Linking Genetic Variation to Phenotypes

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2021

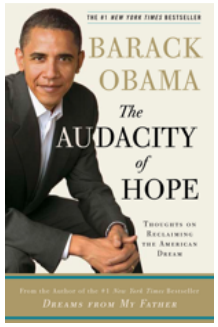
Daifeng Wang

daifeng.wang@wisc.edu

Outline

- How does the genome vary between individuals?
- How do we identify associations between genetic variations and simple phenotypes/diseases?
- How do we identify associations between genetic variations and complex phenotypes/diseases?

How to read sentences/genes for understanding book/genome?



Chapter One

Republicans and Democrats

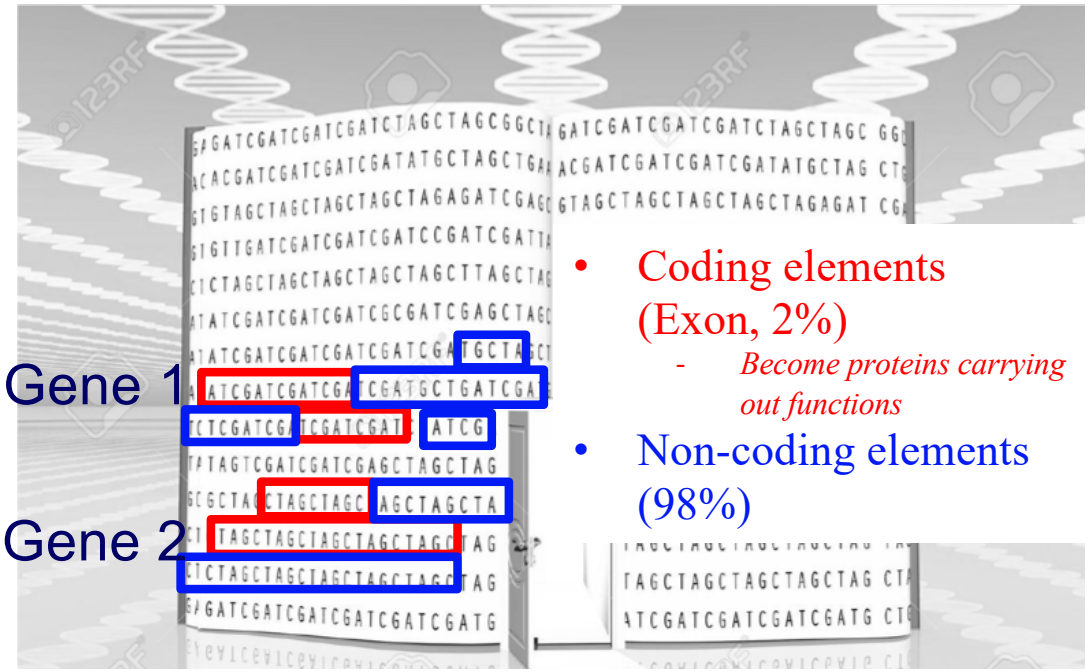
Book	Genome
Chapters	Chromosomes
Sentences	**Genes**
Words	**Elements**
Letters	Bases

"On most days, I enter the Capitol through the basement. A small subway train carries me from the Hart Building, where …"

- Key words
- Non-key words

Overhead, the ceiling forms a creamy white oval, with an American eagle etched in its center. Above the visitors' gallery, the busts of the nation's first twenty vice presidents sit in solemn repose.

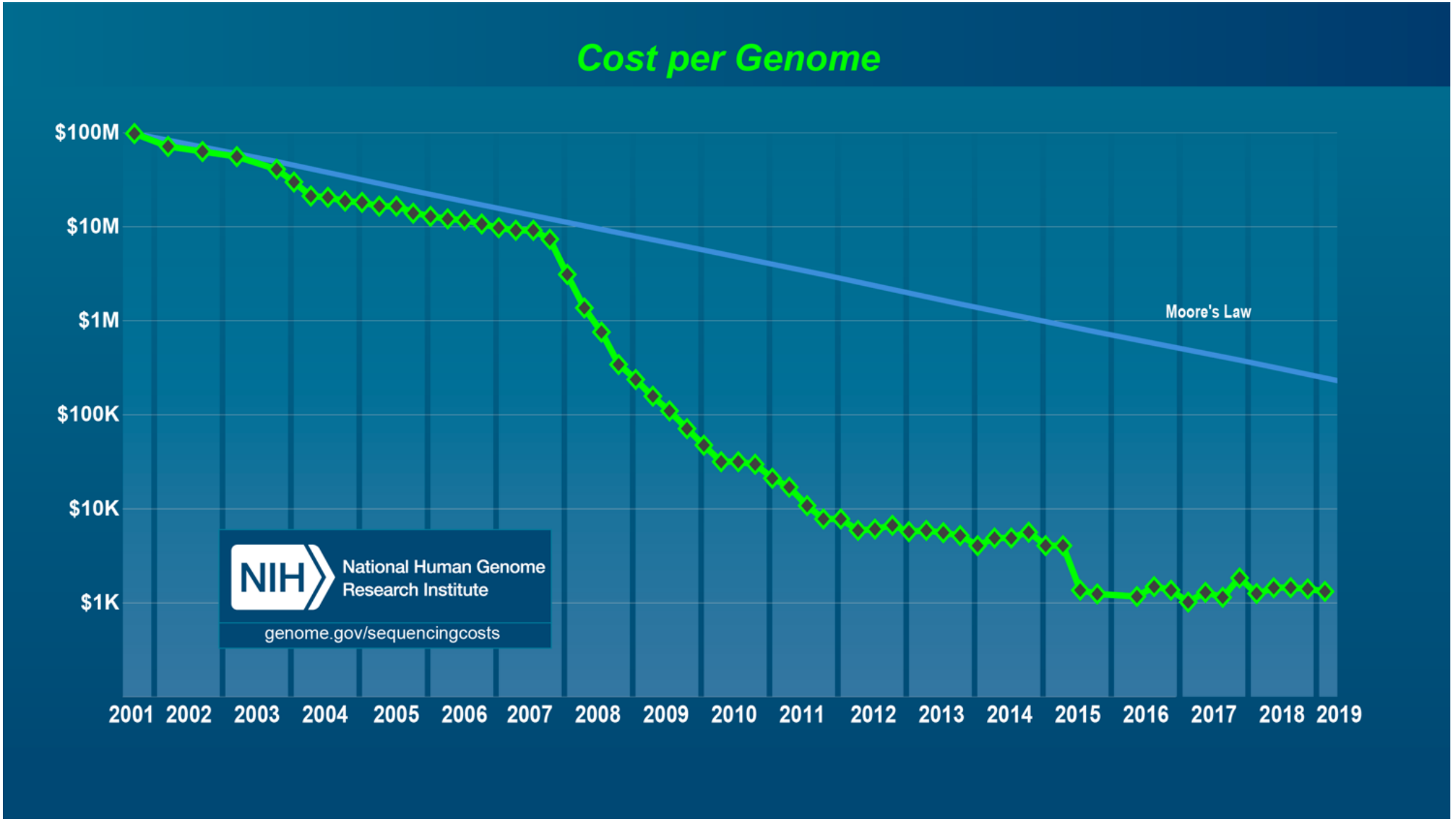
And in gentle steps, one hundred mahogany desks rise from the well of the Senate in four horseshoe-shaped rows. Some of these desks date back to 1819, and atop each desk is a tidy receptacle for inkwells and quills. Open the drawer of any desk, and you will find within the names of the senators who once used it—Taft and Long, Stennis and Kennedy—scratched or penned in the senator's own hand. Sometimes, standing there in



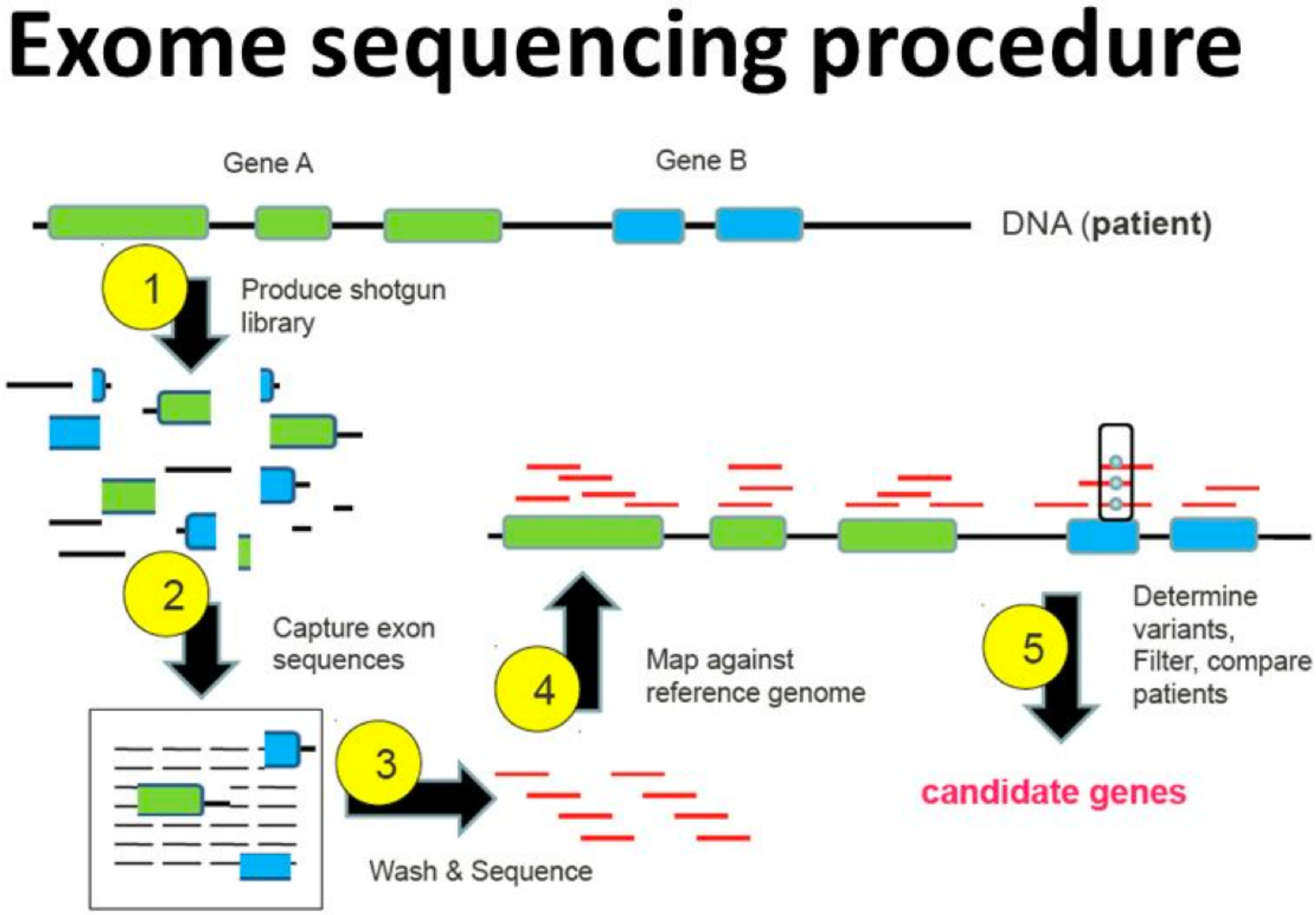
- Coding elements (Exon, 2%)
  - *Become proteins carrying out functions*
- Non-coding elements (98%)

