Computational Biology

Sriram Sankararaman
Computer Science Department

Biology is a data science

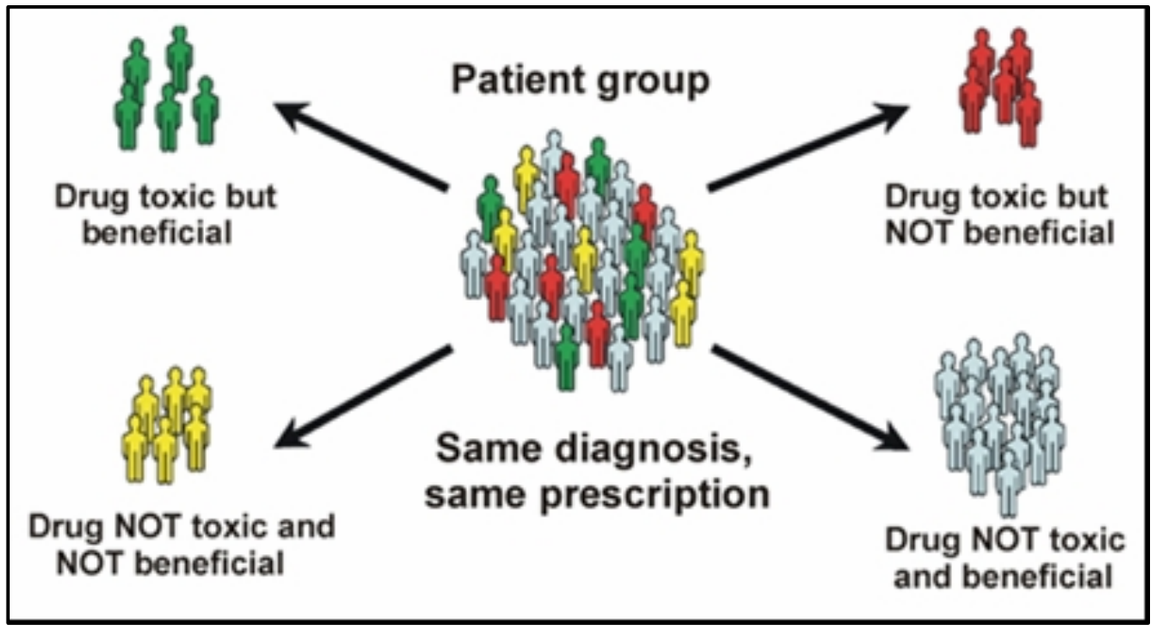
- Genomics
  - Fundamental questions about disease and evolution
  - Massive data from new technologies
  - Bottleneck is computation

Computation key to making sense of genomic data

Bioinformatics

Answering biological questions using computer science, statistics and mathematics.

Personalized medicine

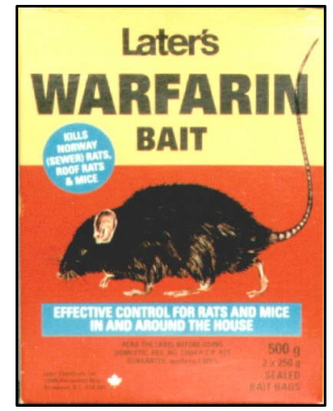


Example: Warfarin

An anticoagulant drug useful in the prevention of thrombosis

Example: Warfarin

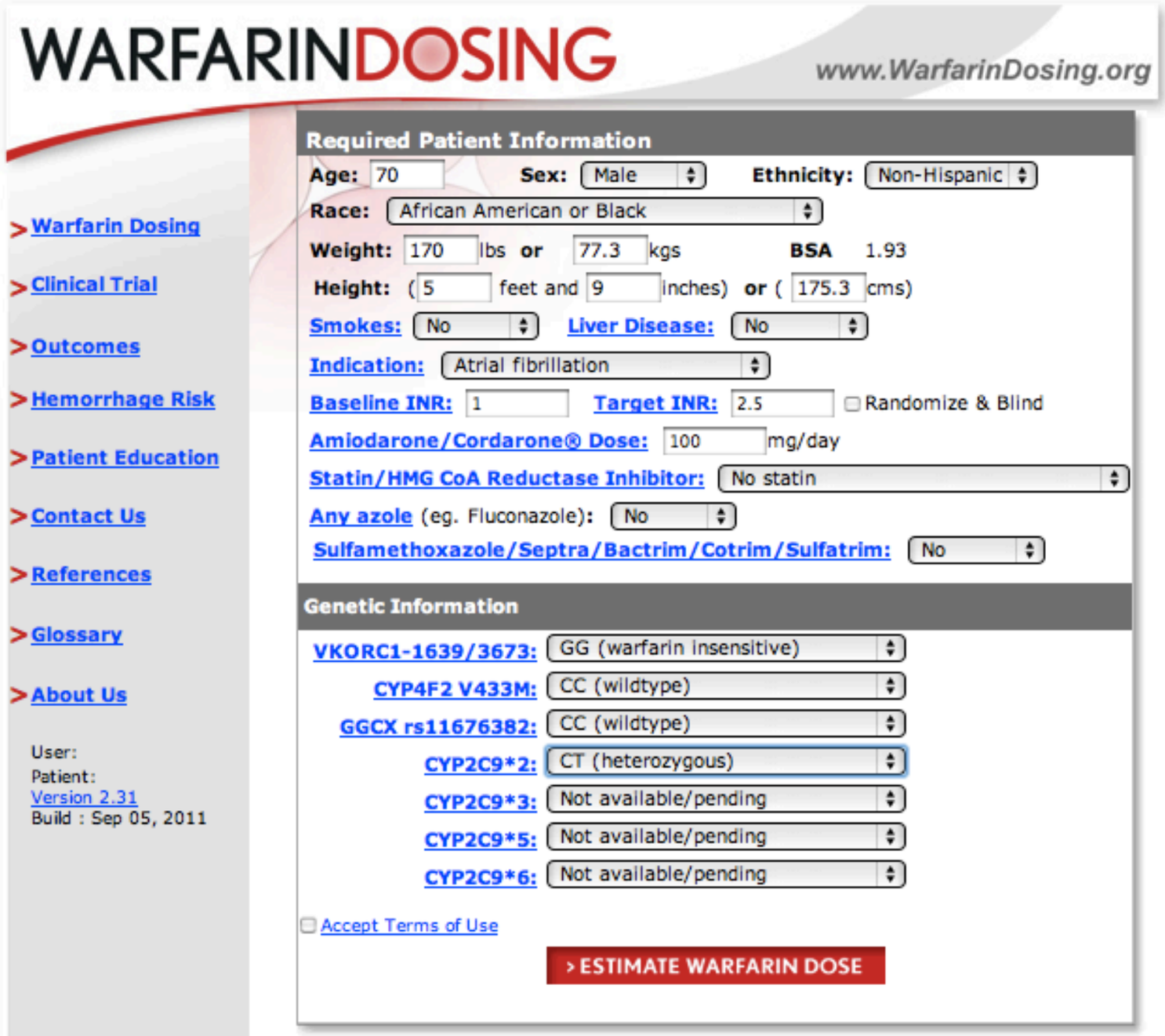
Originally used as rat poison



Optimal dose for treatment varies

Genetic variants (VKORC1 and CYP2C9) affect the optimal dose

Example: Warfarin



Personal genomics

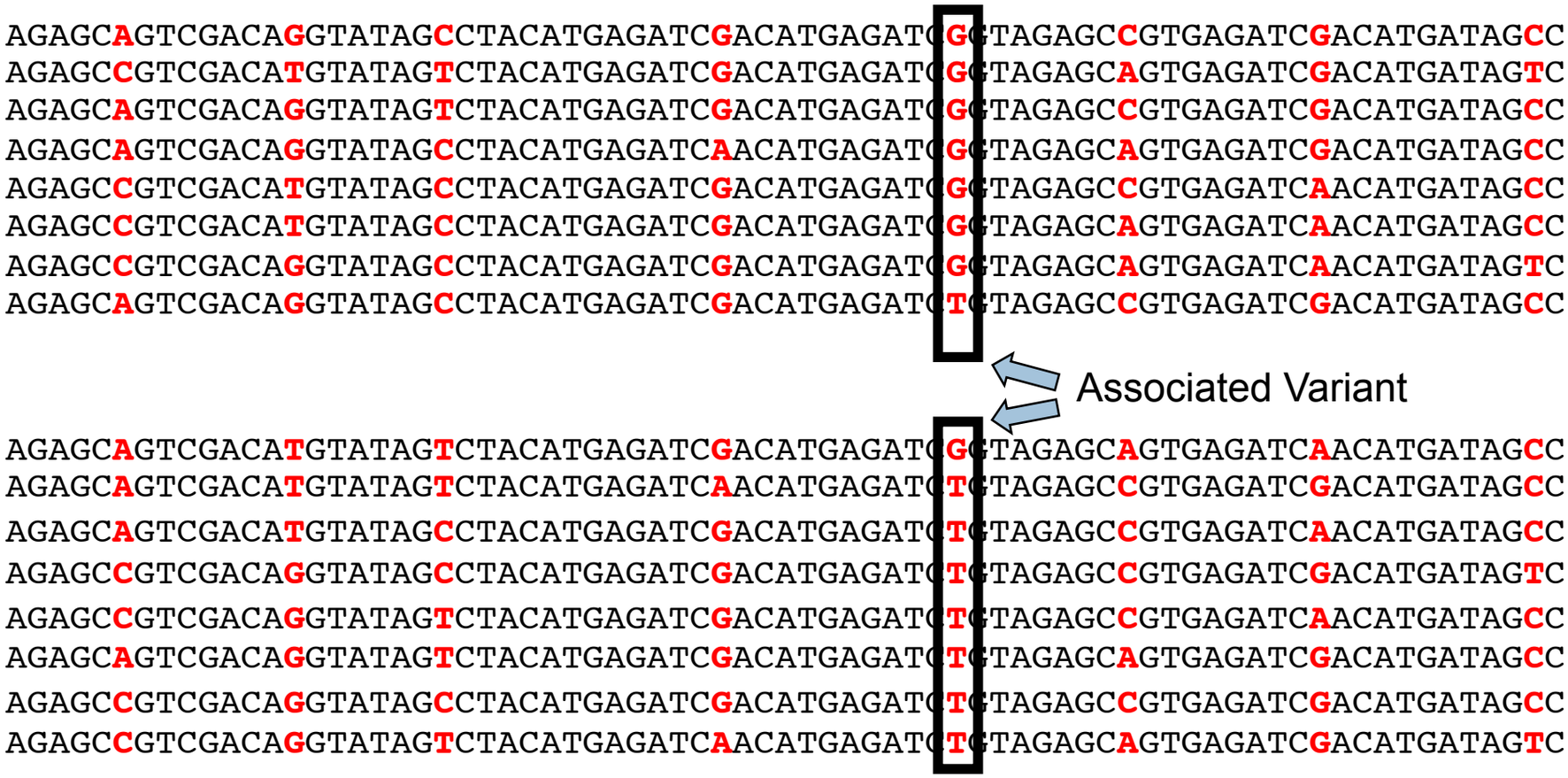
We are the first and only genetic service available directly to you that includes reports that meet FDA standards.

