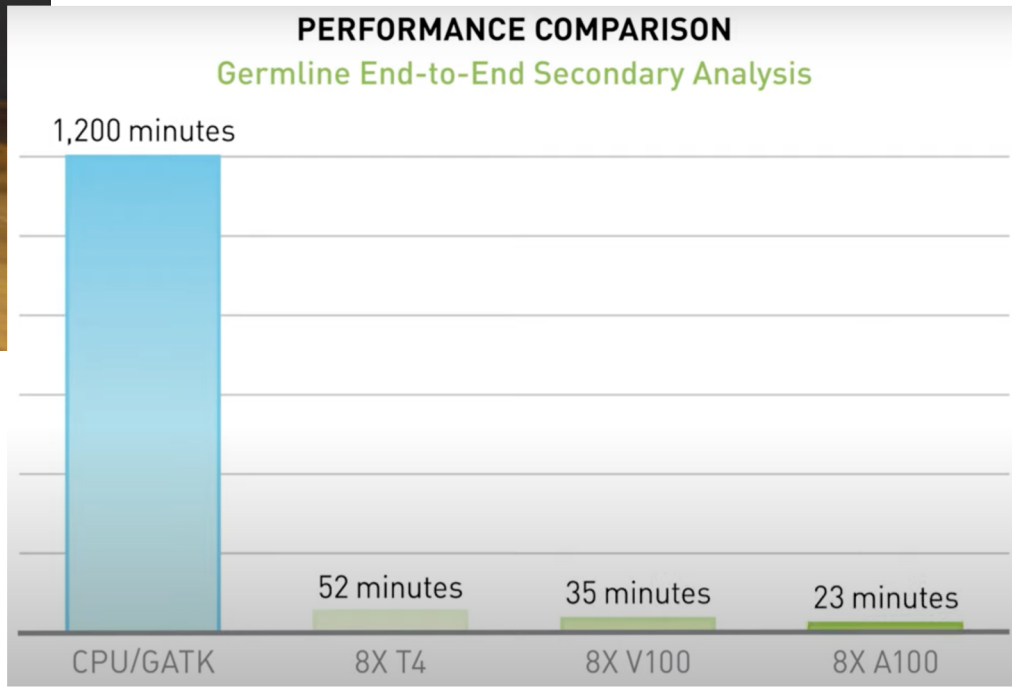
- Processes whole genome at 30x coverage in ~25 minutes with hardware support for data compression



https://emea.illumina.com/products/by-type/informatics-products/dragen-bio-it-platform.html

NVIDIA 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

Where is Read Mapping Going Next?

NO

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

What if we have such a machine?!

Think about comparing large genomes together as in whole-genome alignment, metagenomics, pangenomics, …

Open Questions

How and where to enable

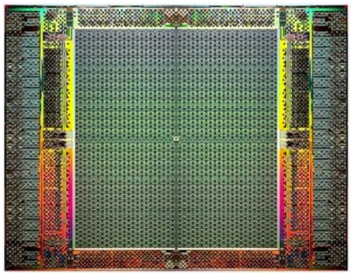
fast, accurate, cheap,

privacy-preserving, and exabyte-scale

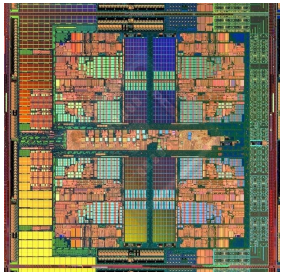
analysis of genomic data?

Processing Genomic Data Where it Makes Sense

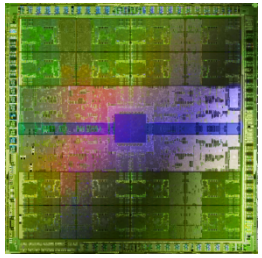
FPGAs



Modern systems

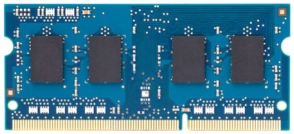


Heterogeneous Processors and Accelerators

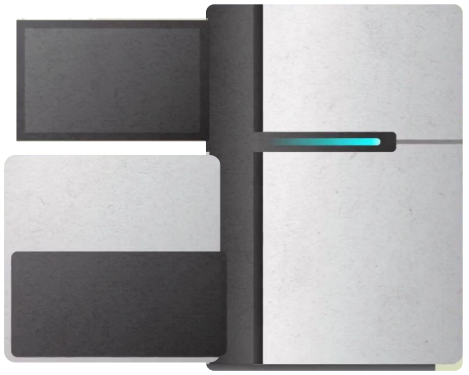


(General Purpose) GPUs

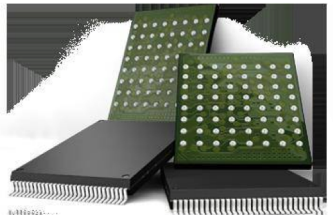
Hybrid Main Memory



?



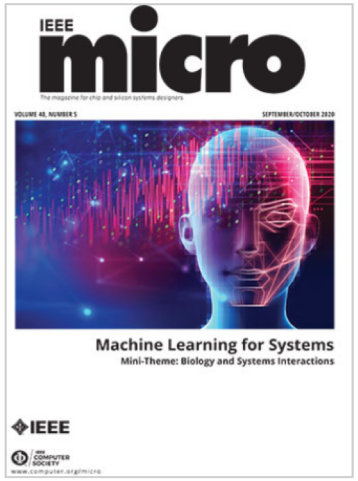
Sequencing Machine



Persistent Memory/Storage

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.



