Genomics primer

Oana Ursu
(Kundaje and Snyder labs)

Outline

Genetics

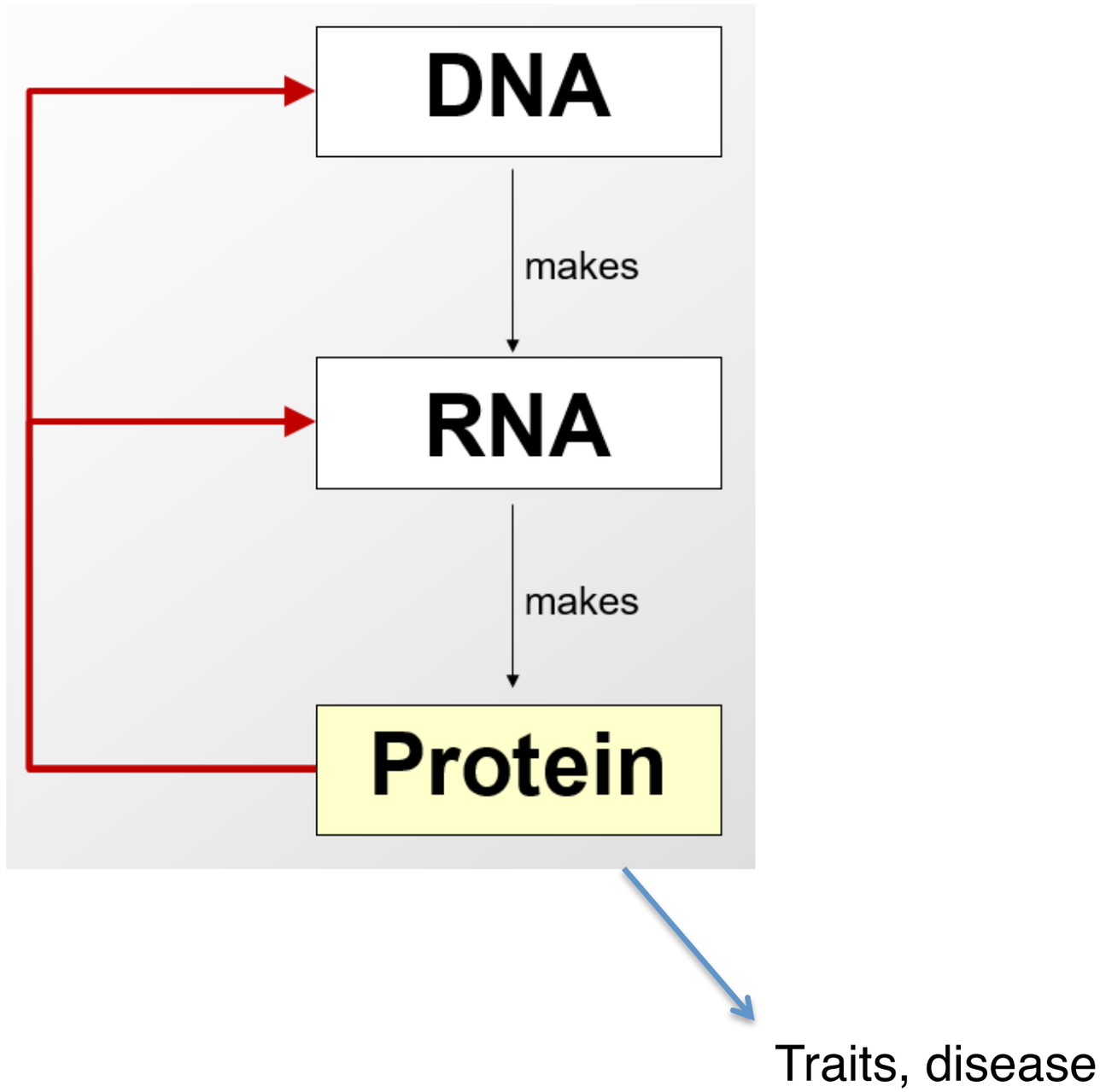
- Intro to genetics
- Diseases and Genome-wide Association Studies (GWAS)

Data types

- Data types and data sources

Genome browsers

Intro to Genetics



Genetic Variation

- Human Genome: 2 x 3 billion bp, 20K genes
- Diploid organism
- Consider genotype at a single locus:
  - Heterozygous: contains two different alleles
  - Homozygous: contains the same allele

Measuring Diversity

- pi: Average pairwise heterozygosity (per bp)
- Human average pi ~ 0.08%: a typical individual is heterozygous in 0.8 sites per kilobase
- What determines the amount of variation:
  - Mutation rate
  - Population size
  - Natural selection
- Genetic Drift: random changes in allele frequencies of neutral variants

Why Genetic Variation?

- Variation of the DNA sequences in our genome:
  - Understand biological processes and mutations
  - Plays a central role in human disease
  - Study human history
  - Natural selection and adaptation

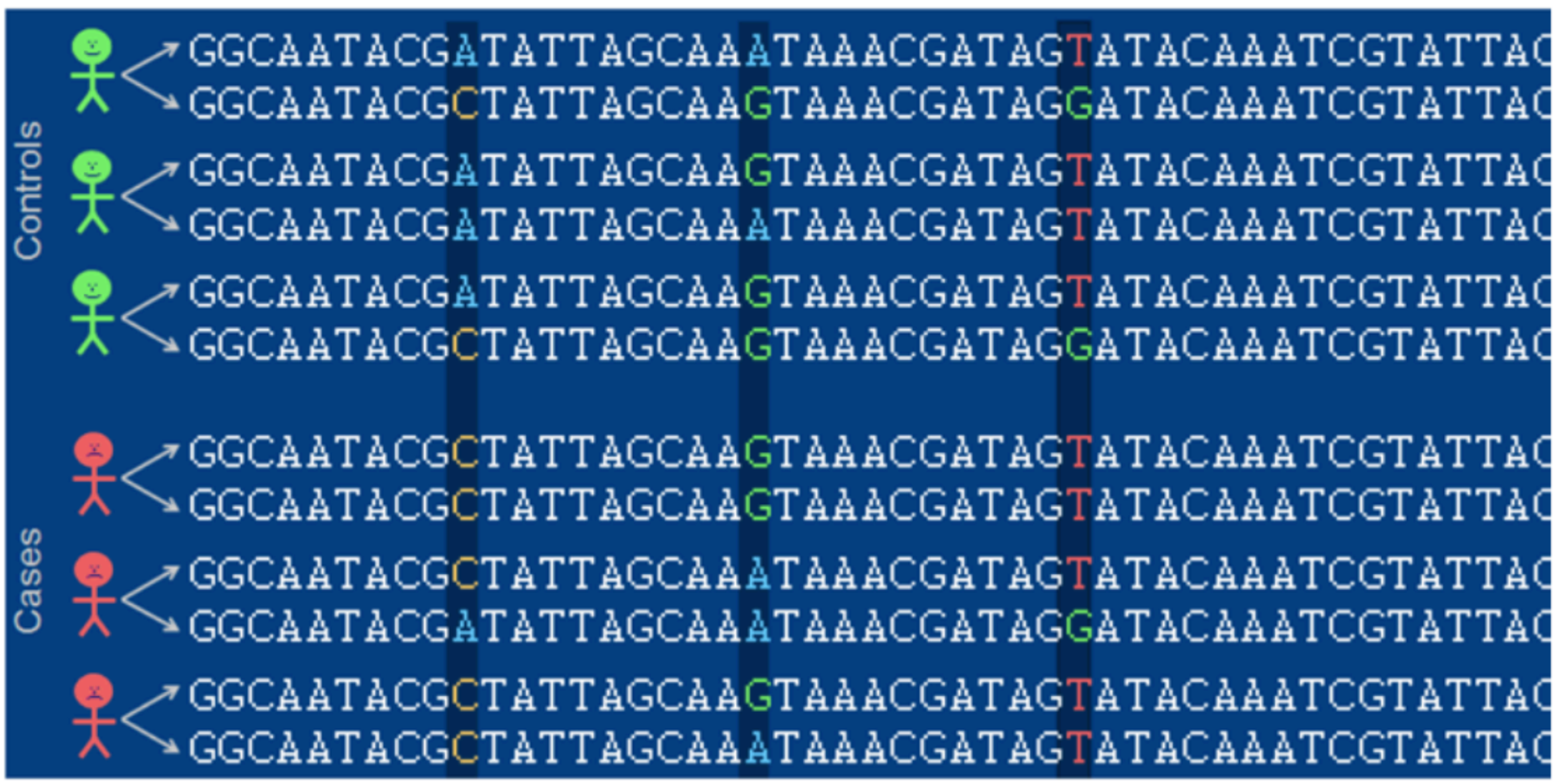
Types of genetic variation

- Mutations: alteration of the nucleotide sequence
  - Small scale:
    - Single Nucleotide Polymorphisms (SNPs):
      - 3x10^6 common SNPs (>5% frequency in human population)
      - Rare SNPs
    - Insertion/Deletion of a few nucleotides
  - Large scale:
    - Copy Number Variations (CNVs)
    - Insertions, Inversions and Translations

SNP = single nucleotide polymorphism

= Single base change

- Sample common SNPs



Minor allele frequency

Say you have a SNP, with possible alleles A and C.
The 3 possible genotypes you get are: AA, AC, CC,
at frequencies 0.5, 0.3 and 0.2

