Welcome Back

Making Lemonade
Understanding Gill

- Roll your “R”s in back of throat
- “A” = “Uh” (i.e. “uhmbulunche”)
- “AF” = “AYE” (i.e. “playee”)
- “Ah” before “TH” = “T” (i.e. “panter”)
- “EY”/”AY” before “TH” = “D” (i.e. “togeder”/“bayding”)
- “O” = “OW” (i.e. “lowck”)
- “OO” = “OH” (i.e. “hook”)
- “Mother/Brother” = “Mader/Brader”
- “H”/”EH” (i.e. “seeter”)

outside resources use encouraged

Plan for the Day

• Overview the entire class

• Test run Zoom gadgetry

Announcements

• Course guidelines, lecture slides, office hours, etc. (posted on Canvas)

• Course communications via Piazza
  – Auditors please sign up too

• Pre-req: CS106 or equiv.
  – Programming knowledge (any language)
    • Give example
    – No Biology background needed

• Grade (see website for details, dates, honor code etc.):
  – Four homework assignments (15% each)
  – Take home exam (40%)
  – No project, no attendance, no videos
Announcements

- Three tutorials:
  - Introductory Biology Primer
  - Introduction to Text Processing
  - UCSC Genome Browser Tools (rec: bring your laptop)
  - Times/locations:
    - Follow website or Piazza
- Lots of genomics research happening on campus
  - If you enjoy this class many labs would love to have you!

Announcements

- Vote for CA Office Hour times on Piazza
  - Once a week for 2 hours starting the week of 1/18
- Homework 1 will be released 1/20 on our website
  - We recommend you start early, read lecture and tutorial slides and your notes carefully, and if need be, utilize office hours, and ask questions on Piazza
What will we study?

The most amazing “Turing Tape” in existence, your genome.

Not a programming language; Not an OS; It’s the OS dev Disk.

Organism – Cell - Genome

10¹⁵ different cells in an adult human.
The cell is the basic unit of life.
DNA = linear molecule inside the cell that carries instructions needed throughout the cell’s life ~ long string(s) over a small alphabet
Alphabet of four (nucleotides/bases) {A,C,G,T}. Strings of length 10⁴-10¹⁵

Genome:

...ACGTACGAC(TGACTAGCATCGACTAC)3ACTAGCAC...
One Cell, One Genome, One Replication

• Every cell holds a copy of all its DNA = its genome.
• The human body is made of ~10^{13} cells.
• All originate from a single cell through repeated cell divisions.

Talk about code reuse

• The same genome "runs" hundreds, likely thousands of very different cell types.
• Nature vs. Nurture – If life is like a river, our genome make your boat’s fabric...

"3*10^9 letters, driving 3*10^{12} cells, for 3*10^9 seconds."

The Genome as an integrated development environment

Read

Write

Understand

[Graph and images related to genome and development environment]
The Biggest Challenge in Genomics…

… is computational:

How does this encode this

\[\text{Program} \rightarrow \text{Output}\]

This “coding” question has profound implications for our lives


The Biggest Challenge in Genomics…

… is computational:

How does this encode this

\[\text{Program} \rightarrow \text{Bugs} \rightarrow \text{Output}\]

What genomic mutations predispose us to disease?

The Biggest Challenge in Genomics…

… is computational:

How does this encode this

Program  Bugs  Patching

What genomic mutations determine our drug response?

The Biggest Challenge in Genomics…

… is computational:

How does this encode this

Program  Bugs  Debugging

We are learning to alter our genome… But what to alter?

The Biggest Challenge in Genomics…

… is computational:

How does this encode this

Program  Output

Where did we come from? How are we different from each other?
The Biggest Challenge in Genomics…

... is computational:

How does this encode this

Program → Output

What in our genomes make us different from other species?

The Biggest Challenge in Genomics…

... is computational:

How does this encode this

Program → Output

Why is our genome full of “memory leaks” and cruft?

Why Genomics?

• Rewriting the book on our understanding of life
• Growing impact on everybody’s life
  • It is saving lives
  • Starting with COVID-19
  • Including in a very large family of (thousands of) disease we will drill into later in the class.
• Genomics is an information/computational science
  • You can change lives from your keyboard
We Change Lives Now

• Rewriting the book on our understanding of life

• Growing impact on everybody’s life
  • It is starting to save lives

• Genomics is an information/computational science
  • You can change lives from your keyboard

• This century is owned by Genomics

• “There is gold in them thar hills”
  • That gold can be yours

Why Genomics?

Genomics is affecting multiple fields of CS

ML
Storage
Encryption
Compression
Databases
HCI etc.

Even if you do not want to be a genomist, some of the most exciting challenges in your field may be at the interface with Genomics.

Most exciting things in Science happen at the interface of very different fields.
Computational Genomics

Genomics is three related fields bundled under one name:
• Technology development – build devices
• Functional genomics – do experiments
• Computational genomics – interpret results

Roles of computational biology (genomics):
• Summarize current experiment
• Discover the most exciting hypothesis / next experiment
• Develop new computational methods (to do the above)

CS273A helps you understand what to ask and why
Make you a better tool developer, and your own explorer

Why understand the why first?

Theoretical CS studies the hardness of questions.
A question is as hard as its easiest solution.
A lot of focus is put on how to answer questions.