Previous
Next
Table of Contents
Past Issues

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

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

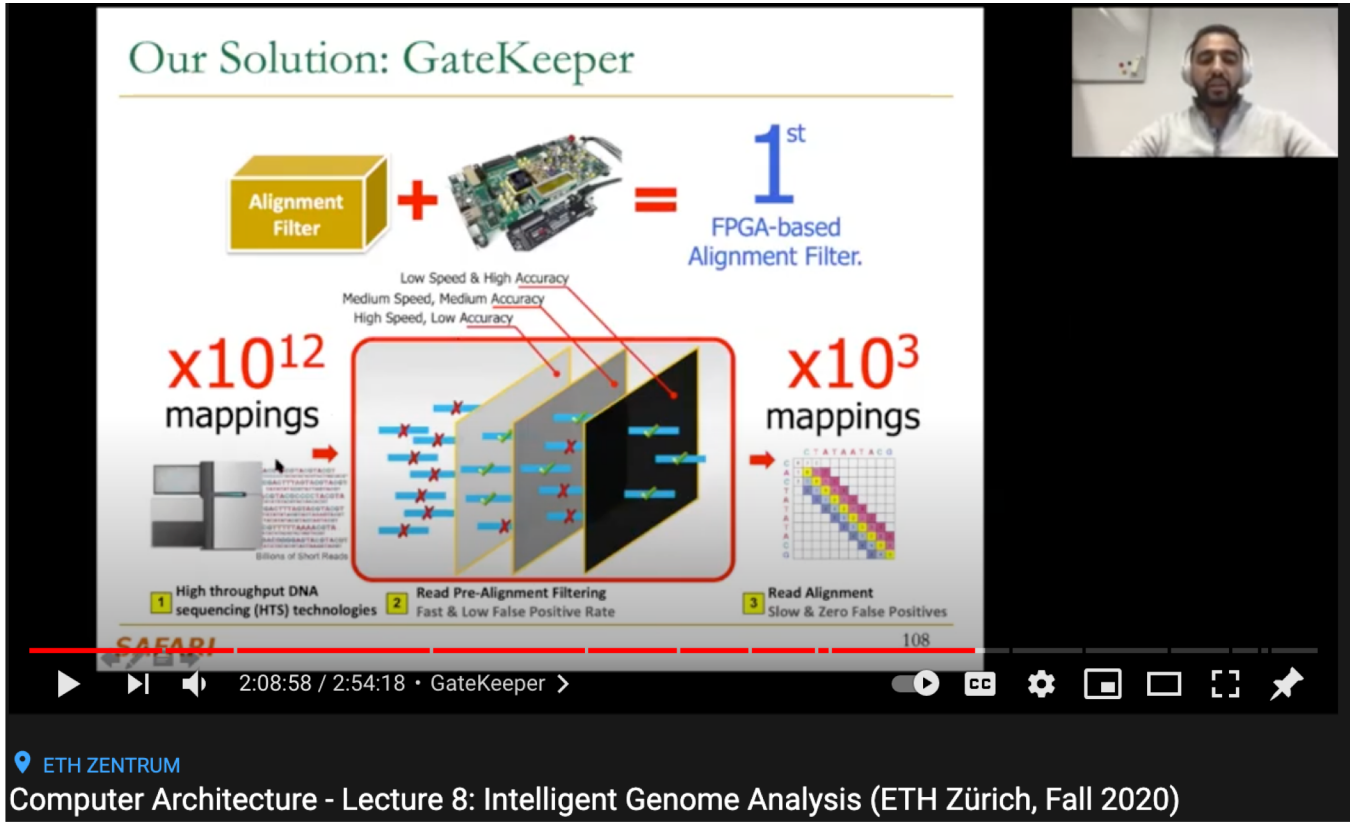
Genome Biology

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*†]

More on Read Mapping



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

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.

Accelerating Genome Analysis

A Primer on an Ongoing Journey

Mohammed Alser

ETH Zurich

@mealser

RECOMB 2021 - Highlights

30 August 2021

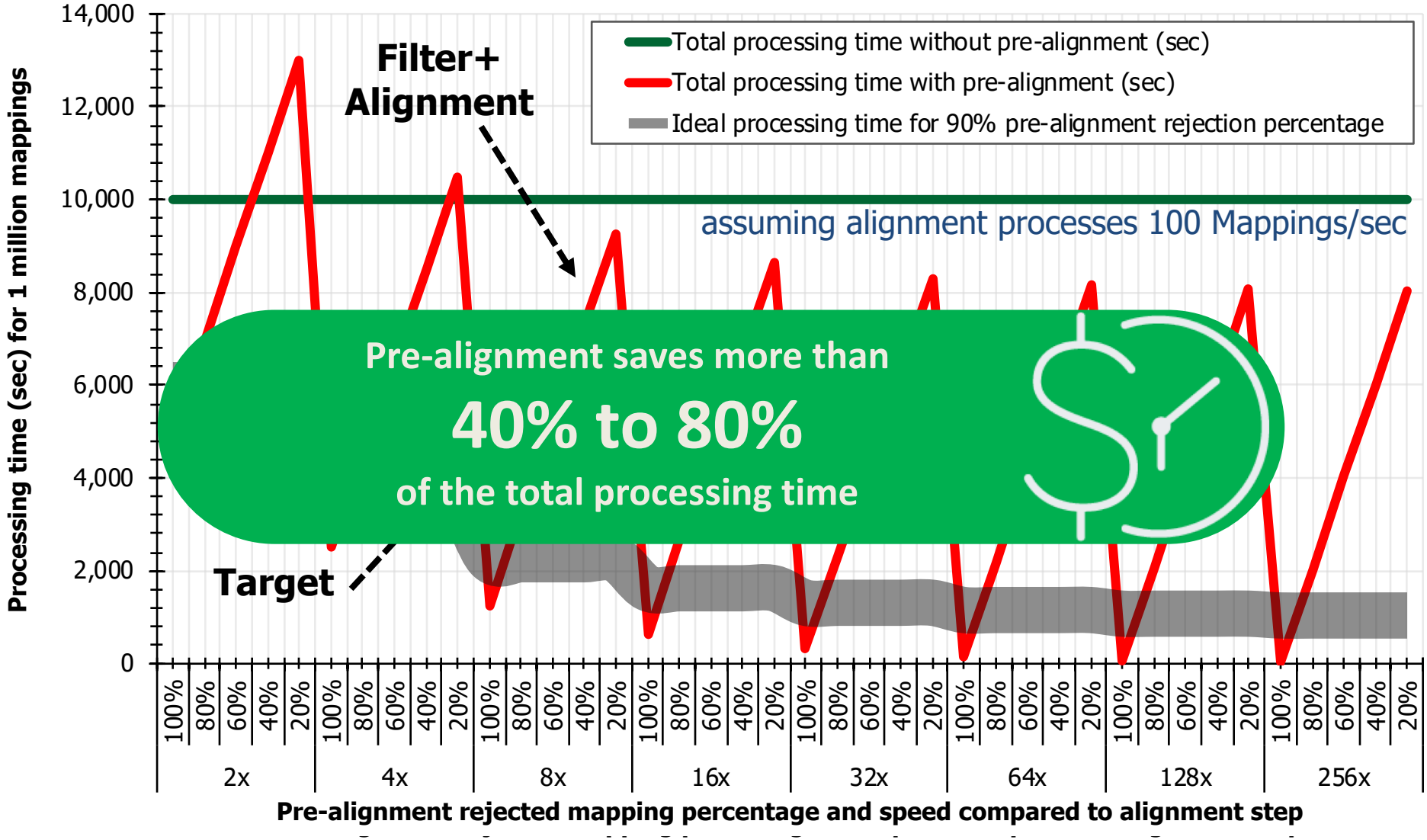
RECOMB 2021
PADOVA (ITALY) AUG 29 - SEP 1