The human genome project

Finding disease genes

Association studies (GWAS)

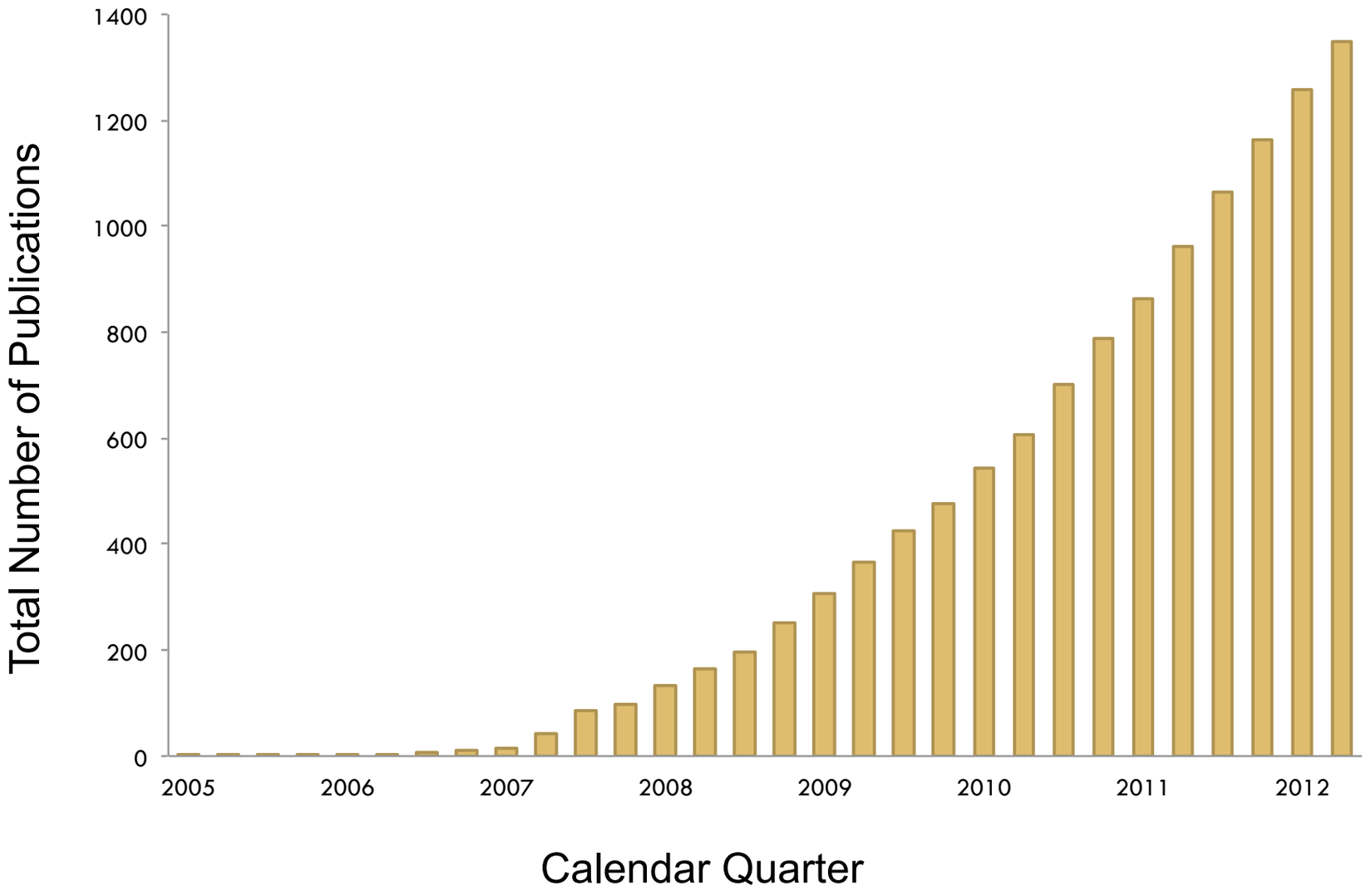
Disease



Normal

Finding disease genes

Association studies (GWAS)

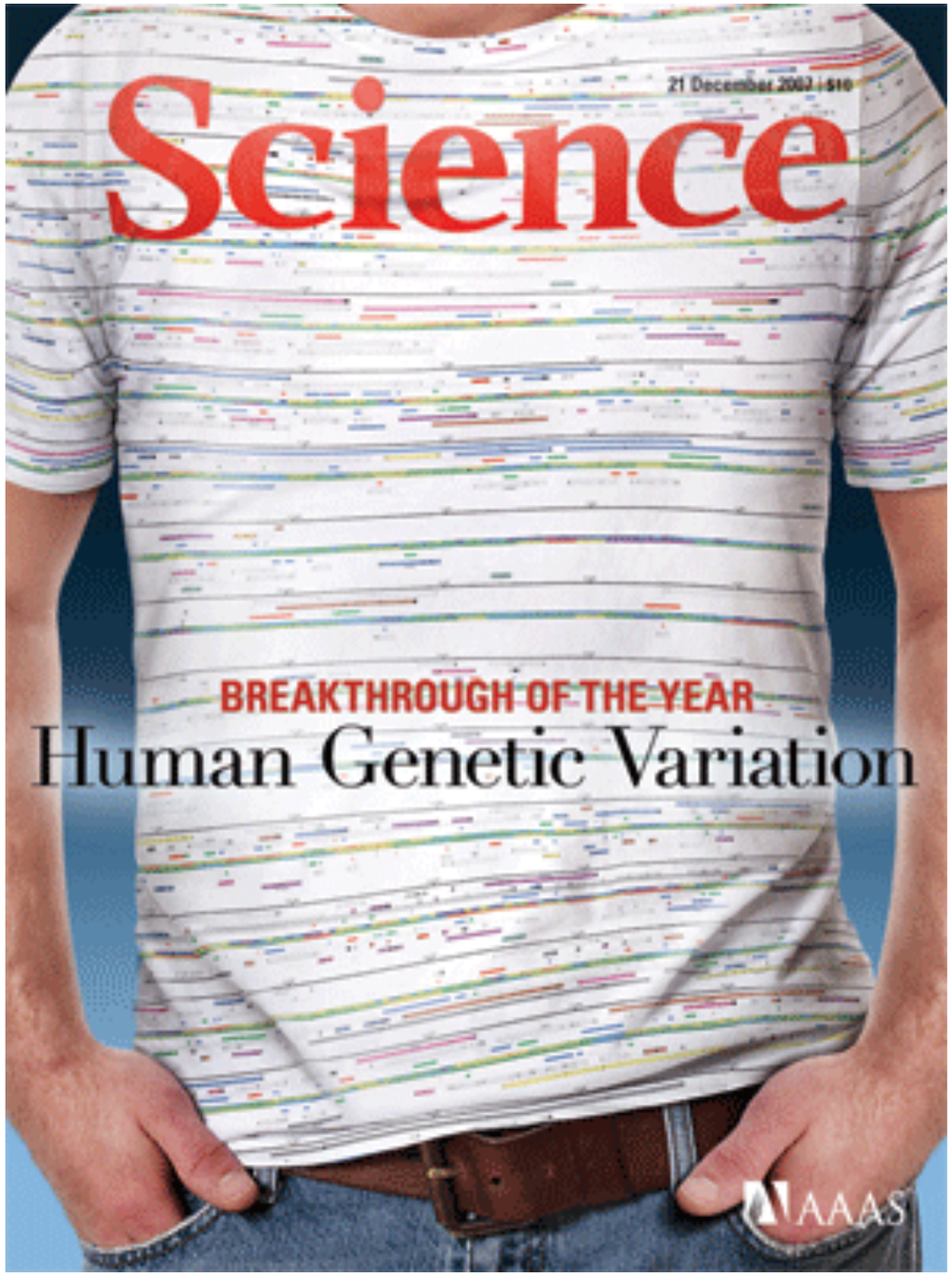


GWAS

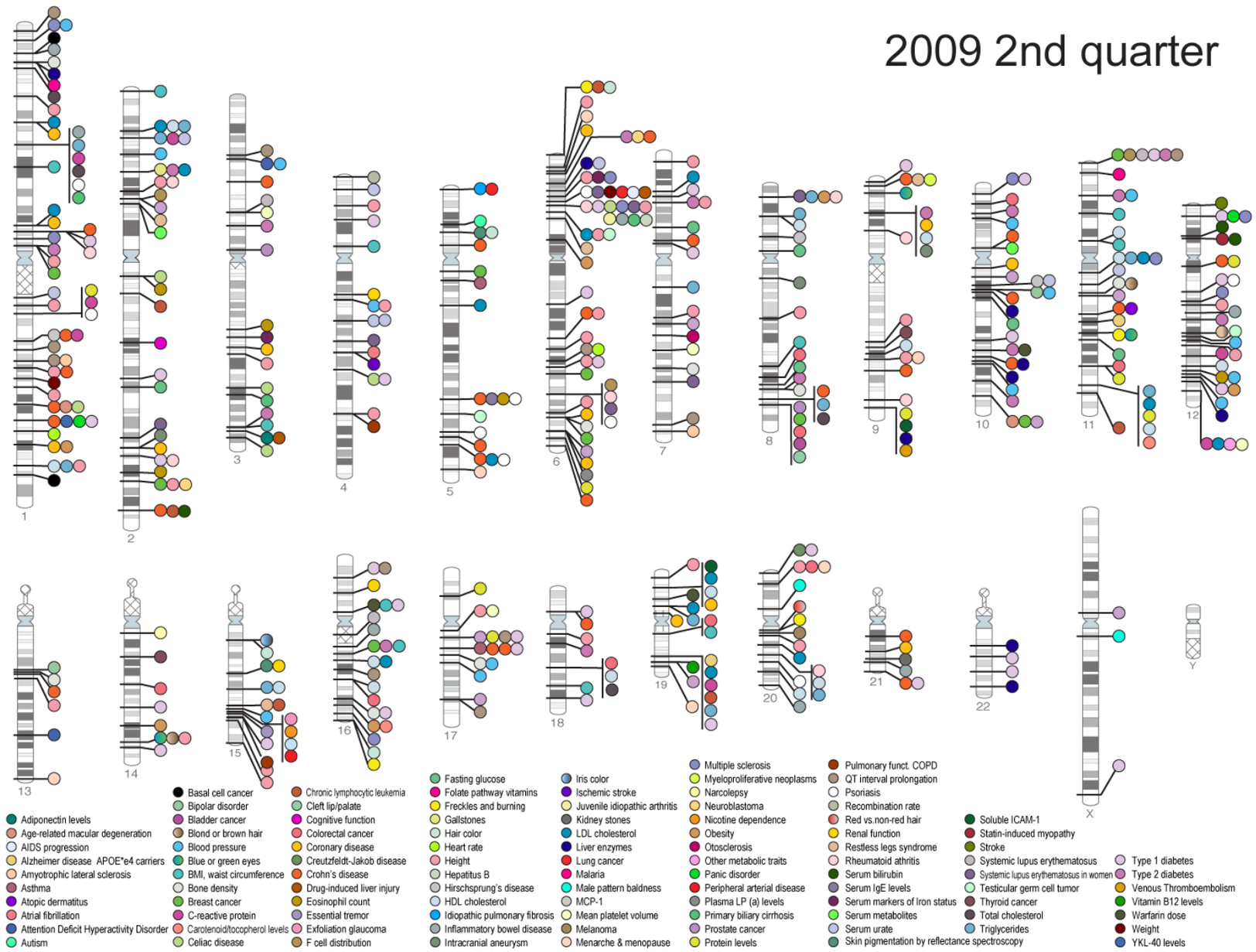
2007 breakthrough of the year

More than 50 genes discovered to affect dozens of common diseases

Reports of “genes causing ….”