https://goo.gl/images/vMaz4T

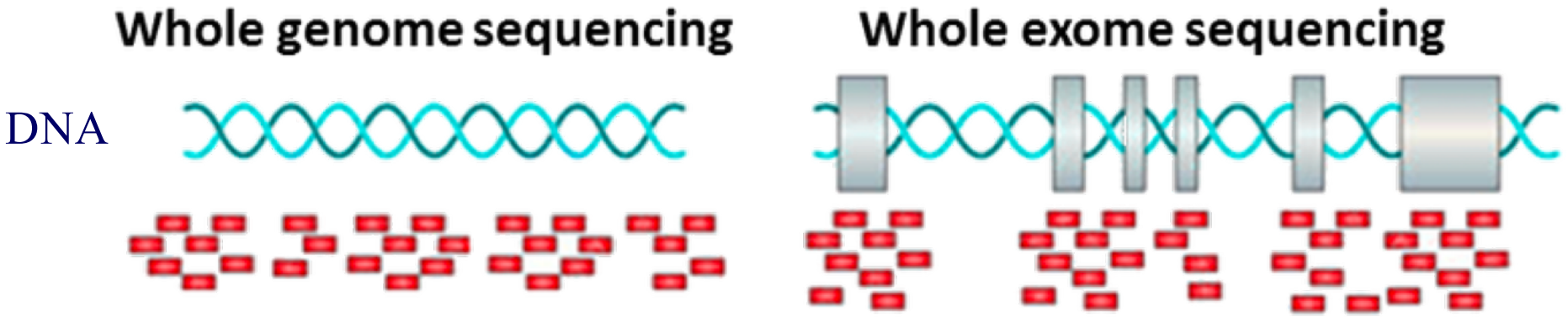
Low sequencing cost enables reading our whole genome



Whole Exome Sequencing (WES) reads 2% coding elements of human genome



Whole Genome Sequencing (WGS) reads 100%!



- Sequencing region : whole genome
- Sequencing Depth: >30X
- Covers everything – can identify all kinds of variants including SNPs, INDELs and SV.

- Sequencing region: whole exome
- Sequencing Depth : >50X ~ 100X
- Identify all kinds of variants including SNPs, INDELs and SV in coding region.
- Cost effective

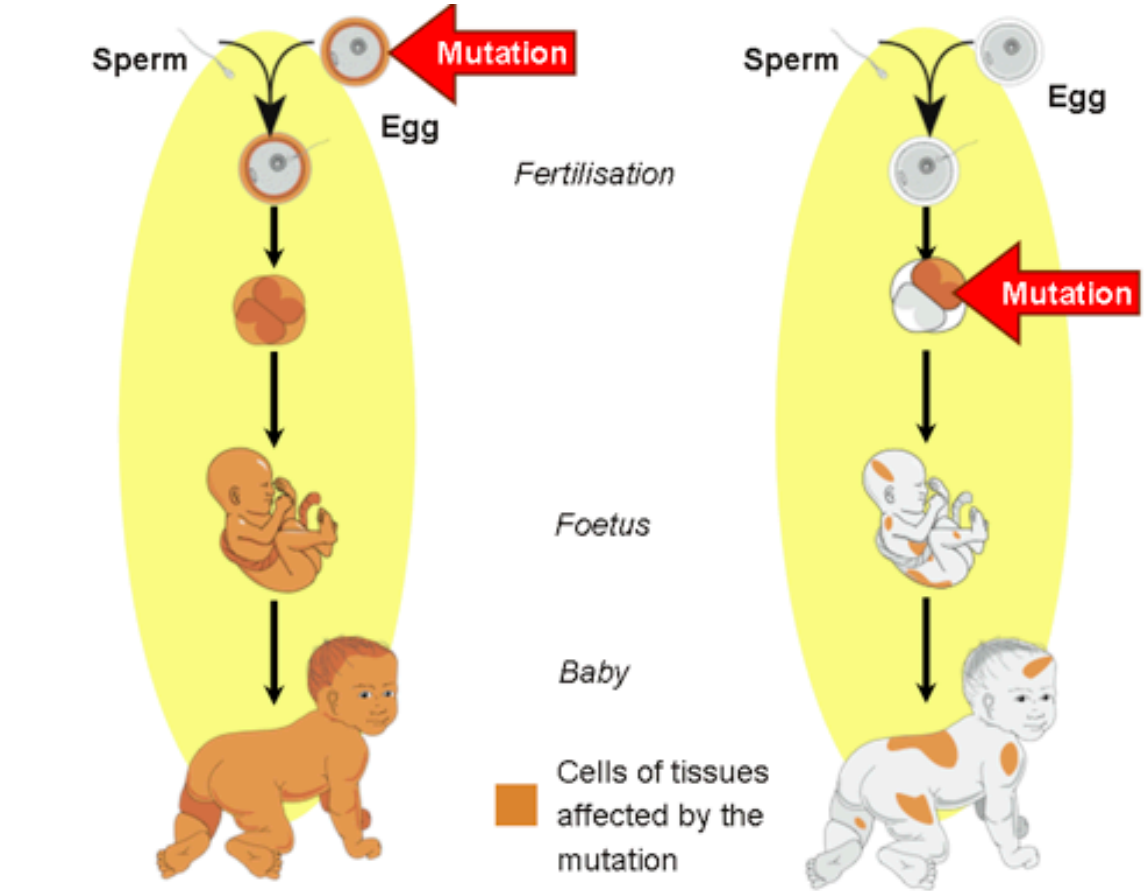
Coding elements

http://www.genomesop.com/somatic-mutations/

Understanding Human Genetic Variation

- The “human genome” was determined by sequencing DNA from a small number of individuals (2001)
- The HapMap project (initiated in 2002) looked at polymorphisms in 270 individuals (Affymetrix GeneChip)
- The 1000 Genomes project (initiated in 2008) sequenced the genomes of 2500 individuals from diverse populations
- 23andMe genotyped its 1 millionth customer in 2015
- Genomics England sequenced 100k whole genomes and linked with medical records (Dec 2018)

Gametic vs. Somatic Mutations



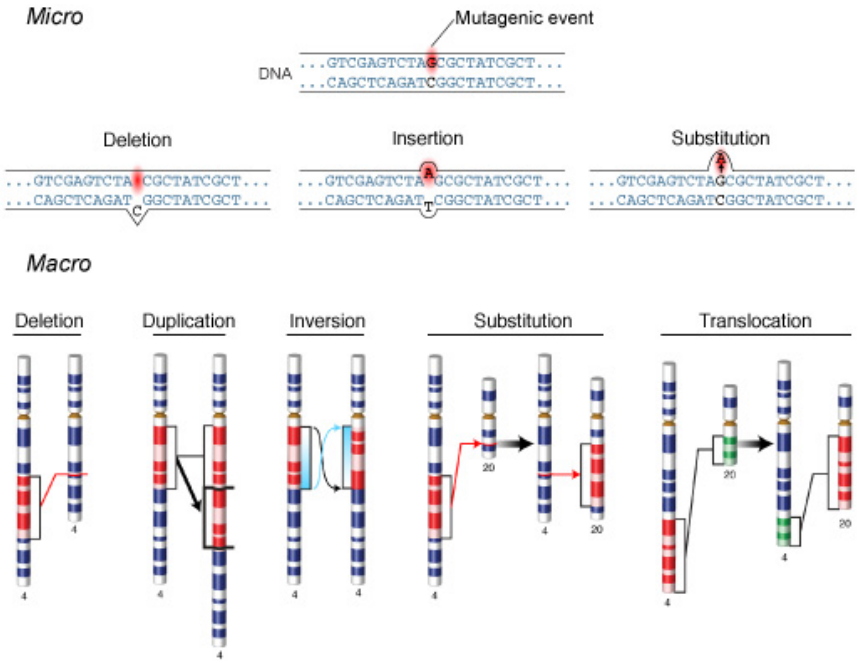
Gametic mutations are inherited and occur in the testes of males and the ovaries of females.

Somatic mutations occur in body cells. They are not inherited but may affect the person during their lifetime.

https://www.pathwayz.org/Tree/Plain/GAMETIC+VS.+SOMATIC+MUTATIONS

Classes of Variants

- Single Nucleotide Polymorphisms (SNPs)
- Indels (insertions/deletions)
- Structural variants



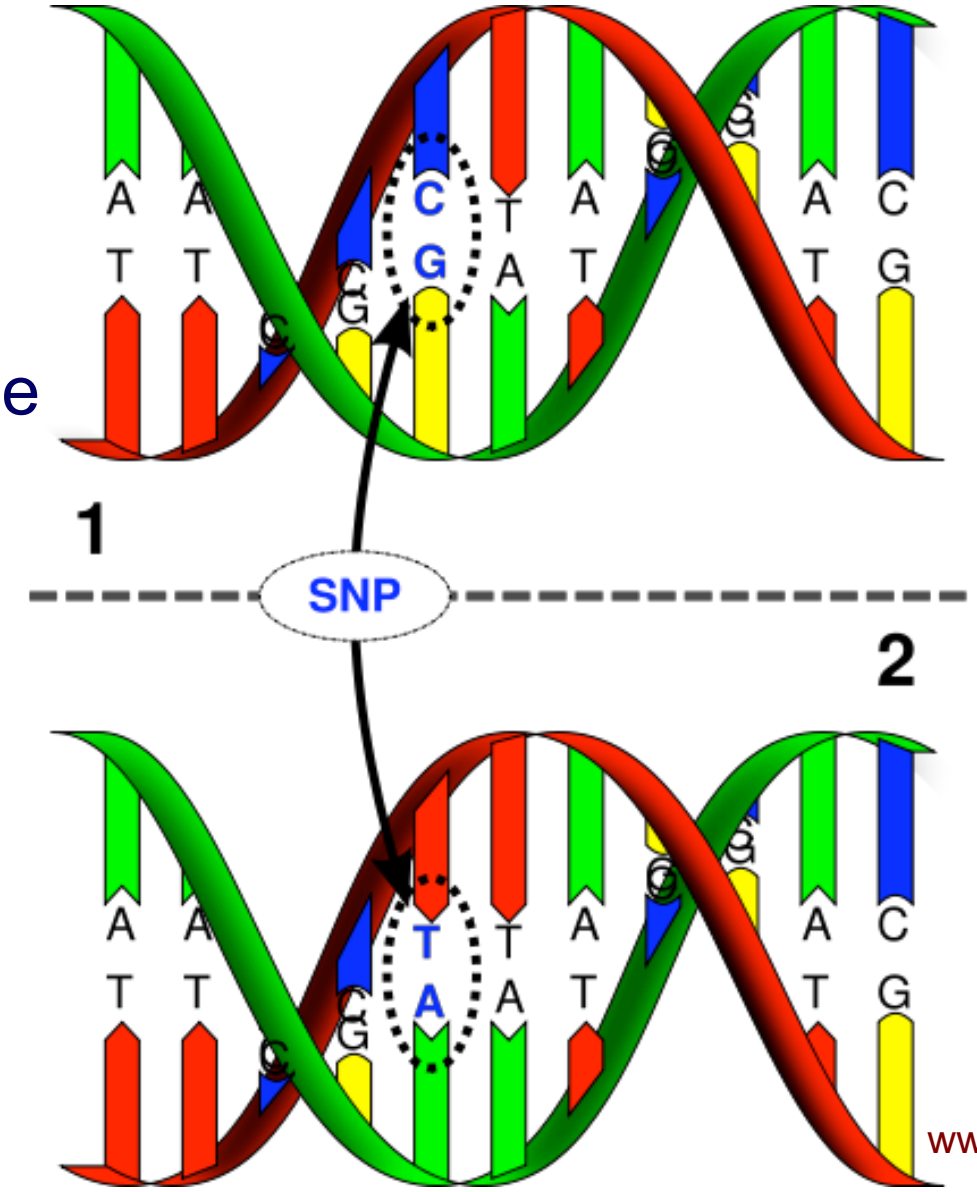
Formal definitions: https://www.snpedia.com/index.php/Glossary

