Computational Genetics
Spring 2014
Lecture 1

Eleazar Eskin

University of California, Los Angeles

Introduction to Computational Genetics and the HapMap

Lecture 1.

March 31st, 2014

Course Requirements

- Prerequisites
  - **Knowledge of a programming language**
  - **A statistics course.**
- Requirements
  - **3 Short Homework Assignments**
  - **5 Paper Responses**
  - **Midterm Exam – April 23rd**
  - **Final Exam – June 9th**
  - **Final Project**
  - **Extra problems and bigger project for graduate students.**
- Grading Basis
  - **Homeworks 20%.**
  - **Midterm Exam 20%.**
  - **Final Exam 20%.**
  - **Paper Responses 10%.**
  - **Final Project 30%.**

Computational Genetics – Part I

- March 31st - Introduction to Computational Genetics + Background in Statistics
- April 2nd - Disease Genetics and Association Analysis + Association Examples
- April 4th - NO DISCUSSION
- April 7th - Indirect Association + Measuring Statistical Power + The HapMap
- April 9th - Multiple Testing Correction + Association Study Design
- April 11th - Introduction to R Statistical Programming (Discussion)
- April 14st - Multi-Variate Normal Distribution for Association Statistics
- April 16th - Meta-Analysis + Imputation + Rare Variants
- April 18th - Association Studies Review (Discussion)
- April 21st - Midterm Review
- April 23rd - MIDTERM
- April 25th - Sequencing Software and Analysis Tools (Discussion)
- April 28th - Sequencing + Read Mapping + Burroughs Wheeler Transform
- April 30th - Sequencing Coverage + Sequence SNP Identification
- May 2nd - Sequencing Software and Analysis Tools (Discussion)

Computational Genetics – Part II

- May 5th - Sequence Assembly
- May 7th - Copy Number Variation + Structural Variation
- May 9th - Final Project Presentations (Discussion)
- May 12th - RNA Sequencing + Meta-Genomics
- May 14th - Population Structure + Mixed Models (Advanced Topic)
- May 16th - Final Project Presentations (Discussion)
- May 19th - Mouse Genetics (Advanced Topic)
- May 21st - Identity-by-Descent Inference + Pedigree Inference
- May 23rd - Final Project Presentations (Discussion)
- May 26th - HOLIDAY
- May 28th - Final Project Presentations
- May 30th - Final Project Presentations (Discussion)
- June 2nd - Final Project Presentations
- June 4th - Final Project Presentations
- June 6th - Final Exam Review (Discussion)
- June 9th - Final Exam (Non-cumulative) – Monday 8:00am-11:00am

Course Goal: Training in Interdisciplinary Computational Research

- Reading papers outside Computer Science with no background.
- Identifying Computational Problems or ways we can contribute.
- Formalizing/abstracting computational problems.

- Open ended Final Project.

Final Projects

- An interdisciplinary Computational Research Project
  - **Important Biological Problem**
  - **Formalize a Computational Problem**
    - Identify Objective Function/Benchmark
    - Identify Competing/Baseline Solutions
  - **IDEA!**
    - Better solution to computational problem.
  - **Evaluate solution compared to benchmarks**
  - **Identify Implications**
- Many problems to choose from.
- Different difficulty levels for grads/undergrads

Final Projects

- 15+ available projects in Association Studies
- 15+ available project in Sequencing
- Need to decide on a project by April 11th.

Final Projects

- 4 levels of difficulty
  - **Easy**
  - **Medium**
  - **Hard**
  - **Very Hard**
- Undergrads can do an easy project.
- Grads must do a medium or harder project.
- Harder projects get more extra credit and later presentation dates.
- No group projects.

Paper Reading Responses

- CourseWeb Discussion Forum
- Mandatory Participation
- 1 Question due on Monday
- 2 Responses due on Wednesday

- This week, both videos (Eric Lander and NOVA).
  - **Post questions by Wednesday and responses due on Friday.**

Eric Lander Video

- Secrets of the Human Genome
- http://hulk03.princeton.edu:8080/WebMedia/flash/lectures/20100419_publect_lander.shtml
- Better version on iTunes.

NOVA
“Cracking your Genetic Code”

- NOVA from March 2012
- http://www.pbs.org/wgbh/nova/body/cracking-your-genetic-code.html

Genomics Options for CS Majors

- Sci-Tech Electives for CS Majors
  - Lower Division Courses in Chemistry and Biology which are Prereqs for Upper Division Biology Courses.
  - No other way to take biology courses!
- Technical Breadth Area in Genomics
  - Mostly upper division courses in "genomics" area
  - Taught by faculty in the Bioinformatics program
  - Many good options and prereqs satisfied by Sci-Tech electives option.

Biology and Chemistry Prereqs

- Main required sequence is Life Sciences 2, 3, and 4.
- These courses also require Chemistry 20A, 20B, 30A and Mathematics 31A.
- Life Sciences 2 and Chemistry 20A can be taken as Engineering + GE requirements.
- Mathematics 31A taken by our students.

Sci-Tech Electives (within CS Major)

1. Life Sciences 3 - Introduction to Molecular Biology

   (prereq Life Sciences 2, Chem 30A)

2. Chem 20B - Chemical Energetics and Change

   (prereq Chem 20A, Math 31A)

3. Chem 30A - Organic Chemistry I: Structure and Reactivity

   (prereq Chem 20B)

Technical Breadth Area in Genomics

3 Courses from this list:

- Life Sciences 4 - Genetics
  - (prereq Life Sciences 2,3, Chemistry 20A, 30A)
- Molecular Cellular and Developmental Biology 144 - Molecular Biology
  - (prereq Life Sciences 3,4)
- Human Genetics 144 - Genomic Technologies
- Ecology and Evolution 135 – Population Genetics
  - (prereq Life Sciences 4)
- Molecular Cellular and Developmental Biology 172 - Genomics and Bioinformatics
  - (prereq Molecular Cellular and Developmental Biology 144)
- Physiological Sciences 125 - Molecular Systems Biology
  - (prereq Life Sciences 2,3,4)

Bioinformatics Minor

- Bioinformatics is an important interdisciplinary research area with tremendous graduate training and industry opportunities.
- Strong group of faculty engaged in active research at UCLA
- Numerous existing course offerings available at UCLA.

- Minor organizes available courses into a coherent undergraduate academic program.
  - Graduating students will be positioned to apply to graduate programs in Bioinformatics.
  - Graduating students will be positioned to enter biotechnology industry.

Bioinformatics Minor Structure

- 8 course minor (5 upper division, 3 lower division)
- Computational Biology Seminar Course
  1. "Introduction to Computational Systems Biology"
     - CS 184 taught by Joe Distefano (lectures by many Bioinformatics faculty)
- Core bioinformatics courses
  2. "Introduction to Bioinformatics"
     - Chem 160A, CS 121 taught by Chris Lee
  3. "Computational Genetics"
     - CS 124, Human Genetics 124 taught by Eleazar Eskin
- Additional required algorithms course
     - CS 180 or Math 182
- Remaining upper division course is an electives
- Additional lower division courses are prerequisites
- Minimum of 20 units in addition to Major
- Up to 8 units of research can be applied to Minor

Bioinformatics Lower Division Courses

- Three required courses are prerequisites for upper division courses
  1. Advanced Programming
     - PIC 10C or CS 32
  2. Linear Algebra and Applications
     - Math 33A
  3. Introduction to Molecular Biology
     - Life Sciences 3, 23

Bioinformatics Upper Division Electives

- Statistics 100B - Introduction to Mathematical Statistics OR Biostatistics 100B - Introduction to Biostatistics
- Computer Science 170A - Mathematical Modeling and Methods for Computer Science
- Electrical Engineering 102 - Systems and Signals
- Electrical Engineering 141 - Principles of Feedback Control
- Computer Science 122 - Algorithms in Bioinformatics and Systems Biology
- Computer Science 229 - Current Topics in Bioinformatics
- Computational and Systems Biology 186 - Computational Systems Biology: Modeling and Simulation of Biological Systems
- Human Genetics 144 - Genomic Technologies
- Ecology and Evolution 135 – Population Genetics
- Molecular Cellular and Developmental Biology 172 - Genomics and Bioinformatics
- Physiological Sciences 125 - Molecular Systems Biology
- Molecular Cellular and Developmental Biology 144 - Molecular Biology OR Microbiology Immunology and Molecular Genetics 132 - Cell Biology of Nucleus OR Chemistry or Biochemistry 153B - Biochemistry: DNA, RNA, and Protein Synthesis

Gateway Course

- Students are required to take 2 unit CS 184 "Introduction to Computational Systems Biology"
  - Seminars by faculty in computational biology (including many Bioinformatics faculty)
- Students encouraged to take seminar course as early as possible.
- Gateway course will be shared with other quantitative biology minors currently being proposed to build undergrad computational biology community.