SAFARI

ETH zürich

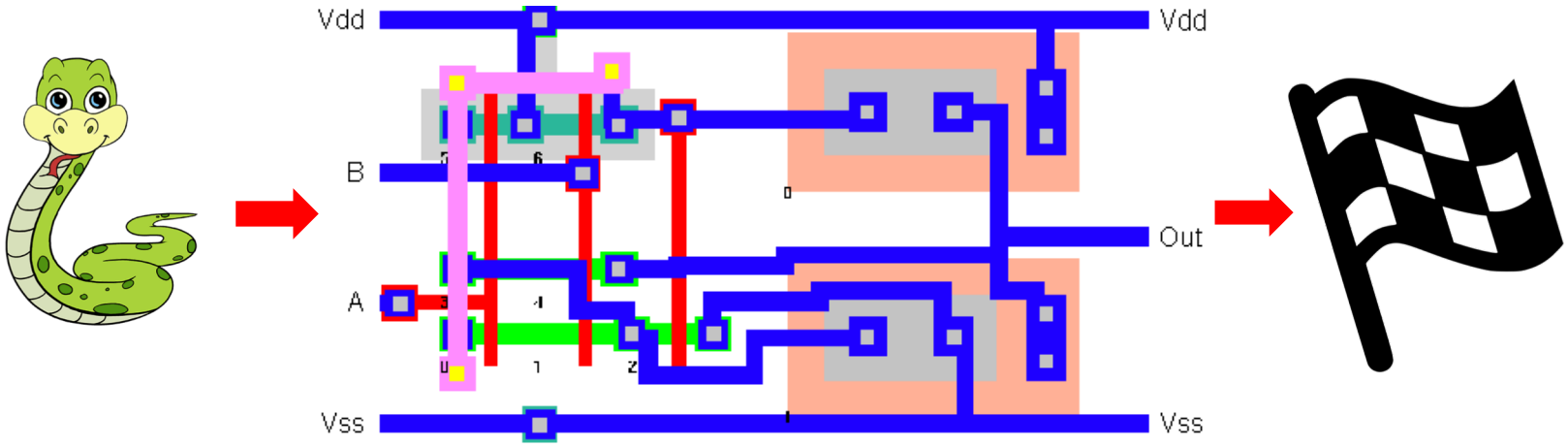
Backup Slides

The Effect of Pre-Alignment (Theoretically)



SneakySnake

- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches.
- **Key idea:**
  - Reduce the approximate string matching problem to the Single Net Routing problem in VLSI chip layout.



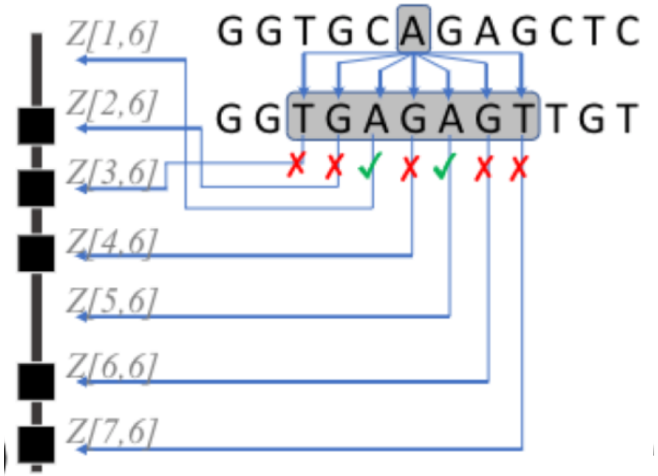
VLSI chip layout

SneakySnake

- **Key observation:**
  - Correct alignment is a sequence of non-overlapping long matches.
- **Key idea:**
  - Reduce the approximate string matching problem to the Single Net Routing problem in VLSI chip layout.
- **Key result:**
  - SneakySnake is up to four orders of magnitude more accurate than Shouji (Bioinformatics'19) and GateKeeper (Bioinformatics'17).
  - SneakySnake greatly accelerates state-of-the-art CPU sequence aligners, Edlib (Bioinformatics'17) and Parasail (BMC Bioinformatics'16)
    - by up to 37.7× and 43.9× (>12× on average), on CPUs
    - by up to 413× and 689× (>400× on average) *with FPGA acceleration*

