

# P&S Genomics

## Lecture 3: Introduction to Sequencing

Dr. Mohammed Alser

ETH Zürich

Spring 2023

16 March 2023

# Agenda for Today

---

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

# Agenda for Today

---

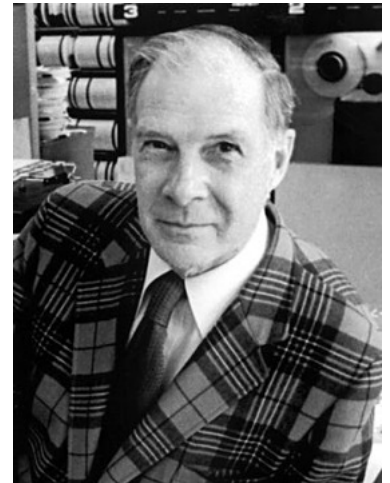
- **What is Genome Analysis?**
- What is Intelligent Genome Analysis?
- How we Analyze Genome?

# What is Data Analysis?

---

“The purpose of **computing** is [to gain] **insight**, not numbers”

Richard Hamming





# What is Genome Analysis?

---



# What is Genome Analysis?



nature research

Search  Login 

nature > subjects > genomic analysis

## Genomic analysis

 Atom  RSS Feed

Genomic analysis is the identification, measurement or comparison of genomic features such as DNA sequence, structural variation, gene expression, or regulatory and functional element annotation at a genomic scale. Methods for genomic analysis typically require high-throughput sequencing or microarray hybridization and bioinformatics.

# DNA Testing



Health + Ancestry  
Service

\$199

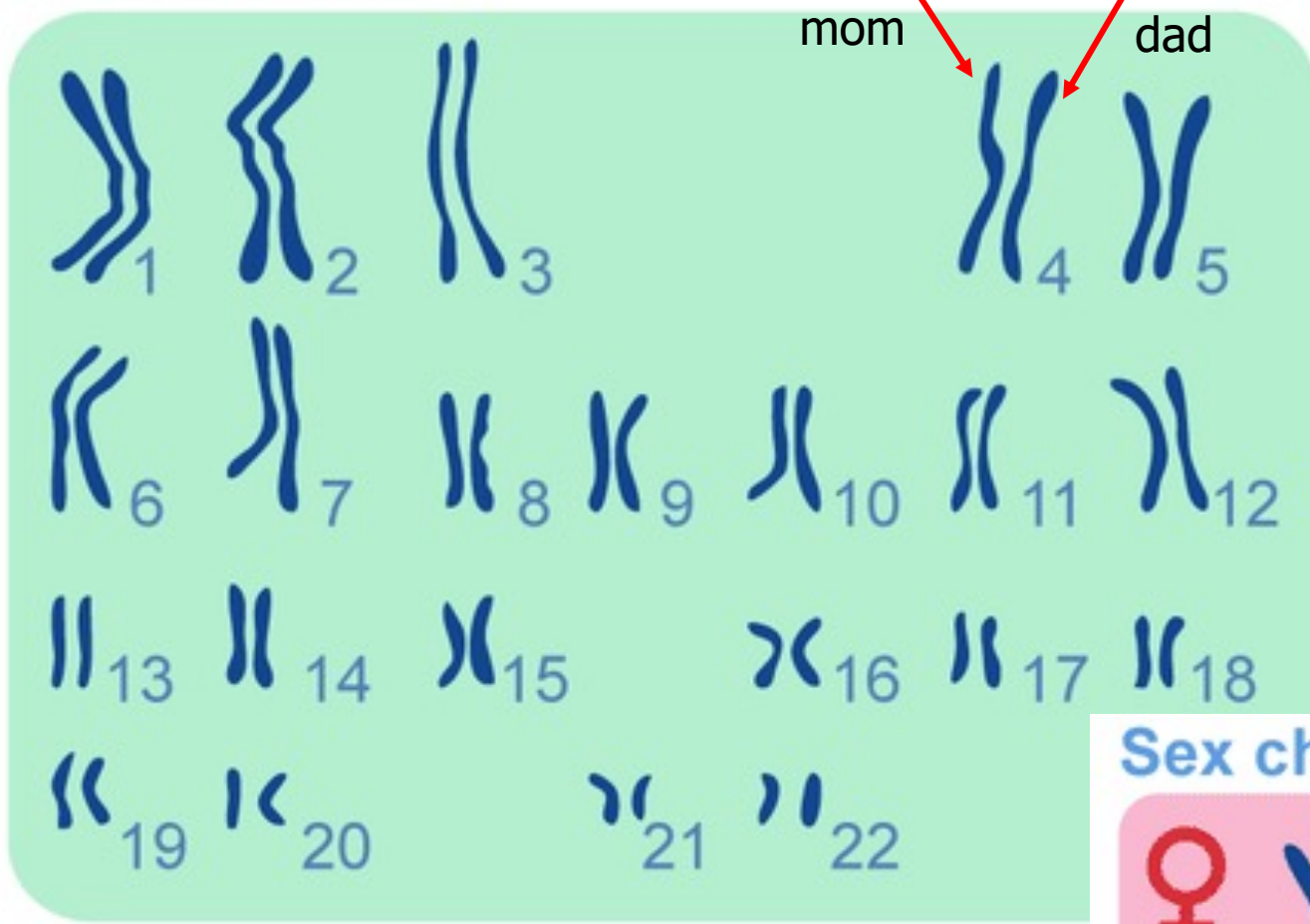
- Includes everything in Ancestry + Traits Service

*PLUS*

- 10+ Health Predisposition reports\*
- 5+ Wellness reports
- 40+ Carrier Status reports\*

