Projects & Seminars
Mobile Genomics
Genome Sequencing on Mobile Devices

Prof. Onur Mutlu

Dr. Mohammed Alser

ETH Zürich

Fall 2020

29 September 2020

The Role of This Course

Projects & Seminars: Mobile Genomics

- We will cover the **basics** of **genome analysis** to understand the **speed-accuracy tradeoff** in using computationally-lightweight heuristics versus accurate computationally-expensive algorithms.

- Students will **experimentally** evaluate different heuristic **algorithms** and observe their effect on **the end results**.

- This evaluation will give the students the chance to carry out a **hands-on project** to implement one or more of these heuristic algorithms in **their smartphones** and **help the society by enabling on-site analysis of genomic data**.

Key Objectives

- Multiple components that are aimed at improving students'
  - Basic knowledge in genome analysis (dry lab)
  - Technical skills in genome analysis and computer architecture
  - Critical thinking and analysis
  - Familiarity with key research directions
  - Technical presentation of your project

Key Goal

(Learn how to)

efficiently implement

one of the key steps in genome

analysis on portable devices

Prerequisites of the Course

- No prior knowledge in bioinformatics or genome analysis is required.
- A good knowledge in C programming language and programming is required.
- Interest in making things efficient and solving problems

Course Info: Who Are We?

- **Onur Mutlu**
  - Full Professor @ ETH Zurich ITET (INFK), since September 2015
  - Strecker Professor @ Carnegie Mellon University ECE/CS, 2009-2016, 2016-…
  - PhD from UT-Austin, worked at Google, VMware, Microsoft Research, Intel, AMD
  - https://people.inf.ethz.ch/omutlu/
  - omutlu@gmail.com (Best way to reach me)
  - https://people.inf.ethz.ch/omutlu/projects.htm

- Research and Teaching in:
  - Computer architecture, computer systems, hardware security, bioinformatics
  - Memory and storage systems
  - Hardware security, safety, predictability
  - Fault tolerance
  - Hardware/software cooperation
  - Architectures for bioinformatics, health, medicine
  - …

Course Info: Who Are We?

- Lead Supervisor:
  - Dr. Mohammed Alser

- Supervisors:
  - Dr. Juan Gomez Luna
  - Jeremie Kim
  - Can Firtina

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

Course Requirements and Expectations

- Attendance required for all meetings

- Study the learning materials

- Each student will carry out a hands-on project
  - Build, implement, code, and design with close engagement from the supervisors

- Participation
  - Ask questions, contribute thoughts/ideas
  - Read relevant papers

We will help the projects with good progress to get published in good venues!

Course Website

- https://safari.ethz.ch/projects_and_seminars/doku.php?id=genome_seq_mobile

- Useful information for the course

- Check your email frequently for announcements

- We will also have Piazza for Q&A, announcements, ..

Next Meetings

- We will announce the projects and their descriptions **next week**.
- We will give you a chance to select a project,
- Then, we will have **1-1 meetings** to match your interests, skills, and background with a suitable project.
- It is important that you **study the learning materials** before our next meeting!

WHAT IS GENOME ANALYSIS?

Genome Analysis

Our goal is to find the complete sequence of A, C, G, T’s in DNA (or RNA).

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

>CCTCCTCAGTGCCACCCAGCCCACTGGCAGCTCCCAAACAGGCTCTTATTAAAACACCCTGTTCCCTGCCCCTTG
GAGTGAGGTGTCAAGGACCTAAACTAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCT
TTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTGCTAAACAGCACTTTTTTGACCATTATTTTG
GATCTGAAAGAAATCAAGAATAAATGAAGGACTTGATACATTGGAAGAGGAGAGTCAAGGACCTACAGAAAAAAAA
AAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTT**C**ATGT
CAAGGACCTAATGTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAAGAAAAAGAATTTAAAATTTAAGTAATTC
TTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAGGCCAAGAGTTGCAAAAAAAAAAAAAAG
AAAAAGAAAAGAAAAAGAATTTAAAATTTA**A**GTAATTCTTTGAAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGA
CCTAATGTAGCCAGAATGGTTGTGGGATGGGAGCCTCTGTGGACCGACCAGGTAGCTCTCTTTTCCACACTGTAGT
CTCAAAGCTTCTTCATGTGGTTTCTCTGAGTGAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAG
TAATTCTTTGAAAAAAACTAATTTCTAAGCTT**T**TTCATGTCAAGGACCTAATGTAGCTATACTGAACGTTATCTAGGG
GAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCACTTCCCAGAAAGCTTCTTCATCCGTTTCTCTCCCACA
……

Cracking the 1st 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.

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, ove deserts. Weather map is on Page

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
G guanine
T thymine

By ordering the base units, scientists hope to locate the genes and determine their functions.

The New York Times

A SHARED SUCC

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

By NICHOLAS WADE

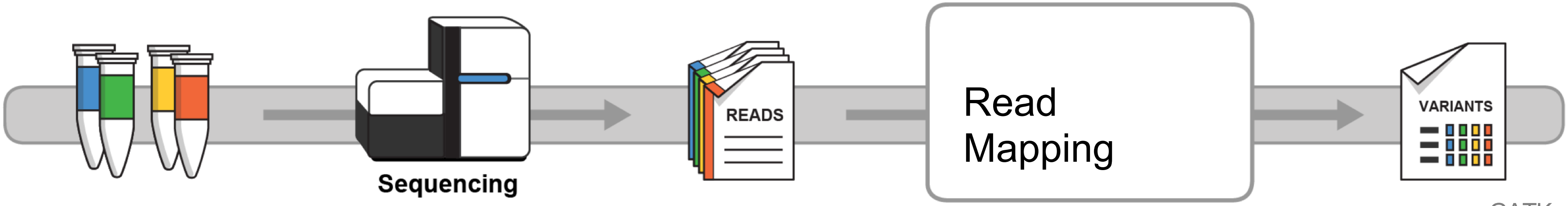
WASHINGTON, June 26 — achievement that represents a nacle of human self-knowledg rival groups of scientists said that they had deciphered the he tary script, the set of instruc that defines the human organis