Course Plan: Computer Science Major

- Courses part of Major required courses:
  - CS 32, Math 33A, CS 180.
- Students will take as Engineering GE:
  - Chem 20A, Life Sciences 2.
- Students will take Sci-Tech Bio option (part of Major):
  - Chem 20B, Chem 30A, Life Sciences 3.
- Students will take CS 184 as an introduction to the area.
- Students can take CS 121 and CS 124 as electives for their CS major.
- Students will take additional bioinformatics elective courses to fulfill the minor requirements including 8 units of research.

- Students who take the optional Technical Breadth Area in Computational Genomics can take prerequisites and electives in the program:
  - Life Sciences 4, + 2 Bioinformatics electives

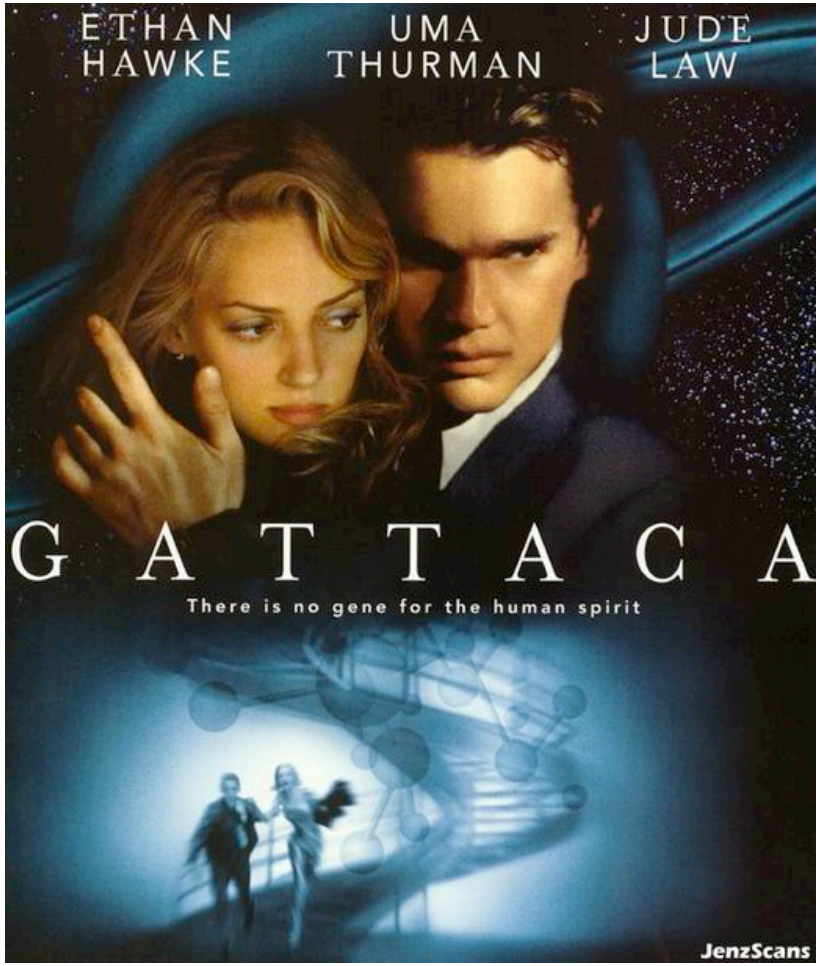
Introduction to Computational Genetics and the HapMap

Lecture 1.

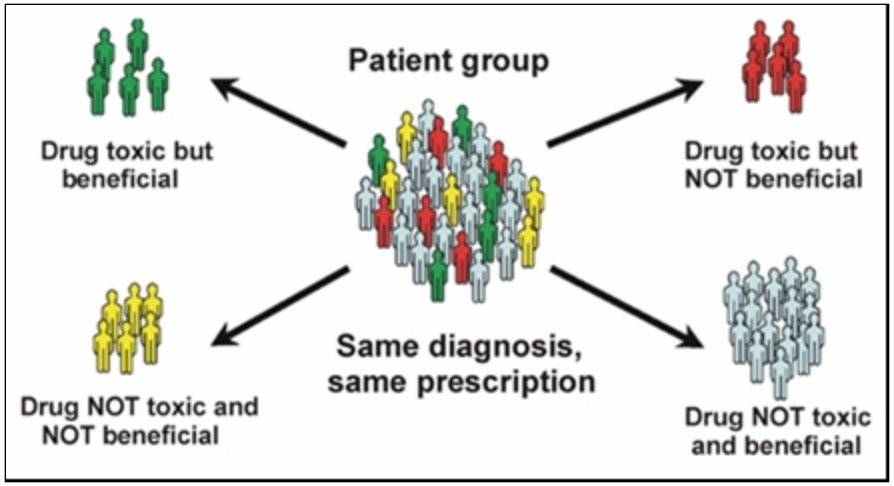
March 31st, 2014

Human Genetics and Applications

- Relate genetics to traits and diseases



The Vision of Personalized Medicine

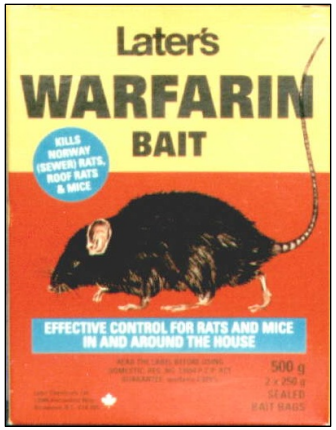


Genetic and epigenetic variants + measurable environmental/behavioral factors would be used for a personalized treatment and diagnosis

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.

WARFARINDOSING

www.WarfarinDosing.org

- Warfarin Dosing
- Clinical Trial
- Outcomes
- Hemorrhage Risk
- Patient Education
- Contact Us
- References
- Glossary
- About Us

User:
Patient:
Version 2.31
Build : Sep 05, 2011

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

Amiodarone/Cordarone® Dose: 100 mg/day

Statin/HMG CoA Reductase Inhibitor: No statin

Any azole (eg. Fluconazole): No

Sulfamethoxazole/Septra/Bactrim/Cotrim/Sulfatrim: 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