# Human Chromosomes (23 Pairs)

## Autosomes



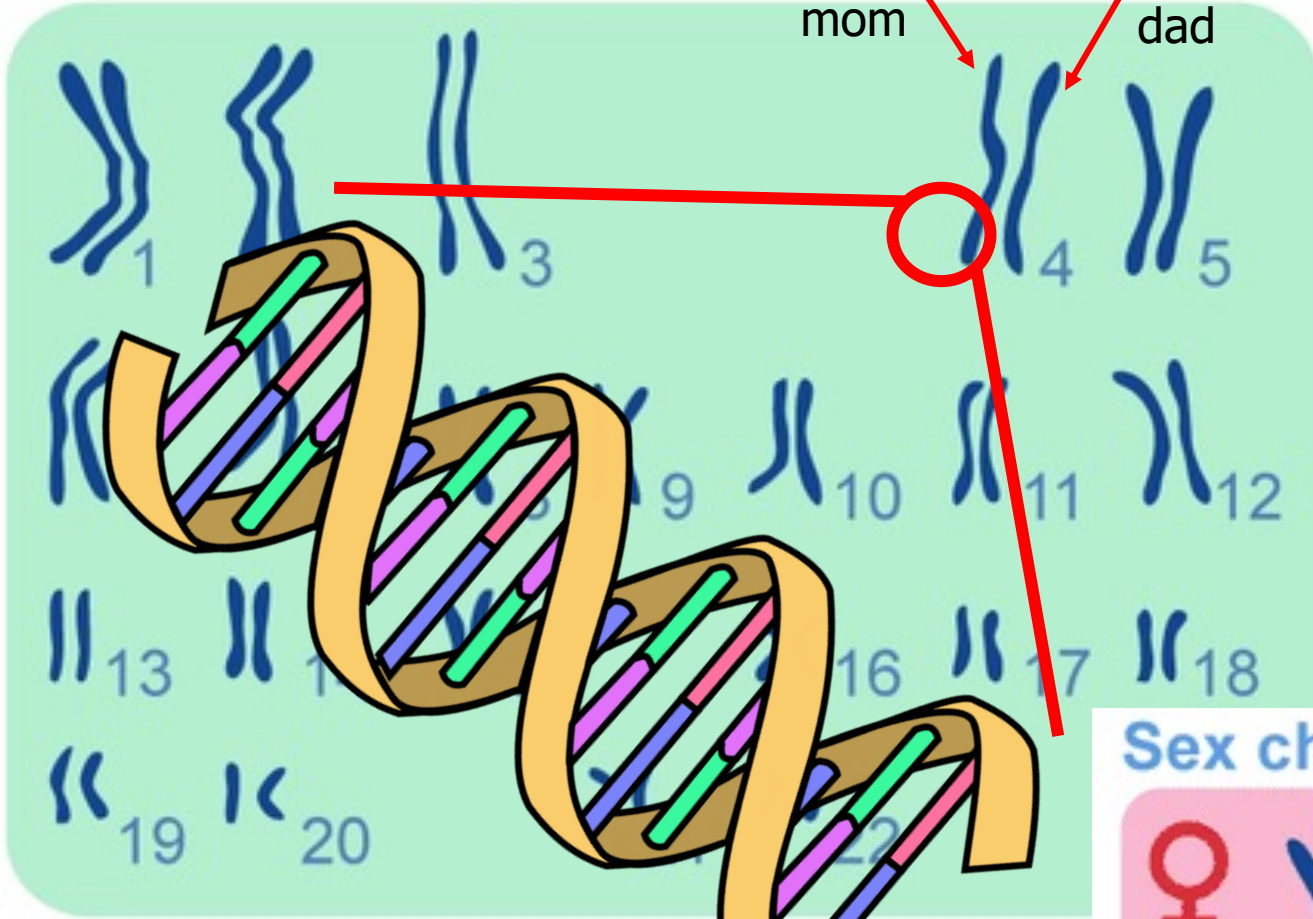
## Sex chromosomes






# Human Chromosomes (23 Pairs)

## Autosomes



-  = Adenine
-  = Thymine
-  = Cytosine
-  = Guanine
-  = Phosphate backbone

## Sex chromosomes



# Finding SNPs Associated with Complex Trait

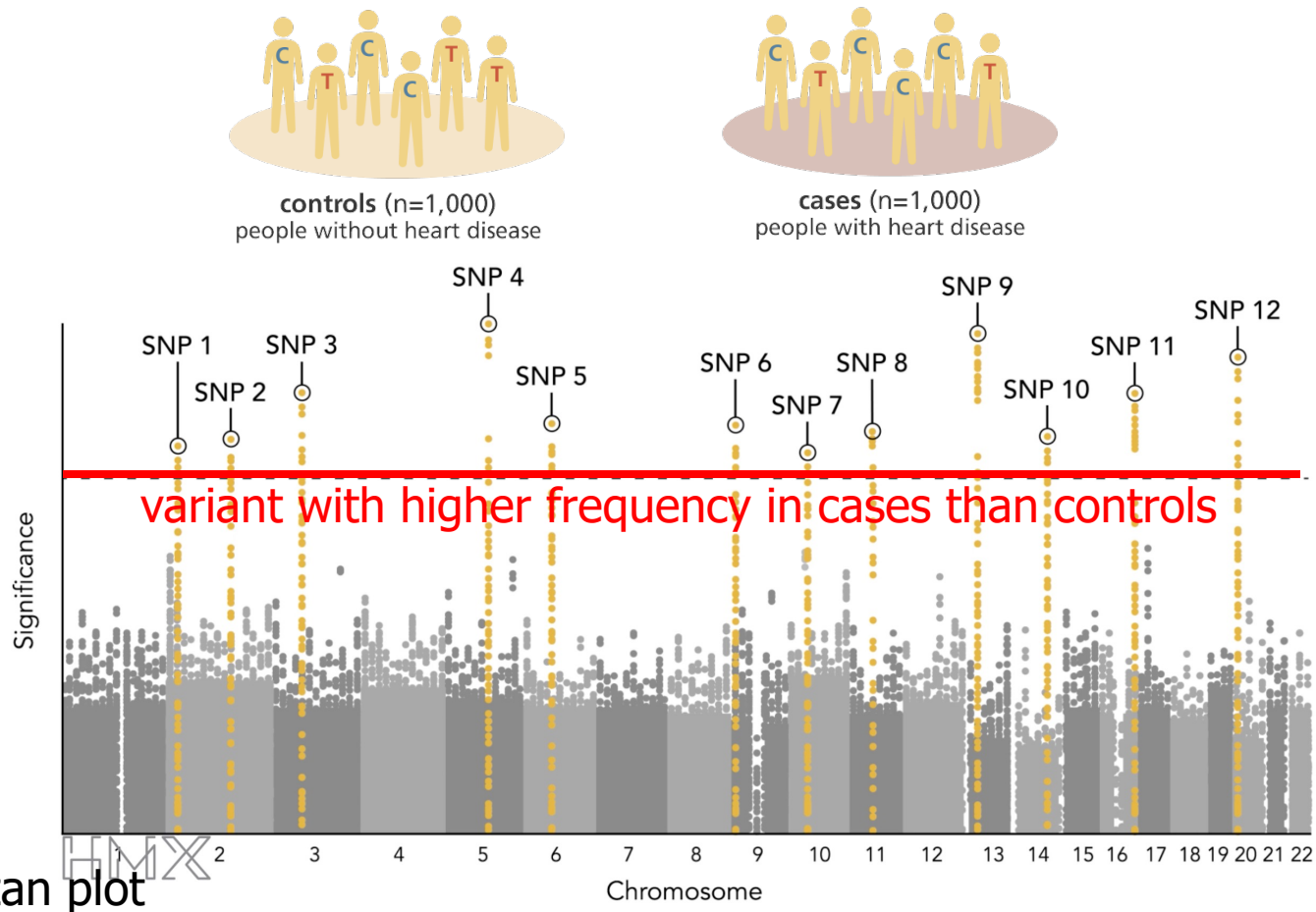
	SNP1	SNP2	Blood Pressure
Individual #1	...ACATG <b>C</b> CGACATTTCATAG <b>G</b> GCC...		<b>180</b>
Individual #2	...ACATG <b>C</b> CGACATTTCATA <b>A</b> GCC...		<b>175</b>
Individual #3	...ACATG <b>C</b> CGACATTTCATAG <b>G</b> GCC...		<b>170</b>
<b>Individual #4</b>	...ACATG <b>C</b> CGACATTTCATA <b>A</b> GCC...		<b>165</b>
Individual #5	...ACATG <b>C</b> CGACATTTCATAG <b>G</b> GCC...		<b>160</b>
Individual #6	...ACATG <b>C</b> CGACATTTCATAG <b>G</b> GCC...		<b>145</b>
Individual #7	...ACATG <b>C</b> CGACATTTCATA <b>A</b> GCC...		<b>140</b>
Individual #8	...ACATG <b>C</b> CGACATTTCATA <b>A</b> GCC...		<b>130</b>
Individual #9	...ACATG <b>T</b> CGACATTTCATAG <b>G</b> GCC...		<b>120</b>
Individual #10	...ACATG <b>T</b> CGACATTTCATA <b>A</b> GCC...		<b>120</b>
Individual #11	...ACATG <b>T</b> CGACATTTCATAG <b>G</b> GCC...		<b>115</b>
Individual #12	...ACATG <b>T</b> CGACATTTCATA <b>A</b> GCC...		<b>110</b>
Individual #13	...ACATG <b>T</b> CGACATTTCATAG <b>G</b> GCC...		<b>110</b>
Individual #14	...ACATG <b>T</b> CGACATTTCATA <b>A</b> GCC...		<b>110</b>
Individual #15	...ACATG <b>T</b> CGACATTTCATAG <b>G</b> GCC...		<b>105</b>
Individual #16	...ACATG <b>T</b> CGACATTTCATA <b>A</b> GCC...		<b>100</b>



SNP: single nucleotide polymorphism

# Genome-Wide Association Study (GWAS)

- Detecting genetic variants associated with phenotypes using two groups of people.



Manhattan plot

# Similar Association Studies

PERSPECTIVE

<https://doi.org/10.1038/s41588-019-0385-z>

nature  
genetics

## Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg<sup>1</sup>, Nasa Sinnott-Armstrong<sup>ID 2</sup>, Nicholas Mancuso<sup>ID 3</sup>, Alvaro N. Barbeira<sup>ID 4</sup>, David A. Knowles<sup>ID 5,6</sup>, David Golan<sup>2</sup>, Raili Ermel<sup>7</sup>, Arno Ruusalepp<sup>7,8</sup>, Thomas Quertermous<sup>ID 9</sup>, Ke Hao<sup>ID 10</sup>, Johan L. M. Björkegren<sup>ID 8,10,11,12\*</sup>, Hae Kyung Im<sup>ID 4\*</sup>, Bogdan Pasaniuc<sup>ID 3,13,14\*</sup>, Manuel A. Rivas<sup>ID 15\*</sup> and Anshul Kundaje<sup>ID 1,2\*</sup>

**Transcriptome-wide association studies (TWAS) integrate genome-wide association studies (GWAS) and gene expression datasets to identify gene-trait associations. In this Perspective, we explore properties of TWAS as a potential approach to prioritize causal genes at GWAS loci, by using simulations and case studies of literature-curated candidate causal genes for schizophrenia, low-density-lipoprotein cholesterol and Crohn's disease. We explore risk loci where TWAS accurately prioritizes the likely causal gene as well as loci where TWAS prioritizes multiple genes, some likely to be non-causal, owing to sharing of expression quantitative trait loci (eQTL). TWAS is especially prone to spurious prioritization with expression data from non-trait-related tissues or cell types, owing to substantial cross-cell-type variation in expression levels and eQTL strengths. Nonetheless, TWAS prioritizes candidate causal genes more accurately than simple baselines. We suggest best practices for causal-gene prioritization with TWAS and discuss future opportunities for improvement. Our results showcase the strengths and limitations of using eQTL datasets to determine causal genes at GWAS loci.**

Wainberg+, "[Opportunities and challenges for transcriptome-wide](#)

[association studies](#)", *Nature genetics*, 2019.