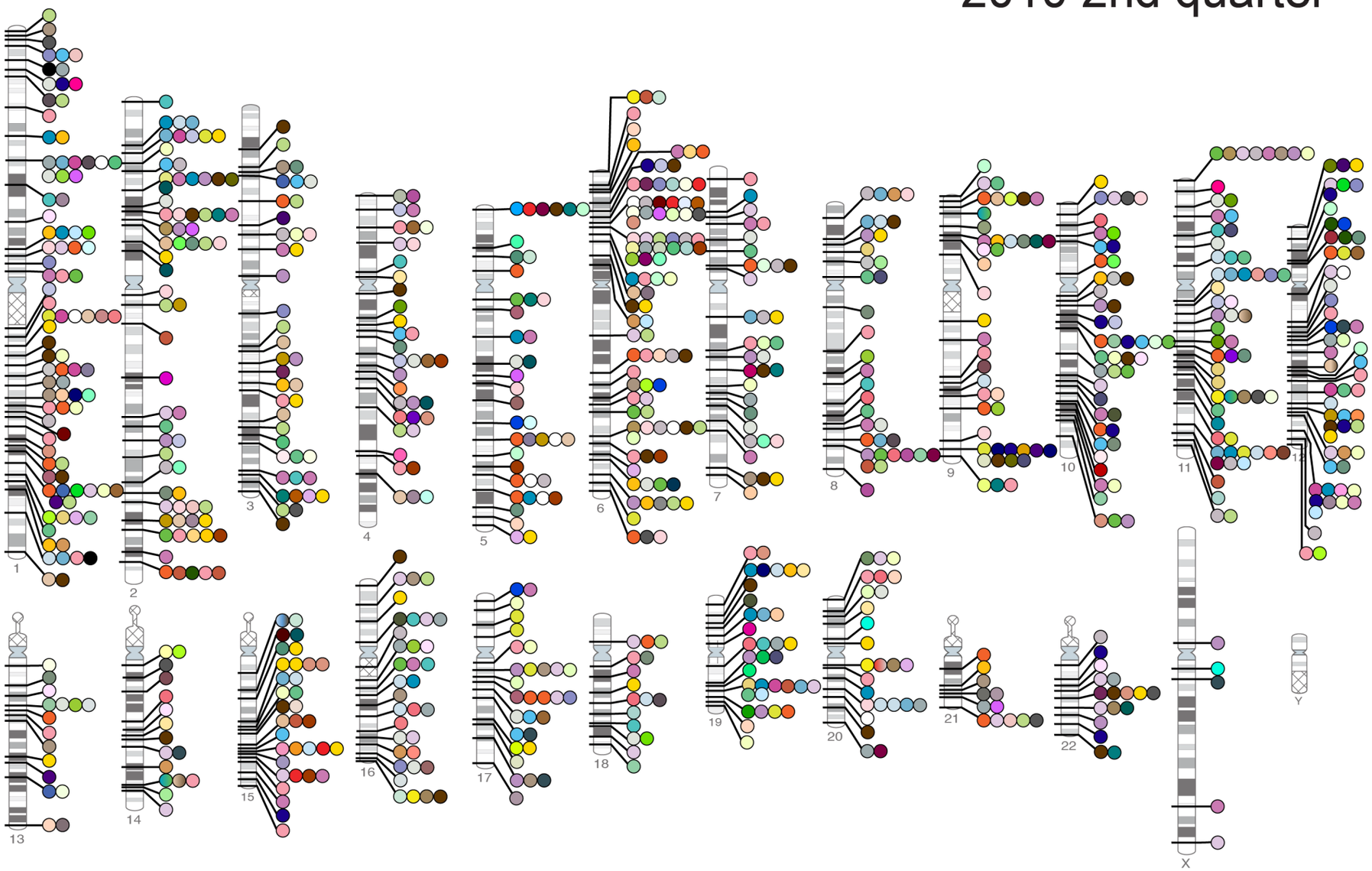


Published GWAS



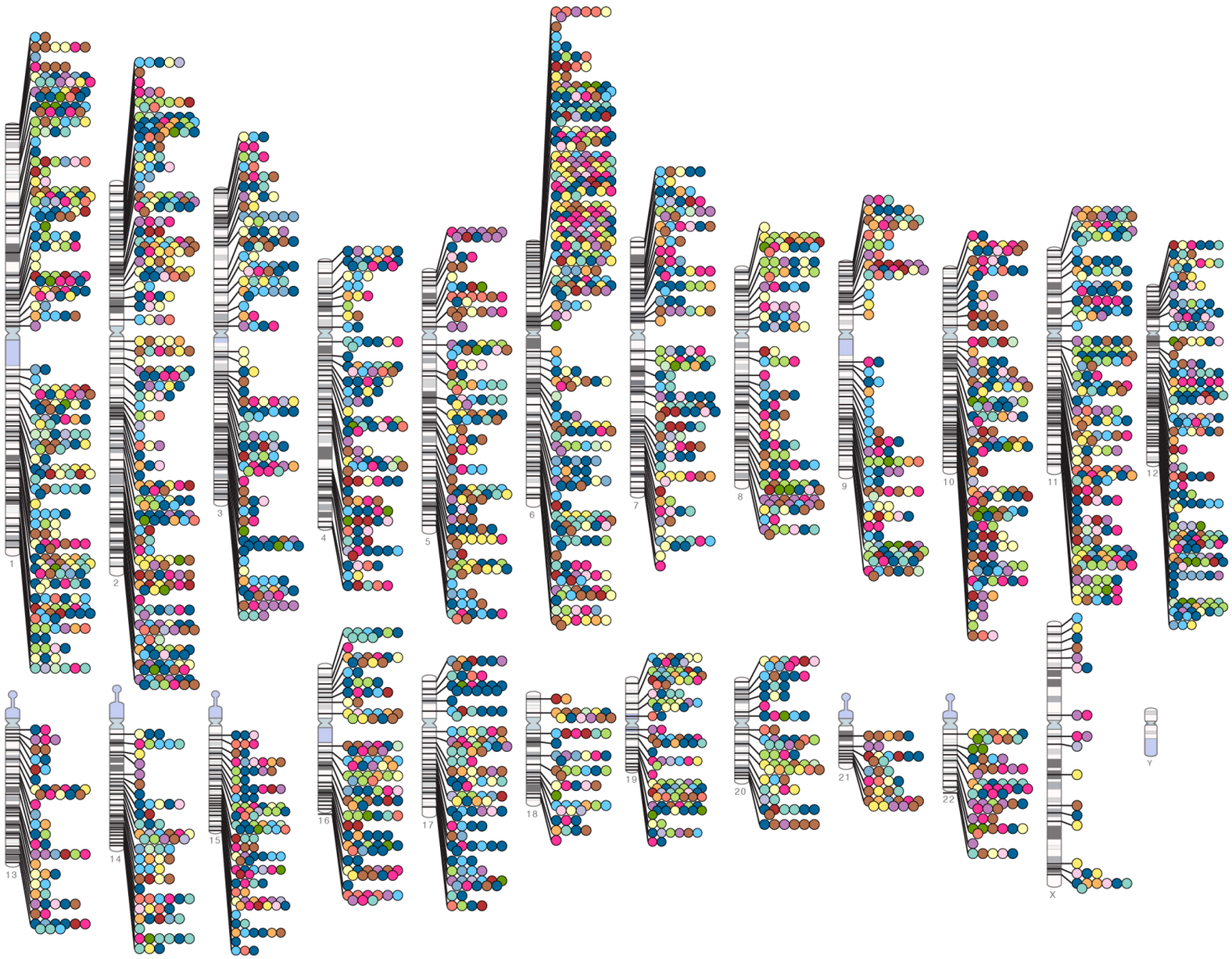
Published GWAS

2010 2nd quarter



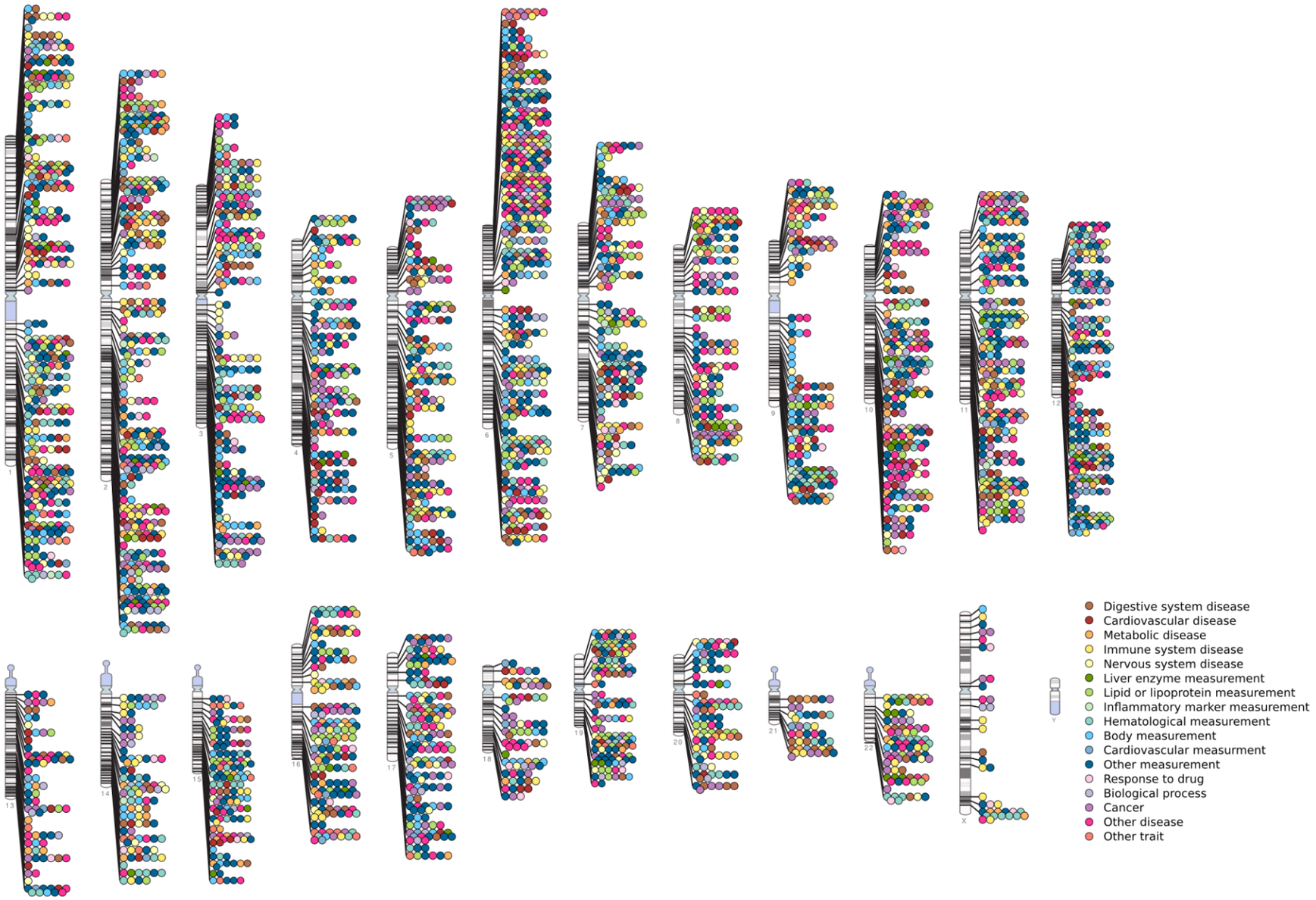
Published GWAS

2012



Published GWAS

2013



What next ?

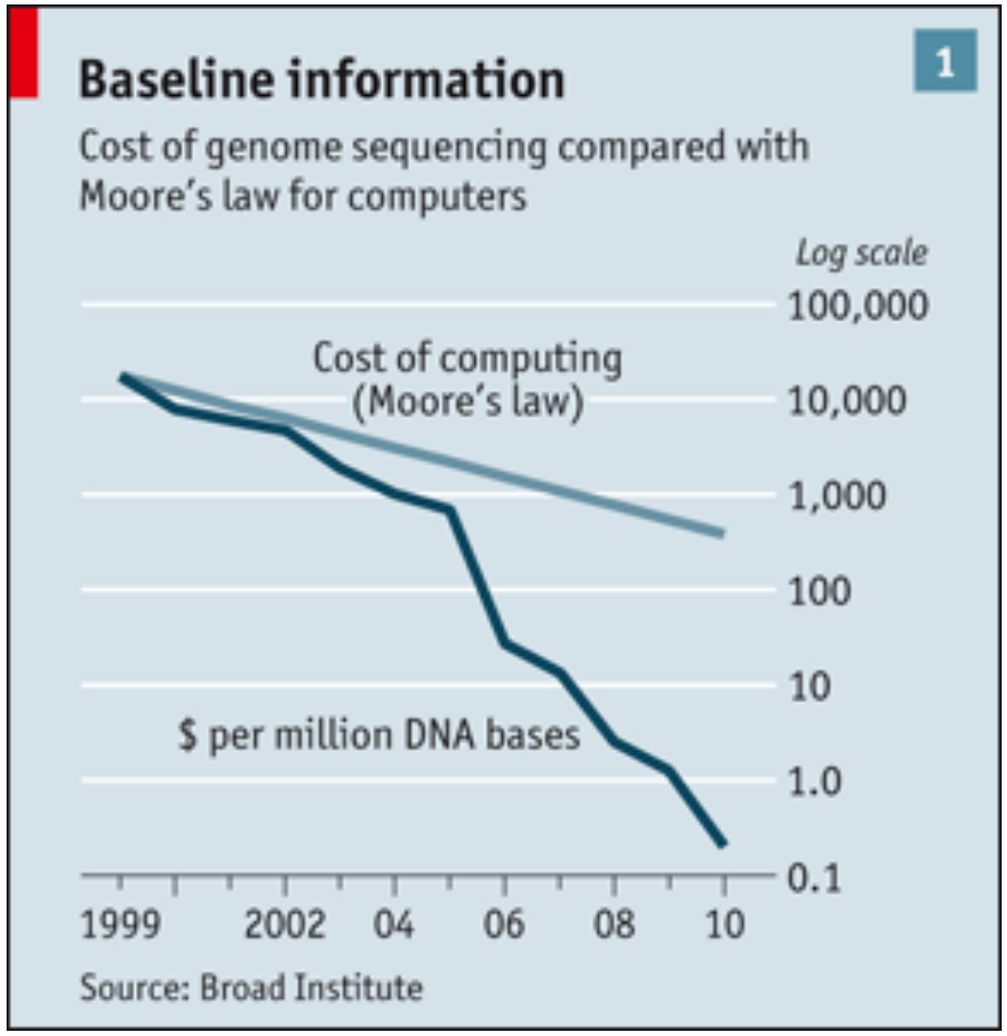
Most genes associated with disease have weak effects

Current GWAS results can predict a small fraction of disease risk

Many areas for computer scientists to contribute

