Prerequisites

Some programming experience (R strongly encouraged)

Familiarity with probability, statistics, linear algebra and algorithms.
What is this course about?

Bioinformatics: Answering **biological questions** using tools from computer science, statistics and mathematics.

Machine Learning: Learning from data
A biological question

Personalized medicine
Cystic fibrosis

Genetic disease

Caused by mutations in the CFTR gene
Ivacaftor is a drug approved for patients with a mutation that changes G to D at position 551 in the protein encoded by CFTR.
Relapse in ALL

Acute Lymphoblastic Leukemia in children

Five-year survival rates ~80%

Poorer survival rates among African-Americans and Hispanics

Higher risk of relapse among children with native American ancestry

Yang et al. Nature Genetics 2011
Precision Medicine Initiative

$215 million investment

- Develop a cohort of 1 million volunteers
- Identify genomic drivers of cancer
- Privacy and standards for data sharing
Path to personalized medicine

Basic biology

Understand the architecture of a disease

Is the disease genetic or environmental?

What are the genetic and environmental factors?

Can we develop drug to target defective genes?

Disease prediction

Predict disease risk for an individual from genetic data
Path to personalized medicine

Basic biology

Understand the architecture of a disease

Is the disease genetic or environmental?
What are the genetic and environmental factors?
Can we develop drug to target defective genes?

Disease prediction

Predict disease risk for an individual from genetic data

Problems of statistical inference
What is this course about?

Bioinformatics: Answering biological questions using tools from computer science, statistics and mathematics.

Machine Learning: Learning from data
What is this course about?

Genomic revolution in biology

2001 Human genome project

2010 1000 genomes project
What is this course about?

Genomic revolution in biology

Number of human genomes

Number of bases

10^9
10^10
10^11
10^12
10^13
10^14
10^15
10^16
10^17
10^18
10^19
10^20

2000
2005
2010
2015
2020
2025

1st Sanger
IHGSC et al
Venter et al

1st Personal Genome
Levy et al

1st Illumina
Burton et al
Wang et al
Levy et al

1st PacBio
Chinnan et al

ExAC

Current Capacity

1000 Genomes

Recorded growth

Double every 7 months (Historical growth rate)

Double every 12 months (Illumina Estimate)

Double every 18 months (Moore’s Law)

What is this course about?

Personal genomics

We are the first and only genetic service available directly to you that includes reports that meet FDA standards.
What does this mean for us?

Statistics and computing important in obvious and subtle ways

Genetic data withdrawn amid privacy concerns

A new method of forensic DNA analysis has created an unexpected headache for scientists investigating the genetic roots of common diseases.

Harvard Professor Re-Identifies Anonymous Volunteers In DNA Study
Course goals

Identify biological questions to which computational and statistical thinking can make a difference.

Abstract these questions into a statistical model (Improve existing models).

Propose inference algorithms (Improve existing algorithms).

Apply the model to appropriate data.

Benchmarking

Are the results “better”? 

Interpret the results

Are they statistically sound? are they biologically meaningful?
Course goals

For CS/Stats students, identify important quantitative problems in Bioinformatics.

For Bioinformatics/Human Genetics students, learn a new set of tools and understand the principles behind tools being used in the field.
Course format

Introduce important problems in Bioinformatics.

Use these problems to motivate the study of learning algorithms.

Will study different aspects of learning algorithms

Statistical

Computational

Issues that arise in practice

Lectures will switch between studying the methodology and the applications.
Course format

**Homework** (5 worth 10% each). Will include programming and data analyses. Turn in typed hard copies (preferably LaTex) on the day it is due in class.

**In-class midterm** (20%) Date: **November 7**

**Project** (30%) Open-ended project and presentation. Development of statistical model/inference algorithm or application of existing methods to a new problem. Will post potential list of projects on CCLE. Welcome to propose your own project.

Will need to decide on a project by end of third week.

*(Optional)* readings to dig deeper into the material
Course format

Use CCLE to discuss and post questions

If emailing me directly, please include CM226 in the subject line
Course format

OH: Wednesday 1-2pm or by appointment

Boelter 4531D
Questions?
Crash-course in genomics

Molecular biology: How does the genome code for function?

Genetics: How is the genome passed on from parent to child?

Genetic variation: How does the genome change when it is passed on?

Population and evolutionary genetics: How does the genome vary across populations and species?

Genome sequencing: How do we read the genome?
Outline

Molecular biology: How does the genome code for function?

Genetics: How is the genome passed on from parent to child?

Genetic variation: How does the genome change when it is passed on?

What can we learn from genetic variation?

Genome sequencing: How do we read the genome?
Traits/Phenotype

Trait/phenotype: Any observable that is inherited

Height, eye color, disease status, cellular measurements, IQ

Instructions that modulate traits found in the genome
Cells and DNA

Genes contain instructions for making proteins.