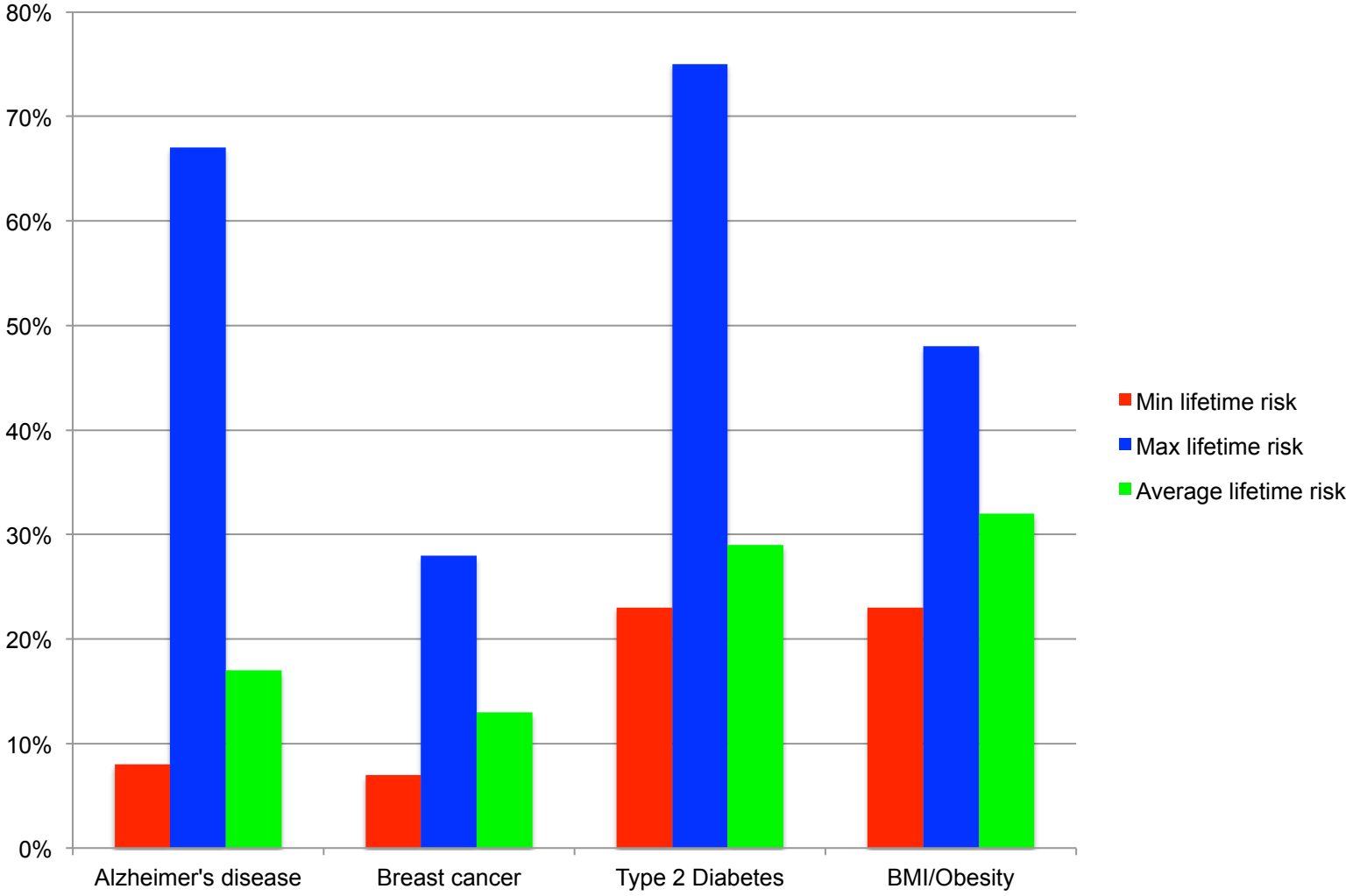
The Human Genome Project

"I would be willing to make a predication that within 10 years, we will have the potential of offering any of you the opportunity to find out what particular genetic conditions you may be at increased risk for…"

"What we are announcing today is that we have reached a milestone … that is covering the genome in…a working draft of the human sequence."

Washington, DC
June, 26, 2000

Effects of Common Variants on Lifetime Risk



Personalized Genomics Road Map

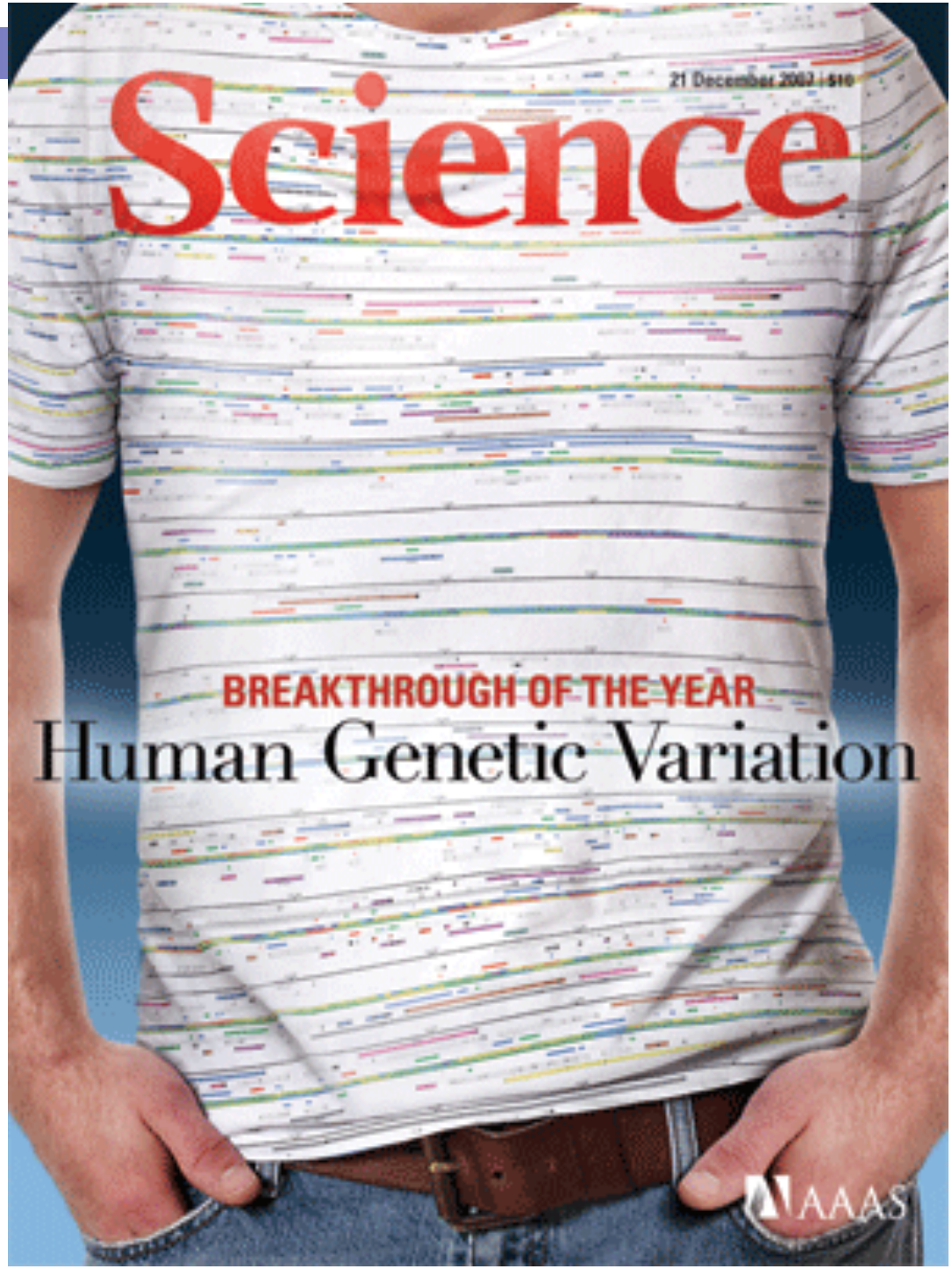
1. Estimate the contribution of the genetic vs. environmental factors to the disease.
2. Find the building blocks of the disease model: the genetic factors, the environmental factors, interactions.
3. Construct a disease model that predicts treatment outcomes and prevents disease.

Personalized Genomics Road Map

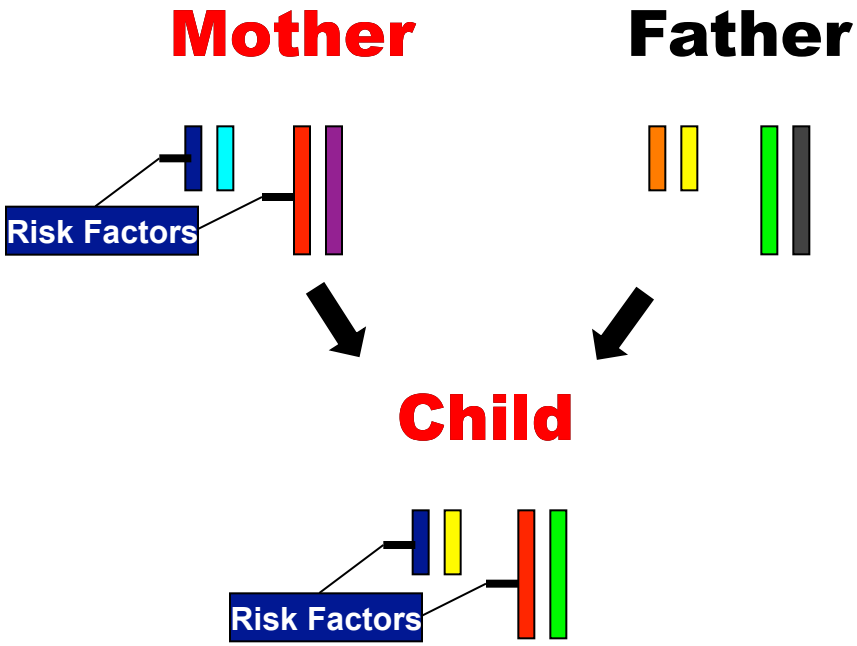
1. Estimate the contribution of the genetic vs. environmental factors to the disease.
2. Find the building blocks of the disease model: the genetic factors, the environmental factors, interactions.
3. Construct a disease model that predicts treatment outcomes and prevents disease.