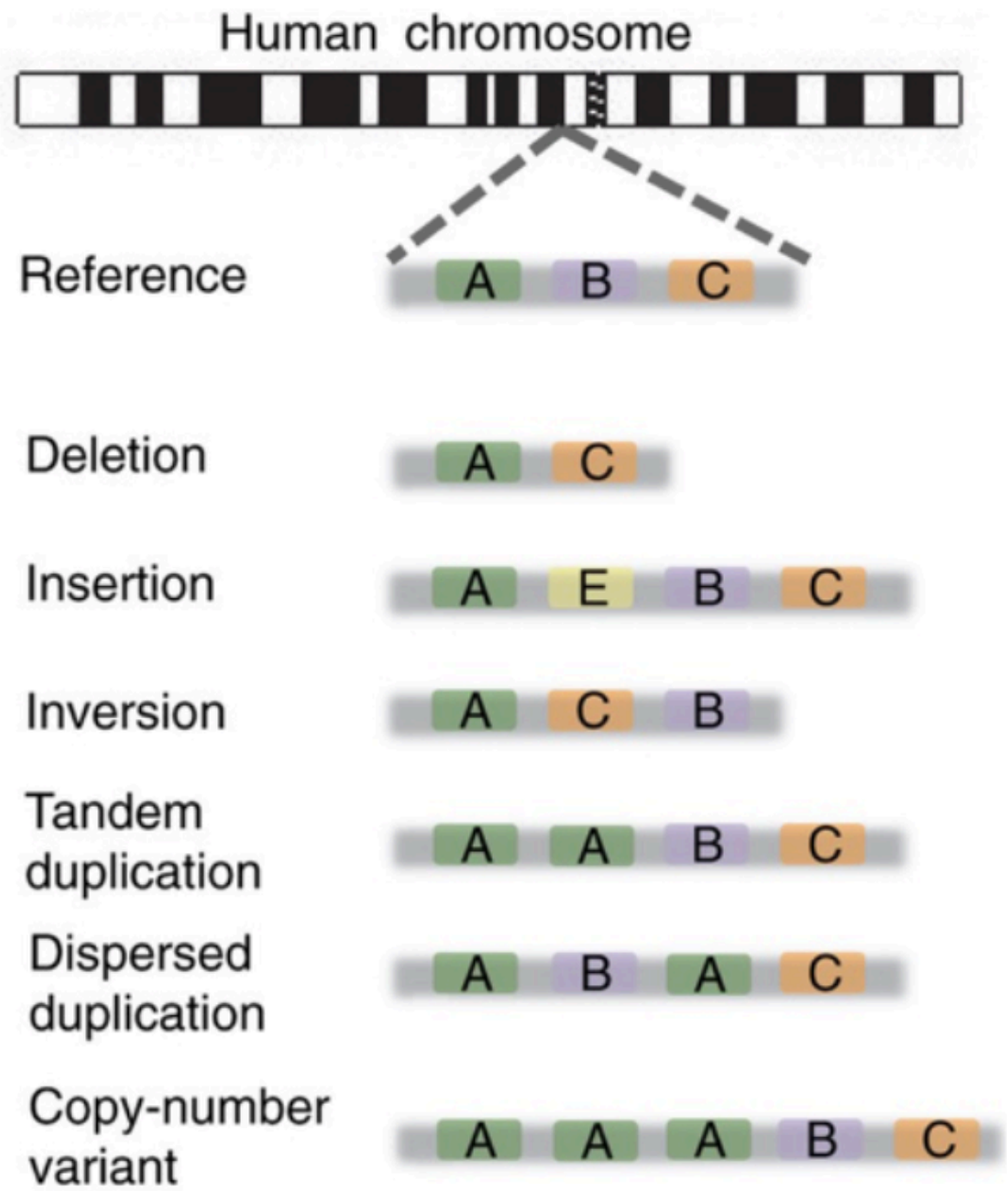
Thus, in the population, we have in total:
[0.5*2 + 0.3*1]/2 = 0.65 A
[0.3*1 + 0.2*2]/2 = 0.35 C

The minor allele frequency
= the proportion of the alleles that are C (the lowest of the 2 alleles)

This is a way to quantify the variation at the SNP.

Structural variation = big changes in the genome

- Structural Variations:
  - Insertion
  - Inversion
  - Translocation
  - CNVs:
    - Duplication
    - Deletion



How do we measure genetic variation?
Route 1: genome sequencing

Whole genome sequencing (WGS)
basic idea:
collect cells from person
sequence DNA => sequencing "reads"

Coming of age: ten years of next-generation sequencing technologies

Sara Goodwin, John D. McPherson & W. Richard McCombie

Affiliations | Corresponding author

Nature Reviews Genetics **17**, 333–351 (2016) | doi:10.1038/nrg.2016.49
Published online 17 May 2016

Genome assembly
Put reads together to obtain a complete genome
(requires deep sequencing)
- many existing methods for this

Align reads to the reference genome=high quality genome
(from Human Genome Project)

Structural variants

Computationally identify genetic variants
Copy number variation

Exome Sequencing

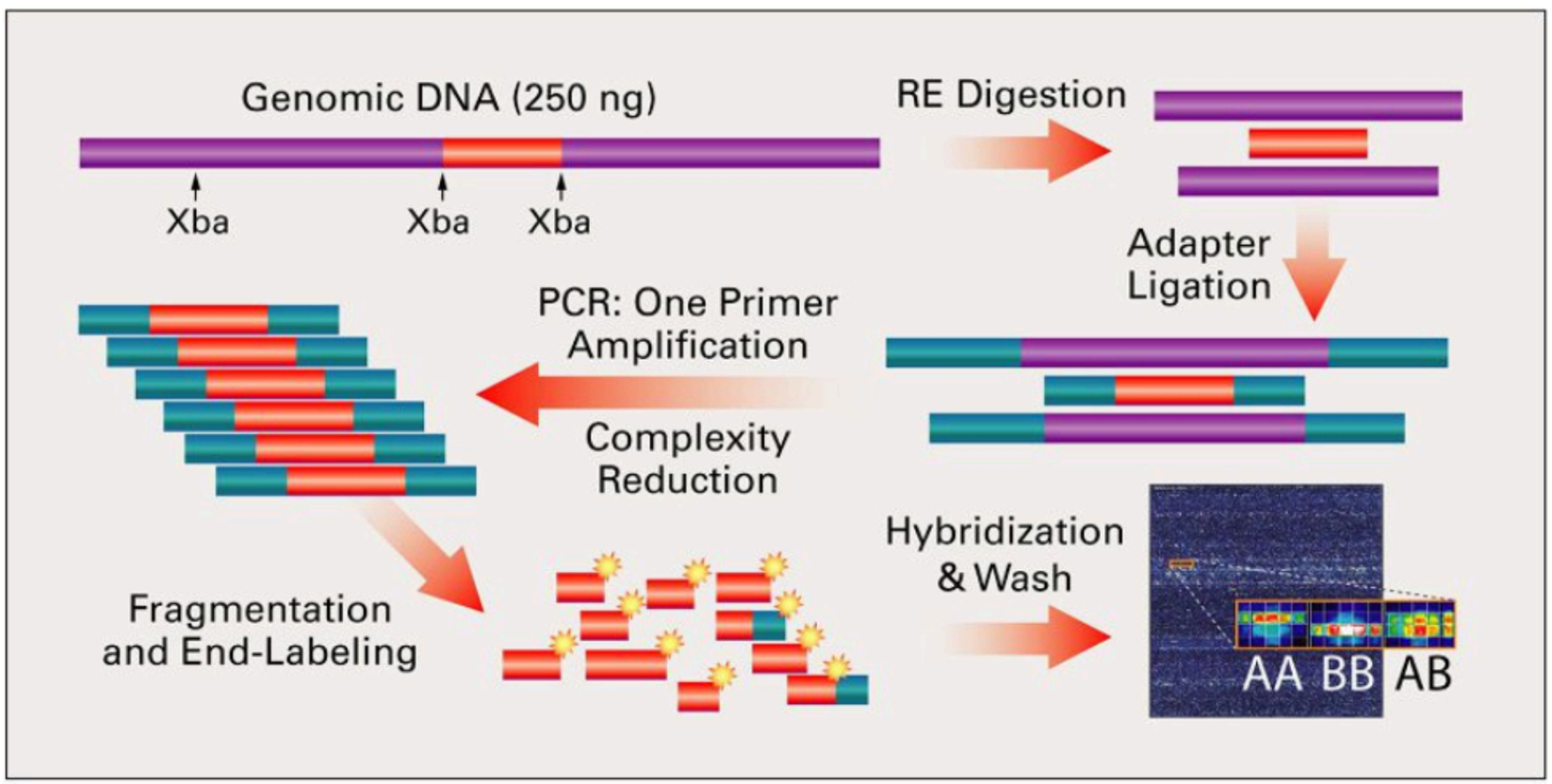
- Exon: segment of DNA containing information coding for a protein
- Exome consists of the exons of all our genes
- WES is technique for sequencing all expressed genes: capture exon after amplification
  - micro-array based capture using cDNA library
  - hybridization based using cDNA that only bind to exons
  - molecular inversion probe
- Goal: avoid high cost of whole genome sequencing

How do we measure genetic variation?
Route 2: measuring specific genetic variants

Idea: Instead of sequencing the whole genome, let's focus on only regions with SNPs.