The case of the missing heritability

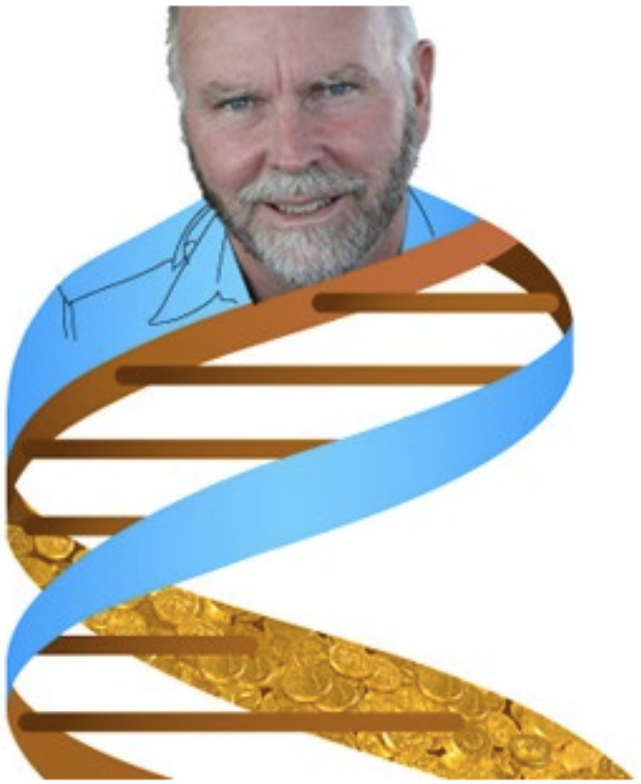
What is driving these studies?



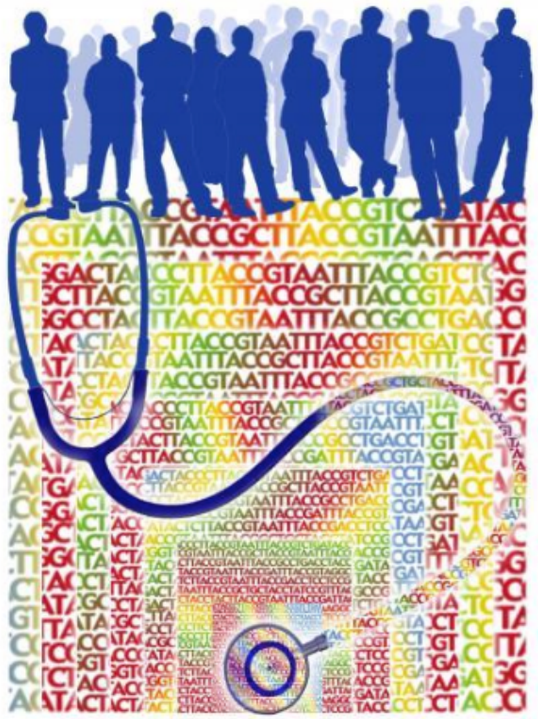
The Economist, 2010

The genomic revolution

2001
Human genome project

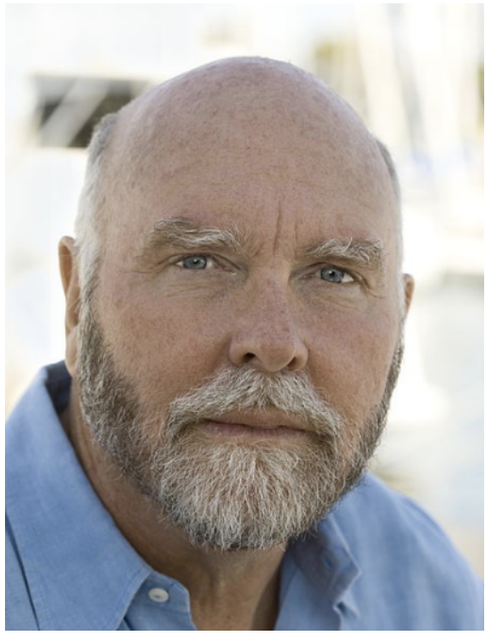


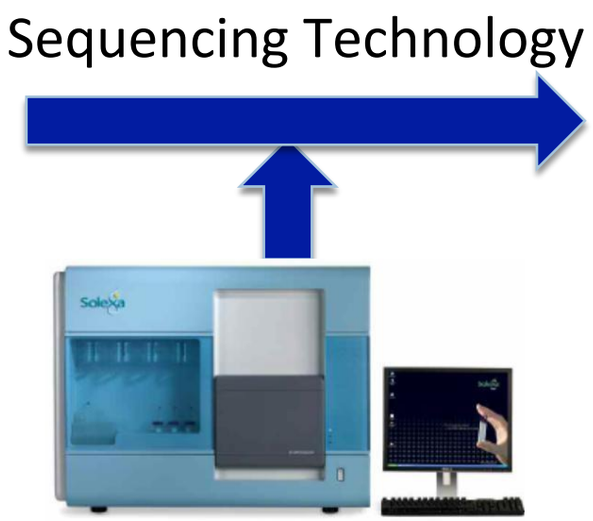
2010
1000 genomes project



Active research problem: Short read resequencing

Where are my mutations ?





