

# P&S Mobile Genomics

## Lecture 4: Read Mapping

Dr. Mohammed Alser

 @meals

ETH Zurich

Fall 2022

7 November 2022

# Agenda for Lecture 2

---

- What is Genome Analysis?
- What is Intelligent Genome Analysis?
- How we Analyze Genome?

# Agenda for Today

---

- What is Read Mapping?
- What Makes Read Mapper Slow?
- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

# Agenda for Today

---

- **What is Read Mapping?**
- What Makes Read Mapper Slow?
- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

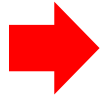
# Read Mapping

---

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



DNA Sample  
"chemical format"



Reads  
"text format"



Subject genome  
"text format"

# 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/>


# Cracking the 1<sup>st</sup> Human Genome Sequence

- 1990-2003:** The Human Genome Project (HGP) provides a complete and accurate sequence of all **DNA base pairs** that make up the human genome and finds 20,000 to 25,000 human genes.



A
C
G
T
 3.2 x10<sup>9</sup> bases


 13 years


 >3x10<sup>9</sup> \$

# Three Decades & Yet to be Complete!

## The complete sequence of a human genome

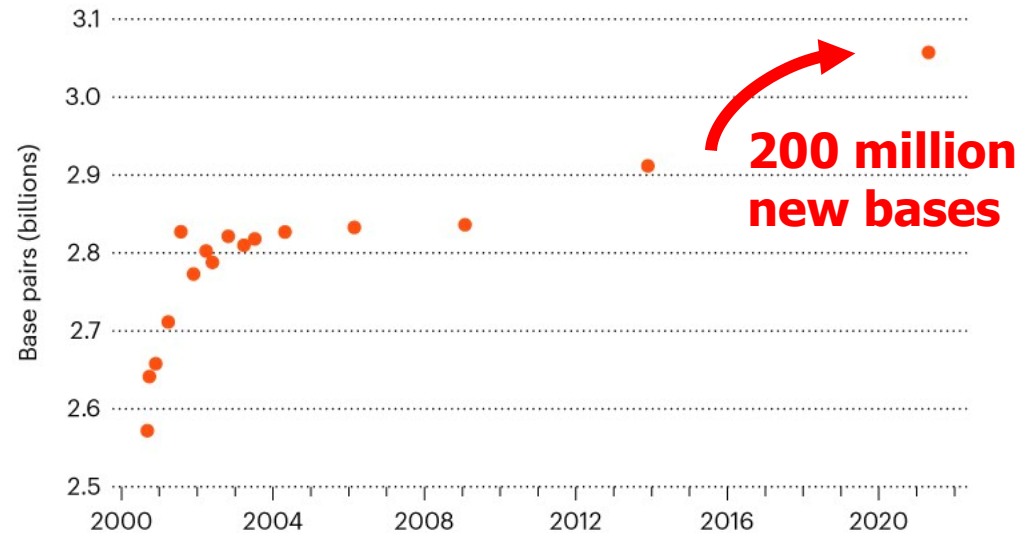
27 May 2021

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen Chen, Chia-Chia Wang, Chuan Li, David G. Klapper, Philip C. Dishuck, Richard Durbin, Tatiana Dvorkina, Arkarachai Fungtammasan, Erik Garrison, Patrick G. Gabrielle A. Hartley, Marina Haukness, Kerstin Howland, Erich D. Jarvis, Peter Kerpedjiev, Melanie Kirsche, M. Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Tamara Potapova, Evgeny I. Rogae, Jeffrey A. Rosenthal, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Tamura, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Frazer A. Wright, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Woollam, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Meaghan E. H. Bryant, Rachel J. O'Neill, Winston Timp, Justin M. Zook, Michael J. Adams, Adam M. Phillippy

doi: <https://doi.org/10.1101/2021.05.26.445798>

### COMPLETING THE HUMAN GENOME

Researchers have been filling in incompletely sequenced parts of the human reference genome for 20 years, and have now almost finished it, with 3.05 billion DNA base pairs.



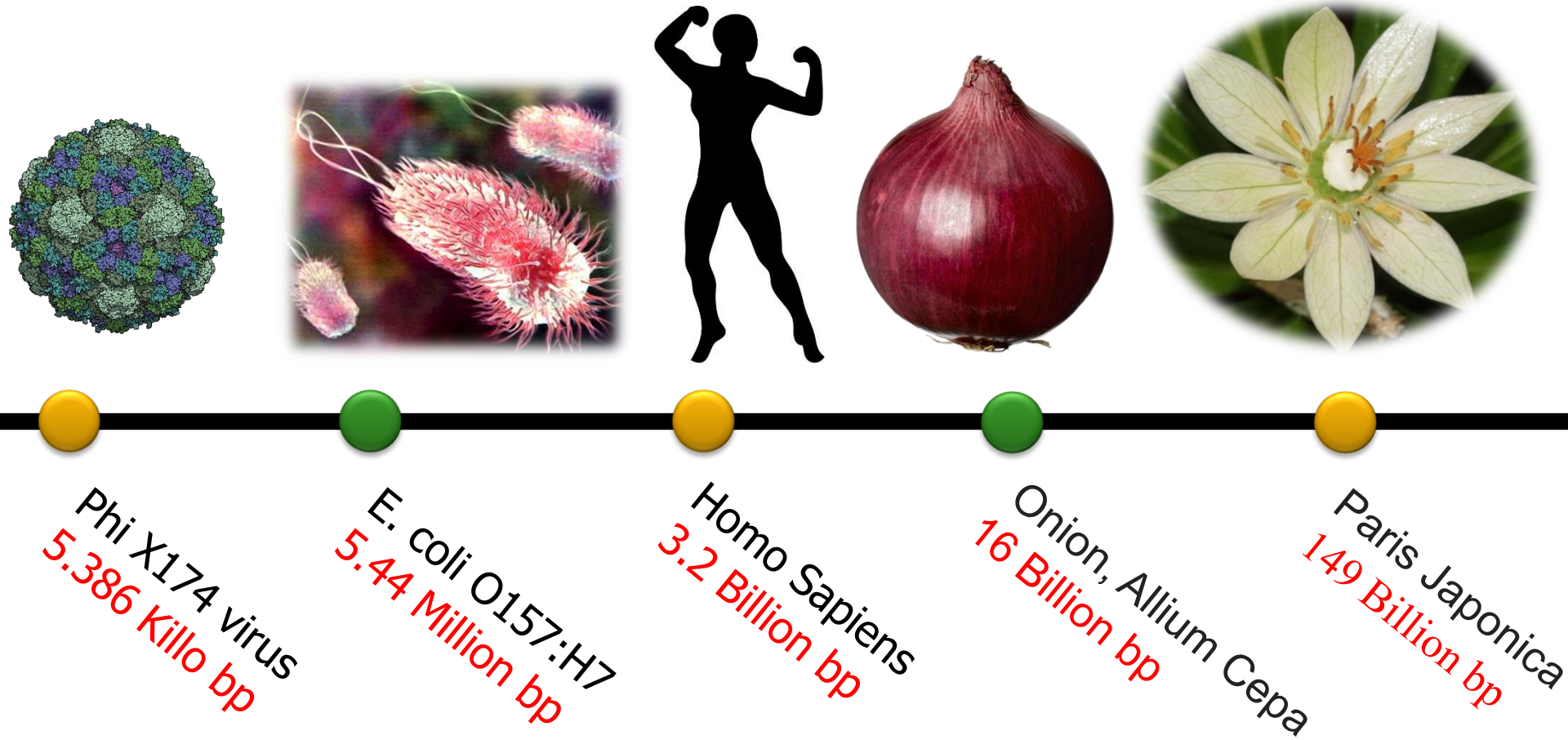
0.3% of sequence might still have errors. Includes X but not Y chromosome. Count excludes mitochondrial DNA.

©nature





# How Long is DNA?



# Obtaining .FASTQ Files

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



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
<a href="#">ERR240727</a>	4,093,747	818.7M	387.2Mb	2013-03-22

---

Let's learn  
how to map a read

# Read Mapping: A Brute Force Algorithm

---

Reference



Read

Very expensive!  
 $O(m^2kn)$

$m$ : read length

$k$ : no. of reads

$n$ : reference genome length

# Read Mapping in 111 pages!

In-depth analysis of 107 read mappers (1988-2020)

**Mohammed Alser**, Jeremy Rotman, Dhriti Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyung 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<sup>1,2,3†</sup>, Jeremy Rotman<sup>4†</sup>, Dhriti Deshpande<sup>5</sup>, Kodi Taraszka<sup>4</sup>, Huwenbo Shi<sup>6,7</sup>, Pelin Icer Baykal<sup>8</sup>, Harry Taegyung Yang<sup>4,9</sup>, Victor Xue<sup>4</sup>, Sergey Knyazev<sup>8</sup>, Benjamin D. Singer<sup>10,11,12</sup>, Brunilda Balliu<sup>13</sup>, David Koslicki<sup>14,15,16</sup>, Pavel Skums<sup>8</sup>, Alex Zelikovsky<sup>8,17</sup>, Can Alkan<sup>2,18</sup>, Onur Mutlu<sup>1,2,3†</sup> and Serghei Mangul<sup>5\*†</sup> 

# 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

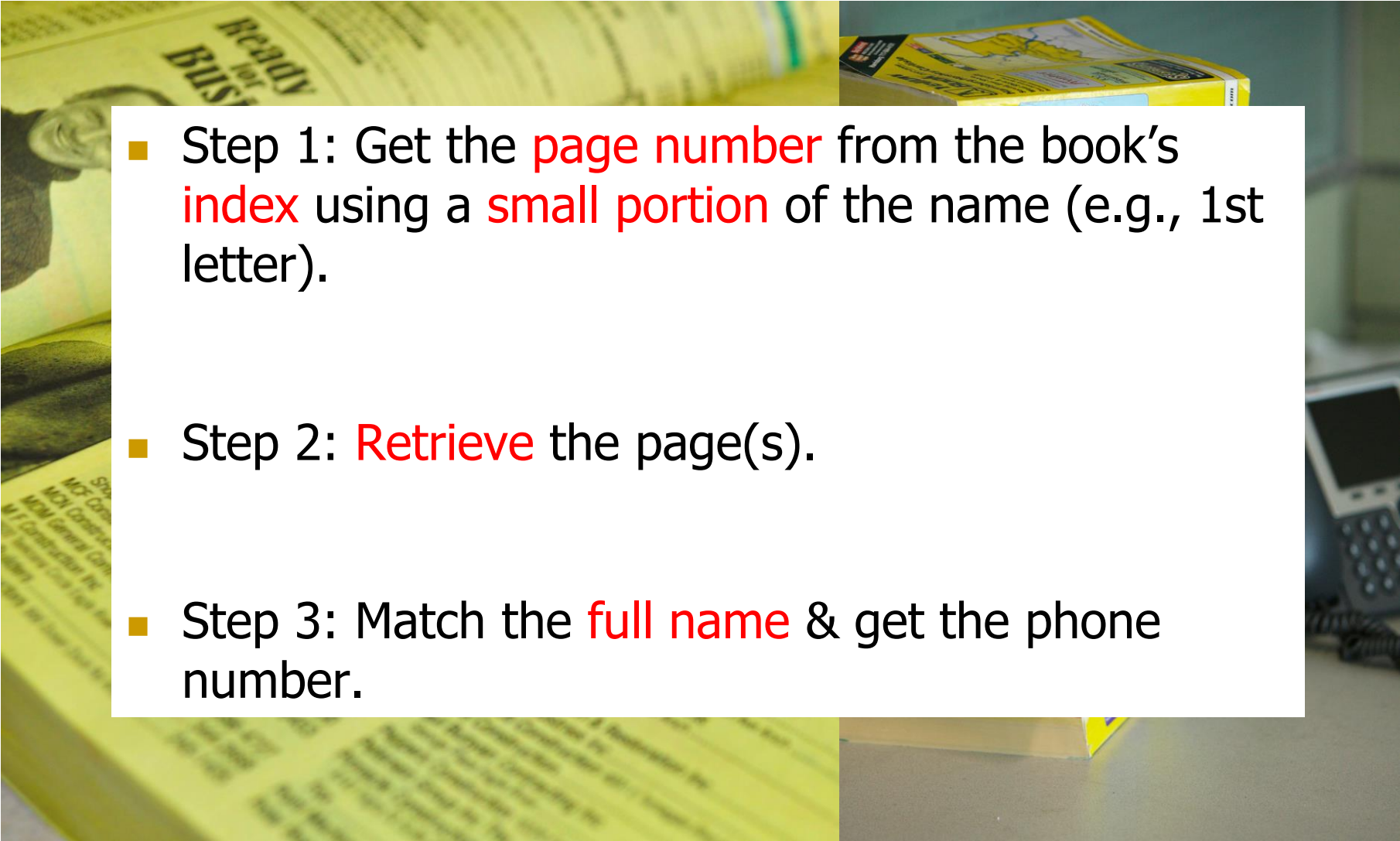
Buen hilo de repaso sobre la evolución de los algoritmos de alineamiento de secuencias a medida que ha mejorado la tecnología de secuenciación

---

Mapping a read is  
similar to querying  
the yellow pages!