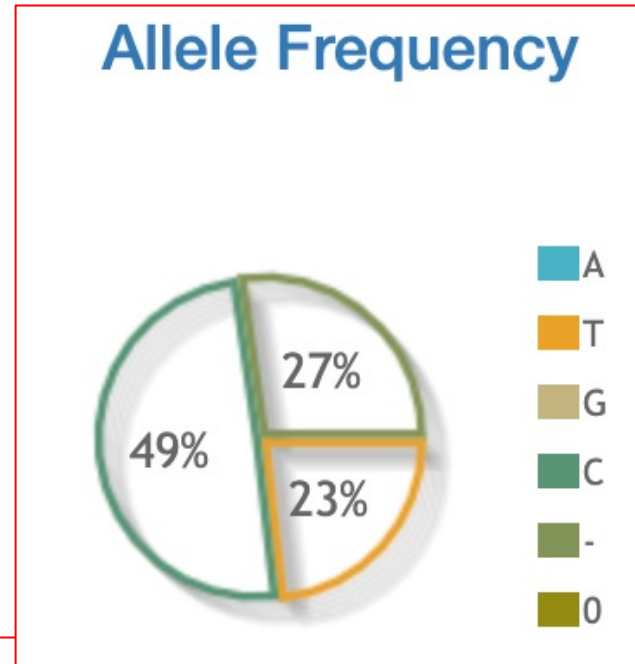
# SNPs and Personalized Medicine

openSNP

## SNP rs12979860

Basic Information

Name	rs12979860
Chromosome	19
Position	39248147
Weight of evidence	926



## Links to SNPedia

Title	Summary
<a href="#">rs12979860 T/T</a>	~20-25% of such hepatitis c patients respond to treatment
<a href="#">rs12979860 C/C</a>	~80% of such hepatitis c patients respond to treatment
<a href="#">rs12979860 C/T</a>	~20-40% of such hepatitis c patients respond to treatment

# Personalized Medicine for Critically Ill Infants

- **rWGS** can be performed in **2-day** (**costly**) or **5-day** time to interpretation.
- Diagnostic **rWGS** for infants
  - Avoids **morbidity**
  - Reduces **hospital stay length** by 6%-69%
  - Reduces **inpatient cost** by \$800,000-\$2,000,000.

Article | [Open Access](#) | Published: 04 April 2018

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

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

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

Article | [Open Access](#) | Published: 05 May 2020

## **Clinical utility of 24-h rapid trio-exome sequencing for critically ill infants**

Huijun Wang, Yanyan Qian, Yulan Lu, Qian Qin, Guoping Lu, Guoqiang Cheng, Ping Zhang, Lin Yang, Bingbing Wu [✉](#) & Wenhao Zhou [✉](#)

*npj Genomic Medicine* **5**, Article number: 20 (2020) | [Cite this article](#)

# Personalized Medicine in UK

---

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



# Much Larger Structural Variations!



## **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)

# Recommended Reading

---

**nature reviews** genetics

Explore our content ▾

Journal information ▾

---

nature > nature reviews genetics > review articles > article

Review Article | Published: 15 November 2019

## Structural variation in the sequencing era

Steve S. Ho, Alexander E. Urban & Ryan E. Mills 

*Nature Reviews Genetics* **21**, 171–189(2020) | [Cite this article](#)

**15k** Accesses | **16** Citations | **309** Altmetric | [Metrics](#)

Ho+, "[Structural variation in the sequencing era](#)", Nature Reviews Genetics, 2020

---

# Agenda for Today

---

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

# What is Intelligent Genome Analysis?

---

- Fast genome analysis

- *Real-time analysis*

Bandwidth

- Using intelligent architectures

- *Specialized HW with less data movement*

Energy-efficiency &  
Latency

- DNA is a valuable asset

- *Controlled-access analysis*

Privacy

- Population-scale genome analysis

- *Sequence anywhere at large scale!*

Scalability

- Avoiding erroneous analysis

- *E.g., your father is not your father*

Accuracy

---

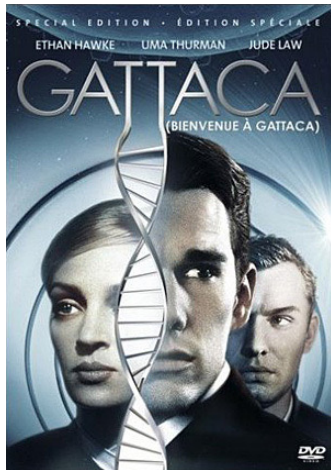
Does intelligent genome  
analysis really matter?



# Fast Genome Analysis?

- **Fast** genome analysis in mere seconds using **limited computational resources** (i.e., personal computer or small hardware).

1997



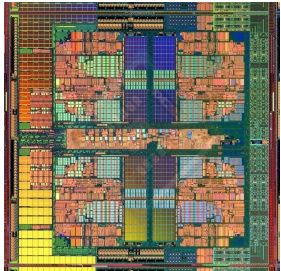
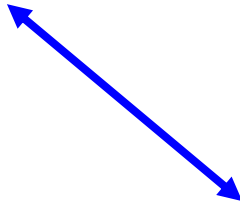
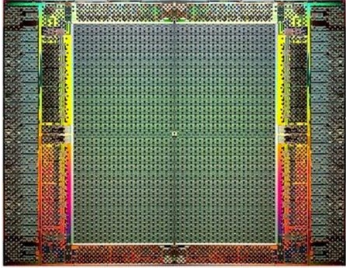
2015



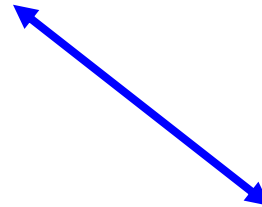
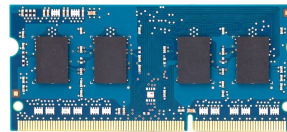
# Intelligent Architecture?

Modern systems

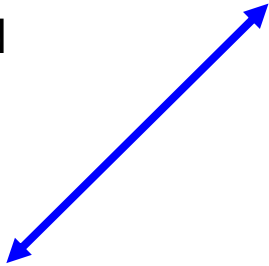
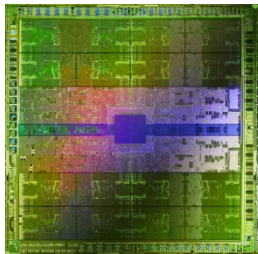
FPGAs



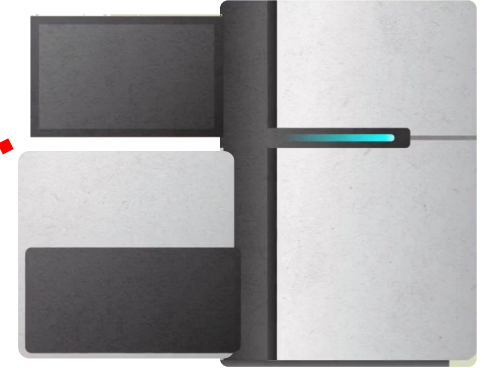
Hybrid Main Memory



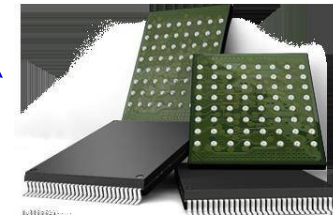
Heterogeneous Processors and Accelerators



(General Purpose) GPUs



Sequencing Machine

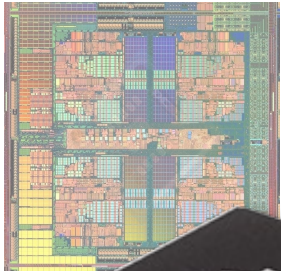
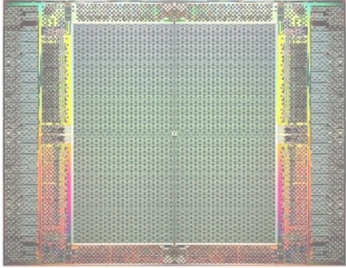


Persistent Memory/Storage

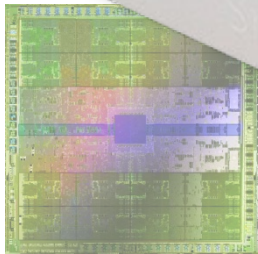
# Intelligent Architecture?