Single Nucleotide Polymorphisms (SNPs)

One nucleotide changes

Variation occurs with some minimal frequency in a population

Pronounced “snip”

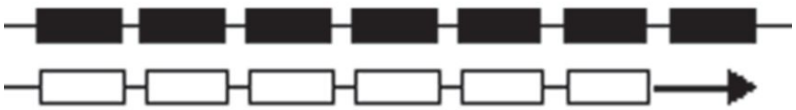


Single Nucleotide Polymorphisms (SNPs) normally happen ~1% on individual human genome.

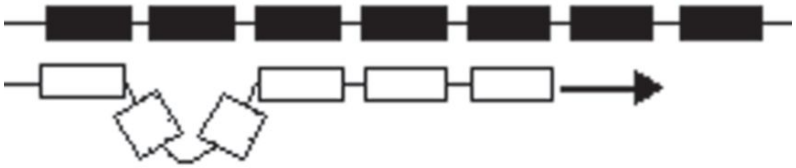
Most SNPs are harmless but some matter

Talking Glossary of Genetic Terms
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
NATIONAL INSTITUTES OF HEALTH | genome.gov
Illustration by Darryl Leja, NHGRI

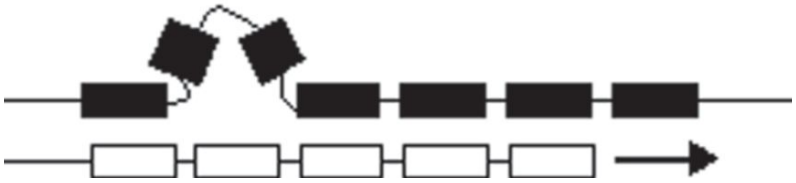
Insertions and Deletions



Black box: DNA template strand
White box: newly replicated DNA



Insertion: slippage inserts extra nucleotides



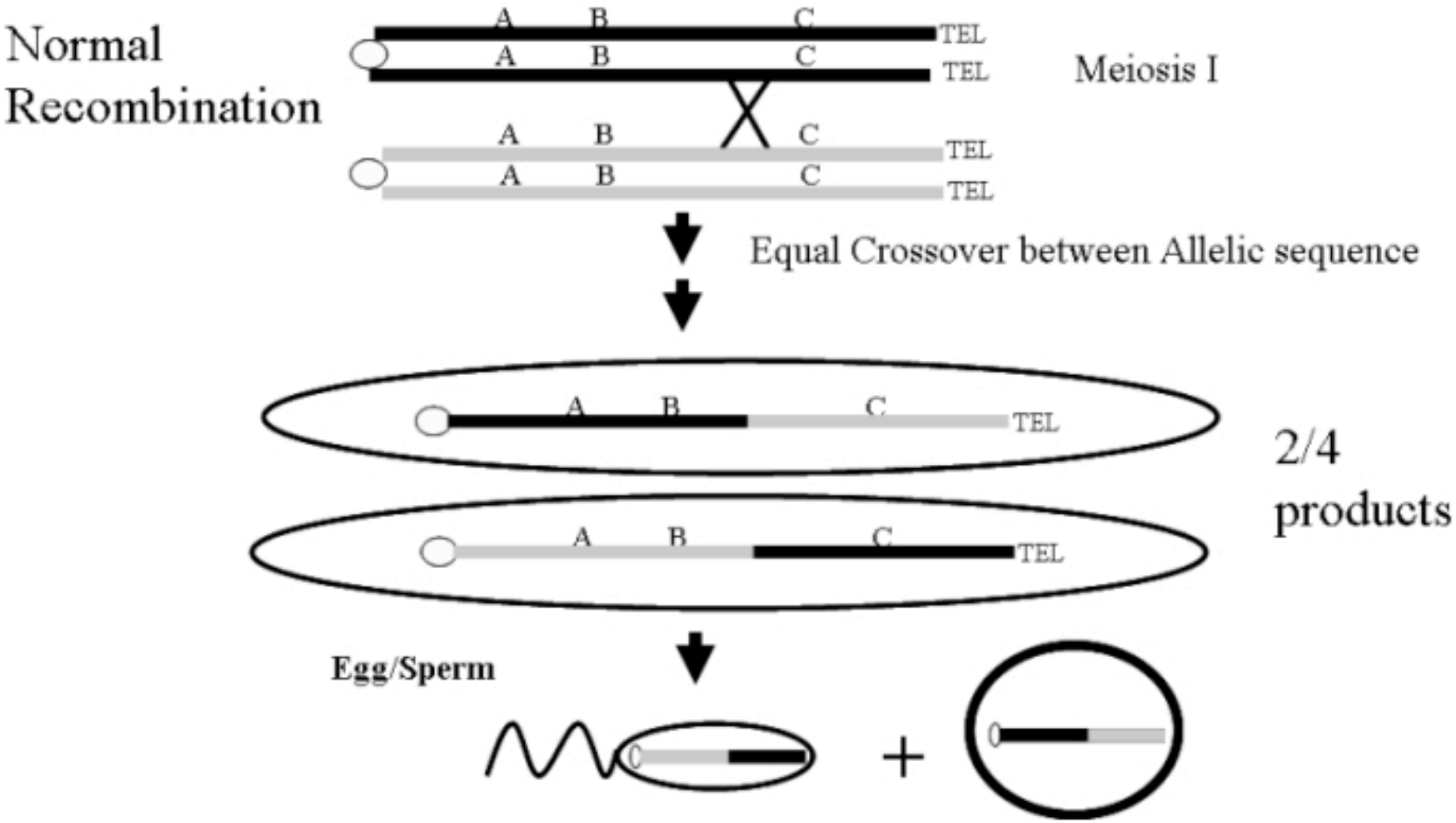
Deletion: slippage excludes template nucleotides

Forster et al. *Proc. R. Soc. B* 2015

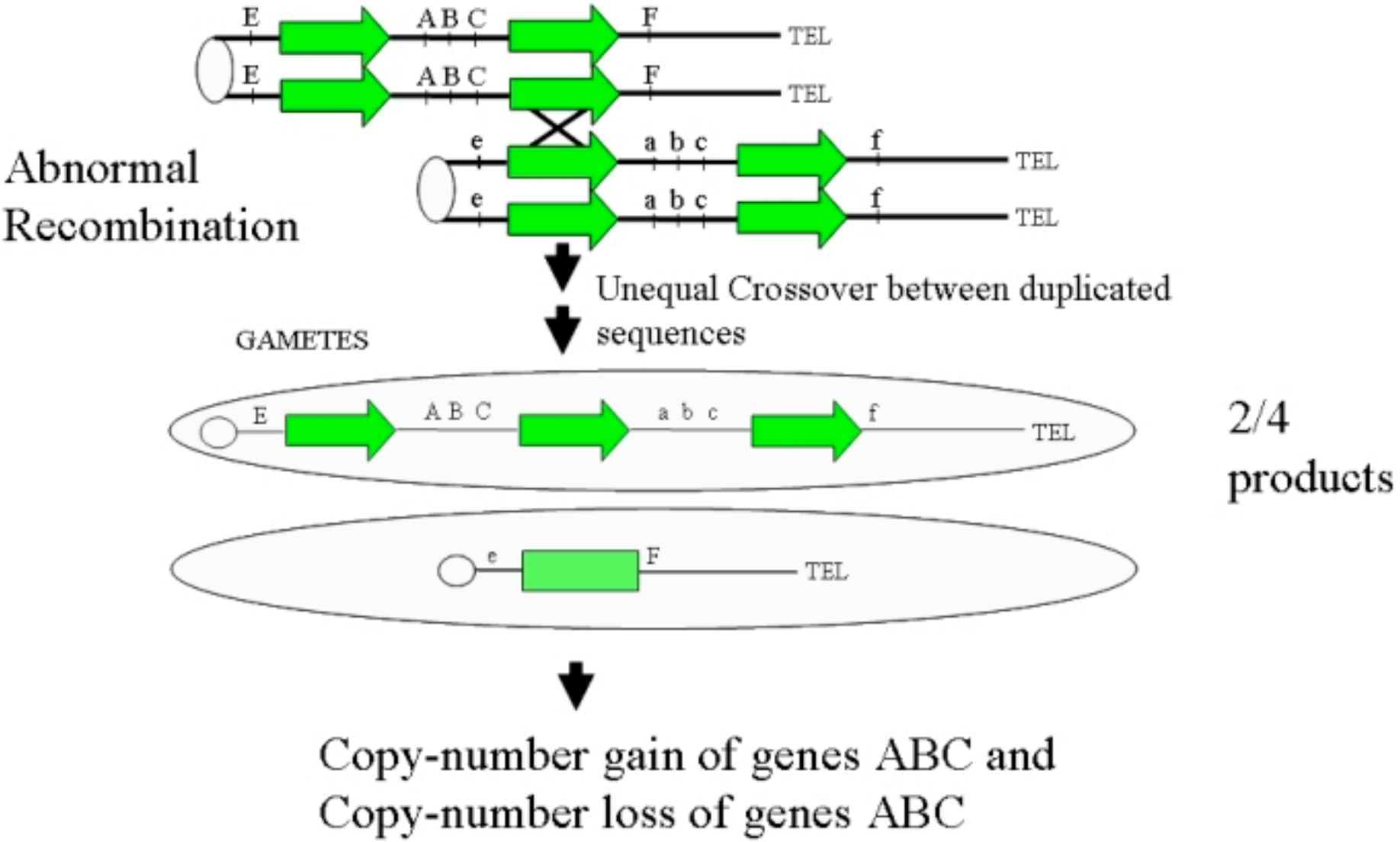
Structural Variants

- Copy number variants (CNVs)
  - Gain or loss of large genomic regions, even entire chromosomes
- Inversions
  - DNA subsequence is reversed
- Translocations
  - DNA subsequence is moved to a different chromosome