# Similar to Searching Yellow Pages!

- 
- Step 1: Get the **page number** from the book's **index** using a **small portion** of the name (e.g., 1st letter).
  - Step 2: **Retrieve** the page(s).
  - Step 3: Match the **full name** & get the phone number.

# Matching Each Read with Reference Genome

---

## .FASTA file:

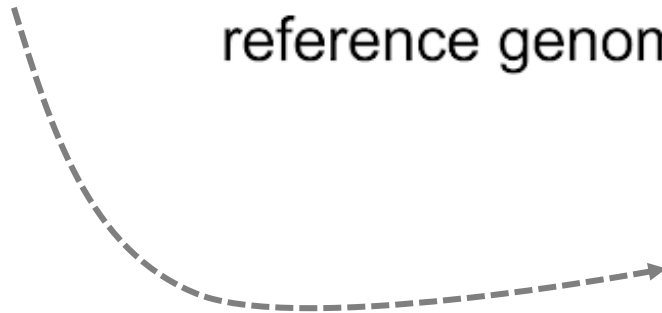
```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCC[redacted]TCATTGACATTTAAACTCTGGGGCAGG[redacted]GAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCC[redacted]CCCCGGCCCGGCTCGGGGCCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCCGCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TCCGAGTGT[redacted]CAAAGTAGCA[redacted]CTCCTA[redacted]TCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTC[redacted]CGCTTGGGAAAG
TCCGTACCCGCGCCT[redacted]AAAGACACCCTGCCGCGGGTTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

## .FASTQ file:

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
T[redacted]AATAAATCT[redacted]TTAGATN[redacted]NNNNNNNNTAG
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcfffffd`feed]` ]_Ba_^__[YBBBBBBBBBRTT
```

# Step 1: Indexing the Reference Genome

---



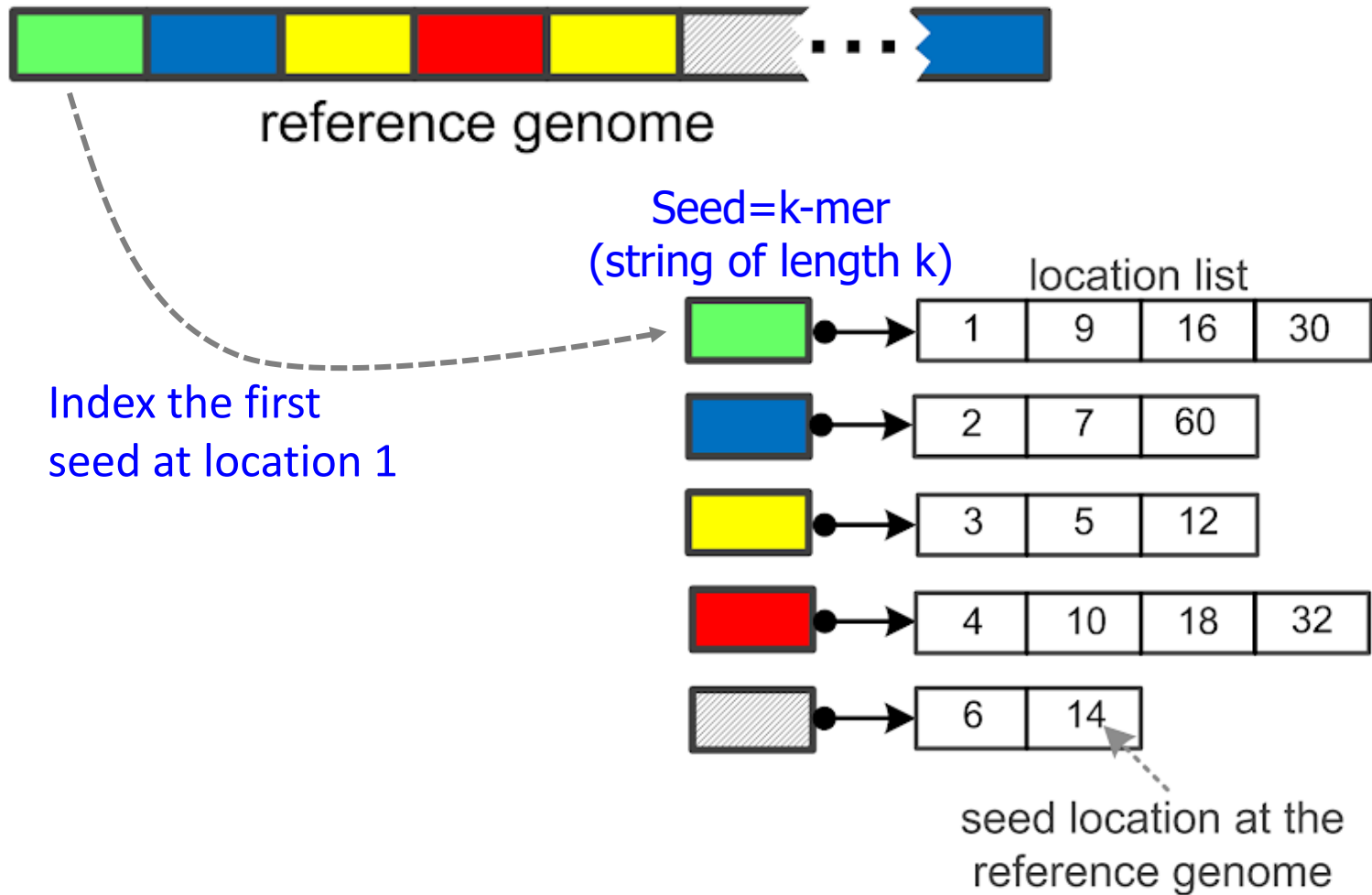
# Popular Indexing Technique

---

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

Alser+, "[Technology dictates algorithms: Recent developments in read alignment](#)",  
Genome Biology, 2021

# Step 1: Indexing the Reference Genome



# Genome Index Properties

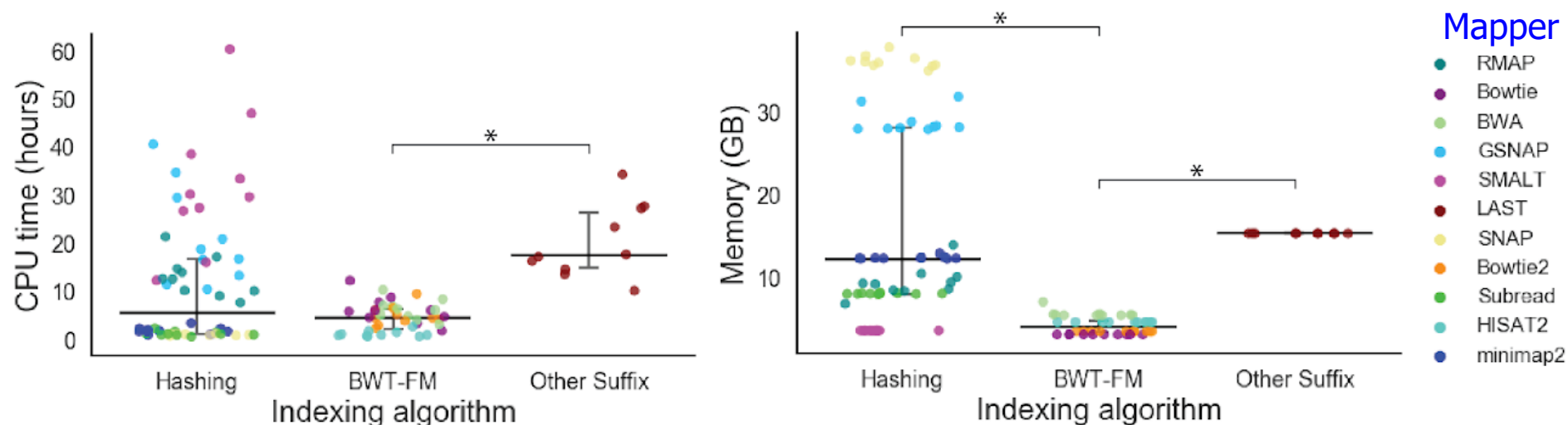
---

- The index is built **only once** for each reference.
- **Seeds** can be overlapping, non-overlapping, spaced, adjacent, non-adjacent, minimizers, compressed, ...

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

\*Human genome = 3.2 GB

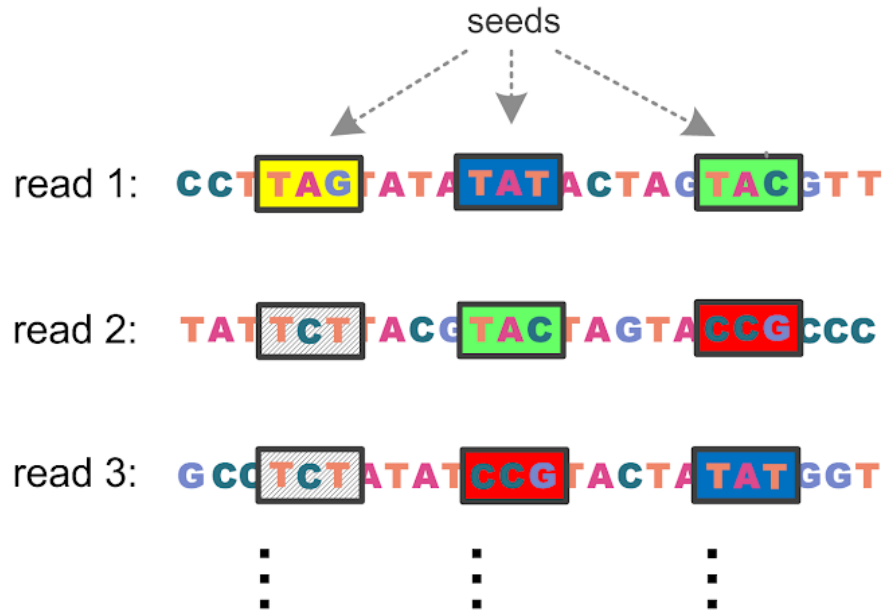
# Performance of Human Genome Indexing



Alser+, "[Technology dictates algorithms: Recent developments in read alignment](#)",  
Genome Biology, 2021

