



# Introduction to Bioinformatics

## Topic 16 GWAS

### Lecturers

Gregory R. Grant

November 8<sup>th</sup>, 2023

Gregory R. Grant

Genetics Department

[ggrant@pennmedicine.upenn.edu](mailto:ggrant@pennmedicine.upenn.edu)

### Teaching Assistants

Chetan Vadali

*ITMAT Bioinformatics Laboratory  
University of Pennsylvania*

# Theory Suggests That All Genes Affect Every Complex Trait

🗨️ 12 | 📄

*The more closely geneticists look at complex traits and diseases, the harder it gets to find active genes that don't influence them.*

## The Goal

- Some phenotypes and diseases involve just one gene.
  - E.g., sickle cell anemia.
- Others involve many, possibly most genes.
- **GOAL:** For every phenotype or disease with a genetic component, find the causative genes.
- (We can figure out how they do it later, first is to simply find the culprits.)

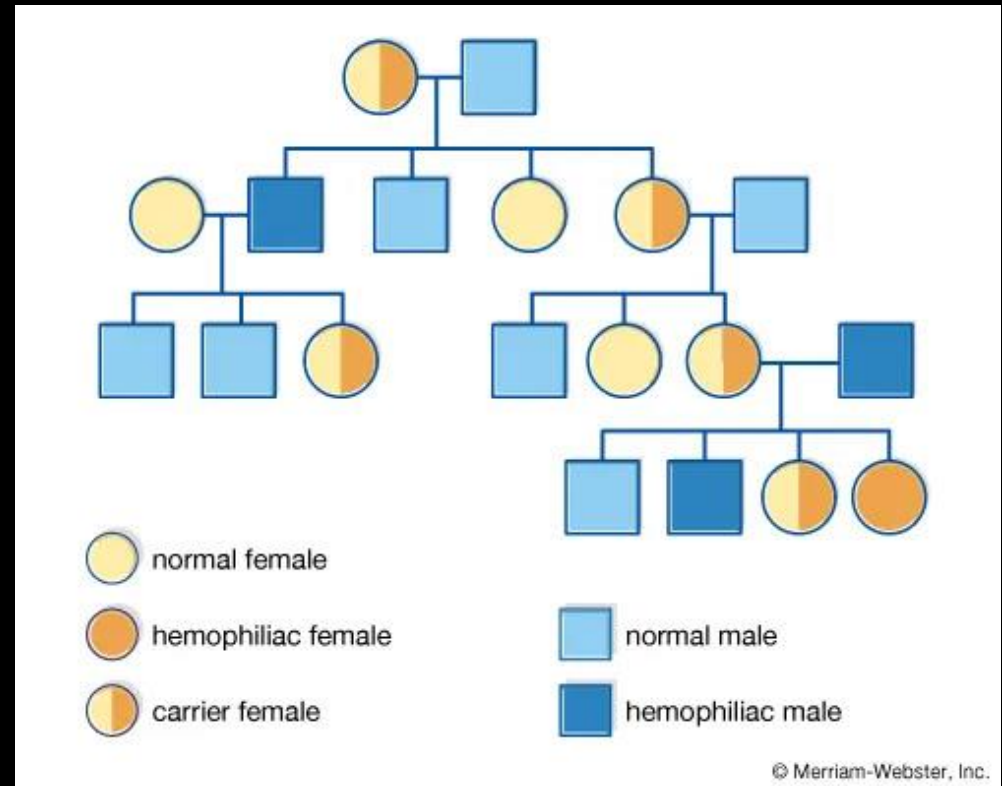


Complex traits such as eye color are the products of many genes working in concert. Just how many genes may be an open question — but the answer could be all of them.

# New Solution to Old Problem

## Pedigrees

- Before “Omics” one had to find *related* individuals with a phenotype.
  - So-called “pedigrees”
- Each affected parent/offspring narrows down the genomic location of a causative gene by  $\frac{1}{2}$ .
- It requires many pedigrees and many years to find the culprit gene or genes this way.
- Works well for monogenic traits.
- The more genes involved, the harder.



# Penetrance

- Another complicating factor is that genotypes do not 100% determine phenotypes or diseases.
  - They interact with the environment.
- The lower the penetrance, the harder it is to find the association and the more subjects are required.
- This is why GWAS often requires thousands or even hundreds of thousands of subjects.
- Fortunately, with big biobanks like the UK Biobank, the Million Veterans DB, 23andMe, etc., there are such well annotated cohorts.

## Dictionary

Definitions from [Oxford Languages](#) · [Learn more](#)



pen·e·trance

*/ˈpenətr(ə)ns/*

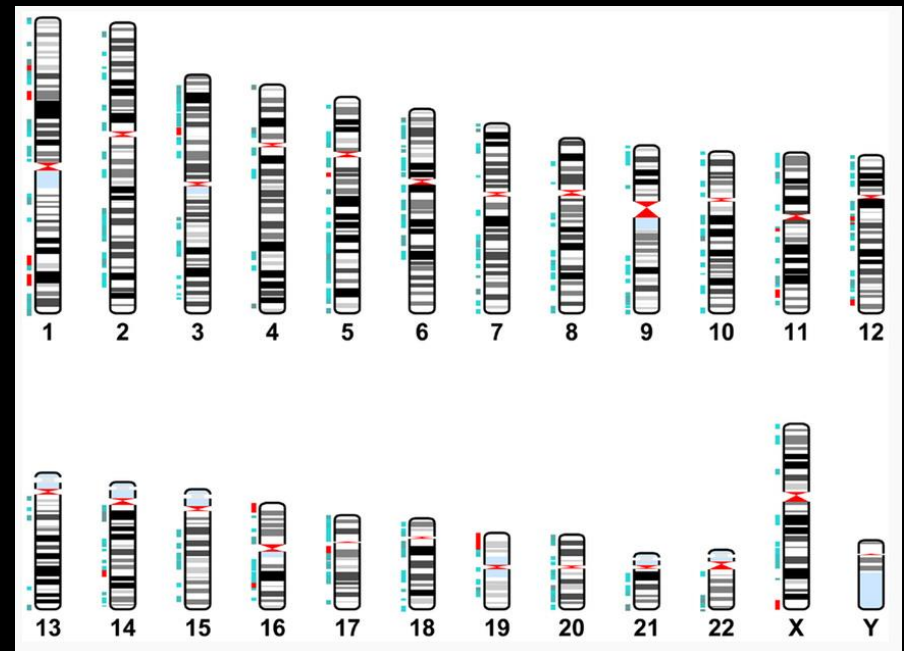
*noun* **GENETICS**

noun: **penetrance**

the extent to which a particular gene or set of genes is expressed in the phenotypes of individuals carrying it, measured by the proportion of carriers showing the characteristic phenotype.

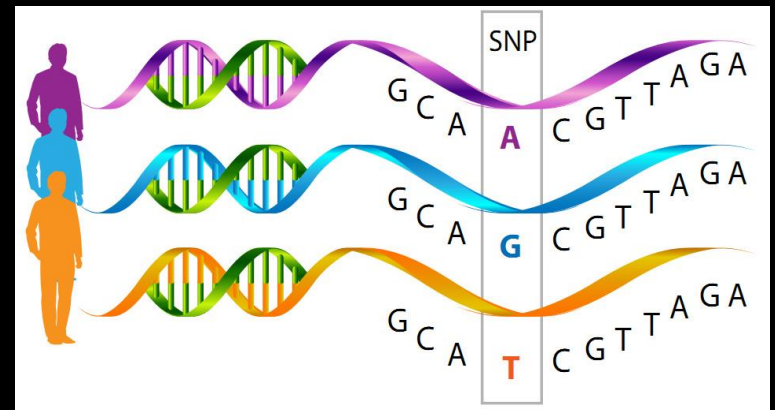
# Genome Wide Association Studies (GWAS)

- Traditionally gene finding involved identifying landmarks in the genome (cytobands) that can be traced through generations.
- Now that we can sequence DNA down to the single base, these landmarks take the form of SNPs.
- And SNPs are more than just landmarks, they are also explanatory.



# Single Nucleotide Polymorphisms (SNPs)

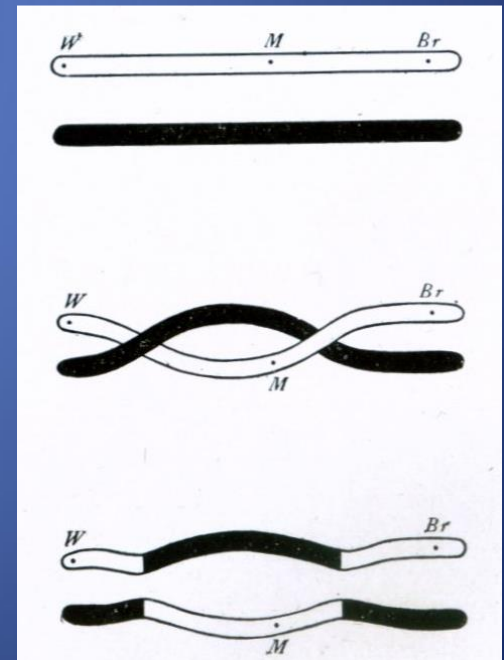
- What ultimately makes us different?
  - To a large extent it's Single Nucleotide Polymorphisms (SNPs)
- SNPs are locations in the genome which differ amongst individuals, *and for which both versions occur in at least 1% of the population.*
- 99.9% of locations in the genome are not SNPs
  - They are the same in pretty much everybody.
- At any given location there are always four possibilities: A, C, G, T
  - But it's almost never the case that there are more than 2 each with  $\geq 1\%$  of the population.
  - You prove this in a population genetics course.