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



Human Genetics



- Disease Risk
  - **"genetic" factors account for 20%-80% of disease risk.**
  - **Many genes contribute to "complex" diseases.**
- Personalized Medicine
  - **Treatment decisions influenced by diagnostics**
- Understanding Disease Biology
  - **New drug targets.**
  - **Understanding of mechanism of disease.**

Where are the risk factors? (Genetic Basis of Disease)

Disease Association Studies
The search for genetic factors

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.

Single Nucleotide Polymorphisms (SNPs)

AGAGC**C**GTCGACA**G**GTATAG**C**CTA
AGAGC**C**GTCGACA**T**GTATAG**T**CTA

AGAGC**A**GTCGACA**G**GTATAG**T**CTA
AGAGC**A**GTCGACA**G**GTATAG**C**CTA

AGAGC**C**GTCGACA**T**GTATAG**C**CTA
AGAGC**A**GTCGACA**T**GTATAG**C**CTA

AGAGC**C**GTCGACA**G**GTATAG**C**CTA
AGAGC**C**GTCGACA**G**GTATAG**C**CTA

- Human Variation
  - **Humans differ by 0.1% of their DNA.**
  - **A significant fraction of this variation is accounted by SNPs.**

Association Analysis

Cases: (Individuals with the disease)

AGAGCAGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCCGTCGACATGTATAGTCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCGACATGATAGTC
AGAGCAGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCAGTCGACAGGTATAGCCTACATGAGATCAACATGAGATCGGTAGAGCAGTGAGATCGACATGATAGCC
AGAGCCGTCGACATGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATCAACATGATAGCC
AGAGCCGTCGACATGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCAACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCAACATGATAGTC
AGAGCAGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGCC

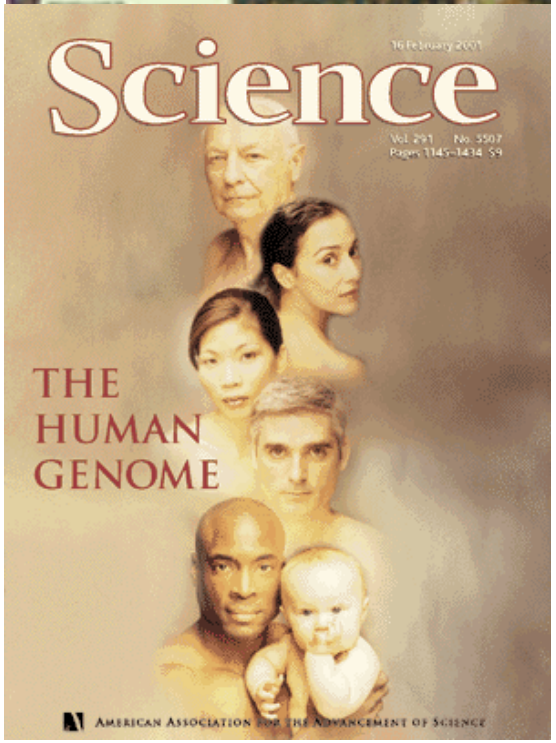
Associated Variant

Controls: (Healthy individuals)

AGAGCAGTCGACATGTATAGTCTACATGAGATCGACATGAGATCGGTAGAGCAGTGAGATCAACATGATAGCC
AGAGCAGTCGACATGTATAGTCTACATGAGATCAACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCAGTCGACATGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCAACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGTC
AGAGCCGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCAACATGATAGCC
AGAGCAGTCGACAGGTATAGTCTACATGAGATCGACATGAGATCTGTAGAGCAGTGAGATCGACATGATAGCC
AGAGCCGTCGACAGGTATAGCCTACATGAGATCGACATGAGATCTGTAGAGCCGTGAGATCGACATGATAGCC
AGAGCCGTCGACAGGTATAGTCTACATGAGATCAACATGAGATCTGTAGAGCAGTGAGATCGACATGATAGTC

Key Ingredient I: The Human Genome

- Human Genome Project
  - **Published in 2001**
  - **"Big Science"**
  - **Two competing projects: NIH and Celera**
  - **Celera sequenced J. Craig Venter**
  - **Worldwide participation**
  - **Sequenced "reference" human genome**
  - **Goal to obtain sequence for consensus human: what we have in common.**
- The Genome Project…
  - **Identified all genes**
  - **What do these genes do?**
  - **How do they influence disease?**



Key Ingredient: Maps of Variation

- The Human Haplotype Map…
  - **Published in 2005**
  - **Worldwide survey of human variation**
  - **270 Individuals**
  - **4 Populations**
  - **4 million genetic polymorphisms**
- informs Association Studies Design
  - **What variation to collect?**
  - **How many people to collect?**
  - **How to analyze the data?**
- Studies with Statistical Power…
  - **Collect 4000+ individuals**
  - **Collect 500,000+ SNPs**

International HapMap Project

- Successor to the Human Genome Project
- International consortium that aims in genotyping the genome of 270 individuals from four different populations.
- Launched in 2002. First phase was finished in October (Nature, 2005).
- Collected genotypes for 3.9 million SNPs.
- Location and correlation structure of many common SNPs.

Public Genotype Data Growth

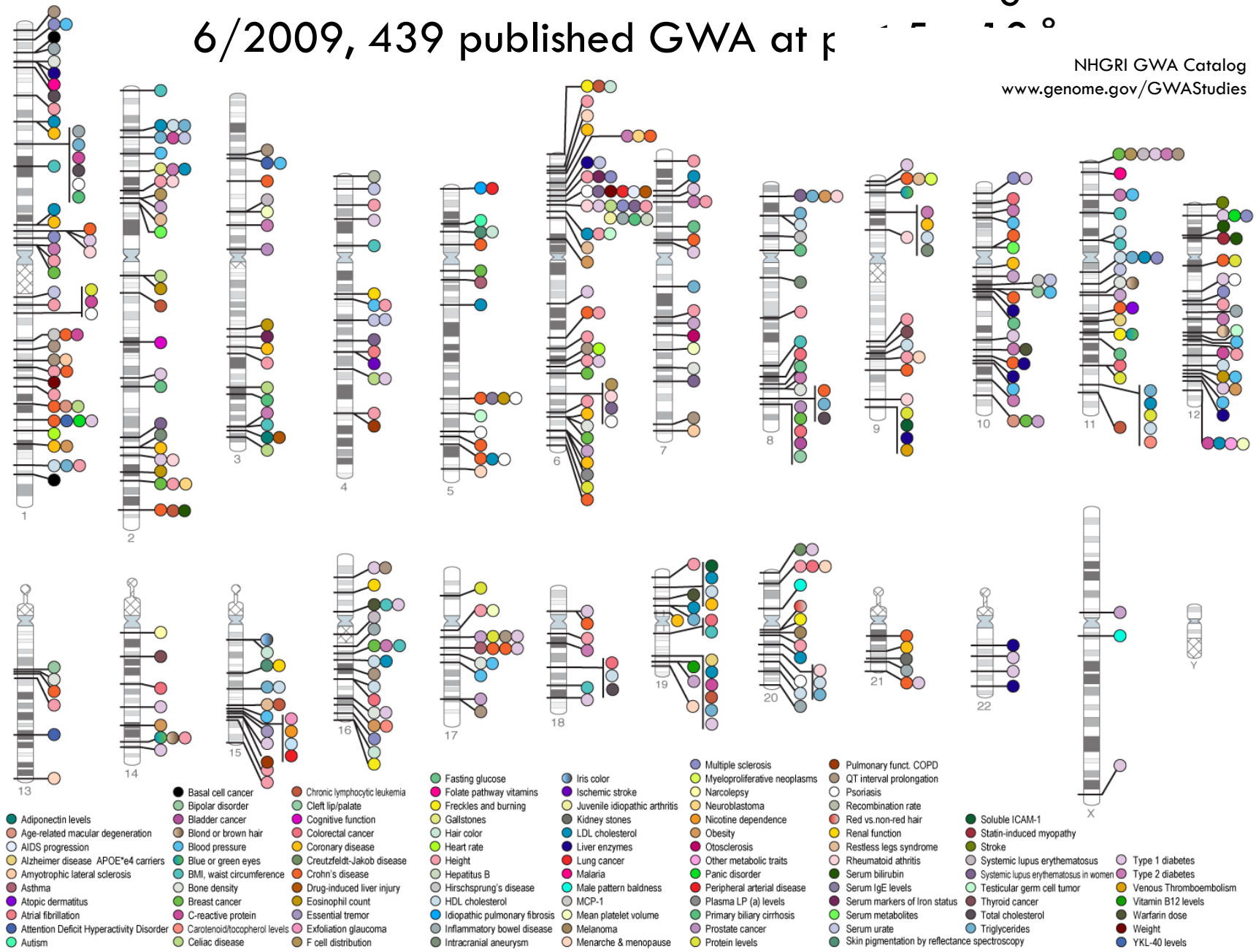
2001	2002	2003	2004	2005	2006
Daly et al. *Nature Genetics* 103 SNPs 40,000 genotypes	Gabriel et al. *Science* 3000 SNPs 400,000 genotypes	TSC Data *Nucleic Acids Research* 35,000 SNPs 4,500,000 genotypes	Perlegen Data *Science* 1,570,000 SNPs 100,000,000 genotypes	NCBI dbSNP *Genome Research* 3,000,000 SNPs 286,000,000 genotypes	HapMap Phase 2 5,000,000+ SNPs 600,000,000+ genotypes