Example: 23andme

The SNP array



Affymetrix.com

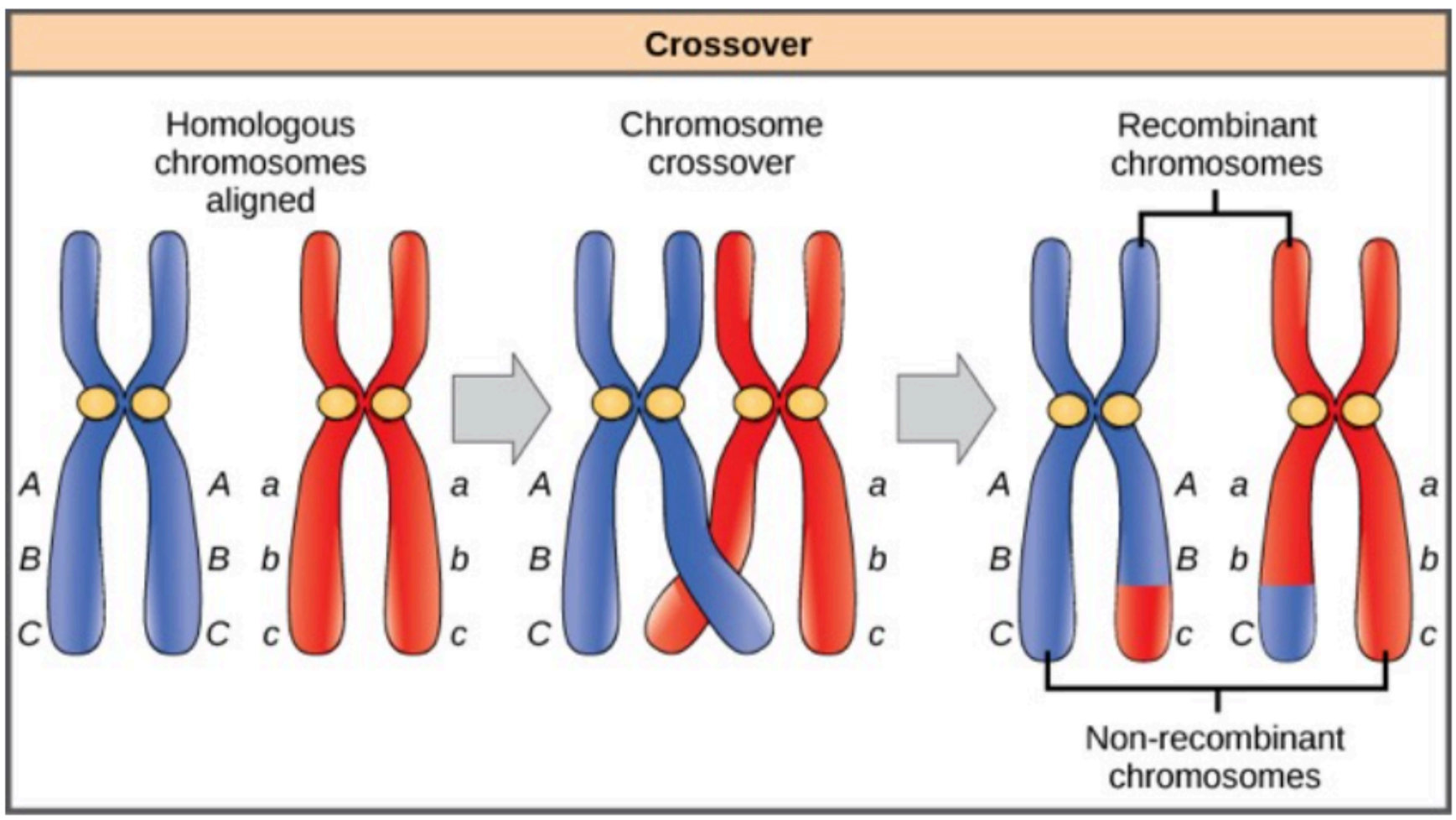
But there are 3M common SNPs! How come 23andme only measures 600k of them?

Linkage disequilibrium!
Turns out we can measure a subset of SNPs and (imperfectly) infer the rest
because there is high correlation between neighboring genotypes

Recombination

- Production of offspring with combinations of traits that differ than those found in the parents

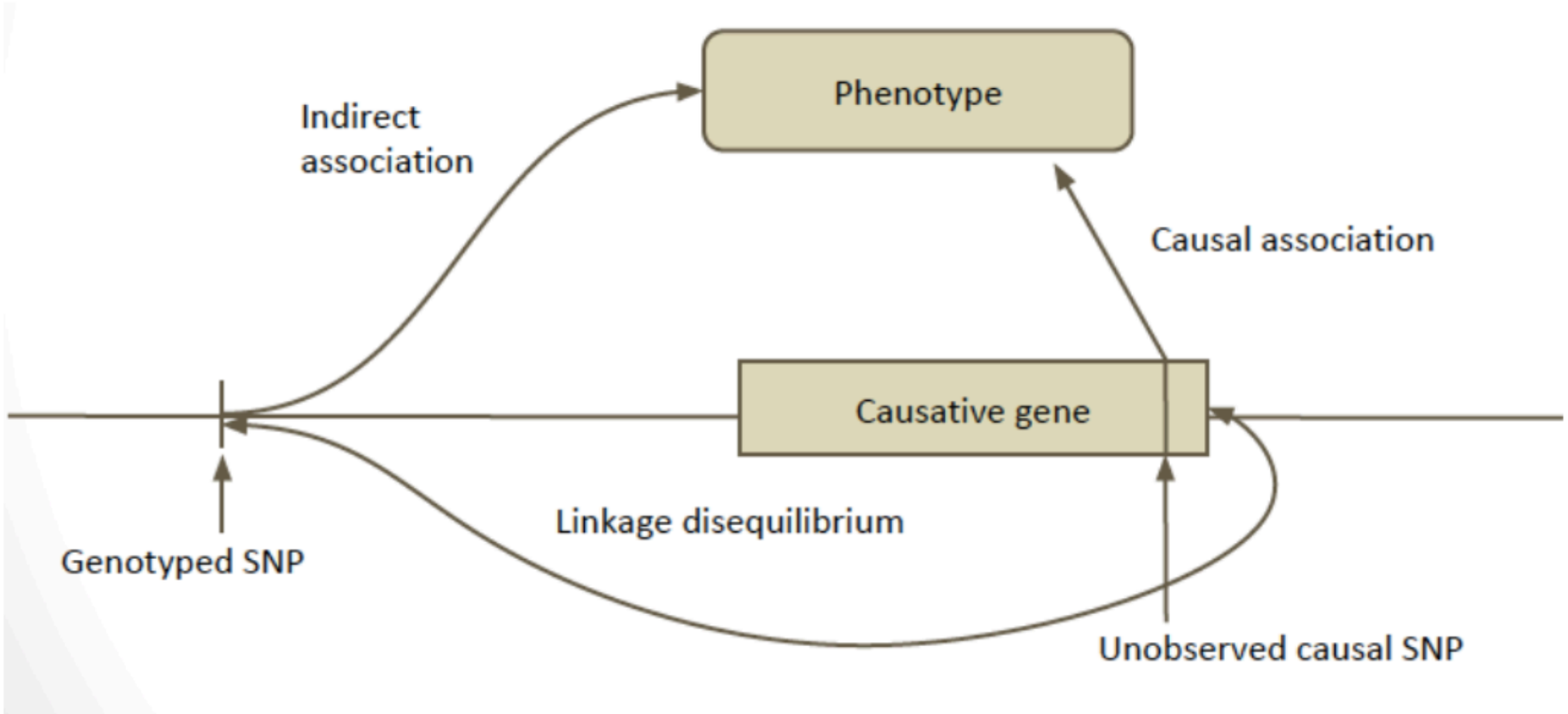
- Genes that typically stay together during recombination are said to be linked



Linkage Disequilibrium

- Alleles are in LE if the frequency of a genotype is equal to the product of the frequencies of the individual alleles
- Measure deviation from LE by comparing the observed and expected genotype frequencies
- Haplotype: set of SNPs that always happen together
- In a short chromosome segment there are only a few distinct haplotypes

Linkage Disequilibrium



- Neighboring markers tend to be inherited together
- Genotypes are redundant because LD causes correlations between the markers

Linkage Disequilibrium

- Basic descriptors:
  - Haplotype frequency of each type of chromosome
  - Common summary measures
    - D
    - D’
    - r^2

LD Measures

- Disequilibrium coefficient: D_AB=p_AB-p_Ap_B
- D_AB is hard to interpret
  - arbitrary sign
  - range depends on allele frequencies

		LocusB		Totals
		B	b	
Locus A	A	p_{AB}	p_{Ab}	p_A
	a	p_{aB}	p_{ab}	p_a
Totals		p_B	p_b	1

LD Measures

$$D'_{AB} = \begin{cases} \dfrac{D_{AB}}{\min(p_A p_B, p_a p_b)} & D_{AB} < 0 \\ \dfrac{D_{AB}}{\min(p_A p_b, p_a p_B)} & D_{AB} > 0 \end{cases}$$

- D'_AB in [-1,1]
- D' = +1/-1 means there is no recombination evidence
- High D' means markers are good surrogates for each other
- Disadvantages: Inflated if sample is small or one allele is rare