SneakySnake Walkthrough

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



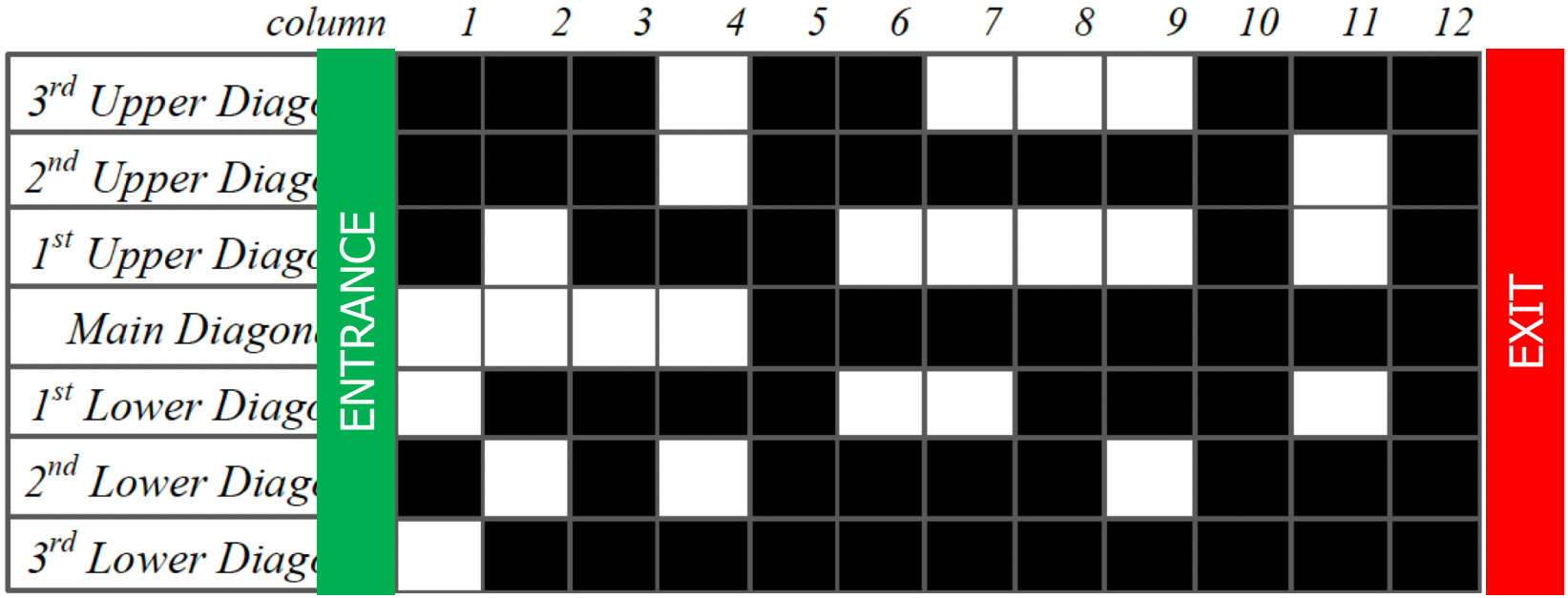
$E = 3$

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

SneakySnake Walkthrough

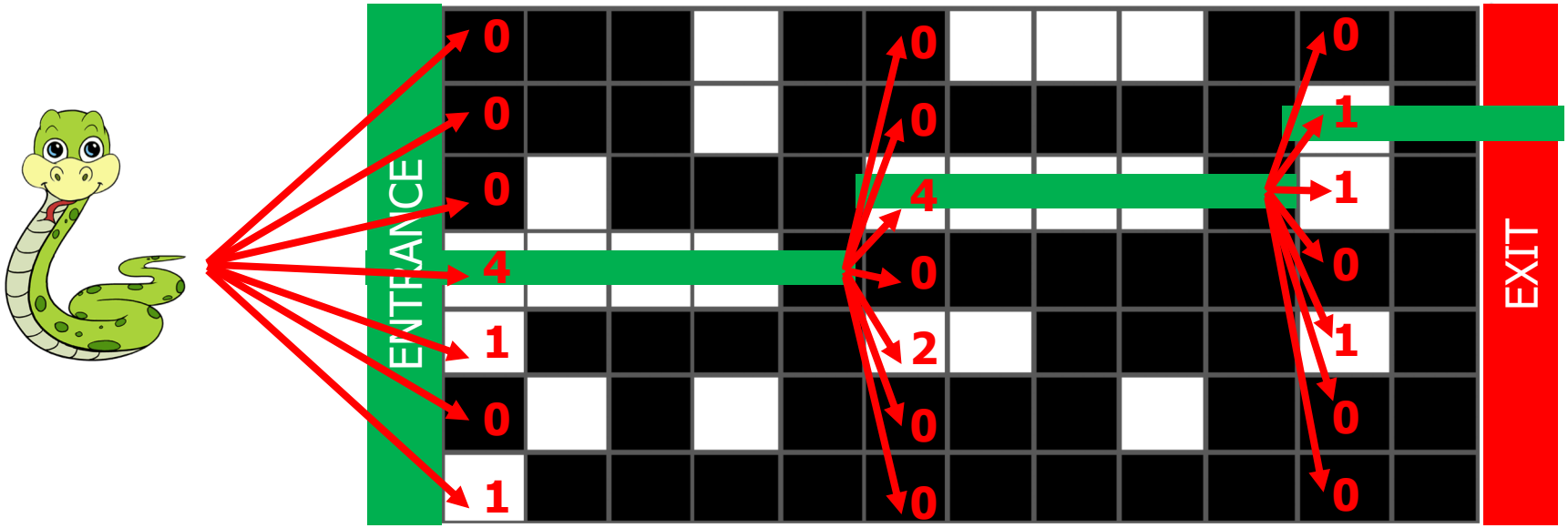
Building Neighborhood Map | Finding the Optimal Routing Path | Examining the Snake Survival

$E = 3$



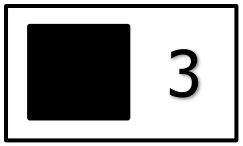
SneakySnake Walkthrough

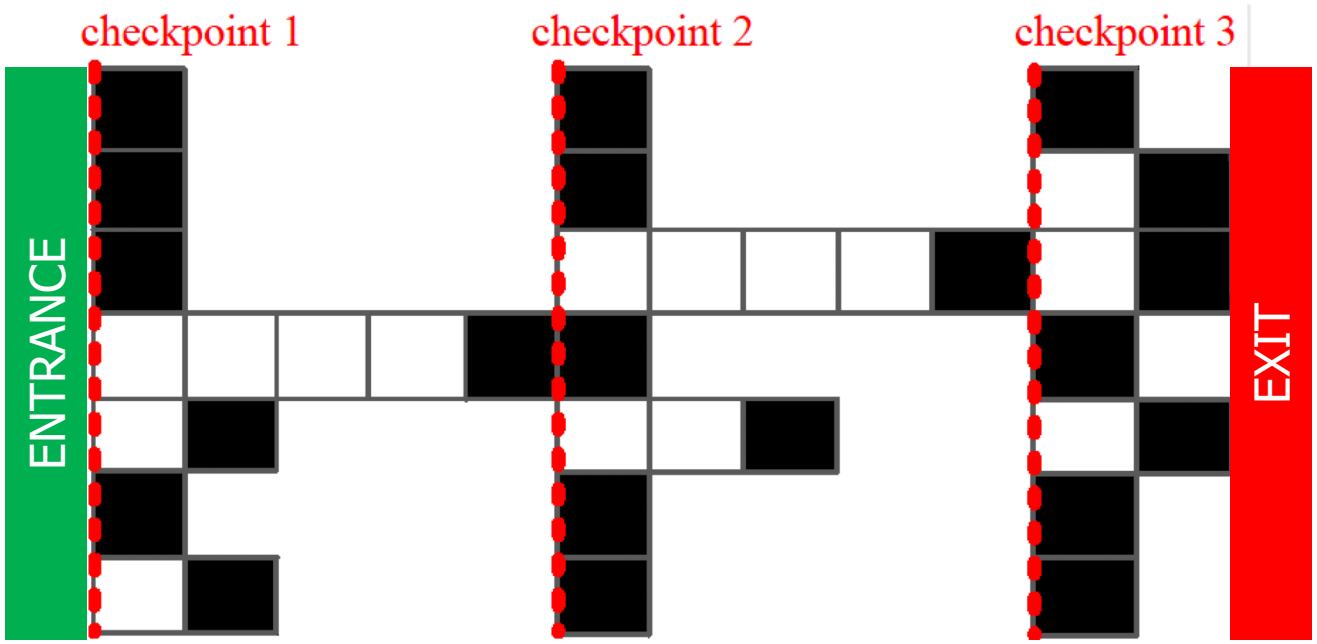
Building Neighborhood Map | Finding the Optimal Routing Path | Examining the Snake Survival



