Introduction
Machine Learning for Bioinformatics
CM229, Spring 2016
Sriram Sankararaman
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.

Wow, biology is impossible.
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

Growth of DNA Sequencing

What is this course about?

Personal genomics

Your DNA. Knowledge about you.

- Receive 60+ personalized genetic reports
- Understand what your DNA says about your health, traits and ancestry
- Access interactive tools to share, compare and discover new with friends and family

We are the first and only genetic service available directly to you that includes reports that meet FDA standards.
What is this course about?

Cancer genomics

Metagenomics

http://www.scq.ubc.ca/metagenomics-the-science-of-biological-diversity/

Single-cell genomics

What does this mean for us?

Statistics and computing going to be even more important in obvious and subtle ways
What does this mean for us?

Statistics and computing going to be even more 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Genetic data withdrawn amid privacy concerns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.

Interpret the results -- are they statistically sound? are they biologically meaningful?
Course goals

For CS/Stats students, identify important quantitative problems.

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

Read primary research papers across diverse fields with little background.

Open-ended research project.
Course format

**Readings**: 1-2 per class (10%). Post a short summary, comments, critiques (<=5 sentences) on the readings to CCLE and respond to questions raised in class.

Each reading should take no more than 30 min. For papers that are mathematical, the aim is to get a general idea of the approach though you are welcome to dig into the details.

**Scribed lecture notes** (10%): One lecture. Please sign up. A LaTeX template will be provided.

**Homework** (3 worth 10% each). Will include programming and data analyses. Turn in hard copies on the day it is due in class.

**Project** (30% project + 20% paper) 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.
Course format

OH: Tuesday 10-11am 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?
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

Galton et al. 1877
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?

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

Genome sequencing: How do we read 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
DNA

Double-stranded
Sequence of \{A,C,G,T\}
Complementarity:
\[ A = T \]
\[ C = G \]
Proteins

Aspartic acid  Glutamic acid  Arginine  Lysine  Histidine  Asparagine  Glutamine  Serine  Threonine  Tyrosine

**Amino Acid**  **Side Chain**  **Polypeptide Backbone**  **Unfolded Polypeptide**

**AMINO ACID**  **SIDE CHAIN**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Side Chain</th>
<th>Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspartic acid</td>
<td>Asp D</td>
<td>negative</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>Glu E</td>
<td>negative</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg R</td>
<td>positive</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys K</td>
<td>positive</td>
</tr>
<tr>
<td>Histidine</td>
<td>His H</td>
<td>positive</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn N</td>
<td>Uncharged polar</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln Q</td>
<td>Uncharged polar</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser S</td>
<td>Uncharged polar</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr T</td>
<td>Uncharged polar</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr Y</td>
<td>Uncharged polar</td>
</tr>
</tbody>
</table>

**AMINO ACID**  **SIDE CHAIN**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Side Chain</th>
<th>Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala A</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly G</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Valine</td>
<td>Val V</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu L</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile I</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro P</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe F</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met M</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp W</td>
<td>Nonpolar</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys C</td>
<td>Nonpolar</td>
</tr>
</tbody>
</table>

**Polar Amino Acids**

**Nonpolar Amino Acids**

Alberts et al. Molecular Biology of the Cell
RNA

Single-stranded
Sequence of \{A,C,G,U\} (instead of T in DNA)
**Messenger RNA (mRNA)** code for proteins.
Others classes of RNA do not (non-coding RNA)

Alberts et al. Molecular Biology of the Cell
DNA, RNA and protein

Sequence of A,C,G,T
Sequence of A,C,G,U
Sequence of amino acids

Translate 4-character string to 20-character string
In groups of 3 characters (4^3 combinations)
DNA, RNA and protein

