Computer Architecture
Lecture 3a: Introduction to Genome Sequence Analysis

Prof. Onur Mutlu
ETH Zürich

Fall 2020

24 September 2020

Four Key Problems + Directions

- Fundamentally Secure/Reliable/Safe Architectures

- Fundamentally Energy-Efficient Architectures
  - Memory-centric (Data-centric) Architectures

- Fundamentally Low-Latency and Predictable Architectures

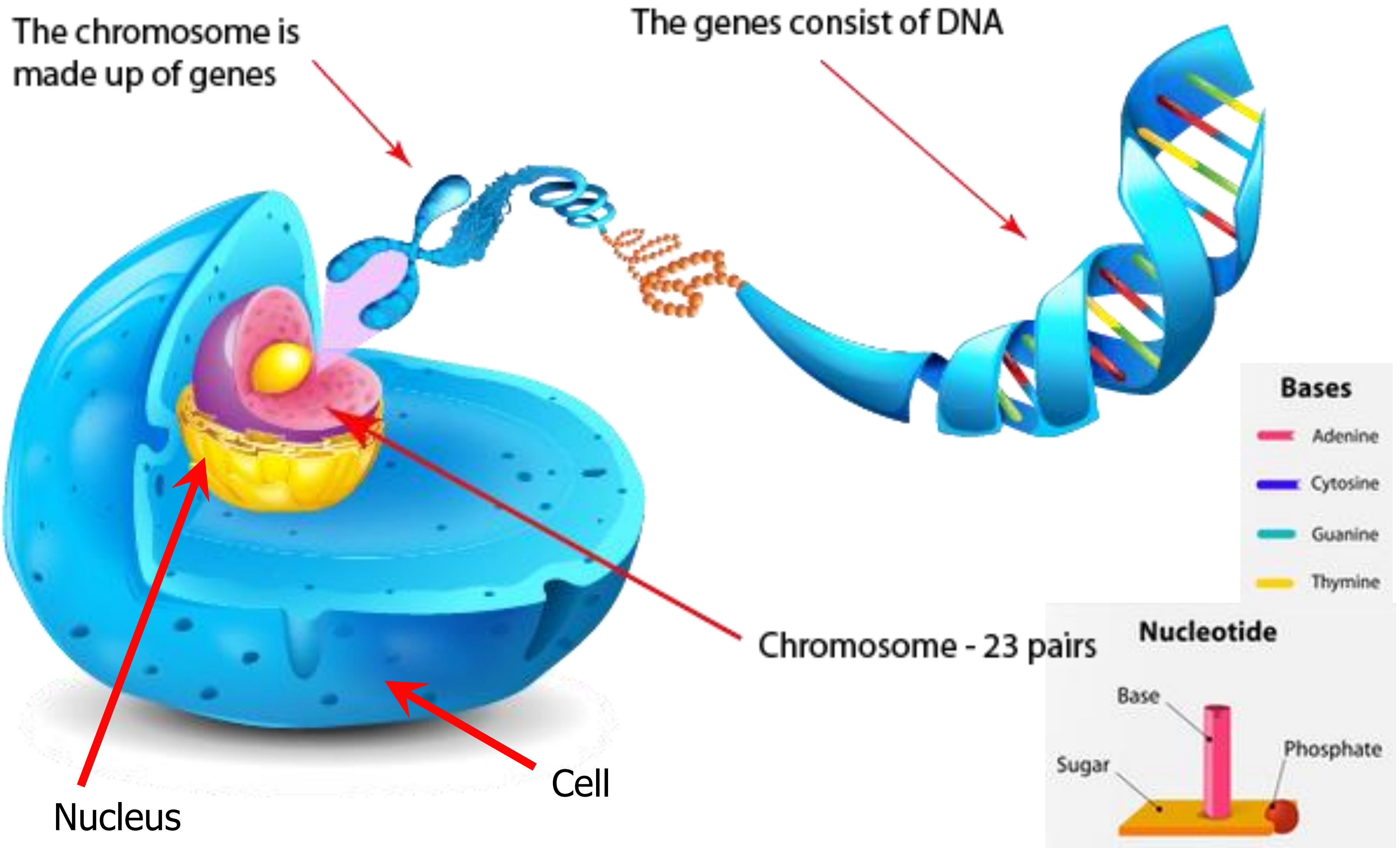
- Architectures for AI/ML, Genomics, Medicine, Health

A Motivating Detour: Genome Sequence Analysis

Our Dream (circa 2007)

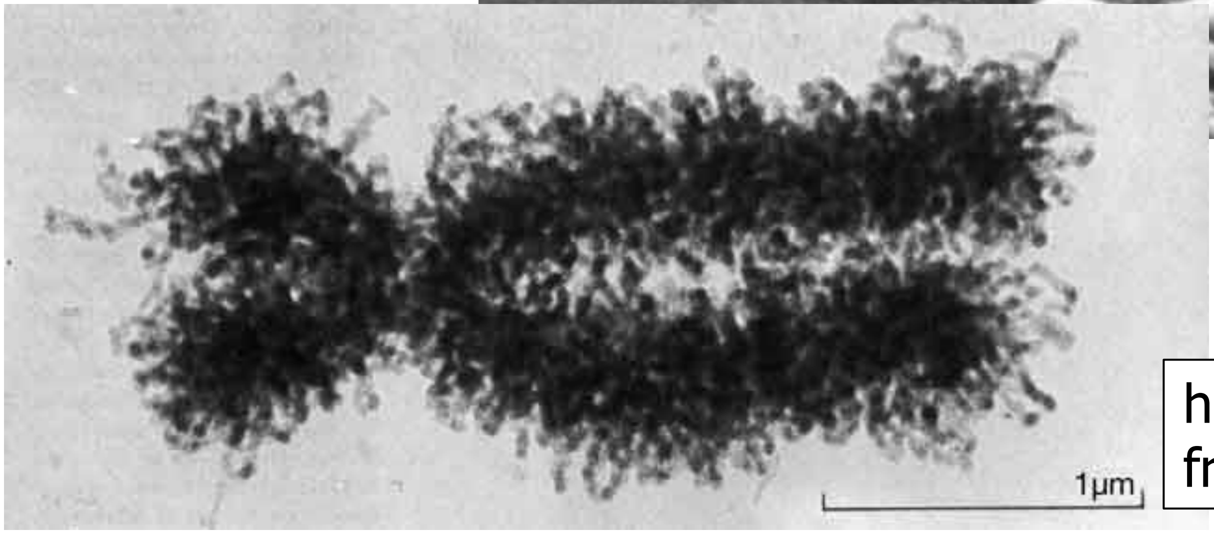
- An embedded device that can perform comprehensive genome analysis in real time (within a minute)
  - Which of these DNAs does this DNA segment match with?
  - What is the likely genetic disposition of this patient to this drug?
  - What disease/condition might this particular DNA/RNA piece associated with?
  - . . .

What Is a Genome Made Of?



The discovery of DNA’s double-helical structure (Watson+, 1953)

DNA Under Electron Microscope



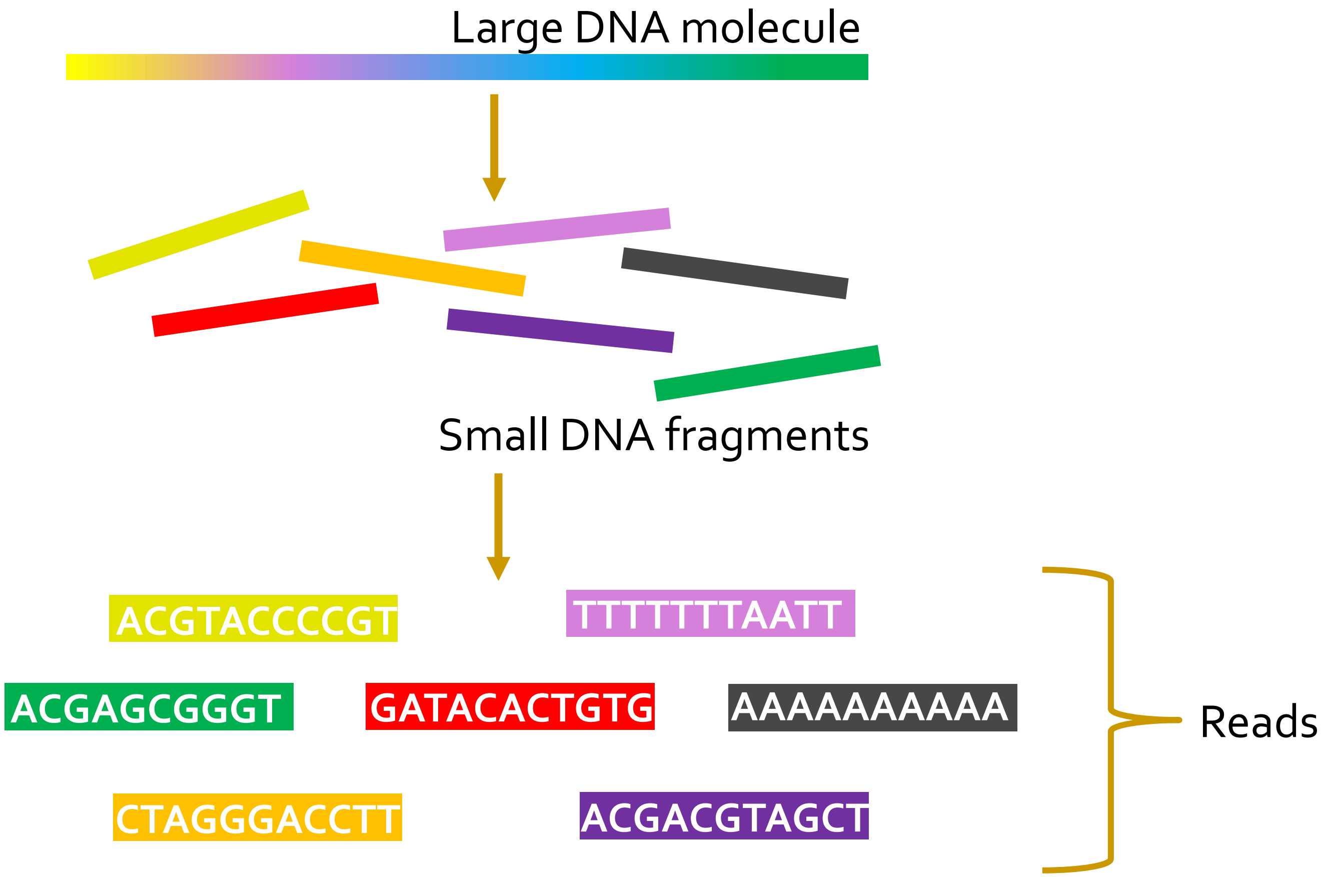
human chromosome #12
from HeLa’s cell

Genome Sequencing

- Goal:
  - Find the complete sequence of A, C, G, T's in DNA (or RNA).

- Challenge:
  - There is no machine that takes long DNA as an input, and gives the complete sequence as output
  - All sequencing machines chop DNA into pieces and identify relatively small pieces (but not how they fit together)

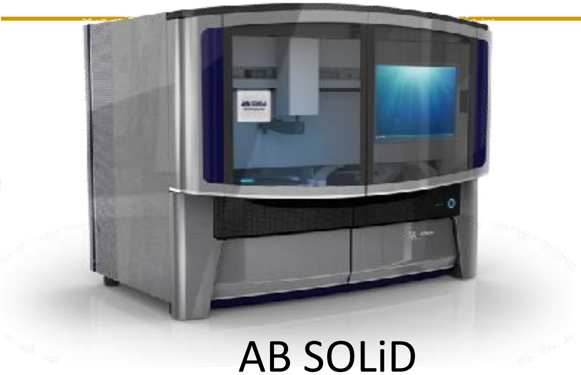
Genome Sequencing



Untangling Yarn Balls & DNA Sequencing

Genome Sequencers

Roche/454


AB SOLiD

Illumina MiSeq

Complete Genomics

Oxford Nanopore MinION

Illumina NovaSeq 6000

Illumina HiSeq2000

Pacific Biosciences RS

Oxford Nanopore GridION

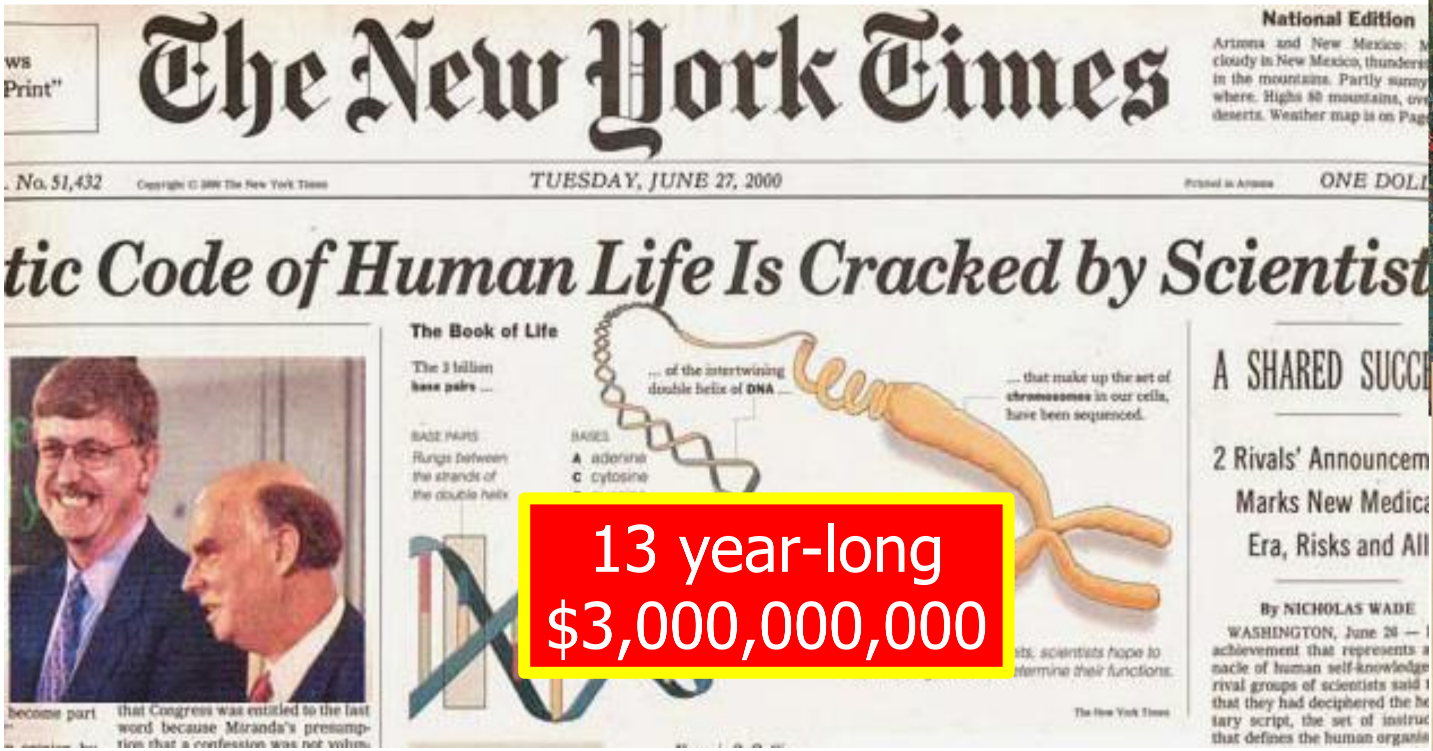
Ion Torrent PGM

Ion Torrent Proton

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

The Genomic Era

- 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.



ws
Print"

The New York Times

National Edition

Arizona and New Mexico: M
cloudy in New Mexico, thunders
in the mountains. Partly sunny
where. Highs 80 mountains, ov
deserts. Weather map is on Pag

No. 51,432 Copyright © 2000 The New York Times TUESDAY, JUNE 27, 2000 Printed in Arizona ONE DOLL

tic Code of Human Life Is Cracked by Scientist

The Book of Life
The 3 billion base pairs ...
... of the intertwining double helix of DNA ...
... that make up the set of chromosomes in our cells, have been sequenced.
BASE PAIRS
Rungs between the strands of the double helix
BASES
A adenine
C cytosine
ts, scientists hope to
etermine their functions.
The New York Times

A SHARED SUCC

2 Rivals' Announcem
Marks New Medica
Era, Risks and All

By NICHOLAS WADE

WASHINGTON, June 26 —
achievement that represents a
nacle of human self-knowledge
rival groups of scientists said t
that they had deciphered the he
tary script, the set of instruc
that defines the human organis

become part

that Congress was entitled to the last word because Miranda's presumption that a confession was not volun-