A C G T 3.2 x10^9 bases

13 years

>3x10^9 $

Vast Improvement in Sequencing



Read Mapping

GATK

CCCCCCTATATATACGTACTAGTACGT

ACGACTTTAGTACGTACGT

TATATATACGTACTAGTACGT

ACGTACG CCCCTACGTA

TATATATACGTACTAGTACGT

ACGACTTTAGTACGTACGT

TATATATACGTACTAAAGTACGT

TATATATACGTACTAGTACGT

ACG TTTTTAAAACGTA

TATATATACGTACTAGTACGT

ACGAC GGGGAGTACGTACGT

$1x10^{12}$ bases*

44 hours*

<1000 $

* NovaSeq 6000

High-Throughput Sequencers

Illumina MiSeq

Pacific
Biosciences
Sequel II

Oxford
Nanopore
PromethION

Oxford Nanopore MinION

Illumina NovaSeq 6000

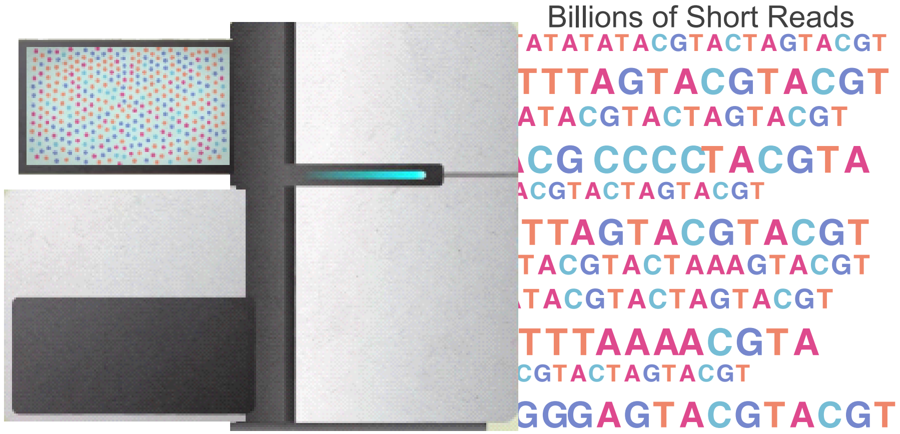
Pacific Biosciences RS II

Oxford
Nanopore
SmidgION

… and more! All produce data with different properties.

How Does HTS Machine Work?

Reads lack information about their order and location (which part of genome they are originated from)



HTS Sequencing Output

Small pieces of a broken vase
short reads

Large pieces of a broken vase
long reads

Which sequencing technology is the best?

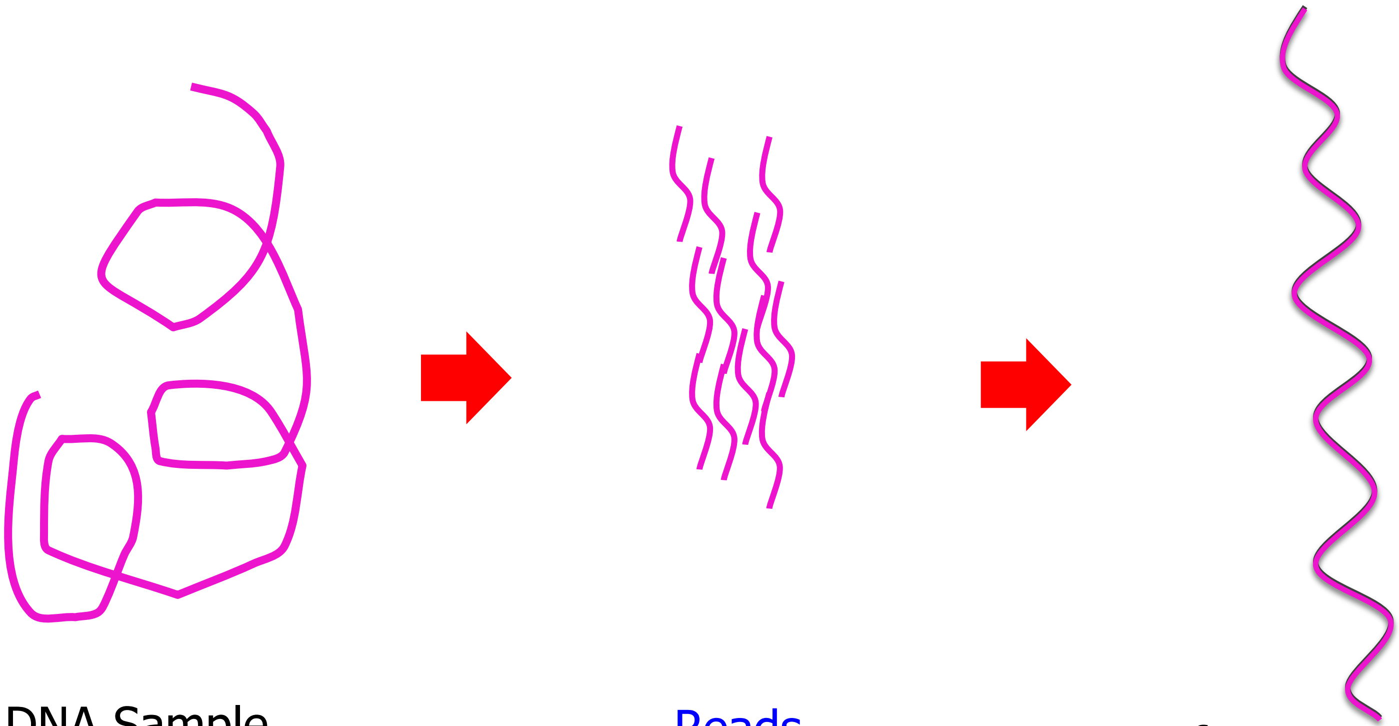
- ❑ 50-300 bp
- ❑ low error rate (~0.1%)

- ❑ 10K-100K bp
- ❑ high error rate (~15%)

Building up the Donor's Genome

Genome Analysis

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



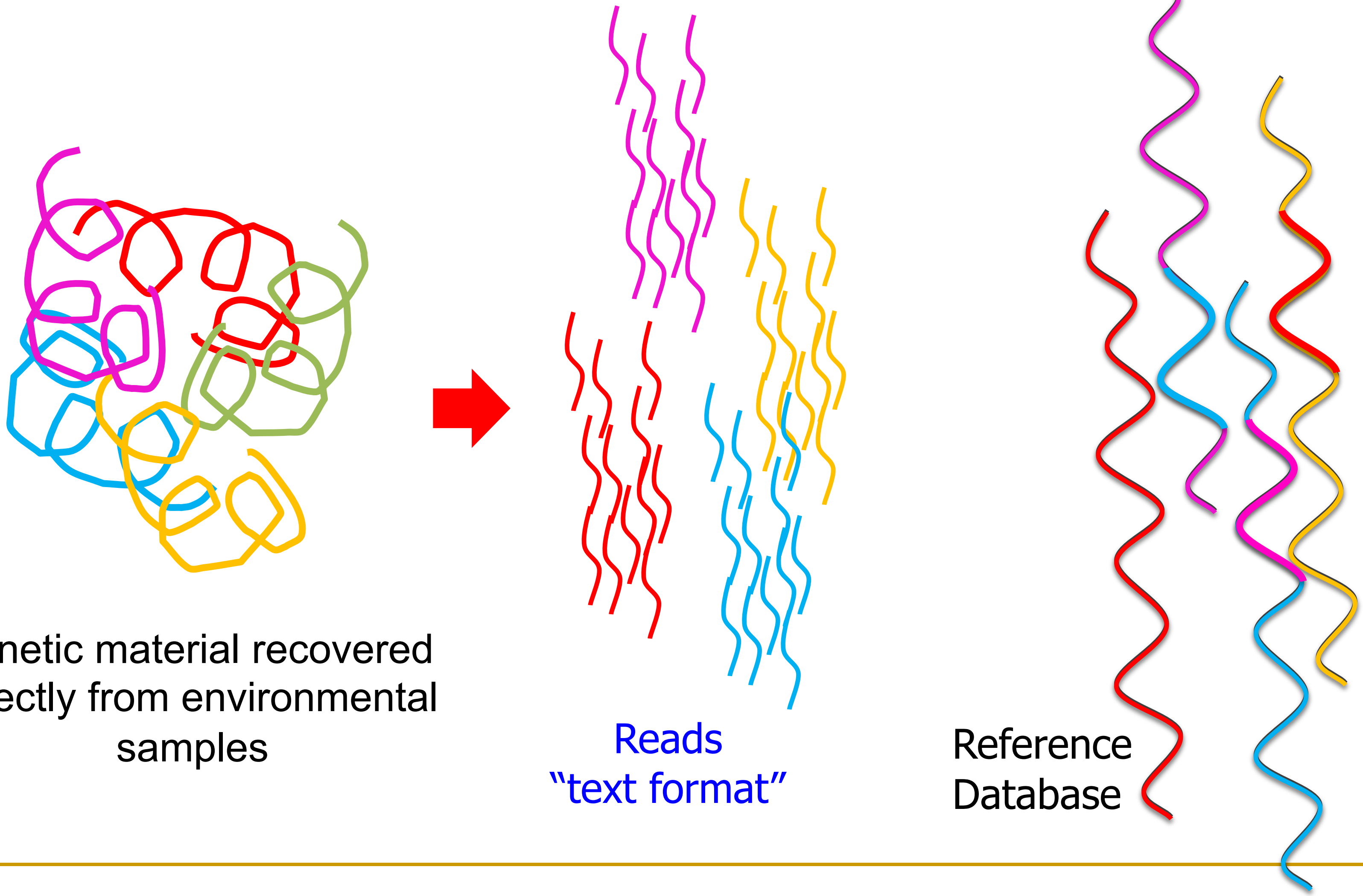
DNA Sample
"chemical format"