Modern systems

FPGAs

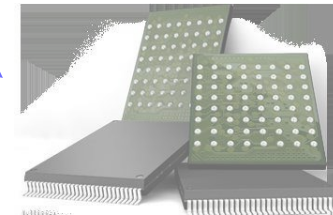


Hetero  
Pro  
Ac



(General Purpose) GPUs

Sequencing  
Machine



Persistent Memory/Storage



# Privacy-Preserving Genome Analysis?

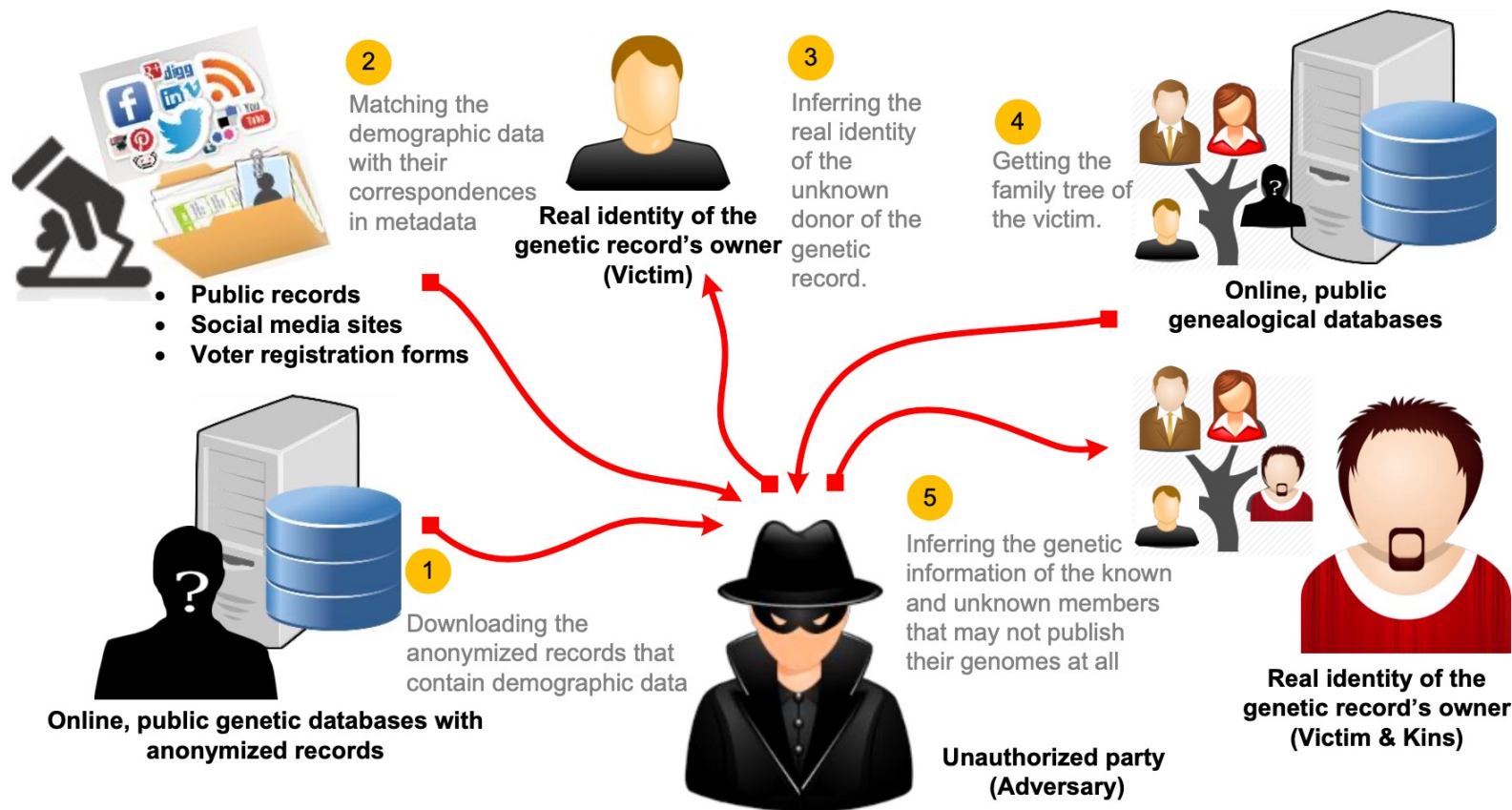


Fig. 5. A completion attack.

Alser+, "[Can you really anonymize the donors of genomic data in today's digital world?](#)" *10th International Workshop on Data Privacy Management (DPM)*, 2015.

# Can you Really Anonymize the Donors?

---

## (Position Paper) Can You Really Anonymize the Donors of Genomic Data in Today's Digital World?

Mohammed Alser, Nour Almadhoun, Azita Nouri, Can Alkan, and Erman Ayday

Computer Engineering Department, Bilkent University, 06800 Bilkent, Ankara, Turkey

**Abstract.** The rapid progress in genome sequencing technologies leads to availability of high amounts of genomic data. Accelerating the pace of biomedical breakthroughs and discoveries necessitates not only collecting millions of genetic samples but also granting open access to genetic databases. However, one growing concern is the ability to protect the privacy of sensitive information and its owner. In this work, we survey a wide spectrum of cross-layer privacy breaching strategies to human genomic data (using both public genomic databases and other public non-genomic data). We outline the principles and outcomes of each technique, and assess its technological complexity and maturation. We then review potential privacy-preserving countermeasure mechanisms for each threat.

**Keywords:** Genomics, Privacy, Bioinformatics

**DPM 2015**

Vienna, Austria  
September 21-22, 2015

Alser+, "[Can you really anonymize the donors of genomic data in today's digital world?](#)" *10th International Workshop on Data Privacy Management (DPM), 2015.*

# Privacy-Preserving DNA Test

## Our DNA Test, Reports, and Technology

- ✓ **Whole Genome Sequencing.** Decode 100% of your DNA with Whole Genome Sequencing and fully unlock your genetic blueprints.
- ✓ **Privacy First DNA Testing.** Begin your journey of discovery without risking the privacy of your most personal information.
- ✓ **Nebula Research Library.** Receive new reports every week that are based on the latest scientific discoveries.
- ✓ **Genome Exploration Tools.** Use powerful, browser-based genome exploration tools to answer any questions about your DNA.
- ✓ **Deep Genetic Ancestry.** Discover more about your ancestry with full Y chromosome and mitochondrial DNA sequencing and analysis.
- ✓ **Genomic Big Data Access.** Download your FASTQ, BAM, and VCF files and dive deeper into your Whole Genome Sequencing data.
- ✓ **Ready for Diagnostics.** Our Whole Genome Sequencing data is of the highest quality and can be used by physicians and genetic counselors.



### 30x Whole Genome Sequencing DNA Test

**\$299**  
Normally ~~\$4000~~  
Save 70%!

A genetic test that decodes 100% of your DNA with very high accuracy. 30x Whole Genome Sequencing offers the best value for money and is the best choice for most people.

### 100x Whole Genome Sequencing DNA Test

**\$999**  
Normally ~~\$3500~~  
Save 70%!

A genetic test that decodes 100% of your DNA with extremely high accuracy. 100x Whole Genome Sequencing is recommended for the discovery of rare genetic mutations.

[Get Sequenced](#)



# Rapid Surveillance of Disease Outbreaks?

Figure 1: Deployment of the portable genome surveillance system in Guinea.



Quick+, "[Real-time, portable genome sequencing for Ebola surveillance](#)", *Nature*, 2016

# Scalable SARS-CoV-2 Testing

medRxiv

THE PREPRINT SERVER FOR HEALTH SCIENCES



BMJ Yale

HOME | ABOUT

[Comments \(1\)](#)

## Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing

[ID](#) Joshua S. Bloom, [ID](#) Eric M. Jones, [ID](#) Molly Gasperini, [ID](#) Nathan B. Lubock, [ID](#) Laila Sathe, [ID](#) Chetan Munugala, [ID](#) A. Sina Booeshaghi, [ID](#) Oliver F. Brandenburg, [ID](#) Longhua Guo, [ID](#) James Boocock, [ID](#) Scott W. Simpkins, [ID](#) Isabella Lin, [ID](#) Nathan LaPierre, [ID](#) Duke Hong, [ID](#) Yi Zhang, [ID](#) Gabriel Oland, [ID](#) Bianca Judy Choe, [ID](#) Sukantha Chandrasekaran, [ID](#) Evann E. Hilt, [ID](#) Manish J. Butte, [ID](#) Robert Damoiseaux, [ID](#) Aaron R. Cooper, [ID](#) Yi Yin, [ID](#) Lior Pachter, [ID](#) Omai B. Garner, [ID](#) Jonathan Flint, [ID](#) Eleazar Eskin, [ID](#) Chongyuan Luo, [ID](#) Sriram Kosuri, [ID](#) Leonid Kruglyak, [ID](#) Valerie A. Arboleda

**doi:** <https://doi.org/10.1101/2020.08.04.20167874>

Bloom+, "[Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing](#)", *medRxiv*, 2020