13 year-long
$3,000,000,000

15 February 2001

nature

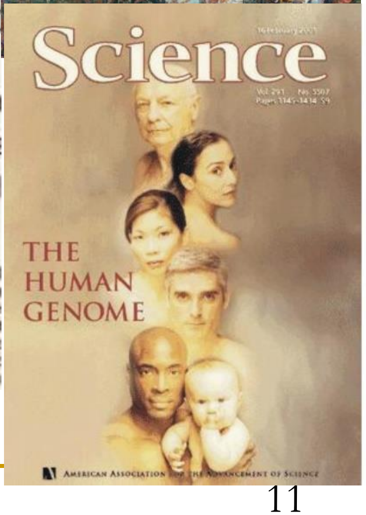
www.nature.com

the
human
genome

Nuclear fission
Five-dimensional energy landscapes

Seafloor spreading
The view from under the Arctic ice

Career prospects
Sequence creates new opportunities



16 February 2001

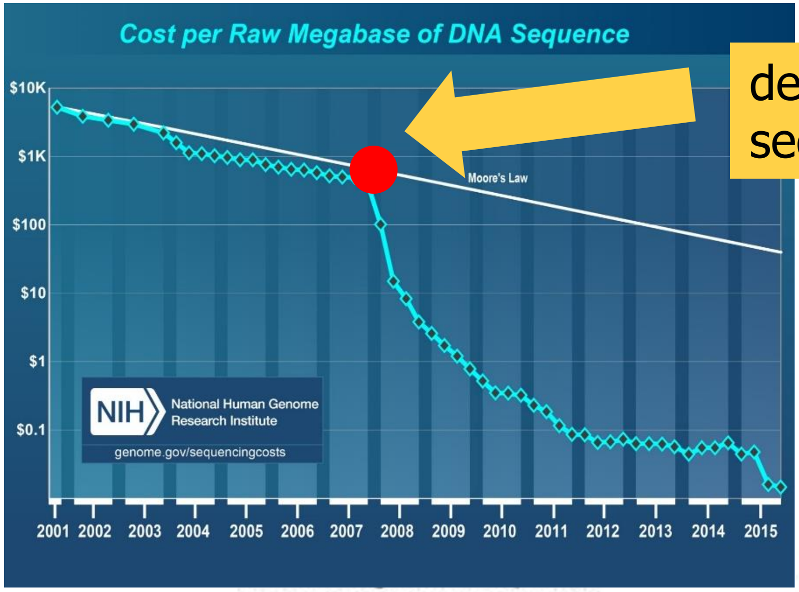
Science

Vol 291 No. 5507
Pages 1145-1434 $9

THE
HUMAN
GENOME

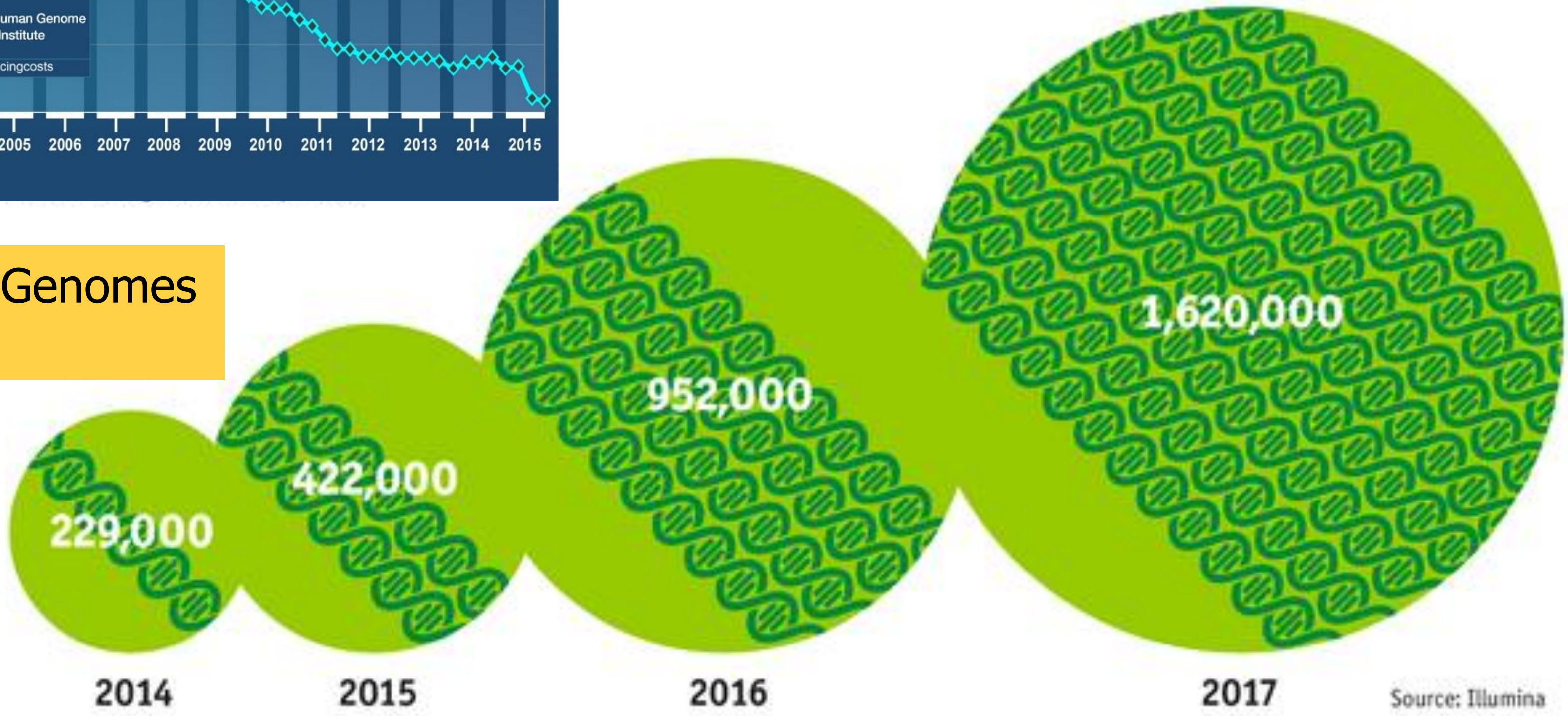
American Association for the Advancement of Science

The Genomic Era (continued)



development of high-throughput sequencing (HTS) technologies

Number of Genomes Sequenced

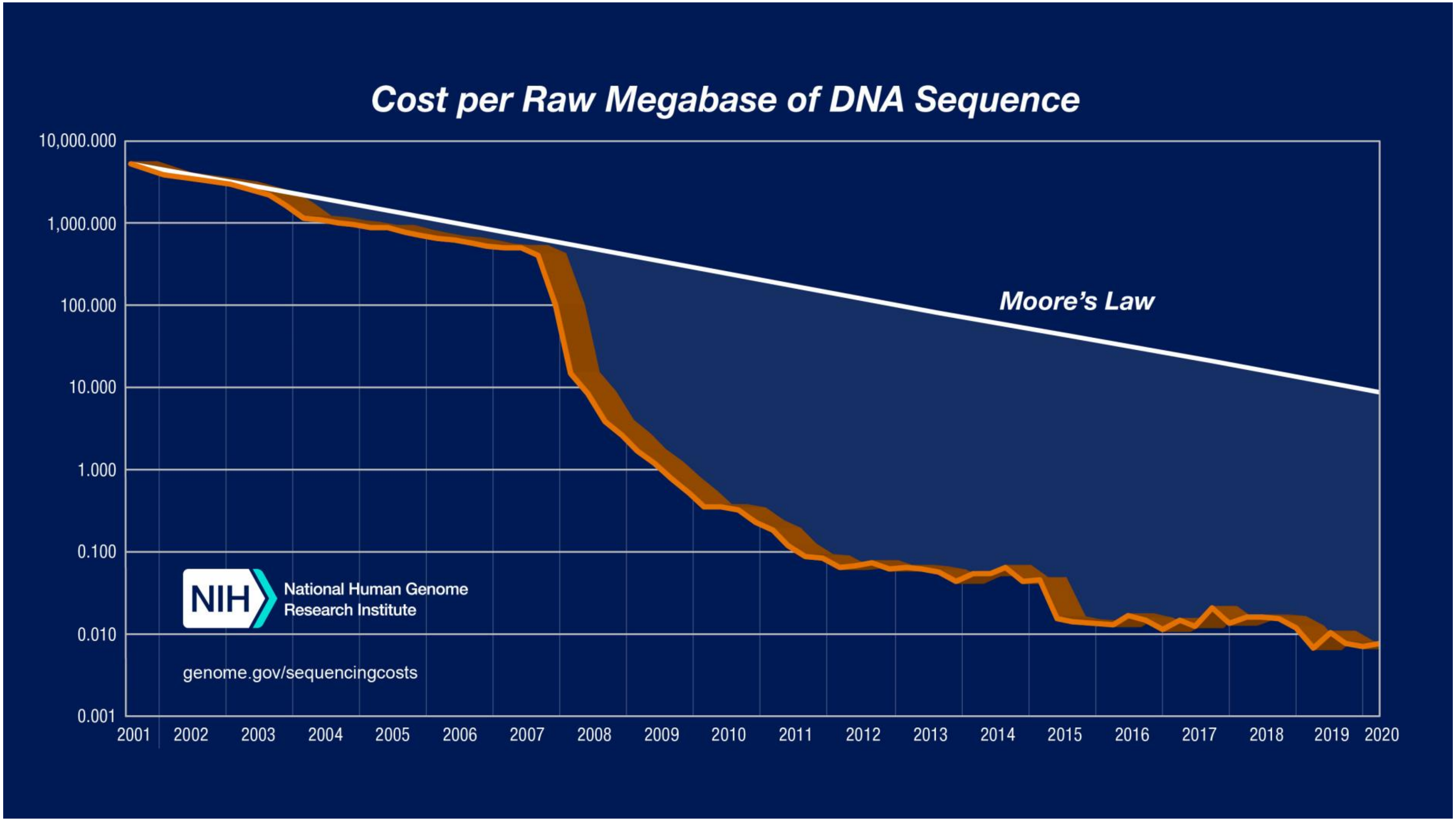


Source: Illumina

The Economist

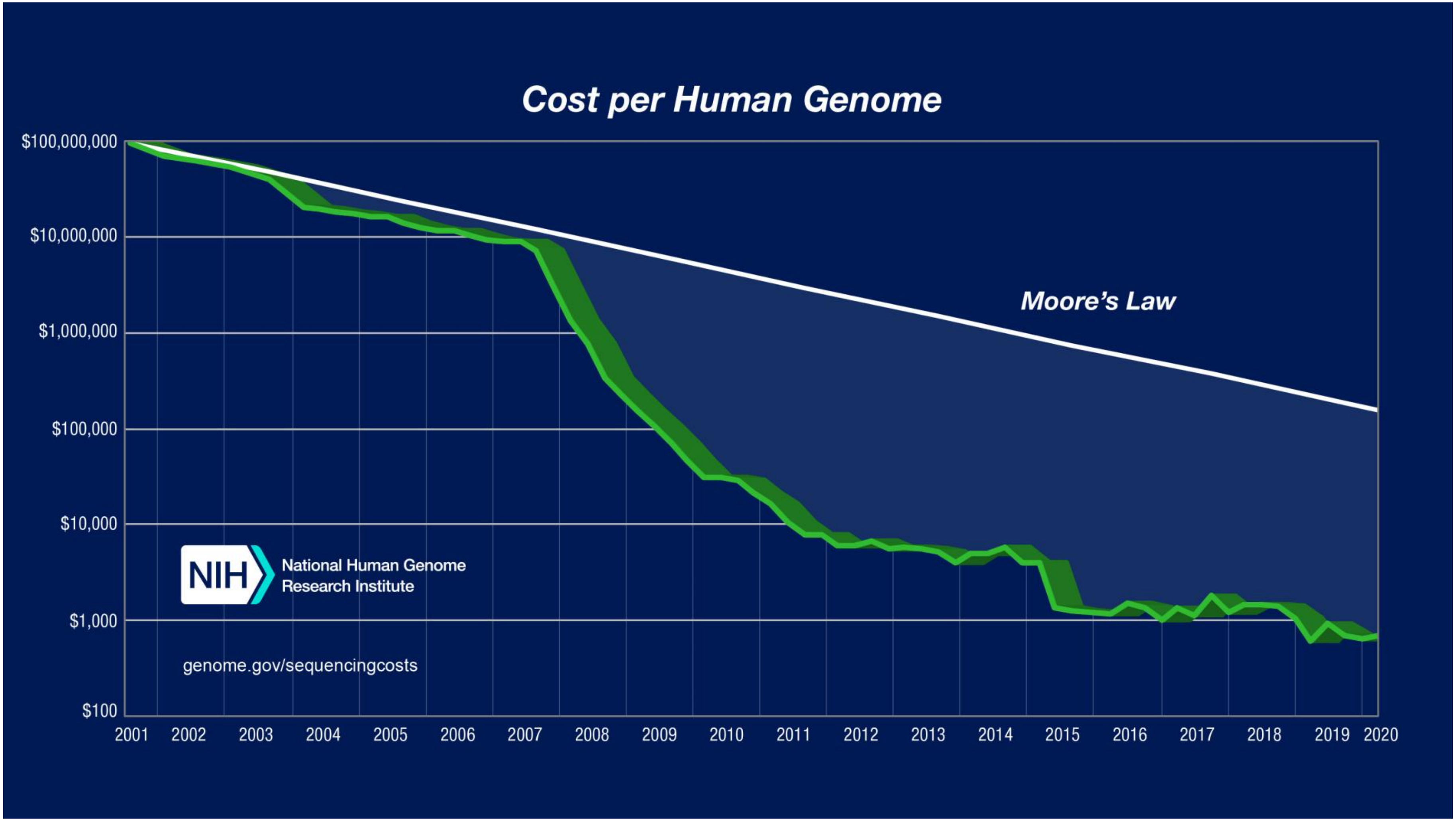
http://www.economist.com/news/21631808-so-much-genetic-data-so-many-uses-genes-unzipped

Cost of Sequencing



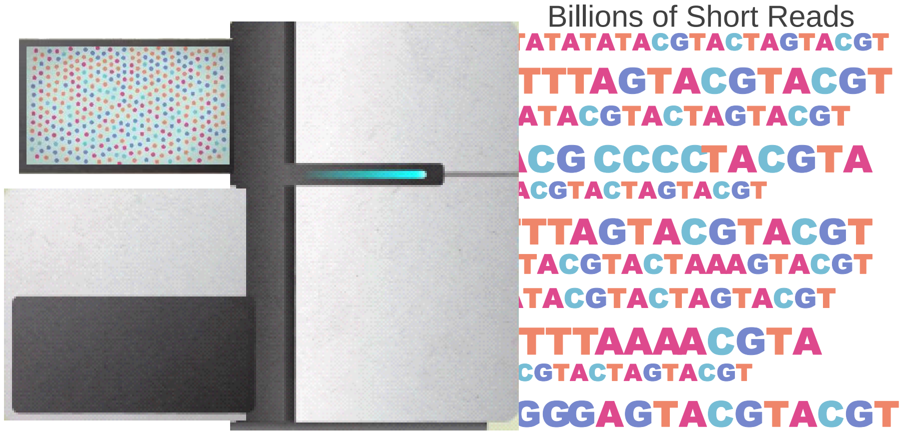
*From NIH (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data)

Cost of Sequencing (cont.)

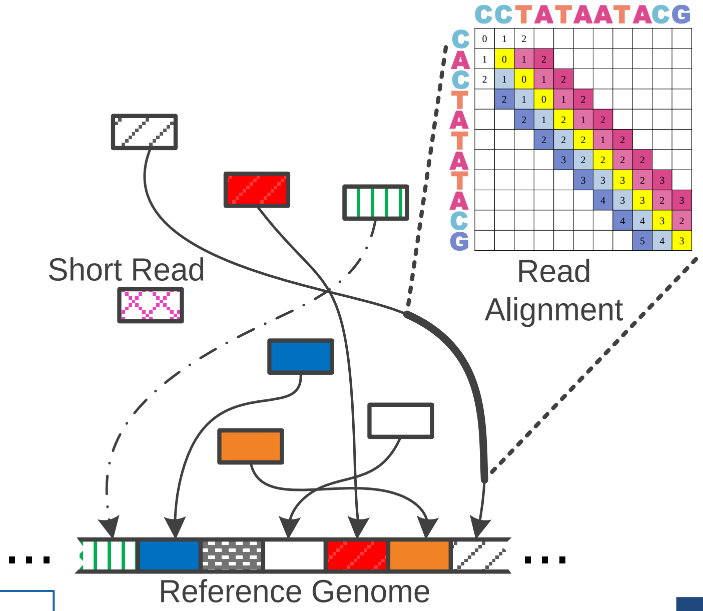


*From NIH (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data)

Genome Analysis



1 Sequencing



	C	C	T	A	T	A	A	T	A	C	G
C	0	1	2								
A	1	0	1	2							
C	2	1	0	1	2						
T		2	1	0	1	2					
A			2	1	2	1	2				
T				2	2	2	1	2			
A					3	2	2	2	2		
T						3	3	3	2	3	
A							4	3	3	2	3
C								4	4	3	2
G									5	4	3