Reads
"text format"

Reference genome
Subject genome
"text format"

Metagenomics Analysis

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



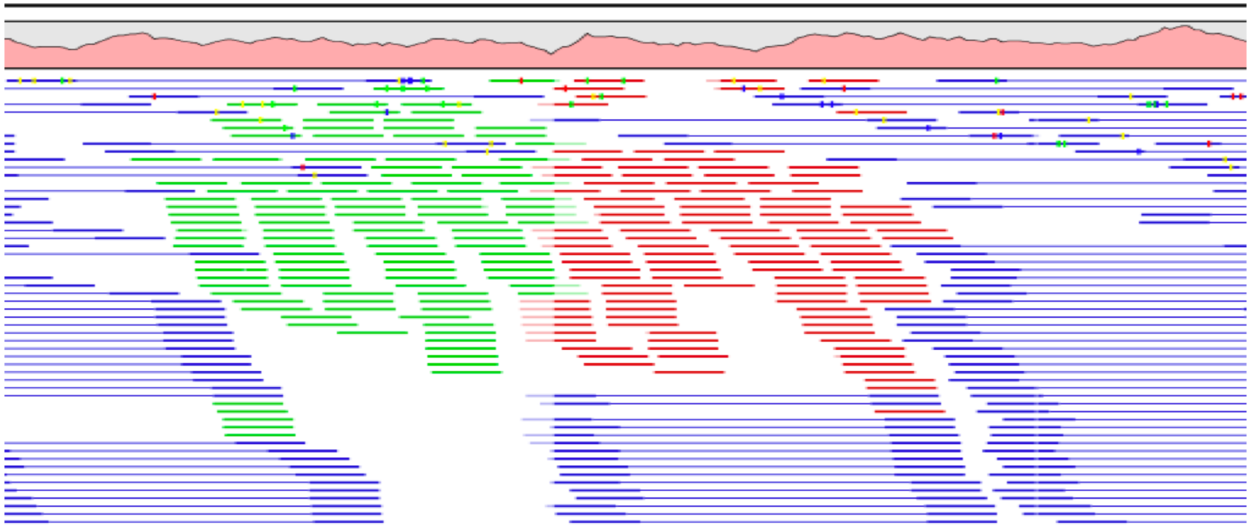
genetic material recovered directly from environmental samples

Reads
"text format"

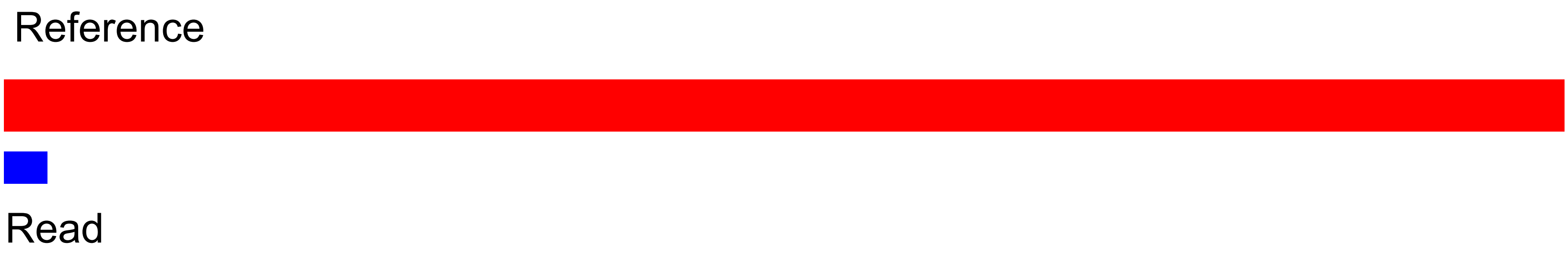
Reference Database

Challenges in Read Mapping

- Need to find many mappings of each read
- Need to tolerate small variances/errors in each read
- Need to map each read very fast (i.e., performance is important, life critical in some cases)



Read Mapping: A Brute Force Algorithm



Reference

Read

Very Expensive!

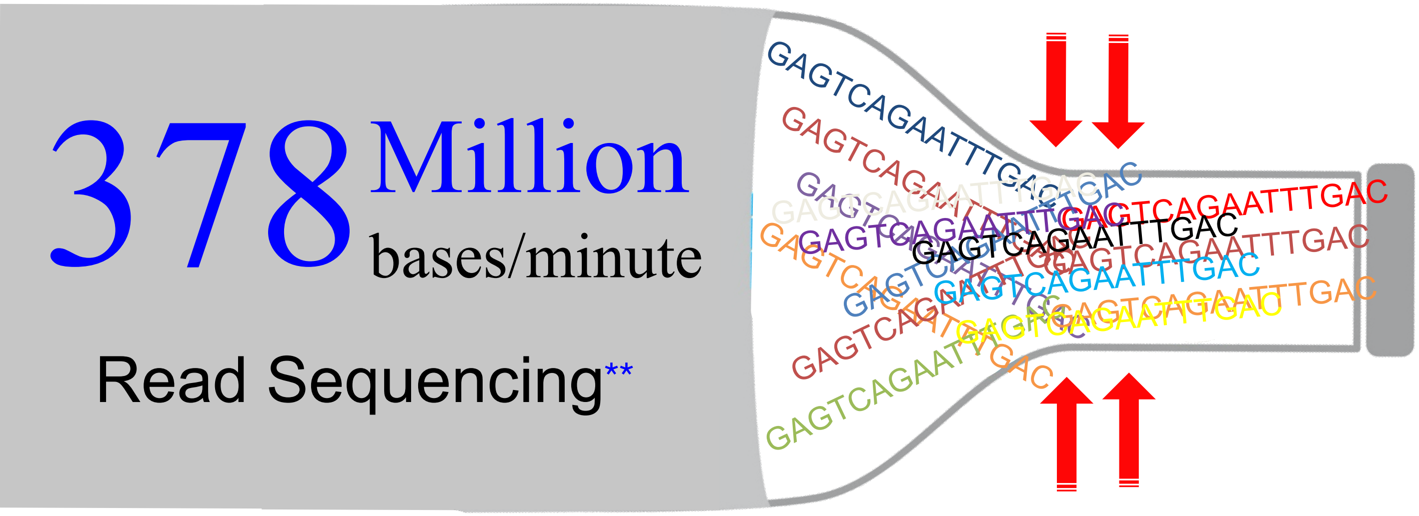
$O(m^2kn)$

m: read length

k: no. of reads

n: reference genome length

Bottlenecked in Read Alignment!!