SneakySnake Walkthrough

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



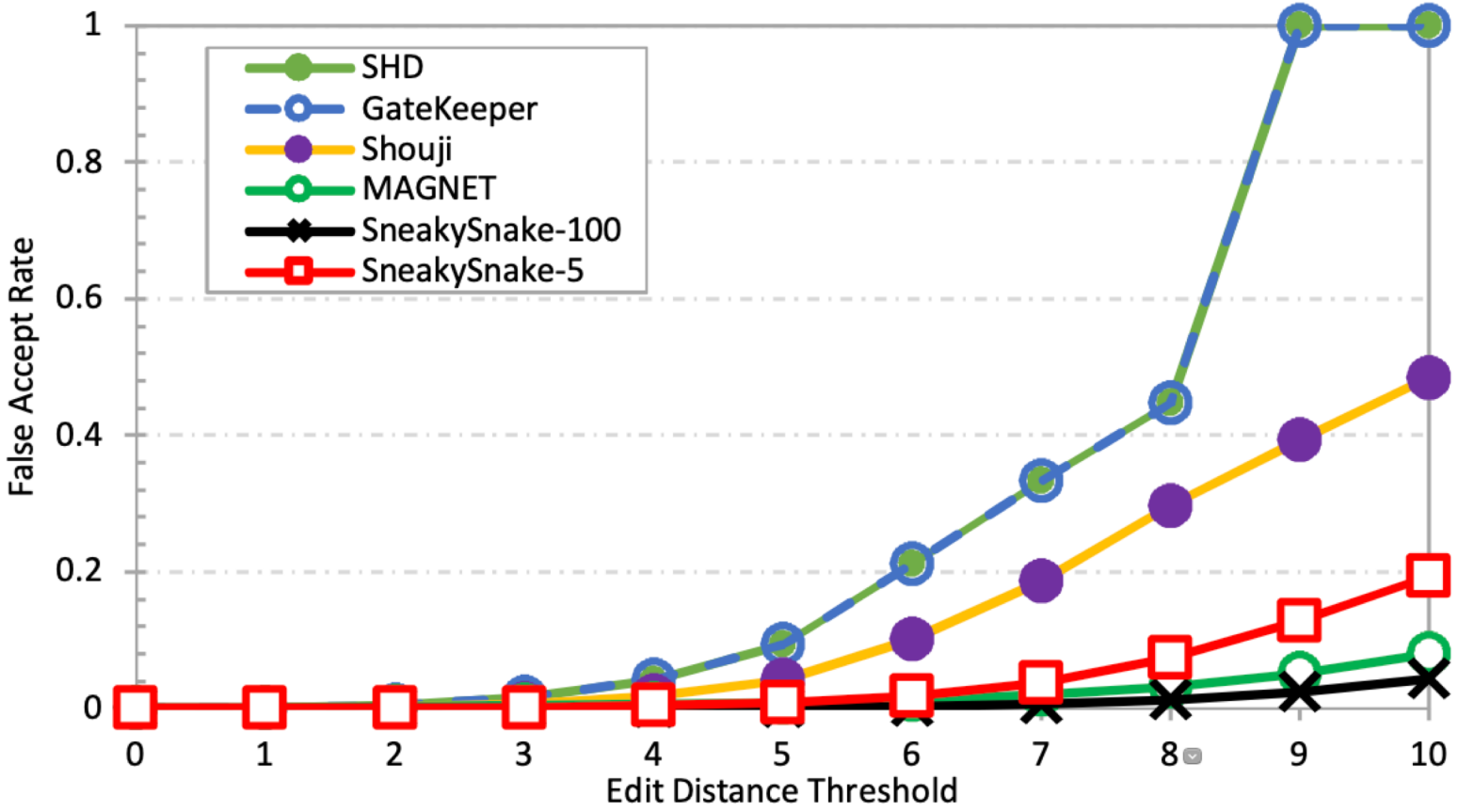


FPGA Resource Analysis

- 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

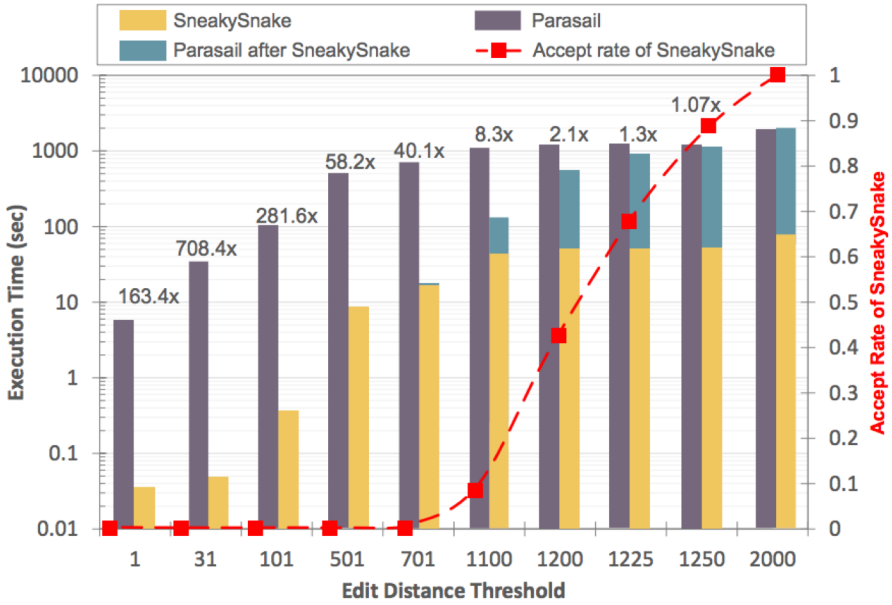
Filtering Accuracy



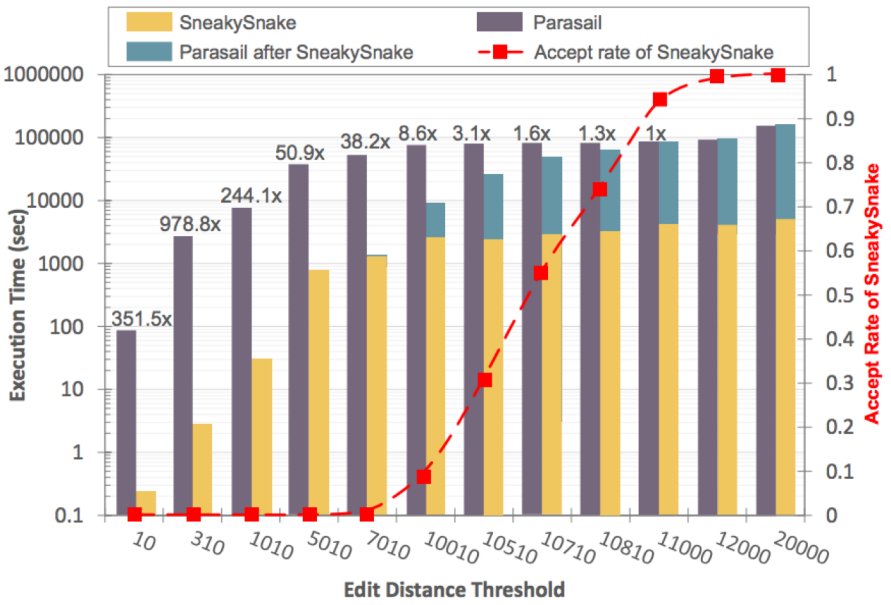
Alser, "Accelerating the Understanding of Life's Code Through Better Algorithms and Hardware Design", *arXiv preprint arXiv:1910.03936,* 2019.

Long Read Mapping (SneakySnake vs Parasail)

10K bp reads **100K bp reads**



(a)



(b)