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUG</td>
<td>Met</td>
</tr>
<tr>
<td>UUU</td>
<td>Leu</td>
</tr>
<tr>
<td>UUA</td>
<td>Leu</td>
</tr>
<tr>
<td>UUA</td>
<td>Leu</td>
</tr>
<tr>
<td>UUG</td>
<td>Leu</td>
</tr>
<tr>
<td>CCA</td>
<td>Pro</td>
</tr>
<tr>
<td>CCA</td>
<td>Pro</td>
</tr>
<tr>
<td>CCG</td>
<td>Pro</td>
</tr>
<tr>
<td>CCC</td>
<td>Pro</td>
</tr>
<tr>
<td>AAA</td>
<td>Lys</td>
</tr>
<tr>
<td>GUA</td>
<td>Arg</td>
</tr>
<tr>
<td>GAG</td>
<td>Arg</td>
</tr>
<tr>
<td>GGG</td>
<td>Arg</td>
</tr>
<tr>
<td>GGC</td>
<td>Arg</td>
</tr>
<tr>
<td>GCC</td>
<td>Arg</td>
</tr>
<tr>
<td>GCA</td>
<td>Arg</td>
</tr>
<tr>
<td>UAA</td>
<td>Lys</td>
</tr>
<tr>
<td>UUA</td>
<td>Lys</td>
</tr>
<tr>
<td>UUG</td>
<td>Lys</td>
</tr>
<tr>
<td>UUC</td>
<td>Phe</td>
</tr>
<tr>
<td>UAC</td>
<td>Ser</td>
</tr>
<tr>
<td>UGU</td>
<td>Cys</td>
</tr>
<tr>
<td>UGC</td>
<td>Cys</td>
</tr>
<tr>
<td>UCU</td>
<td>Cys</td>
</tr>
<tr>
<td>CUC</td>
<td>Cys</td>
</tr>
<tr>
<td>CCA</td>
<td>Cys</td>
</tr>
<tr>
<td>CCC</td>
<td>Cys</td>
</tr>
<tr>
<td>GUC</td>
<td>Cys</td>
</tr>
<tr>
<td>GGU</td>
<td>Cys</td>
</tr>
<tr>
<td>GGA</td>
<td>Cys</td>
</tr>
<tr>
<td>UAU</td>
<td>Asn</td>
</tr>
<tr>
<td>AAG</td>
<td>Lys</td>
</tr>
<tr>
<td>AGA</td>
<td>Arg</td>
</tr>
<tr>
<td>AGG</td>
<td>Arg</td>
</tr>
<tr>
<td>CCG</td>
<td>Pro</td>
</tr>
<tr>
<td>CCA</td>
<td>Pro</td>
</tr>
<tr>
<td>CUC</td>
<td>Pro</td>
</tr>
<tr>
<td>CCC</td>
<td>Pro</td>
</tr>
<tr>
<td>UCA</td>
<td>Ser</td>
</tr>
<tr>
<td>AAC</td>
<td>Asn</td>
</tr>
<tr>
<td>ACG</td>
<td>Asn</td>
</tr>
<tr>
<td>AGC</td>
<td>Asn</td>
</tr>
<tr>
<td>UUA</td>
<td>Lys</td>
</tr>
<tr>
<td>UUG</td>
<td>Lys</td>
</tr>
<tr>
<td>UUU</td>
<td>Leu</td>
</tr>
<tr>
<td>UUC</td>
<td>Phe</td>
</tr>
<tr>
<td>UUA</td>
<td>Lys</td>
</tr>
<tr>
<td>UAU</td>
<td>Asn</td>
</tr>
<tr>
<td>UGA</td>
<td>Gua</td>
</tr>
<tr>
<td>UAA</td>
<td>Lys</td>
</tr>
<tr>
<td>UAG</td>
<td>Gua</td>
</tr>
<tr>
<td>UGU</td>
<td>Cys</td>
</tr>
<tr>
<td>GUC</td>
<td>Cys</td>
</tr>
<tr>
<td>GCG</td>
<td>Cys</td>
</tr>
<tr>
<td>GCC</td>
<td>Cys</td>
</tr>
<tr>
<td>GCA</td>
<td>Cys</td>
</tr>
<tr>
<td>UUG</td>
<td>Lys</td>
</tr>
<tr>
<td>UUA</td>
<td>Lys</td>
</tr>
<tr>
<td>UUC</td>
<td>Phe</td>
</tr>
</tbody>
</table>