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

AGAGC**A**GTCGACA**G**GTA
TAG**T**CTACATGAGATC**G**
ACATGAGATC**G**GTAGAG
C**C**GTGAGATC**G**ACATGA
TAG**C**CAGAGC**A**GTCGAC
A**G**GTATAG**T**CTACATGA
GATC**G**ACATGAGATC**G**G
TAGAGC**C**GTGAGATC**G**A
CATGATAG**C**CAGAGC**A**G
TCGACA**G**GTATAG**T**CTA
CATGAGATC**G**ACATGAG
ATC**G**GTAGAGC**C**GTGAG
ATC**G**ACATGATAG**C**CAG
AGC**A**GTCGACA**G**GTATA
G**T**CTACATGAGATC**G**AC
ATGAGATC**G**GTAGAGC**C**
GTGAGATC**G**ACATGATA
G**C**CAGAGC**A**GTCGACA**G**
GTATAG**T**CTACATGAGA
TC**G**ACATGAGATC**G**GTA
GAGC**C**GTGAGATC**G**ACA
TGATAG**C**CAGAGC**A**GTC
GACA**G**GTATAG**T**CTACA
TGAGATC**G**ACATGAGAT
C**G**GTAGAGC**C**GTGAGAT
C**G**ACATGATAG**C**CAGAG
C**A**GTCGACA**G**GTATAG**T**
CTACATGAGATC**G**ACAT
GAGATC**G**GTAGAGC**C**GT
GAGATC**G**ACATGATAG**C**

Next-generation sequencing

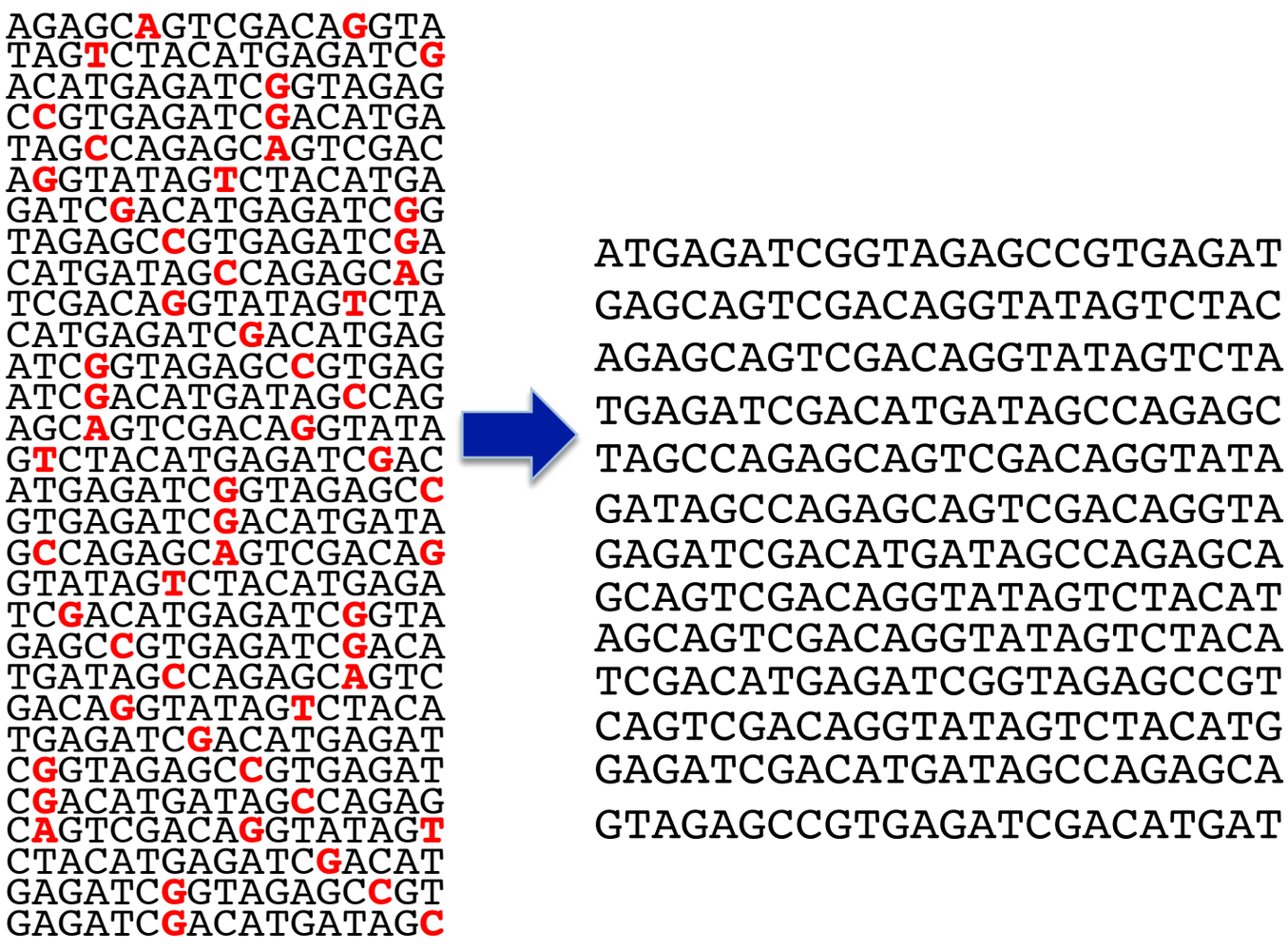
Many short reads

Short read resequencing

A computer science problem

Sequencers generate random short substrings

Problem : recover original sequence



Short read difficulties

We don’t know where each read comes from

Can’t identify where the mutations are

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

Key idea: “re”-sequencing

My genome is close to the human genome

My Genome:
TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAGCCGT

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAGCCGT

Recovered Sequence:
TACATGAGATC**G**ACATGAGATC**G**GTAGAGC**C**GTGAGATC

“Re”-sequencing challenges

Why do we need computer science?

Sequences are long

- Human genome is 3 billion characters long

Sequencers generate many reads

- 100s of millions

Problem: “map” each read to its location

- Trivial algorithm will take thousands of years

Trivial mapping algorithm

Slide read along the genome and count mismatches

If the mismatches are below a threshold, match

The Human Genome:

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

A Sequence Read:

TCGACATGAGATCGGTAGAGCCGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAGCCGT

18 mismatches. Not below threshold. Not a match

Trivial mapping algorithm

Slide read along the genome and count mismatches

If the mismatches are below a threshold, match

The Human Genome:

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
```

A Sequence Read:

```
TCGACATGAGATCGGTAGAGCCGT
```

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
 TCGACATGAGATCGGTAGAGCCGT
```

15 mismatches. Not below threshold. Not a match

Trivial mapping algorithm

Slide read along the genome and count mismatches

If the mismatches are below a threshold, match

The Human Genome:

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
```

A Sequence Read:

