COMPUTATIONAL CHALLENGES OF PERSONAL GENOMES

ELEAZAR ESKIN
COMPUTER SCIENCE DEPARTMENT
The genomic revolution

- Human Genome Facts:
  - Length = 3,000,000,000
  - 10,000,000 variable positions
  - 2 individuals differ by 0.1%.
  - 1,000,000 positions measured per person.

- New technology for measuring variation has transformed Human Genetics:
  - Genetics of Diseases
  - Personalized Medicine
  - Cancer
  - Population genetics
Human Genetics and Applications

- Relate genetics to traits and diseases
Genetic and epigenetic variants + measurable environmental/behavioral factors would be used for a personalized treatment and diagnosis.
Example: Warfarin

An anticoagulant drug, useful in the prevention of thrombosis.
Example: Warfarin

Warfarin was originally used as rat poison.

Optimal dose varies across the population.

Genetic variants (VKORC1 and CYP2C9) affect the variation of the personalized optimal dose.
### Required Patient Information

- **Age:** 70
- **Sex:** Male
- **Ethnicity:** Non-Hispanic
- **Race:** African American or Black
- **Weight:** 170 lbs or 77.3 kgs
- **BSA:** 1.93
- **Height:** 5 feet and 9 inches or 175.3 cms
- **Smokes:** No
- **Liver Disease:** No
- **Indication:** Atrial fibrillation
- **Baseline INR:** 1
- **Target INR:** 2.5
- **Randomize & Blind:** No
- **Amiodarone/Cordarone® Dose:** 100 mg/day
- **Statin/HMG CoA Reductase Inhibitor:** No statin
- **Any azole** (eg. Fluconazole): No
- **Sulfamethoxazole/Septra/Bactrim/Cotrim/Sulfastrim:** No

### Genetic Information

- **VKORC1-1639/3673:** GG (warfarin insensitive)
- **CYP4F2 V433M:** CC (wildtype)
- **GGCX rs11676382:** CC (wildtype)
- **CYP2C9*2:** CT (heterozygous)
- **CYP2C9*3:** Not available/pending
- **CYP2C9*5:** Not available/pending
- **CYP2C9*6:** Not available/pending

[Accept Terms of Use]

[ESTIMATE WARFARIN DOSE]
Effects of Common Variants on Lifetime Risk

- Alzheimer's disease
- Breast cancer
- Type 2 Diabetes
- BMI/Obesity
Personalized Genomics Road Map

1. Estimate the contribution of the genetic vs. environmental factors to the disease.

2. Find the building blocks of the disease model: the genetic factors, the environmental factors, interactions.

3. Construct a disease model that predicts treatment outcomes and prevents disease.
1. Estimate the contribution of the genetic vs. environmental factors to the disease.

2. Find the building blocks of the disease model: the genetic factors, the environmental factors, interactions.

3. Construct a disease model that predicts treatment outcomes and prevents disease.
Genome-Wide Association Study (GWAS)

- 2007 Breakthrough of the Year
- More than 50 genes discovered to affect dozens of common diseases.
- Weekly news reports of “Scientists discovery gene causing ________!”
Comparing the DNA contents of two populations:

- Cases - individuals carrying the disease.
- Controls - background population.

Differences within a gene between the two populations is evidence the gene is involved in the disease.
Association Analysis

Cases: (Individuals with the disease)
AGAGCA<sup>AG</sup>GTCGACAGT<sup>GA</sup>TATAGCT<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCC<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGT<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGTC
AGAGCA<sup>AG</sup>GTCGACAGT<sup>GA</sup>TATAGT<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCG<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGC<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGG<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCA<sup>AG</sup>GTCGACAGT<sup>GA</sup>TATAGCT<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCG<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGG<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC

Controls: (Healthy individuals)
AGAGCA<sup>AG</sup>GTCGACAT<sup>GA</sup>TATAGT<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGT<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGC<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCA<sup>AG</sup>GTCGACAT<sup>GA</sup>TATAGT<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGG<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGG<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC
AGAGCC<sup>GA</sup>GTCGACAT<sup>GA</sup>TATAGG<sup>GA</sup>CTACATGAGATCG<sup>GA</sup>CATGAGATG<sup>GA</sup>TAGAGCA<sup>GA</sup>GTAGAGATCG<sup>GA</sup>CATGATAGCC

Associated Variant
Sequencing Costs

Baseline information
Cost of genome sequencing compared with Moore’s law for computers

Source: The Economist
July 17th, 2010

Source: Broad Institute
Published Genome-Wide Associations through 6/2009, 439
published GWA at $p \leq 5 \times 10^{-8}$

NHGRI GWA Catalog
www.genome.gov/GWAStudies
Published Genome-Wide Associations through 6/2010, 904 published GWA at $p \leq 5 \times 10^{-8}$ for 165 traits

NHGRI GWA Catalog
www.genome.gov/GWAStudies
- The relative risks are normally low; typical Relative Risk is 1.1-1.2.
- There is a lot more that needs to be found.
  - Rare SNPs
  - Gene-gene interactions
  - Gene-environment interactions

The case of the missing heritability
Active Research Problem: Short Read Re-sequencing

Where are my mutations?