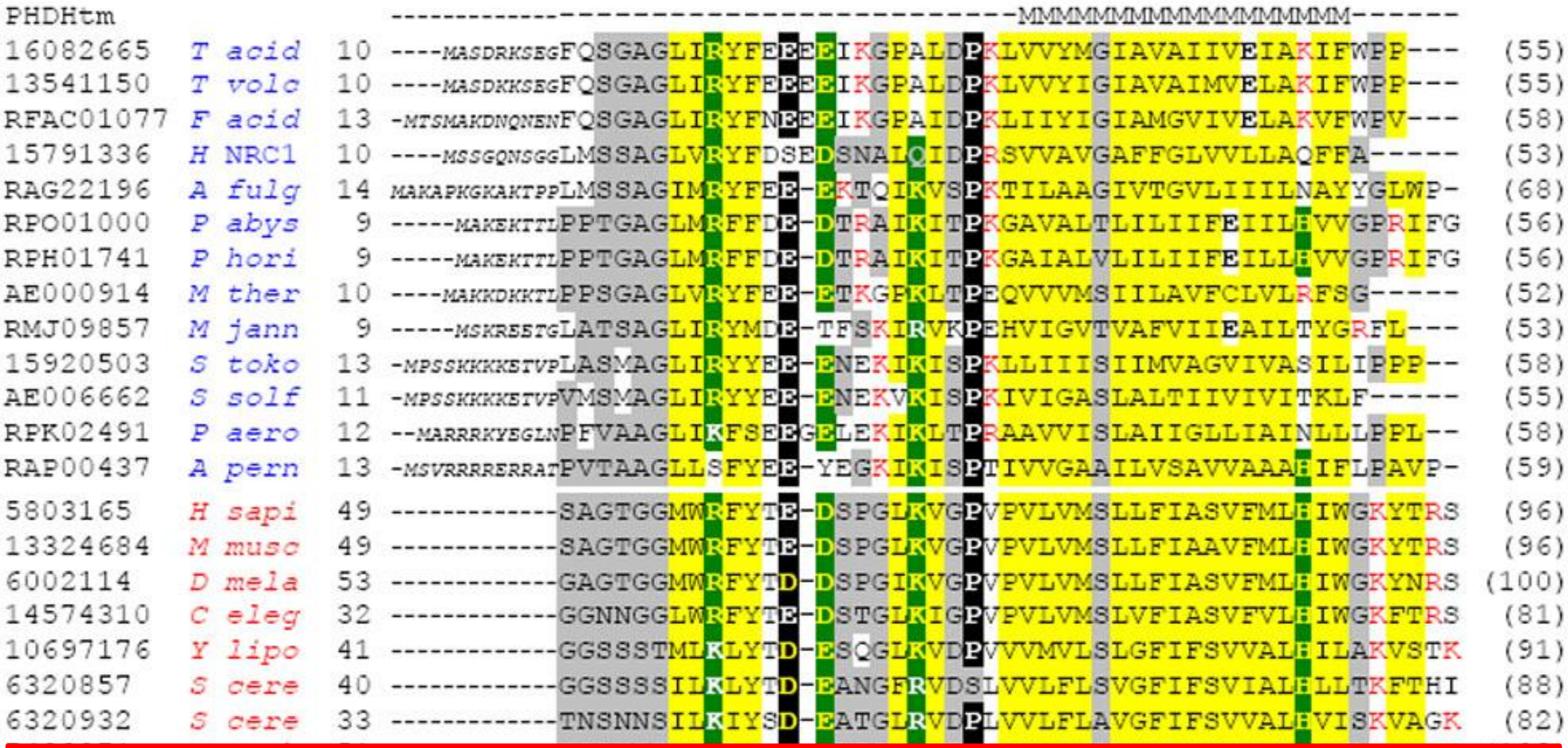
Read Mapping 2

```
reference: TTTATCGCTTCCATGACGCAG
read1:           ATCGCATCC
read2:          TATCGCATC
read3:               CATCCATGA
read4:             CGCTTCCAT
read5:                  CCATGACGC
read6:                TTCCATGAC
```

3 Variant Calling

Scientific Discovery 4

Multiple sequence alignment

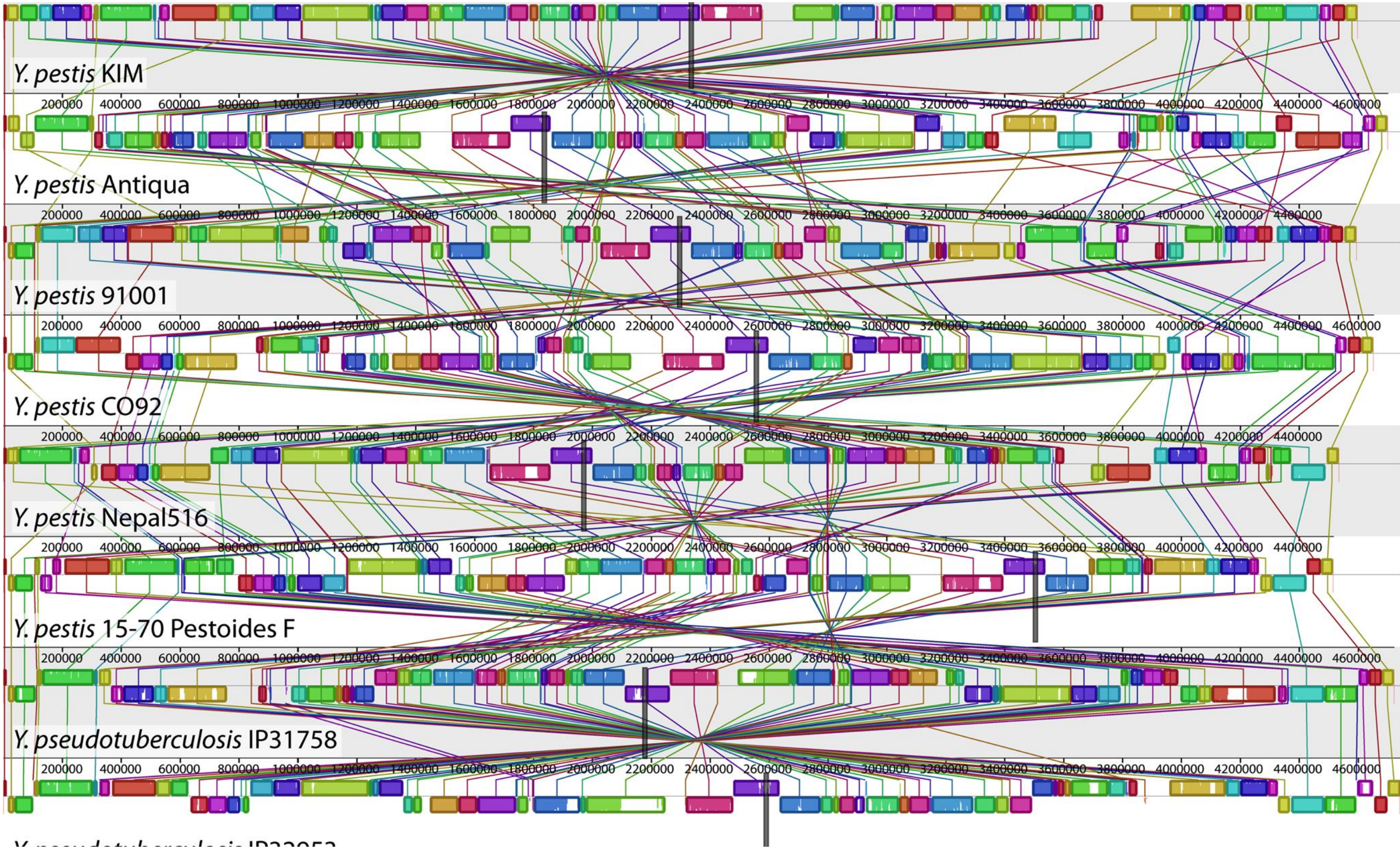


```
PHDHtm                        -------------------------------------MMMMMMMMMMMMMMMMMMM------
16082665   T acid     10  ----MASDRKSEGFQSGAGLIRYFEEEEIKGPALDPKLVVYMGIAVAIIVEIAKIFWPP---   (55)
13541150   T volc     10  ----MASDKKSEGFQSGAGLIRYFEEEEIKGPALDPKLVVYIGIAVAIMVELAKIFWPP---   (55)
RFAC01077  F acid     13  -MTSMAKDNQNENFQSGAGLIRYFNEEEIKGPAIDPKLIIYIGIAMGVIVELAKVFWPV---   (58)
15791336   H NRC1     10  ----MSSGQNSGGLMSSAGLVRYFDSEDSNALQIDPRSVVAVGAFFGLVVLLAQFFA-----   (53)
RAG22196   A fulg     14  MAKAPKGKAKTPPLMSSAGIMRYFEE-EKTQIKVSPKTILAAGIVTGVLIIILNAYYGLWP-   (68)
RPO01000   P abys      9  -----MAKEKTTLPPTGAGLMRFFDE-DTRAIKITPKGAVALTLILIIFEIILHVVGPRIFG   (56)
RPH01741   P hori      9  -----MAKEKTTLPPTGAGLMRFFDE-DTRAIKITPKGAIALVLILIIFEILLHVVGPRIFG   (56)
AE000914   M ther     10  ----MAKKDKKTLPPSGAGLVRYFEE-ETKGPKLTPEQVVVMSIILAVFCLVLRFSG-----   (52)
RMJ09857   M jann      9  -----MSKREETGLATSAGLIRYMDE-TFSKIRVKPEHVIGVTVAFVIIEAILTYGRFL---   (53)
15920503   S toko     13  -MPSSKKKKETVPLASMAGLIRYYEE-ENEKIKISPKLLIIISIIMVAGVIVASILIPPP--   (58)
AE006662   S solf     11  -MPSSKKKKETVPVMSMAGLIRYYEE-ENEKVKISPKIVIGASLALTIIVIVITKLF-----   (55)
RPK02491   P aero     12  --MARRRKYEGLNPFVAAGLIKFSEEGELEKIKLTPRAAVVISLAIIGLLIAINLLLPPL--   (58)
RAP00437   A pern     13  -MSVRRRRERRATPVTAAGLLSFYEE-YEGKIKISPTIVVGAAILVSAVVAAAHIFLPAVP-   (59)

5803165    H sapi     49  -------------SAGTGGMWRFYTE-DSPGLKVGPVPVLVMSLLFIASVFMLHIWGKYTRS   (96)
13324684   M musc     49  -------------SAGTGGMWRFYTE-DSPGLKVGPVPVLVMSLLFIAAVFMLHIWGKYTRS   (96)
6002114    D mela     53  -------------GAGTGGMWRFYTD-DSPGIKVGPVPVLVMSLLFIASVFMLHIWGKYNRS  (100)
14574310   C eleg     32  -------------GGNNGGLWRFYTE-DSTGLKIGPVPVLVMSLVFIASVFVLHIWGKFTRS   (81)
10697176   Y lipo     41  -------------GGSSSTMLKLYTD-ESQGLKVDPVVVMVLSLGFIFSVVALHILAKVSTK   (91)
6320857    S cere     40  -------------GGSSSSILKLYTD-EANGFRVDSLVVLFLSVGFIFSVIALHLLTKFTHI   (88)
6320932    S cere     33  -------------TNSNNSILKIYSD-EATGLRVDPLVVLFLAVGFIFSVVALHVISKVAGK   (82)
```

Example Question: If I give you a bunch of sequences, tell me where they are the same and where they are different.

Genome Sequence Alignment: Example



Source: By Aaron E. Darling, István Miklós, Mark A. Ragan - Figure 1 from Darling AE, Miklós I, Ragan MA (2008). "Dynamics of Genome Rearrangement in Bacterial Populations". PLOS Genetics. DOI:10.1371/journal.pgen.1000128., CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=30550950

The Genetic Similarity Between Species

Human ~ Human

99.9%

Human ~ Chimpanzee

96%

Human ~ Cat

90%

Human ~ Cow

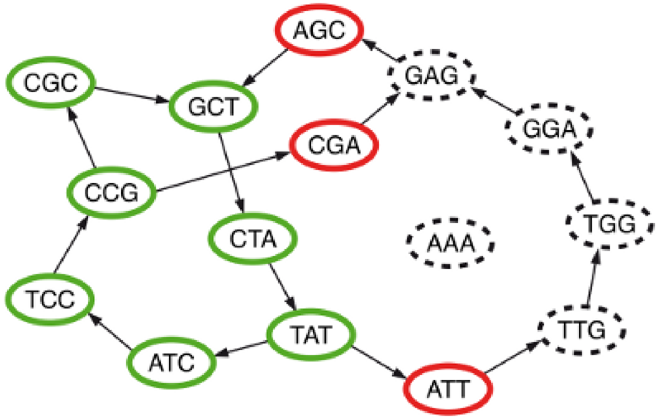
80%

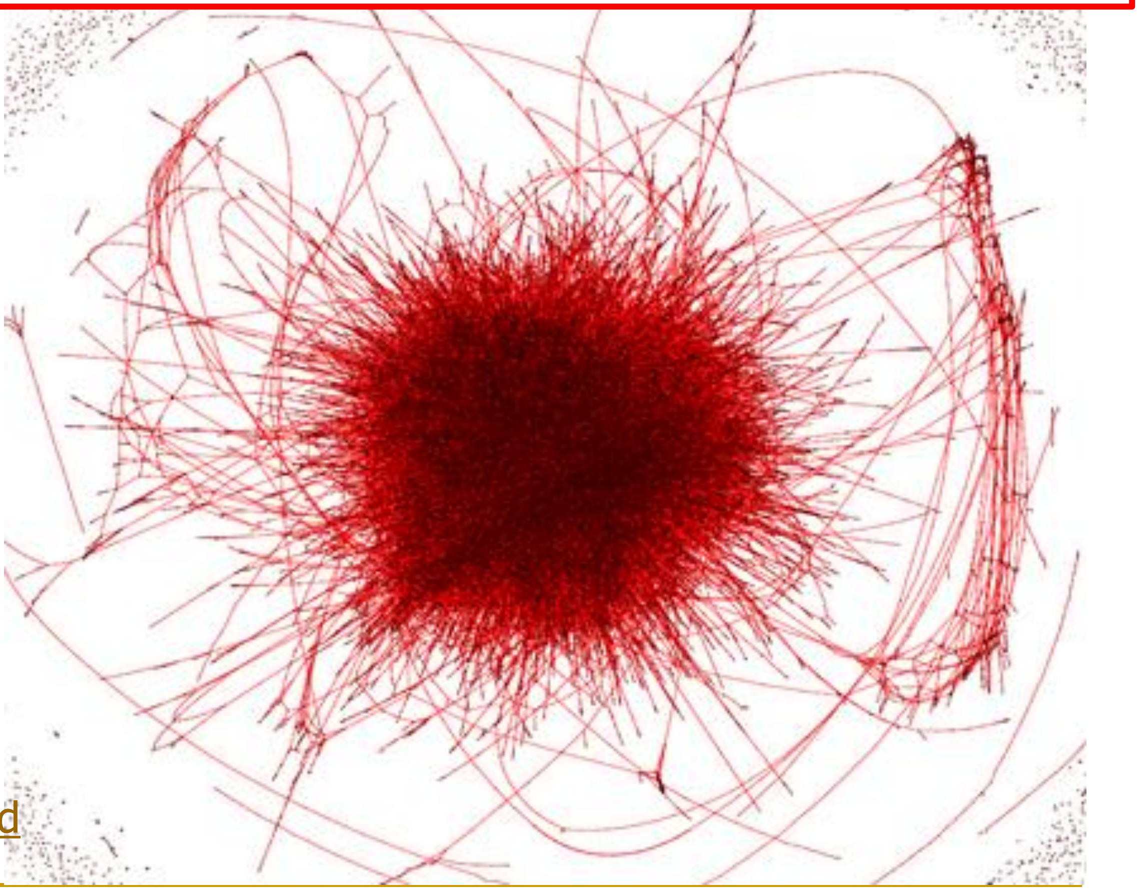
Human ~ Banana

50-60%

Metagenomics, genome assembly, de novo sequencing

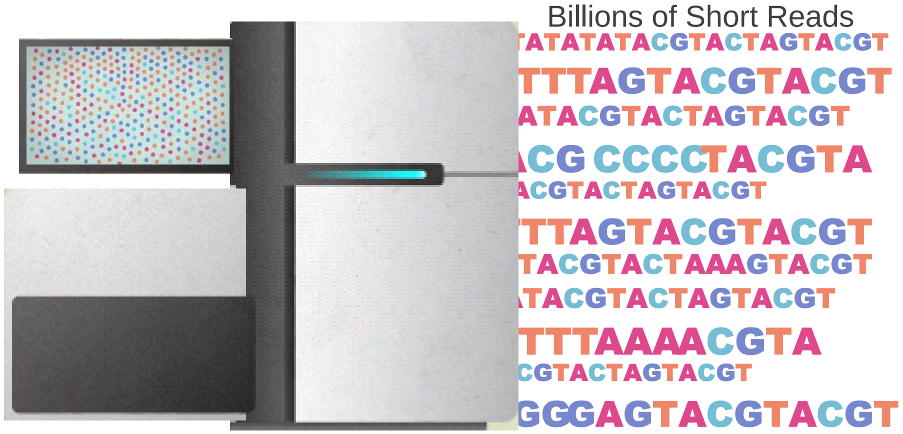
Question 2: Given a bunch of short sequences, Can you identify the approximate species cluster for genomically unknown organisms?



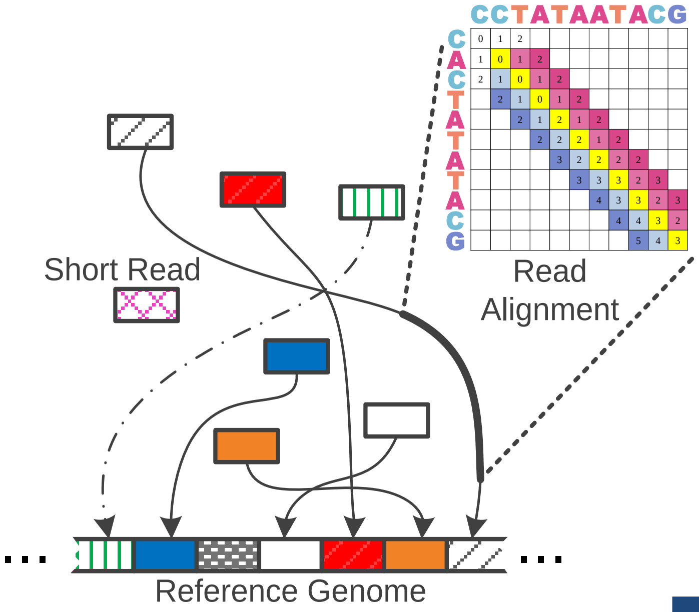


uncleaned de Bruijn graph

http://math.oregonstate.edu/~koslickd

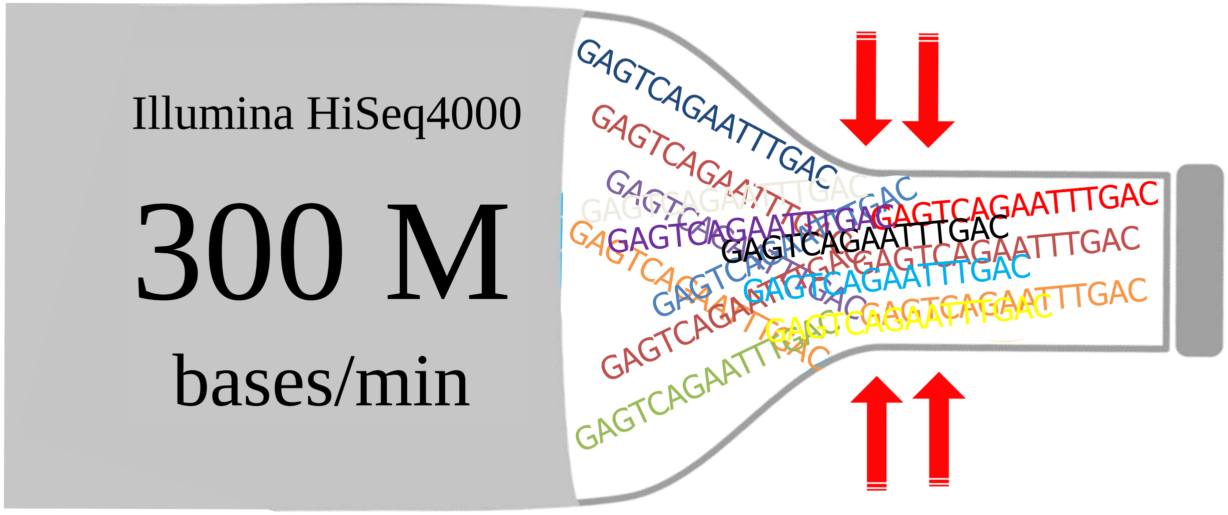


1 Sequencing



Read Mapping 2

Bottlenecked in Mapping!!