# SNPs and Crossovers

- There are approximately 5 million SNPs in the human genome.
  - That's far fewer than people expected, until it was finally counted in the 2000s.
  - We're a very homogeneous species, we're practically clones of each other.
- Due to crossovers, SNPs are transmitted from parent to child in large chunks.
  - There are only one to two crossovers per chromosome per generation.
- For SNPs that are close to each other, it can take many generations for a crossover to happen between them.

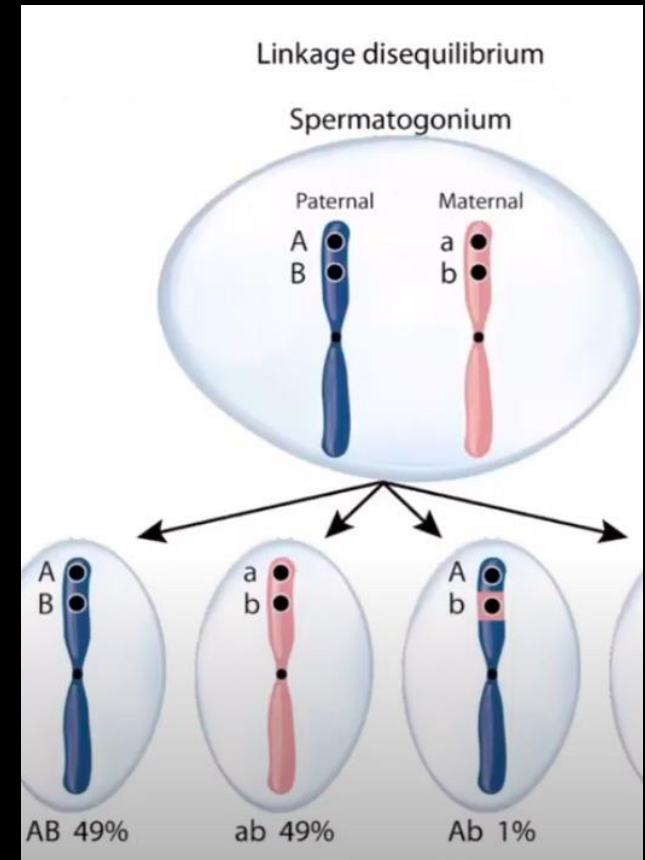


# Linkage Disequilibrium

- SNPs that are on *different chromosomes* transmit largely independently from each other.
- SNPs that are close to each other tend to be transmitted together.
  - Such SNPs are said to be in “Linkage Disequilibrium”

## FORMAL DEFINITION

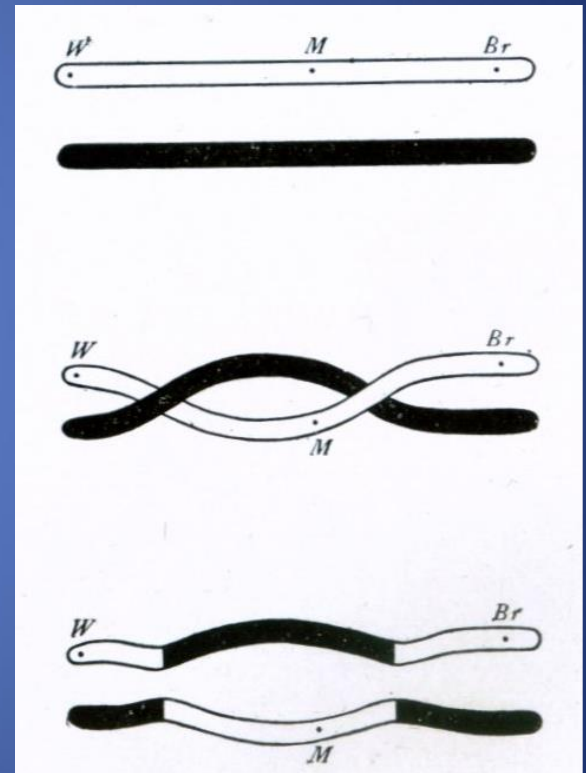
- Suppose SNP variant  $A$  occurs at one location in the genome and variant  $B$  at another location.
- Let  $P_A$  and  $P_B$  be their frequencies in the population.
- Let  $P_{AB}$  be the frequency that they occur together.
  - Same chromosome in same individual.
- If  $P_A P_B = P_{AB}$  then the two SNPs are in linkage *equilibrium*.
- The greater  $|P_A P_B - P_{AB}|$  is, the further they are from equilibrium.





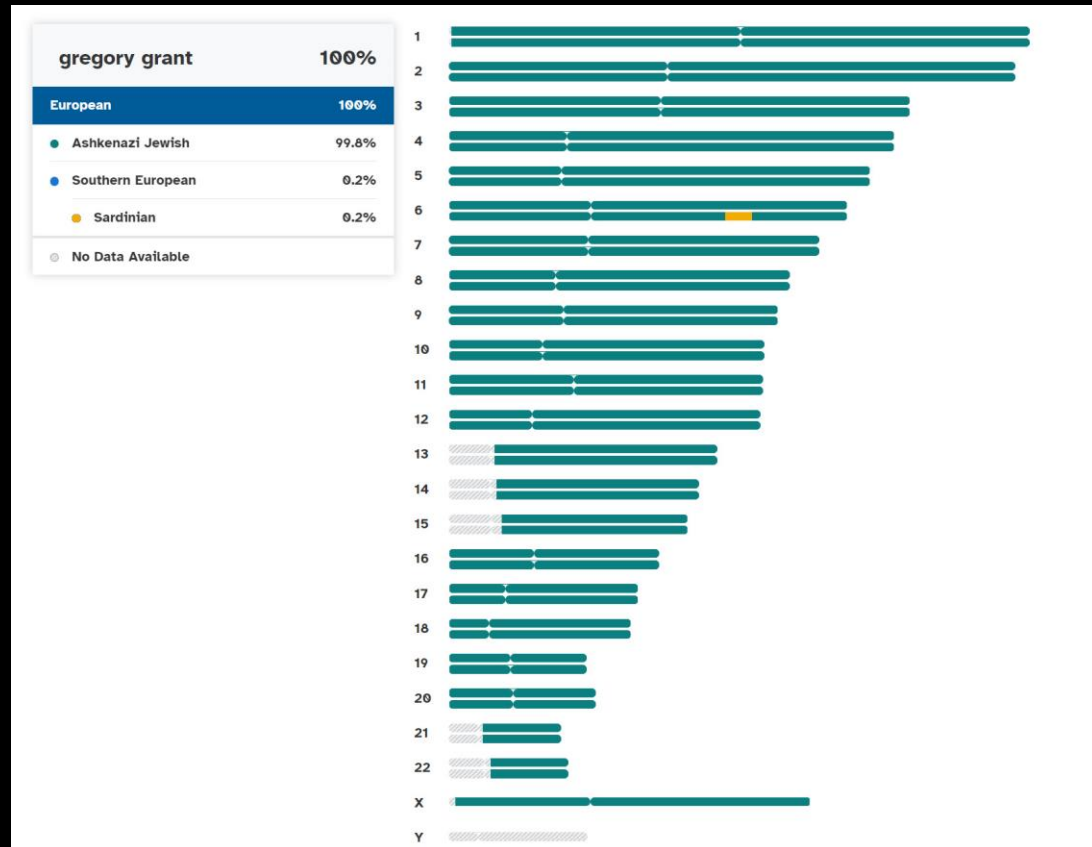
# Linkage

- There are only one to two crossovers per chromosome per generation.
  - So, we inherit large chunks of chromosome in each generation, only *very* coarse scrambling happens.
- When you calculate that you are 3.125% related to a great-great-great grandparent  $\left(\frac{1}{2}\right)^5$  it's not 3.125% spread evenly throughout the genome.
  - It'll be one or a few contiguous pieces on a few chromosomes, and quite *possibly none*.