Next generation sequencing.

Cheap sequencing.

“Short Reads”
Short Read Sequencing Problem (A Computer Science Problem)

- Short read sequencers generate random short substrings from the DNA sequence of a certain length.

How do we recover the original sequence?
Short Reads Difficulties

- We don’t know where each read comes from!
- Can’t identify where the mutations are!
- What do we do?

ATGAGATCGGTAGAGCCGTGAGAT
GAGCAGTCGACAGGTATAGTCTAC
AGAGCAGTCGACAGGTATAGTCTTA
TGAGATCGACATGATAGCCAGAGC
TAGCCAGAGCAGTCGACAGGTTATA
GATAGCCAGAGCAGTCGACAGGTA
GAGATCGACATGATAGCCAGAGCA
GCAGTCGACAGGTATAGTCTACAT
AGCAGTCGACAGGTATAGTCTACA
TCGACATGAGATCGGTAGAGCCGT
CAGTCGACAGGTATAGTCTACATG
GAGATCGACATGATAGCCAGAGCA
GTAGAGCCGTGAGATCGACATGAT
Key Idea: “Re”-Sequencing

We know that my genome is very close to the Human genome.

**My Genome:**
TACATGAGATCGACATGAGATCGGTAGAGCCCGTGAGATC

**A Sequence Read:**
TCGACATGAGATCGGTAGAGCCCGT

**The Human Genome:**
TACATGAGATCGACATGAGATCGGTAGAGCCTGTGAGATC
TCGACATGAGATCGGTAGAGCCCGT

**Recovered Sequence:**
TACATGAGATCGACATGAGATCGGTAGAGCCCGTGAGATC
“Re”-Sequencing Challenges
(Why do we need Computer Science?)

- Sequences are long!
  - Human Genome is 3,000,000,000 long.

- Sequencers generate many reads!
  - A single run generates over 1,000,000,000 reads.

- We need efficient algorithms to “map” each read to its location in the genome.
  - A trivial mapping algorithm will take thousands of years to compute for a genome.
"Re"-Sequencing Problems

The Human Genome:
TACATGAGATCCACATGAGATCTGTGTA

My Genome:
TACATGAGATCGACATGAGATCGGGTACATGAGATCCACAT

Repeated Region

A Sequence Read:
ACATGAGATCGACAT

The Human Genome:
TACATGAGATCCACATGAGATCTGTGTA
ACATGAGATCGACAT
ACATGAGATCGACAT

Recovered Sequence:
TACATGAGATCGACATGAGATCGGGTACATGAGATCGACAT
“Re”-Sequencing Problems

The Human Genome:
TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

My Genome:
TACATGAGGGGGGGGAGATCGGTACATGAGATCCACAT

A Sequence Read:
GAGGGGGGGGG

The Human Genome:
TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT
GAGGGGGGGGG

Too many mismatches to match the read to the reference. Since we don’t know where it came from, we can’t identify the difference in the target sequence.
“Re”-Sequencing: Insertions

**My Genome:**
TACATGAGATCCACATAGAGATCTGTAGAGCTGTGAGATC

**A Sequence Read:**
CCACATAGAGATCTGTAGAGCTGTG

**The Human Genome:**
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGTG

How do we deal with this case?
Family Tree of Humanity

Do I have ancestors from Europe in 1000 AD?

- Adam
- Eve
- Eleazar’s Dad
- Eleazar’s Mom
- Eleazar
- Eleazar’s Sister
- Another Dad
- Another Mom
- Another Person
Family Tree of Humanity

Adam  Eve

Eleazar’s Dad  Eleazar’s Mom

Eleazar  Eleazar’s Sister

Another Dad  Another Mom

Another Person
We don’t need the tree, we need to identify the ancient relationships.
Identifying Related Individuals

Identify shared segments. Shared segments = related!

DNA Sharing:
- Siblings: 50%
- Cousins: 25%
- 2\textsuperscript{nd} Cousins: 12.5%
- 3\textsuperscript{rd} Cousins: 6.25%
- ...
- 7\textsuperscript{th} Cousins: .4%

New technology allows identifying small shared regions.
(Corresponding to Ancient Relationships)
Maternal Tree (Mitochondria)

How are we related? Who are our common ancestors? Can we count them?
Modeling Mutation Frequency

Spain: 0.8
Germany: 0.5
Russia: 0.2
Genomes and geography
Genomes and Geography

- 1000 individuals
- Each with ancestry from single country in Europe
Genomes and World Geography

Africa
Europe
Middle East
Central South Asia
East Asia
Oceania
America
Closing…

- The most exciting phrase to hear in science - the one that heralds new discoveries - is not "Eureka!" but "That's funny...”.

- I. Asimov