Key Ingredient: High-Throughput Genotyping Technology

- Collects SNP information from DNA
  - **2 Major Companies: Affymetrix and Illuminia**
  - **Based on hybridization technology**
- Significant Cost Savings
  - **Reduces cost of collecting genotype information from 14 cents per genotype to .02 cents per genotype.**
  - **The HapMap originally cost over 100 million dollars.**
  - **Today the HapMap would cost $20,000.**
  - **Associations studies now cost in the low millions.**

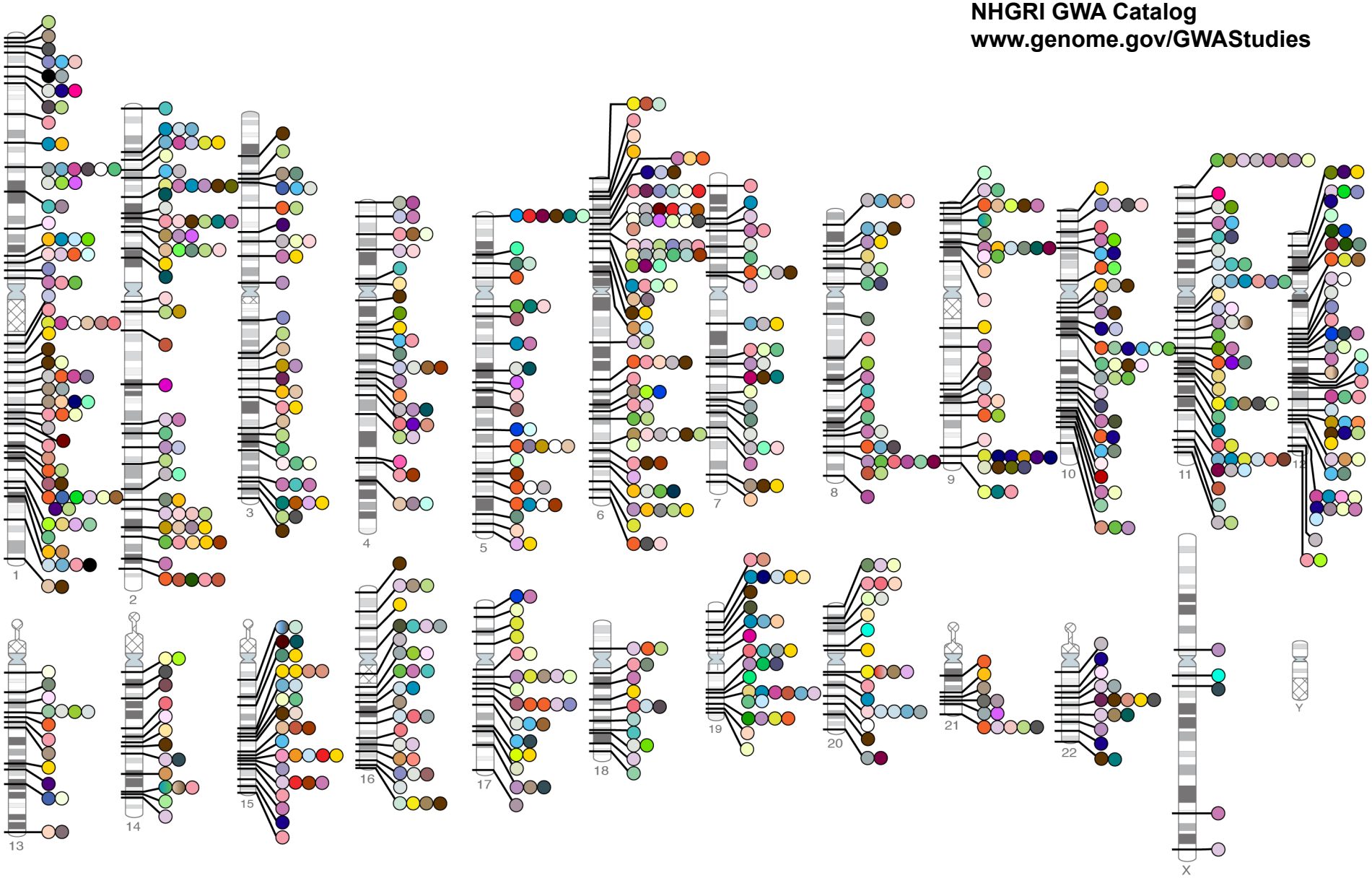
Published Genome-Wide Associations through 6/2009, 439 published GWA at p