# Example

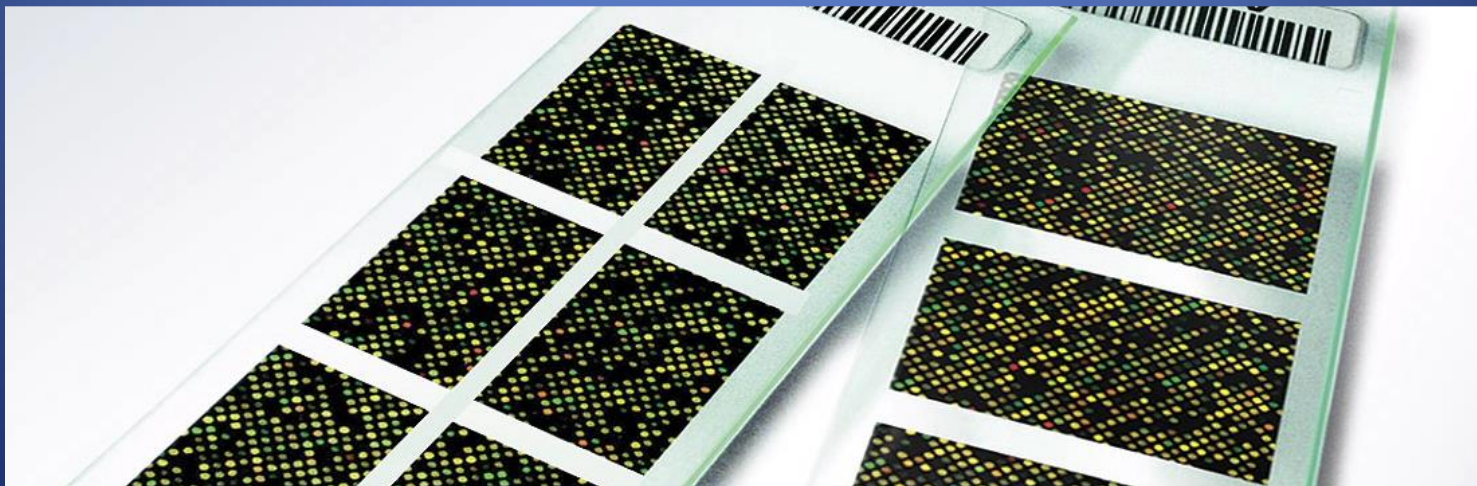
- This is my genome, assayed by SNP array.
- I'm 0.2% Sardinian, but the entirety of what I've inherited from that individual is just one contiguous stretch of chromosome 6.
- And there's some 75% chance any child of mine would end up with none of it.
  - They could get the wrong chromosome or the wrong half of the right chromosome.
- In fact, if you go back just 10 generations then you likely have no DNA left at all from more than half of your direct ancestors.



The point is, SNPs that are in LD tend to stay together for a very long time, many generations.

# SNPs

- SNPs are typically measured genome-wide with microarrays.
  - For SNP calling, microarrays have been slow to be replaced by sequencing.
- The array has probes for each of the two variants, at a large set of known SNP locations.
  - The rest are inferred from population/haplotype information.
- We talked about calling SNPs from DNA-Seq but it's still too expensive for GWAS which needs thousands of subjects.

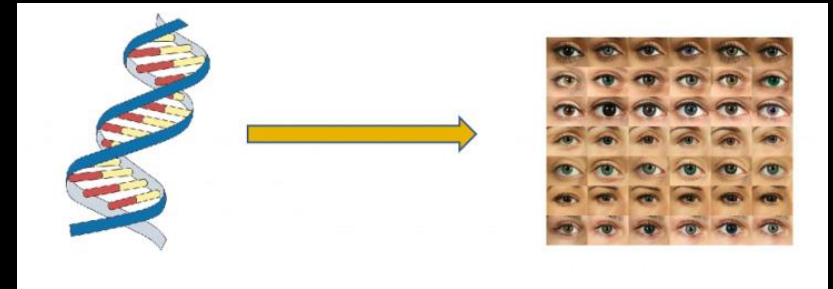


# Haplotypes and other variants

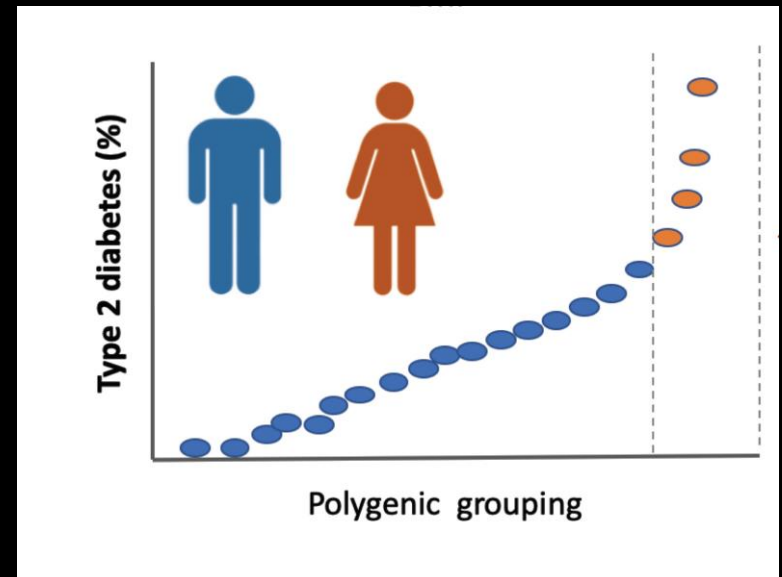
- Phasing and population info about haplotypes are generally used to infer SNP's not on the array.
- Modern GWAS studies involve upwards of 5M SNPs.
- But SNPs are not the only types of variants in the genome.
- Sometimes what makes you different is not a substituted amino acid but an additional one.
- Therefore, indels also associate with phenotypes. And they can also be included in GWAS studies.
  - But we'll focus on SNPs today.

# Uses of GWAS

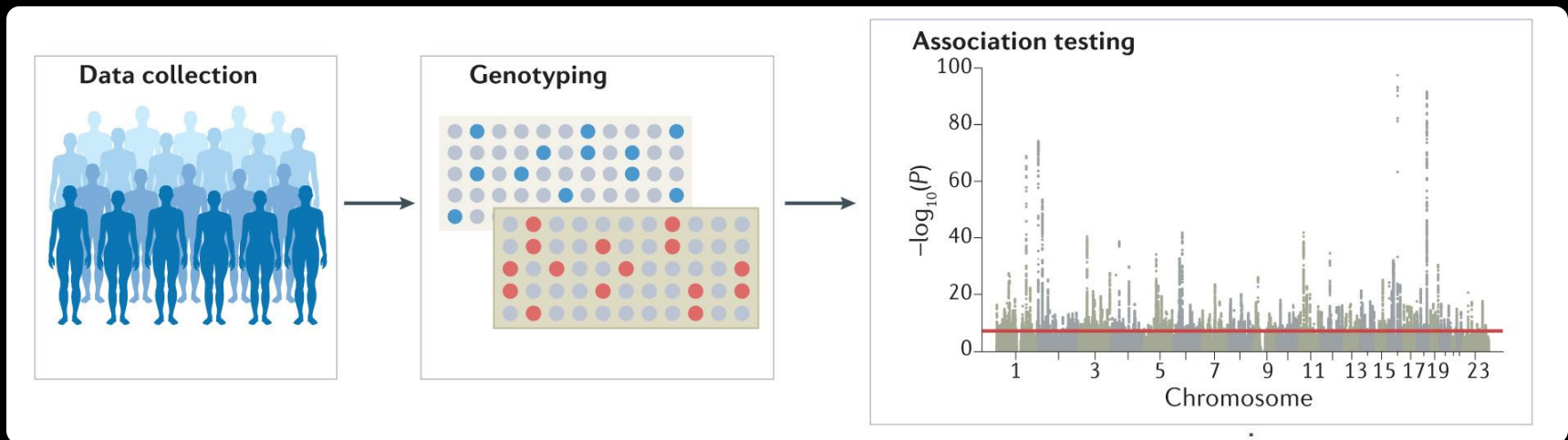
- GWAS results are used in multiple ways. For example:
  - To gain insight on the mechanism that leads to a phenotype.
  - To predict an individual's risk for a particular disease based on their genetic profile.
    - So-called Polygenic Risk Scores (PRS).
- Thousands of GWAS studies have been performed on thousands of traits and diseases, some studies involving over a million participants.
  - 23andme for example.



Genotype/Phenotype



Polygenic Risk Scores



# GWAS in a Nutshell

GWAS associates phenotypes or diseases with SNP genotypes

1. Obtain a population of individuals some of which have phenotype 1 and the rest with phenotype 2.
2. Consider one SNP with two variants A and B.
3. Each subject is one of three genotypes at the SNP
  - A/A, A/B and B/B.
4. Do a statistical test for association between genotype and phenotype.
5. Do this for all SNPs and correct for multiple testing.



# Manhattan Plots

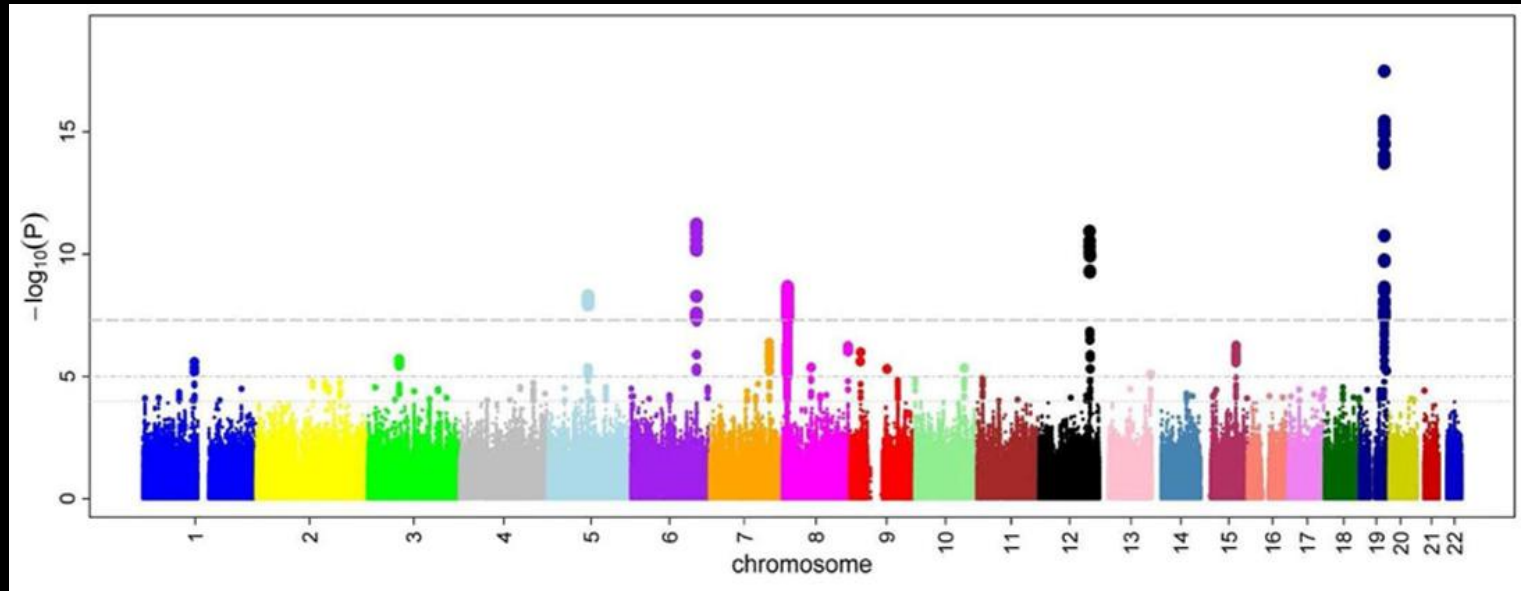
Results are presented in Manhattan Plots

Each point is a SNP.

