Agenda for Today

- Introduction to Genomics
  - Today: The ways we analyze genomes
  - Intelligent genome analysis
The purpose of **computing** is [to gain] **insight**, not numbers

Richard Hamming

We need to gain insights and observations much more efficiently than ever before.
Big Data is Everywhere

- **Astronomy**: 25 zetta-bytes/year
- **Twitter (now X)**: 0.5-15 billion tweets/year
- **YouTube**: 500-900 million hours/year
- **Genomics**: 1 zetta-bases/year

Angstrom (10^{-10} m) Era of Semiconductors

Intel Process Technology Innovations

*Graphic is for illustrative purposes only and is not to scale

https://siliconangle.com/2021/07/26/...
Problems with Data Analysis Today

Special-Purpose Machine for Data Generation

General-Purpose Machine for Data Analysis

FAST

SLOW

Slow and inefficient processing capability
Large amounts of data movement
Intelligent Data Analysis

- The **science and art** of revealing previously unknown and potentially **valuable information** or knowledge from **data**

- While **meeting certain goals**: 
  - Performance & latency
  - Energy consumption
  - Cost...

- Tailored for many **important applications**
  - AI/ML, Genomics, Medicine, Health...

Adapted from the definition of data mining
Four Key Current 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
Faster, Scalable & Accurate Genome Analysis

Understanding **genetic variations, species, evolution, ...**

Predicting the **presence and relative abundance of microbes** in a sample

Rapid surveillance of **disease outbreaks**

Developing **personalized medicine**

**SAFARI**

And, many, many other applications …
What is a Genome?

https://onlinelearning.hms.harvard.edu/hmx/courses/genetic-testing/
What is a **Genome**?

The entire set of DNA sequences in a cell

[Online Learning](https://onlinelearning.hms.harvard.edu/hmx/courses/genetic-testing/)

[Genome Glossary](https://www.genome.gov/genetics-glossary/)
How Large is a (Human) Genome?

Andreaturm, Zurich

~3.2 billion genomic bases*

https://www.gigon-guyer.ch/de/project/andreaturm/

*~100 meters = 80 characters per line, 40 lines per A4 page, each page is 0.1mm thick
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.
Constructing the Human Reference Genome

Now: Complete human reference genome

The complete sequence of a human genome

Abstract

In 2001, Celera Genomics and the International Human Genome Sequencing Consortium published their initial drafts of the human genome, which revolutionized the field of genomics. While these drafts and the updates that followed effectively covered the euchromatic fraction of the genome, the heterochromatin and many other complex regions were left unfinished or erroneous. Addressing this remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium has finished the first truly complete 3.055 billion base pair (bp) sequence of a human genome, representing the largest improvement to the human reference genome since its initial release. The new T2T-CHM13 reference includes gapless assemblies for all 22 autosomes plus Chromosome X, corrects numerous errors, and introduces nearly 200 million bp of novel sequence containing 2,226 paralogous gene copies, 115 of which are predicted to be protein coding. The newly completed regions include all centromeric satellite arrays and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies for the first time.

SAFARI https://genomics.ucsc.edu/2021/06/09/the-complete-sequence-of-a-human-genome/
https://time.com/collection/100-most-influential-people-2022/...
How About Other Species?

- Phi X174 virus: 5.386 Kilo bp
- E. coli O157:H7: 5.44 Million bp
- Homo Sapiens: 3.2 Billion bp
- Onion, Allium Cepa: 16 Billion bp
- Paris Japonica: 149 Billion bp
Human Chromosomes (23 Pairs)

Autosomes

Parent #1

Parent #2

Sex chromosomes

XX or XY

SAFARI
Human Chromosomes (23 Pairs)

Autosomes

Parent #1

Parent #2

= Adenine

= Thymine

= Cytosine

= Guanine

= Phosphate backbone

Sex chromosomes

XX or XY
DNA Under Electron Microscope

human chromosome #12 from HeLa’s cell
The Central Dogma of Molecular Biology

DNA Genotypes → Transcription → RNA → Translation → Protein Phenotypes

SAFARI
Most **cells** in a person's body have almost the same DNA and the same genes