Genetic Recombination

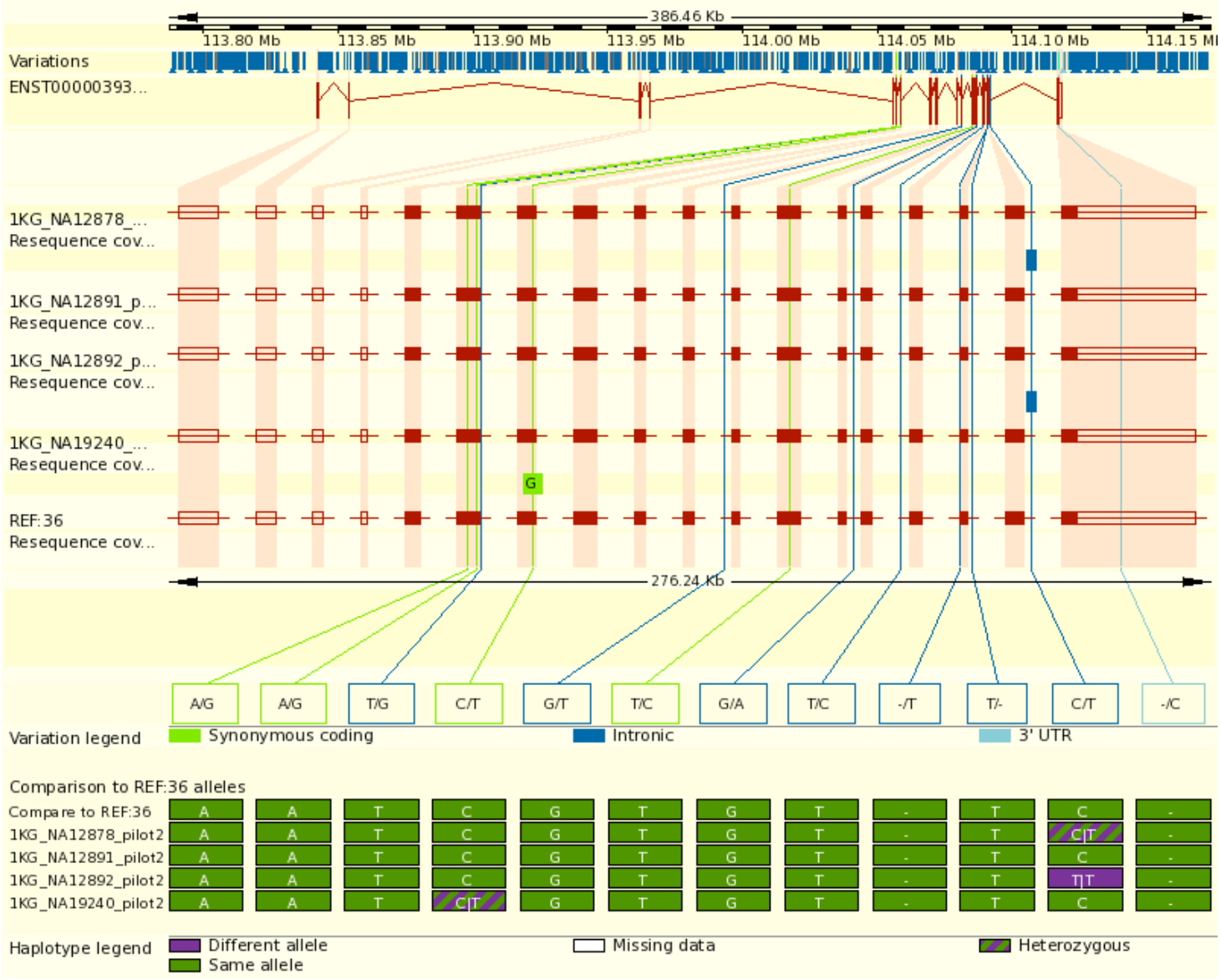


Recombination Errors Lead to Copy Number Variants (CNVs)

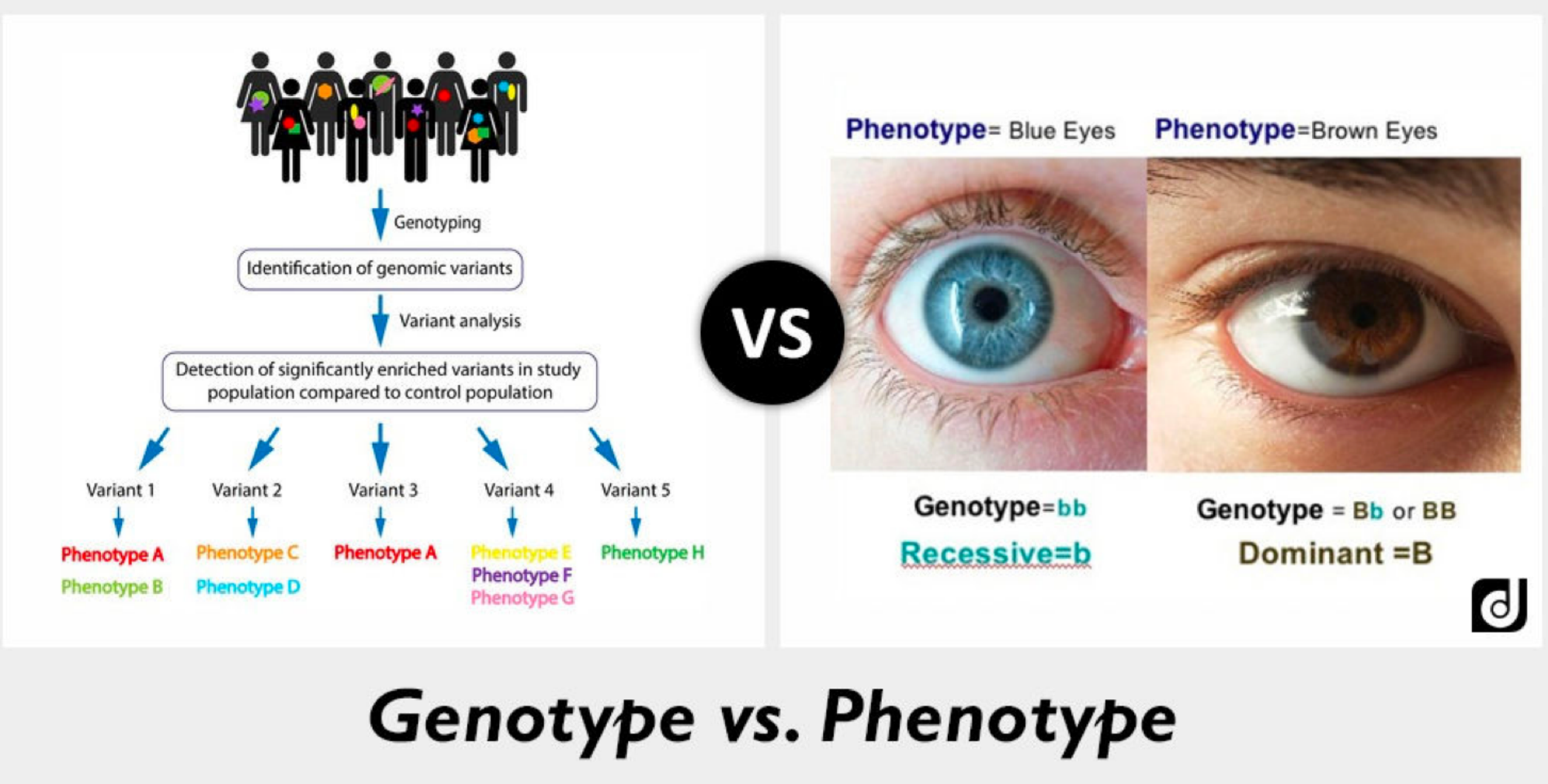


1000 Genomes Project

Project goal: produce a catalog of human variation down to variants that occur at >= 1% frequency over the genome



Genotype to Phenotype



Understanding Associations Between Genetic Variation and Disease

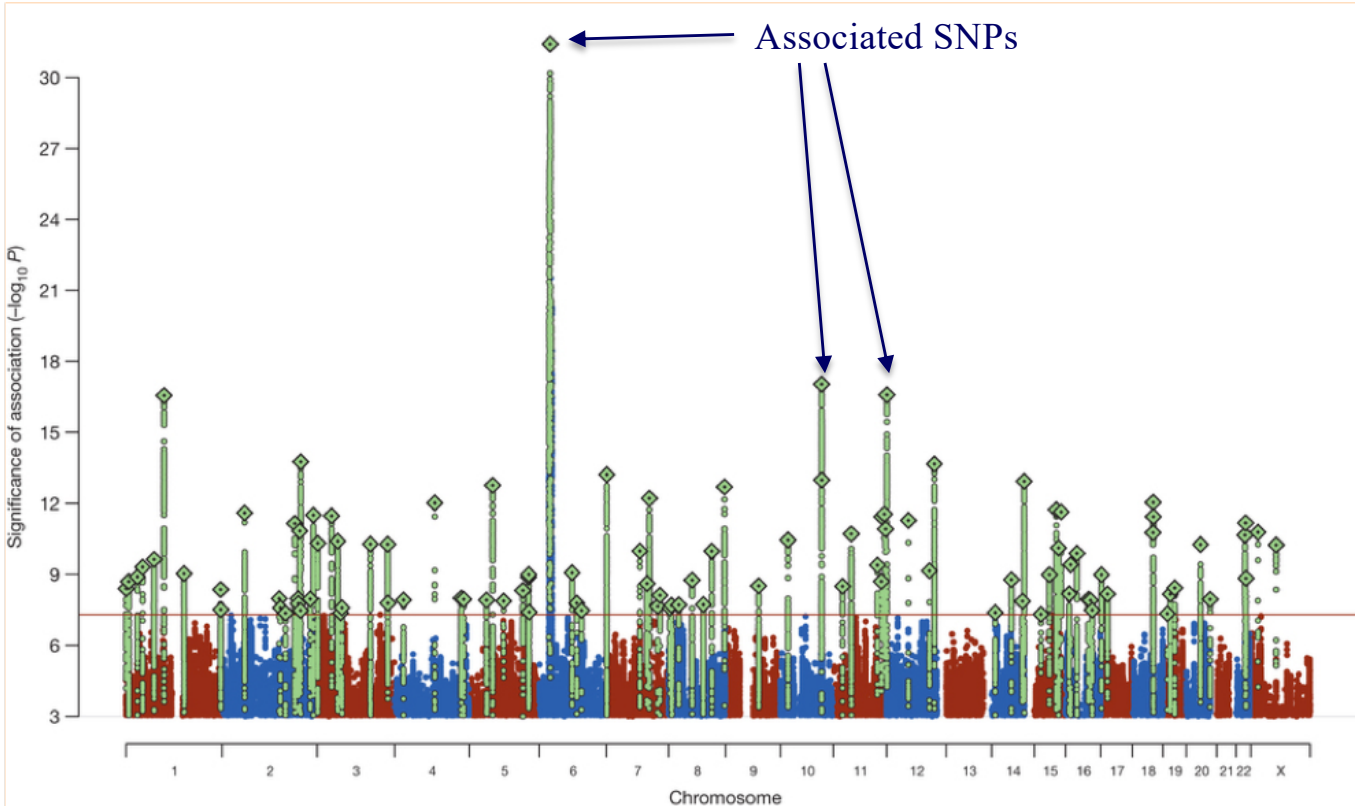
Genome-wide association study (GWAS)