Colored by chromosome.

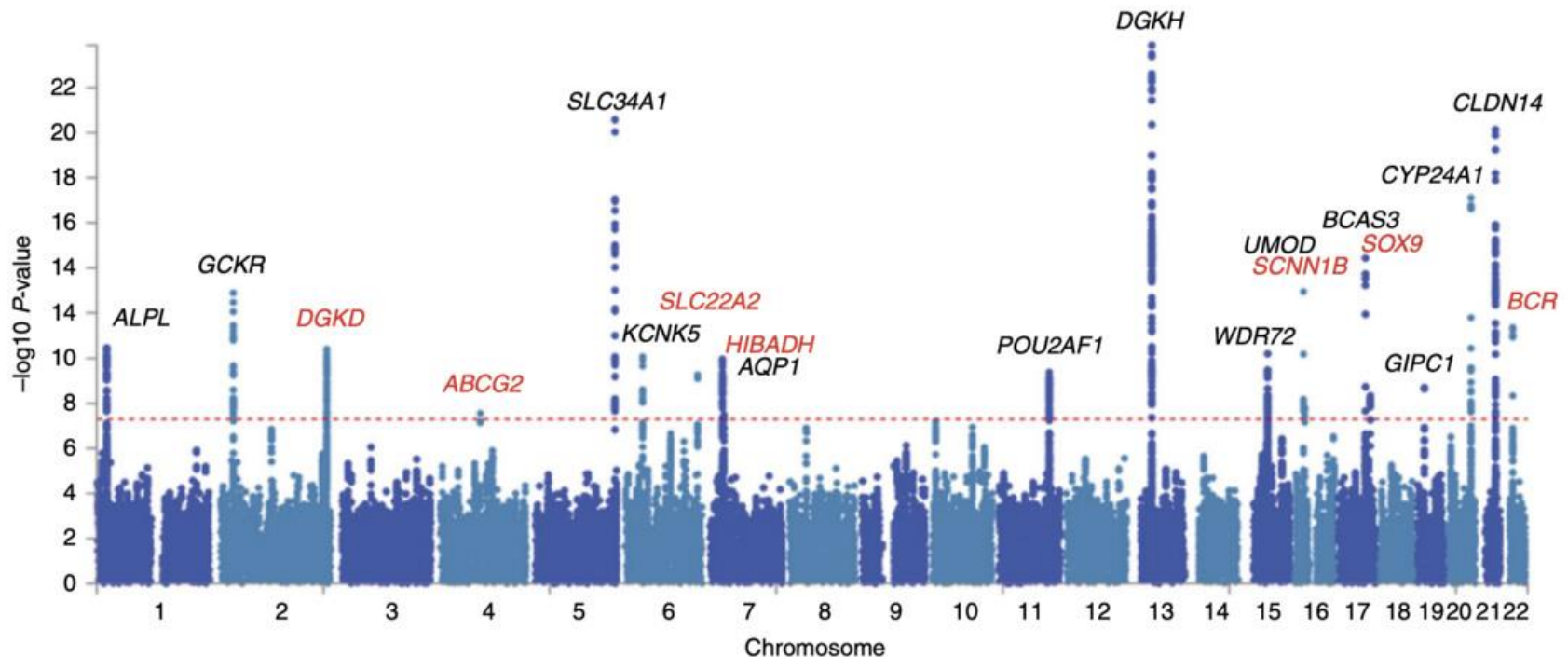
The top dashed line is a Bonferroni correction cutoff.

The middle is an FDR correction cutoff.

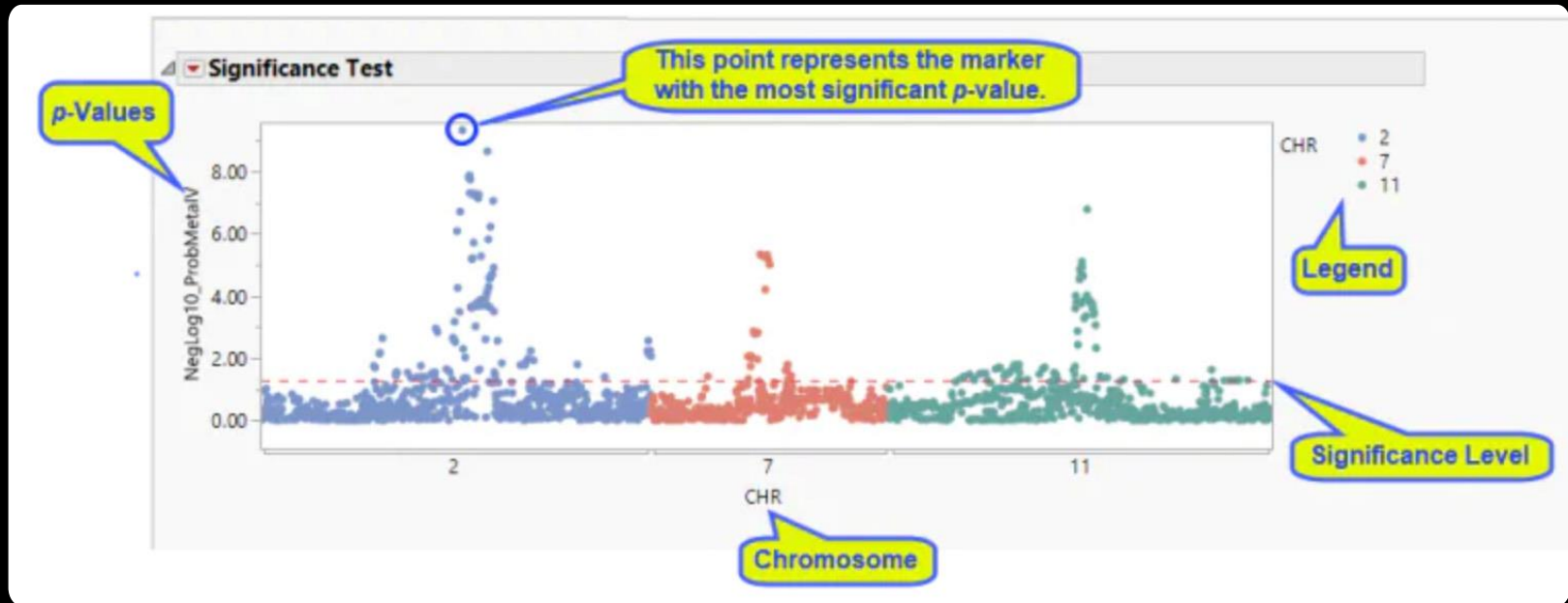


# Manhattan Plots

- It looks like a bunch of SNPs are in the exact same location.
- That's just because the x-axis is so compressed.
- Each peak is a block of SNPs under high linkage disequilibrium.
- One (or more) of those SNPs should be the causative one(s).
- But it's not usually the most significant one..



The gene names shown are of the most significant gene in that locus. But each locus has many genes, and any could be the causative ones



# Fine Mapping

- Figuring out the exact causative SNP is called the Fine Mapping Problem.
- There are usually dozens to hundreds of SNPs in a significance block.
- We'll return to this problem.

# Association Tests

	Cases	Controls
A/A	4292	157
B/B	312	3417

How is the statistical test of association done?

We'll start by considering a Simplified approach just to get the idea: *Focus only on the homogeneous subjects at a given SNP.*

- Record the homogeneous subjects' genotypes at the SNP in a 2-by-2 table.
- The data above indicates a strong association between genotype and case/control status.
- The test for association can be as simple as Fisher exact test (see next slide).
- The null hypothesis is that there is no association between the rows and the columns.

	Class I	Class II	Row Total
Blue	$a$	$b$	$a + b$
Red	$c$	$d$	$c + d$
Column Total	$a + c$	$b + d$	$a + b + c + d (=n)$

## The Fisher Exact Test

- The null hypothesis is that there is no association between the rows and the columns.
- Imagine the two genotypes are two colored balls, red and blue.
  - Imagine there's an urn with  $a+b$  blue and  $c+d$  red balls (the row totals)
- Imagine drawing  $a+c$  balls (the first column total) from the urn at random.
- We want to know if there are a small number of balls of one color in the drawn sample. (A *significantly* small number)
- This number of red balls follows the hypergeometric distribution.
  - Just like in pathway analysis.
- And same for the number of blue balls.
  - So, we use the hypergeometric to calculate  $p$ -values for significant association.
  - You're not responsible for the details.
- It is called the "Exact Test" because it doesn't involve any approximations, it's an exact probability of the 2-by-2 association.
  - If we consider all three genotypes A/A, A/B and B/B then we need a different test that's not exact.