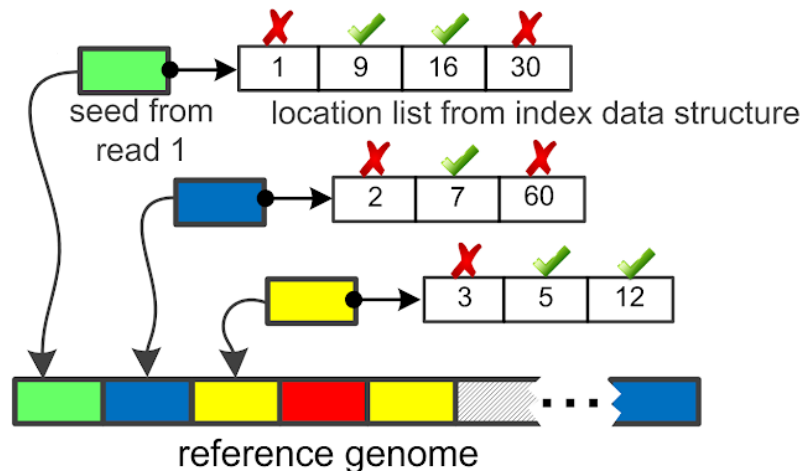
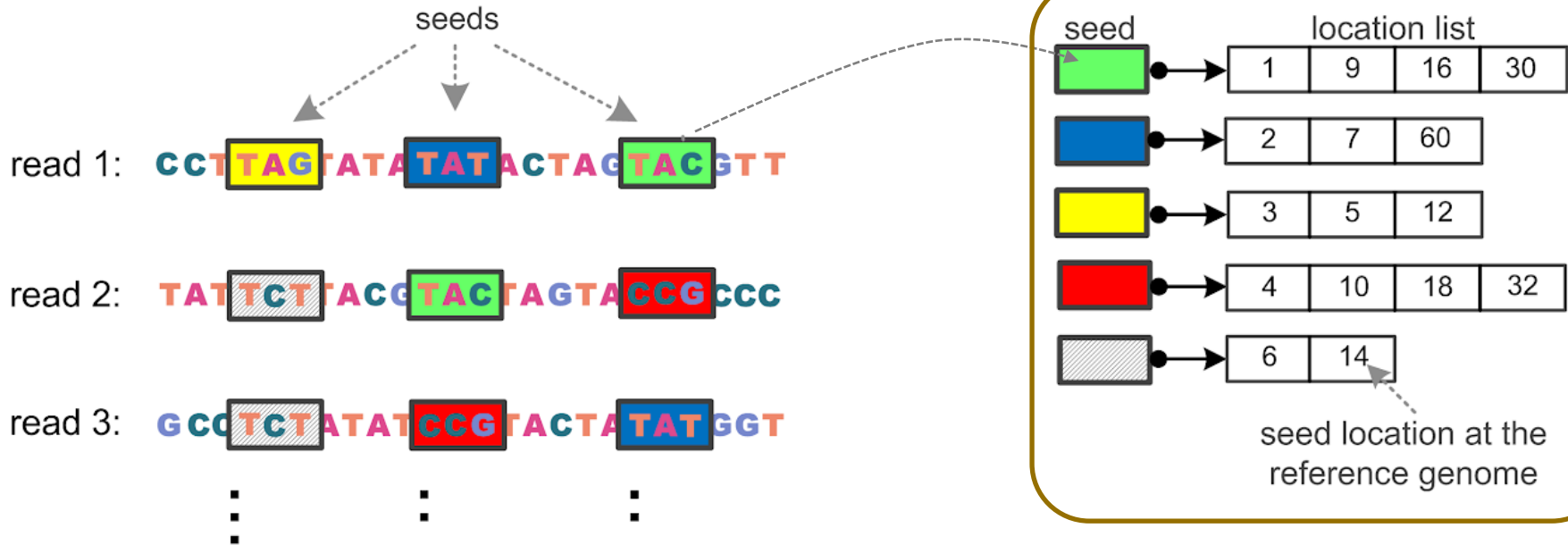
# Step 2: Query the Index Using Read Seeds

---





# Step 2: Query the Index Using Read Seeds



# Step 2: Query the Index Using Read Seeds

seeds

seed

location list

1	9	16	30
---	---	----	----

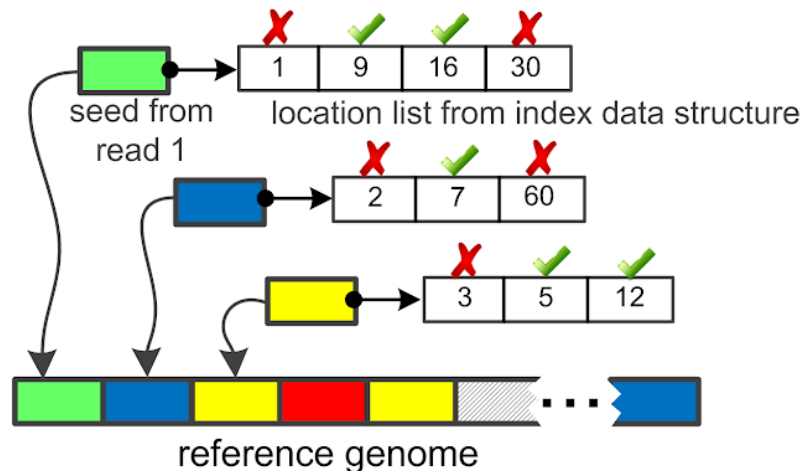
We can query the Hash table with substrings from reads to quickly find a list of possible mapping locations

read

read

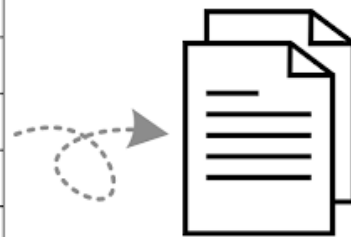
read 3: **GCC** **TCT** **TATAT** **CCG** **FACTA** **TAT** **GGT**  
⋮ ⋮ ⋮

seed location at the reference genome



# Step 3: Sequence Alignment (Verification)

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



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

# Step 3: Sequence Alignment (Verification)

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

organization x operation

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

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

Edit distance = 7

organization x translation

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

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

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

Edit distance = 4

match
deletion
insertion
mismatch

# Popular Algorithms for Sequence Alignment

---

**Smith-Waterman** remains  
the **most popular** algorithm  
since 1988

**Hamming distance** is  
the **second most popular** technique  
since 2008

# An Example of Hash Table Based Mappers

---

- + Guaranteed to find *all* mappings → very sensitive
- + Can tolerate up to *e* errors

nature  
genetics

<https://github.com/BilkentCompGen/mrfast>

---

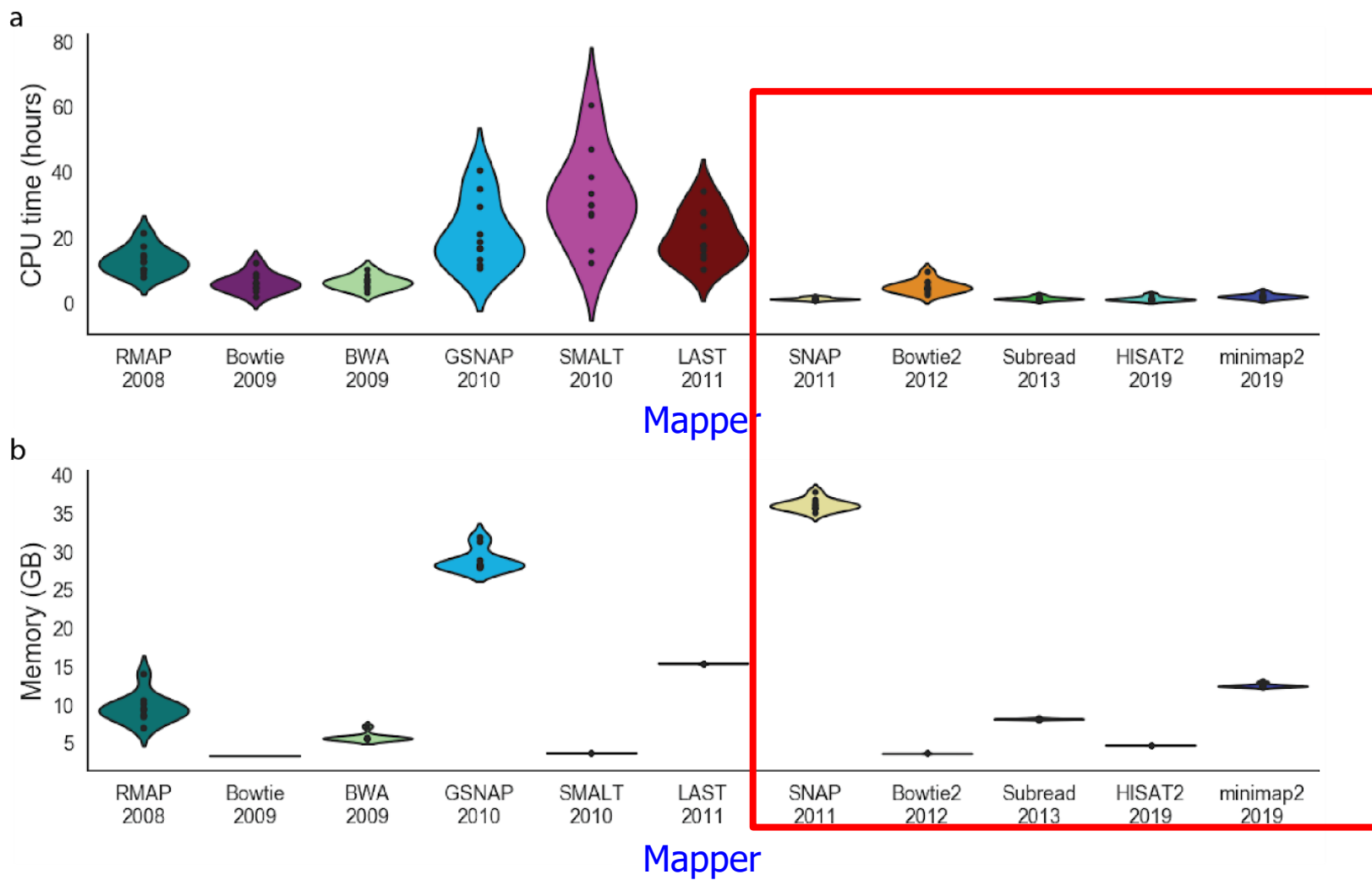
## Personalized copy number and segmental duplication maps using next-generation sequencing

Can Alkan<sup>1,2</sup>, Jeffrey M Kidd<sup>1</sup>, Tomas Marques-Bonet<sup>1,3</sup>, Gozde Aksay<sup>1</sup>, Francesca Antonacci<sup>1</sup>, Fereydoun Hormozdiari<sup>4</sup>, Jacob O Kitzman<sup>1</sup>, Carl Baker<sup>1</sup>, Maika Malig<sup>1</sup>, Onur Mutlu<sup>5</sup>, S Cenk Sahinalp<sup>4</sup>, Richard A Gibbs<sup>6</sup> & Evan E Eichler<sup>1,2</sup>

Alkan+, "[Personalized copy number and segmental duplication maps using next-generation sequencing](#)", Nature Genetics 2009.