378 Million bases/minute

Read Sequencing**

2 Million bases/minute

Read Mapping*

150x slower

* BWA-MEM

** NovaSeq 6000, MinION

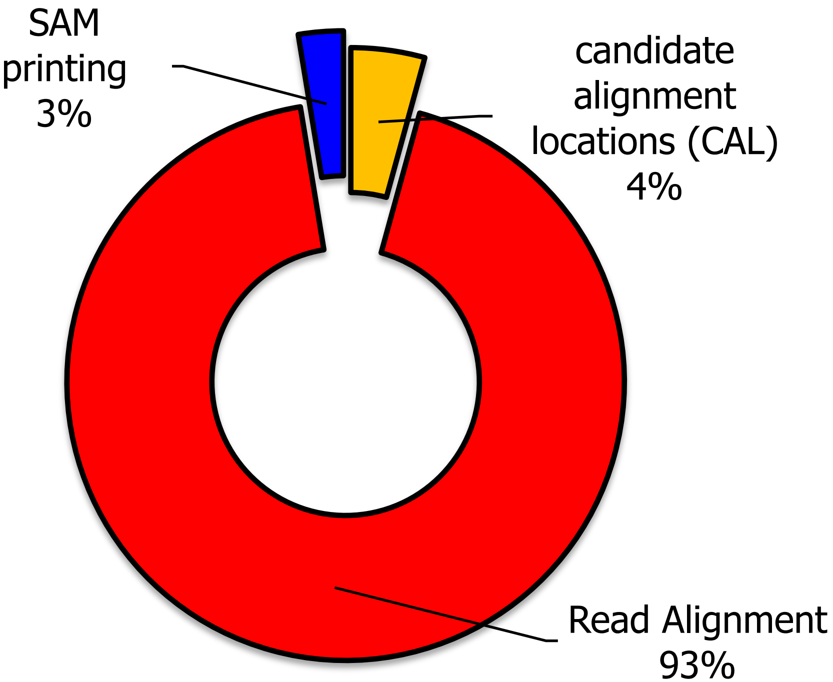
ACCELERATING GENOME ANALYSIS

What Makes Read Mapper Slow?

Key Observation # 1

70-90%

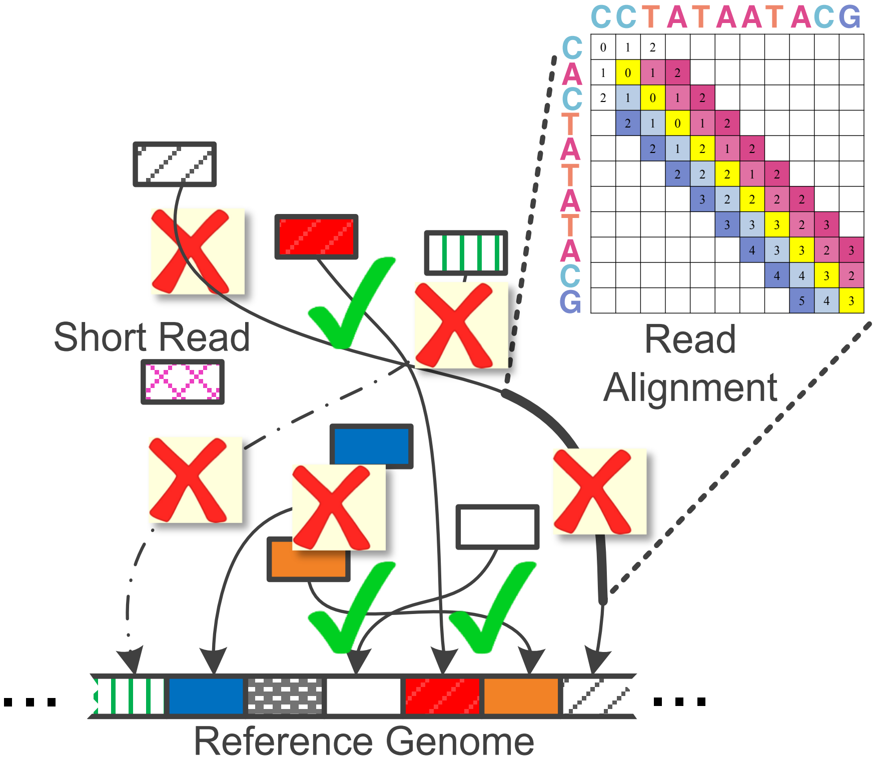
of the read mapper's execution time is spent in read alignment.



Alser et al, Bioinformatics (2017)

What Makes Read Mapper Slow? (cont’d)

Key Observation # 2



	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

98%

of candidate locations have high dissimilarity with a given read.

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

What Makes Read Mapper Slow? (cont’d)

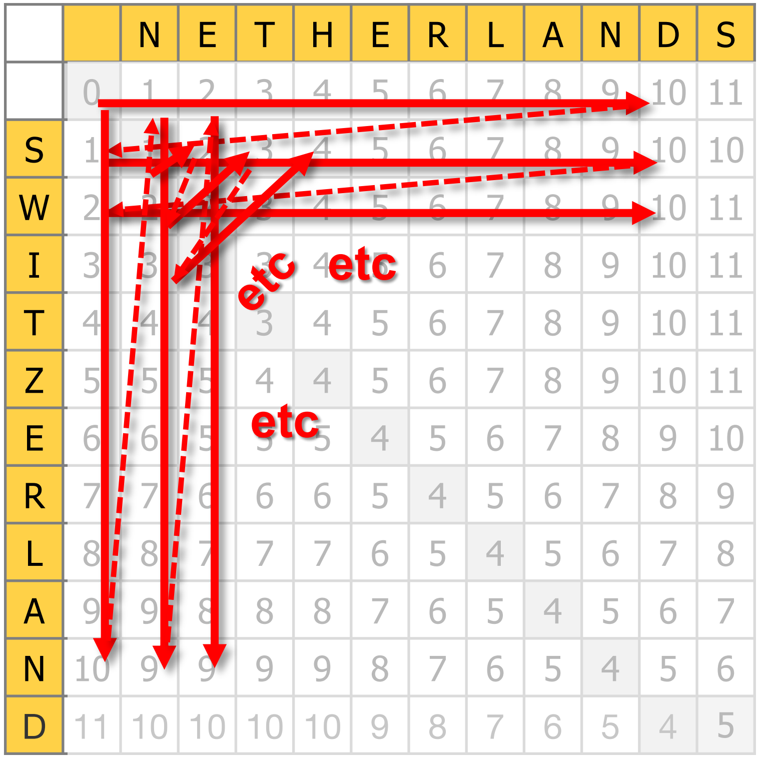
Key Observation # 3

- 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	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

