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?

https://onlinelearning.hms.harvard.edu/hmx/courses/genetic-testing/
https://www.nature.com/subjects/genomic-analysis
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*

https://www.myheritage.ch/dna  https://www.23andme.com/
Human Chromosomes (23 Pairs)

Autosomes

From mom

From dad

Sex chromosomes

or

SAFARI
Human Chromosomes (23 Pairs)

Autosomes

From mom

From dad

= Adenine
= Thymine
= Cytosine
= Guanine
= Phosphate backbone

Sex chromosomes: XX or XY
Finding SNPs Associated with Complex Trait

<table>
<thead>
<tr>
<th>Individual #1</th>
<th>SNP1</th>
<th>SNP2</th>
<th>Blood Pressure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>180</td>
</tr>
<tr>
<td>Individual #2</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>175</td>
</tr>
<tr>
<td>Individual #3</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>170</td>
</tr>
<tr>
<td>Individual #4</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>165</td>
</tr>
<tr>
<td>Individual #5</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>160</td>
</tr>
<tr>
<td>Individual #6</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>145</td>
</tr>
<tr>
<td>Individual #7</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>140</td>
</tr>
<tr>
<td>Individual #8</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>...ACATG**CCGACATTTCATAGGCC...</td>
<td>130</td>
</tr>
<tr>
<td>Individual #9</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>120</td>
</tr>
<tr>
<td>Individual #10</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>120</td>
</tr>
<tr>
<td>Individual #11</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>115</td>
</tr>
<tr>
<td>Individual #12</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>110</td>
</tr>
<tr>
<td>Individual #13</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>110</td>
</tr>
<tr>
<td>Individual #14</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>110</td>
</tr>
<tr>
<td>Individual #15</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>105</td>
</tr>
<tr>
<td>Individual #16</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>...ACATG**TCGACATTTCATAGGCC...</td>
<td>100</td>
</tr>
</tbody>
</table>

SNP: single nucleotide polymorphism

Eleazar Eskin: Discovering the Causal Variants Involved in GWAS Studies, CGSI 2018, UCLA
Genome-Wide Association Study (GWAS)

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

Manhattan plot

variant with higher frequency in cases than controls

https://onlinelearning.hms.harvard.edu/hmx/courses/genetic-testing/
Opportunities and challenges for transcriptome-wide association studies

Michael Wainberg¹, Nasa Sinnott-Armstrong², Nicholas Mancuso³, Alvaro N. Barbeira⁴, David A. Knowles⁵,⁶, David Golan², Raili Ermel⁷, Arno Ruusalepp⁷,⁸, Thomas Quertermous⁹, Ke Hao¹⁰, Johan L. M. Björkegren¹⁰,¹¹,¹², Hae Kyung Im¹⁰,⁴*, Bogdan Pasaniuc¹³,¹⁴*, Manuel A. Rivas¹⁵* and Anshul Kundaje¹,²*

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.

SNPs and Personalized Medicine

SNP rs12979860

Basic Information

<table>
<thead>
<tr>
<th>Name</th>
<th>rs12979860</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome</td>
<td>19</td>
</tr>
<tr>
<td>Position</td>
<td>39248147</td>
</tr>
<tr>
<td>Weight of evidence</td>
<td>926</td>
</tr>
</tbody>
</table>

Allele Frequency

- A: 49%
- T: 27%
- G: 23%
- C: 49%
- - : 27%
- O: 23%

Links to SNPpedia

<table>
<thead>
<tr>
<th>Title</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs12979860 T/T</td>
<td>~20-25% of such hepatitis c patients respond to treatment</td>
</tr>
<tr>
<td>rs12979860 C/C</td>
<td>~80% of such hepatitis c patients respond to treatment</td>
</tr>
<tr>
<td>rs12979860 C/T</td>
<td>~20-40% of such hepatitis c patients respond to treatment</td>
</tr>
</tbody>
</table>

https://opensnp.org/snps/rs12979860
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.