Illumina HiSeq4000

300 M

bases/min

on average

2 M

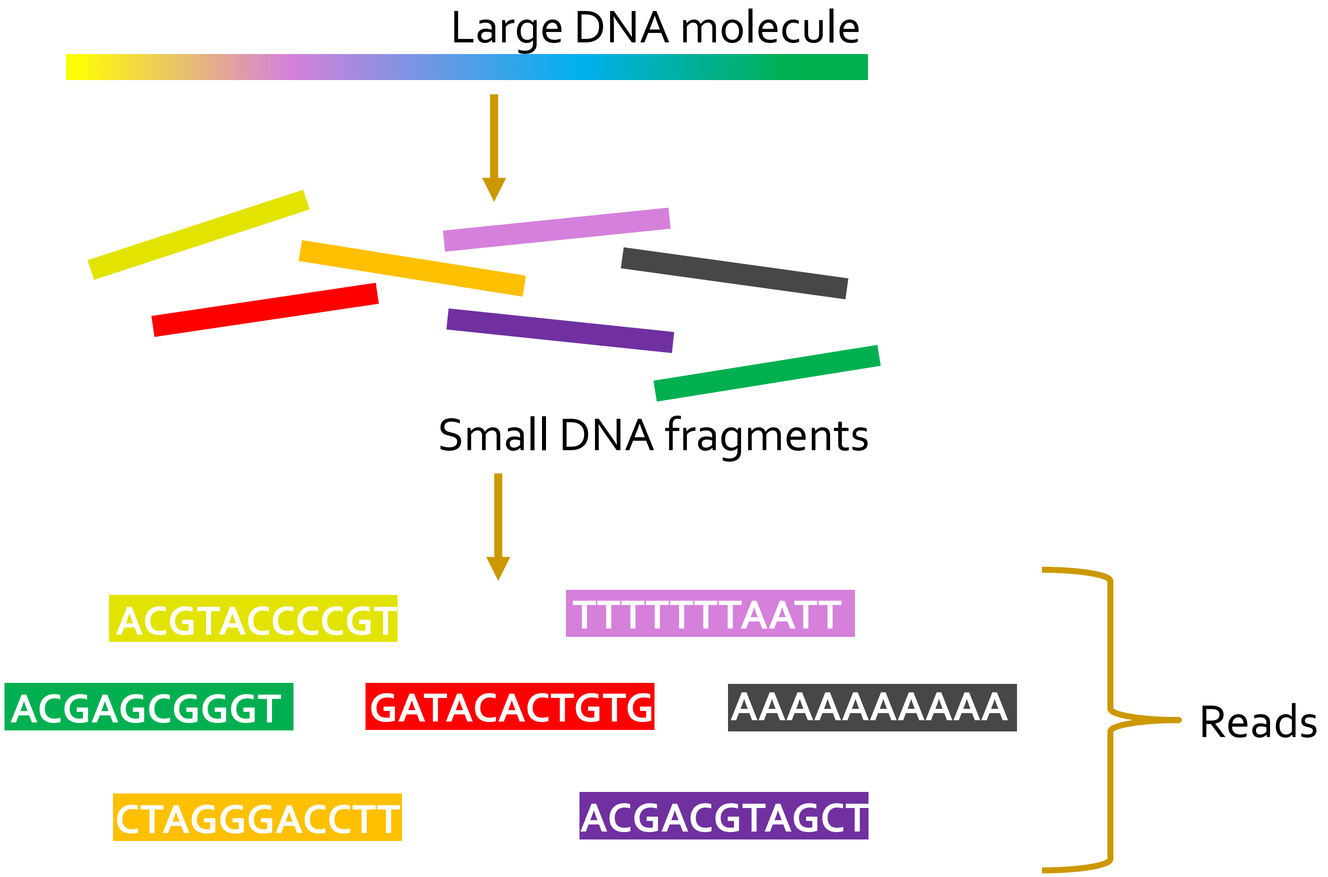
bases/min

(0.6%)

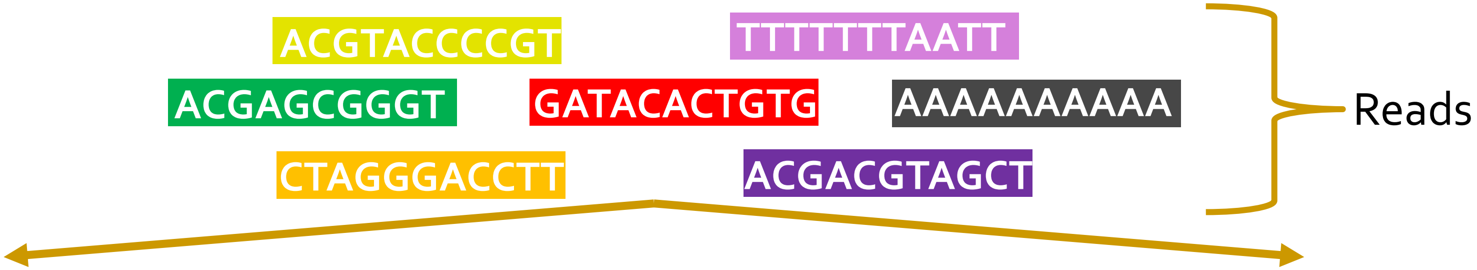
Problem

Need to construct the entire genome from many reads

Genome Sequencing



Genome Sequence Analysis



ACGTACCCCGT

TTTTTTTAATT

ACGAGCGGGT

GATACACTGTG

AAAAAAAAAA

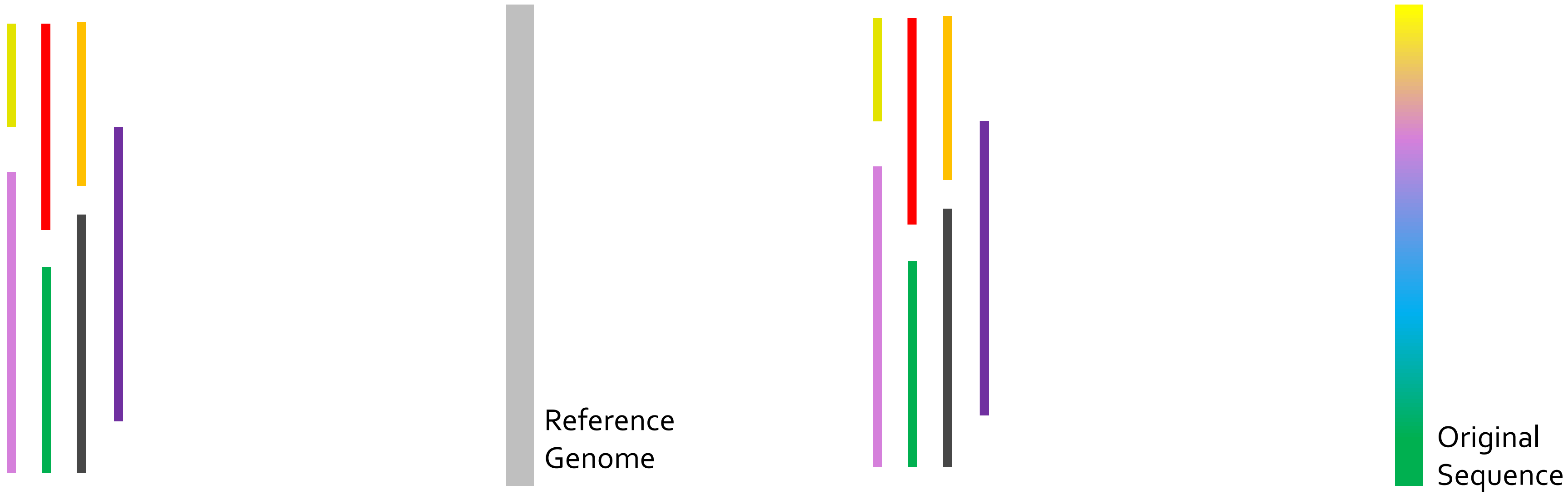
CTAGGGACCTT

ACGACGTAGCT

Reads

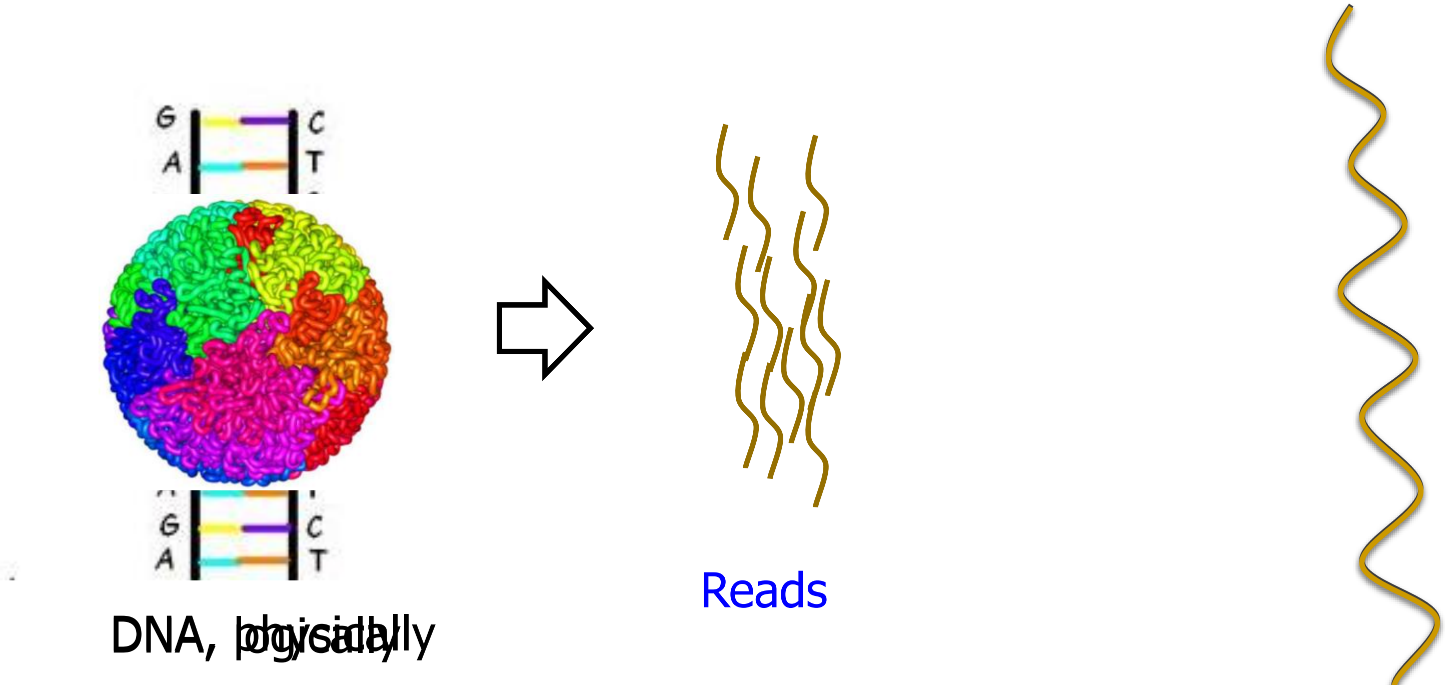
Read Mapping, method of aligning the reads against a known reference genome to **detect matches and variations.**

***De novo* Assembly,** method of merging the reads in order to **construct** the original sequence.



Read Mapping

- Map many short DNA fragments (reads) to a known reference genome with some differences allowed



Reference genome

Reads

DNA, logically

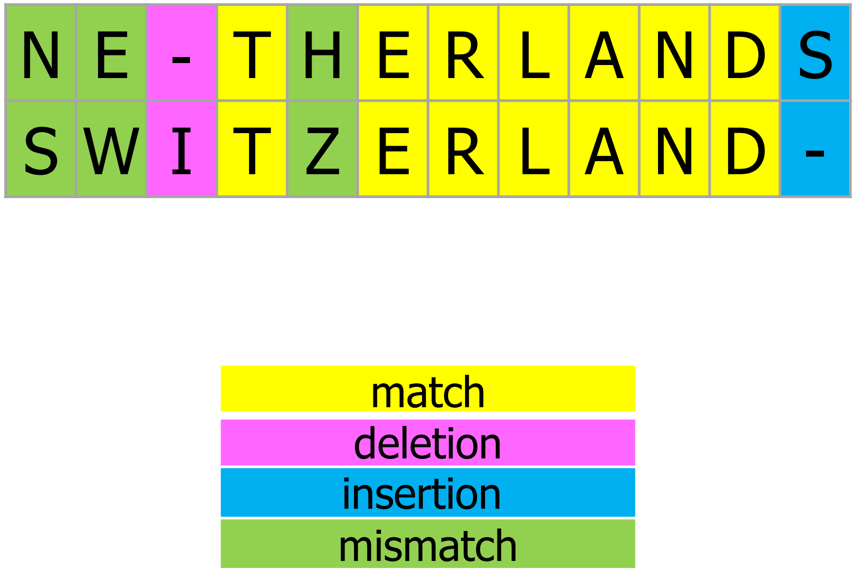
DNA, physically

Mapping short reads to reference genome is challenging (billions of 50-300 base pair reads)

Read 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.

NETHERLANDS x SWITZERLAND



N	E	-	T	H	E	R	L	A	N	D	S
S	W	I	T	Z	E	R	L	A	N	D	-

match
deletion
insertion
mismatch

Challenges in Read Mapping

- Need to find many mappings of each read
    - How can we find all mappings efficiently?

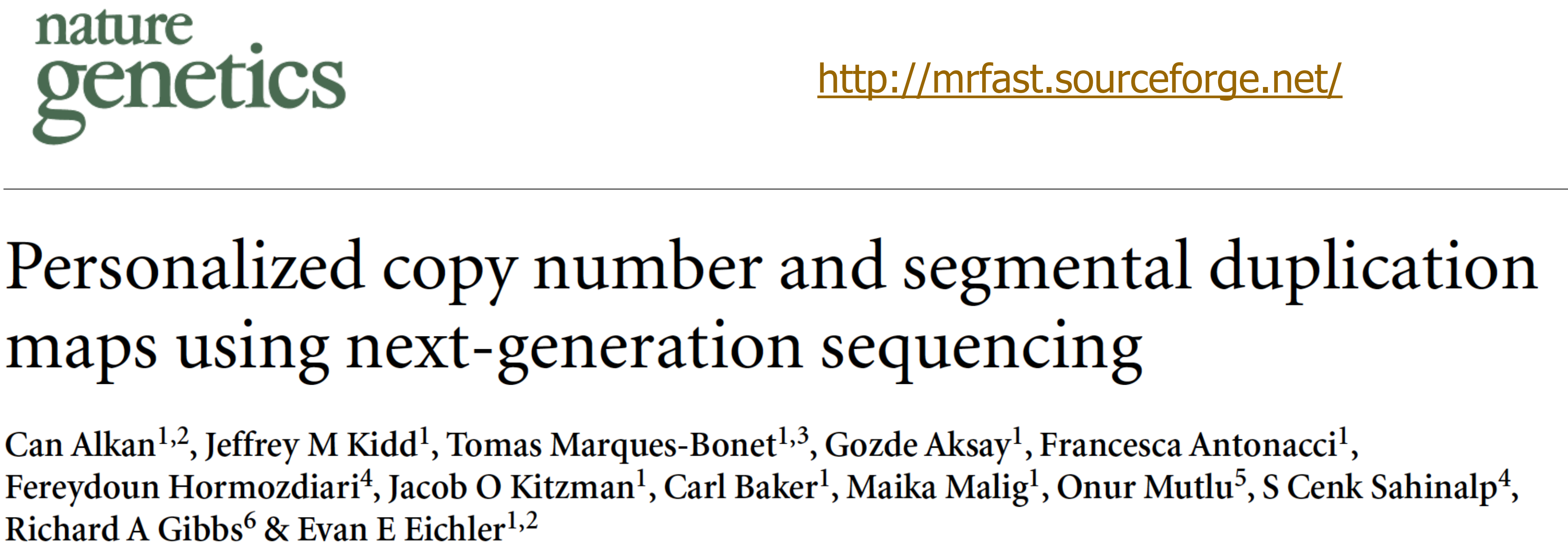
- Need to tolerate small variances/errors in each read
    - Each individual is different: Subject’s DNA may slightly differ from the reference (Mismatches, insertions, deletions)
    - How can we efficiently map each read with up to *e* errors present?

- Need to map each read very fast (i.e., performance is important)
    - Human DNA is 3.2 billion base pairs long → Millions to billions of reads (State-of-the-art mappers take weeks to map a human’s DNA)
    - How can we design a much higher performance read mapper?