- Gather some population of individuals
- Genotype each individual at polymorphic markers (usually SNPs)
- Test association between *state* at marker and some variable of interest (say disease)
- Adjust for multiple comparisons

- Phenotypes: observable traits

Example: Genome-Wide Association Study (GWAS) identifies disease associated genetic variants

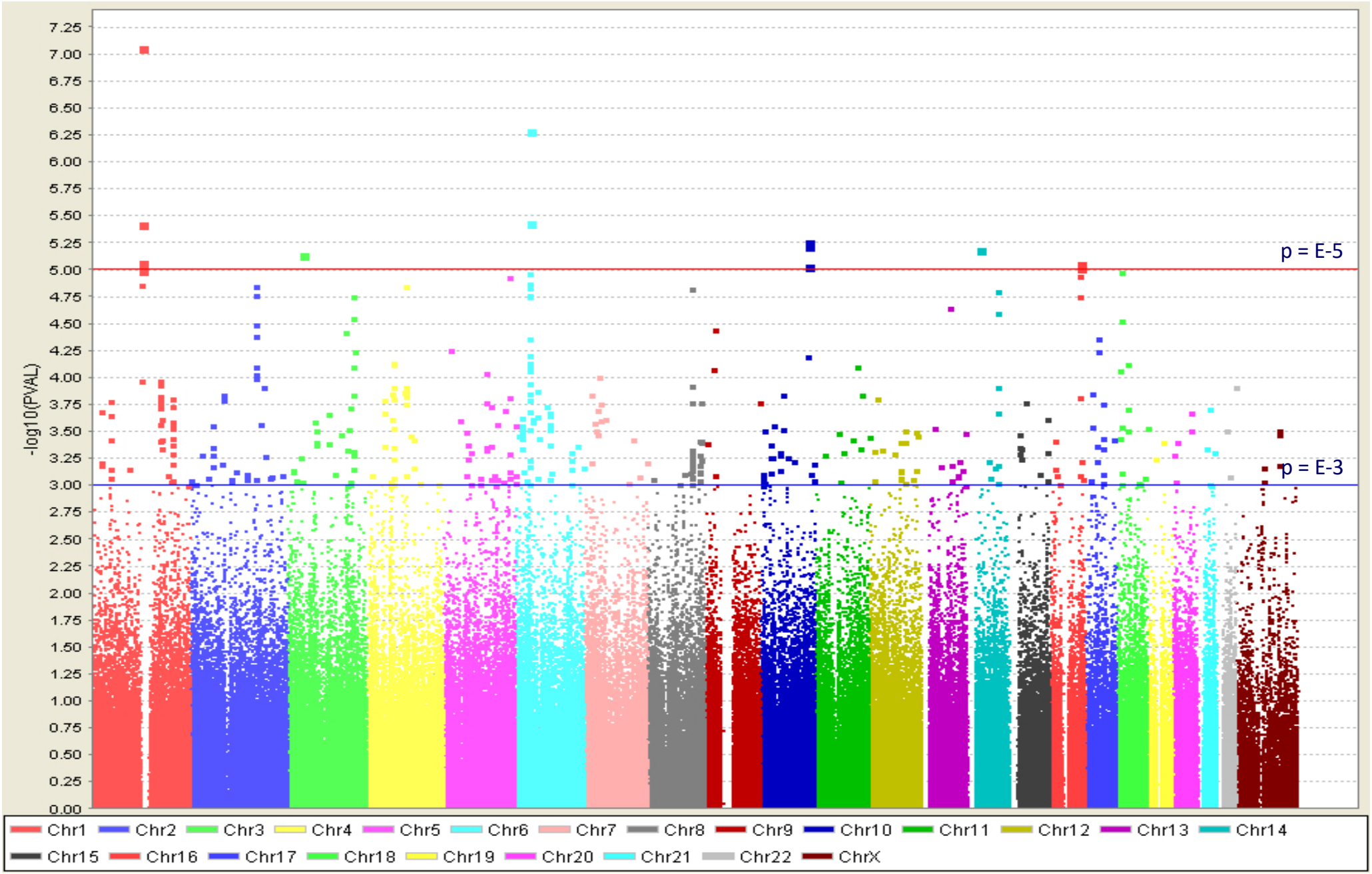
36,989 schizophrenia cases and 113,075 controls
in Psychiatric Genomics Consortium



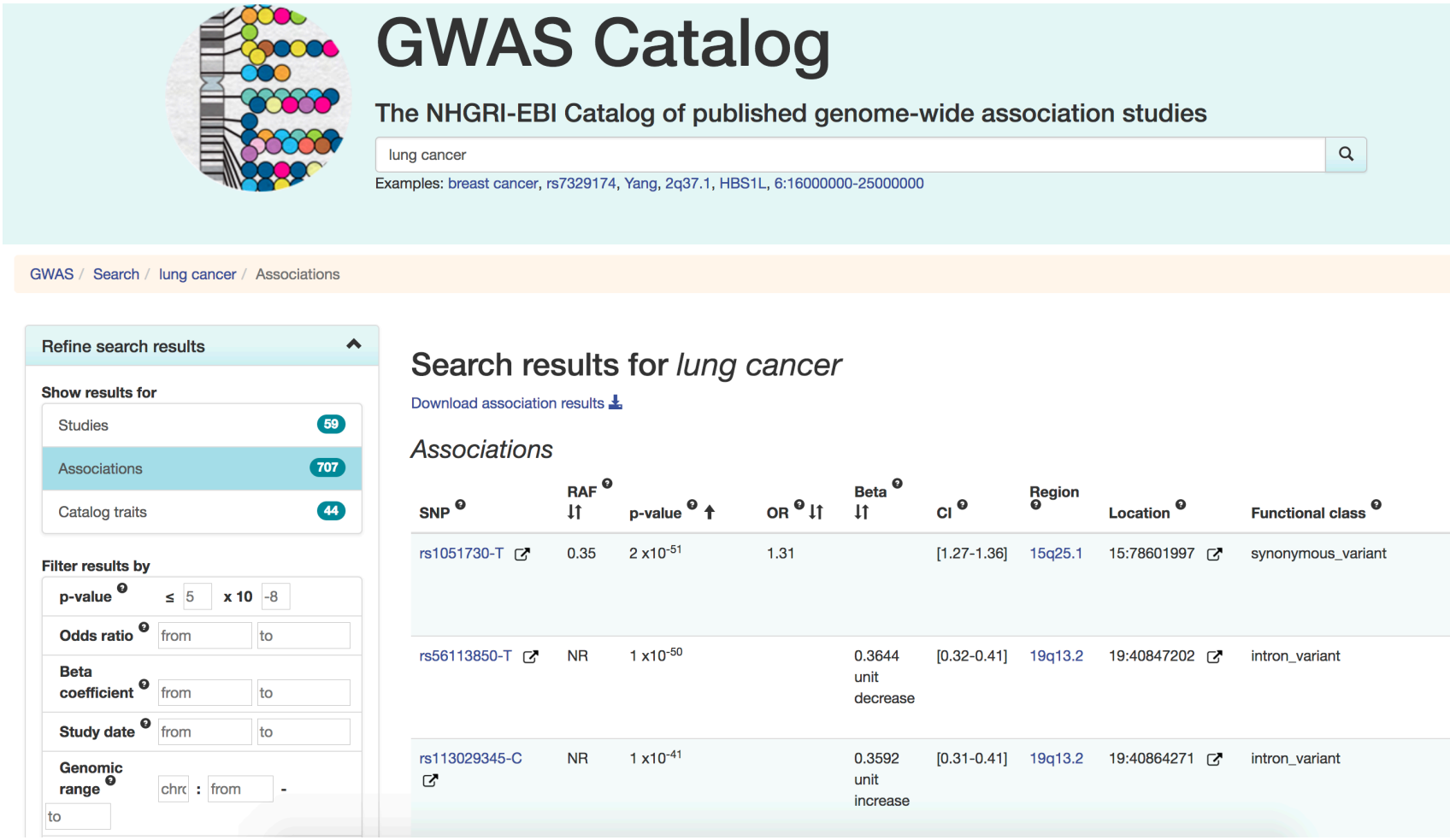
$P=5*10^{-8}$

Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nature (2014)