Farnaes+, “Rapid whole-genome sequencing decreases infant morbidity and cost of hospitalization”, NPJ Genomic Medicine, 2018
“From 2019, all seriously ill children in UK will be offered whole genome sequencing as part of their care”
Much Larger Structural Variations!

**AUTISM**
Deletion of 593 kb

**OBESITY**
Walters, *Nature* 2010
Deletion of 593 kb

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

**UNDERWEIGHT**
Duplication of 593 kb

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

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

CNV: copy number variation
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

- Using intelligent architectures
  - Specialized HW with less data movement

- DNA is a valuable asset
  - Controlled-access analysis

- Population-scale genome analysis
  - Sequence anywhere at large scale!

- Avoiding erroneous analysis
  - E.g., your father is not your father

Bandwidth

Energy-efficiency & Latency

Privacy

Scalability

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

![Poster of the movie GATTACA](image1.png)

2015

![Poster of the movie TOMORROWLAND](image2.png)
Intelligent Architecture?

Modern systems

FPGAs

Heterogeneous Processors and Accelerators

(General Purpose) GPUs

Hybrid Main Memory

Sequencing Machine

Persistent Memory/Storage
Intelligent Architecture?

Modern systems

FPGAs

(General Purpose) GPUs

Heterogeneous Processors and Accelerators

Hybrid Main Memory

Persistent Memory/Storage

https://nanoporetech.com/products/smidgion
Privacy-Preserving Genome Analysis?

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 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

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

https://nebula.org/whole-genome-sequencing/
Rapid Surveillance of Disease Outbreaks?

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

Scalable SARS-CoV-2 Testing

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


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

https://blog.wego.com/7-crowded-places-and-events-that-you-will-love/
City-Scale Microbiome Profiling

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

A global metagenomic map of urban microbiomes and antimicrobial resistance

Danko+, "A global metagenomic map of urban microbiomes and antimicrobial resistance", Cell, 2021
Plague (Yersinia Pestis)

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 (Yersiniose)

What Is It?

Published: December, 2018

Plague is caused by Yersinia pestis and has been present for at least 2,000 years. Plague has caused skin sores that form black dots and can be fatal in up to one-third of the population, leading to death within days.

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

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
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
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?
NO machine can read the entire content of a genome
Genome Analysis

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

Why?!
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

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

**Genome Analysis**

**Sequencing**

```
CCCCCCTATATATACGTACGTACGT
ACGACTTTTAGTACGTACGT
TATATATACGTACGTACGT
ACGTACGCCCCTACGTACGT
TATATATACGTACGTACGT
ACGACTTTTAGTACGTACGT
TATATATACGTACGTACGT
ACGTACGCCCCTACGTACGT
TATATATACGTACGTACGT
ACGACTTTTAGTACGTACGT
TATATATACGTACGTACGT
ACGACTTTTAGTACGTACGT
TATATATACGTACGTACGT
ACGACTTTTAGTACGTACGT
TATATATACGTACGTACGT
ACGACTTTTAGTACGTACGT
TATATATACGTACGTACGT
```

1x10^{12} bases*

44 hours*

<1000 $

* NovaSeq 6000
High-Throughput Sequencers

... and more! All produce data with different properties.
<table>
<thead>
<tr>
<th></th>
<th>MinION Mk1B</th>
<th>MinION Mk1C</th>
<th>GridION Mk1</th>
<th>PromethION 24</th>
<th>PromethION 48</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read length</strong></td>
<td>&gt; 2Mb</td>
<td>&gt; 2Mb</td>
<td>&gt; 2Mb</td>
<td>&gt; 2Mb</td>
<td>&gt; 2Mb</td>
</tr>
<tr>
<td><strong>Yield per flow cell</strong></td>
<td>50 Gb</td>
<td>50 Gb</td>
<td>50 Gb</td>
<td>220 Gb</td>
<td>220 Gb</td>
</tr>
<tr>
<td><strong>Number of flow cells per device</strong></td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>24</td>
<td>48</td>
</tr>
<tr>
<td><strong>Yield per device</strong></td>
<td>&lt;50 Gb</td>
<td>&lt;50 Gb</td>
<td>&lt;250 Gb</td>
<td>&lt;5.2 Tb</td>
<td>&lt;10.5 Tb</td>
</tr>
<tr>
<td><strong>Starting price</strong></td>
<td>$1,000</td>
<td>$4,990</td>
<td>$49,995</td>
<td>$195,455</td>
<td>$327,455</td>
</tr>
</tbody>
</table>