---

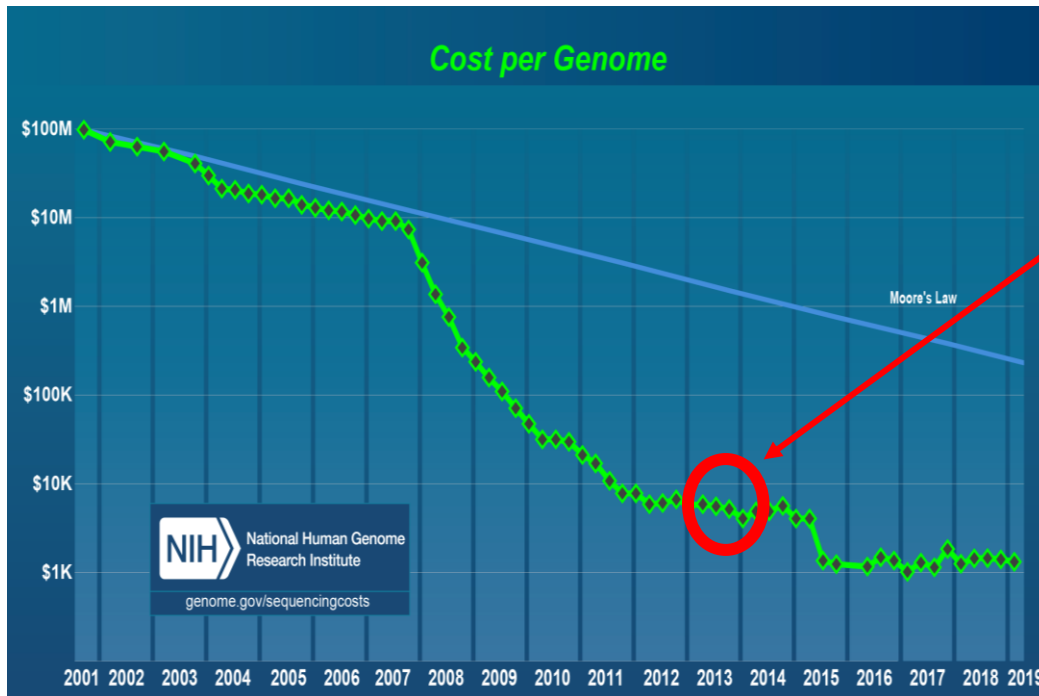
# Performance of Read Mapping



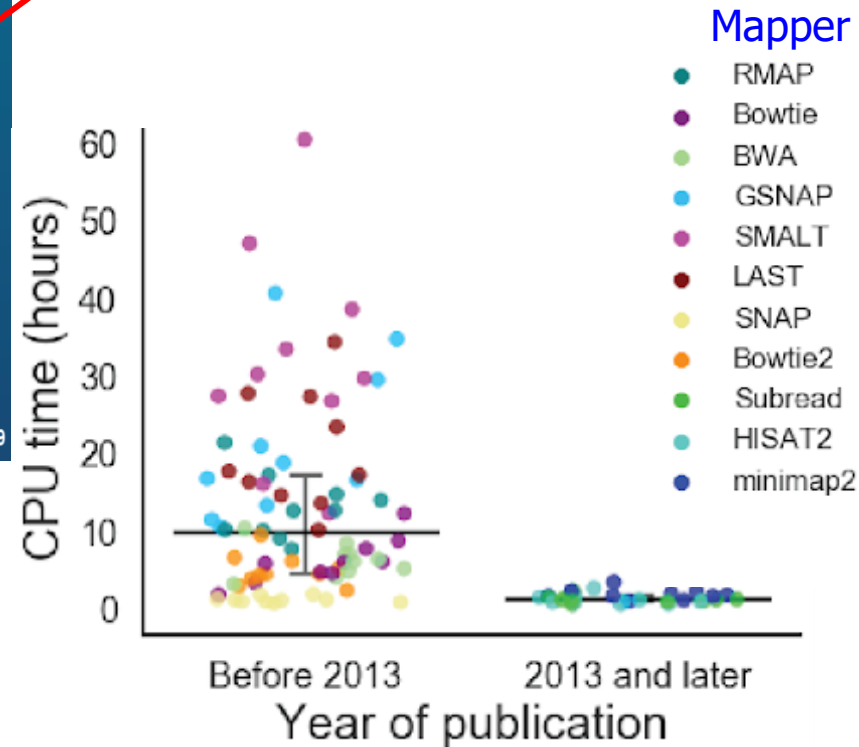
Alser+, "[Technology dictates algorithms: Recent developments in read alignment](#)",

Genome Biology, 2021

# The Need for Speed



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



Alser+, "[Technology dictates algorithms: Recent developments in read alignment](#)",  
Genome Biology, 2021



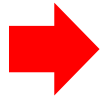
# Read Mapping

---

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



DNA Sample  
"chemical format"



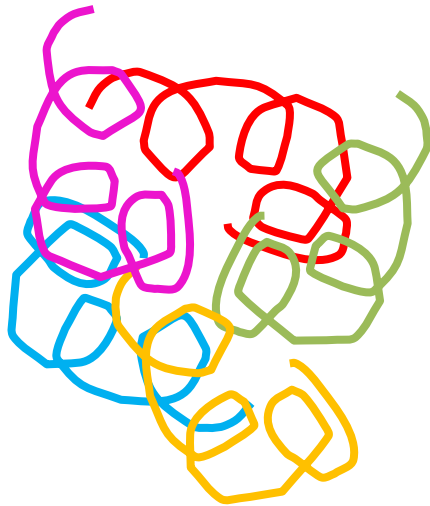
Reads  
"text format"



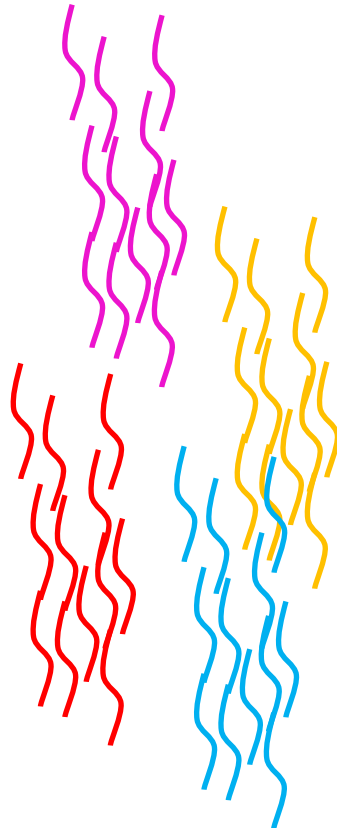
Subject genome  
"text format"

# Metagenomics Analysis

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

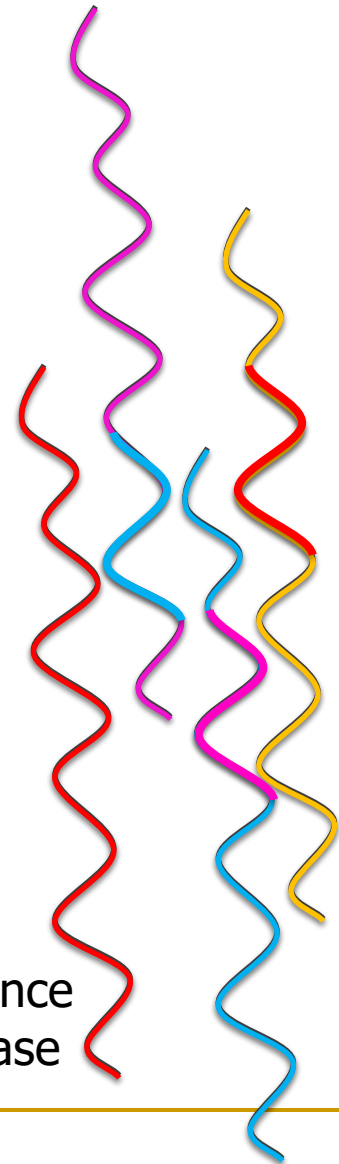


genetic material recovered directly from environmental samples



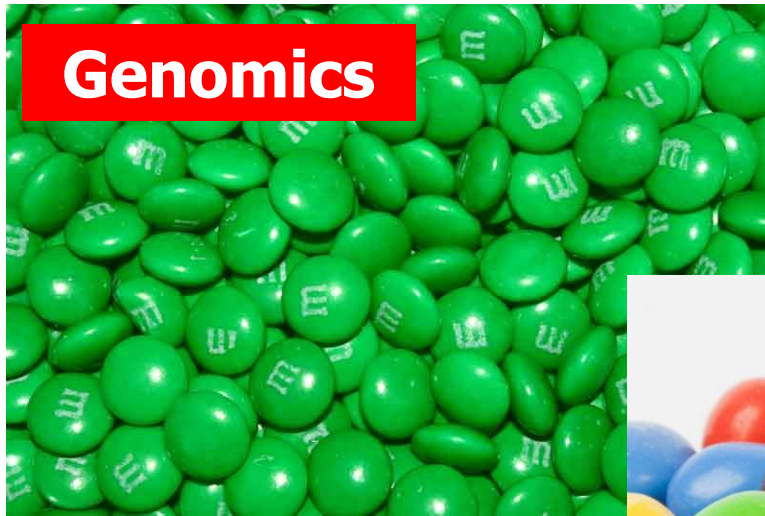
Reads "text format"

Reference Database



# Genomics vs. Metagenomics

---



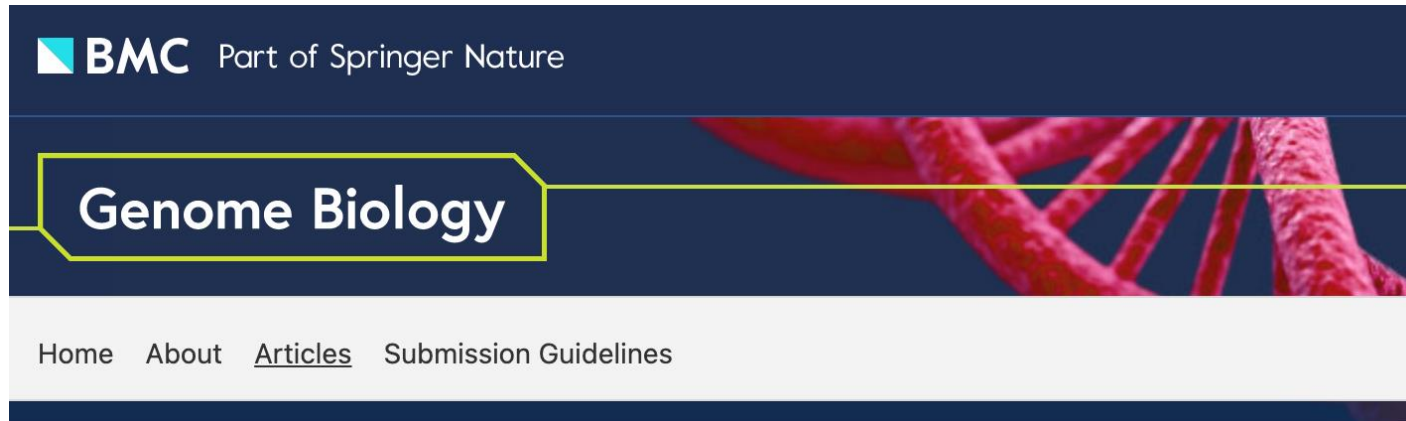
# More on Metagenomic Profiling: Metalign

Nathan LaPierre, Mohammed Alser, Eleazar Eskin, David Koslicki, Serghei Mangu  
“[Metalign: efficient alignment-based metagenomic profiling via containment min hash](#)”

**Genome Biology**, September 2020.

[[Talk Video](#) (7 minutes) at ISMB 2020]

[[Source code](#)]



Software | [Open Access](#) | [Published: 10 September 2020](#)

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

[Nathan LaPierre](#) ✉, [Mohammed Alser](#), [Eleazar Eskin](#), [David Koslicki](#) ✉ & [Serghei Mangu](#) ✉

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

# Check Also CAMI II Paper

---

F. Meyer, A. Fritz, Z.L. Deng, D. Koslicki, A. Gurevich, G. Robertson, Mohammed Alser, and others

[“Critical Assessment of Metagenome Interpretation - the second round of challenges”](#)

bioRxiv, 2021

[\[Source Code\]](#)