	Cases	Controls
A/A	2915	157
A/B	1151	516
B/B	312	3417

# Chi-Square Association Test

- If the table has more than 2 rows and/or 2 columns, then we use a different association test.
  - Published in 1900 by Karl Pearson.
- We don't have time to go into the details, but association tests are relatively straightforward.
- Conceptually it should be clear what they are testing for.



# Multiple Testing

The association tests give a  $p$ -value for each SNP.

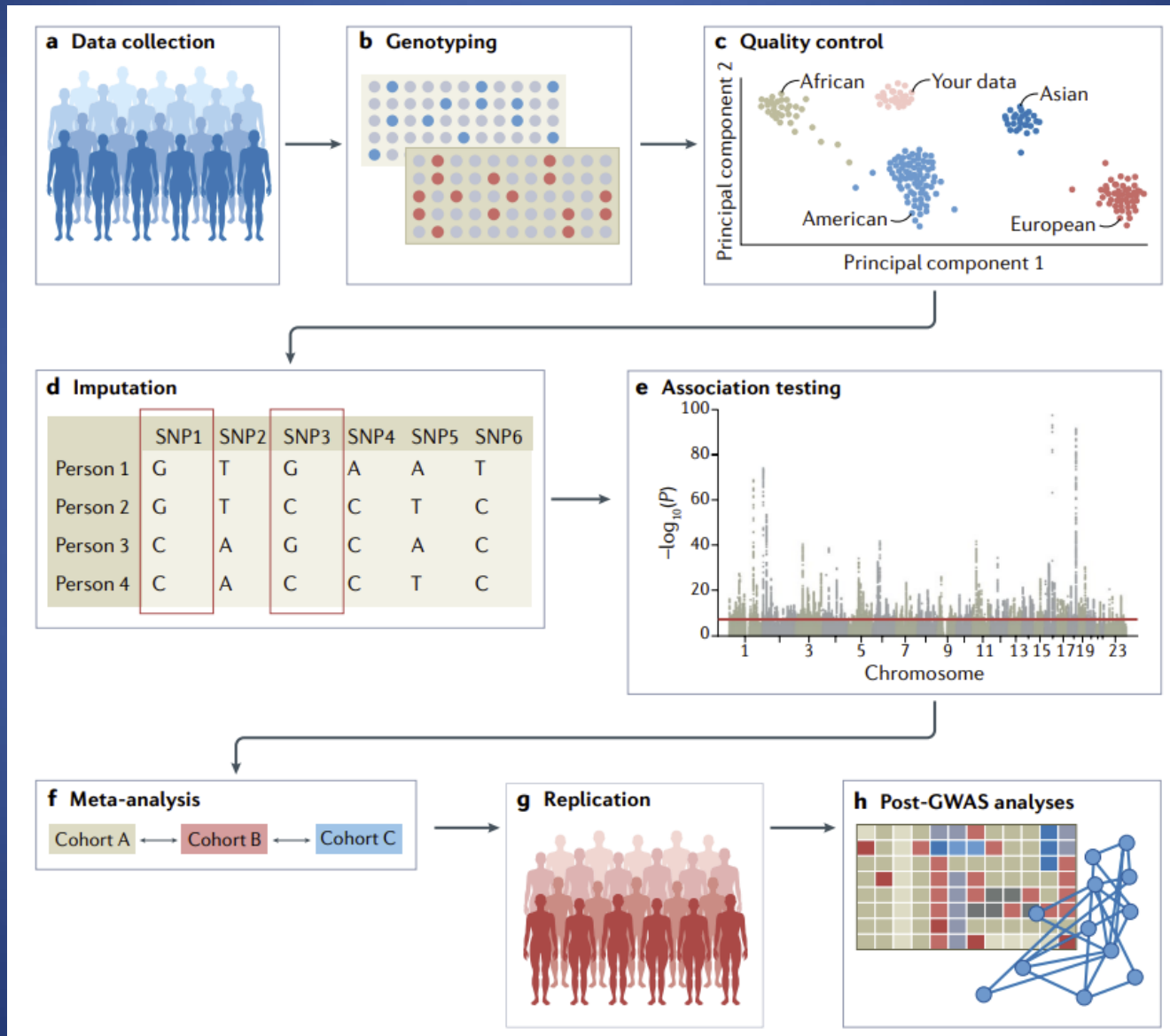
Since there are millions of SNPs they need to be corrected for multiple testing.

The correction can be made less severe by working with haplotype blocks as units of analysis, instead of individual SNPs.

The best way to correct is an ongoing debate.

Both FDWR or FDR approaches are common.

# Workflow



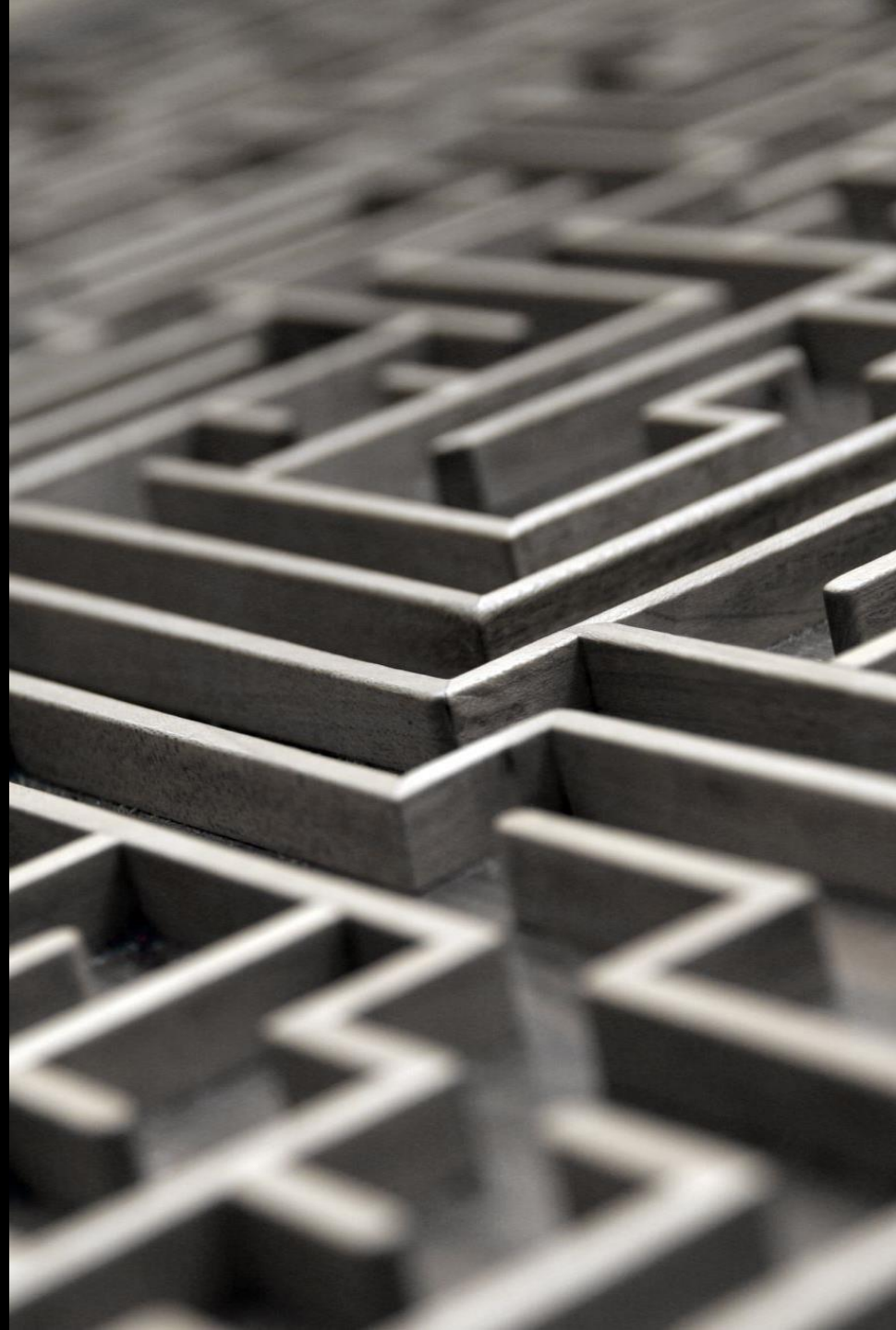
# GWAS Promises

- Around the turn of the century tremendous promises were made about GWAS.
- A few good GWAS studies will reveal the genes responsible for everything.
- And once we know the genes, an understanding of mechanism will soon follow, and from such understanding will flow miracle cures.
- As a result, tremendous amounts of money were allocated to perform GWAS studies.



# GWAS Hard Lessons

- Mother Nature didn't turn out to be as cooperative as people had hoped.
- All the easy diseases, e.g. the ones caused by a single mutation, had already been found the hard way.
- Several hard lessons presented themselves as the industry matured.



# Height

---

Even a phenotype as simple (seeming) as height involves at least 1/3 of all genes.