Genetic code: RNA triplets to amino acids
Degenerate: Many to one
**Codon**: RNA triplet
**Stop codon**: signals end to translation
**Start codon**: signals start of translation (AUG)

Alberts et al. Molecular Biology of the Cell
DNA to RNA to protein

Code to guide the cell to where to start, stop and splice. Exons joined together and introns are removed in eukaryotes (splicing)

Alberts et al. Molecular Biology of the Cell
DNA to RNA to protein

Transcription

DNA to RNA to protein

Translation

Gene regulation

Which genes are turned on?

Function of cell type, time and environmental state

Many mechanisms

Alberts et al. Molecular Biology of the Cell
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?

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

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

**Ploidy**: number of sets of homologs

- Human cells mostly diploid
- Sex cells (egg and sperm) haploid
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
Genetics and inheritance

https://en.wikipedia.org/wiki/Chromosome
Diploid cell divides to form four haploid cells (gametes)

Alberts et al. Molecular Biology of the Cell
Genetics and inheritance

Haploid gametes combine to form diploid zygote
Genetics and inheritance

Meiosis contributes to variation

Daughter cells can differ from parents

Alberts et al. Molecular Biology of the Cell
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?

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

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

DNA not always inherited accurately

Mutations: changes in DNA

Causes of mutations

  DNA fails to copy accurately (error rates of 1 in a million to 1 in 100 million)

  Exposure to radiation/chemicals
Classes of variants

Single nucleotide mutations/point mutations

Replace one nucleotide with another

~30 per meiosis
Classes of variants

Structural variants

More bases of the genome affected by structural variants than point mutations

Harder to measure/assay

Alkan et al. Nature Reviews Genetics 2011
Classes of variants

Microsatellites or Short Tandem Repeats

Can be thought of structural variants

Operationally, these are short repetitive sequences that vary in the number of repeats

Have higher mutation rates than SNPs
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>
<th>Maternal</th>
<th>Paternal</th>
</tr>
</thead>
<tbody>
<tr>
<td>A T C C T T T A G G A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A T C T T T T C A G A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A T C T T T T C A G A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A T C T T T T C A A A</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

**Individual 2**: (1,TT), (2,GA)

If the two alleles at a locus are same, **homozygous**. Otherwise, **heterozygous**.
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)

Locus 1  Locus 2
A T C  C  T  T  T  A  G  G  A  Maternal
A T C  T  T  T  T  C  A  G  A  Paternal
A T C  T  T  T  T  C  A  G  A
A T C  T  T  T  C  A  A  A

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)

Identical twins (same genotype) can have different phenotypes

~30% are concordant for asthma, depression

People with same phenotype can have different genotypes

Huang et al. Genetics in Medicine 2000
Genotype and phenotype

Some mutations have no effect on a phenotype

A locus with two alleles a/A can affect a phenotype in many ways

**Additive:** $f(aA) = 0.5(f(AA) + f(aa))$

**Dominant:** A is dominant over a if $f(Aa) = f(AA)$

a is *recessive* relative to A
Genotype and phenotype

Example: Sickle-cell anaemia

Genetic disease with severe symptoms (pain, anaemia, stroke)

Caused by a mutation in the hemoglobin gene

Effect on DNA

- **NORMAL DNA:** GAG
- **MUTATION:** GAG to GTG
- **SICKLE CELL DNA:** GTG

Effect on protein

- **NORMAL PROTEIN:** GLU
- **MUTANT PROTEIN:** VAL

Effect on cell

- **NORMAL HEMOGLOBIN:**
- **CLumped HEMOGLOBIN:**

http://evolution.berkeley.edu/evolibrary/article/sicklecase_01
Genotype and phenotype