Proteins act alone or in complexes to perform many cellular functions.

U.S. DEPARTMENT OF ENERGY
Outline

Molecular biology: How does the genome code for function?

Genetics: How is the genome passed on from parent to child?

Genetic variation: How does the genome change when it is passed on?

What can we learn from genetic variation?

Genome sequencing: How do we read the genome?
Genetics and inheritance

Typical human cell has 46 chromosomes

- 22 pairs of homologous chromosomes (autosomes)
- 1 pair of sex chromosomes

The chromosome painting collective
Genetics and inheritance

One member of each pair of homologous chromosomes comes from the father (paternal) and the other from the mother (maternal).

In males, Y from father and X from mother.

The chromosome painting collective
Outline

Molecular biology: How does the genome code for function?

Genetics: How is the genome passed on from parent to child?

Genetic variation: How does the genome change when it is passed on?

What can we learn from genetic variation?

Genome sequencing: How do we read the genome?
Causes of genetic variation

DNA not always inherited accurately

**Mutations**: changes in DNA

Changes at a single base (single nucleotide)

Can have more complex changes
More definitions

**Locus**: position along the chromosome (could be a single base or longer).

**Allele**: set of variants at a locus

**Genotype**: sequence of alleles along the loci of an individual

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>Locus 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A T C</td>
<td>C T T T A G G A</td>
</tr>
<tr>
<td>Individual 1</td>
<td>Paternal</td>
</tr>
<tr>
<td>A T C</td>
<td>T T T C A G A</td>
</tr>
<tr>
<td>Individual 2</td>
<td></td>
</tr>
<tr>
<td>A T C</td>
<td>T T T C A G A</td>
</tr>
<tr>
<td>A T C</td>
<td>T T T C A A A</td>
</tr>
</tbody>
</table>

Individual 1: (1,CT),(2,GG)

Individual 2: (1,TT), (2,GA)
Single Nucleotide Polymorphism (SNP)

Form the basis of most genetic analyses

Easy to study in high-throughput (million at a time)

Common (80 million SNPs discovered in 2500 individuals)

Two human chromosomes have a SNP every ~1000 bases
Most SNPs are biallelic.

Pick one allele as the reference allele.

Can represent a genotype as the number of copies of the reference allele.

Each genotype at a single base can be 0/1/2

Locus 1:C is reference
Individual 1 has genotype 1
Individual 2 has genotype 0
Single Nucleotide Polymorphism (SNP)

Form the basis of most genetic analyses

Easy to study in high-throughput

SNP arrays have millions of common SNPs

Common (80 million SNPs discovered in 2500 individuals)
Genotype and phenotype

Phenotype = function(Genotype, Environment)

Twins have similar phenotype

Identical twins (same genotype) can have different phenotypes

~30% are concordant for asthma, depression

Huang et al. Genetics in Medicine 2000
Segregation (Mendel’s first law)

\[
p(A) = 1 \quad A \quad a \quad p(A) = 0
\]

\[
\begin{array}{c}
\text{Generation 0} \\
\text{Gametes} \\
\text{Generation 1} \\
\text{Gametes}
\end{array}
\]

\[
p(A) = 0.5
\]
Back to genetic inheritance

Segregation (Mendel’s first law)
Back to genetic inheritance

Assortment (Mendel’s second law)

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>Locus 2</th>
<th>Generation 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>ab</td>
<td>0.25</td>
</tr>
<tr>
<td>Ab</td>
<td>aB</td>
<td>0.25</td>
</tr>
<tr>
<td>Ab</td>
<td>ab</td>
<td>0.25</td>
</tr>
<tr>
<td>ab</td>
<td>Ab</td>
<td>0.25</td>
</tr>
<tr>
<td>ab</td>
<td>Ab</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Back to genetic inheritance

Assortment (Mendel’s second law)

Not quite

<table>
<thead>
<tr>
<th>AaBb</th>
<th>Generation 0</th>
<th>Gametes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>Ab</td>
<td>aB</td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Back to genetic inheritance
Assortment (Mendel’s second law)

Not quite. **Crossover recombination**
Back to genetic inheritance

Assortment (Mendel’s second law)

Not quite. **Crossover recombination**

\[
\begin{align*}
AB & \quad \text{ab} & \quad Ab & \quad aB \\
(1-r)/2 & \quad (1-r)/2 & \quad r/2 & \quad r/2 \\
\end{align*}
\]

\(r:\) recombination fraction \((0<=r<=1/2)\)
Back to genetic inheritance

Assortment (Mendel’s second law)

Linkage: Positions nearby inherited together.

Important idea for mapping disease genes.
Back to genetic inheritance

Mutation and recombination (among other forces that we will learn about later) produce genetic variation

Mutation produces differences

Recombination shuffles these differences
Outline

Molecular biology: How does the genome code for function?