Fig. 10: The execution time of SneakySnake, Parasail, and SneakySnake integrated with Parasail using long sequences, (a) 10Kbp and (b) 100Kbp, and 40 CPU threads. The left y-axes of (a) and (b) are on a logarithmic scale. For each edit distance threshold value, we provide in the right y-axes of (a) and (b) the rate of accepted pairs (out of 100,000 pairs for 10Kbp and out of 74,687 pairs for 100Kbp) by SneakySnake that are passed to Parasail. We present the end-to-end speedup values obtained by integrating SneakySnake with Parasail.

Long Read Mapping (SneakySnake vs KSW2)

10K bp reads **100K bp reads**

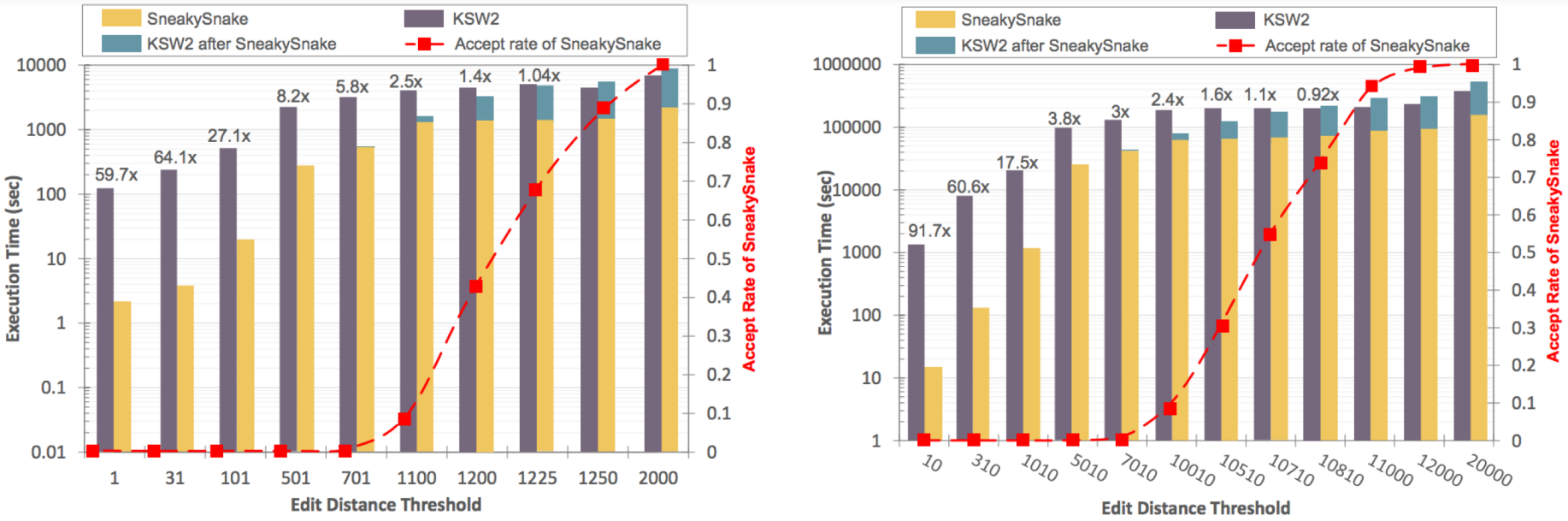
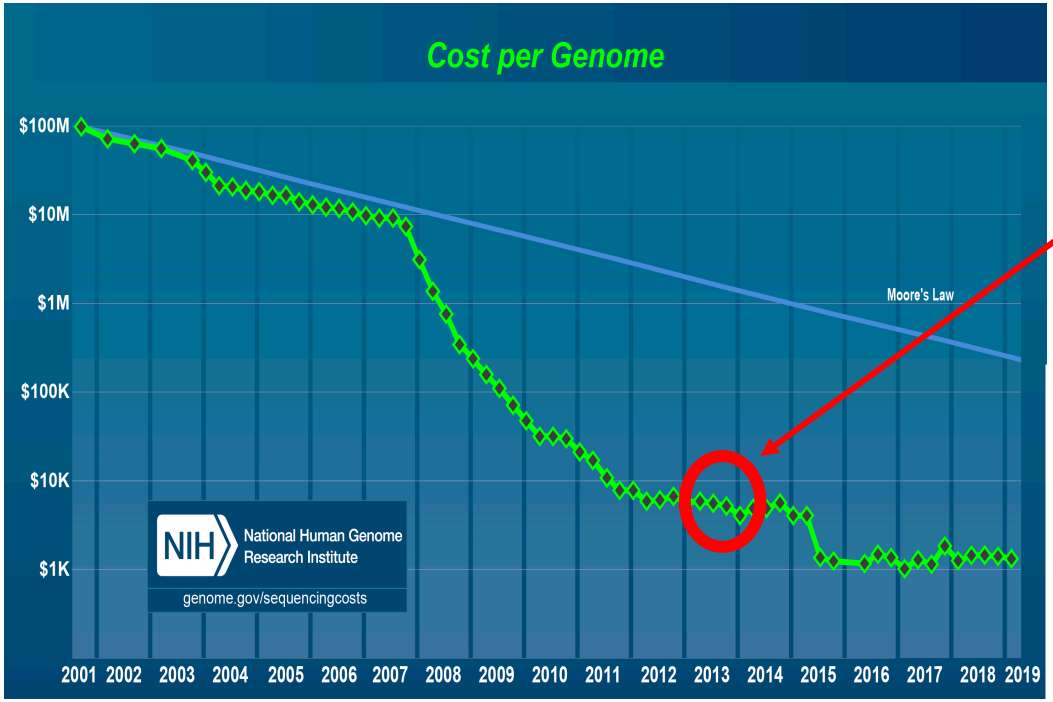