What Makes Read Mapper Slow? (cont'd)

Key Observation # 3

- 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

Finding SNPs Associated with Complex Trait

Different individuals

	SNP1		SNP2		Blood Pressure
...ACATG	C	CGACATTTCATA	G	GCC...	180
...ACATG	C	CGACATTTCATA	A	GCC...	175
...ACATG	C	CGACATTTCATA	G	GCC...	170
...ACATG	C	CGACATTTCATA	A	GCC...	165
...ACATG	C	CGACATTTCATA	G	GCC...	160
...ACATG	C	CGACATTTCATA	G	GCC...	145
...ACATG	C	CGACATTTCATA	A	GCC...	140
...ACATG	C	CGACATTTCATA	A	GCC...	130
...ACATG	T	CGACATTTCATA	G	GCC...	120
...ACATG	T	CGACATTTCATA	A	GCC...	120
...ACATG	T	CGACATTTCATA	G	GCC...	115
...ACATG	T	CGACATTTCATA	A	GCC...	110
...ACATG	T	CGACATTTCATA	G	GCC...	110
...ACATG	T	CGACATTTCATA	A	GCC...	110
...ACATG	T	CGACATTTCATA	G	GCC...	105
...ACATG	T	CGACATTTCATA	A	GCC...	100

Eleazar Eskin: Discovering the Causal Variants Involved in GWAS Studies, CGSI 2018, UCLA

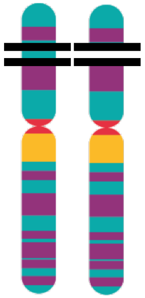
Mirror Phenotypes of 593 Kb CNVs

AUTISM
Weiss, *N Eng J Med* 2008
Deletion of 593 kb

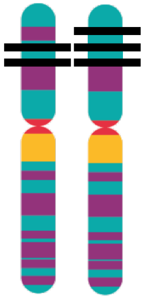
SCHIZOPHRENIA
McCarthy, *Nat Genet* 2009
Duplication of 593 kb

OBESITY
Walters, *Nature* 2010
Deletion of 593 kb

UNDERWEIGHT
Jacquemont, *Nature* 2011
Duplication of 593 kb



Deletion in the short arm
of chromosome 16 (16p11.2)

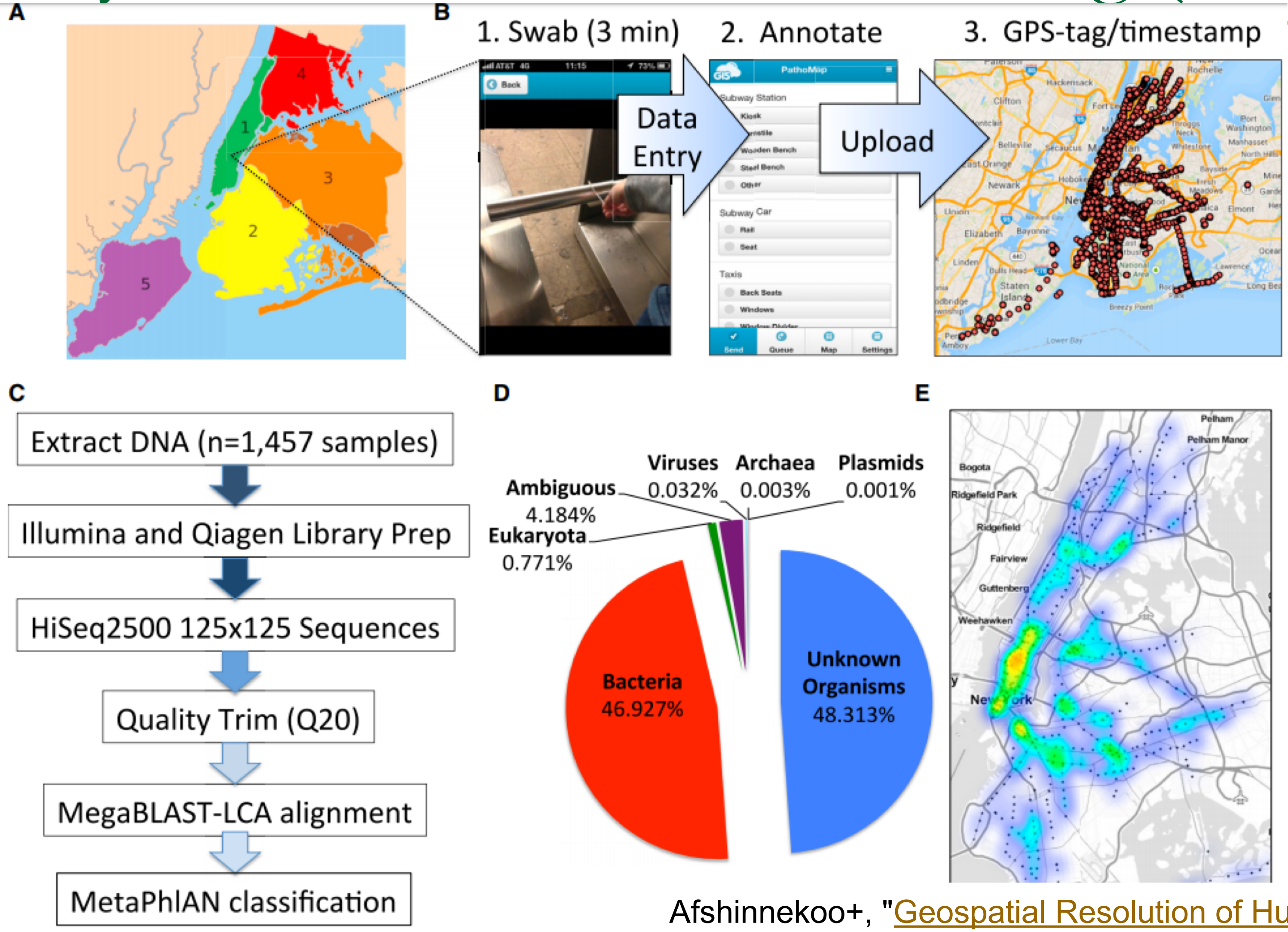


Duplication in the short arm
of chromosome 16 (16p11.2)

City-Scale Microbiome Profiling

https://gothamist.com/news/the-subway-is-probably-bubonic-plague-free-after-all

City-Scale Microbiome Profiling (cont'd)



Afshinnekoo+, "Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics", Cell Systems, 2015