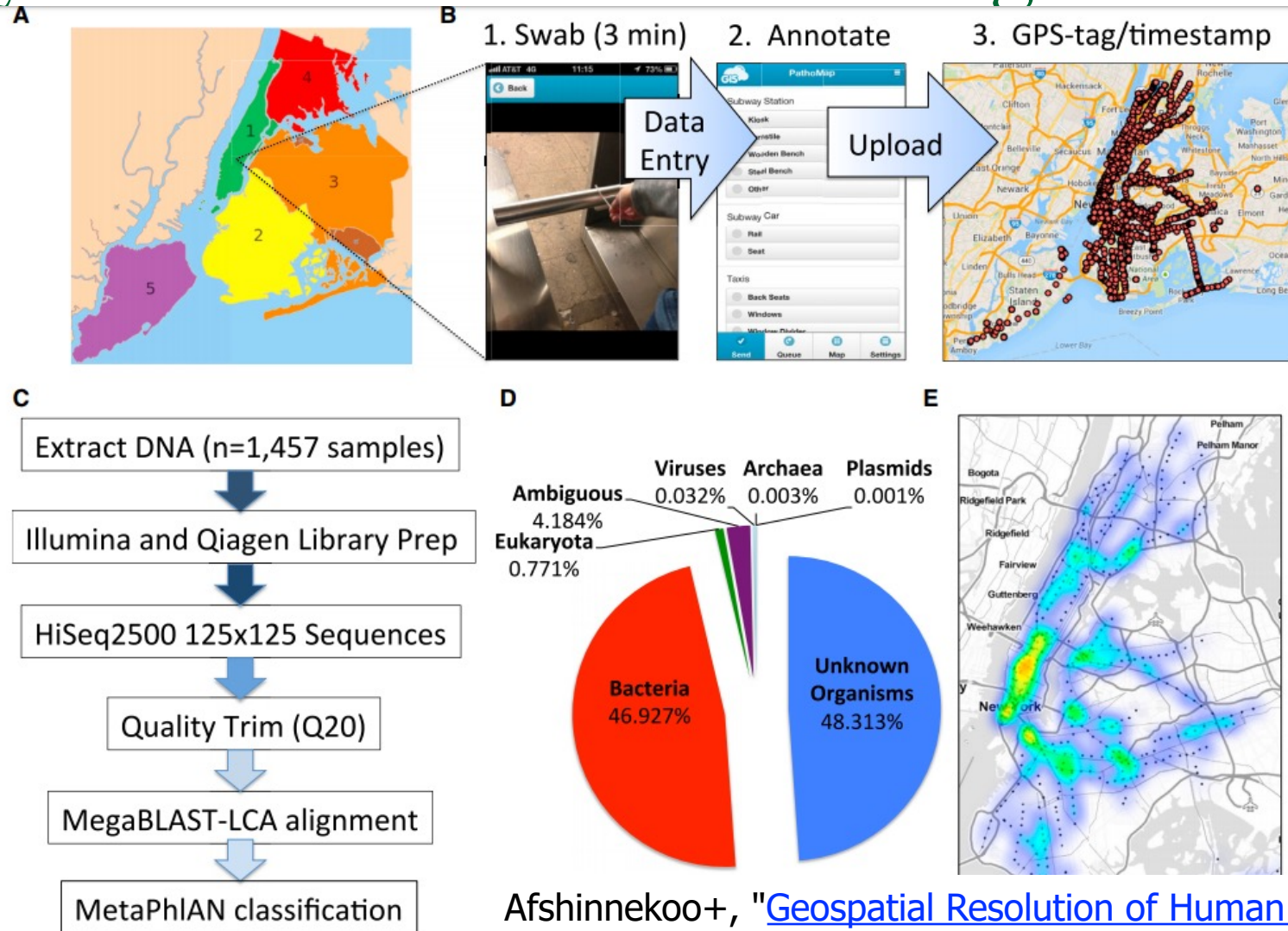


# Population-Scale Microbiome Profiling





# City-Scale Microbiome Profiling



**Figure 1. The Metagenome of New York City**

(A) The five boroughs of NYC include (1) Manhattan (green)

(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

Afshinnekoo+, "[Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics](#)", Cell Systems, 2015

# Population-Scale Microbiome Profiling



Log in Register Su

ARTICLE | ONLINE NOW

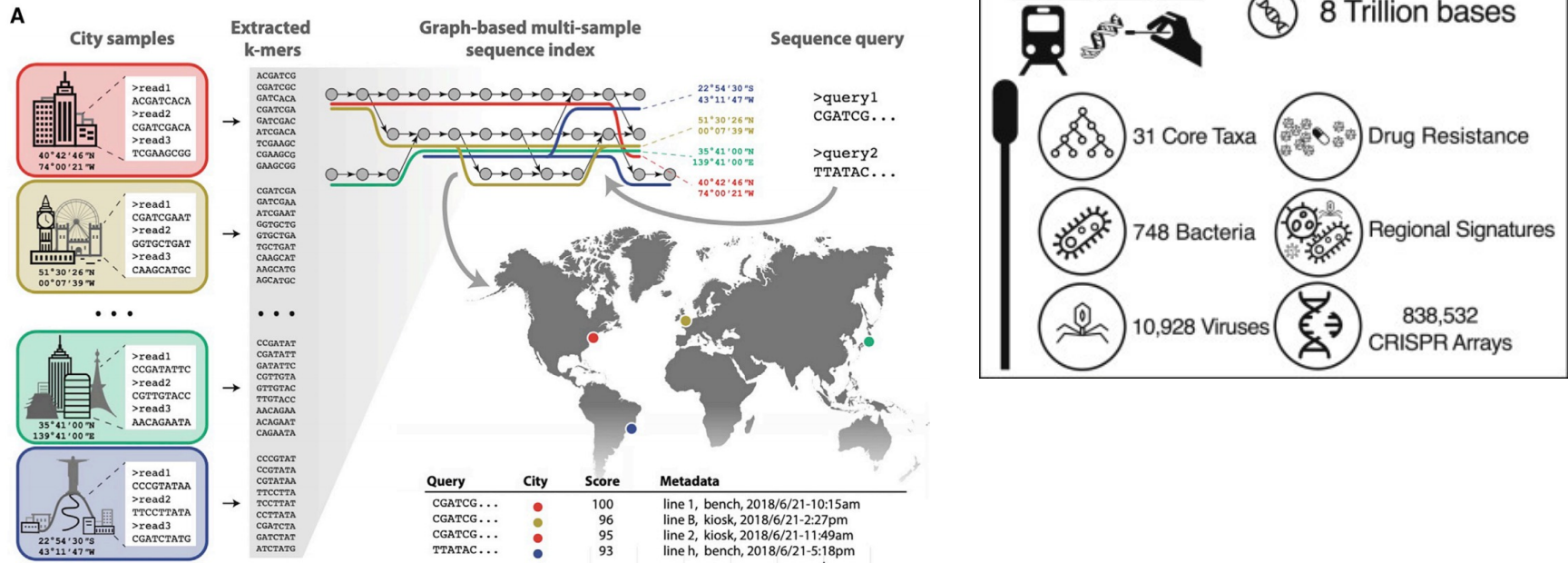
PDF [9 MB] Figures Save

## A global metagenomic map of urban microbiomes and antimicrobial resistance

David Danko <sup>68</sup> • Daniela Bezdán <sup>68</sup> • Evan E. Afshin • ... Sibó Zhu • Christopher E. Mason <sup>69</sup> ✉

The International MetaSUB Consortium • Show all authors • Show footnotes

Open Access • Published: May 26, 2021 • DOI: <https://doi.org/10.1016/j.cell.2021.05.002>



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

# 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 (Yersinia)

### What Is It?

Published: December, 2018

Plague is caused by Yersinia treated promptly. Plague has last 2,000 years. Plague has cause skin sores that form b than one-third of the popul the population died within

*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

---



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.

# Achieving Intelligent Genome Analysis?

---

How and where to enable

fast, accurate, cheap,

privacy-preserving, and exabyte scale

analysis of genomic data?



# Agenda for Today

---

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

# Genome Analysis

---



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



```
>CCTCCTCAGTGCCACCCAGCCCCTGGCAGCTCCCAAACAGGCTCTTATTAAAACACCCTGTTCCCTGCCCTTGGAGTGAGGTGTCAAG
GACCTAAACTAAAAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTCTT
CATGTCAAGGACCTAATGTGCTAAACAGCACTTTTTTGACCATTATTTTGGATCTGAAAGAAATCAAGAATAAATGAAGGACTTGATACATTG
GAAGAGGAGAGTCAAGGACCTACAGAAAAAAAAAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTAAAATTTAAGTAATTCTTTGAAAAAA
ACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCTGTGTTGCAGGTCTTCTTGCATTTCCCTGTCAAAGAAAAAGAATTTAAAATTT
AAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTCAAGGCCAAGAGTTGCAAAAAAAAAAAAAAAAAAGAAAA
GAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTCTTCATGTCAAGGACCTAATGTAGCCAGAATGG
TTGTGGGATGGGAGCCTCTGTGGACCGACCAGGTAGCTCTCTTTCCACACTGTAGTCTCAAAGCTTCTTCATGTGGTTTTCTCTGAGTGAAA
AAAAAAAAAAGAAAAAGAAAAGAAAAAGAATTTAAAATTTAAGTAATTCTTTGAAAAAACTAATTTCTAAGCTTTTCATGTCAAGGACC
TAATGTAGCTATACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCCTTCCCAGAAAGCTTCTTCA.....
```

# Genome Analysis

---



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



Why?!

>CCT  
GACC  
CATGT  
GAAG  
ACTA  
AAGT  
GAAA  
TTGTC  
AAAAAAAAAAGAAAAAGAAAAGAAAAGAATTTAAAATTTAAGTAATTTCTTTGAAAAAACTAATTTCTAAGCTTTTTCATGTCAAGGACC  
TAATGTAGCTATACTGAACGTTATCTAGGGGAAAGATTGAAGGGGAGCTCTAAGGTCAACACACCACCACTTCCCAGAAAGCTTCTTCA.....