```
TCGACATGAGATCGGTAGAGCCGT
```

```
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
          TCGACATGAGATCGGTAGAGCCGT
```

3 mismatches. Below threshold. Match

Complexity of trivial mapping algorithm

3,000,000,000 genome length (N)

300,000,000 reads (M)

30 read length (L)

Number of mismatches allowed 2 (D)

Each comparison takes 1/1,000,000 seconds (t)

Total time = NMLt = 27,000 billion seconds = ~850,000 years

Some observations

Most positions in the genome match poorly

We are looking for small mismatches (D is small)

A substring will match perfectly

Finding a perfect substring match

Create an index

Can look up an entry quickly

If L=30, each key of length 10

Will contain $N/4^{10}$ values on average (~3K)

If L=45, key length = 15

Will contain values (~3)

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

Complexity of indexing algorithm

Look up each third of read

For L=30, index will contain keys of length 10. Each key will contain an average of $N/4^{L/3}$ or 3000 positions

For each position, we need to compute number of mismatches.

Running time is $L \times M \times 3 \times N / 4^{L/3} \times T$ = 81 million seconds = 937 days

L=45, time = 81 thousand seconds = 22.5 hours

More problems

Sequencing errors

Each sequence can have some random errors

My Genome:
TACATGAGATCGACATGAGATCGGTAGAGCCGTGAGATC

A Sequence Read:
TCGACATGAGATCGGTAGAACCGT

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
TCGACATGAGATCGGTAGAACCGT

Recovered Sequence:
TACATGAGATCGACATGAGATCGGTAGAACCGTGAGATC

More problems

Repeat sequences

The Human Genome:

TACATGAGATCCACATGAGATCTGTACATGAGATCCACAT

Repeated Region

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

More problems

Insertions

My Genome:
TACATGAGATCCACAT**A**GAGATCTGTAGAGCTGTGAGATC

A Sequence Read:
CCACATAGAGATCTGTAGAGCTGT

The Human Genome:
TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC

CCACATAGAGATCTGTAGAGCTGT

TACATGAGATCCACATGAGATCTGTAGAGCTGTGAGATC
CCACATAGAGATCTGTAGAGCTGT

Many other challenges

Coverage of sequence reads is not uniform

- Some places have many reads while some have few.

Large memory requirements

- Need to fit index into our RAM (~10s of Gb)

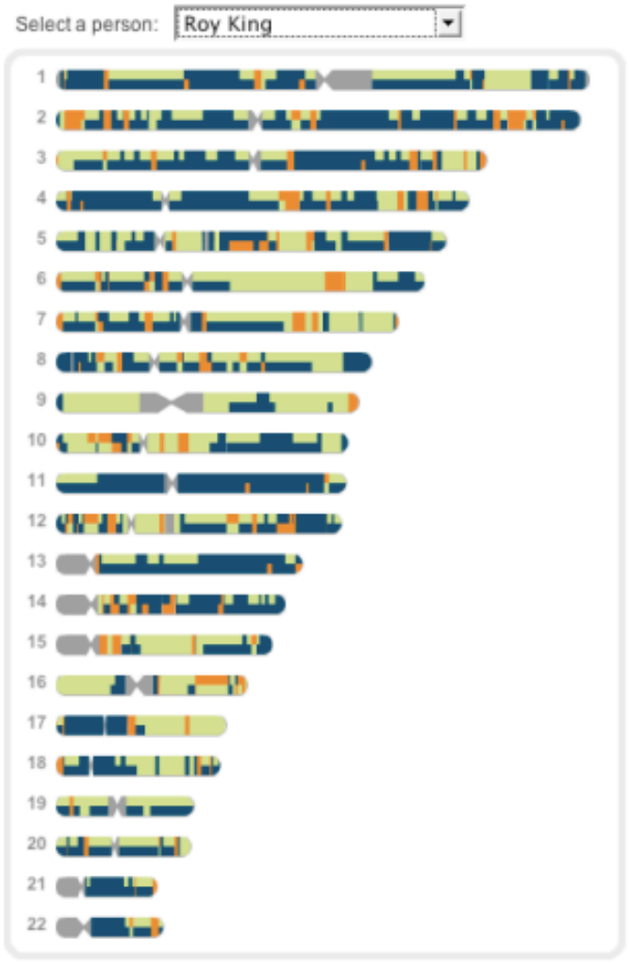
Bioinformatics meets machine learning

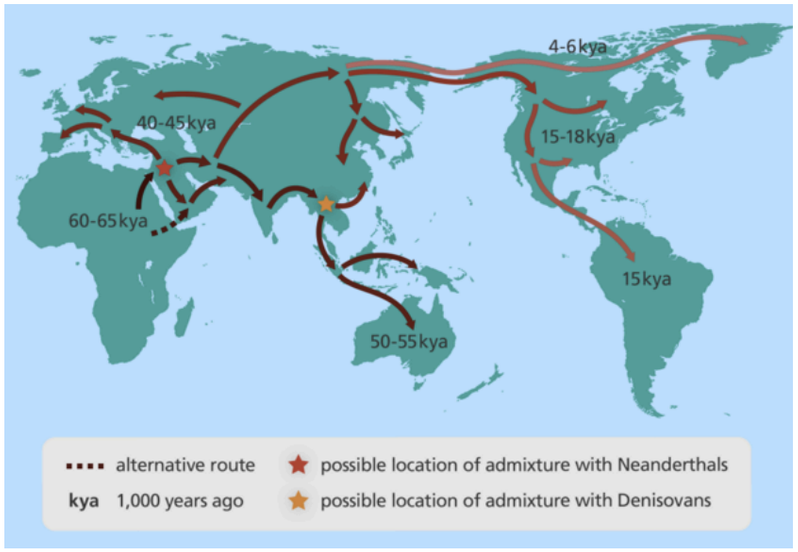
Bioinformatics: Answering biological questions using tools from computer science, statistics and mathematics.

Machine Learning: Learning from data

Inferring human history and ancestry

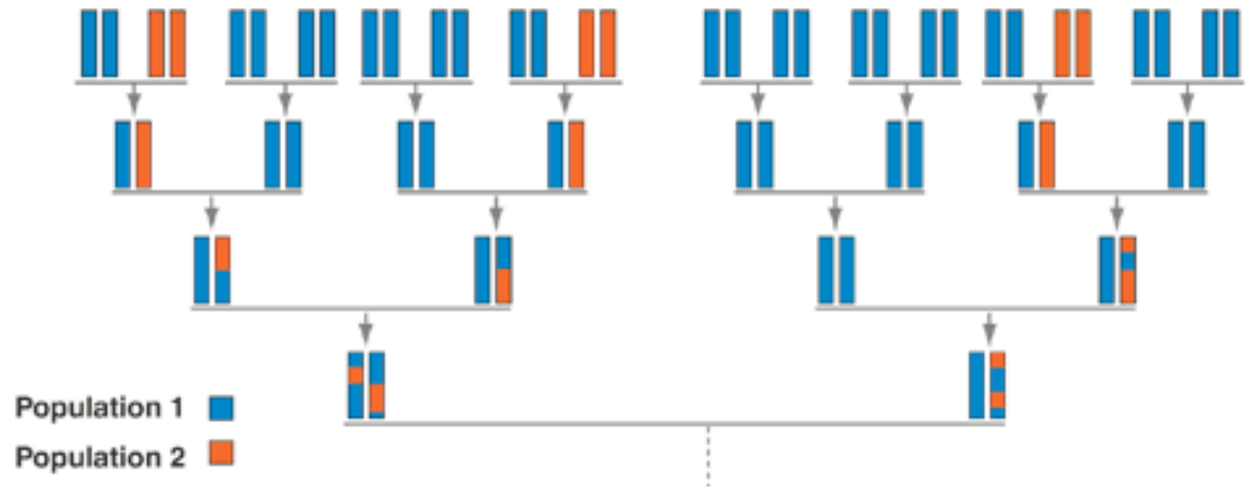
History of human populations learned from genetics





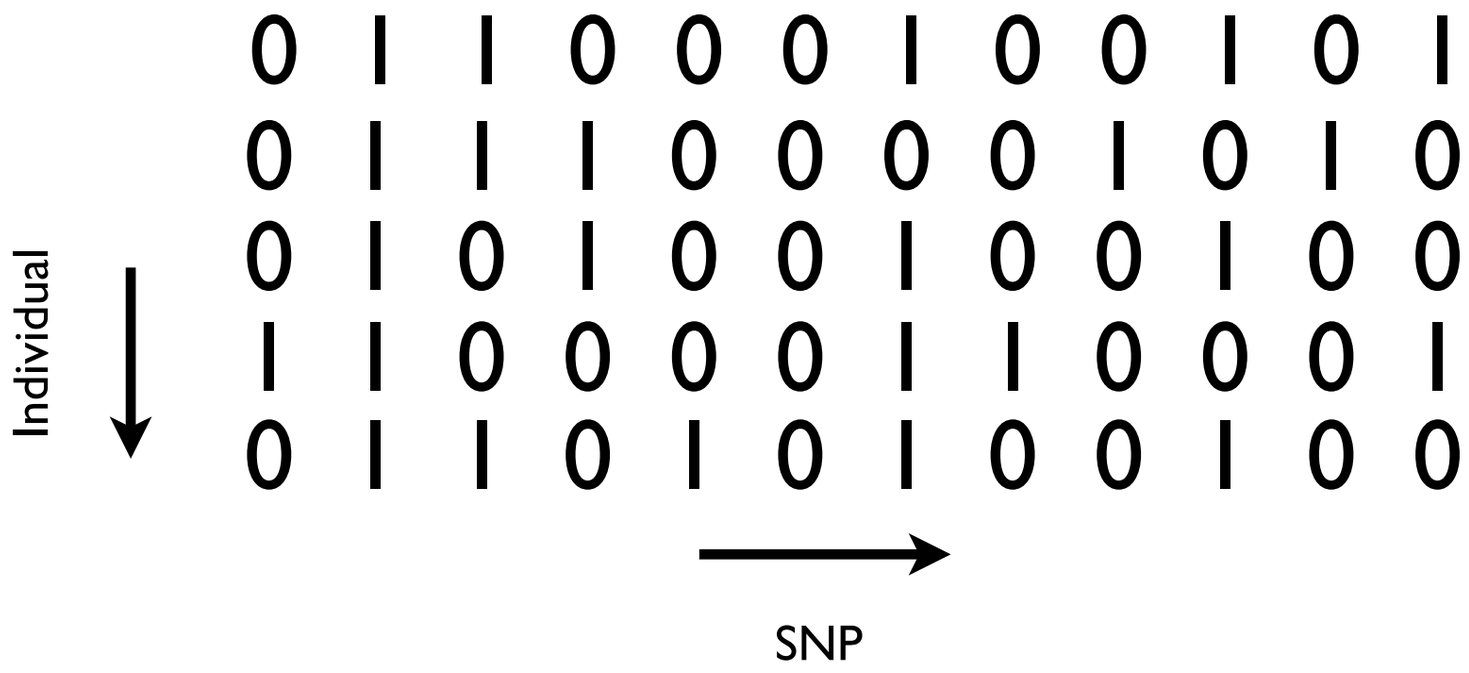