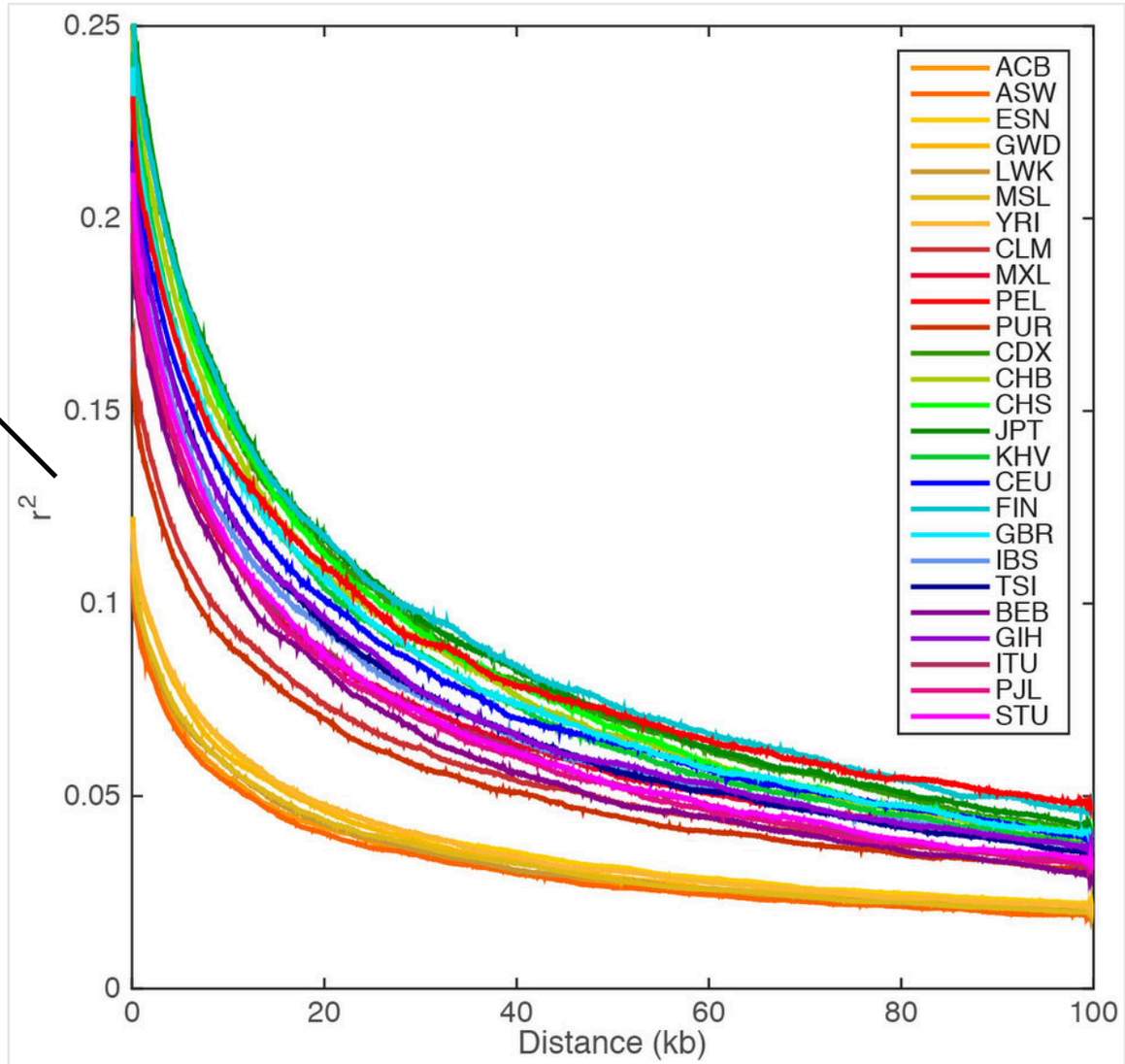
LD Measures

$$r^2 = \frac{(D_{AB})^2}{p_A(1-p_A)p_B(1-p_B)}$$

- Ranges in [0,1]: 0 if in equilibrium, 1 if identical information
- Measures loss in efficiency when marker A is replaced by marker B in an association study

Correlation between genotypes decreases with genomic distance

Measure of linkage disequilibrium



A global reference for human genetic variation

The 1000 Genomes Project Consortium

Nature **526**, 68–74 (01 October 2015) | doi:10.1038/nature15393

Why should you care about recombination and linkage disequilibrium?

SNP: rs2476601 (A or G) on PTPN22 (chromosome 1)

	AA	Moderately higher odds of developing hypothyroidism.
	AG	Slightly higher odds of developing hypothyroidism.
Greg Mendel	**GG**	**Typical odds of developing hypothyroidism.**

You're seeing your data because you're currently logged in to 23andMe. Don't worry — your results aren't visible to anyone else!

This content has not been approved by 23andMe and is for informational purposes only.

Although this SNP correlates with hypothyroidism, we cannot be certain that is causally relatived to this condition. All SNPs in LD with this SNP could contribute to hypothyroidism.

=> to distinguish between the different SNPs, you would need to do an experiment where you selectively mutate each SNP separately and see which one is related to the trait.

Machine learning problem:
Predict which SNP is the causal one for a given trait. For instance, causal SNPs might occur more frequently in regulatory elements.

Association studies: a powerful idea in genetics

Say you observe a strong correlation between a SNP and a trait, e.g. height

This could be because:
- the SNP affects height
- height affects the SNP
- something else affects both height and the SNP

Given that the DNA is (mostly) not changeable, it is likely that the SNP really affects the trait, in a causal way.

Association studies measure genotypes and some trait of interest in a population of people, and then identify which SNPs might affect the trait.

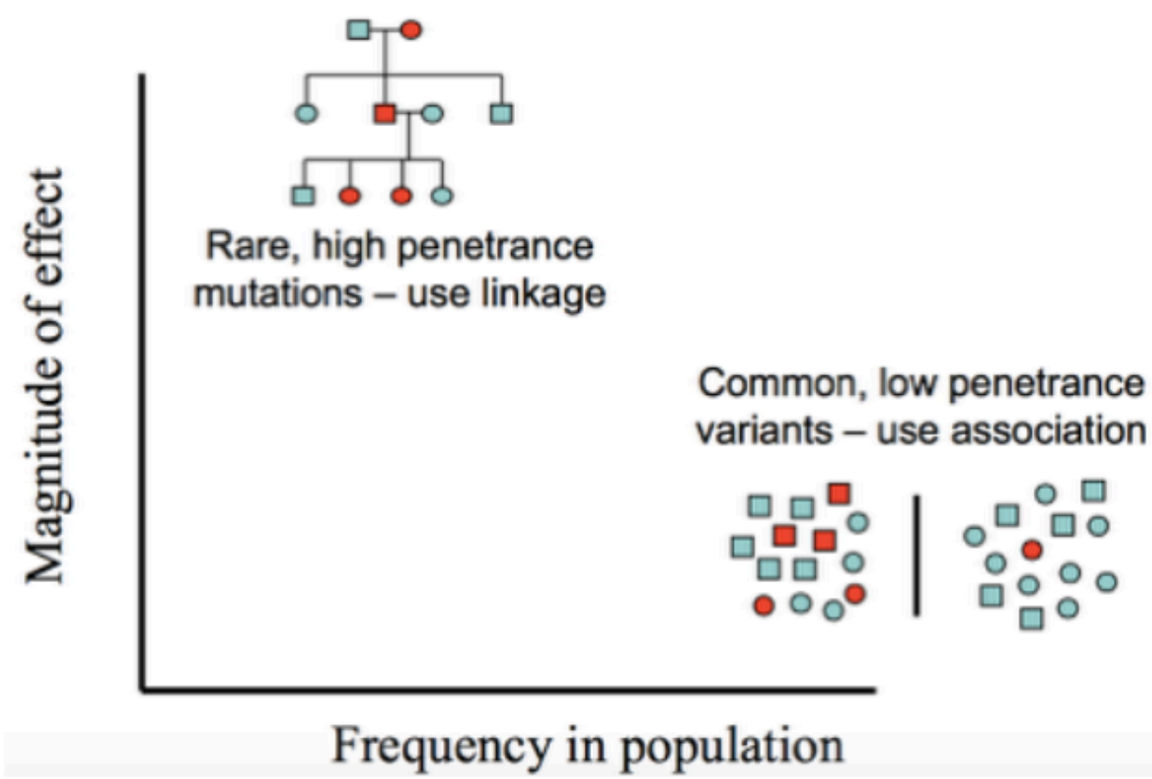
Such results serve as powerful hypotheses to be followed-up in the lab for understanding the mechanisms determining traits in general, and in particular disease.

Natural Disease Variation

- Interest in finding genetic factors underlying disease
  - personalized treatment
  - identify druggable targets
  - insight into biological pathways of disease
- Two main classes of diseases:
  - Mendelian: mutations in a single disease gene produce phenotype (eg cystic fibrosis)
  - Complex: multifactorial, many genes and environmental factors (eg diabetes)