# Suggested Readings

---

## nature methods

Explore content ▾

About the journal ▾

Publish with us ▾

---

[Published: November 2009](#)

### **Next-generation sequencing library preparation: simultaneous fragmentation and tagging using *in vitro* transposition**

[Fraz Syed](#) , [Haiying Grunenwald](#) & [Nicholas Caruccio](#)

[Nature Methods](#) **6**, i–ii (2009) | [Cite this article](#)

**16k** Accesses | **4** Citations | **5** Altmetric | [Metrics](#)

<https://www.nature.com/articles/nmeth.f.272>

---

# Suggested Readings

---

## nature biotechnology

---

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

---

[nature](#) > [nature biotechnology](#) > [review articles](#) > [article](#)

[Published: 09 October 2008](#)

## Next-generation DNA sequencing

[Jay Shendure](#)  & [Hanlee Ji](#) 

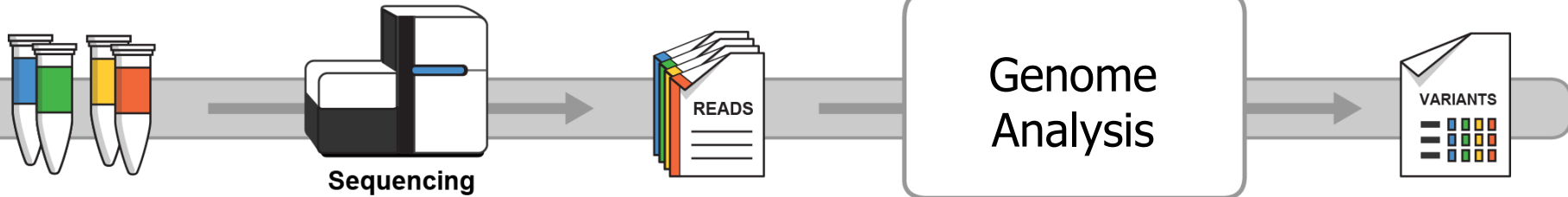
[Nature Biotechnology](#) **26**, 1135–1145 (2008) | [Cite this article](#)

**149k** Accesses | **2645** Citations | **79** Altmetric | [Metrics](#)

<https://www.nature.com/articles/nbt1486>

---

# Genome Sequencer is a Chopper



CCCCCTATATACGTACTAGTACGT  
ACGACTTTAGTACGTACGT  
TATATACGTACTAGTACGT  
ACGTACGCCCCTACGTA  
TATATACGTACTAGTACGT  
ACGACTTTAGTACGTACGT  
TATATACGTACTAAAGTACGT  
TATATACGTACTAGTACGT  
ACGTTTTTAAACGTA  
TATATACGTACTAGTACGT  
ACGACGGGGAGTACGTACGT



$1 \times 10^{12}$  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.**

# Oxford Nanopore Sequencers



**MinION Mk1B**



**MinION Mk1C**



**GridION Mk1**



**PromethION 24/48**

	<b>MinION Mk1B</b>	<b>MinION Mk1C</b>	<b>GridION Mk1</b>	<b>PromethION 24</b>	<b>PromethION 48</b>
<b>Read length</b>	> 2Mb	> 2Mb	> 2Mb	> 2Mb	> 2Mb
<b>Yield per flow cell</b>	50 Gb	50 Gb	50 Gb	220 Gb	220 Gb
<b>Number of flow cells per device</b>	1	1	5	24	48
<b>Yield per device</b>	<50 Gb	<50 Gb	<250 Gb	<5.2 Tb	<10.5 Tb
<b>Starting price</b>	\$1,000	\$4,990	\$49,995	\$195,455	\$327,455