Fig. 11: The execution time of SneakySnake, KSW2, and SneakySnake integrated with KSW2 using long sequences, (a) 10Kbp and (b) 100Kbp, and a single CPU thread. The left y-axes of (a) and (b) are on a logarithmic scale. For each edit distance threshold value, we provide in the right y-axes of (a) and (b) the rate of accepted pairs (out of 100,000 pairs for 10Kbp and out of 74,687 pairs for 100Kbp) by SneakySnake that are passed to KSW2. We present the end-to-end speedup values obtained by integrating SneakySnake with KSW2.

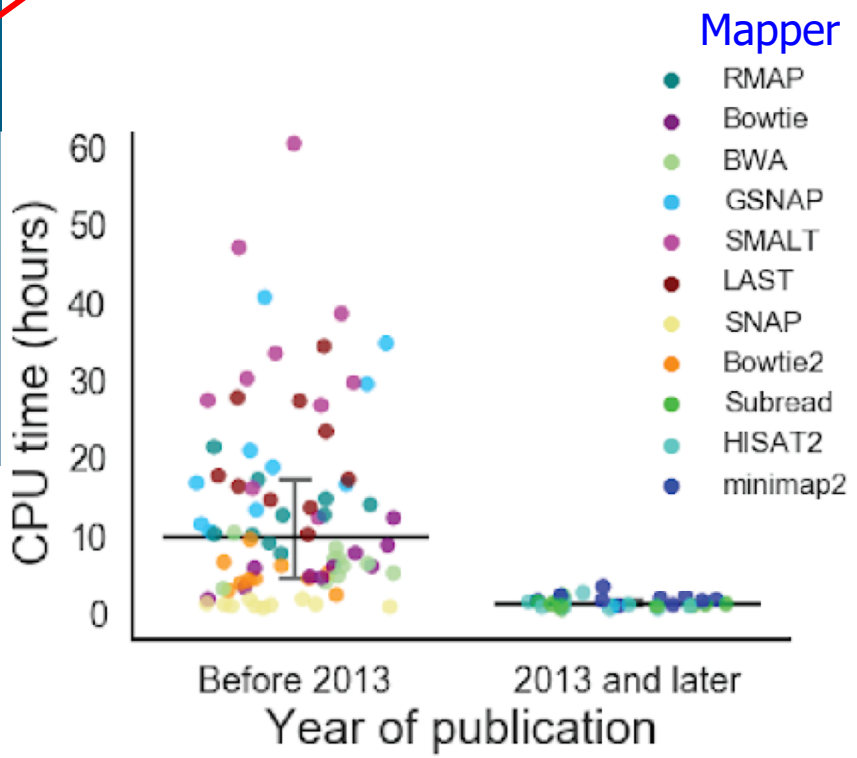
Ongoing Directions

- **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.