Two types of studies

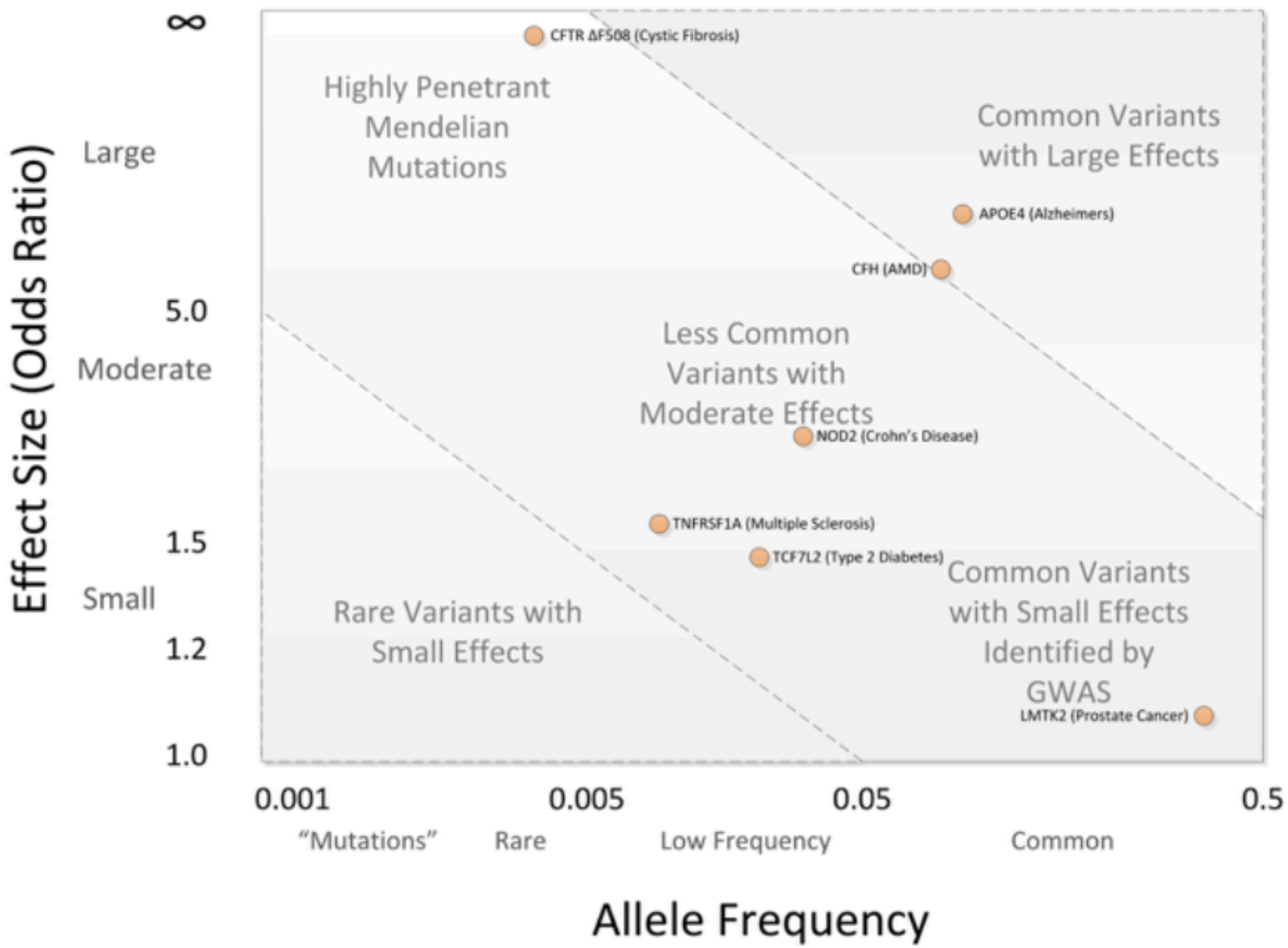
- Linkage: find markers that are transmitted with disease in families (powerful for Mendelian cases)
- Association: identify markers with frequency differences between cases and controls (more common)



Genome-wide association studies

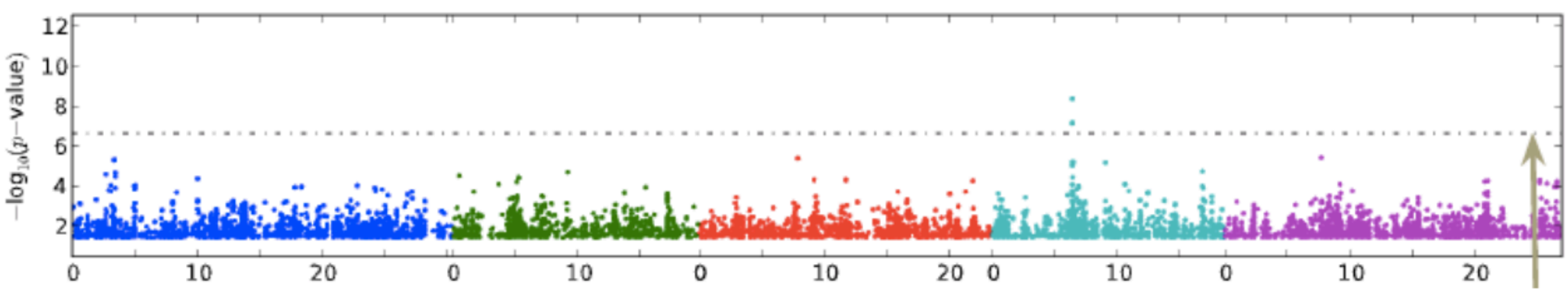
- Examine set of genetic variants in many individuals to associate them with a trait
- Identify large amounts of associations efficiently to understand genetics of diseases and traits
- Focused on associations between SNPs and diseases
- If one type of the variant is more frequent in people with the disease it is associated with the disease
- Use summary association statistics in conjunction with linkage disequilibrium

Disease Allele Effects



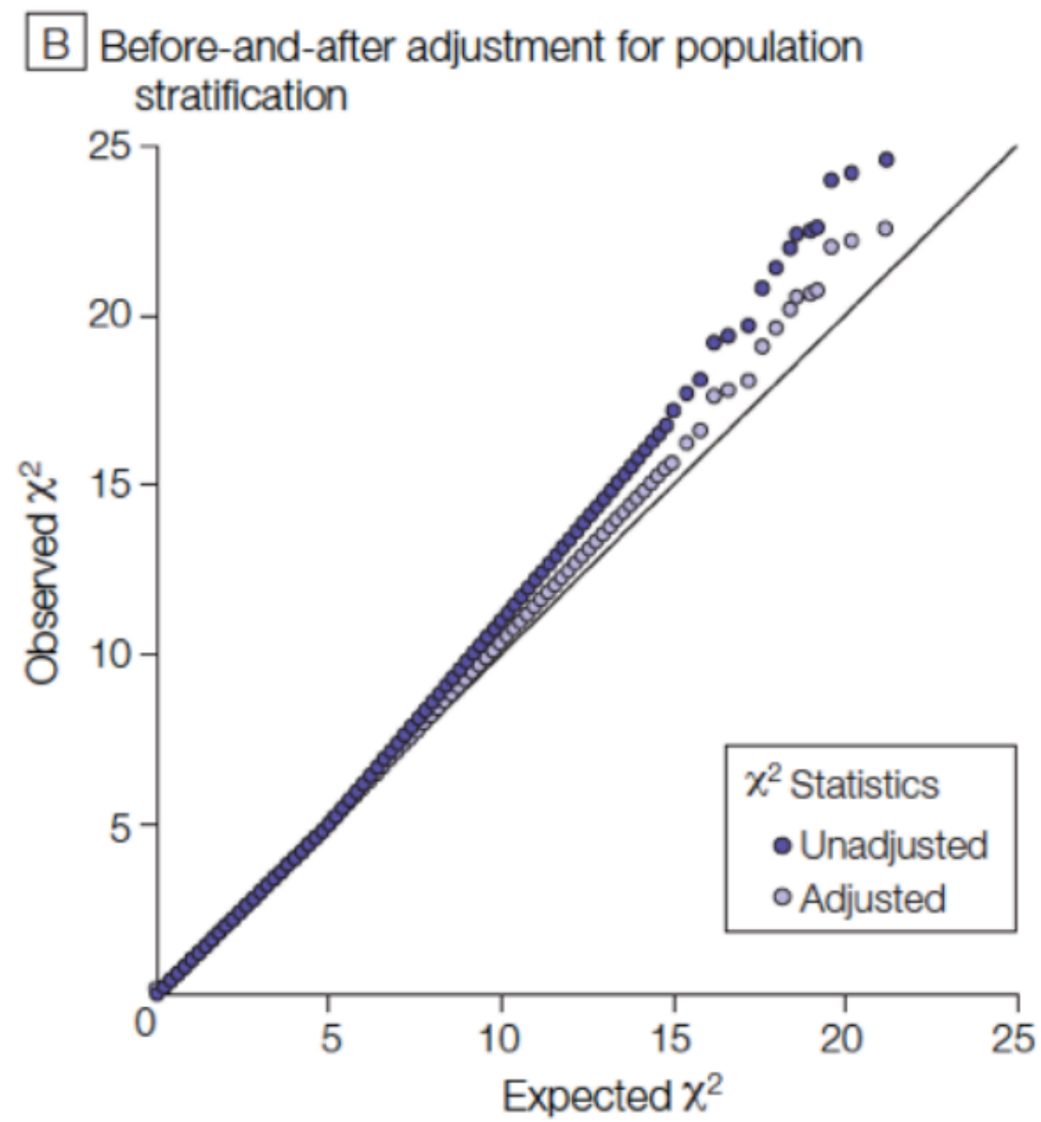
GWAS Example

- Manhattan plot:
  - x-axis: SNP locations
  - y-axis: -log(p-value)
- Multiple testing problem:
  - At 5% significance threshold, will expect 5% of markers that have true effect of 0 to be significant
  - Bonferroni correction



GWAS Confounders

- Careful about variables that could be different between the cases and controls other than the disease itself:
  - Population structure is the most common confounder. Example: marker for skin color might be associated with malaria resistance
  - Use QQ plot to show confounders aren't at work



So, you have a bunch of SNPs. What's next for understanding disease mechanisms?