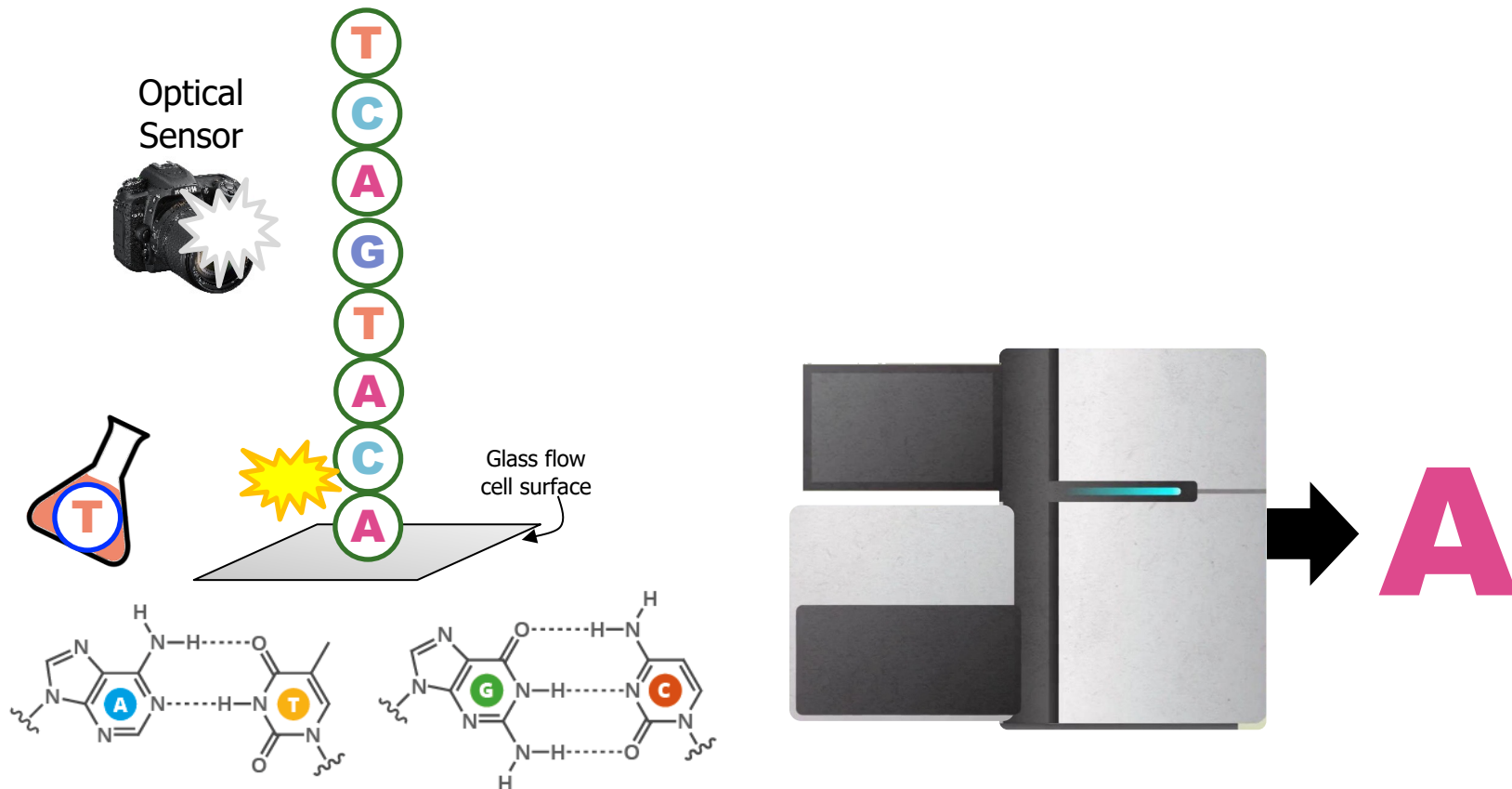


# Illumina Sequencers

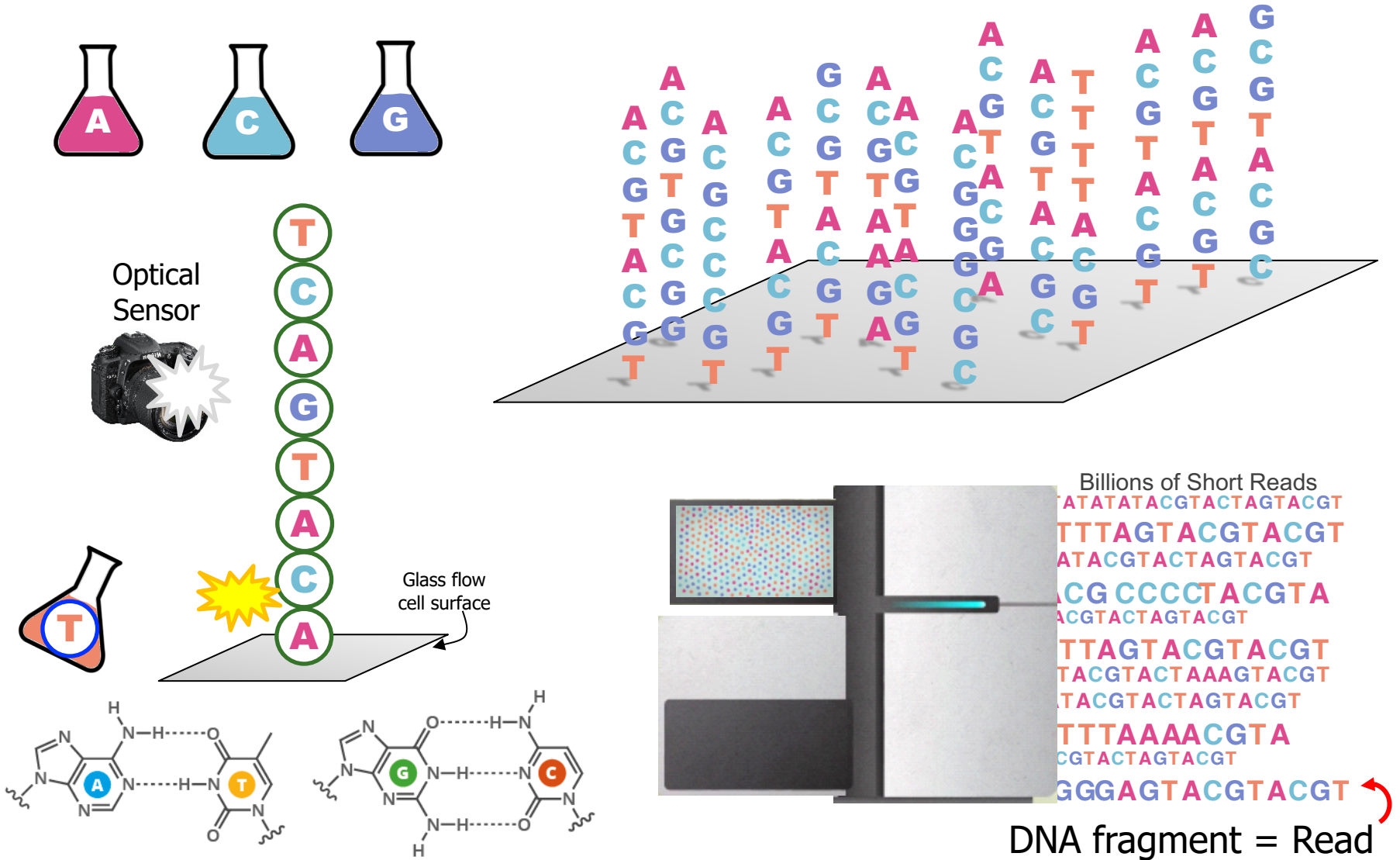


<b>Run time</b>	9.5–19 hrs	4–24 hrs	4–55 hrs	12–30 hrs	24-48 hrs	13-44 hrs
<b>Max. reads per run</b>	4 million	25 million	25 million	400 million	1 billion	20 billion
<b>Max. read length</b>	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 x 250
<b>Max. output</b>	1.2 Gb	7.5 Gb	15 Gb	120 Gb	300 Gb	6000 Gb
<b>Estimated price</b>	\$19,900	\$49,500	\$128,000	\$275,000	\$335,000	\$985,000

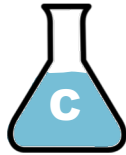
# How Does Illumina Machine Work?



# How Does Illumina Machine Work?

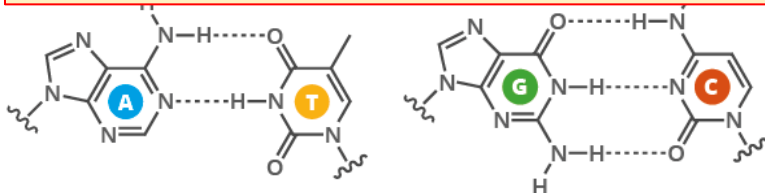


# How Does Illumina Machine Work?



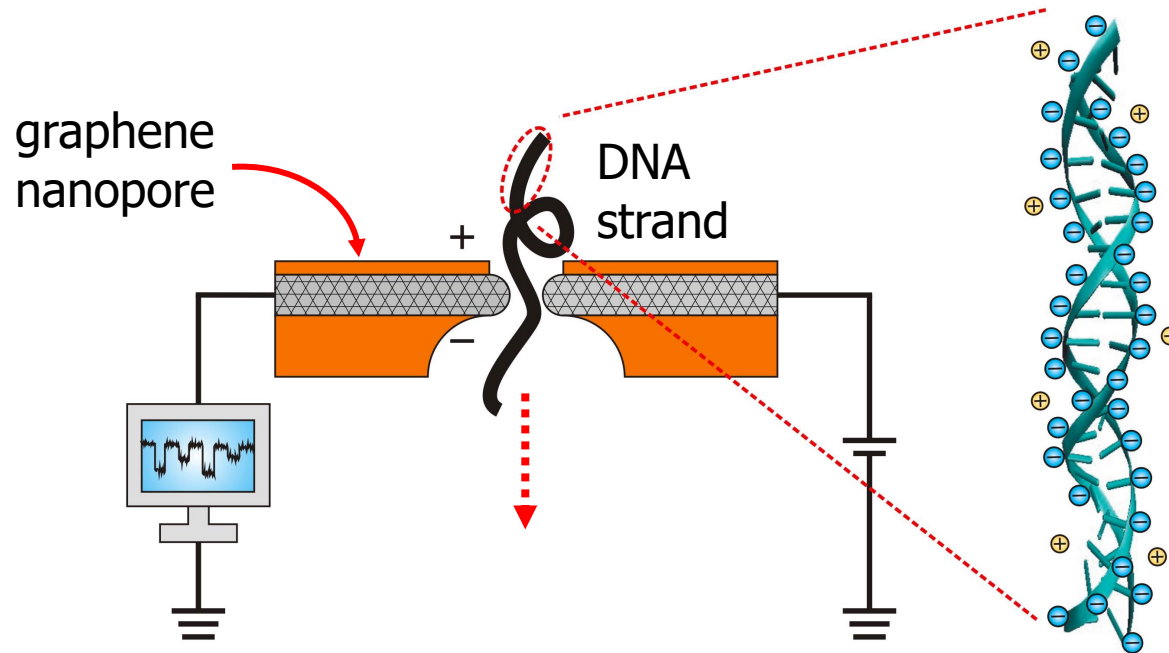
Check Illumina virtual tour:

<https://emea.illumina.com/systems/sequencing-platforms/iseq/tour.html>



DNA fragment = Read

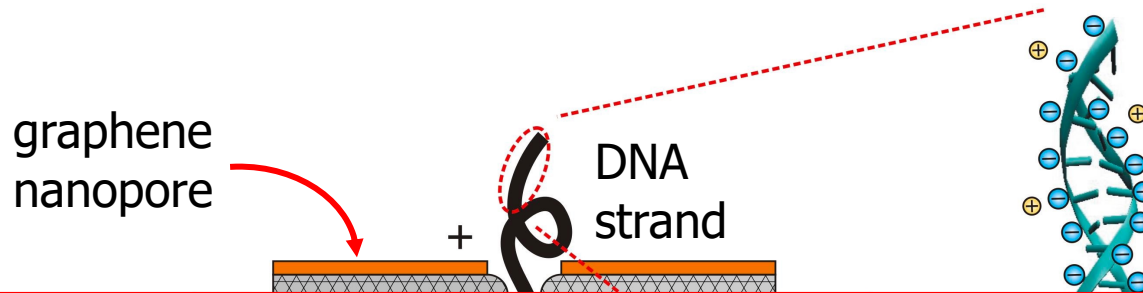
# How Does Nanopore Machine Work?



- **Nanopore** is a nano-scale hole (<20nm).
- In nanopore sequencers, an **ionic current** passes through the nanopores
- When the DNA strand passes through the nanopore, the sequencer measures the **change in current**
- This change is used to identify the bases in the strand with the help of **different electrochemical structures** of the different bases



# How Does Nanopore Machine Work?



Check Nanopore virtual tour:

<https://nanoporetech.com/resource-centre/minion-video>

measures the the **change in current**

- This change is used to identify the bases in the strand with the help of **different electrochemical structures** of the different bases

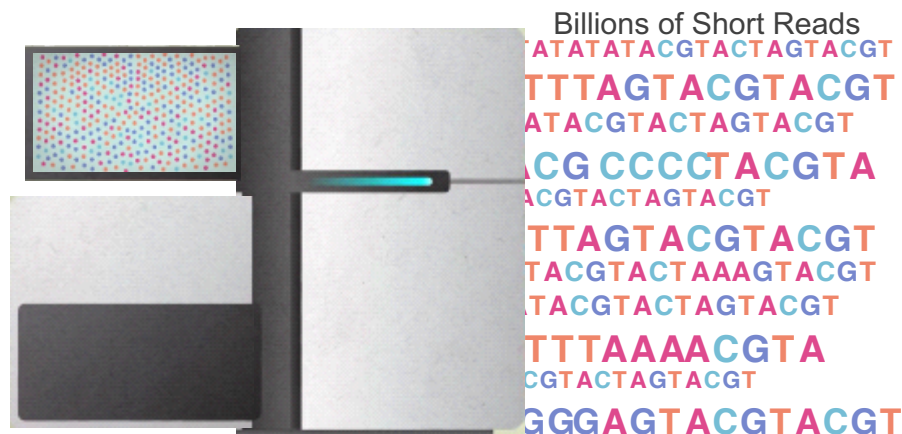
# Common Disadvantages!