- **Expression** of the genes **differs** between cells
- But **not all genes** are used or expressed by those cells
- Mutations may occur over time

---

**Cells of Different Organs and Tissues**

20,000-25,000 human genes

NIH 2009 National DNA Day
### Finding SNPs Associated with Complex Trait

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

SNP: single nucleotide polymorphism

**SAFARI** 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-Armstrong ², Nicholas Mancuso ³, Alvaro N. Barbeira ⁴, David A. Knowles ⁵,⁶, David Golan², Raili Ermel⁷, Arno Ruusalepp⁷,², Thomas Quertermous²,⁹, Ke Hao²,¹⁰, Johan L. M. Björkegren²,⁸,¹⁰,¹¹,¹², Hae Kyung Im²,¹⁴,*, Bogdan Pasaniuc ³,¹³,¹⁴,*, Manuel A. Rivas²,¹⁵,*, and Anshul Kundaje ¹,²,*. 

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**

<table>
<thead>
<tr>
<th>Basic Information</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>rs12979860</td>
</tr>
<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: 3%
- -: 4%
- 0: 0%

**Links to SNPedia**

<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
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

- Introduction to Genomics
  - Today: The ways we analyze genomes
  - Intelligent genome analysis
Does intelligent genome analysis really matter?
Intelligent Genome Analysis

- Fast genome analysis
  - Real-time analysis

- Large scale
  - Analyze the entire population

- Accurate analysis
  - Incorrect diagnosis of disease

- Using intelligent architectures
  - Specialized hardware with less data movement

- DNA is a valuable asset
  - Controlled-access analysis

Latency & Throughput

Scalability

Precision & Accuracy

Energy-efficiency & Bandwidth

Privacy
Intelligent Genome Analysis

Mohammed Alser, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu

“From Molecules to Genomic Variations: Intelligent Algorithms and Architectures for Intelligent Genome Analysis”
Computational and Structural Biotechnology Journal, 2022
[Source code]

Review

From molecules to genomic variations: Accelerating genome analysis via intelligent algorithms and architectures

Mohammed Alser*, Joel Lindegger, Can Firtina, Nour Almadhoun, Haiyu Mao, Gagandeep Singh, Juan Gomez-Luna, Onur Mutlu*

ETH Zurich, Gloriastrasse 35, 8092 Zürich, Switzerland
Fast Genome Analysis

Fast genome analysis in mere seconds using limited computational resources (e.g., personal computer or a mobile device).
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.

---

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

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

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

---

**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

---

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"
Rapid Surveillance of Disease Outbreaks

Real-time, portable genome sequencing for Ebola surveillance

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

University spinout's portable DNA sequencer has proved invaluable in tracking the global spread of coronavirus

COVID-19 Outbreak and PCR Testing

- Outbreaks
  - Urgent need for testing

- PCR Testing
  - Reliable results for **known regions to target**
  - High latency to get the right answer
  - Hard to customize to any region (e.g., COVID-19 variants)
Massively scaled-up testing for SARS-CoV-2 RNA via next-generation sequencing of pooled and barcoded nasal and saliva samples

Joshua S. Bloom, Laila Sathe, [..] Valerie A. Arboleda

*Nature Biomedical Engineering* 5, 657–665 (2021) | Cite this article

4675 Accesses | 110 Altmetric | Metrics

Bloom+, "Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing", *Nature Biomedical Engineering*, 2021
Large Scale Analysis

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

Goal: What organisms are present in a given environment and how abundant are they?

https://blog.wego.com/7-crowded-places-and-events-that-you-will-love/
Characterizing genomic variations of 49,962 Icelanders took **4.15 million CPU hours** or 83 CPU hours per sample on average.

"GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs", Nature Communications, 2019
Petabase-scale Viral Discovery

Building and Profiling 3,500 genomic assemblies needs 28,000 virtual AWS CPUs.