Problems
- what is the relevant cell type for the disease?
- what is the downstream effect of the SNP?
  - TF binding change
  - downstream changes in gene expression
  - gene networks => disease

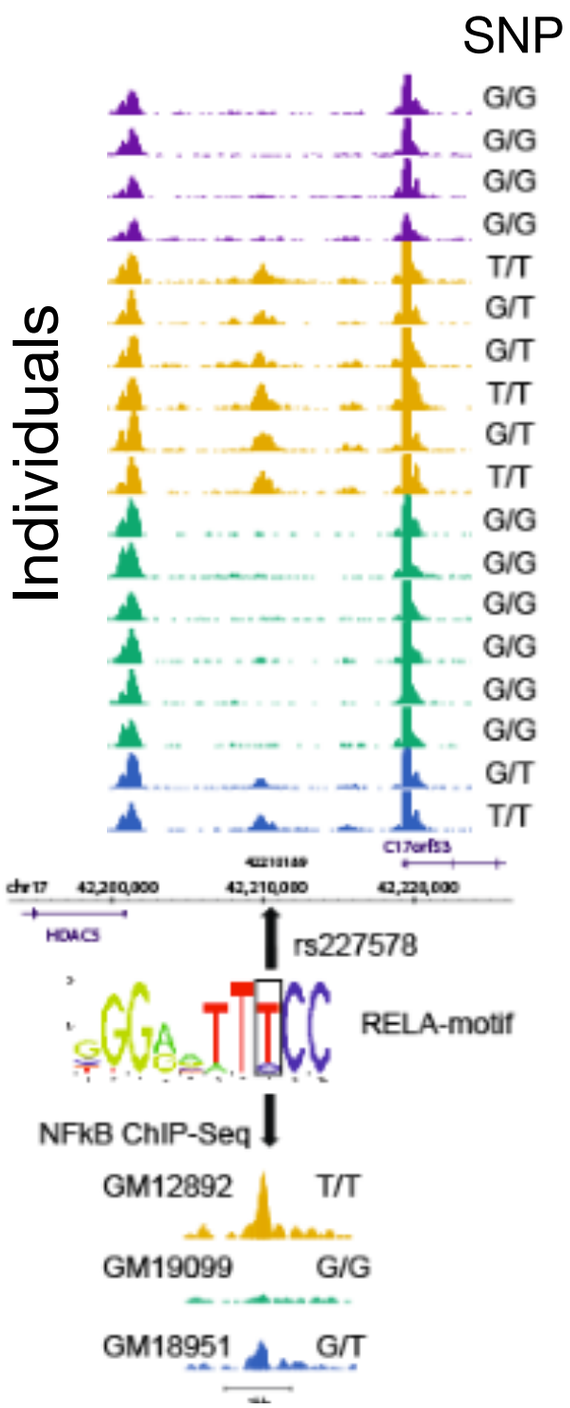
The fun part is that you can use association studies to understand the intermediate steps above.

For instance:
- eQTL = SNP correlates with gene expression
- binding QTL = SNP correlates with TF binding

etc.

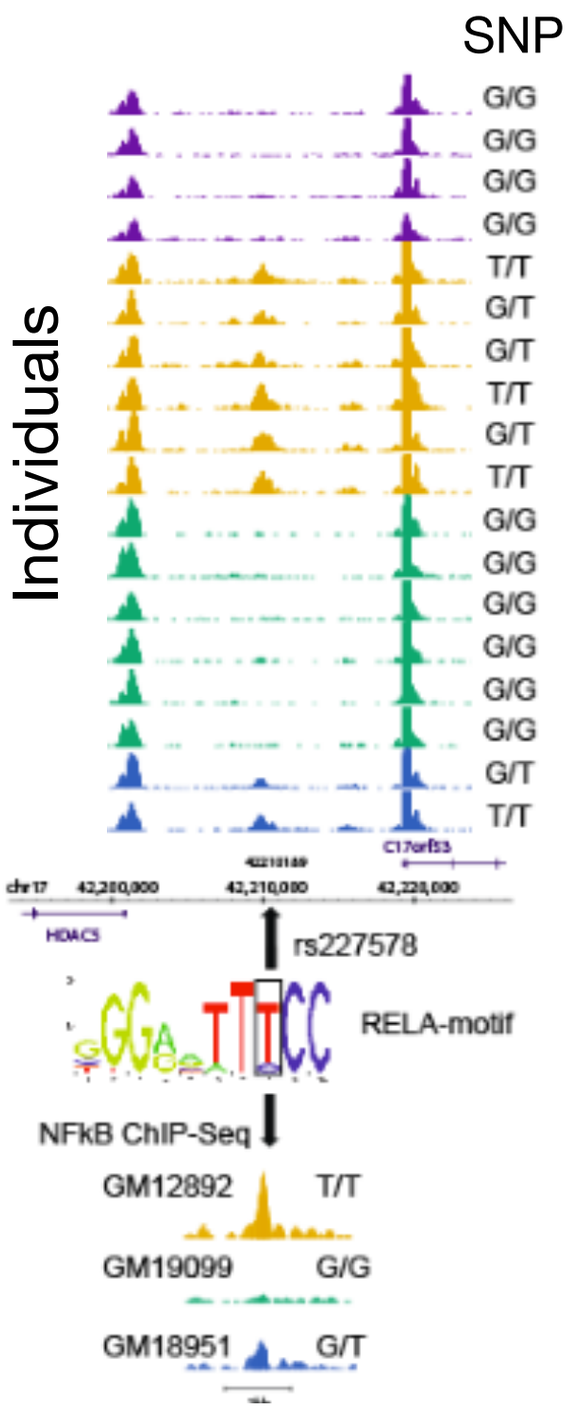
Some examples:
Grubert et al. 2015, Waszak et al. 2015, Kasowski et al., 2013, McVicker et al., 2013, Kilpinen et al, 2013, Degner, 2012

Studying gene regulation by measuring histone marks across individuals



Kasowski et al., 2013, McVicker et al., 2013, Kilpinen et al, 2013, Degner, 2012

Studying gene regulation by measuring histone marks across individuals

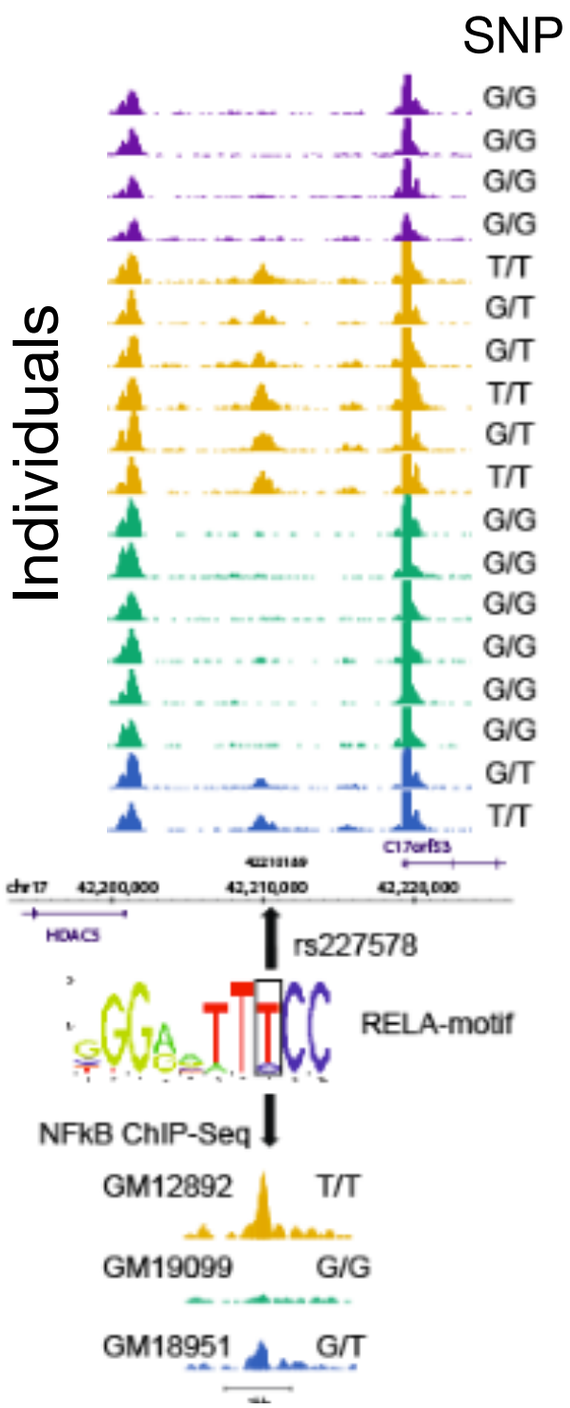


Hypothesis
SNP => TF binding => histone mark

Quantitative trait locus (QTL)
= SNP associated with histone mark

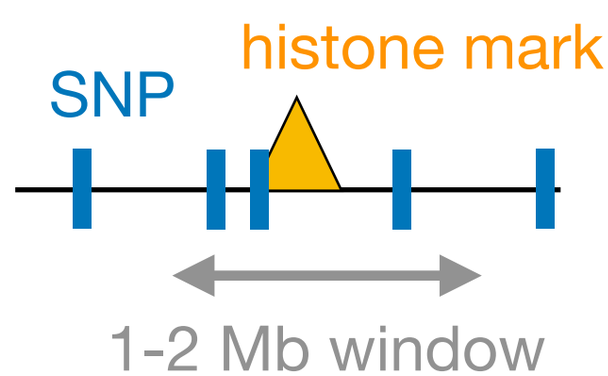
Kasowski et al., 2013, McVicker et al., 2013, Kilpinen et al, 2013, Degner, 2012

Studying gene regulation by measuring histone marks across individuals



Hypothesis
SNP => TF binding => histone mark

Quantitative trait locus (QTL)
= SNP associated with histone mark



Kasowski et al., 2013, McVicker et al., 2013, Kilpinen et al, 2013, Degner, 2012

Population structure is very important for 1) understanding disease and 2) personalized medicine

1) For understanding disease

- different populations have different variable SNPs and LD structures

=> results are not directly translatable between populations

2) For personalized medicine

Algorithms predicting treatment regimes trained on a specific population can perform sub-optimally when applied to a different population, due to different allele frequencies.

Genotype-Guided Dosing of Vitamin K Antagonists

Roxana Daneshjou, B.S., Teri E. Klein, Ph.D., and Russ B. Altman, M.D., Ph.D.

Author information ▶ Copyright and License information ▶

The publisher's final edited version of this article is available at N Engl J Med
See other articles in PMC that cite the published article.