Type 2 Diabetes Results: 386,731 markers

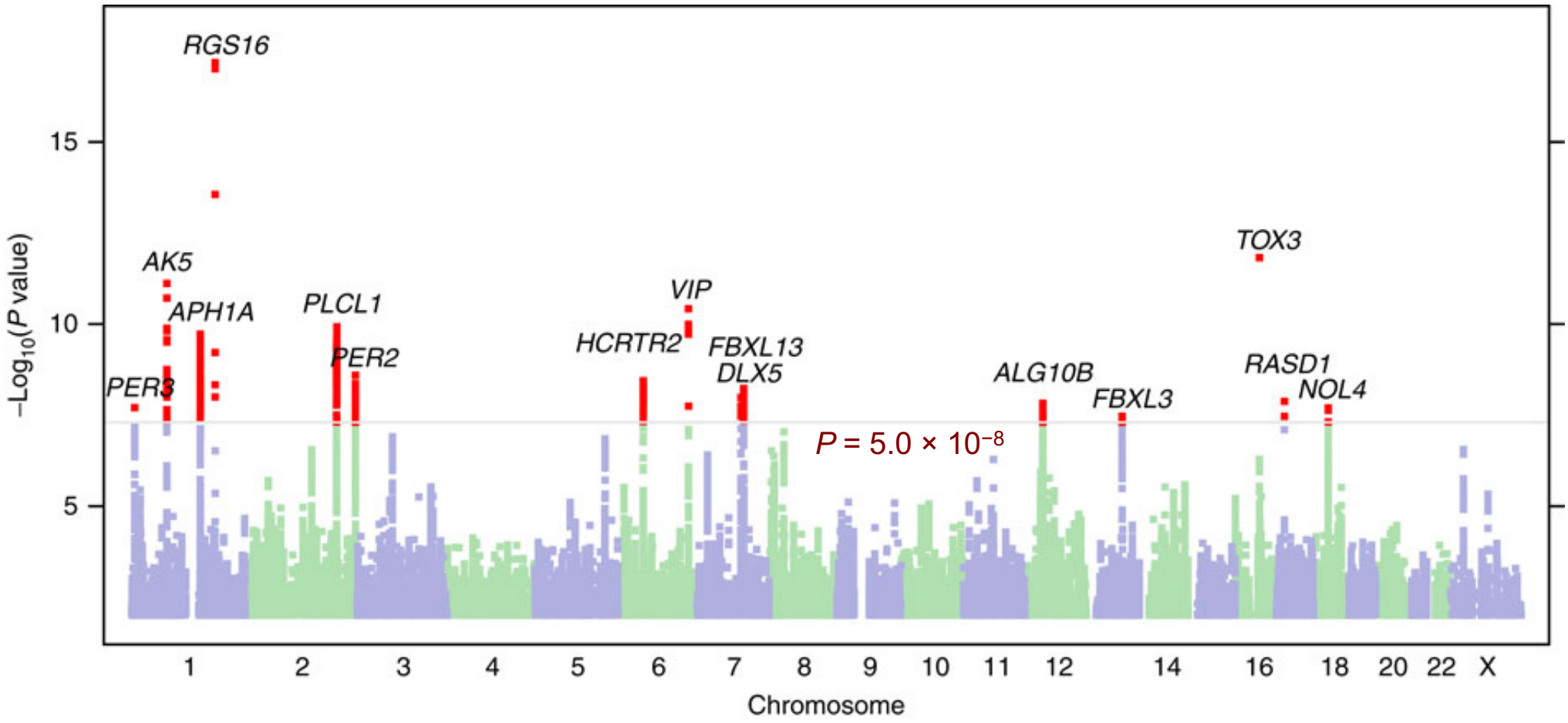


Type 2 diabetes association P values by chromosome (386,731 markers). The x-axis is the genomic position by chromosome 1-22 and X (by color), and the y-axis is the negative base 10 logarithm of the P value.



- https://www.ebi.ac.uk/gwas/

Morning Person GWAS



Hu et al. *Nature Communications* 2016

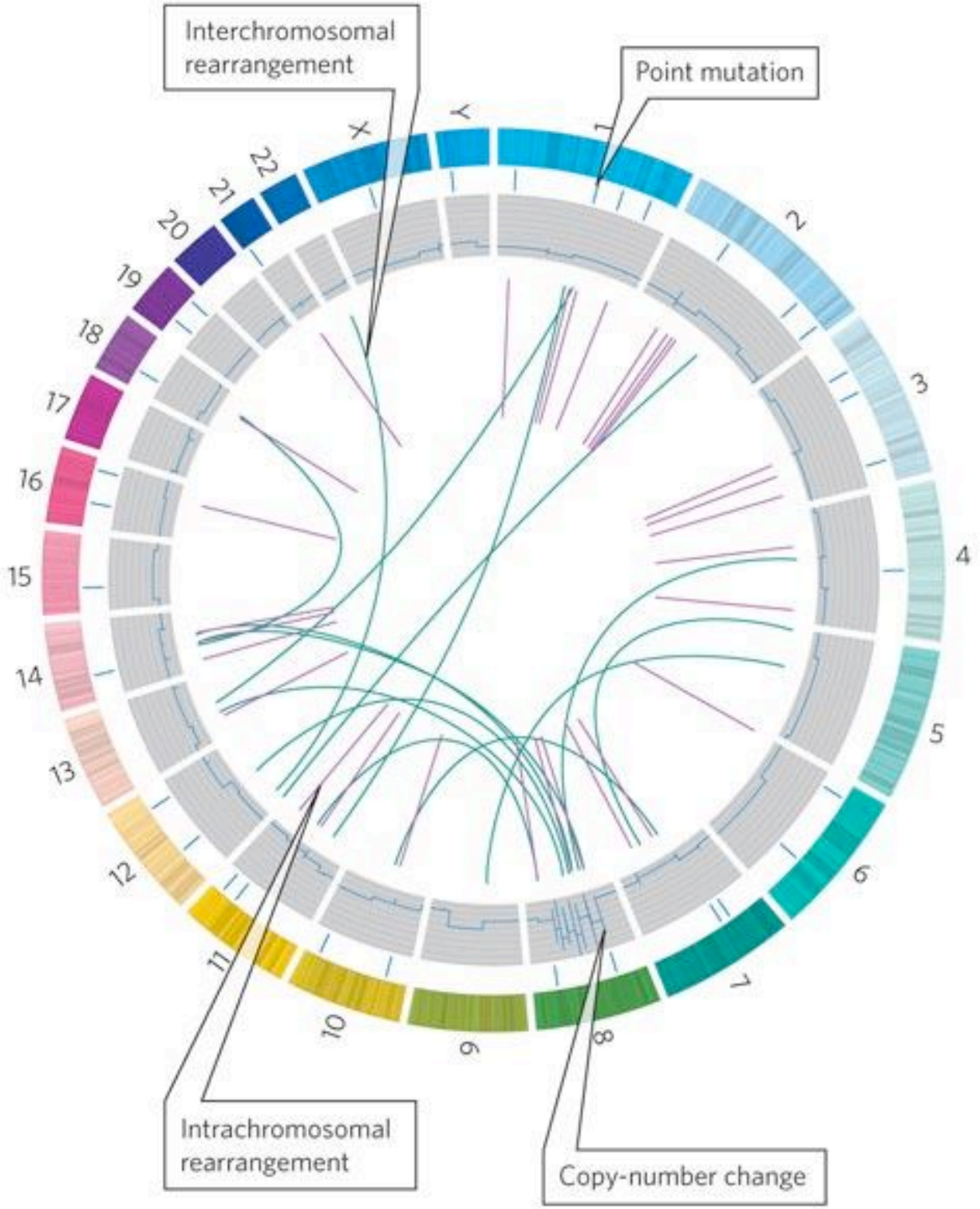
Understanding Associations Between Genetic Variation and Disease

International Cancer Genome Consortium

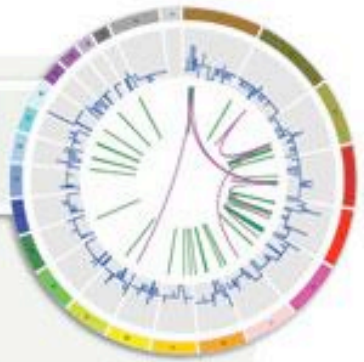
- Includes NIH's *The Cancer Genome Atlas*
- Sequencing DNA from 500 tumor samples for each of 50 different cancers

- Goal is to distinguish *drivers* (mutations that cause and accelerate cancers) from *passengers* (mutations that are byproducts of cancer's growth)

A Circos Plot

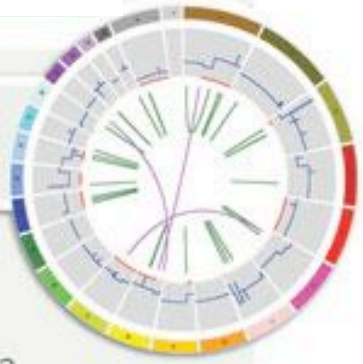


Some Cancer Genomes



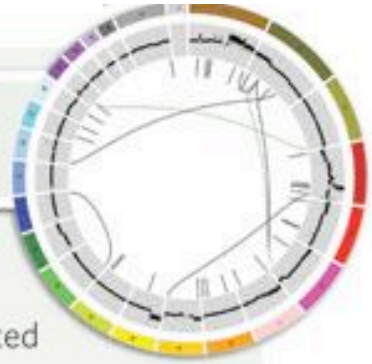
LUNG CANCER
Cancer: small-cell lung carcinoma

- Sequenced: full genome
- Source: NCI-H209 cell line
- Point mutations: 22,910
- Point mutations in gene regions: 134
- Genomic rearrangements: 58
- Copy-number changes: 334



SKIN CANCER
Cancer: metastatic melanoma

- Sequenced: full genome
- Source: COLO-829 cell line
- Point mutations: 33,345
- Point mutations in gene regions: 292
- Genomic rearrangements: 51
- Copy-number changes: 41



BREAST CANCER
Cancer: basal-like breast cancer

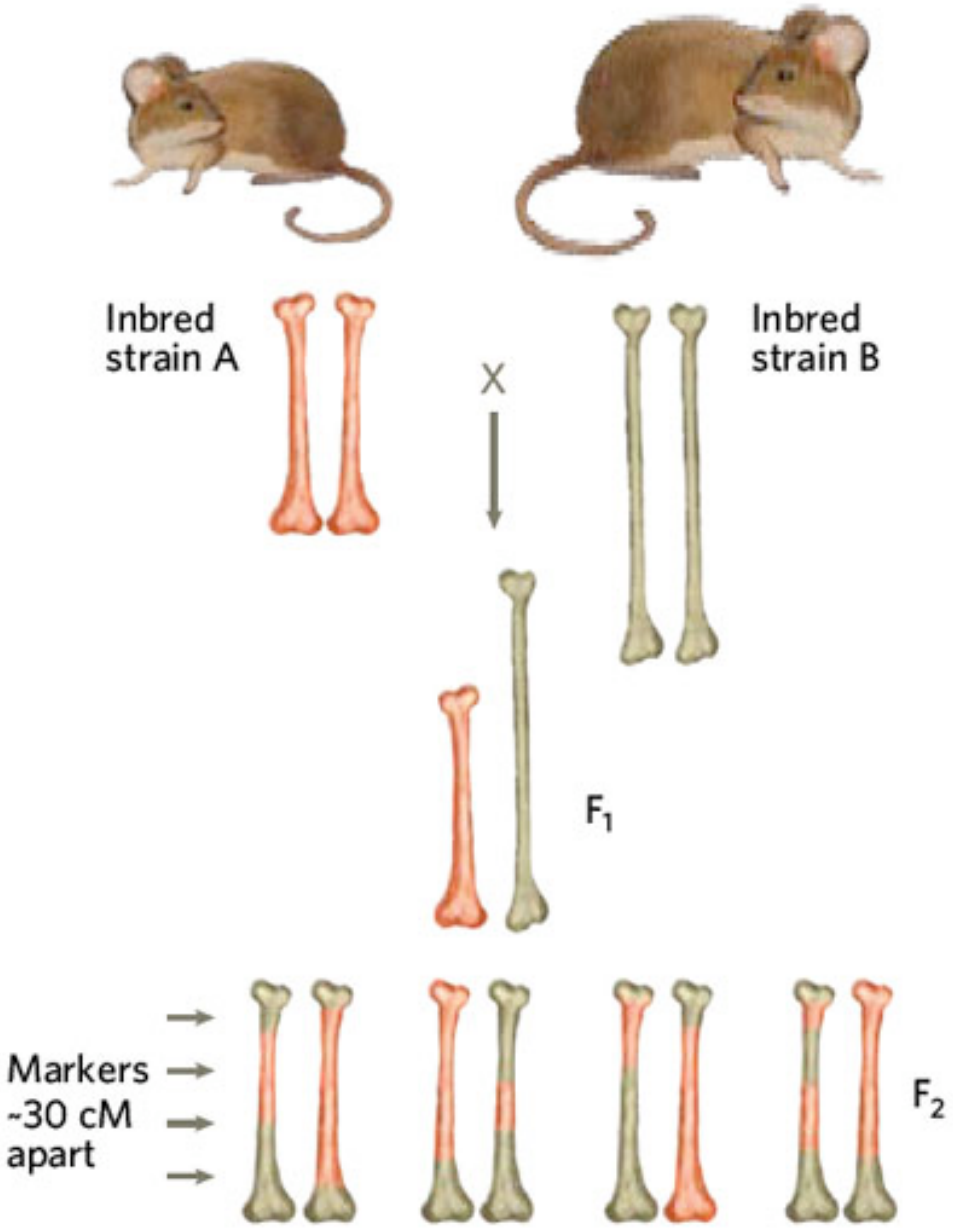
- Sequenced: full genome
- Source: primary tumour, brain metastasis, and tumours transplanted into mice
- Point mutations: 27,173 in primary, 51,710 in metastasis and 109,078 in transplant
- Point mutations in gene regions: 200 in primary, 225 in metastasis, 328 in transplant
- Genomic rearrangements: 34
- Copy-number changes: 155 in primary, 101 in metastasis, 97 in transplant