NHGRI GWA Catalog
www.genome.gov/GWAStudies



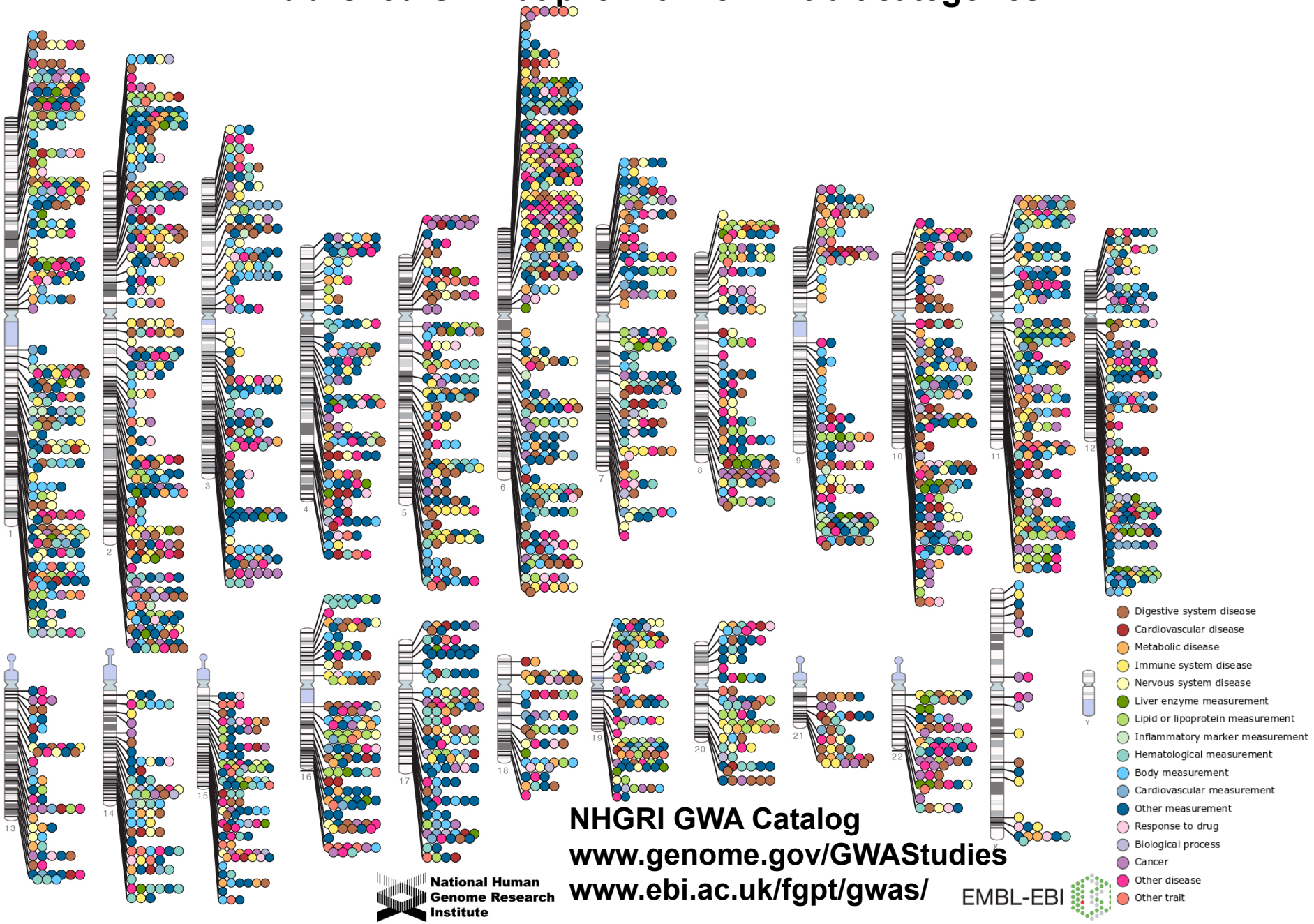
Published Genome-Wide Associations through 6/2010, 904 published GWA at $p \leq 5 \times 10^{-8}$ for 165 traits

NHGRI GWA Catalog
www.genome.gov/GWAStudies



Published Genome-Wide Associations through 12/2012

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

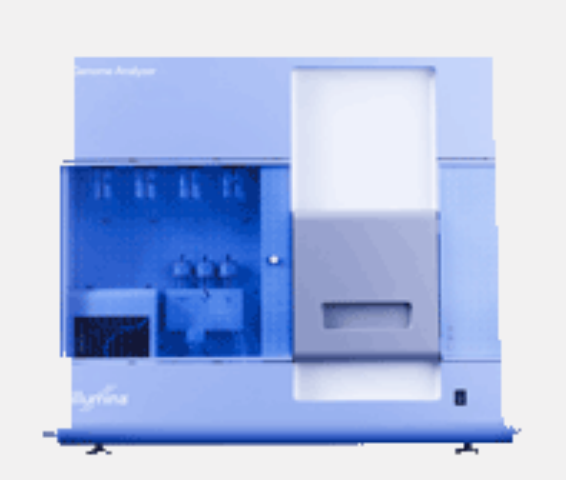


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

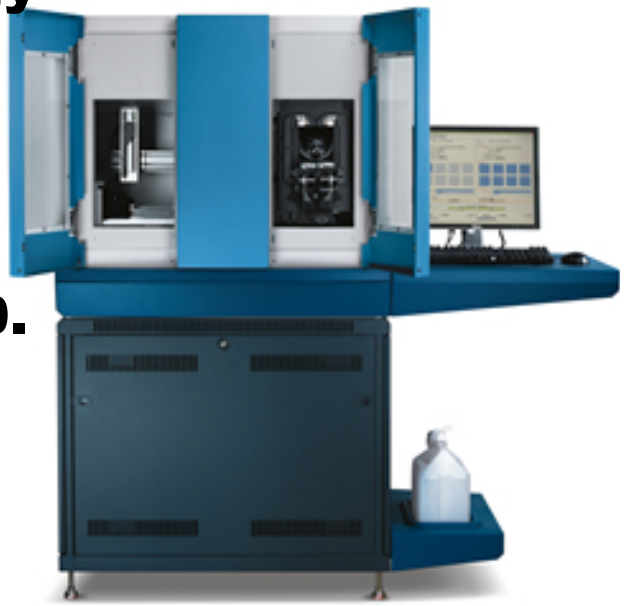
National Human Genome Research Institute

EMBL-EBI

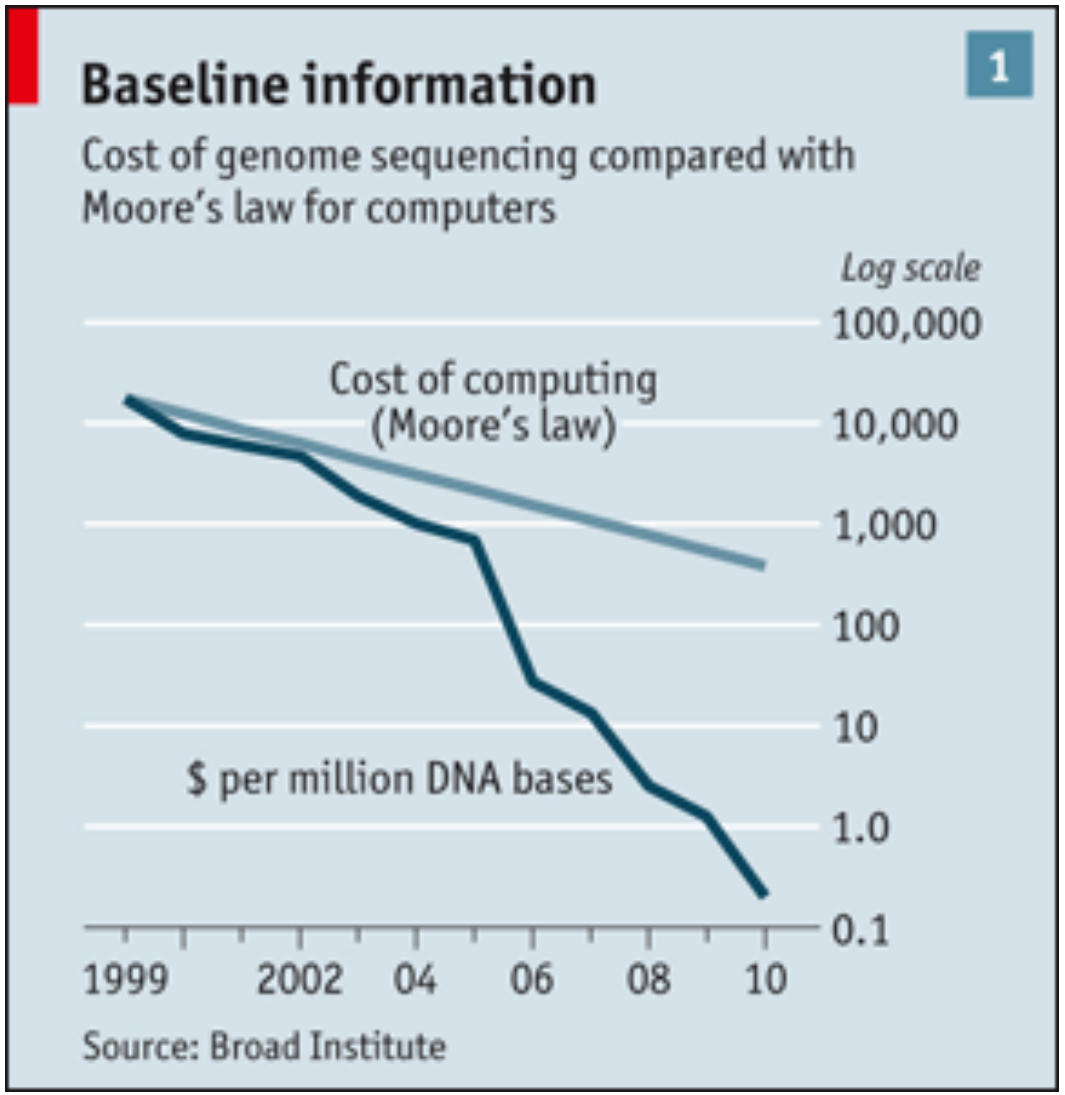
Key Ingredient: High-Throughput Sequencing Technology