The Need for Speed

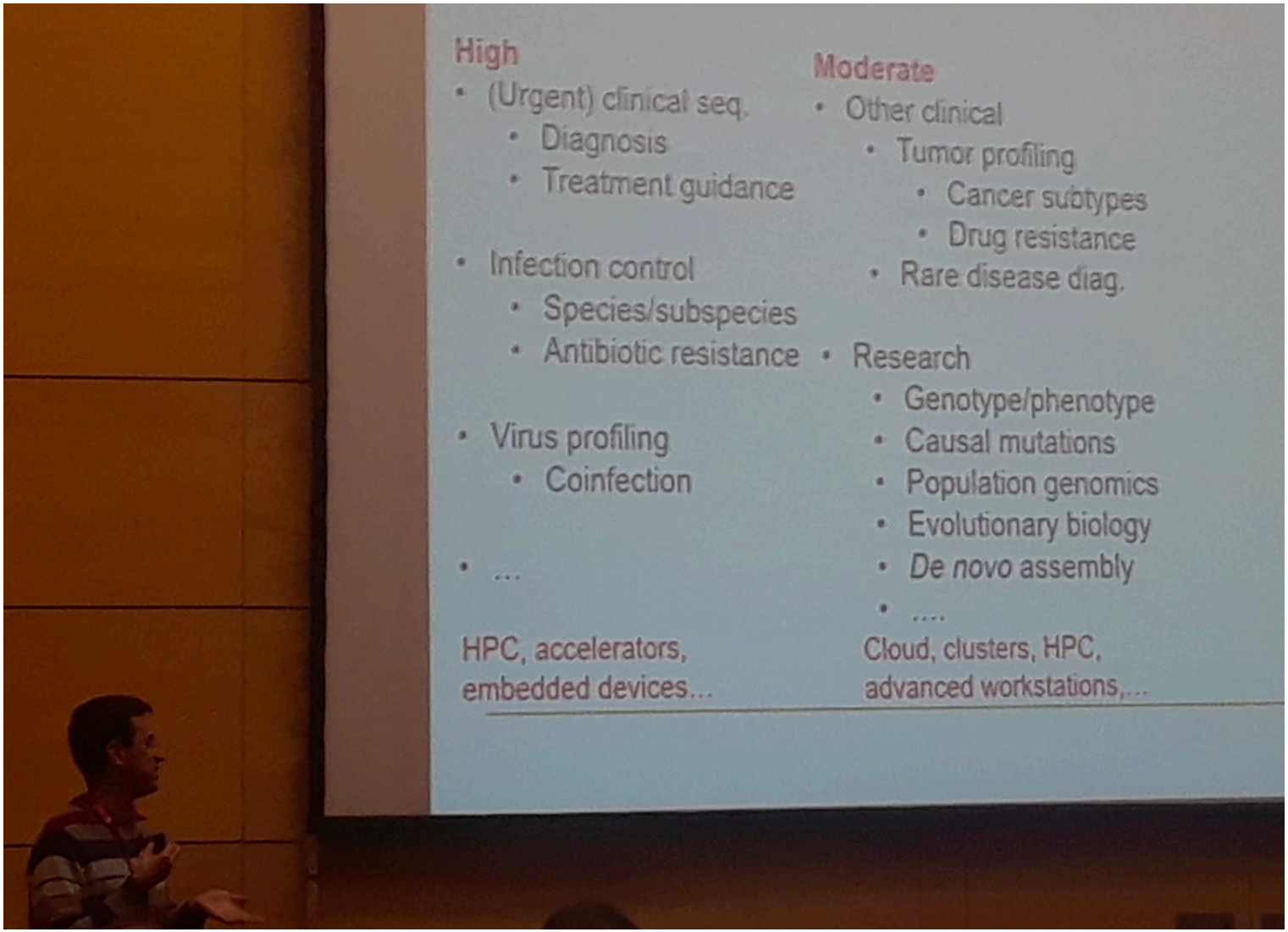


Did we realize the **need** for **faster** genome analysis?



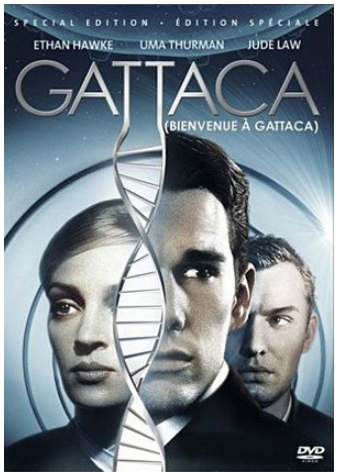
Alser+, "Technology dictates algorithms: Recent developments in read alignment", arXiv, 2020

The Need for Speed



Did we Achieve Our Goal?

- Our goal is to significantly reduce the time spent on calculating the optimal alignment in genome analysis from hours to mere seconds using both **new algorithms & hardware accelerators**, given limited computational resources (i.e., personal computer or small hardware).



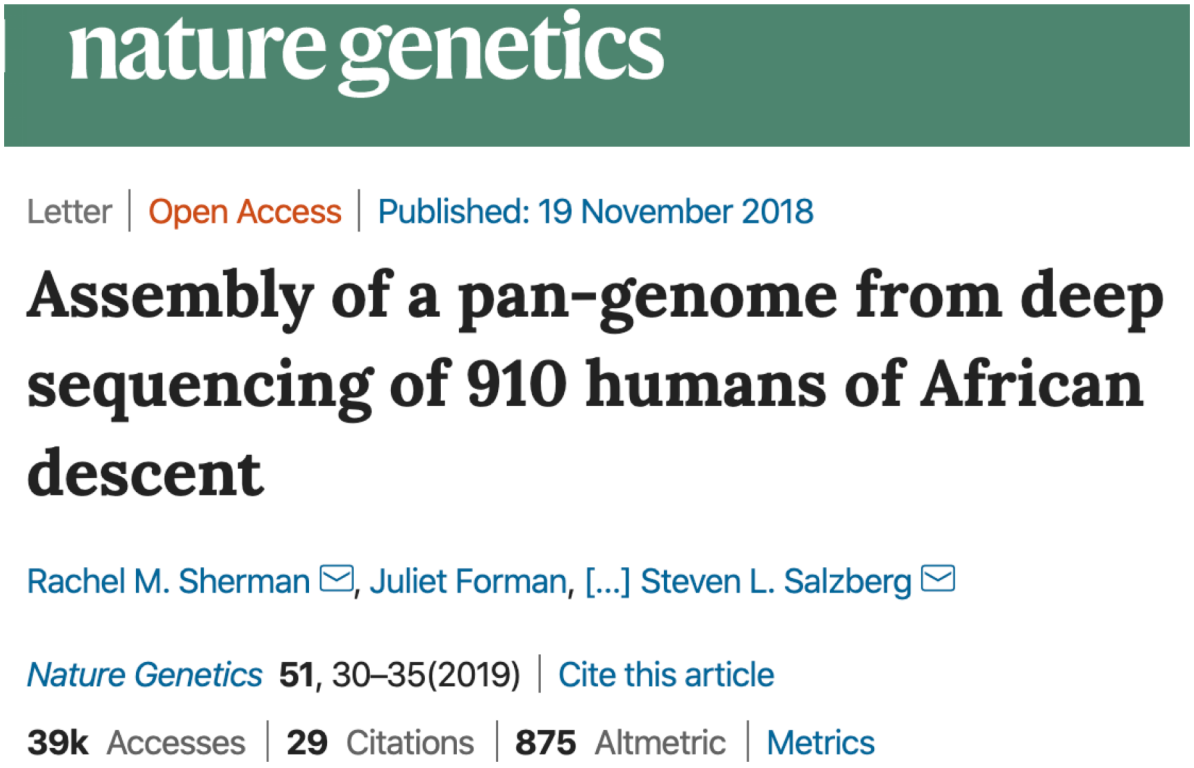
1997

2015

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*

Where is Read Mapping Going Next?



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

39k Accesses | **29** Citations | **875** Altmetric | Metrics

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

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