Example: Sickle-cell anaemia

Genetic disease with severe symptoms (pain, anaemia, stroke)

Caused by a mutation in the hemoglobin gene

Effect on individual
Negative effects: symptoms of disease
Positive effects: resistance to malaria

http://evolution.berkeley.edu/evolibrary/article/sicklecase_01
Genotype and phenotype

Example: Sickle-cell anaemia

Genetic disease with severe symptoms (pain, anaemia, stroke)

Recessive trait
Back to genetic inheritance

Segregation (Mendel’s first law)

\[ p(A) = 1 \]
\[ p(A) = 0 \]
\[ p(A) = 0.5 \]
Back to genetic inheritance

Segregation (Mendel’s first law)

<table>
<thead>
<tr>
<th>Parent Combination</th>
<th>Generation 0</th>
<th>Generation 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA x aa</td>
<td>AA x Aa</td>
<td>Generation 0</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>Generation 1</td>
</tr>
<tr>
<td></td>
<td>Aa</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parent Combination</th>
<th>Generation 0</th>
<th>Generation 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aa x Aa</td>
<td>AA</td>
<td>Generation 0</td>
</tr>
<tr>
<td></td>
<td>Aa</td>
<td>Generation 1</td>
</tr>
<tr>
<td></td>
<td>aa</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
</tr>
</tbody>
</table>
Back to genetic inheritance

Assortment (Mendel’s second law)

<table>
<thead>
<tr>
<th>Locus 1</th>
<th>Locus 2</th>
<th>Gametes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AaBb</td>
<td></td>
<td>Generation 0</td>
</tr>
<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

<table>
<thead>
<tr>
<th>AaBb</th>
<th>Generation 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>0.25</td>
</tr>
<tr>
<td>Ab</td>
<td>0.25</td>
</tr>
<tr>
<td>aB</td>
<td>0.25</td>
</tr>
<tr>
<td>ab</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Gametes
Back to genetic inheritance
Assortment (Mendel’s second law)

Not quite. **Crossover recombination**

![Diagram of chromosome pairing and crossover](image)

- **Parental**:
  - $AB$
  - $ab$

- **Recombinant**:
  - $Ab$
  - $aB$

1 meiosis with a single crossover
Back to genetic inheritance

Assortment  (Mendel’s second law)

Not quite. Crossover recombination

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

r: recombination fraction \((0 \leq r \leq 1/2)\)
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?

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

Genome sequencing: How do we read the genome?
Population genetics

Study of genetic variation in populations

What are the forces that affect variation? Mutation, recombination, others?

How different are genomes of two individuals within a population/across populations?

Where did our ancestors come from?

Why are some genetic variants more common than others? e.g. why is the sickle cell mutation common?

Applications

Understanding the map of genotype to phenotype.

Personalized medicine (pharmacogenomics)

Forensics

Conservation

Will return these topics in future lectures.
Population genetics

History of human populations learned from genetics

Li et al. *Science* 2008

Population genetics
Genes under selection

Genetic variant for lactase persistence

Gerbault et al. PTRSB 2011
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?
Reading the genome

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

**Template**

- 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

Major findings

Humans have genes not found in flies and worms

Lots of repetitive DNA
(15% is duplication of long sequences)
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

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

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.

International HapMap Project
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)
Illumina’s HiSeq X Ten: 1000-dollar genome

“It is a major human accomplishment on par with the development of the telescope or the microprocessor”—Michael Schatz
Maps of genetic variation

2500 individuals from 26 populations

Discover ~90 million SNPs, 3.6 million indels, 60k structural variants

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
Some questions to think about:

1. What are the problems unique to genomic data which would need tailored methods?

2. In addition to genomic data, it will now be possible to also collect high-dimensional phenotypes, e.g. data from FitBit. How does that change the things we can learn?

3. In the four aspects of data referred, which are the most interesting to you?
Relevant courses

CS269: Scalable Machine Learning by Prof. Ameet Talwalkar

TR 4:00-5:50pm