Our First Step: Comprehensive Mapping

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

http://mrfast.sourceforge.net/



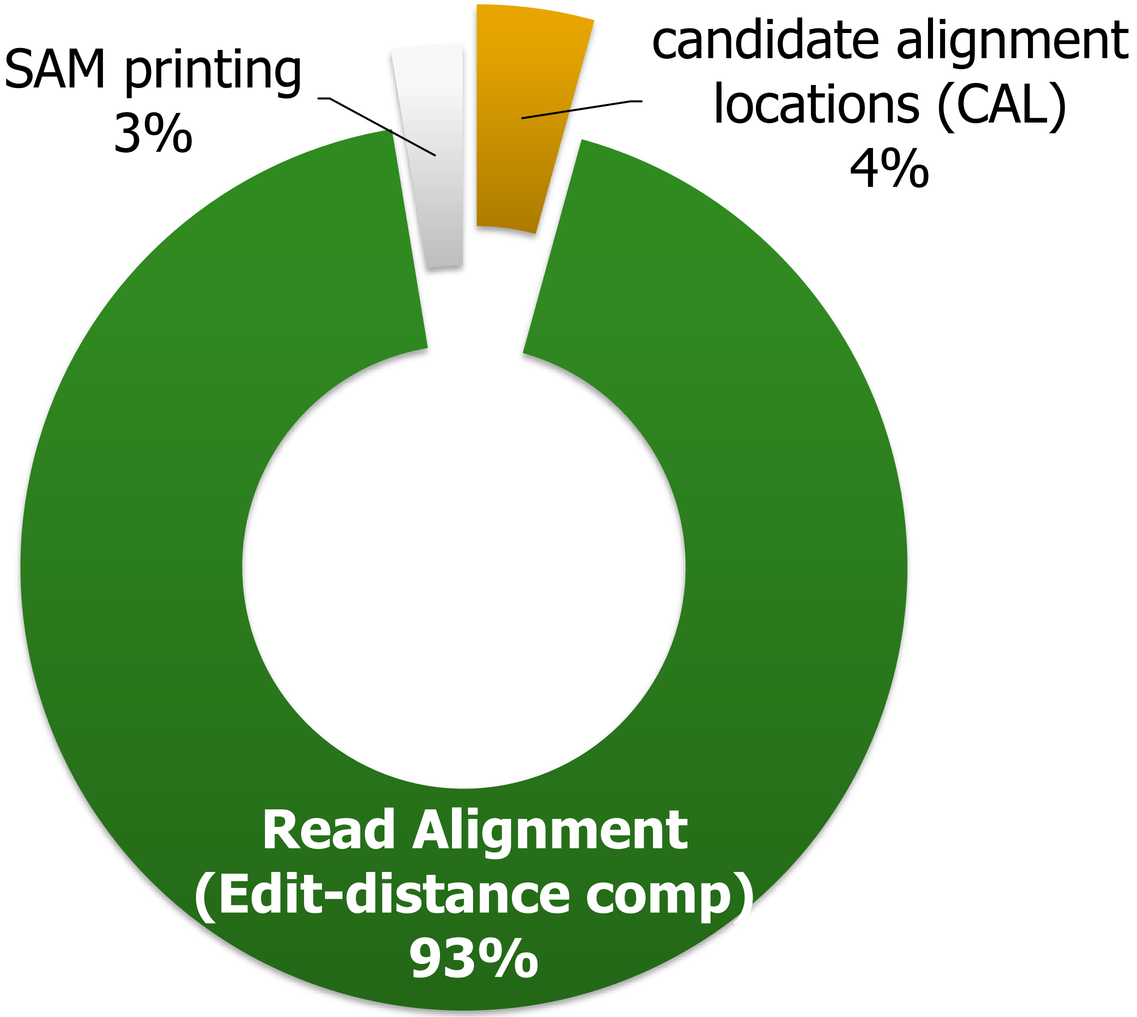
nature genetics

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

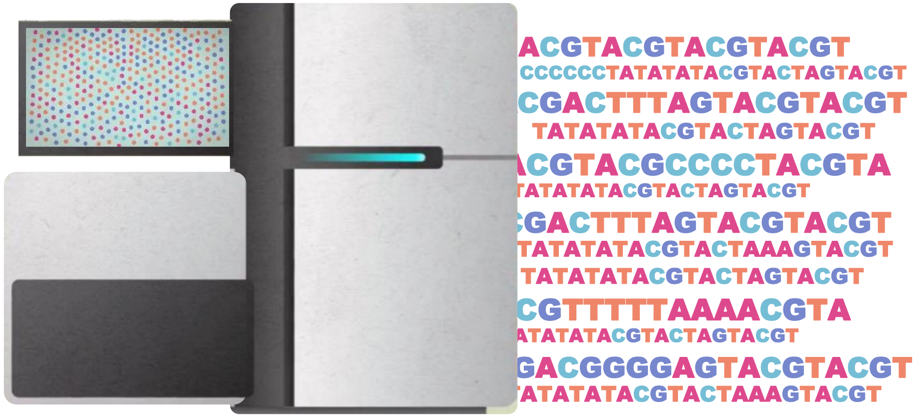
Can Alkan[1,2], Jeffrey M Kidd[1], Tomas Marques-Bonet[1,3], Gozde Aksay[1], Francesca Antonacci[1], Fereydoun Hormozdiari[4], Jacob O Kitzman[1], Carl Baker[1], Maika Malig[1], Onur Mutlu[5], S Cenk Sahinalp[4], Richard A Gibbs[6] & Evan E Eichler[1,2]

Alkan+, **"Personalized copy number and segmental duplication maps using next-generation sequencing"**, Nature Genetics 2009.

Read Mapping Execution Time Breakdown

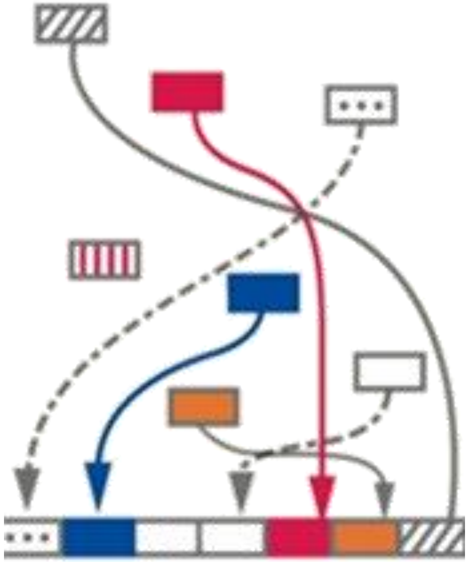


The Read Mapping Bottleneck



Illumina HiSeq4000

300 Million bases/minute



2 Million bases/minute

150X slower

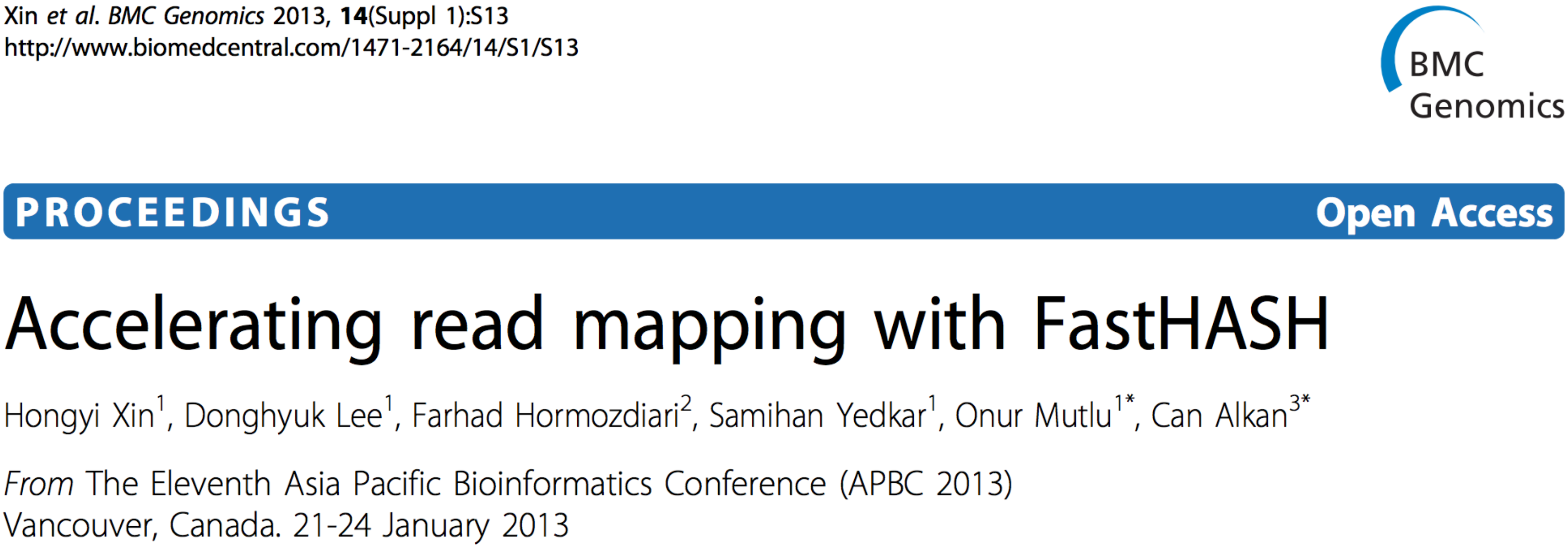
Idea

Filter fast before you align

Minimize costly
"approximate string comparisons"

Our First Filter: Pure Software Approach

- Download the source code and try for yourself
  - Download link to FastHASH



Xin *et al. BMC Genomics* 2013, **14**(Suppl 1):S13
http://www.biomedcentral.com/1471-2164/14/S1/S13

BMC Genomics

PROCEEDINGS | Open Access

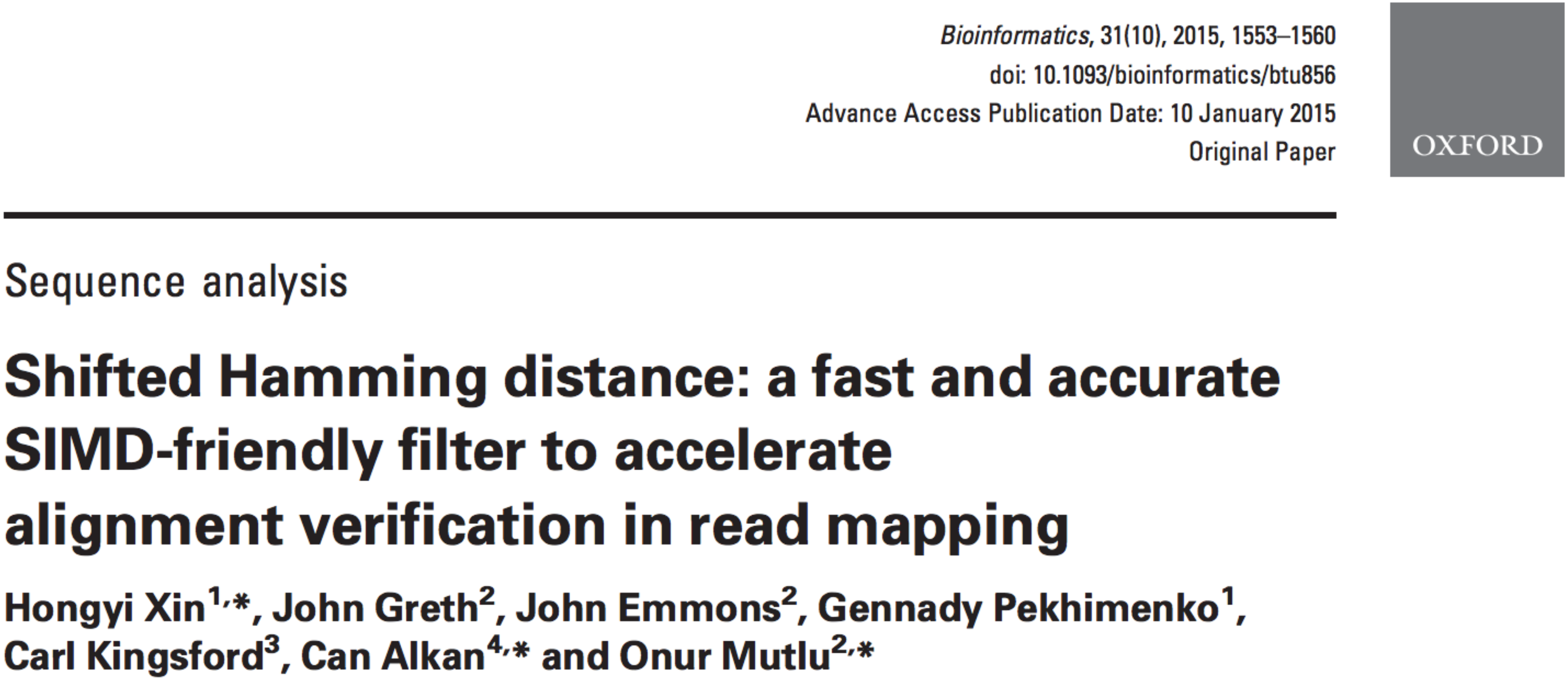
Accelerating read mapping with FastHASH

Hongyi Xin[1], Donghyuk Lee[1], Farhad Hormozdiari[2], Samihan Yedkar[1], Onur Mutlu[1*], Can Alkan[3*]

From The Eleventh Asia Pacific Bioinformatics Conference (APBC 2013)
Vancouver, Canada. 21-24 January 2013

Shifted Hamming Distance: SIMD Acceleration

https://github.com/CMU-SAFARI/Shifted-Hamming-Distance



Bioinformatics, 31(10), 2015, 1553–1560
doi: 10.1093/bioinformatics/btu856
Advance Access Publication Date: 10 January 2015
Original Paper

OXFORD

Sequence analysis

Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping

Hongyi Xin[1,*], John Greth[2], John Emmons[2], Gennady Pekhimenko[1], Carl Kingsford[3], Can Alkan[4,*] and Onur Mutlu[2,*]

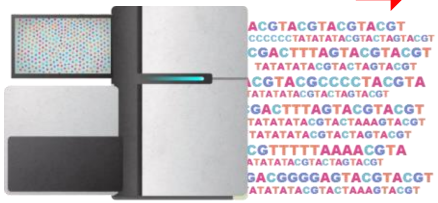
Xin+, **"Shifted Hamming Distance: A Fast and Accurate SIMD-friendly Filter to Accelerate Alignment Verification in Read Mapping", Bioinformatics 2015.**

GateKeeper: FPGA-Based Alignment Filtering

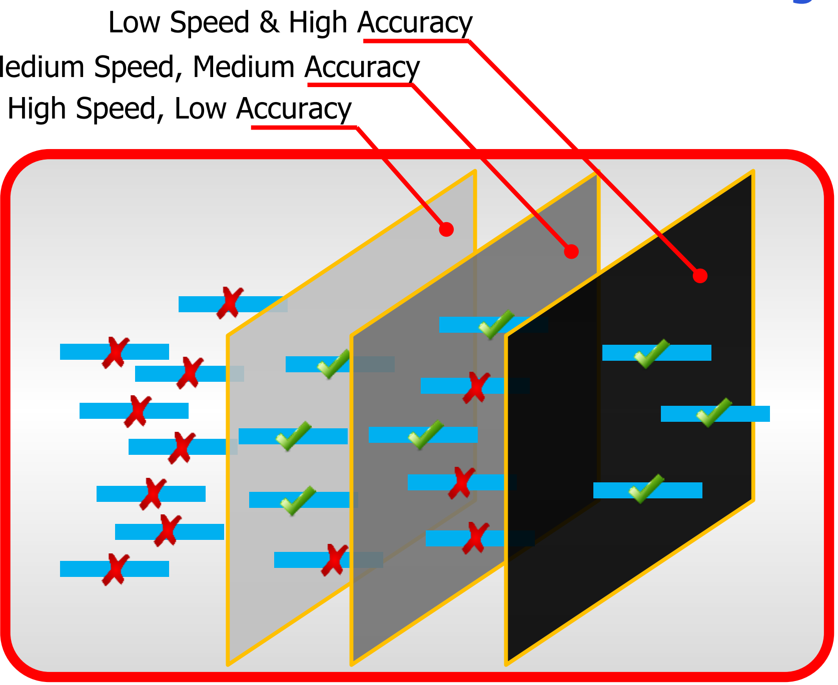
Alignment Filter +

= 1st FPGA-based Alignment Filter.

x10^{12} mappings



Billions of Short Reads



x10^{3} mappings

		C	T	A	T	A	A	T	A	C	G
C	0	1	2								
A	1	0	1	2							
C	2	1	0	1	2						
T		2	1	0	1	2					
A			2	1	2	1	2				
T				2	2	2	1	2			
A					3	2	2	2	2		
T						3	3	3	2	3	
A							4	3	3	2	3
C								4	4	3	2
G									5	4	3

1 **High throughput DNA sequencing (HTS) technologies**

2 **Read Pre-Alignment Filtering**
Fast & Low False Positive Rate

3 **Read Alignment**
Slow & Zero False Positives

GateKeeper: FPGA-Based Alignment Filtering

- Mohammed Alser, Hasan Hassan, Hongyi Xin, Oguz Ergin, Onur Mutlu, and Can Alkan
  "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping"
  Bioinformatics, [published online, May 31], 2017.
  [Source Code]
  [Online link at Bioinformatics Journal]