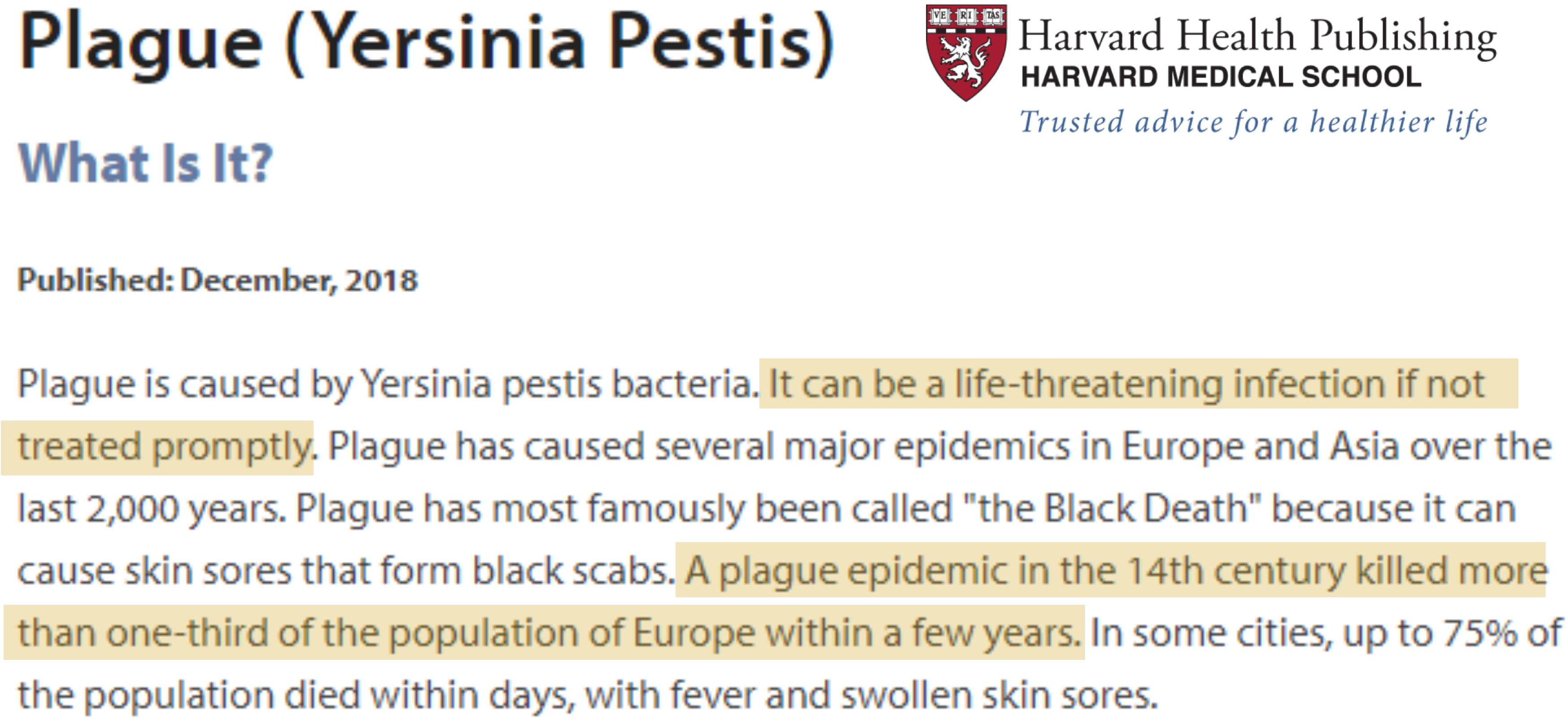
Figure 1. The Metagenome of New York City

(A) The five boroughs of NYC include (1) Manhattan (green), (2) Brooklyn (yellow), (3) Queens (orange), (4) Bronx (red), (5) Staten Island (lavender).

(B) The collection from the 466 subway stations of NYC across the 24 subway lines involved three main steps: (1) collection with Copan Elution swabs, (2) data entry into the database, and (3) uploading of the data. An image is shown of the current collection database, taken from http://pathomap.giscloud.com.

(C) Workflow for sample DNA extraction, library preparation, sequencing, quality trimming of the FASTQ files, and alignment with MegaBLAST and MetaPhlAn to discern taxa present.

Plague in New York Subway System?



Plague (Yersinia Pestis)

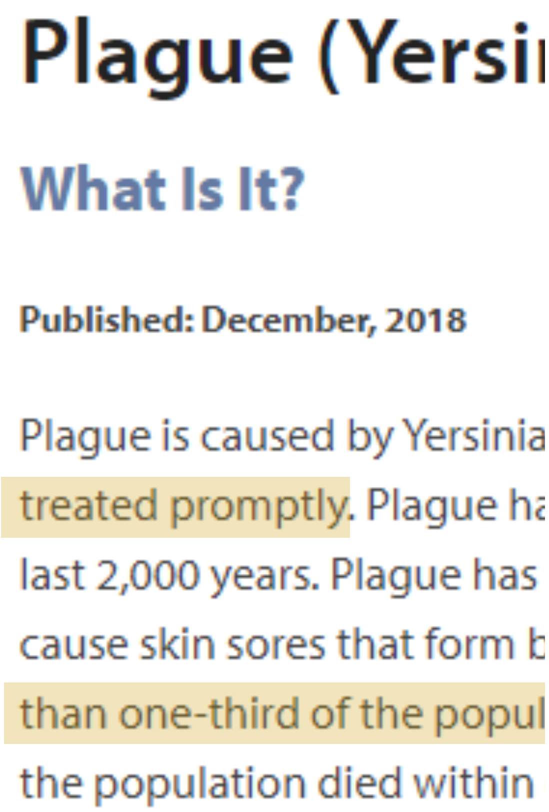
Harvard Health Publishing
HARVARD MEDICAL SCHOOL
Trusted advice for a healthier life

What Is It?

Published: December, 2018

Plague is caused by Yersinia pestis bacteria. It can be a life-threatening infection if not treated promptly. Plague has caused several major epidemics in Europe and Asia over the last 2,000 years. Plague has most famously been called "the Black Death" because it can cause skin sores that form black scabs. A plague epidemic in the 14th century killed more than one-third of the population of Europe within a few years. In some cities, up to 75% of the population died within days, with fever and swollen skin sores.

Plague in New York Subway System?



Plague (Yersi

What Is It?