## **Critical Assessment of Metagenome Interpretation - the second round of challenges**

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

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

# Check Also MiCoP

Nathan LaPierre, Serghei Mangul, Mohammed Alser, Igor Mandric, Nicholas C. Wu, David Koslicki & Eleazar Eskin

[“MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples”](#)

**BMC Genomics**, June 2019.

[\[Source code\]](#)

 **BMC** Part of Springer Nature

**BMC Genomics**

Research | [Open Access](#) | Published: 06 June 2019

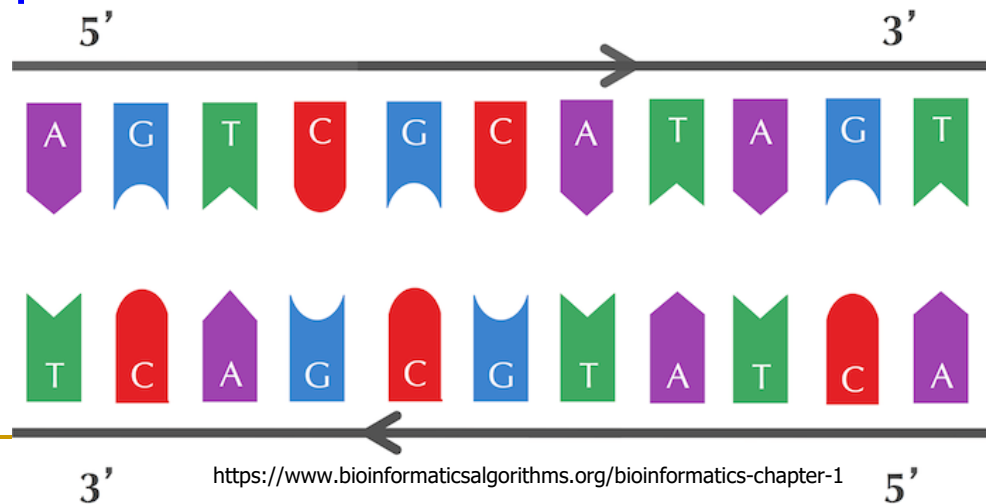
## MiCoP: microbial community profiling method for detecting viral and fungal organisms in metagenomic samples

[Nathan LaPierre](#), [Serghei Mangul](#) , [Mohammed Alser](#), [Igor Mandric](#), [Nicholas C. Wu](#), [David Koslicki](#) & [Eleazar Eskin](#)

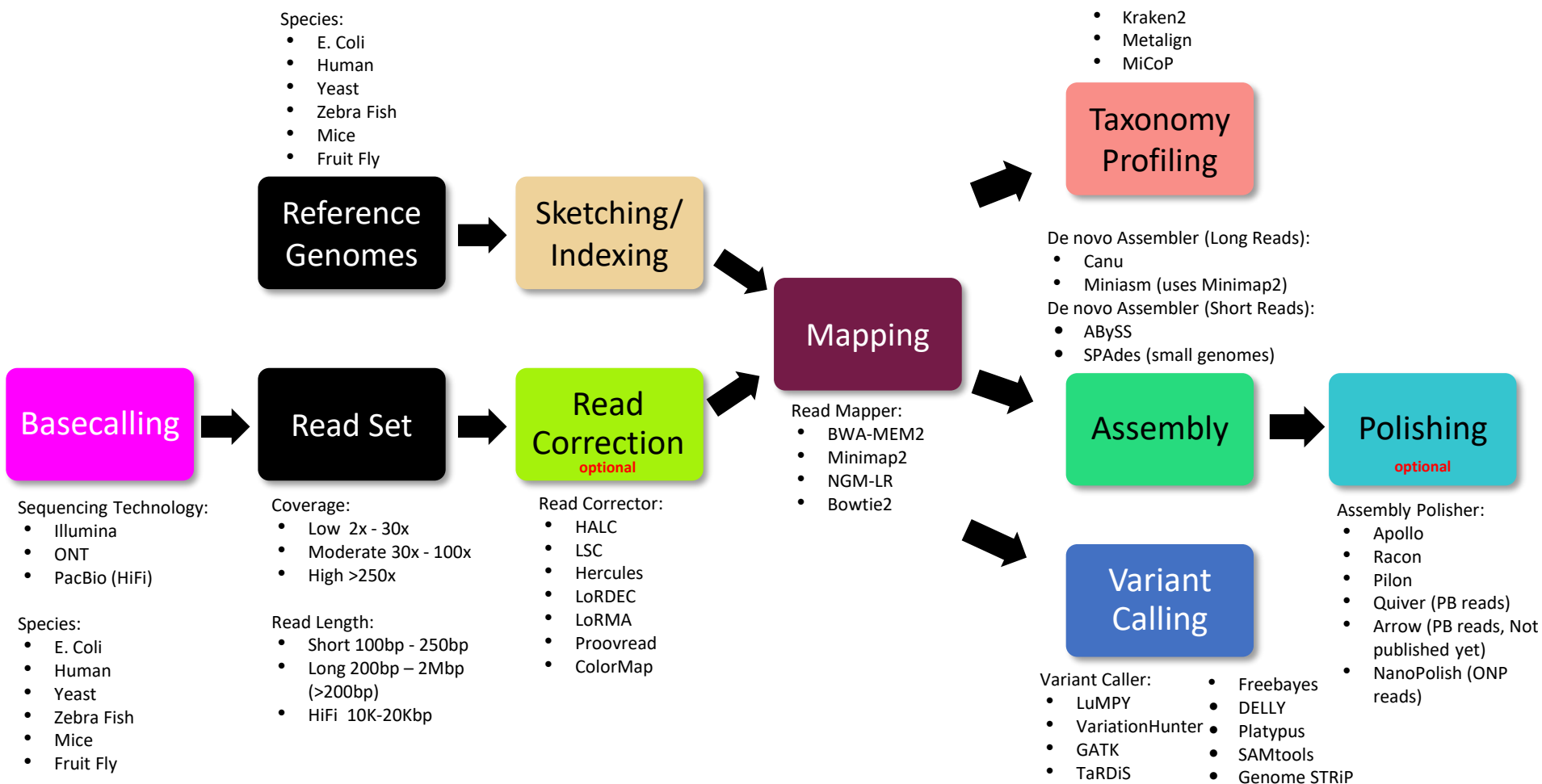
[BMC Genomics](#) **20**, Article number: 423 (2019) | [Cite this article](#)

# Challenges in Read Mapping

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



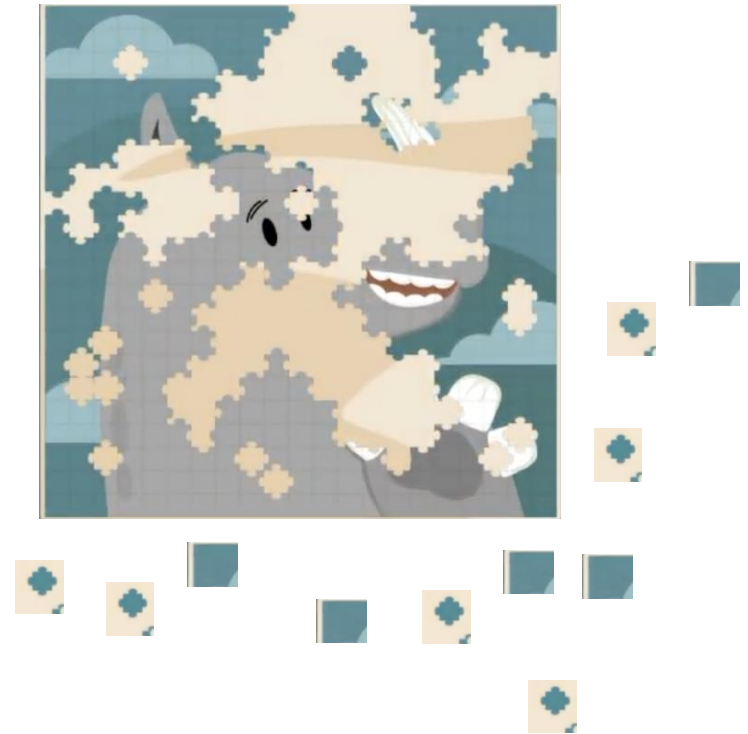
# Several Genome Analysis Pipelines





# Revisiting the Puzzle

---



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

## nature genetics

Letter | [Open Access](#) | Published: 19 November 2018

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

Rachel M. Sherman [✉](#), Juliet Forman, [...] Steven L. Salzberg [✉](#)

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

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

# Time to Change the Reference Genome

## Genome Biology

[Home](#) [About](#) [Articles](#) [Submission Guidelines](#)

Opinion | [Open Access](#) | Published: 09 August 2019

## Is it time to change the reference genome?

[Sara Ballouz](#), [Alexander Dobin](#) & [Jesse A. Gillis](#) 

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

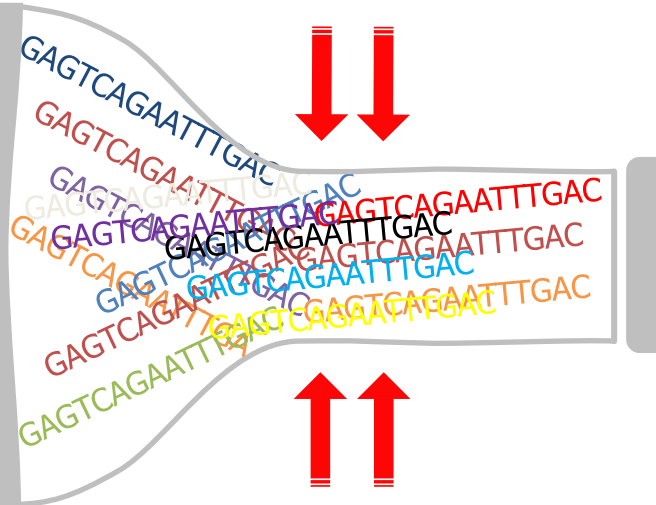
**12k** Accesses | **11** Citations | **45** Altmetric | [Metrics](#)

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

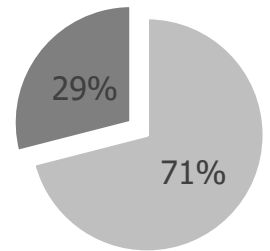
# Analysis is Bottlenecked in Read Mapping!!