[SAFARI](https://nanoporetech.com/products/comparison)
# Illumina Sequencers

<table>
<thead>
<tr>
<th></th>
<th>iSeq 100</th>
<th>MiniSeq</th>
<th>MiSeq</th>
<th>NextSeq 550</th>
<th>NextSeq 2000</th>
<th>NovaSeq 6000</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Run time</strong></td>
<td>9.5–19 hrs</td>
<td>4–24 hrs</td>
<td>4–55 hrs</td>
<td>12–30 hrs</td>
<td>24-48 hrs</td>
<td>13-44 hrs</td>
</tr>
<tr>
<td><strong>Max. reads per run</strong></td>
<td>4 million</td>
<td>25 million</td>
<td>25 million</td>
<td>400 million</td>
<td>1 billion</td>
<td>20 billion</td>
</tr>
<tr>
<td><strong>Max. read length</strong></td>
<td>$2 \times 150$ bp</td>
<td>$2 \times 150$ bp</td>
<td>$2 \times 300$ bp</td>
<td>$2 \times 150$ bp</td>
<td>$2 \times 150$ bp</td>
<td>$2 \times 250$ bp</td>
</tr>
<tr>
<td><strong>Max. output</strong></td>
<td>1.2 Gb</td>
<td>7.5 Gb</td>
<td>15 Gb</td>
<td>120 Gb</td>
<td>300 Gb</td>
<td>6000 Gb</td>
</tr>
<tr>
<td><strong>Estimated price</strong></td>
<td>$19,900</td>
<td>$49,500</td>
<td>$128,000</td>
<td>$275,000</td>
<td>$335,000</td>
<td>$985,000</td>
</tr>
</tbody>
</table>

[SAFARI](https://www.illumina.com/systems/sequencing-platforms.html)
How Does Illumina Machine Work?
How Does Illumina Machine Work?

Glass flow cell surface

Optical Sensor

DNA fragment = Read

Billions of Short Reads
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.

Figure is adapted from: [https://phys.org/news/2013-12-gene-sequencing-future.html](https://phys.org/news/2013-12-gene-sequencing-future.html)
How Does Nanopore Machine Work?

Nanopore is a nanoscale 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.

Check Nanopore virtual tour:
https://nanoporetech.com/resource-centre/minion-video

Figure is adapted from: https://phys.org/news/2013-12-gene-sequencing-future.html
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

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%)

HiFi Reads (PacBio)

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

https://labs.wsu.edu/genomicscore/illumina-sequencing/
https://pacbio.gs.washington.edu/
HiFi Reads (PacBio)

SMRTbell

Subread errors

Full-length (passes)

Subreads

Generate consensus read

HiFi read

https://ccs.how/how-does-ccs-work.html
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, Dhrithi Deshpande, Kodi Taraszka, Huwenbo Shi, Pelin Icer Baykal, Harry Taegyun Yang, Victor Xue, Sergey Knyazev, Benjamin D. Singer, Brunilda Balliu, David Koslicki, Pavel Skums, Alex Zelikovsky, Can Alkan, Onur Mutlu, Serghei Mangul

"Technology dictates algorithms: Recent developments in read alignment"

Genome Biology, 2021

[Source code]
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
Replied 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

https://twitter.com/mealser/status/1435223377644503040
Looking forward,
Will we be able to read
the entire genome sequence?