Published: December, 2018

Plague is caused by Yersinia

treated promptly. Plague ha

last 2,000 years. Plague has

cause skin sores that form b

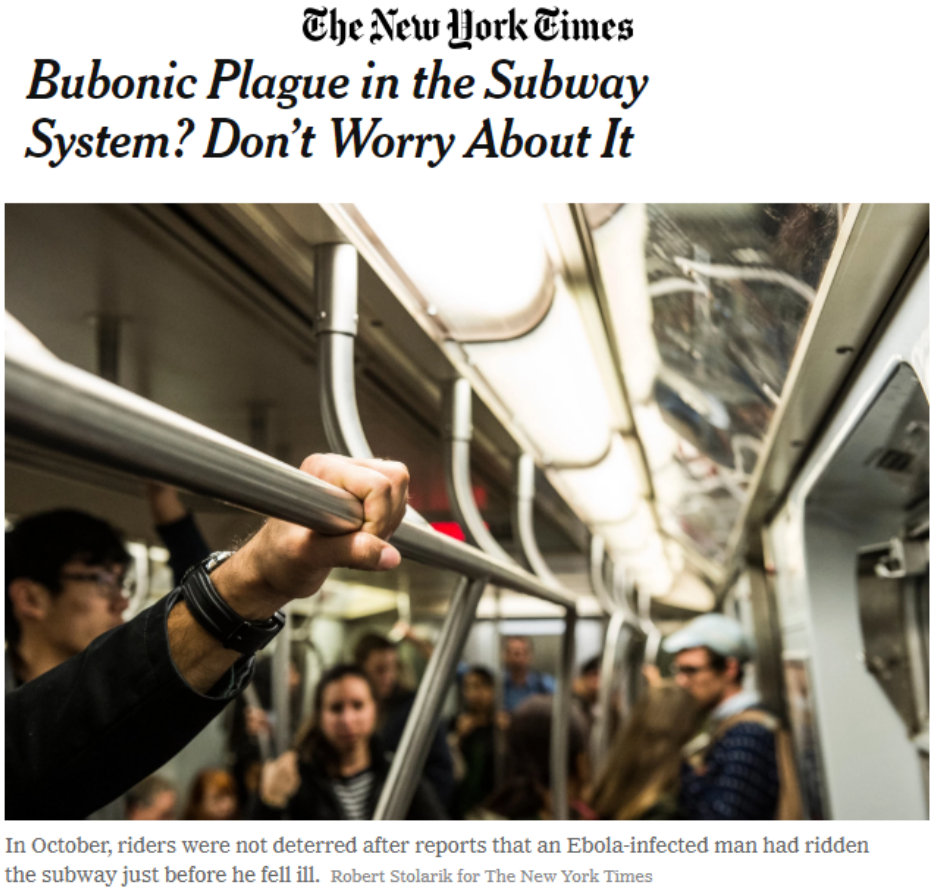
than one-third of the popul

the population died within

he

ore

s of



The New York Times

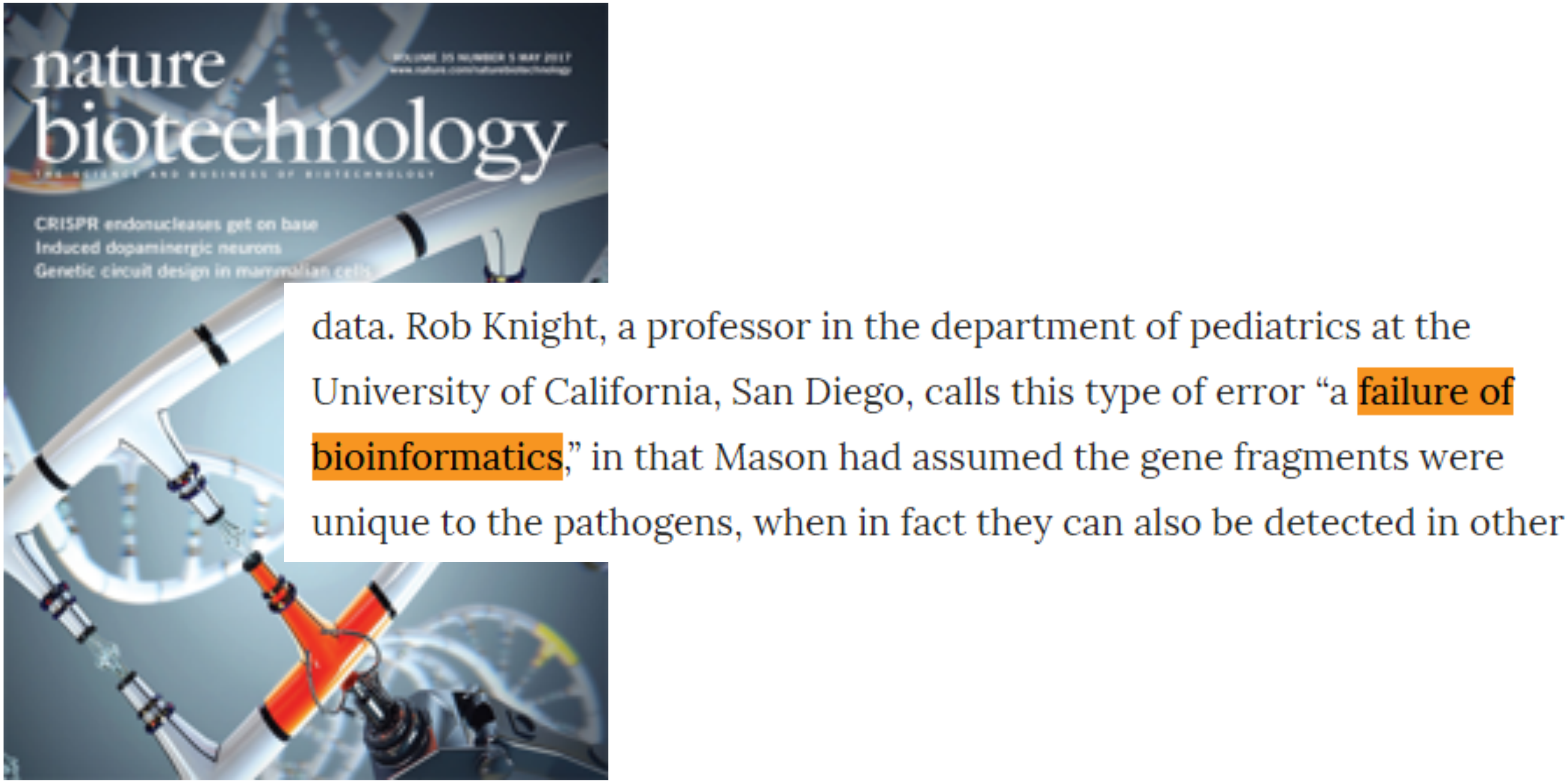
Bubonic Plague in the Subway System? Don't Worry About It

In October, riders were not deterred after reports that an Ebola-infected man had ridden the subway just before he fell ill. Robert Stolarik for The New York Times

https://www.nytimes.com/2015/02/07/nyregion/bubonic-plague-in-the-subway-system-dont-worry-about-it.html

The findings of Yersinia Pestis in the subway received wide coverage in the lay press, causing some alarm among New York residents

Failure of Bioinformatics


nature biotechnology
VOLUME 35 NUMBER 5 MAY 2017
www.nature.com/naturebiotechnology
THE SCIENCE AND BUSINESS OF BIOTECHNOLOGY
CRISPR endonucleases get on base
Induced dopaminergic neurons
Genetic circuit design in mammalian cells

data. Rob Knight, a professor in the department of pediatrics at the University of California, San Diego, calls this type of error "a failure of bioinformatics," in that Mason had assumed the gene fragments were unique to the pathogens, when in fact they can also be detected in other

Living in a microbial world
Charles Schmidt
Nature Biotechnology, **volume 35**, pages401–403 (2017)
https://www.nature.com/articles/nbt.3868

There is a critical need for **fast** and **accurate** genome analysis.