**Largest genome-wide association study ever uncovers nearly all genetic variants linked to height**

By analyzing data from nearly 5.4 million people, Broad researchers have identified more than 12,000 genetic variants that influence height.



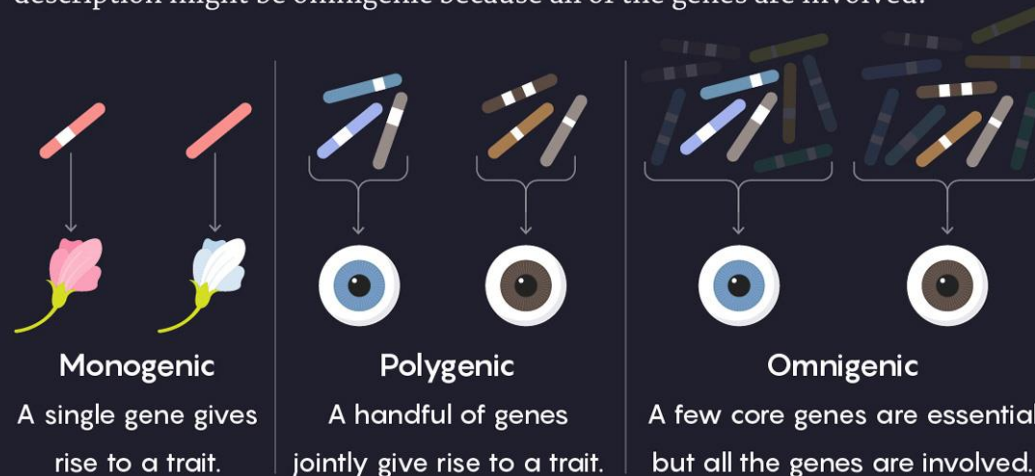
*Credit: Colette A. Zylstra, Broad Communications*

# Complex Traits and Diseases

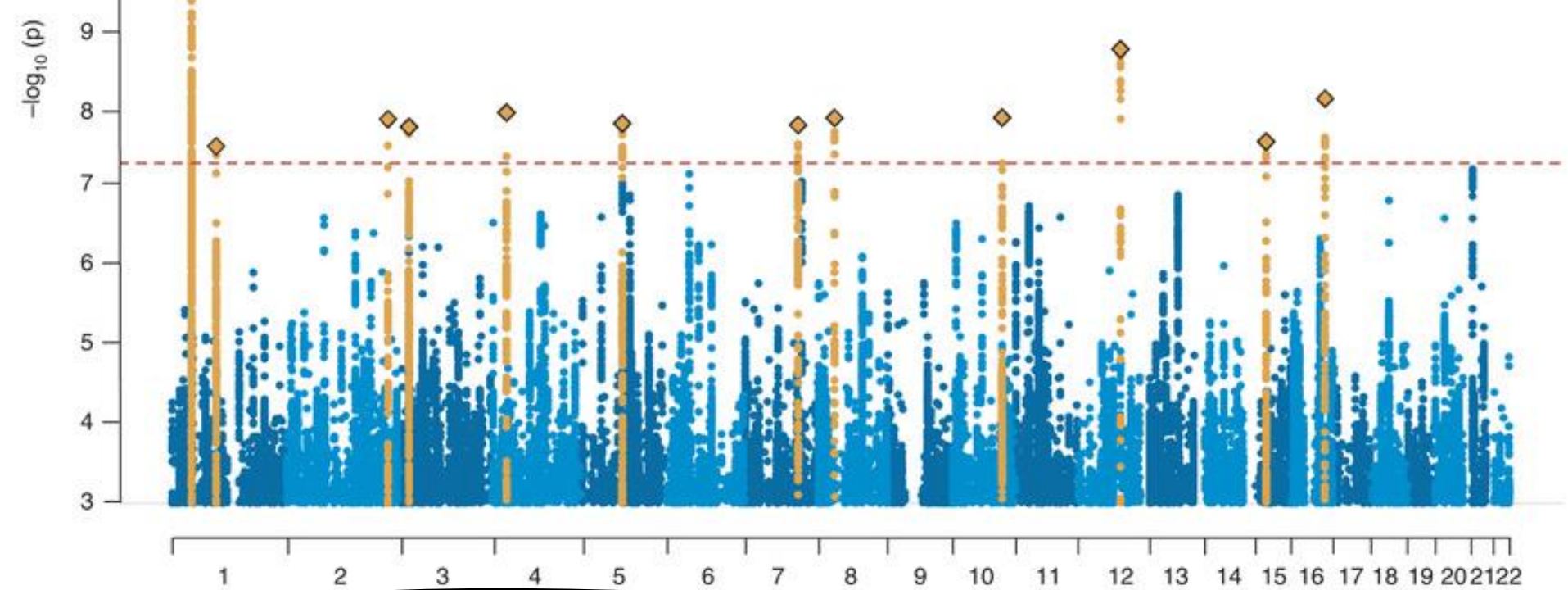
- The hard diseases are not caused by rare variants, but rather by rare combinations of common variants.
  - Alzheimer's, Parkinson, Heart Disease, Cancer, Diabetes, etc.
- Complex diseases involve thousands of genes, most of which involve many different SNPs that can lead to the same phenotypes.
- And the function of genes is tightly linked to environmental factors, which can be extremely difficult to study in controlled experiments.

## How Many Genes Are at Work?

Simple traits may be controlled by just one gene (monogenic). More complex traits are usually considered polygenic, but a new theory suggests that a better description might be omnigenic because all of the genes are involved.







- GWAS sounds simple enough.
  1. Find some individuals with a trait and some without.
  2. Profile their SNPs and find the ones that are different between the two groups.

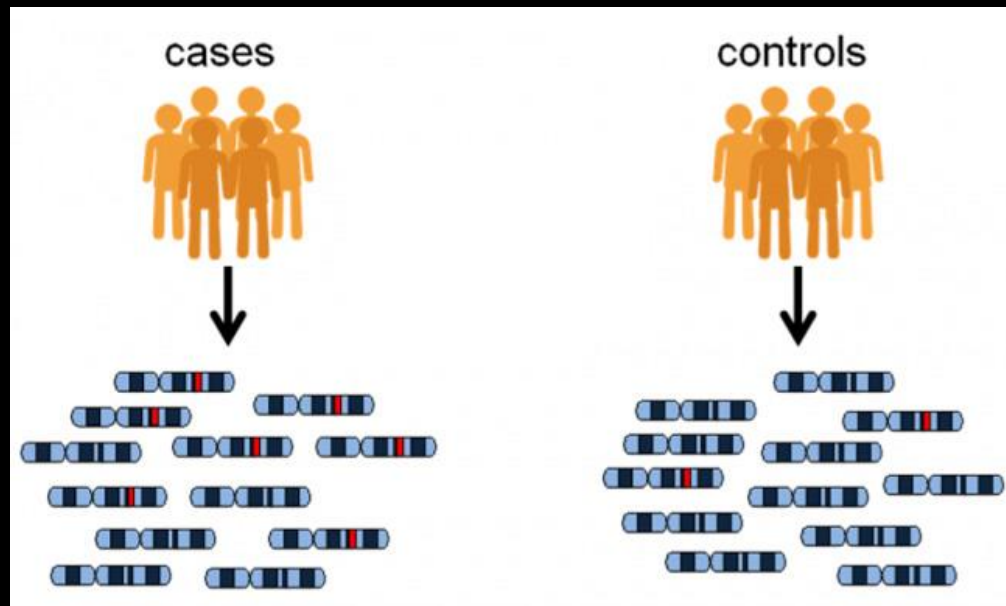
## Power

### Easy to say...

- For complex diseases, it requires thousands of subjects to achieve statistical significance. Sometimes tens or hundreds of thousands.
- There's also a huge multiple-testing problem here.
  - Each SNP is a test so we're doing millions of tests in parallel.
  - So, this is a much bigger problem than RNA-Seq with 30K tests.

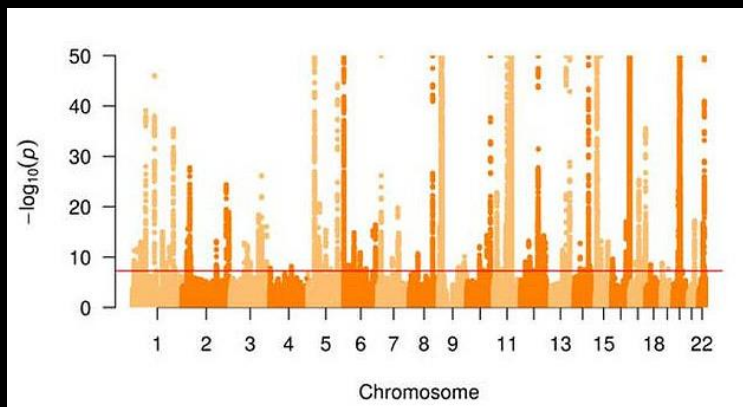
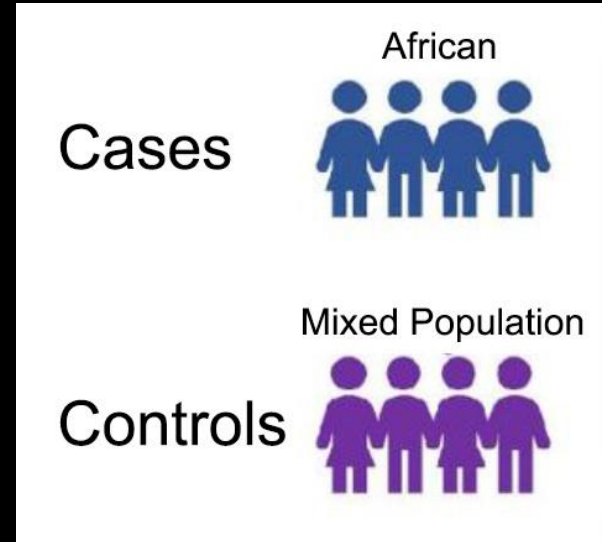
# Difficulties Choosing Populations

- Suppose we perform a GWAS for sickle cell anemia, a disease that only affects people of African descent.
  - Having one copy of the sickle cell variant makes people immune to malaria. This is called “heterozygous advantage”.
  - Unfortunately, having two copies causes a blood disorder.
- Suppose without thinking too hard about it, we find 100 people with and 100 people without the disease.