Understanding Associations Between Genetic Variation and Complex Phenotypes

Quantitative trait loci (QTL) mapping

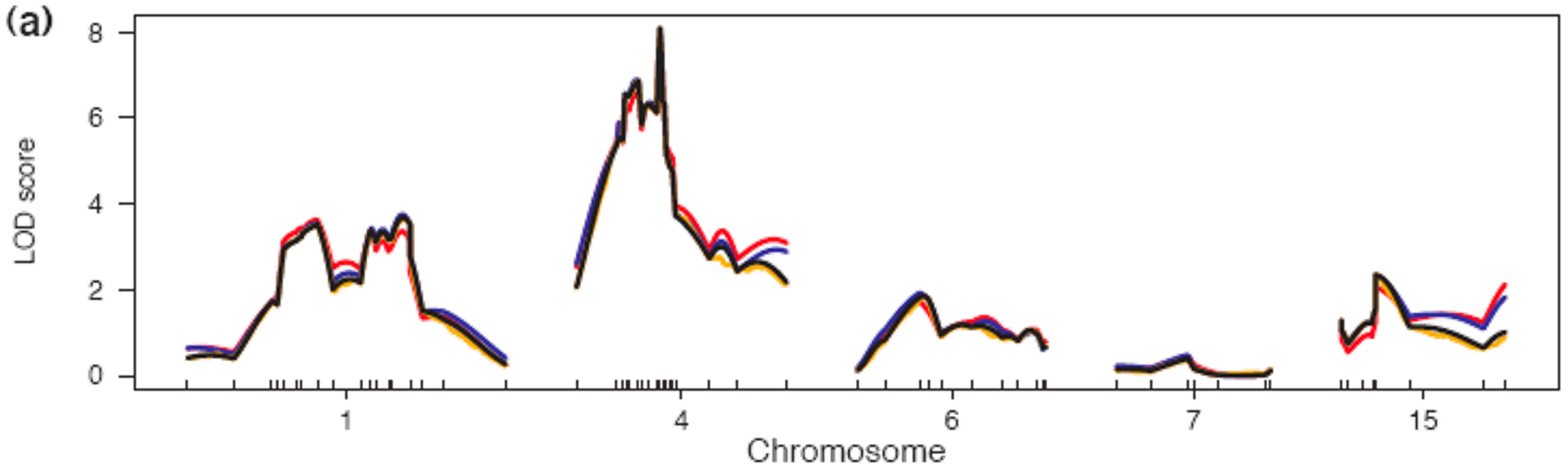
- Gather some population of individuals
- Genotype each individual at polymorphic markers
- Map quantitative trait(s) of interest to chromosomal locations that seem to explain variation in trait

QTL Mapping Example



QTL Mapping Example

QTL mapping of mouse blood pressure, heart rate [Sugiyama et al., Broman et al.]



Logarithm of Odds

$$\mathrm{LOD}(q) = \log_{10} \frac{P(q \mid \text{QTL at } m)}{P(q \mid \text{no QTL at } m)}$$

quantitative trait

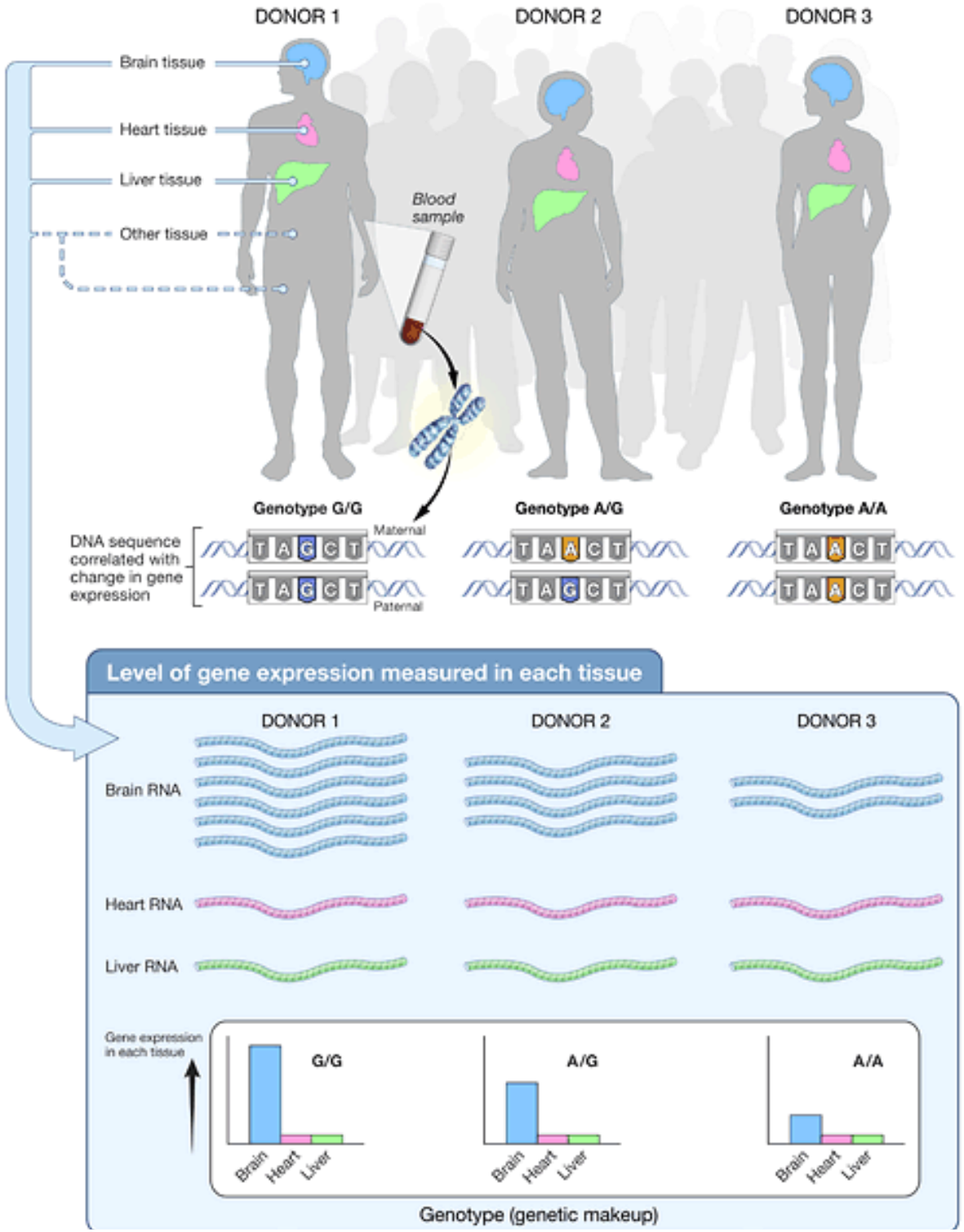
position in the genome

QTL Example: Genotype-Tissue Expression Project (GTEx)