**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



■ Read Mapping ■ Others

# Agenda for Today

---

- What is Read Mapping?
- **What Makes Read Mapper Slow?**
- Algorithmic & Hardware Acceleration
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

---

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	<b>Kilo</b> = 1,000
Bacterial genomes	1995	<b>Mega</b> = 1,000,000
Human genome	2000	<b>Giga</b> = 1,000,000,000
Human microbiome	2005	<b>Tera</b> = 1,000,000,000,000
50K Microbiomes	2015	<b>Peta</b> = 1,000,000,000,000,000
what is expected for the next 15 years ? (a Giga?)		
200K Microbiomes	2020	<b>Exa</b> = 1,000,000,000,000,000,000
1M Microbiomes	2025	<b>Zetta</b> = 1,000,000,000,000,000,000,000
Earth Microbiome	2030	<b>Yotta</b> = 1,000,000,000,000,000,000,000,000

Source:  
[@kyrpides](#)

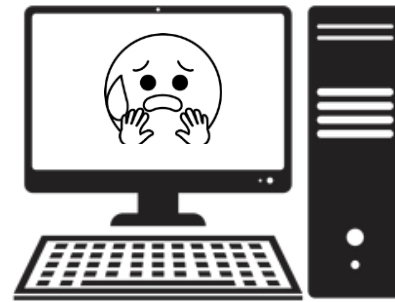
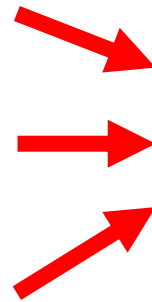
# Lack of Specialized Compute Capability

---



**Specialized** Machine  
for Sequencing

**FAST**



**General-Purpose** Machine  
for Analysis

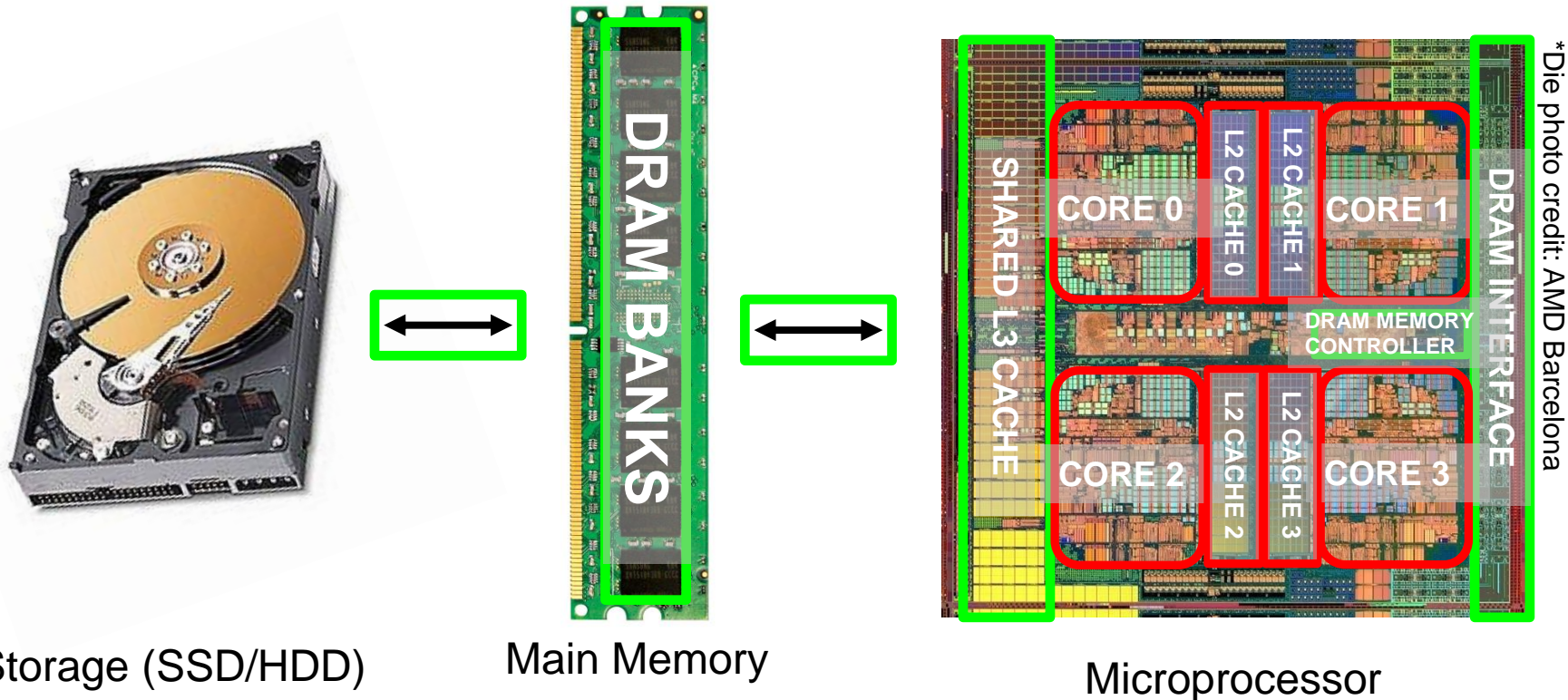
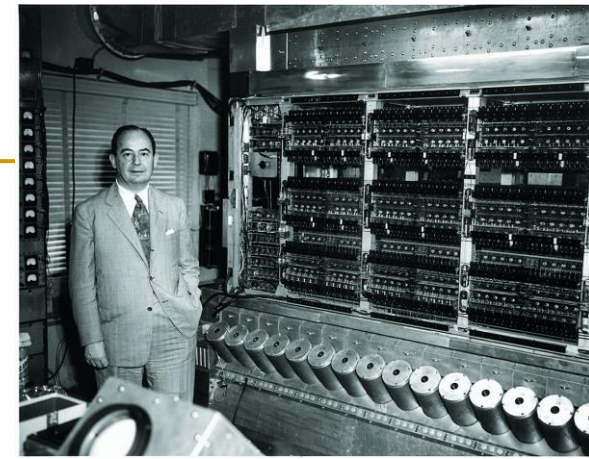
**SLOW**



# Today's Computing Systems

von Neumann model, 1945

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



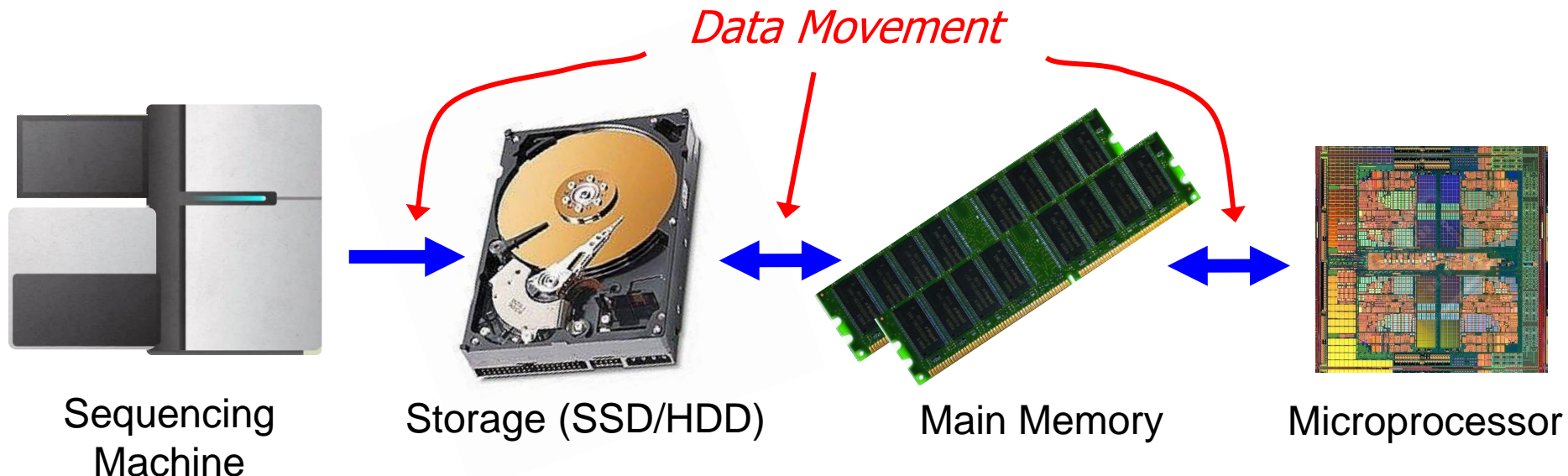
# The Problem

---

Data analysis  
is performed  
far away from the data

# Data Movement Dominates Performance

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



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

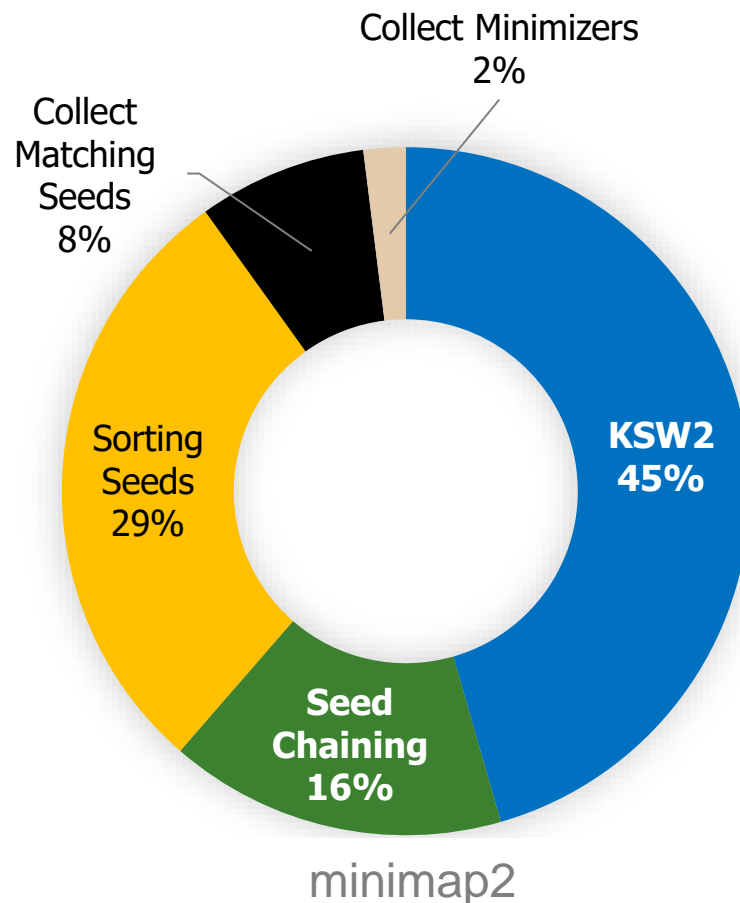
\* Kestor et al., "Quantifying the Energy Cost of Data Movement in Scientific Applications," IISWC 2013

\* 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

# Sequence Alignment in Unavoidable

- **Quadratic-time** dynamic-programming algorithm **WHY?!**

Enumerating all possible prefixes

- NETHERLANDS x SWITZERLAND
- NETHERLANDS x S
- NETHERLANDS x SW
- NETHERLANDS x SWI
- NETHERLANDS x SWIT
- NETHERLANDS x SWITZ
- NETHERLANDS x SWITZE
- NETHERLANDS x SWITZER
- NETHERLANDS x SWITZERL
- NETHERLANDS x SWITZERLA
- NETHERLANDS x SWITZERLAN
- NETHERLANDS x SWITZERLAND

		N	E	T	H	E	R	L	A	N	D	S
	0	1	2	3	4	5	6	7	8	9	10	11
S	1	1	2	3	4	5	6	7	8	9	10	10
W	2	1	2	3	4	5	6	7	8	9	10	11
I	3	3	4	5	6	7	8	9	10	11	11	11
T	4	4	4	3	4	5	6	7	8	9	10	11
Z	5	5	5	4	4	5	6	7	8	9	10	11
E	6	6	5	5	5	4	5	6	7	8	9	10
R	7	7	6	6	6	5	4	5	6	7	8	9
L	8	8	7	7	7	6	5	4	5	6	7	8
A	9	9	8	8	8	7	6	5	4	5	6	7
N	10	9	9	9	9	8	7	6	5	4	5	6
D	11	10	10	10	10	9	8	7	6	5	4	5

# Sequence Alignment in Unavoidable

- **Quadratic-time** dynamic-programming algorithm

Enumerating all possible prefixes

- **Data dependencies** limit the computation parallelism

Processing row (or column) after another

- **Entire matrix** is computed even though strings can be dissimilar.

Number of differences is computed only at the backtraking step.

		N	E	T	H	E	R	L	A	N	D	S
	0	1	2	3	4	5	6	7	8	9	10	11
S	1	1	2	3	4	5	6	7	8	9	10	10
W	2	2	2	3	4	5	6	7	8	9	10	11
I	3	3	3	3	4	5	6	7	8	9	10	11
T	4	4	4	3	4	5	6	7	8	9	10	11
Z	5	5	5	4	4	5	6	7	8	9	10	11
E	6	6	5	5	5	4	5	6	7	8	9	10
R	7	7	6	6	6	5	4	5	6	7	8	9
L	8	8	7	7	7	6	5	4	5	6	7	8
A	9	9	8	8	8	7	6	5	4	5	6	7
N	10	9	9	9	9	8	7	6	5	4	5	6
D	11	10	10	10	10	9	8	7	6	5	4	5

# Computational Cost is Mathematically Proven

arXiv.org > cs > arXiv:1412.0348

Search...