TO THE EDITOR Go to:

Kimmel and colleagues note that a genotype-guided dosing algorithm using *CYP2C9*2, CYP2C9*3*, and *VKORC1* (–1639G→A) is statistically inferior to a clinical-dosing algorithm in patients of African descent. However, the single-nucleotide polymorphisms (SNPs) used in the study's pharmacogenetic dosing algorithm are known to occur at significantly lower frequencies in persons of African descent than in persons of European descent (Table 1).[1–4] The authors' ability to draw appropriate conclusions about the usefulness of genetics when determining dosages of warfarin for patients of African descent is thus very limited, and the benefits for this population have not been adequately tested. Physicians should not assume that self-reported race is an accurate proxy for the influence of genetic ancestry.[5] Rather, studies testing the usefulness of pharmacogenetics in a specific population should test variants with high frequency and measurable effect in that population.

Frequencies of Tested Warfarin Variants in Persons of European and African Descent.[*]

Variant	Genotype Frequencies		P Value[*]
	Persons of European Descent	Persons of African Descent	
*CYP2C9*2*[‡]	TT = 0.0175	TT = 0.00182	$P<2.2\times10^{-16}$
	TC = 0.227	TC = 0.0499	
	CC = 0.756	CC = 0.948	
*CYP2C9*3*[‡]	CC = 0.00349	CC = 0.000454	$P<2.2\times10^{-16}$
	CA = 0.125	CA = 0.0277	
	AA = 0.872	AA = 0.972	
VKORC1 (3673G→A)[‡]	AA = 0.195	AA = 0.061	$P = 1.564\times10^{-7}$
	GA = 0.407	GA = 0.082	
	GG = 0.398	GG = 0.857	

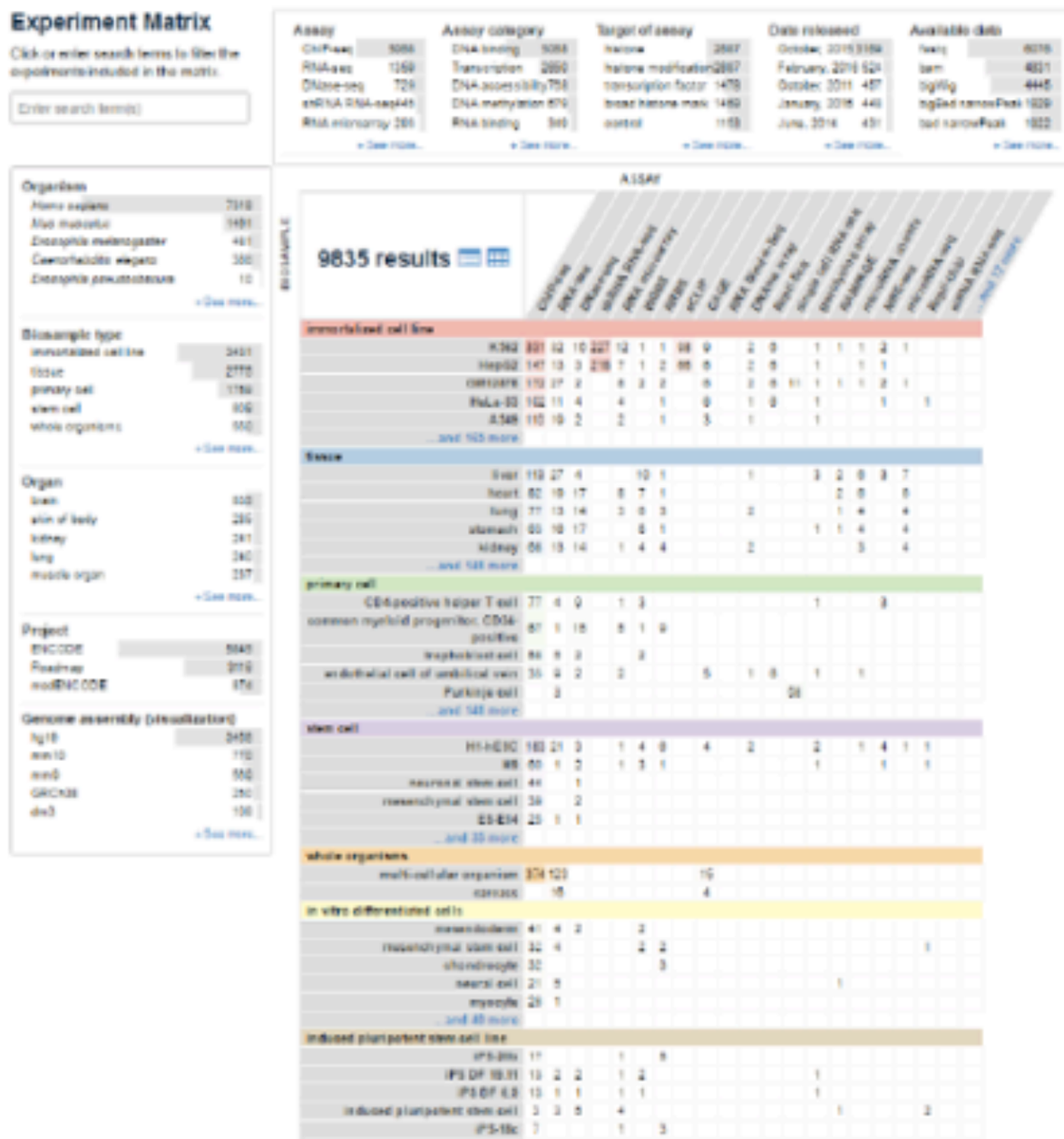
Data

Data Sources

- ENCODE
- Roadmap epigenomics project
- Cancer Genome Atlas
- 1000 Genomes project
- Data Deluge

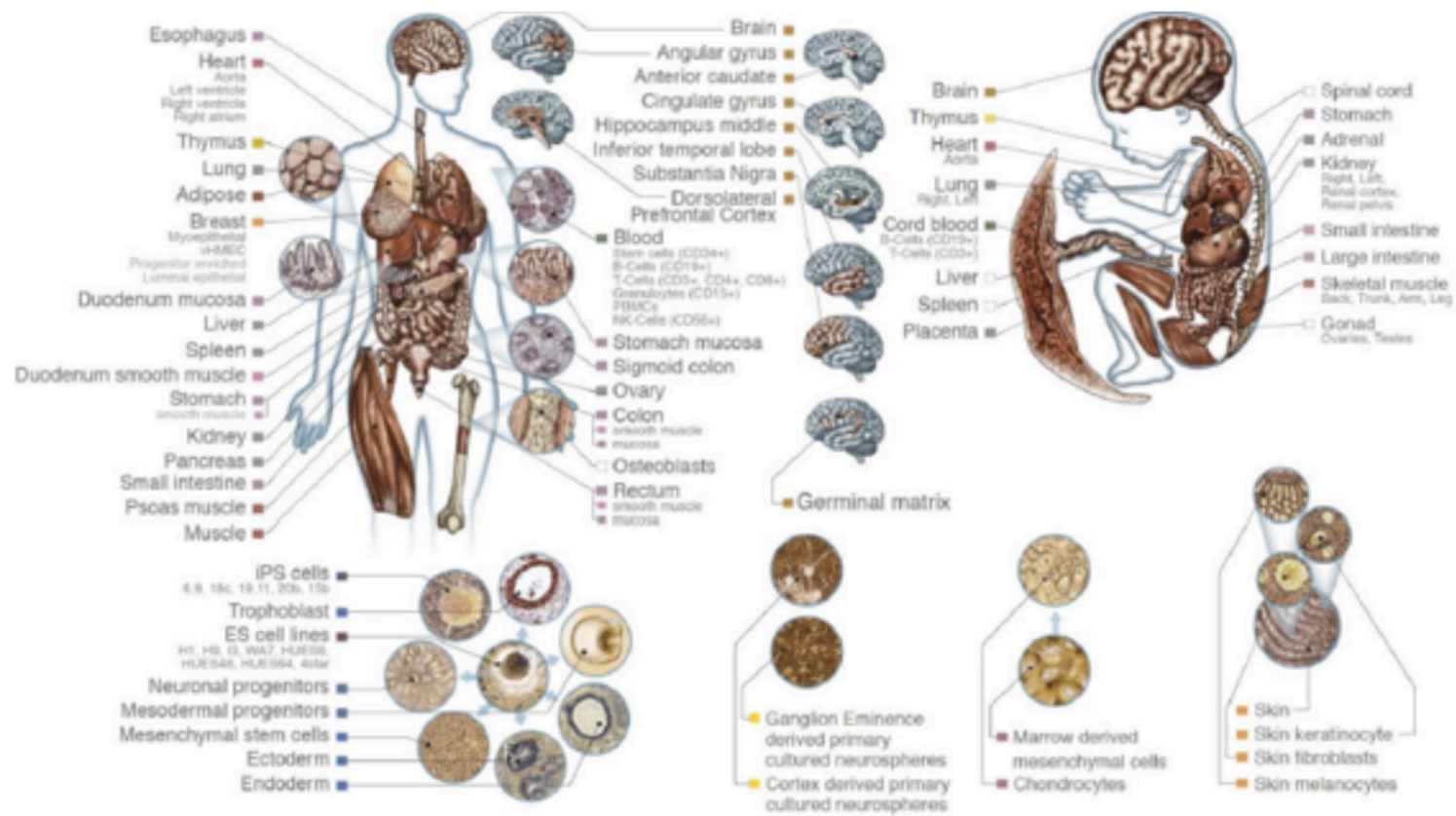
Encyclopedia of DNA elements (ENCODE)

- https://www.encodeproject.org
- https://www.nature.com/encode



Roadmap Epigenomics project

- >150 Primary cells / tissues
  - 6 histone marks
  - open chromatin
  - DNA methylation
  - Gene Expression