Edgar+, "Petabase-scale sequence alignment catalyses viral discovery", Nature 2022

https://serratus.io/
City-Scale Microbiome Profiling

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

Danko+, "A global metagenomic map of urban microbiomes and antimicrobial resistance", Cell, 2021
Accurate genome analysis to make life-critical decisions and improving the quality of life.
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?

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


Plague (Yersinia Pestis)

What Is It?

Published: December, 2018

Plague is caused by Yersinia and has lasted 2,000 years. Plague has caused skin sores that form before one-third of the population died within the subway just before he fell ill. Robert Stolarik for The New York Times

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


The New York Times

Bubonic Plague in the Subway System? Don’t Worry About It
data. Rob Knight, a professor in the department of pediatrics at the University of California, San Diego, calls this type of error “a \textit{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 ...
CAMI Consortium

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

“Critical Assessment of Metagenome Interpretation - the second round of challenges”, *Nature Methods*, 2022

[Source Code]
Nathan LaPierre, Mohammed Alser, Eleazar Eskin, David Koslicki, Serghei Mangul

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

Genome Biology, September 2020.

[Talk Video (7 minutes) at ISMB 2020]

[Source code]
Using Intelligent Architectures & Reliability

Challenging Environment in Outer Space

DNA sequencing at the picogram level to investigate life on Mars and Earth

Jyothi Basapathi Raghavendra, Maria-Paz Zorzano, Deepak Kumaresan & Javier Martin-Torres

Scientific Reports 13, Article number: 15277 (2023) | Cite this article

Abstract

DNA is an incontrovertible biosignature whose sequencing aids in species identification, genome functionality, and evolutionary relationships. To study life within the rocks of Earth and Mars, we demonstrate, in an ISO5 clean room, a procedure based on nanopore technology that correctly identifies organisms at picogram levels of DNA without amplification. Our study with E. coli and S. cerevisiae DNA samples showed that MinION sequencer (Oxford Nanopore Technologies) can unequivocally detect and characterise microbes with as little as 2 pg of input with just 50 active nanopores. This result is an excellent advancement in sensitivity, immediately applicable to investigating low biomass samples. This value is also at the level of possible background contamination associated with the reagents and the environment. Cultivation of natural and heat-treated Martian analogue (MMS-2) regolith samples, exposed to atmospheric water vapour or in increasing water concentrations, led to the extraction of 600–1000 pg of DNA from 500 mg of soil. Applying the low detectability technology enabled through MinION sequencer for a natural low biomass setting, we characterised the dry MMS-2 and found few soil-related organisms and airborne contaminants. The picogram detection level and the procedure presented here, may be of interest for the future Mars sample Return program, and the life research and planetary protection studies that will be implemented through the sample safety assessment.
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

(General Purpose) GPUs

Heterogeneous Processors and Accelerators

Hybrid Main Memory

Persistent Memory/Storage

Sequencing Machine

https://nanoporetech.com/products/smidgion
Worried about the 23andMe hack? Here’s what you can do.

A bad actor offered to sell information on 23andMe’s users, calling out Jewish people specifically

By Tahm Hustler
Updated October 13, 2023 at 3:18 p.m. EDT | Published October 12, 2023 at 7:00 a.m. EDT

Fourteen million people have shared their genetic information with 23andMe in hopes of learning more about their heritage. After a hack that appeared to target people with Jewish ancestry, some might be wondering how to cut ties with the company.
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.

Fig. 5. A completion attack.
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.

---

SAFARI  [https://nebula.org/whole-genome-sequencing/](https://nebula.org/whole-genome-sequencing/)
Faster, Scalable & Accurate Genome Analysis

Understanding genetic variations, species, evolution, ...