- Expression QTL (eQTL): traits are expression levels of various genes

- Map genotype to gene expression in different human tissues

QTL Example: GTEx



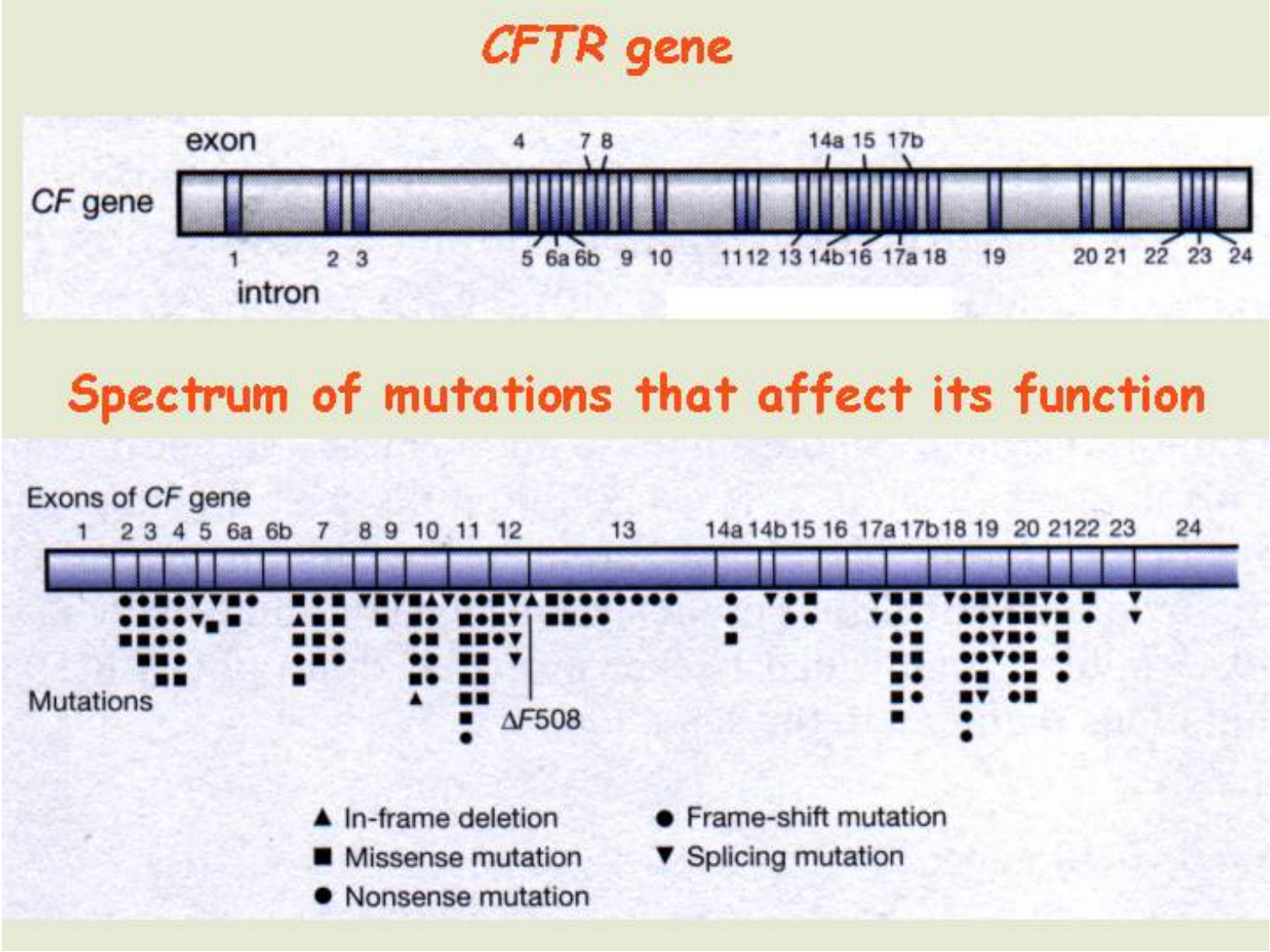
https://www.genome.gov/27543767/

GWAS Versus QTL

- Both associate genotype with phenotype
- GWAS pertains to discrete phenotypes
  - For example, disease status is binary
- QTL pertains to quantitative (continuous) phenotypes
  - Height
  - Gene expression
  - Splicing events
  - Metabolite abundance

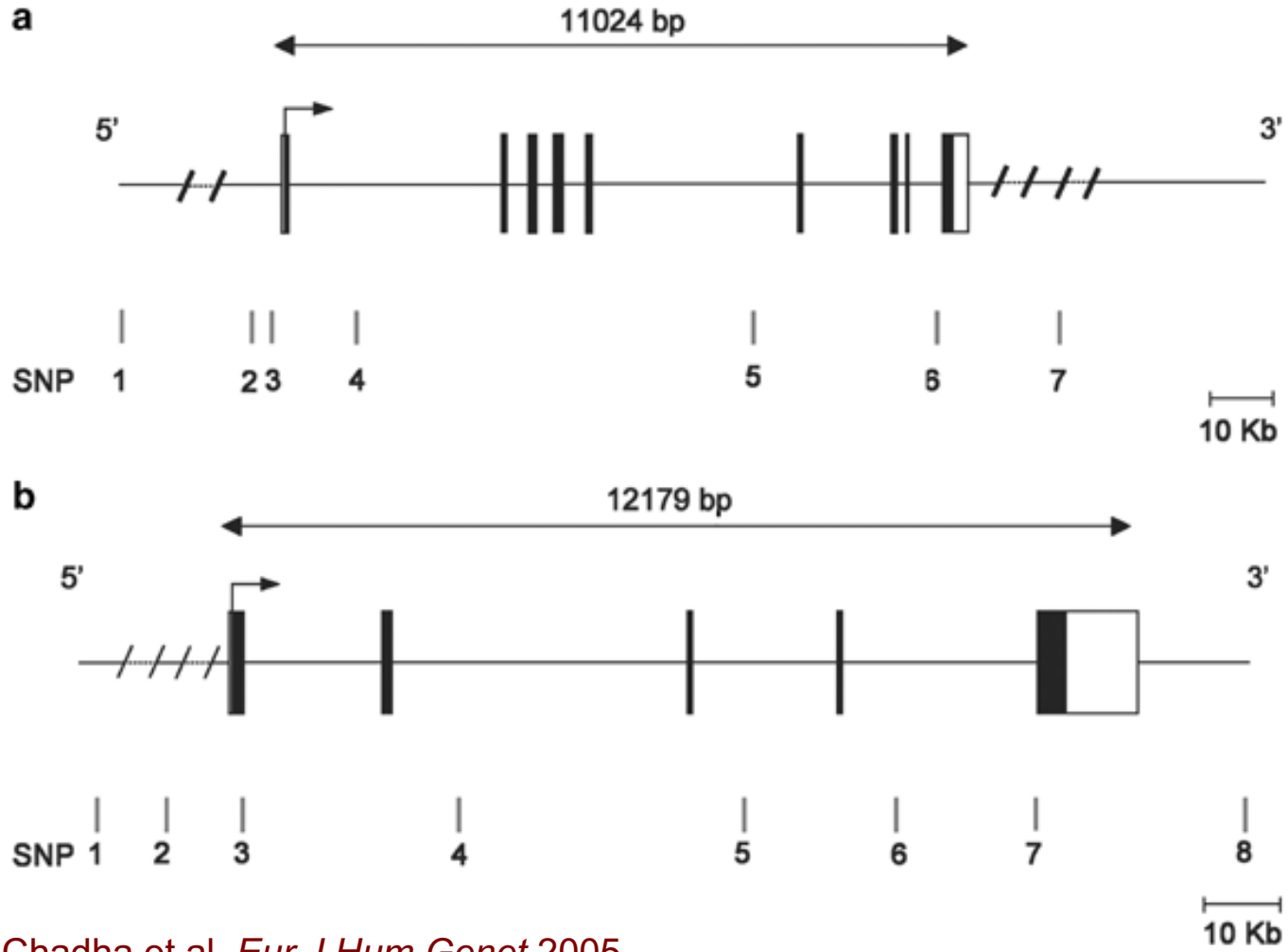
Determining Association is Not Enough

A simple case: CFTR (Cystic Fibrosis Transmembrane Conductance Regulator)



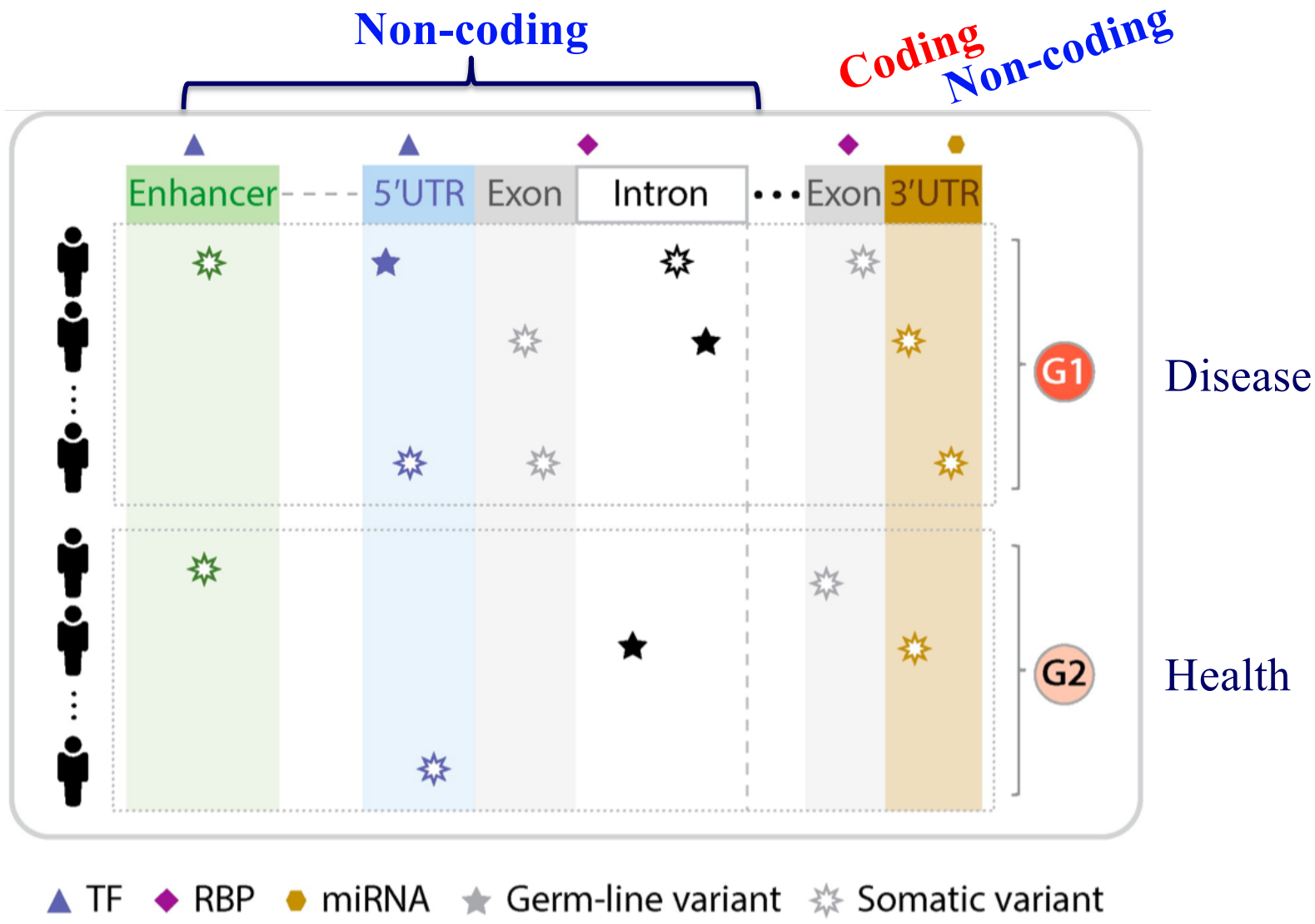
Many Measured SNPs Not in Coding Regions

- Genes encoding CD40 and CD40L with relative positions of the SNPs studied



Chadha et al. *Eur J Hum Genet* 2005

Non-coding variants



McGillivray P, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:153–80

Computational Problems

- Assembly and alignment of thousands of genomes
- Detecting large structural variants
- Data structures to capture extensive variation
- Identifying functional roles of markers of interest (which genes/pathways does a mutation affect and how?)
- Identifying interactions in multi-allelic diseases (which combinations of mutations lead to a disease state?)
- Identifying genetic/environmental interactions that lead to disease
- Inferring network models that exploit all sources of evidence: genotype, expression, metabolic, etc.