Help | Advanced

Computer Science > Computational Complexity

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

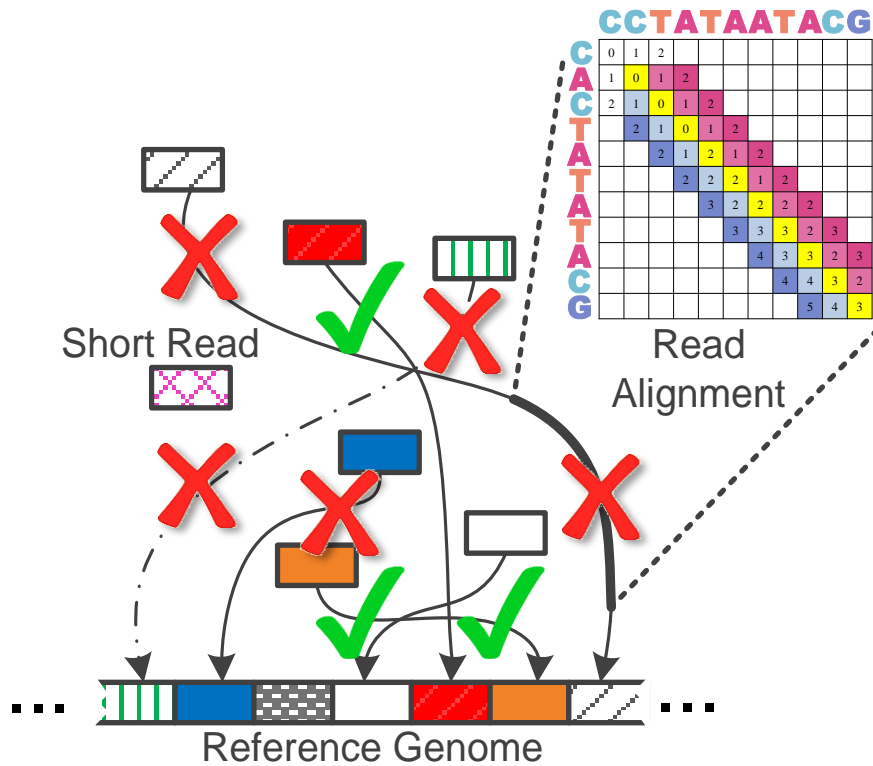
## Edit Distance Cannot Be Computed in Strongly Subquadratic Time (unless SETH is false)

Arturs Backurs, Piotr Indyk

The edit distance (a.k.a. the Levenshtein distance) between two strings is defined as the minimum number of insertions, deletions or substitutions of symbols needed to transform one string into another. The problem of computing the edit distance between two strings is a classical computational task, with a well-known algorithm based on dynamic programming. Unfortunately, all known algorithms for this problem run in nearly quadratic time.

In this paper we provide evidence that the near-quadratic running time bounds known for the problem of computing edit distance might be tight. Specifically, we show that, if the edit distance can be computed in time  $O(n^{2-\delta})$  for some constant  $\delta > 0$ , then the satisfiability of conjunctive normal form formulas with  $N$  variables and  $M$  clauses can be solved in time  $M^{O(1)}2^{(1-\epsilon)N}$  for a constant  $\epsilon > 0$ . The latter result would violate the Strong Exponential Time Hypothesis, which postulates that such algorithms do not exist.

# Large Search Space for Mapping Location



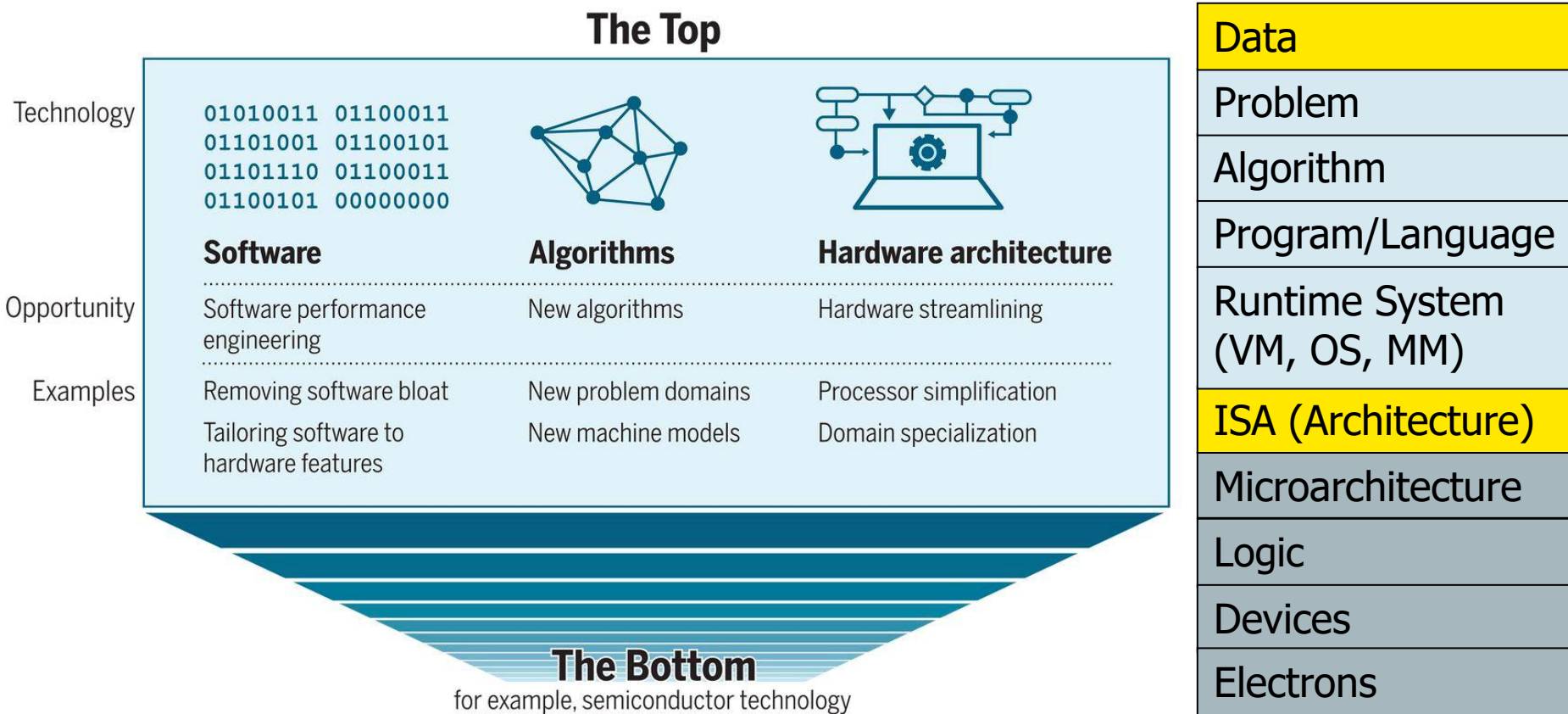
**98%**  
of candidate locations  
have high dissimilarity  
with a given read

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



# Computing System

Leiserson+, "[There's plenty of room at the Top: What will drive computer performance after Moore's law?](#)", Science, 2020



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.

# Software & Hardware Optimizations

## Multiplying Two 4096-by-4096 Matrices

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

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \times \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix} = \begin{bmatrix} & 58 \\ & \end{bmatrix}$$

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

Leiserson+, "[There's plenty of room at the Top: What will drive computer performance after Moore's law?](#)", Science, 2020

# FASTQ Parsing

Program	Language	t <sub>gzip</sub> (s)	t <sub>plain</sub> (s)	Comments
<a href="#">fqcnt_rs2_needletail.rs</a>	Rust	9.3	0.8	<a href="#">needletail</a> ; fasta/4-line fastq
<a href="#">fqcnt_c1_kseq.c</a>	C	9.7	1.4	multi-line fasta/fastq
<a href="#">fqcnt_cr1_klib.cr</a>	Crystal	9.7	1.5	kseq.h port
<a href="#">fqcnt_nim1_klib.nim</a>	Nim	10.5	2.3	kseq.h port
<a href="#">fqcnt_jl1_klib.jl</a>	Julia	11.2	2.9	kseq.h port
<a href="#">fqcnt_js1_k8.js</a>	Javascript	17.5	9.4	kseq.h port
<a href="#">fqcnt_go1.go</a>	Go	19.1	2.8	4-line only
<a href="#">fqcnt_lua1_klib.lua</a>	LuaJIT	28.6	27.2	partial kseq.h port
<a href="#">fqcnt_py2_rfq.py</a>	PyPy	28.9	14.6	partial kseq.h port
<a href="#">fqcnt_py2_rfq.py</a>	Python	42.7	19.1	partial kseq.h port

---

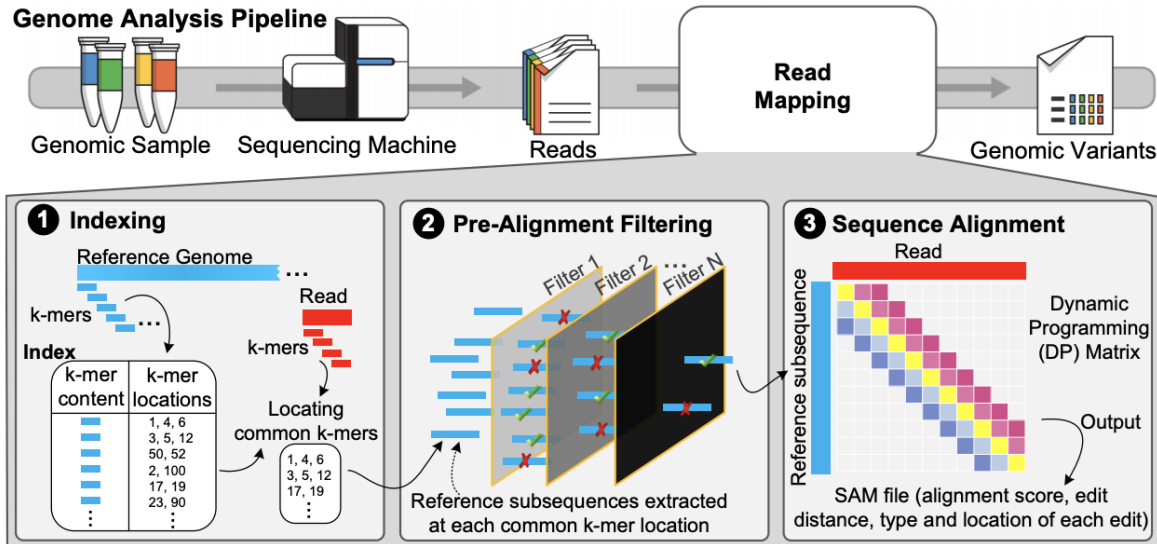
We need intelligent algorithms  
and intelligent architectures  
that handle data well

# Agenda for Today

---

- What is Read Mapping?
- What Makes Read Mapper Slow?
- **Algorithmic & Hardware Acceleration**
  - Seed Filtering Technique
  - Pre-alignment Filtering Technique
  - Read Alignment Acceleration