---

Regardless the sequencing machine,

reads still lack information about their order and location

(which part of genome they are originated from)



# Solving the Puzzle

---



Reference genome



Reads

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

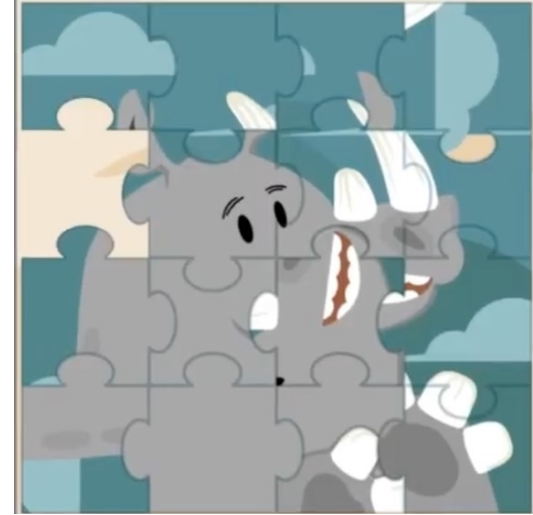
# HTS Sequencing Output

---

Small pieces of a puzzle  
**short reads (Illumina)**



Large pieces of a puzzle  
**long reads (ONT & PacBio)**



Which sequencing technology is the best?

100-300 bp

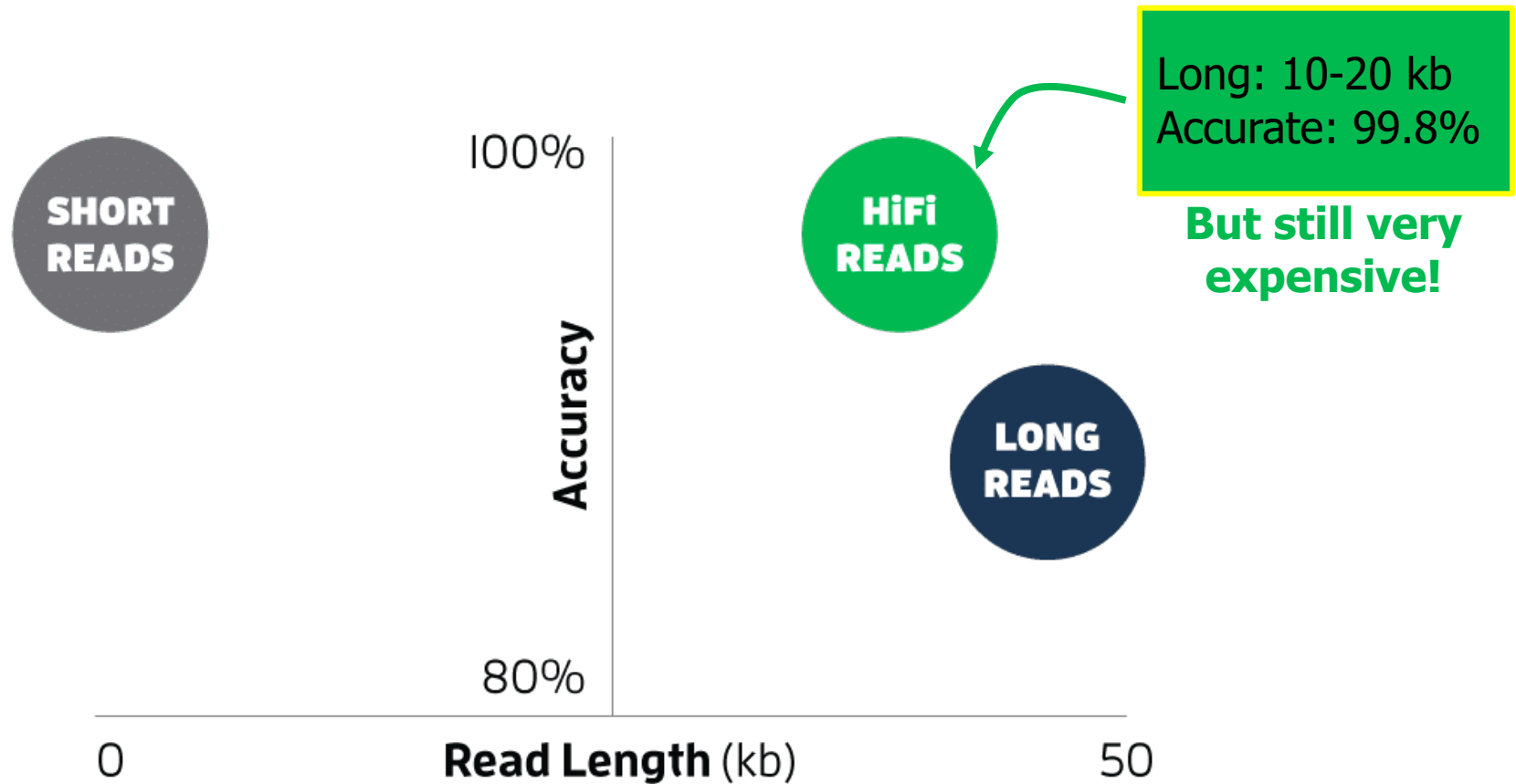
low error rate (~0.1%)

500-2M bp

high error rate (~15%)

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

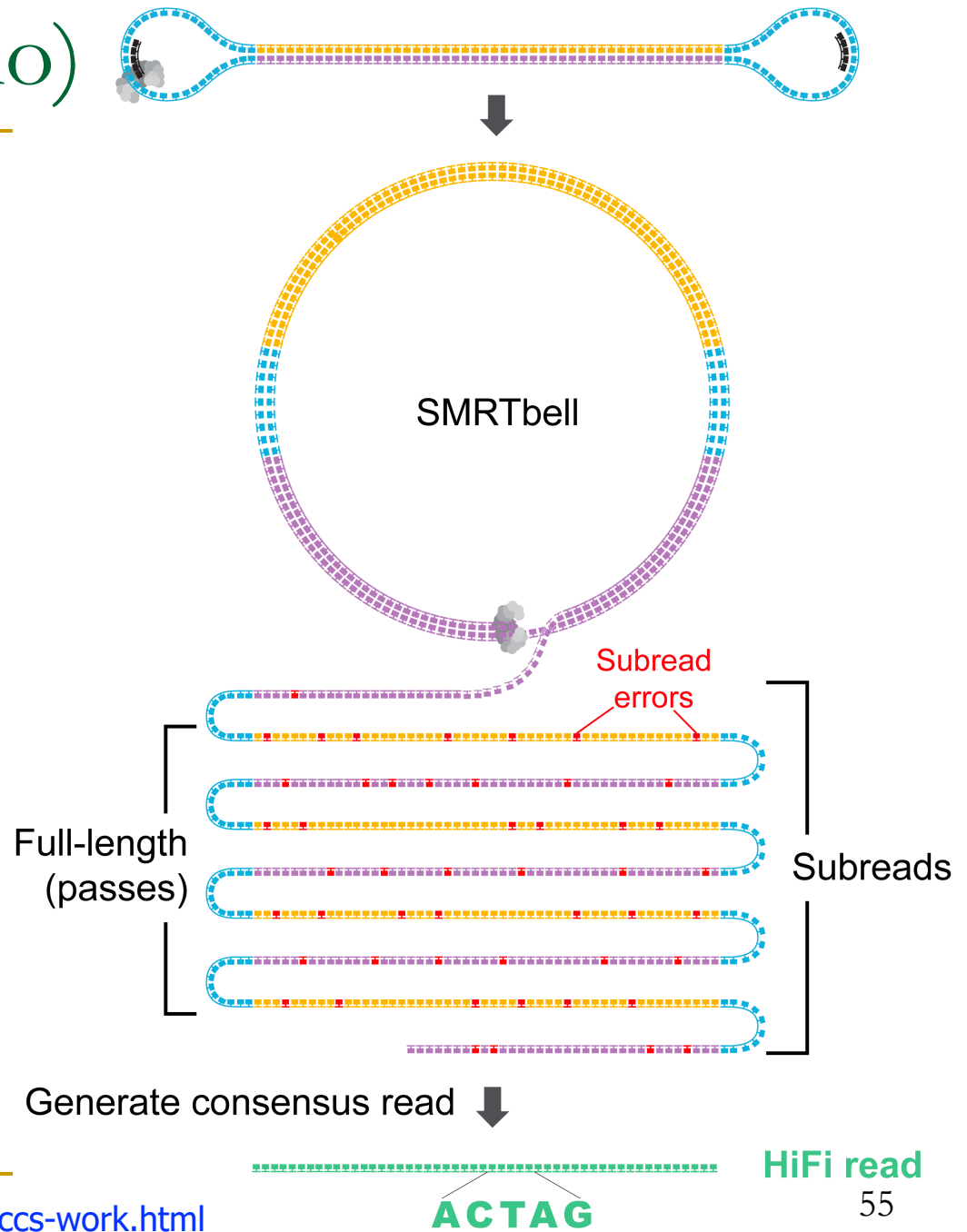
# HiFi Reads (PacBio)



Wenger+, "[Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome](#)", *Nature Biotechnology*, 2019



# HiFi Reads (PacBio)



---

Changes in sequencing technologies  
can render some  
read mapping algorithms irrelevant

# 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

---

Looking forward,  
Will we be able to read  
the entire genome sequence?



# P&S Genomics

## Lecture 3: Introduction to Sequencing

Dr. Mohammed Alser

ETH Zürich

Spring 2023

16 March 2023