- Collects sequence information from DNA
  - **Many companies developing technology**
  - **Allows discovery of rare variation**
- Significant Cost Savings
  - **Reduces cost of collecting individuals genome from $1,000,000,000 to $5,000.**
  - **Within 2 years, cost of genome will be under $1,000.**
  - **Many projects leveraging this technology**
    - **Thousand Genomes Project**



Sequencing Costs



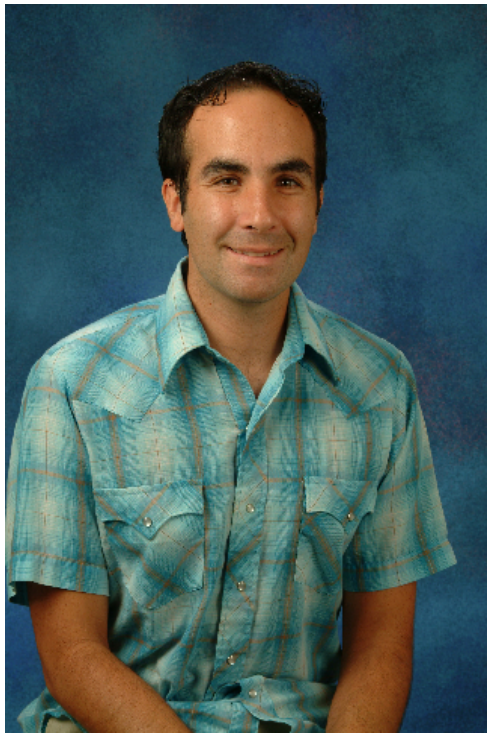
Source: The Economist
July 17th, 2010

Many Computational Problems

- Genotype-Phenotype Problems
  - **Design and Analysis of Association Studies**
  - **Combining Association Studies**
  - **Integrating Prior Information**
  - **Population Structure**
- New Technology Problems
  - **Sequence Assembly**
  - **Read Mapping**
  - **Identifying Structural Variation**
- Population Genetics Problems
  - **Inference of Human Genetic History**
  - **Admixed Populations**

Break!

How do we get someone's DNA sequence? Where are my mutations?



Sequencing Technology

Illumina / Solexa
Genetic Analyzer 1G
1000 Mb/run, 35bp reads

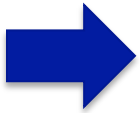
AGAGCAGTCGAC
AGGTATAGTCTA
CATGAGATCGAC
ATGAGATCGGTA
GAGCCGTGAGAT
CGACATGATAGC
CAGAGCAGTCGA
CAGGTATAGTCT
ACATGAGATCGA
CATGAGATCGGT
AGAGCCGTGAGA
TCGACATGATAG
CCAGAGCAGTCG
ACAGGTATAGTC
TACATGAGATCG
ACATGAGATCGG
TAGAGCCGTGAG
ATCGACATGATA
GCCAGAGCAGTC
GACAGGTATAGT
CTACATGAGATC
GACATGAGATCG
GTAGAGCCGTGA
GATCGACATGAT

- Next generation sequencing.
  - **Cheap sequencing.**
  - **"Short Reads"**

Short Read Sequencing Problem (A Computer Science Problem)

Full DNA Sequence

AGAGCAGTCGAC
AGGTATAGTCTA
CATGAGATCGAC
ATGAGATCGGTA
GAGCCGTGAGAT
CGACATGATAGC
CAGAGCAGTCGA
CAGGTATAGTCT
ACATGAGATCGA
CATGAGATCGGT
AGAGCCGTGAGA
TCGACATGATAG
CCAGAGCAGTCG
ACAGGTATAGTC
TACATGAGATCG
ACATGAGATCGG
TAGAGCCGTGAG
ATCGACATGATA
GCCAGAGCAGTC
GACAGGTATAGT
CTACATGAGATC



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

ATGAGATCGGTAGAGCCGTGAGAT
GAGCAGTCGACAGGTATAGTCTAC
AGAGCAGTCGACAGGTATAGTCTA
TGAGATCGACATGATAGCCAGAGC
TAGCCAGAGCAGTCGACAGGTATA
GATAGCCAGAGCAGTCGACAGGTA
GAGATCGACATGATAGCCAGAGCA
GCAGTCGACAGGTATAGTCTACAT
AGCAGTCGACAGGTATAGTCTACA
TCGACATGAGATCGGTAGAGCCGT
CAGTCGACAGGTATAGTCTACATG
GAGATCGACATGATAGCCAGAGCA
GTAGAGCCGTGAGATCGACATGAT

Short Reads Difficulties

```
ATGAGATCGGTAGAGCCGTGAGAT
GAGCAGTCGACAGGTATAGTCTAC
AGAGCAGTCGACAGGTATAGTCTA
TGAGATCGACATGATAGCCAGAGC
TAGCCAGAGCAGTCGACAGGTATA
GATAGCCAGAGCAGTCGACAGGTA
GAGATCGACATGATAGCCAGAGCA
GCAGTCGACAGGTATAGTCTACAT
AGCAGTCGACAGGTATAGTCTACA
TCGACATGAGATCGGTAGAGCCGT
CAGTCGACAGGTATAGTCTACATG
GAGATCGACATGATAGCCAGAGCA
GTAGAGCCGTGAGATCGACATGAT
```

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

Key Idea: "Re"-Sequencing

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

My Genome:

TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGAGCCGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAGCCGT

Recovered Sequence:

TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC

"Re"-Sequencing Output

Resequencing provides a list of changes to make from the reference to change it to the target. Similar to unix "diff".

My Genome:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

The Human Genome:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

A list of changes

Recovered Sequence:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

"Re"-Sequencing Problems

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

My Genome:

TACATGAGATC**G**ACATGAGATCGGTACATGAGATCCACAT

A Sequence Read:

ACATGAGATC**G**ACAT

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT
ACATGAGATC**G**ACAT ACATGAGATC**G**ACAT

Error!

Recovered Sequence:

TACATGAGATC**G**ACATGAGATCGGTACATGAGATC**G**ACAT

"Re"-Sequencing Problems

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

My Genome:

TACATGAG**GGGGGGGG**GAGATCGGTACATGAGATCCACAT

A Sequence Read:

GAG**GGGGGGGG**

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

GAG**GGGGGGG**G

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 seqeunce.

Key Question: When does resequencing work?

- We must be able to map a substring from the target to its corresponding place in the reference.
- Why can this not happen?
  - **Reference has repeated sequences. In this case reads from target will map to multiple places.**
  - **Target sequence differs that resemblance to reference sequence is lost.**

Formalizing the Problem

- Target sequence – Sequence of the genome that we are analyzing and collecting reads from.
- Reference sequence – Sequence of the similar genome which we have available.
- Constraints on the reference sequence
  - **Non repetitive sequences (or non-repetitive portion)**
- Constrains on difference between the target and reference.
  - **Assume that there are a small number of structured differences.**

Simple Resequencing Formulation

- Assume that the reference sequence is of length N.
- Assume target sequence is of length N.
- Constraint on Mutations - Assume that target sequence differs from reference by less than D mutations in any window of L.
- Unique Sequence Assumption – Assume that any 2 positions in the reference sequence differ by more than D+1 mismatches.

Algorithmic “Re”-Sequencing Challenges

- Sequences are long!
  - **Human Genome is 3,000,000,000 long.**
- Sequencers generate many reads!
  - **A single run generates over 300,000,000 reads.**
- We need efficient algorithms to “map” each read to its location in the genome.

There are other challenges which we are not mentioning.

Trivial Mapping Algorithm

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT

- We can slide our read along the genome and count the total mismatches between the read and the genome.
- If the mismatches are below a threshold, we say that it is a match.

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAGCCGT

Total of 18 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGAGCCGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAGCCGT

Total of 15 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
  TCGACATGAGATCGGTAGAGCCGT

Total of 23 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAGCCGT

Total of 23 mismatches. Not below threshold. Not a match.