In genomics (an empirical science) the temporal order is:
• What to ask and why?
• Can available/acquirable data answer it and if so how?
• Got data. Computed. What does my answer mean?

Advice: Reject the “us” (CS) and “them” (BIO) dichotomy
• Read what you need
• Develop your own taste for questions and answers
⇒ Be a better methods developer, discoverer

Field Goals
Class Goals

• Meet your genome (learn to surf, learn the surf)
• Understand genomic tools (theory, applications)
• DIY (pose questions, write & run tools, understand answers)
• Try to entice you to learn more / do more in Genomics

Class Topics

(0) Genome context:
  - cells, DNA, central dogma
(1) Genome content / genome function:
  - genes, gene regulation, epigenetics, repeats, SARS-CoV-2
(2) Genome sequencing:
  - technologies, assembly/analysis, technology dependence
(3) Genome evolution:
  - evolution = mutation + selection, main forces of evolution:
    - Neutral evolution, Negative selection, Positive selection
(4) Population genomics:
  - Human migration, paternity testing, forensics, cryptogenomics
(5) Genomics of human disease:
  - personal genomics, GxE disease types, deep dive monogenics
(6) Comparative Genomics:
  - Genomics of amazing animal adaptations, ultraconservation

Computational Genomics is (also) a Data Science
Genomes, Genes & Proteins

The most visible instructions in our genome are Genes. Genes explain exactly HOW to synthesize any protein. Proteins are the work horses of every living cell.

Genome:

...ACGTACGACTGACTAGCATCGACTACGACTAGCAC...

Genes & Gene Regulation

- Human genome encodes 20-25,000 genes (2% genome), >1,000,000 genomic switches that control genes (>10%).
- Gene = genomic substring that encodes HOW to make a protein.
- Genomic switch = genomic substring that encodes WHEN, WHERE & HOW MUCH of a protein to make.
**Epigenomics: transient writing “on” the genome**

---

**Repeats / Mobile Elements (“selfish/junk DNA”)**

---

**Bottom line**

- The Genome is really simple
  - Genes
  - Gene regulation

- The Genome is really fascinating
  - One code used in many contexts
  - Lots of reuse in the code
  - Output (you) is breathtaking

- Biology is a vast and deep ocean, but
  - Humanity’s biological knowledge is shallow
  - Dive anywhere and you quickly reach the frontier of human knowledge
SARS-CoV-2 & COVID-19

How can 30,000bp bring 3,000,000,000bp to their knees?


It's already a decade of Genomics

How can 30,000bp bring 3,000,000,000bp to their knees?
DNA sequencing

Genomes are awesome. Let's get'em.

[DNA double helix image]

Genomic Drama

HGC

Celera

2001

Getting the "blueprint of life"

DNA Sequencing Costs & Generations

First genome draft cost $1-3 billion ($1/bp)
1st Generation Genome Assembly

Some Terminology

- **read**: a 500-900 long word that comes out of sequencer
- **mate pair**: a pair of reads from two ends of the same insert fragment
- **contig**: a contiguous sequence formed by several overlapping reads with no gaps
- **supercontig** (scaffold): an ordered and oriented set of contigs, usually by mate pairs
- **consensus sequence**: sequence derived from multiple alignment of reads in a contig

Output: ACAGTTACAGGTT...

---

2nd Gen: Next Generation (re)Sequencing

Output = massive amounts of short, lower quality reads.
New Technologies + New Algorithms = New Opportunities

---

3rd Gen: cost effective, long reads

Just one example technology:

**The Envisioned Device: A SOLID STATE NANOPORE WITH EMBEDDED NANOTUBE SENSOR**

Output: very long reads of 10,000-100,000 basepairs each. Sequence "anything" you like. In a lab. Easy assembly.
Genome Browser Database

Primary table: positions, names, etc.

Auxiliary table: related data

Underlying Database (MySQL)

Visualize

Search & download

Genome Evolution

"Nothing in Biology Makes Sense Except in the Light of Evolution"

Theodosius Dobzhansky
Every Genome is Different

DNA Replication is imperfect – between individuals of the same species, even between the cells of an individual.

\[
\text{junk} \rightarrow \text{functional}
\]

\[
\ldots \text{ACGTACGACGACTGACTAGCATCGACTACGA} \ldots
\]

This has bad implications – disease, and good implications – evolution.

Genome mutation types: anything you can do to a string

- Deletion
- Inversion
- Translocation
- Duplication

Modes of evolution = Mutation + Selection

Genome similarity – mostly neutral DNA

Think paternity testing, think forensics

From neutral evolution alone
Cancer is a disease of the genome

that makes a cell veer off plan
(whatever its particular plan is)
and start dividing uncontrollably
ultimately throwing the organism off balance

Single Base Changes Can Be Detrimental

“Non-coding” mutations can be detrimental

Monogenic Diseases

Find patient variant that according to literature causes patient phenotypes

We’ll deep dive into a combined Genomics, ML, NLP (even Crypto) system that automates patient diagnosis — as an example of the future of Medicine.
Gene/Cell Therapy: Curing Genomics Defects

1. Get 'em
2. Fix 'em
3. Put 'em back

Evolution is not all bad! By far
We can sample ancient genomes (tens of thousands of years old).
We can reconstruct ancestral genomes (tens of millions of years old).
We learn from these ancient genomes what their owners were like.