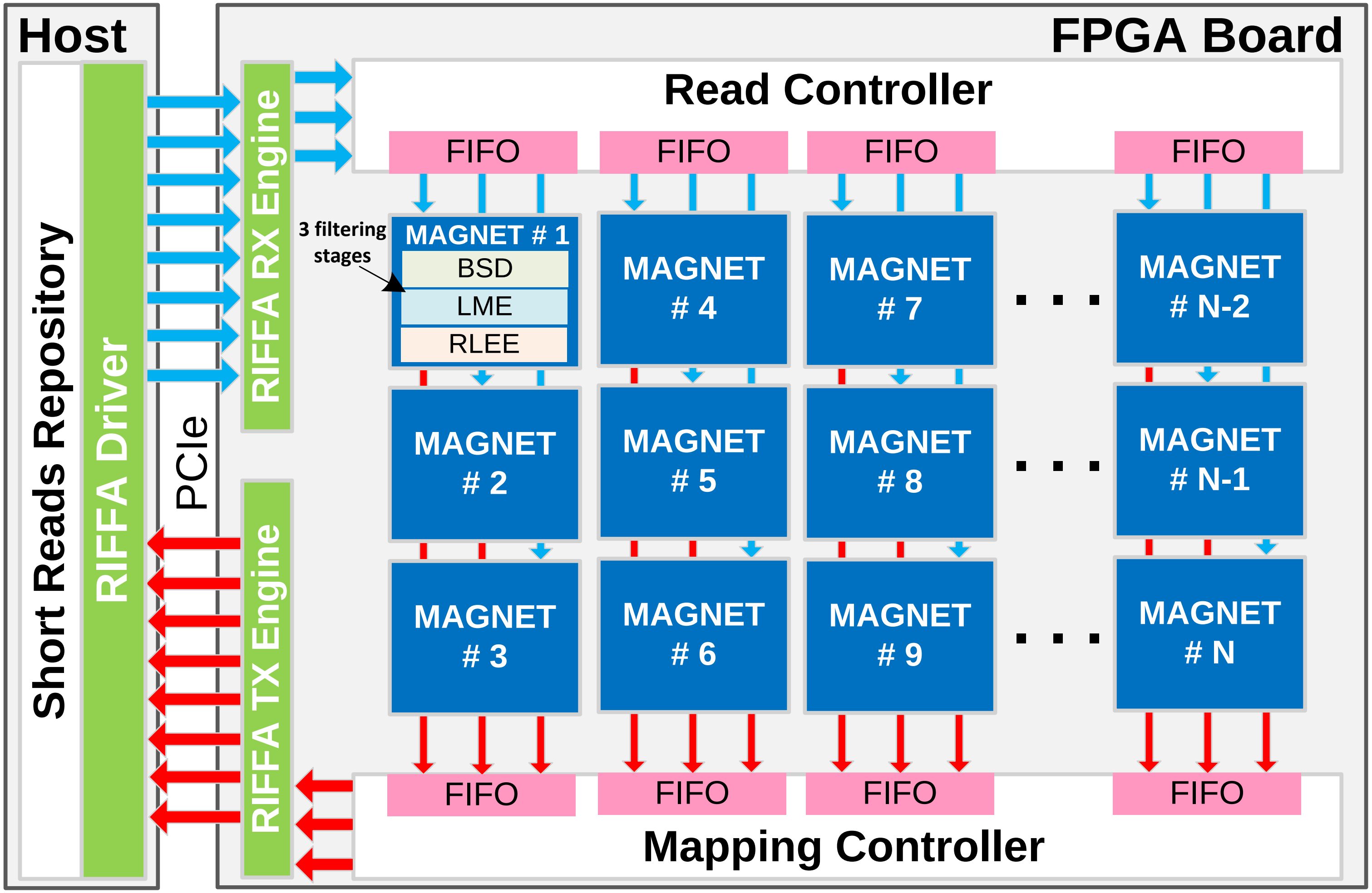
GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping

Mohammed Alser ✉, Hasan Hassan, Hongyi Xin, Oğuz Ergin, Onur Mutlu ✉, Can Alkan ✉

Bioinformatics, Volume 33, Issue 21, 1 November 2017, Pages 3355–3363,
https://doi.org/10.1093/bioinformatics/btx342

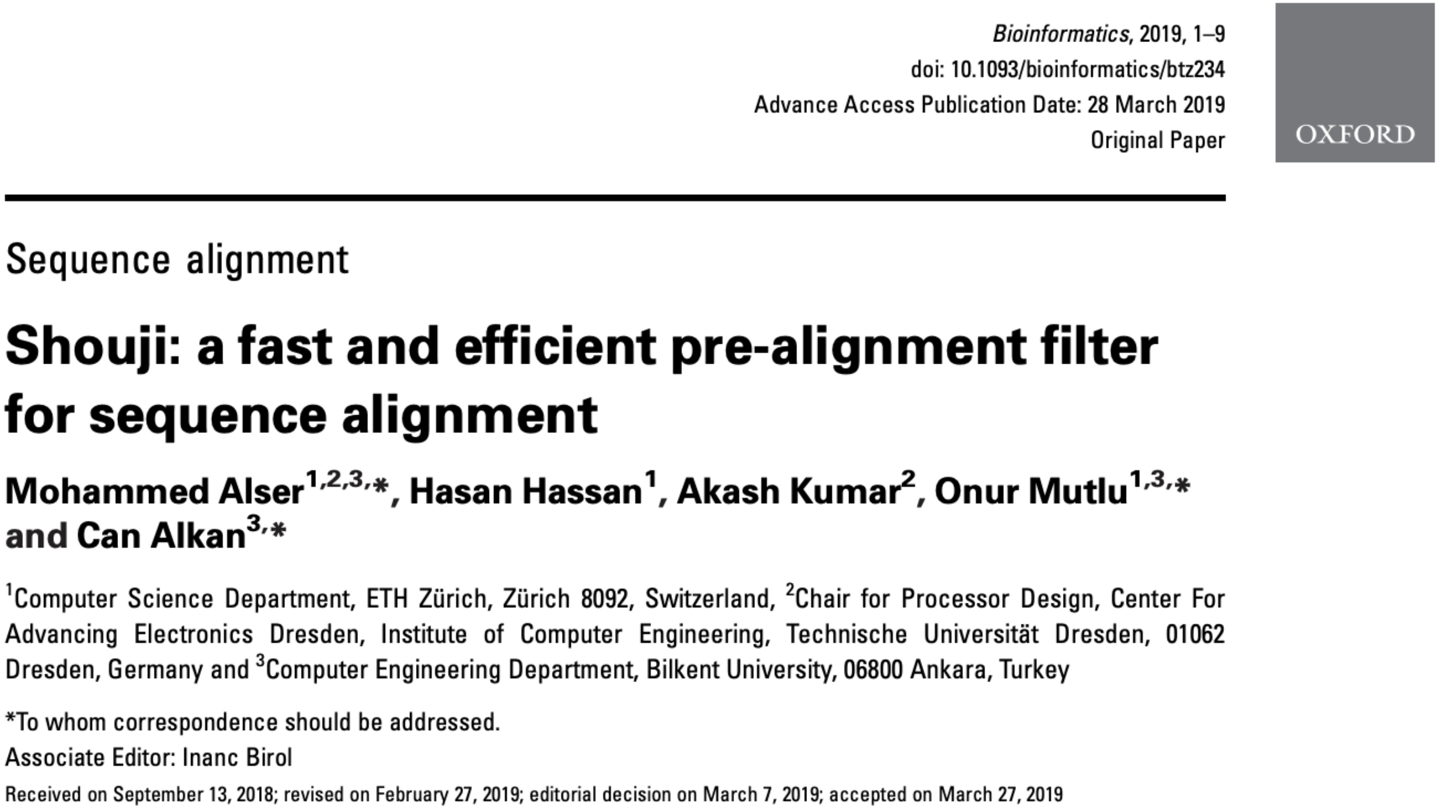
Published: 31 May 2017 **Article history** ▾

MAGNET Accelerator [Alser+, TIR 2017]



Newest Work: Shouji [Alser+, Bioinformatics 2019]

Mohammed Alser, Hasan Hassan, Akash Kumar, Onur Mutlu, and Can Alkan,
"Shouji: A Fast and Efficient Pre-Alignment Filter for Sequence Alignment"
Bioinformatics, [published online, March 28], 2019.
[Source Code]
[Online link at Bioinformatics Journal]



Bioinformatics, 2019, 1–9
doi: 10.1093/bioinformatics/btz234
Advance Access Publication Date: 28 March 2019
Original Paper

Sequence alignment

Shouji: a fast and efficient pre-alignment filter for sequence alignment

Mohammed Alser[1,2,3,*], Hasan Hassan[1], Akash Kumar[2], Onur Mutlu[1,3,*] and Can Alkan[3,*]

[1]Computer Science Department, ETH Zürich, Zürich 8092, Switzerland, [2]Chair for Processor Design, Center For Advancing Electronics Dresden, Institute of Computer Engineering, Technische Universität Dresden, 01062 Dresden, Germany and [3]Computer Engineering Department, Bilkent University, 06800 Ankara, Turkey

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on September 13, 2018; revised on February 27, 2019; editorial decision on March 7, 2019; accepted on March 27, 2019

DNA Read Mapping & Filtering

- Problem: **Heavily bottlenecked by Data Movement**

- GateKeeper FPGA performance limited by DRAM bandwidth [Alser+, Bioinformatics 2017]

- Ditto for SHD on SIMD [Xin+, Bioinformatics 2015]

- Solution: Processing-in-memory can alleviate the bottleneck

- However, we need to design mapping & filtering algorithms to fit processing-in-memory

In-Memory DNA Sequence Analysis

- Jeremie S. Kim, Damla Senol Cali, Hongyi Xin, Donghyuk Lee, Saugata Ghose, Mohammed Alser, Hasan Hassan, Oguz Ergin, Can Alkan, and Onur Mutlu, **"GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies"**
  BMC Genomics, 2018.
  *Proceedings of the 16th Asia Pacific Bioinformatics Conference (**APBC**)*, Yokohama, Japan, January 2018.
  arxiv.org Version (pdf)

GRIM-Filter: Fast seed location filtering in DNA read mapping using processing-in-memory technologies

Jeremie S. Kim[1,6*], Damla Senol Cali[1], Hongyi Xin[2], Donghyuk Lee[3], Saugata Ghose[1], Mohammed Alser[4], Hasan Hassan[6], Oguz Ergin[5], Can Alkan[4*] and Onur Mutlu[6,1*]

From The Sixteenth Asia Pacific Bioinformatics Conference 2018
Yokohama, Japan. 15-17 January 2018

Quick Note: Key Principles and Results

- Two key principles:
  - Exploit the structure of the genome to minimize computation
  - Morph and exploit the structure of the underlying hardware to maximize performance and efficiency

- **Algorithm-architecture co-design** for DNA read mapping
  - **Speeds up** read mapping by **~300X (sometimes more)**
  - **Improves accuracy** of read mapping in the presence of errors

Xin et al., "Accelerating Read Mapping with FastHASH," BMC Genomics 2013.

Xin et al., "Shifted Hamming Distance: A Fast and Accurate SIMD-friendly Filter to Accelerate Alignment Verification in Read Mapping," Bioinformatics 2015.

Alser et al., "GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping," Bioinformatics 2017.

Kim et al., "Genome Read In-Memory (GRIM) Filter," BMC Genomics 2018.

New Genome Sequencing Technologies

Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions

Damla Senol Cali ✉, Jeremie S Kim, Saugata Ghose, Can Alkan, Onur Mutlu

Briefings in Bioinformatics, bby017, https://doi.org/10.1093/bib/bby017

Published: 02 April 2018 **Article history** ▾

Oxford Nanopore MinION

Senol Cali+, "**Nanopore Sequencing Technology and Tools for Genome Assembly: Computational Analysis of the Current State, Bottlenecks and Future Directions**," Briefings in Bioinformatics, 2018.
[Preliminary arxiv.org version]

Future of Genome Sequencing & Analysis

MinION from ONT

SmidgION from ONT

Nanopore Genome Assembly Pipeline

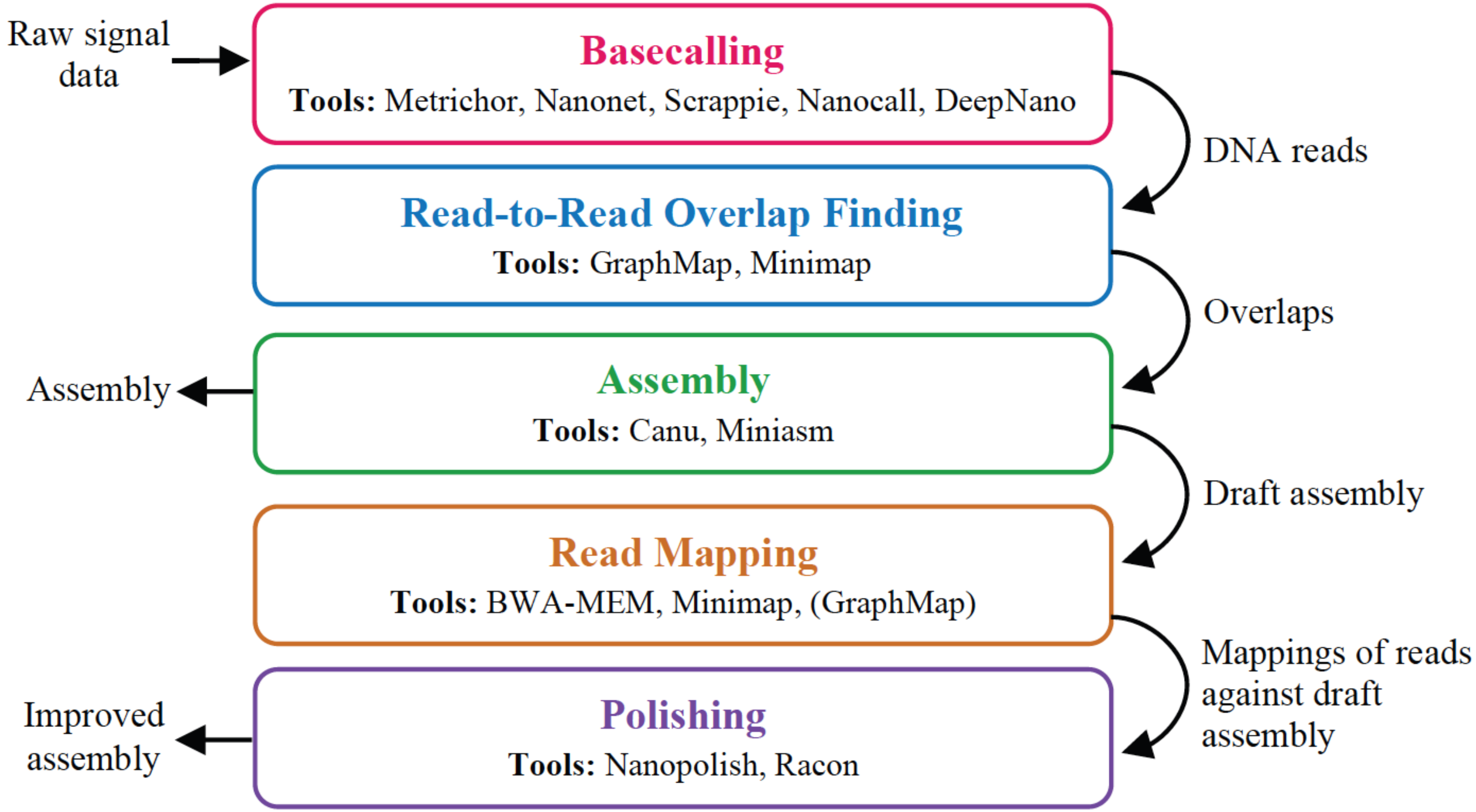


Figure 1. The analyzed genome assembly pipeline using nanopore sequence data, with its five steps and the associated tools for each step.

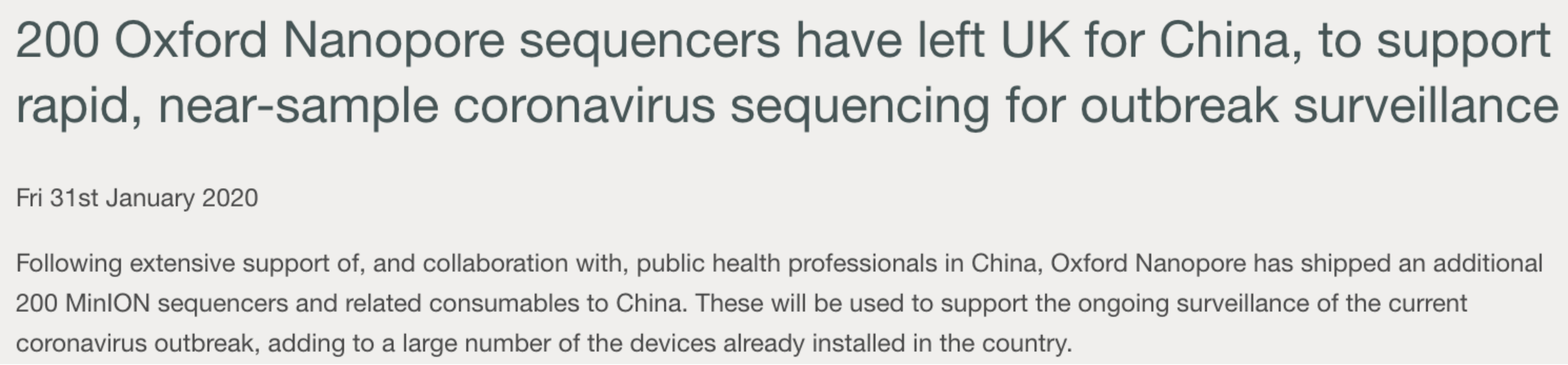
Senol Cali+, "**Nanopore Sequencing Technology and Tools for Genome Assembly,**" Briefings in Bioinformatics, 2018.