Open Questions

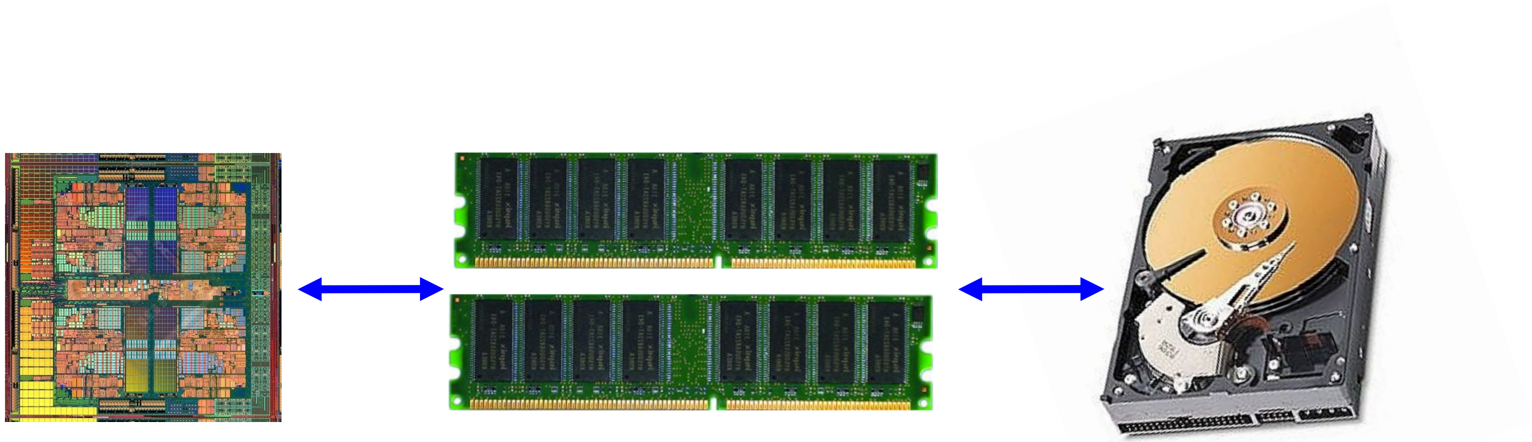
How and where to enable

fast, accurate, cheap,

privacy-preserving, and exabyte scale

analysis of genomic data?

Pushing Towards New Architectures



Microprocessor

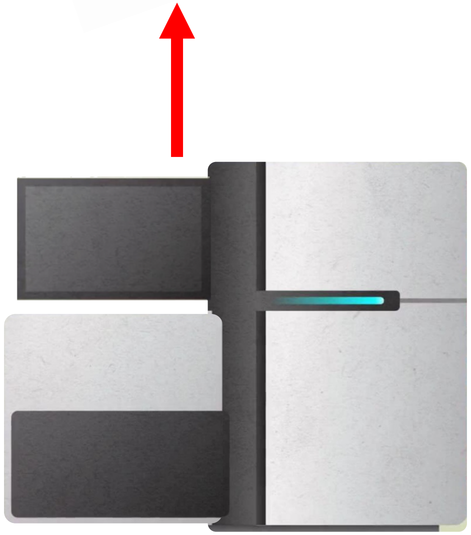
Main Memory

Storage (SSD/HDD)

Single memory request consumes

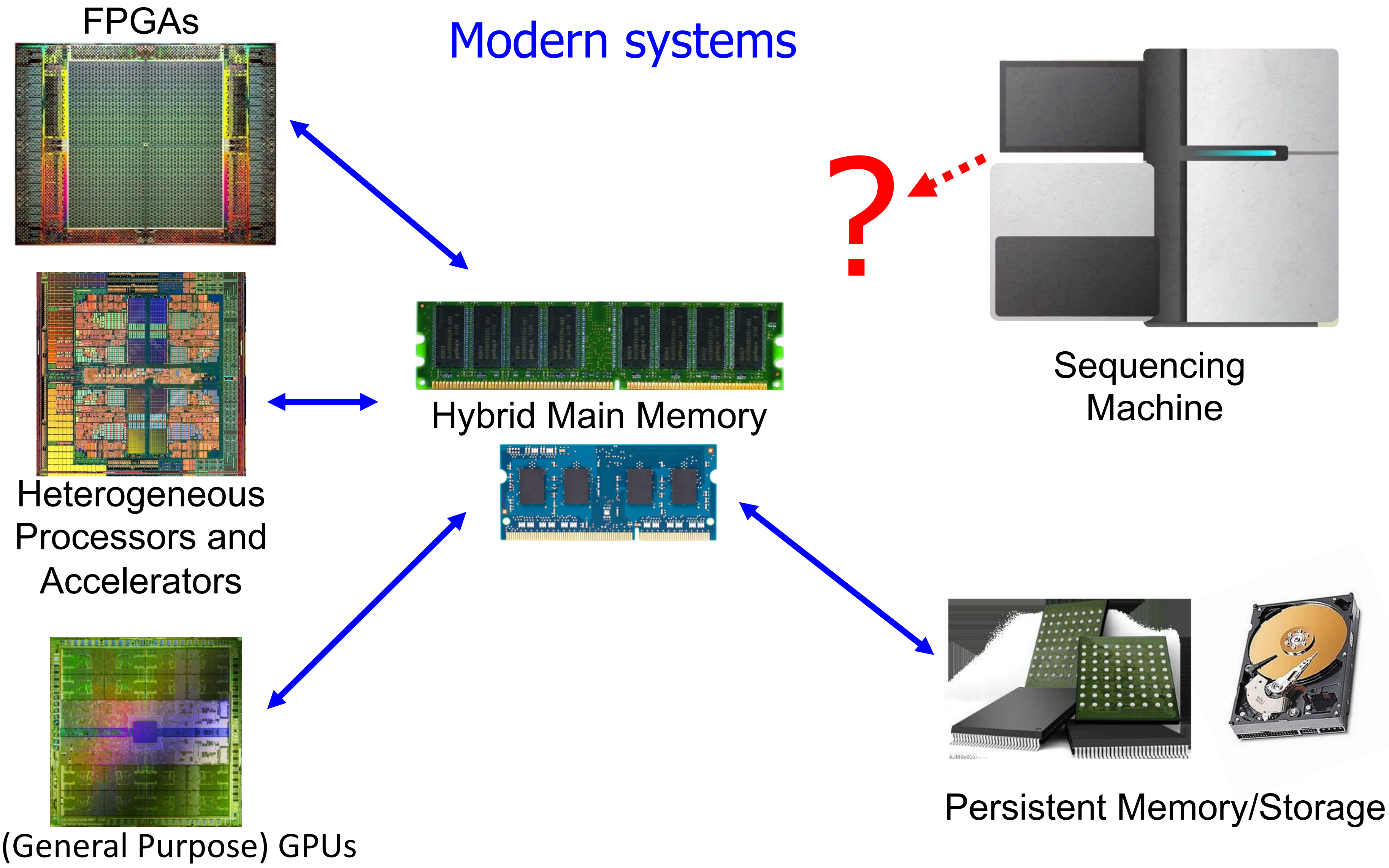
>160x-800x more energy compared to

performing a complex add operation



Sequencing
Machine

Processing Genomic Data Where it Makes Sense



Modern systems

FPGAs

Heterogeneous Processors and Accelerators

(General Purpose) GPUs

Hybrid Main Memory

Sequencing Machine

Persistent Memory/Storage

Key Takeaways

Most speedup comes from parallelism enabled by novel architectures and algorithms

Projects & Seminars Mobile Genomics

Genome Sequencing on Mobile Devices

Prof. Onur Mutlu

Dr. Mohammed Alser

ETH Zürich

Fall 2020

29 September 2020