Predicting the presence and relative abundance of microbes in a sample

Rapid surveillance of disease outbreaks

Developing personalized medicine

And, many, many other applications ...
Applications are only limited by our imagination
Genome Editing

The Nobel Prize in Chemistry 2020
awarded "for the development of a method of genome editing"

DNA Computing

Massive parallelism to solve (hard) problems!

https://electronicsforyou.in/seminar-report-on-dna-computing/
New Personalized Shopping Paradigm

https://www.dnanudge.com/
Achieving Intelligent Genome Analysis

How and where to enable fast, accurate, cheap, privacy-preserving, and exabyte scale analysis of genomic data?
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?
- . . .
Algorithm-Arch-Device Co-Design is Critical

Computer Architecture (expanded view)

- Problem
- Algorithm
- Program/Language
- System Software
- SW/HW Interface
- Micro-architecture
- Logic
- Devices
- Electrons
Onur Mutlu and Can Firtina,
"Accelerating Genome Analysis via Algorithm-Architecture Co-Design"
[arXiv version]
A Bright Future for Intelligent Genome Analysis

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


Accelerating Genome Analysis: A Primer on an Ongoing Journey
DOI Bookmark: 10.1109/MM.2020.3013728

FPGA-Based Near-Memory Acceleration of Modern Data-Intensive Applications
DOI Bookmark: 10.1109/MM.2021.3088396

MinION from ONT

SmidgION from ONT
Genome Analysis in Real Life

**Sample Collection**

**Chopped DNA Fragments**

**Library Preparation**

**Sequencing**

**Computational Steps**

**Raw Sequencing Data**

**Genomic Analyses**
Many Genome Analysis Tools

**Sequencing Technology:**
- Illumina
- ONT
- PacBio (HiFi)

**Species:**
- E. Coli
- Human
- Yeast
- Zebra Fish
- Mice
- Fruit Fly

**Reference Genomes**

**Read Set**

**Read Correction (optional)**

**Sketching/Indexing**

**Mapping**
- Read Mapper:
  - BWA-MEM2
  - Minimap2
  - NGM-LR
  - Bowtie2
- Read Corrector:
  - HALC
  - LSC
  - Hercules
  - LoRDEC
  - LoRMA
  - Proovread
  - ColorMap

**Assembly**
- De novo Assembler (Long Reads):
  - Canu
  - Miniasm (uses Minimap2)
- De novo Assembler (Short Reads):
  - ABysS
  - SPAdes (small genomes)
- Assembly Polisher (optional):
  - Apollo
  - Racon
  - Pilon
  - Quiver (PB reads)
  - Arrow (PB reads, Not published yet)
  - NanoPolish (ONP reads)

**Variant Calling**
- Variant Caller:
  - LuMPY
  - VariationHunter
  - GATK
  - TaRDiS
  - Freebayes
  - DELLY
  - Platypus
  - SAMtools
  - Genome STRiP

**Polishing (optional)**

**Taxonomy Profiling**
- Kraken2
- Metalign
- MiCoP

**Coverage:**
- Low 2x - 30x
- Moderate 30x - 100x
- High >250x

**Read Length:**
- Short 100bp - 250bp
- Long 200bp – 2Mbp (>200bp)
- HiFi 10K-20Kbp

**Basecalling**

**Variant Calling**

**Assembly**

**Mapping**

**Sketching/Indexing**

**Read Set**

**Read Correction (optional)**

**Reference Genomes**
Next Week: Obtaining Sequencing Data

Generating Sequencing Data

- **DNA Extraction**
  - DNA Extraction
  - DNA Fragmentation
  - Library Preparation

- **Sequencing**
  - Illumina: Multiple images, .BCL/.CBCL
  - ONT: Squiggle, .FASTQ
  - PacBio: 30-hour movie, .BAM

Downloading Real Sequencing Data

- FASTQ/.FNA

Simulating Sequencing Data

- FASTQ/.FNA

---

Alser+, “Going From Molecules to Genomic Variations to Scientific Discovery: Intelligent Algorithms and Architectures for Intelligent Genome Analysis”, CSBJ, 2022
P&S Genomics

Lecture 2:
Introduction to Genome Analysis

Can Firtina

ETH Zürich
Fall 2023
26 October 2023