Recall Our Dream (from 2007)

- An embedded device that can perform comprehensive genome analysis in real time (within a minute)

- Still a long ways to go
  - Energy efficiency
  - Performance (latency)
  - Security
  - **Huge memory bottleneck**

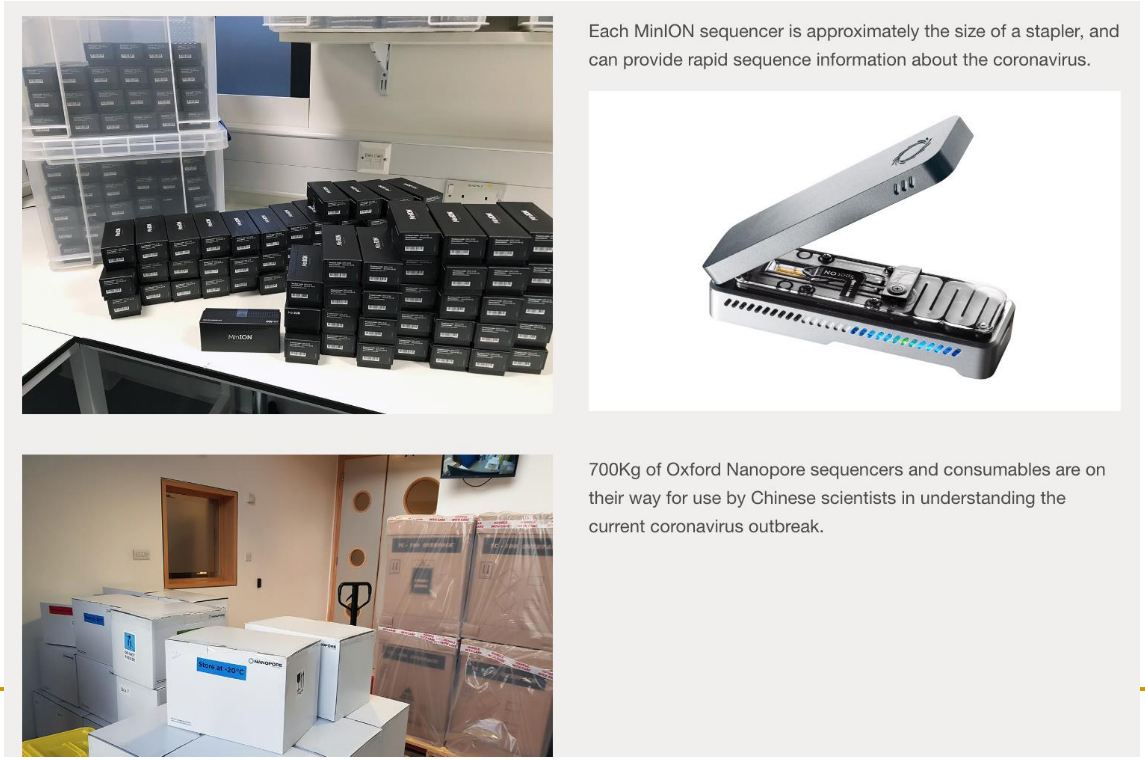
Why Do We Care? An Example from 2020



200 Oxford Nanopore sequencers have left UK for China, to support rapid, near-sample coronavirus sequencing for outbreak surveillance

Fri 31st January 2020

Following extensive support of, and collaboration with, public health professionals in China, Oxford Nanopore has shipped an additional 200 MinION sequencers and related consumables to China. These will be used to support the ongoing surveillance of the current coronavirus outbreak, adding to a large number of the devices already installed in the country.



Each MinION sequencer is approximately the size of a stapler, and can provide rapid sequence information about the coronavirus.

700Kg of Oxford Nanopore sequencers and consumables are on their way for use by Chinese scientists in understanding the current coronavirus outbreak.

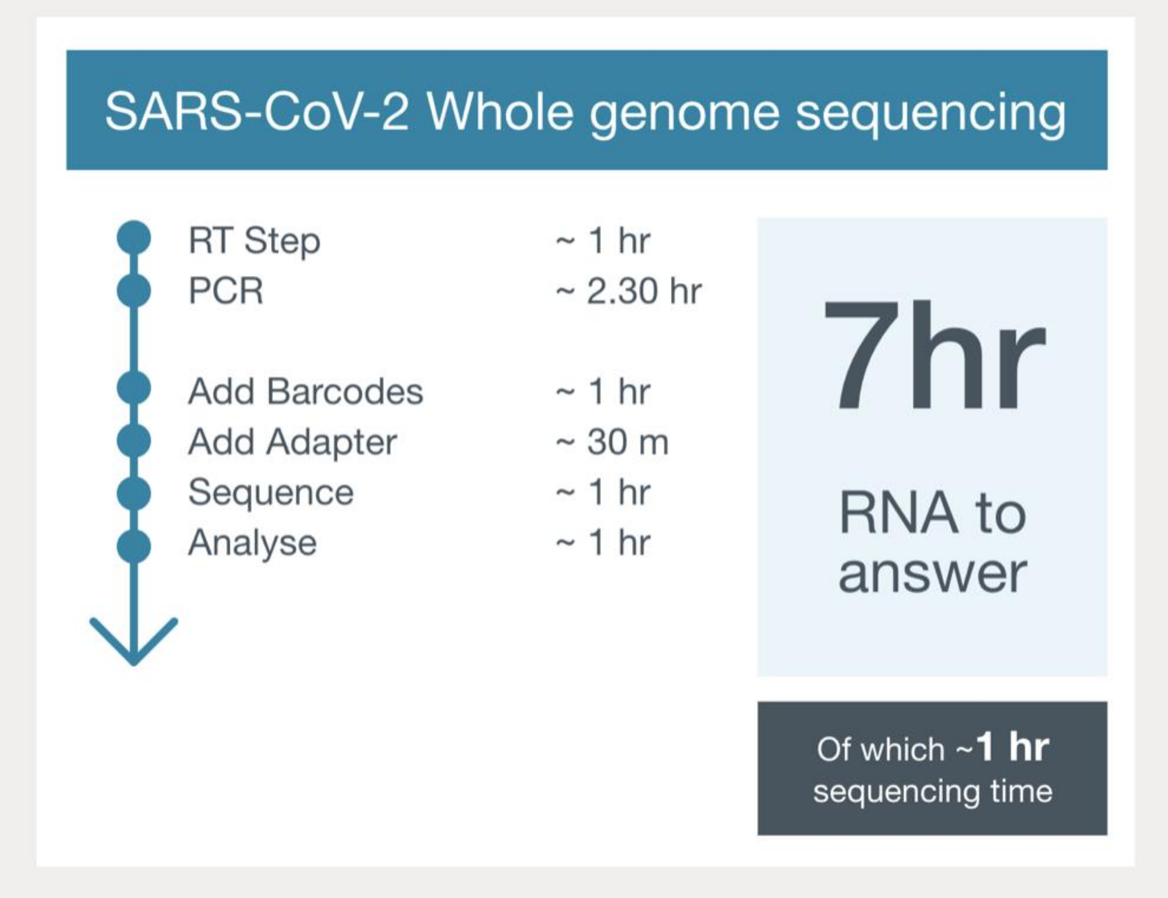
Source: https://nanoporetech.com/about-us/news/200-oxford-nanopore-sequencers-have-left-uk-china-support-rapid-near-sample

Sequencing of COVID-19

- **Whole genome sequencing (WGS) and sequence data analysis are important**
  - To detect the virus from a human sample such as saliva, Bronchoalveolar fluid etc.
  - To understand the sources and modes of transmission of the virus
  - To discover the genomic characteristics of the virus, and compare with better-known viruses (e.g., 02-03 SARS epidemic)
  - To design and evaluate the diagnostic tests

- **Two key areas of COVID-19 genomic research**
  - To sequence the genome of the virus itself, COVID-19, in order to track the mutations in the virus.
  - To explore the genes of infected patients. This analysis can be used to understand why some people get more severe symptoms than others, as well as, help with the development of new treatments in the future.

COVID-19 Nanopore Sequencing (I)



SARS-CoV-2 Whole genome sequencing

RT Step	~ 1 hr
PCR	~ 2.30 hr
Add Barcodes	~ 1 hr
Add Adapter	~ 30 m
Sequence	~ 1 hr
Analyse	~ 1 hr

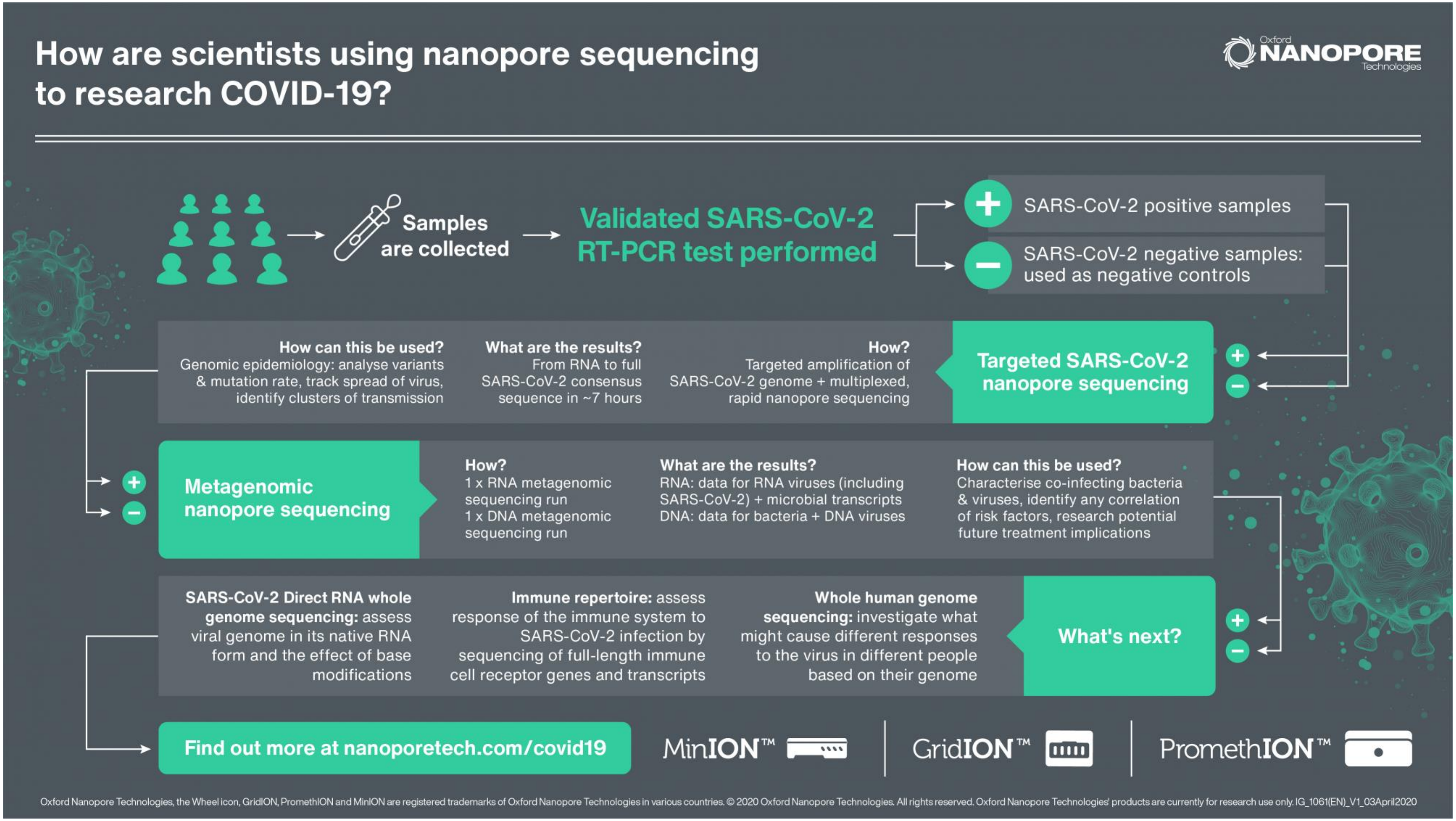
7hr

RNA to answer

Of which ~**1 hr** sequencing time

- From ONT (https://nanoporetech.com/covid-19/overview)

COVID-19 Nanopore Sequencing (II)



How are scientists using nanopore sequencing to research COVID-19?

Samples are collected

Validated SARS-CoV-2 RT-PCR test performed

+ SARS-CoV-2 positive samples

− SARS-CoV-2 negative samples: used as negative controls

Targeted SARS-CoV-2 nanopore sequencing

How? Targeted amplification of SARS-CoV-2 genome + multiplexed, rapid nanopore sequencing

What are the results? From RNA to full SARS-CoV-2 consensus sequence in ~7 hours

How can this be used? Genomic epidemiology: analyse variants & mutation rate, track spread of virus, identify clusters of transmission

Metagenomic nanopore sequencing

How? 1 x RNA metagenomic sequencing run 1 x DNA metagenomic sequencing run

What are the results? RNA: data for RNA viruses (including SARS-CoV-2) + microbial transcripts DNA: data for bacteria + DNA viruses

How can this be used? Characterise co-infecting bacteria & viruses, identify any correlation of risk factors, research potential future treatment implications

What's next?

Whole human genome sequencing: investigate what might cause different responses to the virus in different people based on their genome

Immune repertoire: assess response of the immune system to SARS-CoV-2 infection by sequencing of full-length immune cell receptor genes and transcripts

SARS-CoV-2 Direct RNA whole genome sequencing: assess viral genome in its native RNA form and the effect of base modifications

Find out more at nanoporetech.com/covid19

MinION™ GridION™ PromethION™

Oxford Nanopore Technologies, the Wheel icon, GridION, PromethION and MinION are registered trademarks of Oxford Nanopore Technologies in various countries. © 2020 Oxford Nanopore Technologies. All rights reserved. Oxford Nanopore Technologies' products are currently for research use only. IG_1061(EN)_V1_03April2020

- From ONT (https://nanoporetech.com/covid-19/overview)

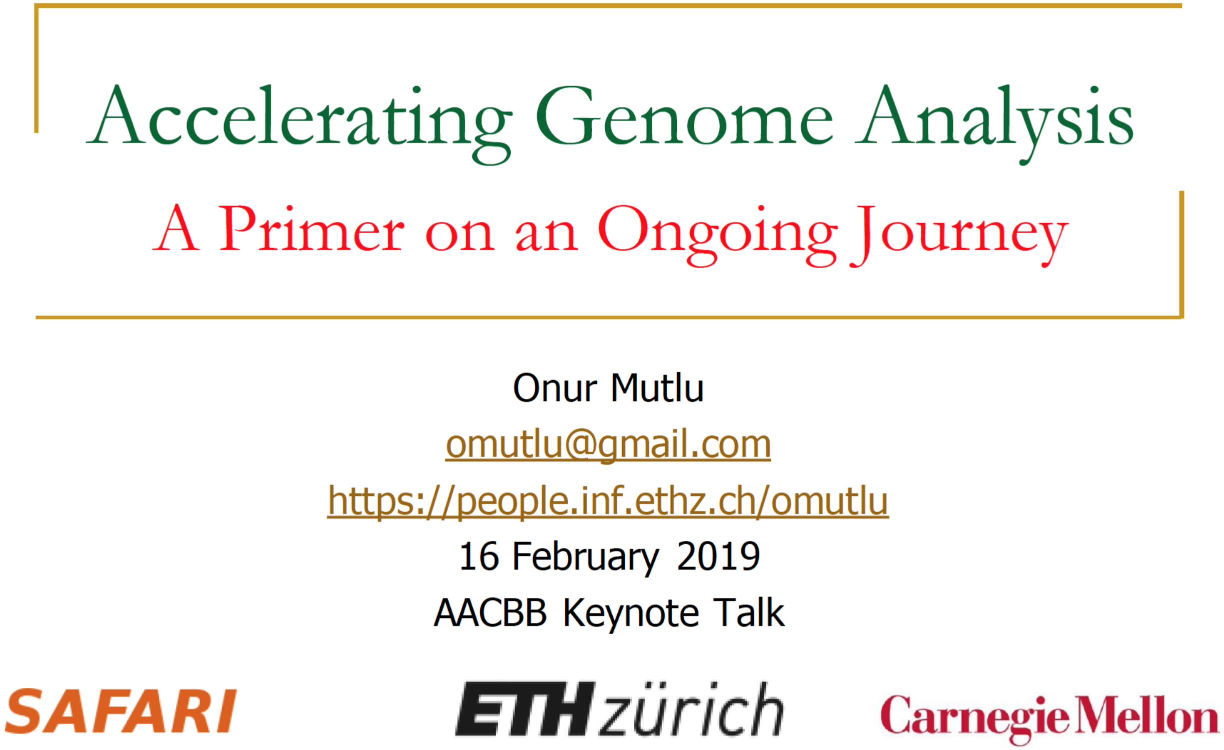
Future of Genome Sequencing & Analysis

MinION from ONT

SmidgION from ONT

More on Genome Analysis: Another Talk

- Onur Mutlu,
  "Accelerating Genome Analysis: A Primer on an Ongoing Journey"
  *Keynote talk at 2nd Workshop on Accelerator Architecture in Computational Biology and Bioinformatics (**AACBB**)*, Washington, DC, USA, February 2019.
  [Slides (pptx)(pdf)]
  [Video]



Accelerating Genome Analysis

A Primer on an Ongoing Journey

Onur Mutlu

omutlu@gmail.com

https://people.inf.ethz.ch/omutlu

16 February 2019

AACBB Keynote Talk

SAFARI ETHzürich Carnegie Mellon

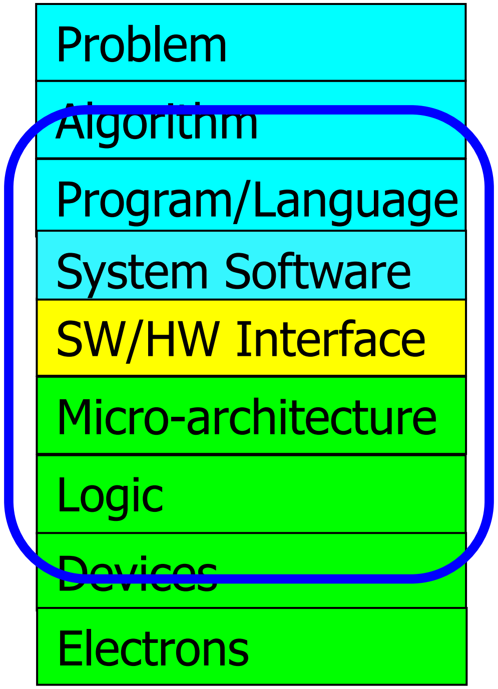
https://www.youtube.com/watch?v=hPnSmfwu2-A