Genetics: How is the genome passed on from parent to child?

Genetic variation: How does the genome change when it is passed on?

What can we learn from genetic variation?

Genome sequencing: How do we read the genome?
What can we learn from genetic variation?

Evolution and history

Biological function and disease
What can we learn from genetic variation?

History of human populations learned from genetics

What can we learn from genetic variation?

Evolution and history

Biological function and disease
Genome-wide Association Studies (GWAS)
Outline

Molecular biology: How does the genome code for function?

Genetics: How is the genome passed on from parent to child?

Genetic variation: How does the genome change when it is passed on?

What can we learn from genetic variation?

Genome sequencing: How do we read the genome?
**Reading the genome**

**Goal**: Determine the sequence of bases along each chromosome

- Fragment the chromosomes
- Read each fragment
- **Assemble** the fragments
- Details depend on the technology

**Computationally hard**
The human genome project

Goals

Sequence an accurate reference human genome

Find the set of all genes

Draft published in 2001

High-quality version completed in 2003

Cost: ~$3 billion.

Time: ~13 years.

Two competing groups (public and private)
The human genome project

Major findings

Fewer genes than previously thought (~20K)

Pertea and Salzberg, *Genome Biology* 2010
The human genome project

Other outcomes

International collaborations

Power of computing
Reading the genome

Human genome provides a reference

Humans share most of their genome (~99.9%)

Can focus on reading the differences
Reading the genome

High-throughput genotyping

Hybridization of DNA molecules

Nucleotides bind to their complementary bases

A=T, C=G

Can be used to get the genotype at a chosen set of SNPs
Maps of genetic variation

International HapMap Project

**Goals:** Describe common patterns of genetic variation in human populations

Phase 1: Genotyped ~1 million SNPs from 270 individuals in 4 populations. Aims to capture all SNPs with a frequency of >5%.

Phase 3: 7 additional populations included

All data publicly available.
Reading the genome

Limitations of genotyping

Can only read SNPs that are on the chip

Biased by how these SNPs are chosen (e.g. common SNPs)
Reading the genome

High-throughput (or next-gen) sequencing

Technologies: Illumina, IonTorrent, PacBio

Can read small pieces of the genome (~100bp)

Two major differences

  Sequence hundreds of thousands of fragments in parallel

  Use the reference human genome to **find** the locations of the reads (and to infer mutations)
Reading the genome

Cost of genome sequencing

Moore's Law

NIH National Human Genome Research Institute

genome.gov/sequencingcosts
Maps of genetic variation

2500 individuals from 26 populations

Discover ~90 million SNPs

Includes >99% of SNPs with frequency >1%

All data publicly available
Maps of genetic variation

Many more such efforts underway

Example: Simons Genome Diversity Project: 260 genomes from 127 populations

Also publicly available
Other interesting data

EXAC data: Exomes from ~60,000 individuals
  Also publicly available

UK Biobank: 500,000 individuals with 200 phenotypes
  Not publicly available
Computational and statistical problems

Recurring theme:

How to better model a biological problem?
Computational and statistical problems

Recurring theme:

How to scale up our favorite algorithms to massive datasets?

Hundreds of thousands of individuals at hundreds of millions of SNPs measured for thousands of phenotypes

Tradeoffs

Computational efficiency vs statistical power/flexibility

Computational efficiency vs biological reality

Constraints from technology

Constraints from the data generation. e.g. privacy
Computational and statistical problems

Learning the genotype-phenotype map

Genome-wide association studies

Inferring genetic architecture

Correcting for confounding

Multiple hypothesis testing

Predicting disease

Causality

Inferring pathways

Population genomics

Inference of human history

Admixture inference
This class

Overview of the biological questions

Basic concepts in genetics

Genomes are inherited according to well-known rules (Mendel’s laws)

Genomes change

Genetic variation forms the starting point for inference.

Possible inferences: history, disease risk and many more

Advances in technology are allowing us to read many more genomes
Basic concepts in statistics

Given data, how do we perform inference?

What are the broad set of inferential tasks we are interested in?

How do we derive procedures for these tasks?

How do we evaluate these procedures?
What is this course about?

Biology is largely solved. DNA is the source code. For our bodies, now that gene sequencing is easy, we just have to read it.

It's not just "source code." There's a ton of feedback and external processing.

But even if it were, DNA is the result of the most aggressive optimization process in the universe, running in parallel at every level, in every living thing, for four billion years.

It's still just code.

Ok, try opening Google.com and clicking "view source."

Ok, I--... Oh my God.

That's just a few years of optimization by Google devs. DNA is thousands of times longer and way, way worse.

Jeez, biology is impossible.

https://xkcd.com/1605/