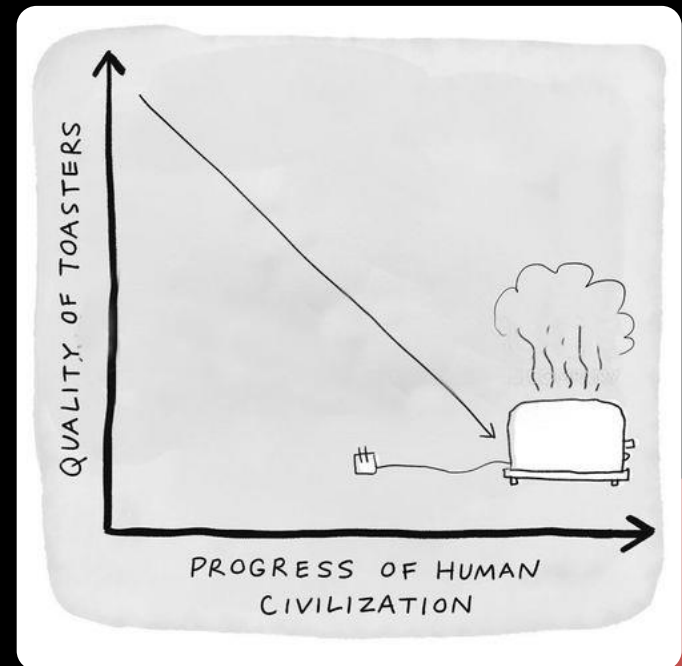
# Confounding

- The set of people with the disease will be of African Descent.
- The set of people without will be from everybody else.
- Therefore, we'll find a significant association between sickle cell anemia and genes that encode for skin pigmentation.
- That will lead us to investigating a lot of genes that have nothing to do with the disease.
- Conclusion: the two populations must be similar on all factors except the trait in question.



# Causality

- The key word in GWAS is “Association”
  - These studies identify *associations* between phenotypes.
  - But association does not imply causation.
- For example, if you collect data, you’ll find that more *ex-smokers* die of lung cancer than active smokers.
  - Does this therefore imply that quitting smoking raises your chance of lung cancer?
- It’s exactly the opposite: Being diagnosed with lung cancer causes most smokers to quit.
  - So, most people who die of lung cancer are ex-smokers.
  - Quitting does not raise the incidence.
  - But it is highly *associated* with it.



# Populations vs. Experiments

Sleeping with shoes on is correlated with waking up with a headache.

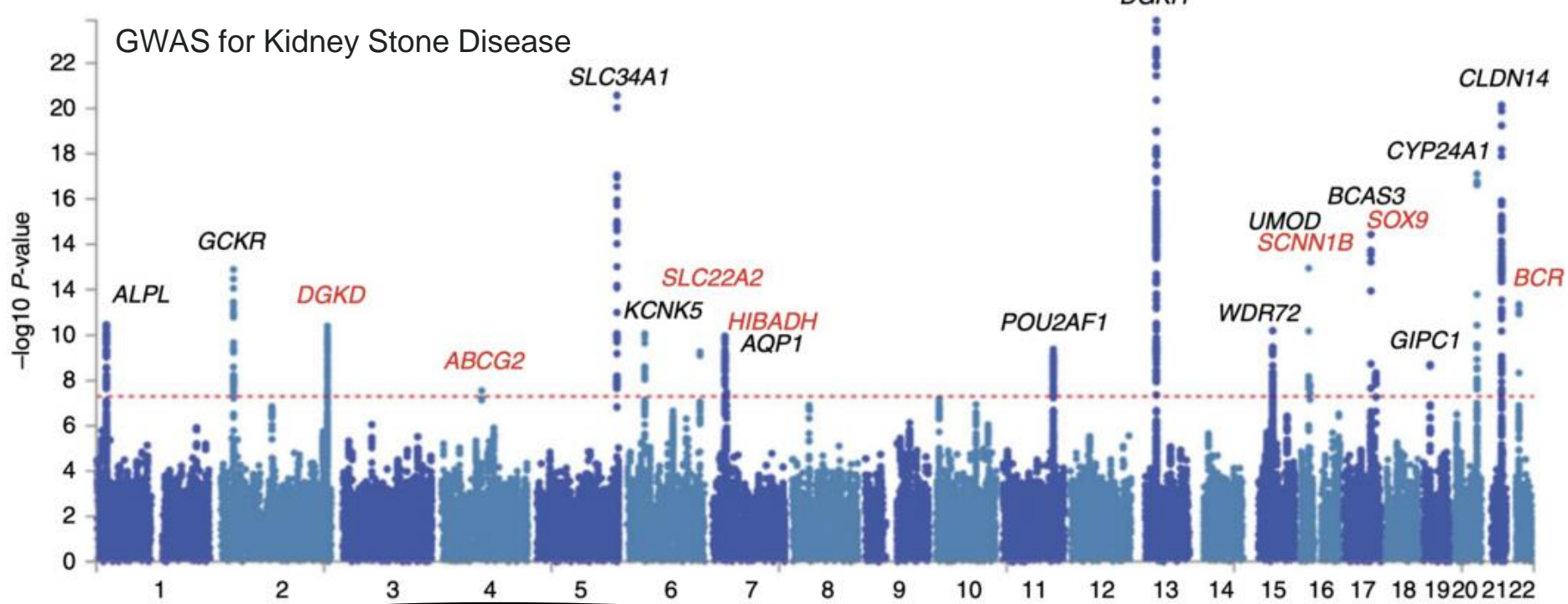
- But sleeping with shoes on is also associated with falling asleep drunk.

What exactly is causing what? And what is coming along for the ride?

- We could easily get to the bottom of this one by designing a controlled experiment.

But we cannot design experiments for populations and phenotypes.

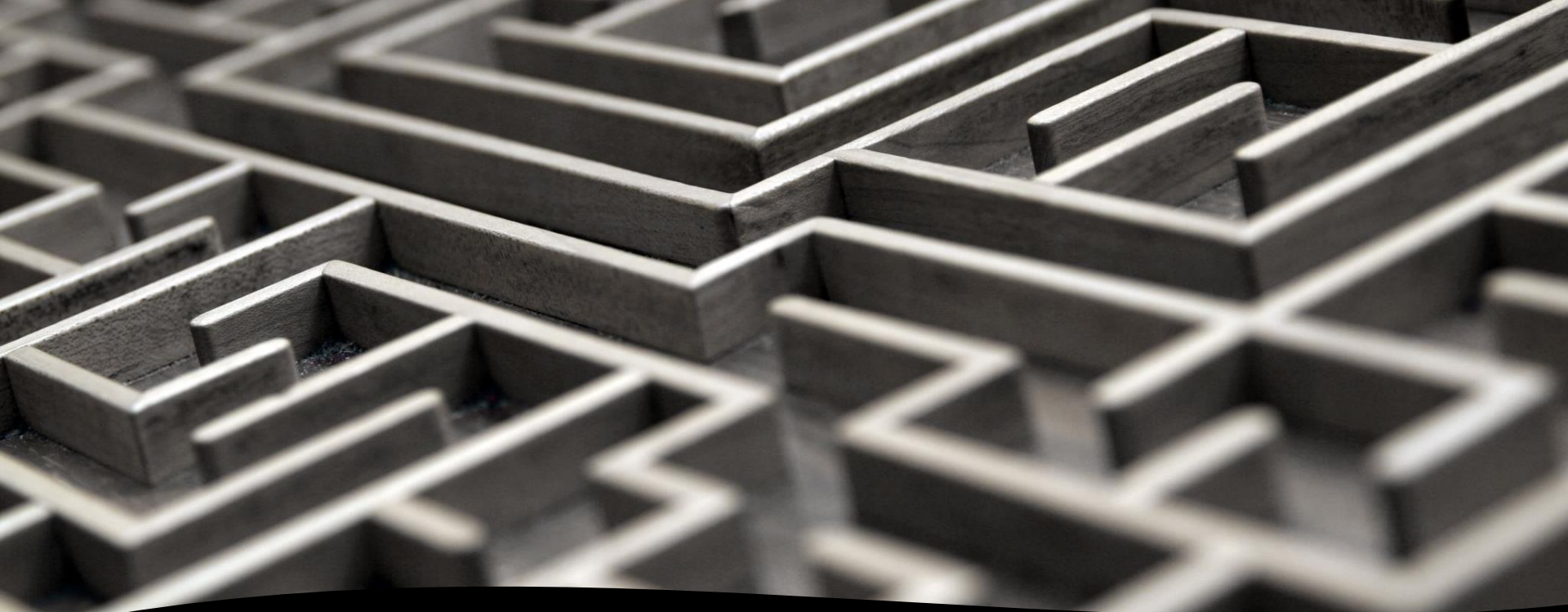
- We achieve the same thing by careful selection of individuals.