Recall Our Axiom

To achieve the highest energy efficiency and performance:

we must take the expanded view

of computer architecture



Problem

Algorithm

Program/Language

System Software

SW/HW Interface

Micro-architecture

Logic

Devices

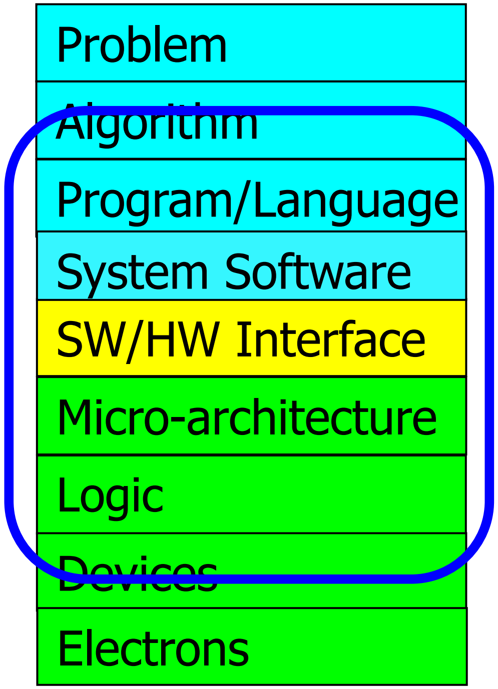
Electrons

Co-design across the hierarchy:
Algorithms to devices

Specialize as much as possible
within the design goals

Our Axiom Applies Well to Genome Analysis

Computer Architecture (expanded view)



Problem

Algorithm

Program/Language

System Software

SW/HW Interface

Micro-architecture

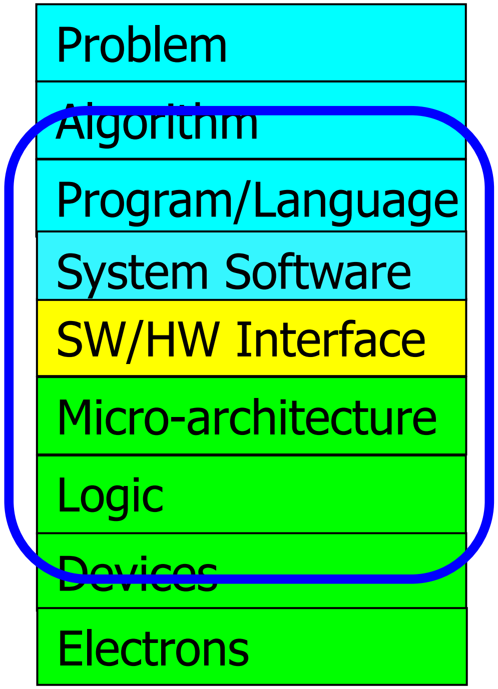
Logic

Devices

Electrons

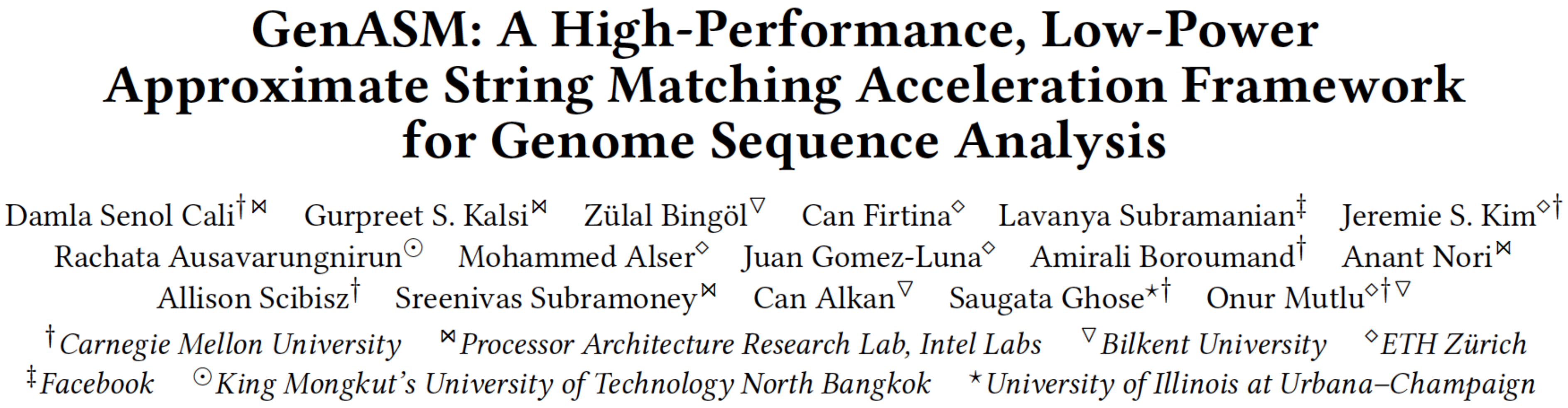
Algorithm-Arch-Device Co-Design is Critical

Computer Architecture (expanded view)



GenASM [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.
  [ARM Research Summit Talk Video (21 minutes)]
  [ARM Research Summit Short Talk Video (15 minutes)]
  [ARM Research Summit Short Talk Video and Q&A (31 minutes)]
  [ARM Research Summit Talk Slides (pptx) (pdf)]
  [ARM Research Summit Short Talk 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*

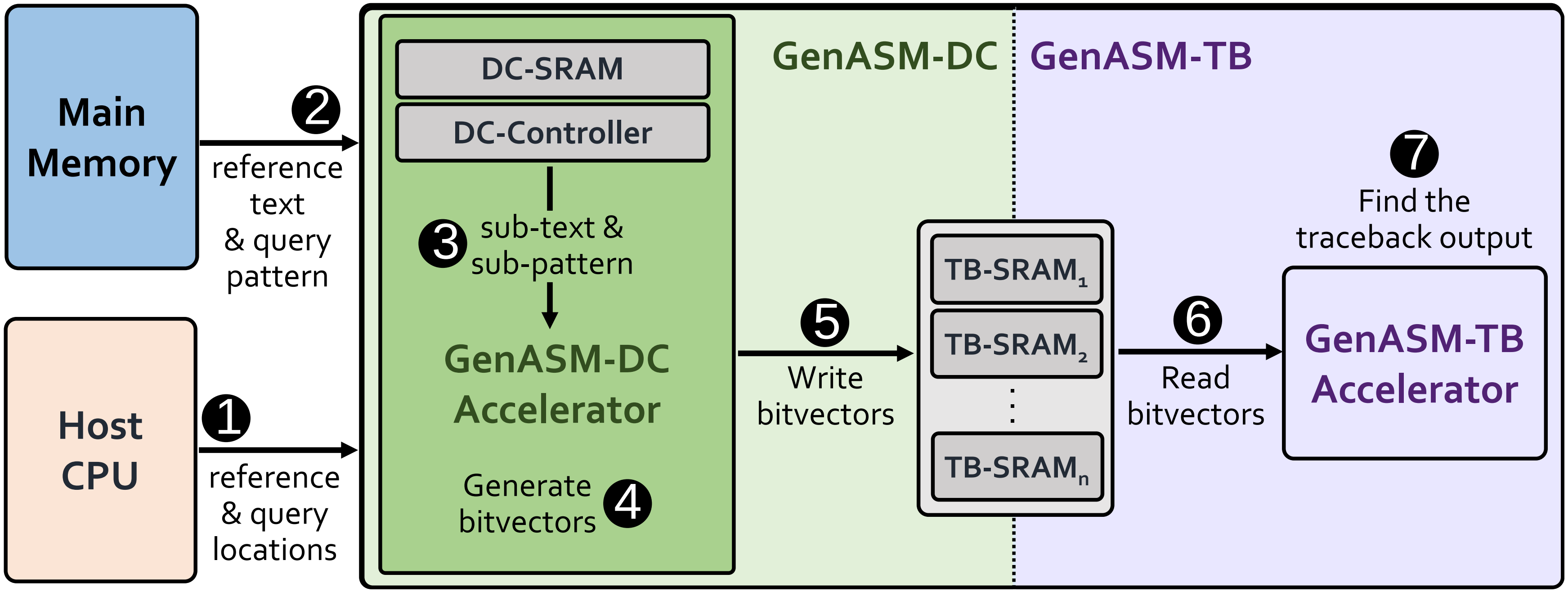
Problem & Our Goal

- ❑ Multiple steps of read mapping require *approximate string matching*
  - ○ ASM enables read mapping to account for sequencing errors and genetic variations in the reads
- ❑ ASM makes up a significant portion of read mapping (more than 70%)
- ❑ One of the major bottlenecks of genome sequence analysis

Our Goal:

Accelerate approximate string matching by designing a fast and flexible framework, which can be used to accelerate *multiple steps* of the genome sequence analysis pipeline

GenASM: Hardware Design

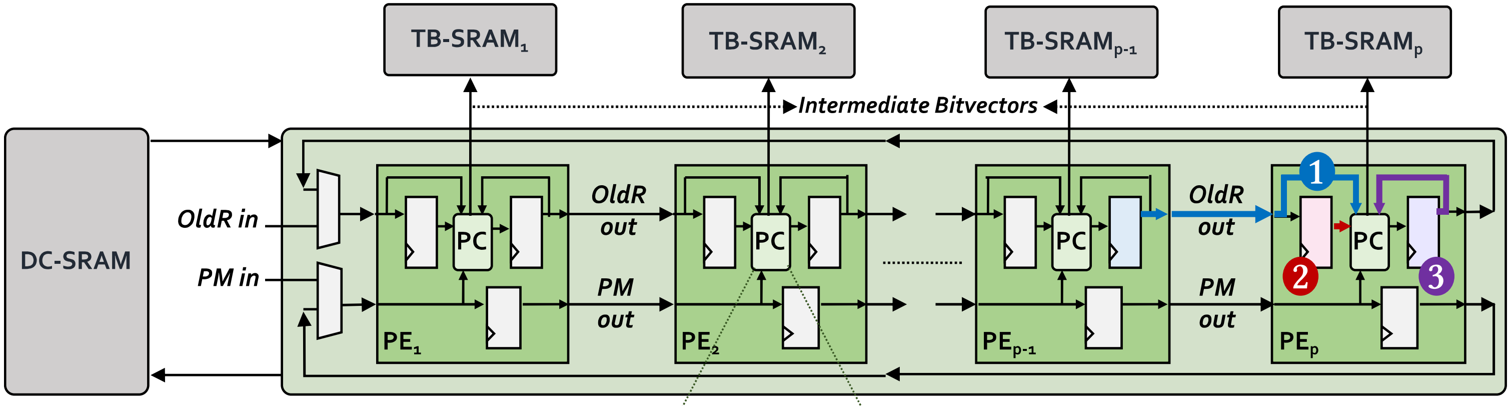


GenASM-DC:
generates bitvectors and performs edit **D**istance **C**alculation

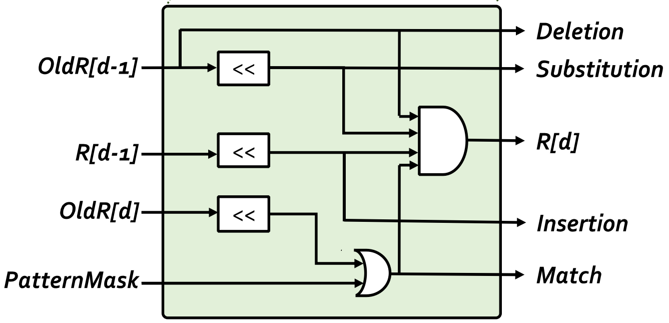
GenASM-TB:
performs **T**race**B**ack and assembles the optimal alignment

GenASM-DC: Hardware Design

- **Linear cyclic systolic array** based accelerator
  - Optimized to reduce memory bandwidth and memory footprint

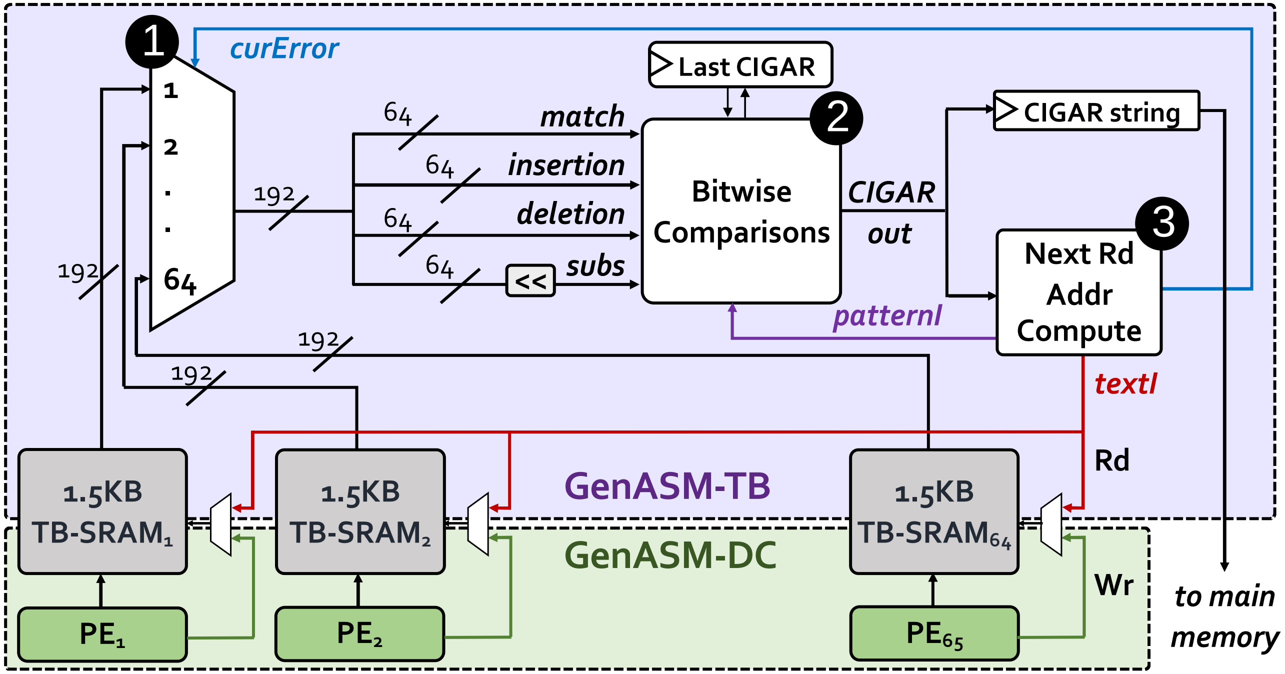


Processing Block (PB), DC-SRAM and TB-SRAMs for each PE



Processing Core (PC)

GenASM-TB: Hardware Design



- Very simple logic:

1) Reads the bitvectors from one of the TB-SRAMs using the computed address

2) Performs the required bitwise comparisons to find the traceback output for the current position

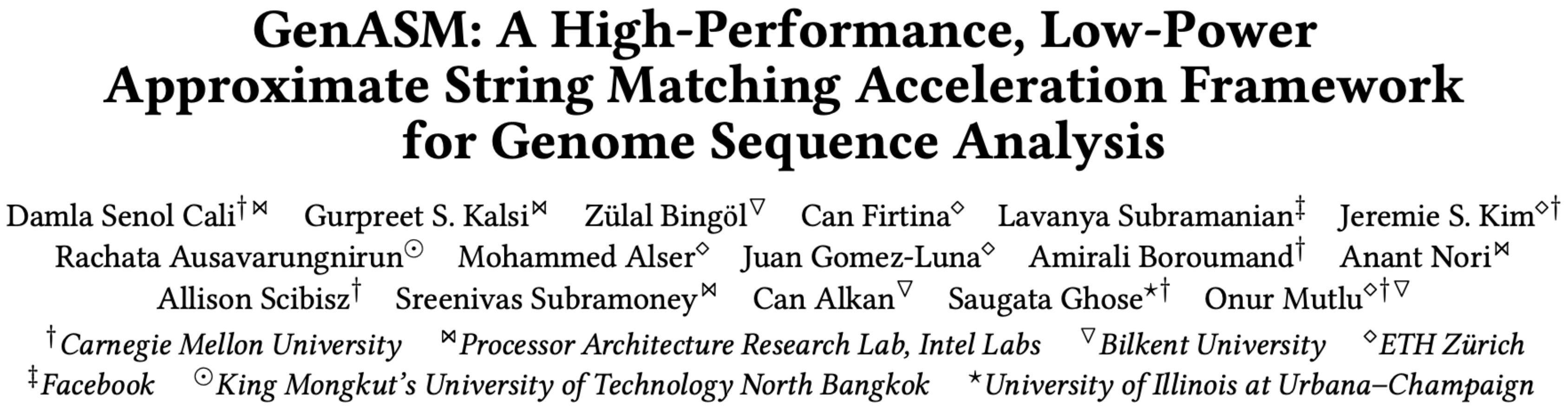
3) Computes the next TB-SRAM address to read the new set of bitvectors

GenASM [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.



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*

Recall Our Dream (from 2007)

- An embedded device that can perform comprehensive genome analysis in real time (within a minute)

- Still a long ways to go
  - Energy efficiency
  - Performance (latency)
  - Security
  - **Huge memory bottleneck**

Four Key Directions

- Fundamentally Secure/Reliable/Safe Architectures

- Fundamentally Energy-Efficient Architectures
  - Memory-centric (Data-centric) Architectures

- Fundamentally Low-Latency and Predictable Architectures

- Architectures for AI/ML, Genomics, Medicine, Health

Memory & Storage

Computer Architecture

Lecture 3a: Introduction to Genome Sequence Analysis

Prof. Onur Mutlu

ETH Zürich

Fall 2020

24 September 2020