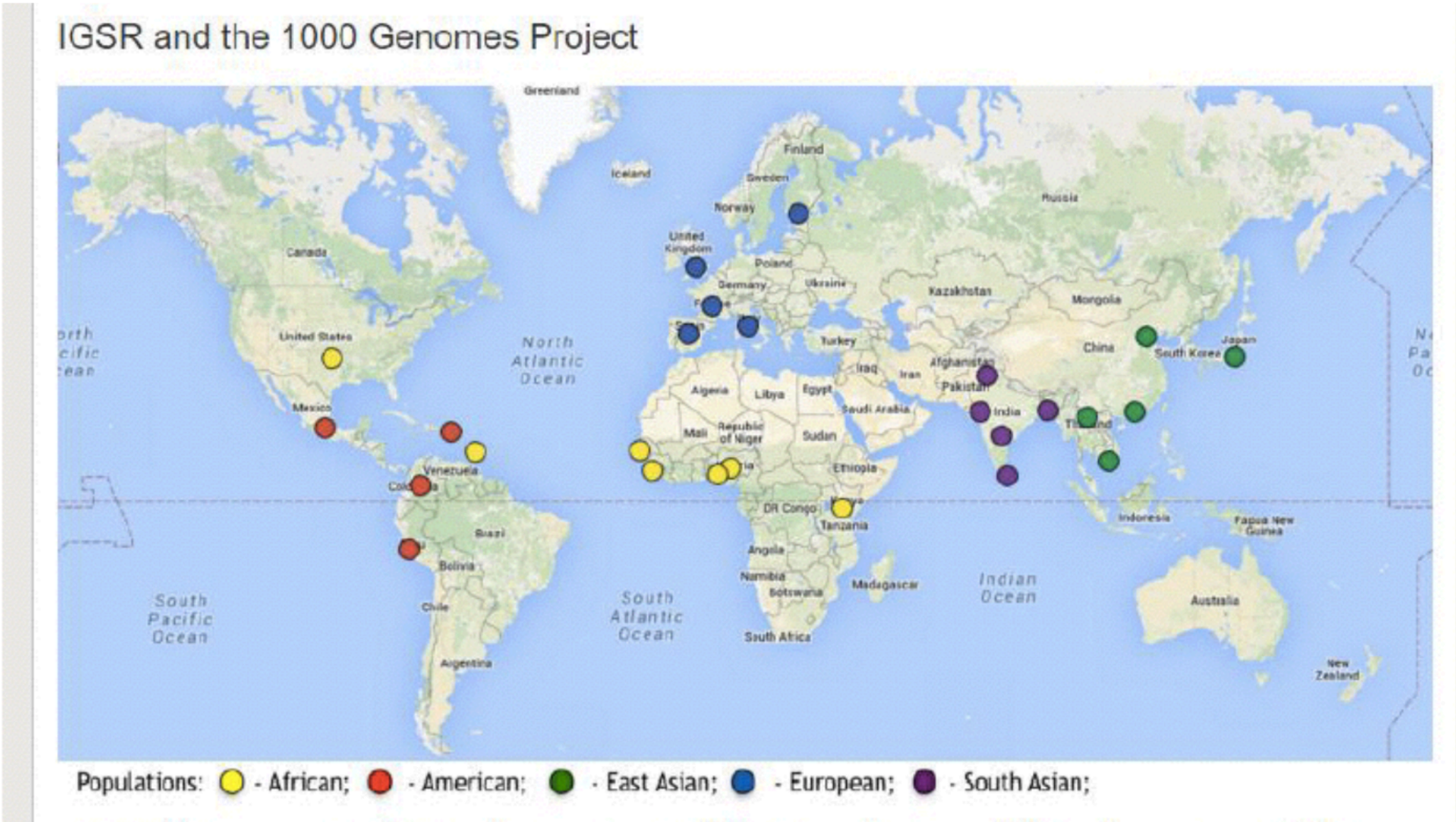
http://egg2.wustl.edu/roadmap/web_portal/

The Cancer Genome Atlas (TCGA)

- Provides platform for researchers to search, download and analyze datasets generated by TCGA
- Contains clinical information, genomic data characterization and high level analysis of tumor genomes

Available Cancer Types	# Cases Shipped by BCR*	# Cases with Data*	Date Last Updated (mm/dd/yy)
Acute Myeloid Leukemia [LAML]	200	200	03/23/16
Adrenocortical carcinoma [ACC]	80	80	03/14/16
Bladder Urothelial Carcinoma [BLCA]	412	412	03/14/16
Brain Lower Grade Glioma [LGG]	516	516	03/14/16
Breast invasive carcinoma [BRCA]	1100	1097	03/14/16
Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC]	308	307	03/14/16
Cholangiocarcinoma [CHOL]	36	36	03/14/16
Colon adenocarcinoma [COAD]	461	461	03/16/16
Esophageal carcinoma [ESCA]	185	185	03/14/16
FFPE Pilot Phase II [FPPP]	38	38	01/25/16
Glioblastoma multiforme [GBM]	529	528	03/16/16

1000 Genomes Project

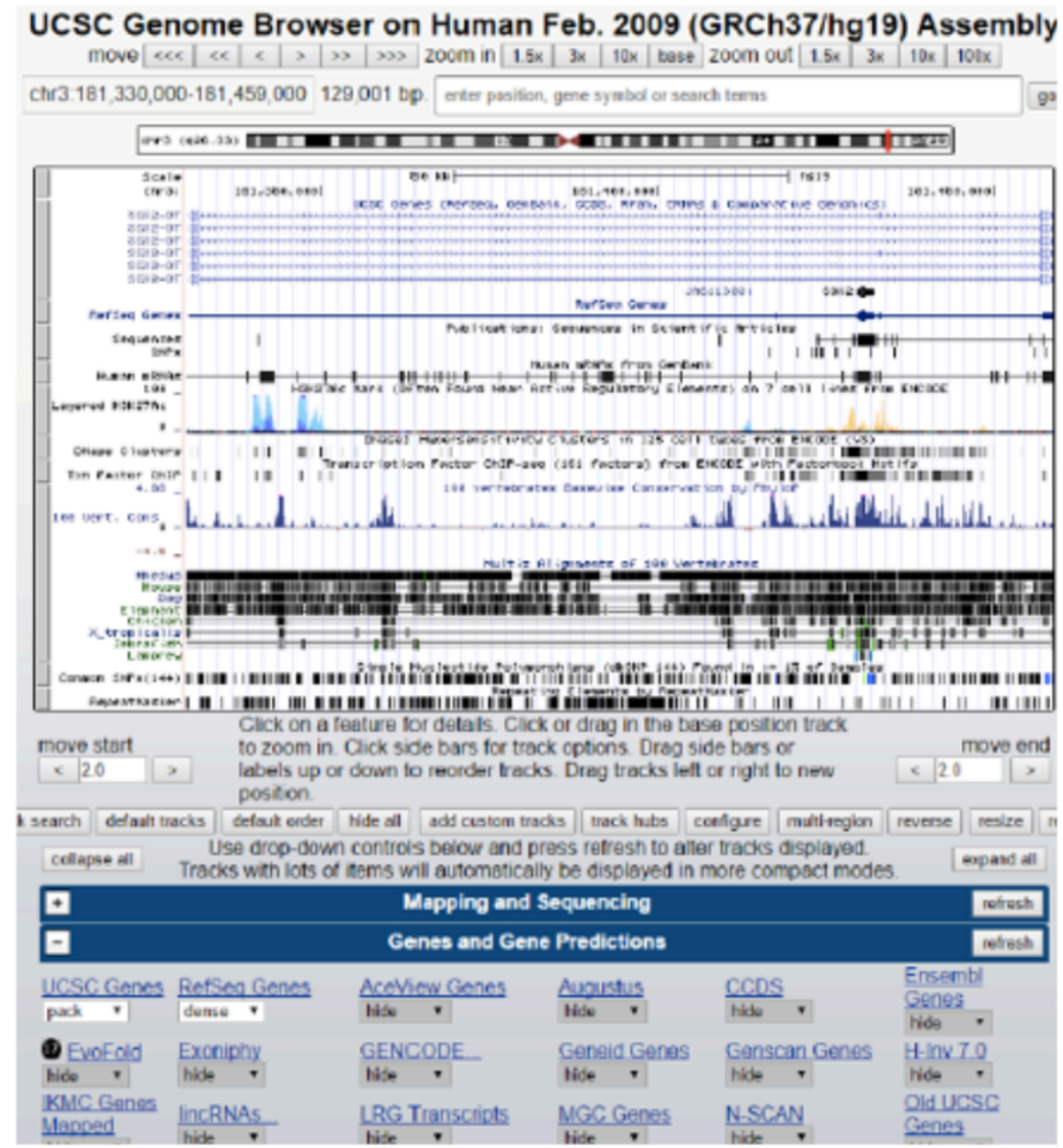


- http://www.1000genomes.org

Genome browsers

USCS Genome Browser

- Old browser: clunky and a bit ugly
- High utility: largest unified collection of built in data tracks



WashU Epigenome Browser

- Next-gen browser: dynamic, pretty, more responsive and can handle large-scale data.
- Visualizing multiple loci simultaneously
- Visualizing long-range genome interaction data
- Widgets for built-in data analysis (scatter plots, correlations aggregate plots)

References

- Kundaje and Pritchard lecture notes, GENE245
- Linkage Disequilibrium Part 1, University of Washington
- Statistical Challenges in genome-wide associations, Lecture14 University of Oslo
- How to interpret a Genome-wide association study, by Thomas Pearson and Teri Manolio
- Hands-on tutorial to Genome-wide Association studies, by Umit Seren