Trivial Mapping Algorithm

The Human Genome:

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
```

A Sequence Read:

```
TCGACATGAGATCGGTAGAGCCGT
```

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
          TCGACATGAGATCGGTAGAGCCGT
```

Total of 3 mismatches. Below threshold. A match!

Complexity of Trivial Algorithm

- 3,000,000,000 length genome (N)
- 300,000,000 reads to map (M)
- Reads are of length 30 (L)
- Number of mismatches allowed is 2 (D).
- Each comparison of match vs. mismatch takes 1/1,000,000 seconds (t).

Total Time = N*M*L*t = 27,000,000,000,000 seconds or 864,164 years!

- Important: Trivial algorithm only solves problem under assumptions.

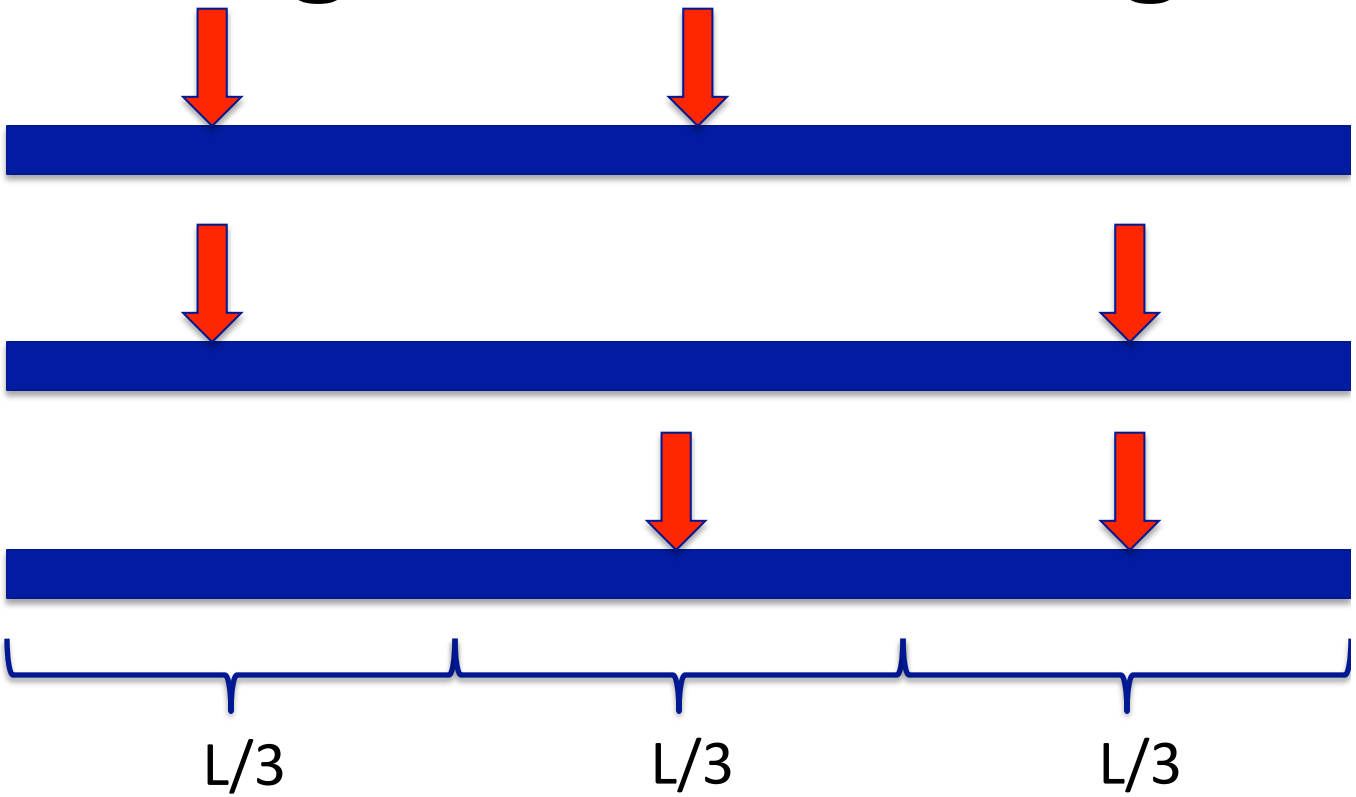
Some observations

- Most positions in the genome match very poorly.
- We are looking for only a few mismatches.

  (D is small)
- A substring of our read will match perfectly.

Perfect Matching Read Substrings

Three "worst" possible cases for placement of mutations.



- In each case, there is a perfect match of L/3.

Finding a perfect match of length L/3

- Intuition: Create an index (or phone book) for the genome.
- We can look up an entry quickly.

Sequence	Positions
AAAAAAAAAA	32453, 64543, 76335
AAAAAAAAAC	64534, 84323, 96536
AAAAAAAAAG	12352, 32534, 56346
AAAAAAAAAT	23245, 54333, 75464
AAAAAAAACA	
AAAAAAAACC	43523, 67543
…	
CAAAAAAAAA	32345, 65442
CAAAAAAAAC	34653, 67323, 76354
…	
TCGACATGAG	54234, 67344, 75423
TCGACATGAT	11213, 22323
…	
TTTTTTTTTG	64252
TTTTTTTTTT	64246, 77355, 78453

If L=30, each entry will have a key of length 10. Each entry will contain on average $N/4^{10}$ positions. (Approximately 3,000).

If L=45, each entry will have a key of length 15. Each entry will contain on average 3 positions.

Complexity of Indexing Algorithm

- We need to look up each third of the read in the index.
- For L=30, our index will contain entries of length 10. Each entry will contain on average $N/(4^{L/3})$ or 3,000 positions.
- For each position, we need to compute the number of mismatches.
- Our running time is $L * M*3*N/(4^{L/3})*T$=81,000,000 seconds or 937 days.
- If L=45, then the time is 81,000 seconds or 22.5 hours.

More problems: Sequencing Errors

- Each sequence read can have some random errors.

My Genome:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGA**A**CCGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAACCGT

Recovered Sequence:

TACATGAGATC**G**ACATGAGATC**G**GTAGA**A**C**C**GTGAGATC

Sequencing Errors: Solution

- Collect redundant data.

My Genome:

TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

Sequence Reads:

TCGACATGAGATCGGTAGA**A**CCGT
GACA**A**GAGATCGGTAGAGCCGTGA
TGAGATCGG**T**AGAGCCGTGAGATC

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TC**G**ACATGAGATC**G**GTAGA**A**C**C**GT
GACA**A**GAGATC**G**GTAGAGC**C**GTGA
TGAGATC**G**G**T**AGAGC**C**GTGAGATC

Recovered Sequence:

TACATGAGATC**G**ACATGAGATC**G**GTAGAAC**C**GTGAGATC

How much coverage do we need?

- If error rate is *e*, and we are going to predict the consensus sequence, what is the error rate if the coverage is 3.
- We will make a prediction with an error if two out of our three reads have an error in the same place.

$$e^3 + \binom{3}{2}(1-e)e^2$$

- This is approximately $3e^2$.

Diploid Sequencing

- Humans have 2 chromosomes.
- Each chromosome may have a different SNP.
- Some reads come from 1 chromosome, some come from other chromsome.
- Why does consensus method not work?
- How do we address this problem?

"Re"-Sequencing: Insertions

My Genome:

TACATGAGATCCACAT**A**GAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

How do we deal with this case?

"Re"-Sequencing: Insertions

My Genome:

TACATGAGATCCACATAGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

CCACATAGAGATCTGTAGAGCTGT

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Solution: Add Insertion to the Human Genome

TACATGAGATCCACAT-GAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Difficulties for handling insertions

- Requires “Alignment” of reads to genome.
- Much more computational intensive

- Need to change assumptions for “sequence uniqueness” to use edit distance.

Many other challenges

- Repeated regions in the genome.
  - **When we align a read, we get two positions that it matches!**
- Coverage of sequence reads is not uniform
  - **Some places we have many reads, while some we have fewer. How do we design an approach so we can always recover the sequence.**
- Large memory requirements
  - **We need to fit our index into RAM. Often tens of Gigabytes or greater.**