https://blog.23andme.com/23andme-and-you/genetics-101/a-beautiful-ancestry-painting/

Admixture

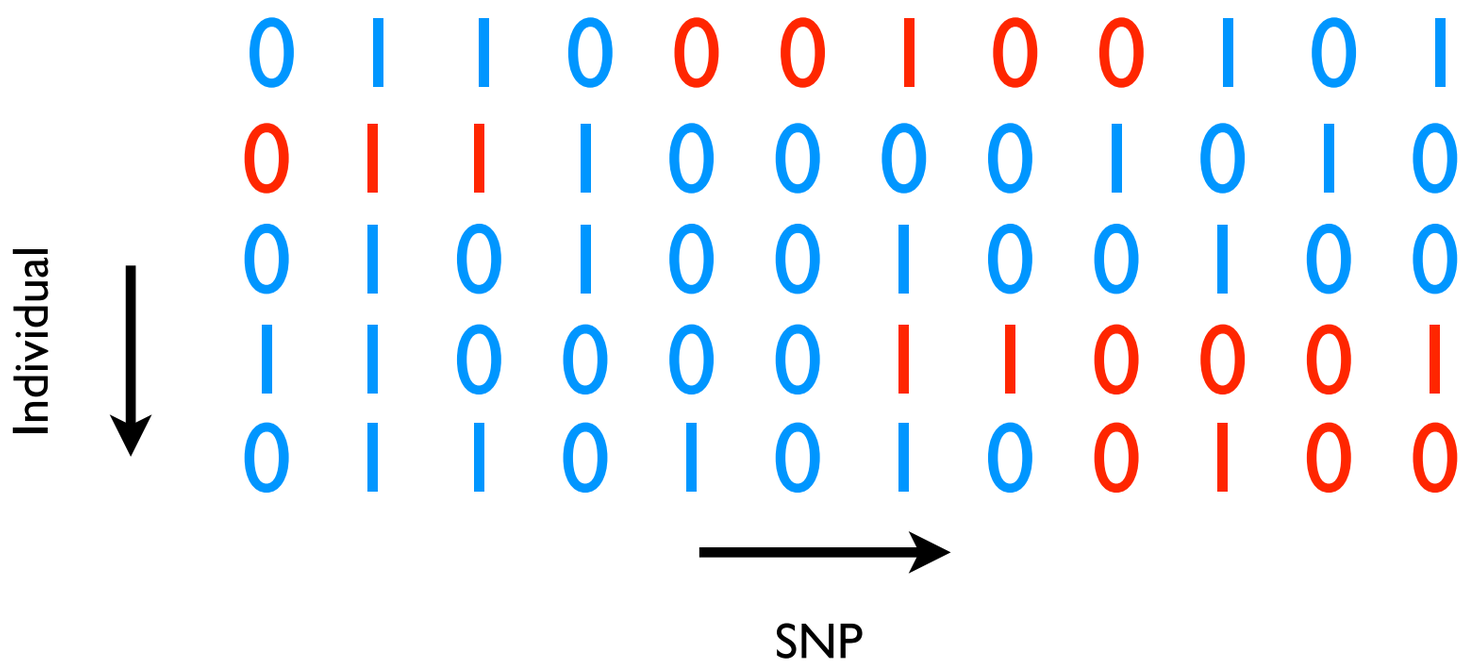


Darvasi and Shifman, *Nature Genetics*,2005

Active research problem: Local ancestry inference

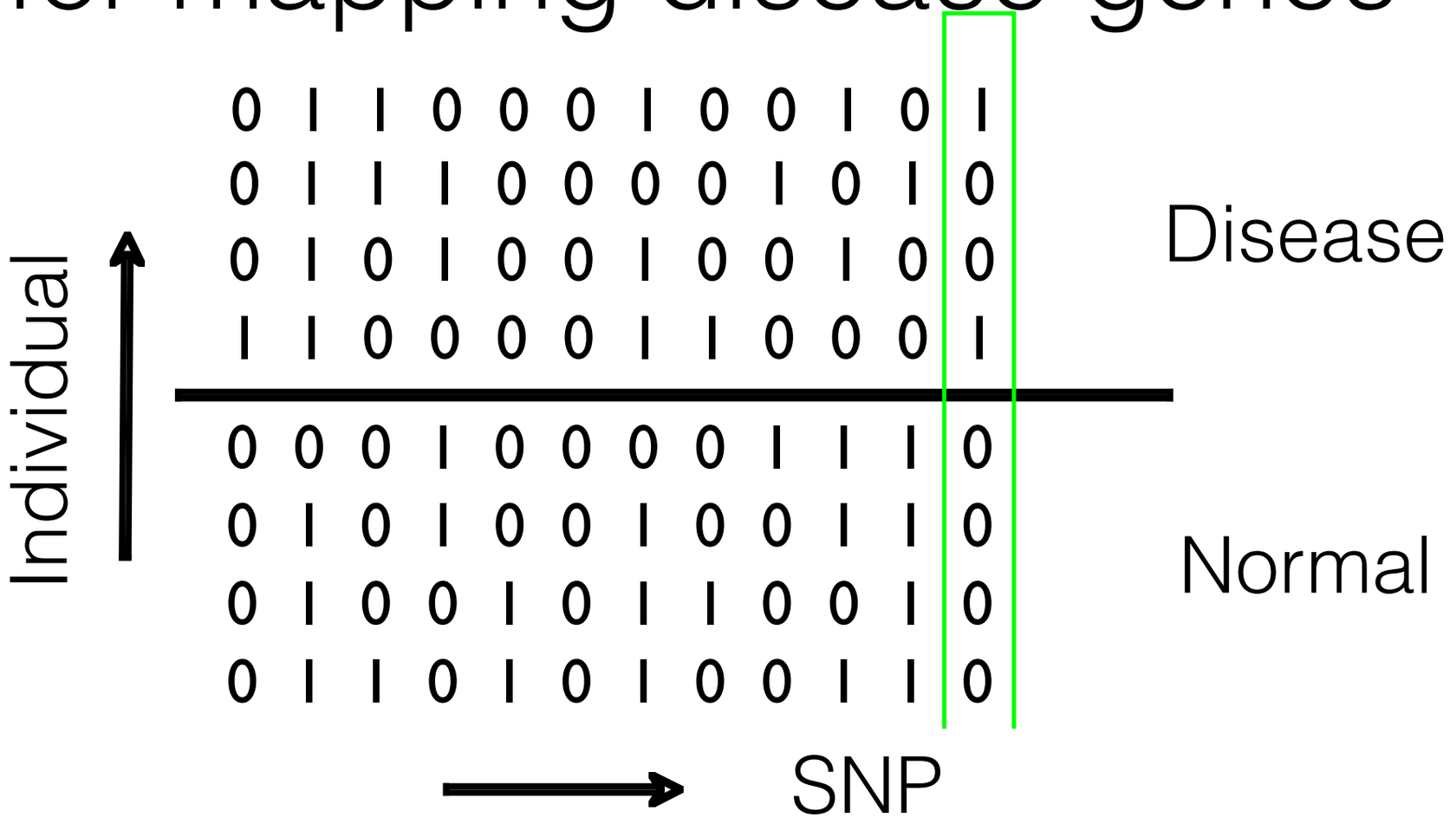


Active research problem: Local ancestry inference

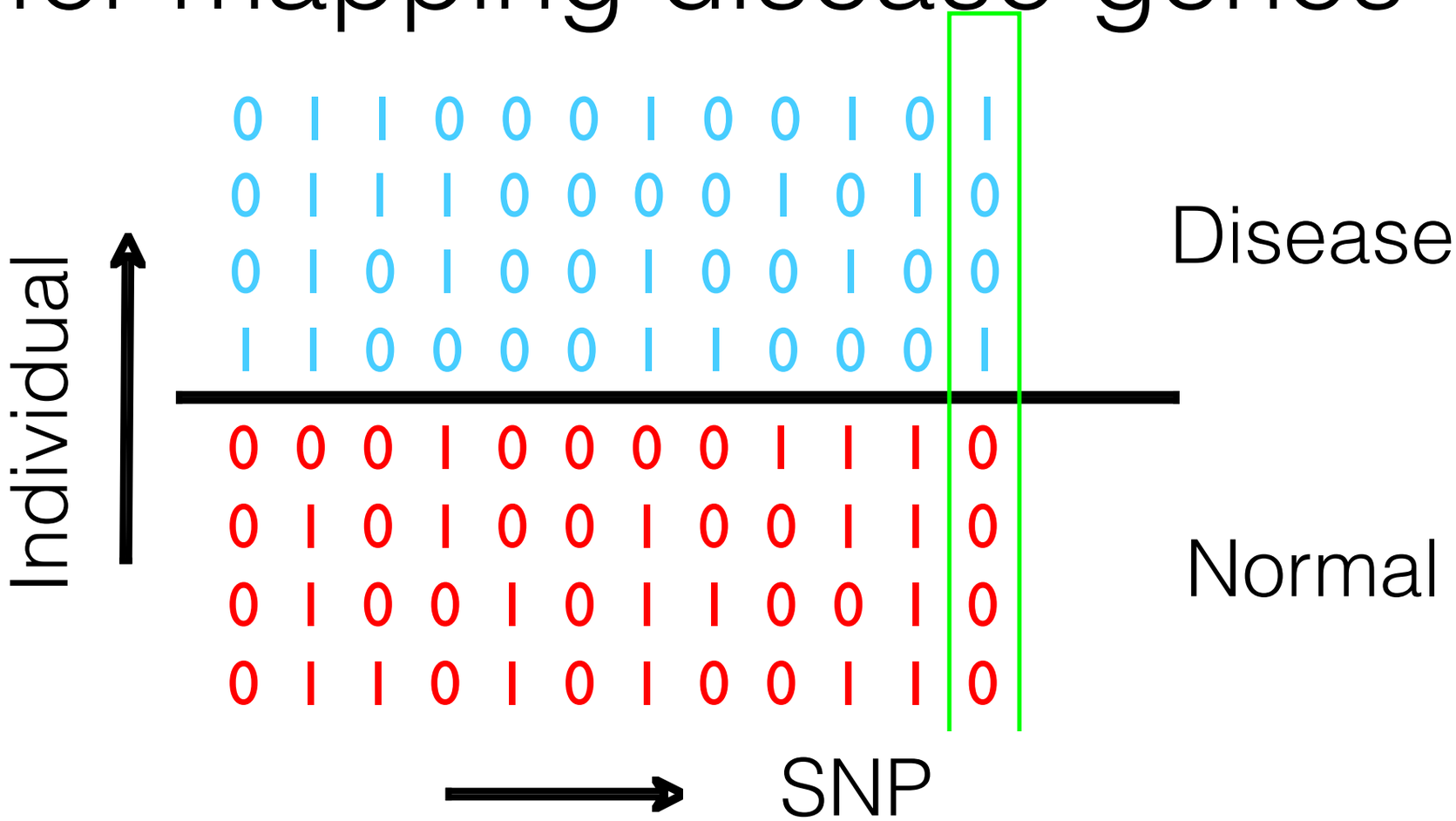


The ancestry of an admixed individual is described by its local ancestries.

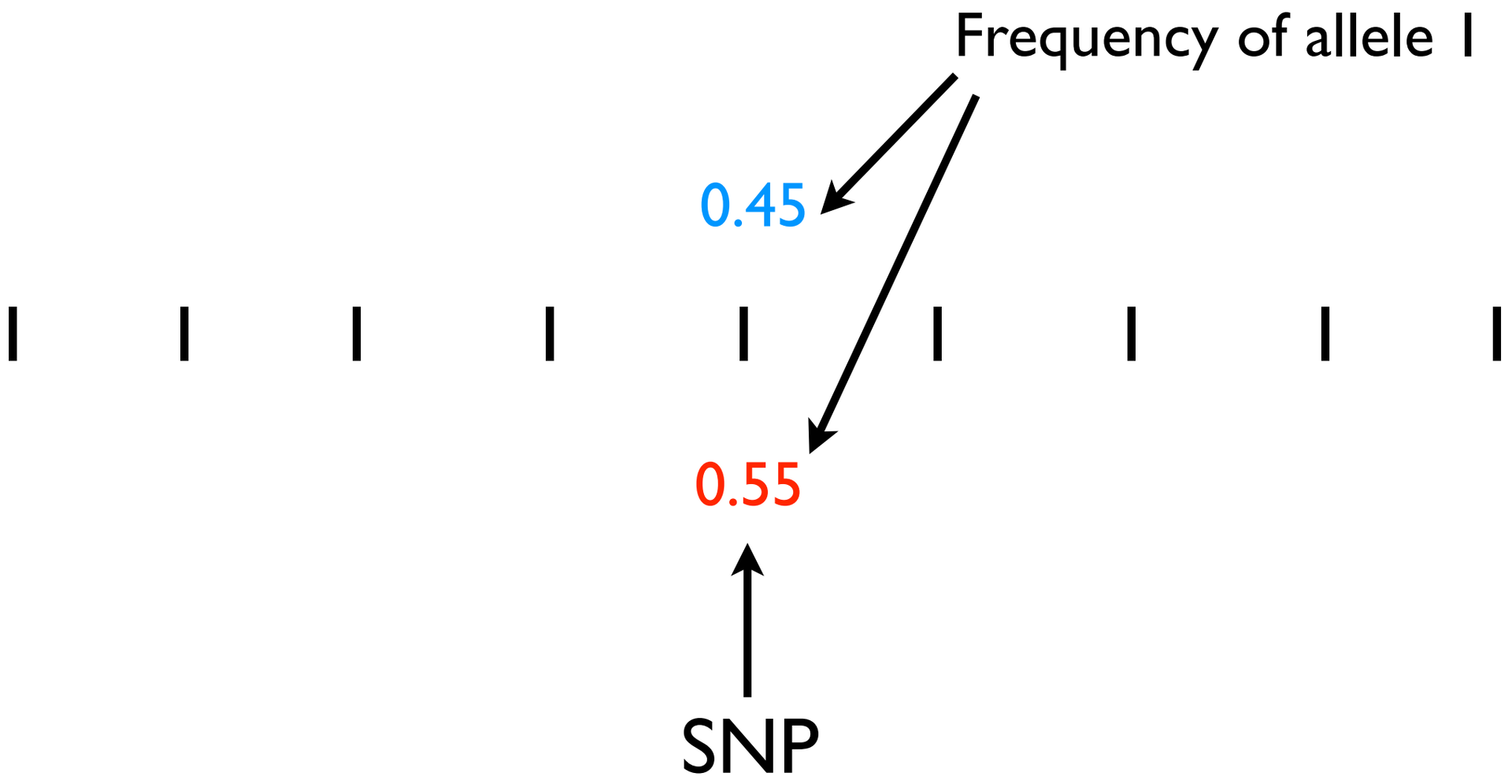
Understanding ancestry important for mapping disease genes



Understanding ancestry important for mapping disease genes



Local ancestry inference



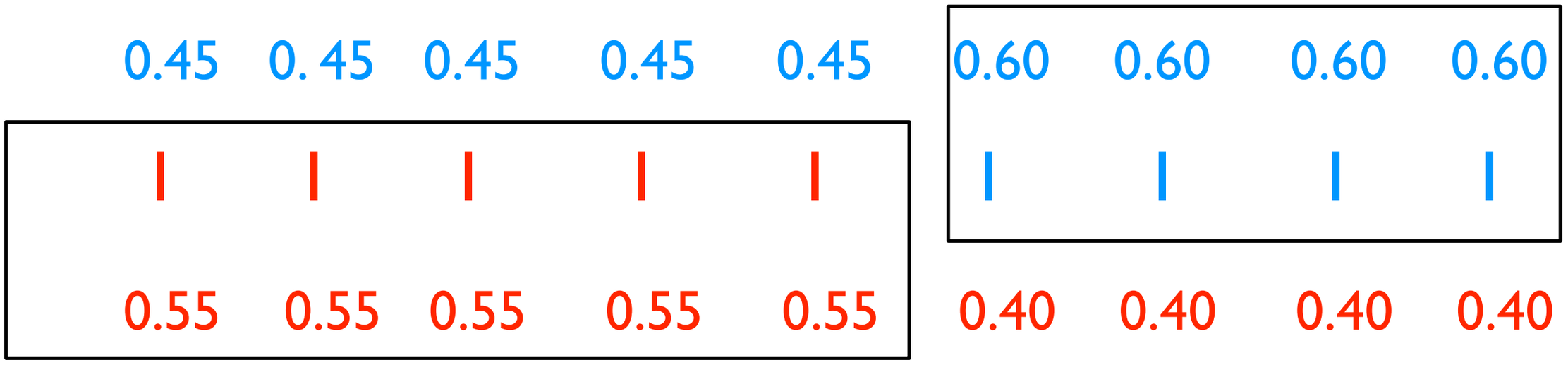
Local ancestry inference

0.45 0. 45 0.45 0.45 0.45 0.60 0.60 0.60 0.60

0.55 0.55 0.55 0.55 0.55 0.40 0.40 0.40 0.40

Local ancestry inference

Machine learning algorithms to infer ancestry



Difficulties

- Ancestry switches not known
- Ancestral allele frequencies unknown or poorly estimated
- Non-independence of SNPs

Inferences used to find disease genes in admixed populations (African-Americans, Latinos)

Ovarian cancer (Bojensen et al. Nature Genetics 2013, Pharoah et al. Nature Genetics, 2013)

Triglyceride levels (Weissglass-Volkov et al. Journal of Medical Genetics, 2013)

Asthma (Torgerson et al. American Journal of Human Genetics, 2012)

Colorectal cancer (Wang et al. Human Molecular Genetics, 2013)

Type 2 Diabetes (Vassy et al. Pediatrics 2012)

Bipolar disorder (Smith et al. Molecular Psyciatry 2009)

Tuberculosis risk (Chimusa et al. Human Molecular Genetics 2013)

Bioinformatics @UCLA

Genomics options of 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!

https://www.seasoasa.ucla.edu/wp-content/uploads/seasoasa/CS-Sci-Tech-List-current.pdf

Technical breadth area in computational genomics

Mostly upper division courses in genomics

Taught by faculty in Bioinformatics program

https://www.seasoasa.ucla.edu/wp-content/uploads/seasoasa/TBA.pdf

Undergraduate Bioinformatics Minor

A way to organize bioinformatics courses into a coherent academic program.

Undergraduate minors since 2013

42 total since 2013

Applied Mathematics (3)
Biochemistry (6)
Bioengineering (2)
Biology (1)
Biophysics (1)
Cognitive Science (1)
Computer Science (3)
Electrical Engineering (1)
Mathematics (1)
Mathematics of Computation (1)
Microbiology, Immunology, and Molecular Genetics (5)
Molecular, Cell, and Developmental Biology (7)
Neuroscience (5)
Statistics (1)

Current Bioinformatics minor enrollments

24 total enrollment (underestimate; based on Fall 2017 snapshot)

Biochemistry (4)
Geology/Paleobiology (1)
Microbiology, Immunology, and Molecular Genetics (4)
Molecular, Cell, and Developmental Biology (4)