# 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.

# 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.

# Our Contributions

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

**GRIM-Filter [BMC Genomics'18]**

**GenASM [MICRO 2020]**

**SneakySnake [IEEE Micro'21]**

## Near-memory Sequence Alignment

**GenASM [MICRO 2020]**

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

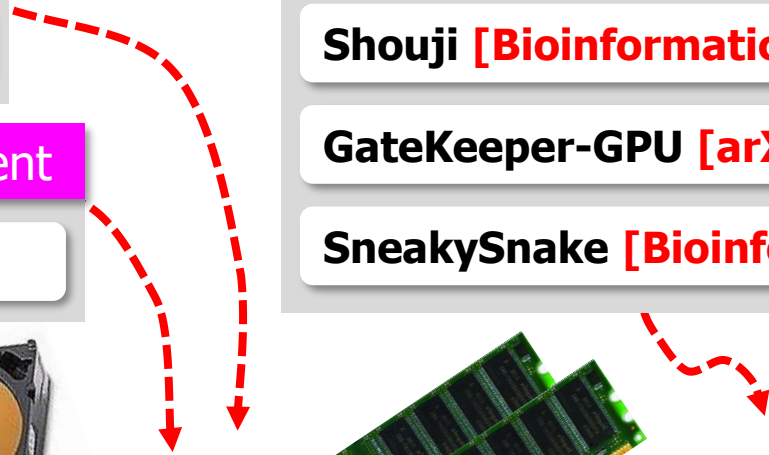
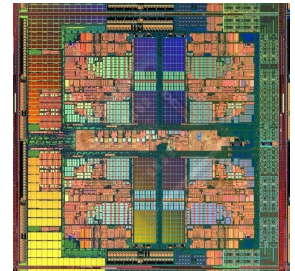
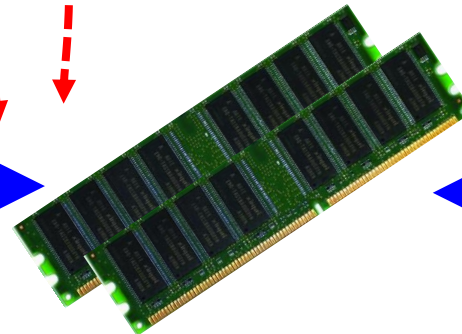
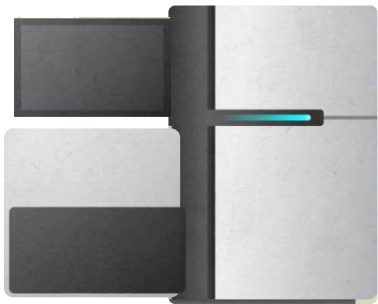
**GateKeeper [Bioinformatics'17]**

**MAGNET [AACBB'18]**

**Shouji [Bioinformatics'19]**

**GateKeeper-GPU [arXiv'21]**

**SneakySnake [Bioinformatics'20]**



Sequencing Machine

Storage (SSD/HDD)

Main Memory

Microprocessor



# Read Mapping in 111 pages!

In-depth analysis of 107 read mappers (1988-2020)

**Mohammed Alser**, Jeremy Rotman, Dhriti Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyung 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



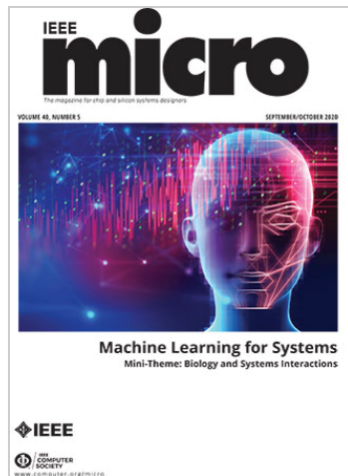
Mohammed Alser<sup>1,2,3†</sup>, Jeremy Rotman<sup>4†</sup>, Dhriti Deshpande<sup>5</sup>, Kodi Taraszka<sup>4</sup>, Huwenbo Shi<sup>6,7</sup>, Pelin Icer Baykal<sup>8</sup>, Harry Taegyung Yang<sup>4,9</sup>, Victor Xue<sup>4</sup>, Sergey Knyazev<sup>8</sup>, Benjamin D. Singer<sup>10,11,12</sup>, Brunilda Balliu<sup>13</sup>, David Koslicki<sup>14,15,16</sup>, Pavel Skums<sup>8</sup>, Alex Zelikovsky<sup>8,17</sup>, Can Alkan<sup>2,18</sup>, Onur Mutlu<sup>1,2,3†</sup> and Serghei Mangul<sup>5\*†</sup> 

# Detailed Analysis of Tackling the Bottleneck

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

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

IEEE Micro, August 2020.



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

*IEEE Micro*

## Accelerating Genome Analysis: A Primer on an Ongoing Journey

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

DOI Bookmark: [10.1109/MM.2020.3013728](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

◀	▶
<a href="#">Previous</a>	<a href="#">Next</a>
☰	<a href="#">Table of Contents</a>
📄	<a href="#">Past Issues</a>

# 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](https://doi.org/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

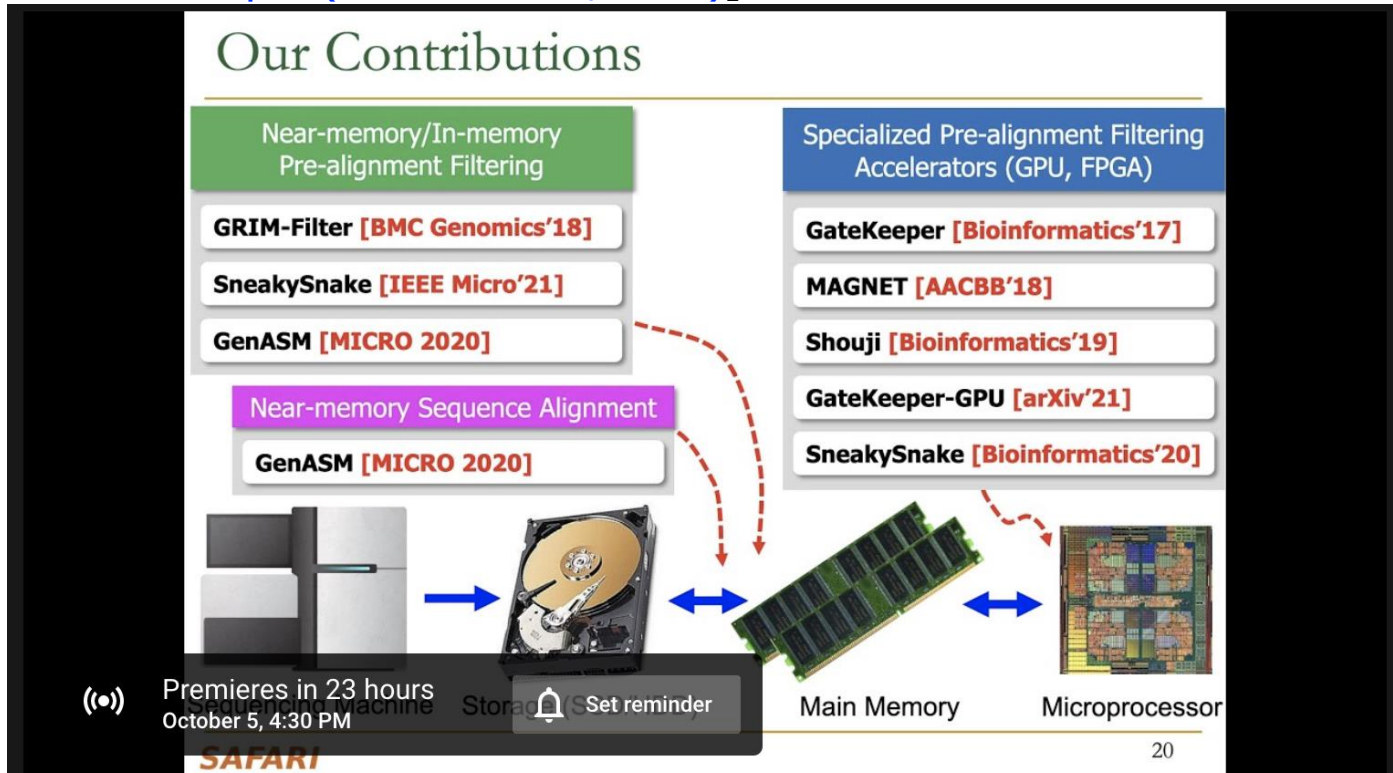
[Henk Corporaal](#), Eindhoven University of Technology, Eindhoven, The Netherlands

[Onur Mutlu](#), ETH Zürich, Zürich, Switzerland

<a href="#">◀</a>	<a href="#">▶</a>
<a href="#">Previous</a>	<a href="#">Next</a>
<a href="#">☰</a>	<a href="#">Table of Contents</a>
<a href="#">📄</a>	<a href="#">Past Issues</a>

# 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)]



# More on Intelligent Genome Analysis ...

- Mohammed Alser,  
**"Computer Architecture - Lecture 8: Intelligent Genome Analysis"**  
*ETH Zurich, Computer Architecture Course, Lecture 8, Virtual, 15 October 2021.*  
[[Slides \(pptx\) \(pdf\)](#)]  
[[Talk Video](#) (2 hour 54 minutes, including Q&A)]  
[[Related Invited Paper \(at IEEE Micro, 2020\)](#)]

**Our Solution: GateKeeper**

Alignment Filter + FPGA-based Alignment Filter = 1<sup>st</sup> FPGA-based Alignment Filter.

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

$\times 10^{12}$  mappings

$\times 10^3$  mappings

1 High throughput DNA sequencing (HTS) technologies  
2 Read Pre-Alignment Filtering Fast & Low False Positive Rate  
3 Read Alignment Slow & Zero False Positives

108

ETH ZENTRUM

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

# More on Fast Genome Analysis ...

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

Insight: Shifting a String Helps Similarity Search

7 matches 1 mismatch

ISTANBUL

ISTNBUL

ISTNBUL

81

46:08 / 1:37:37

Onur Mutlu

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

566 views · Premiered Feb 6, 2021

31 0 SHARE SAVE ...

 Onur Mutlu Lectures  
13.9K subscribers

ANALYTICS EDIT VIDEO

# 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=rqjl8ZyLsAg&list=PL5Q2soXY2Zi9E2bBVAgCqLgwiDRQDTyId>

# Prior Research on Genome Analysis (1/2)

---

- Alser+, "[Technology dictates algorithms: Recent developments in read alignment](#)", *Genome Biology*, 2021.
- Alser + "[SneakySnake: A Fast and Accurate Universal Genome Pre-Alignment Filter for CPUs, GPUs, and FPGAs.](#)", *Bioinformatics*, 2020.
- Senol Cali+, "[GenASM: A High-Performance, Low-Power Approximate String Matching Acceleration Framework for Genome Sequence Analysis](#)", *MICRO* 2020.
- Kim+, "[AirLift: A Fast and Comprehensive Technique for Translating Alignments between Reference Genomes](#)", *arXiv*, 2020
- Alser+, "[Accelerating Genome Analysis: A Primer on an Ongoing Journey](#)", *IEEE Micro*, 2020.



# Prior Research on Genome Analysis (2/2)

---

- Firtina+, "[Apollo: a sequencing-technology-independent, scalable and accurate assembly polishing algorithm](#)", *Bioinformatics*, 2019.
- Alser+, "[Shouji: a fast and efficient pre-alignment filter for sequence alignment](#)", *Bioinformatics* 2019.
- Kim+, "[GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies](#)", *BMC Genomics*, 2018.
- Alser+, "[GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping](#)", *Bioinformatics*, 2017.
- Alser+, "[MAGNET: understanding and improving the accuracy of genome pre-alignment filtering](#)", *IPSI Transaction*, 2017.

# P&S Mobile Genomics

## Lecture 4: Read Mapping

Dr. Mohammed Alser

 @meals

ETH Zurich

Fall 2022

7 November 2022