Neuroscience (3)
Pre-Computational and Systems Biology (2)
Pre-Microbiology, Immunology, and Molecular Genetics (2)
Psychobiology (2)
Statistics (2)

Grad school placements

MINORS

Elizabeth Chin - Stanford
Andrea Castro - Stanford
Linqing Wei - Berkeley
Alec Chiu - UCLA
Brandon Jew - UCLA
Leah Briscoe - UCLA
Michael Thompson - UCLA
Ruthie Johnson - UCLA
Xingyi Shi - Boston University
Greg Darnell - Princeton

“UNOFFICIAL” MINORS

Jeremy Rotman - UCLA
Ariel Wu - UCLA
Shayna Stein - Harvard
Harry Yang - UCLA

Undergrad Bioinformatics minor

8 Course Minor

- 2 Bioinformatics core courses
- 1 Bioinformatics elective course
- 1 seminar course
- 1 year of programming
- 1 upper division algorithms course
- 1 molecular biology course
- 1 linear algebra course

Many opportunities for research

Undergrad Bioinformatics minor

Undergraduate Core Curriculum (3 courses)

- Co-scheduled with graduate Bioinformatics Ph.D. Core
- CM 121 – Introduction to Bioinformatics (taught by Chis Lee)
- CM 122 – Algorithms in Bioinformatics (taught by Eleazar Eskin)
- CM 124 – Computational Genetics (taught by Eran Halperin)

Undergraduate Research Program

- Enables undergraduates to find available research opportunities
- http://bioinformatics.ucla.edu/undergraduate-research/

Bioinformatics lower division courses

Three required courses are prerequisites for upper division courses

Advanced programming

PIC10C or CS32

Linear Algebra and its applications

Math 33A

Introduction to molecular biology

Life Science 3

Bioinformatics upper division electives

Statistcs 100A, 100B - Introduction to Mathematical Statistics OR Biostatistics 110A,110B - 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

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

Research opportunities

Minor courses can be part of Major program giving additional electives to complete Minor.

Research can help complete minor:

- 8 units of research is available as part of Minor
- with 2 additional units from CM 184 leaves only 10 units required to complete Minor

Undergraduate Research Program hosts:

- bioinformatics.ucla.edu/undergraduate-research/
- Many available projects with Bioinformatics Faculty
- Leads to a few Bioinformatics Ph.D. students each year

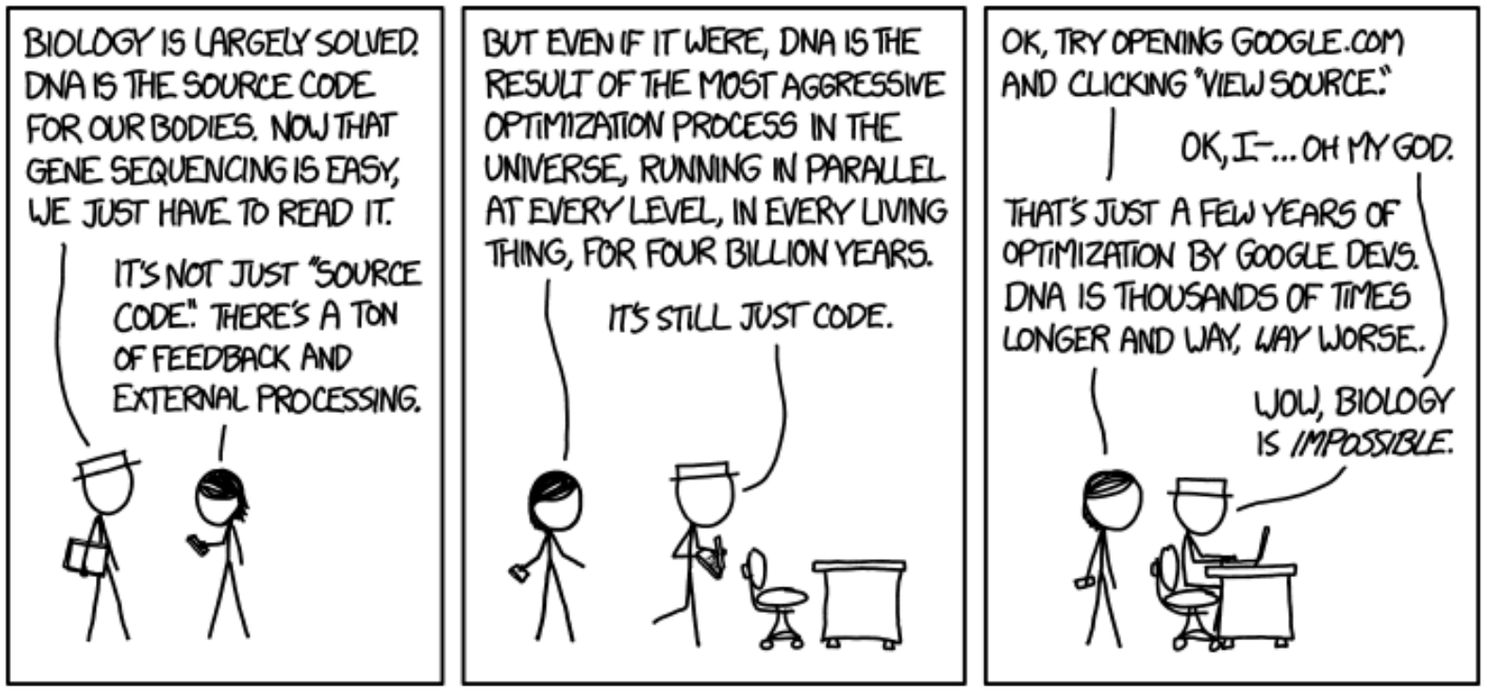
Questions ?

Visit the Bioinformatics Minor webpage:
bioinformatics.ucla.edu/undergradute-bioinformatics-minor/

Course requirements:
bioinformatics.ucla.edu/bioinformatics-minor-course-requirements/

Minor FAQs:
bioinformatics.ucla.edu/bioinformatics-minor-faqs/

Research Opportunities:
bioinformatics.ucla.edu/undergraduate-research/



https://xkcd.com/1605/