## Linkage and Fine Mapping

- Linkage makes blocks of proximal SNPs travel together from generation to generation.
- Often these blocks contain hundreds of SNPs.
- If only one of them is the causative ‘functional’ SNP that underlies the phenotype, it could be extremely difficult to sort it out from the rest of the SNPs coming along for the ride.
- This is the so-called “fine-mapping” problem.





## Progress

- Around 20 years ago biology experienced the introduction of high-throughput methods, such as microarrays, sequencing, and GWAS.
- For the first 10 years or so, a huge amount of effort was put into generating data and finding associated SNPs.
- Less effort was put into finding mechanism, interpretation and translation into medicine.
- As a result, for a while it was being claimed that not a single useful result had come from all the GWAS efforts and promises.
- Things are better now. But getting from associated SNPs to real biology is hard.

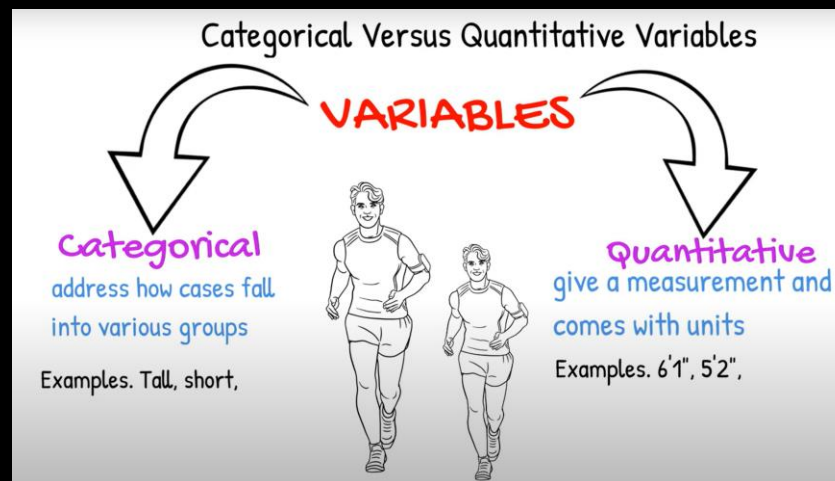
## Progress

Slowly progress is being made and GWAS continues to be a thriving industry.

Slowly investigators are picking through the published associations and working out causality and mechanism.

# Two Types of Traits: Categorical and Continuous

- The design of a GWAS study depends on the type of trait being studied.
- Some traits are qualitative
  - E.g. Blood Type A vs. Blood Type B
- Some traits are quantitative
  - E.g. Height
- If the trait is qualitative with two possibilities, we need individuals from both categories.
- If the trait is quantitative, then we need individuals with a range of values.
  - And a different statistical test.



# Cohorts

- The number of subjects needed to achieve sufficient power is usually in the thousands.
  - Both phenotype and genotype info must be obtained from all subjects.
- The resources required to obtain such data are greater than most labs can afford.
- There are public resources with large cohorts with both genotypic and phenotypic information.
  - E.g. the UK Biobank
- Most studies use these resources.
  - And may add some in-house data.

Review Article | [Published: 29 May 2018](#)

## From genome-wide associations to candidate causal variants by statistical fine-mapping

[Daniel J. Schaid](#) , [Wenan Chen](#) & [Nicholas B. Larson](#)

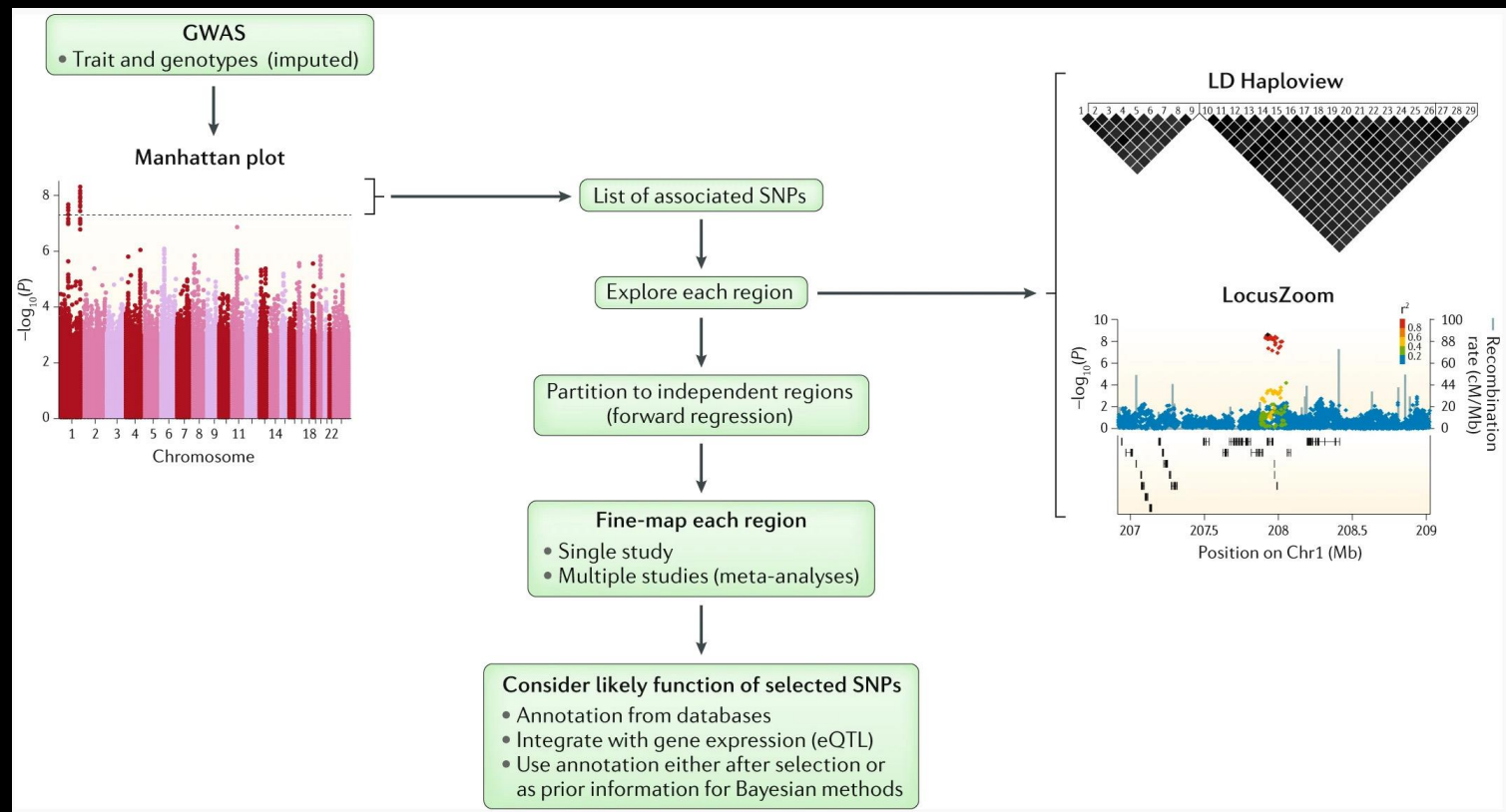
[Nature Reviews Genetics](#) **19**, 491–504 (2018) | [Cite this article](#)

## Fine Mapping

Big subject, which uses many techniques we've studied.


# Fine Mapping

Almost every technique we've studied comes to bear on the fine mapping problem. We'll look at just a couple.




# eQTL's

- eQTL stands for “Expression Quantitative Trait Locus”.
- Which is a fancy way of saying a gene whose expression is affected by a SNP.
- Just as height or lifespan are traits that can be associated with SNPs, so is “expression level of Gene X in Tissue Y”.



Biochimica et Biophysica Acta (BBA) -  
Molecular Basis of Disease  
Volume 1842, Issue 10, October 2014, Pages 1896-1902



Review

## From genome to function by studying eQTLs ☆

[Harm-Jan Westra](#) ✉, [Lude Franke](#) 👤 ✉



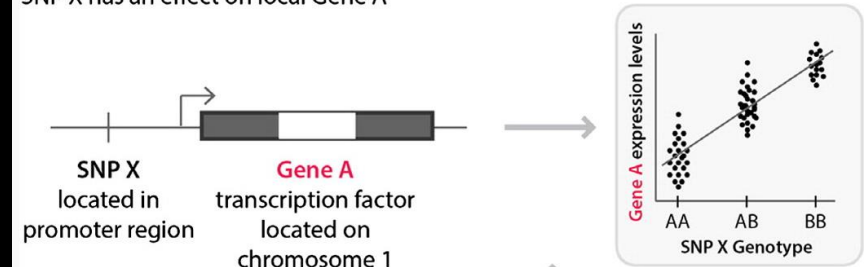
# CIS and Trans

- Expression across the cohort is plotted for the three genotypes at a SNP.
- SNPs can affect the expression of genes near them, or far away.
- So, this is an even bigger multiple testing problem:

Number of SNPS X Number of Genes.

## Cis-eQTL

SNP X has an effect on local Gene A

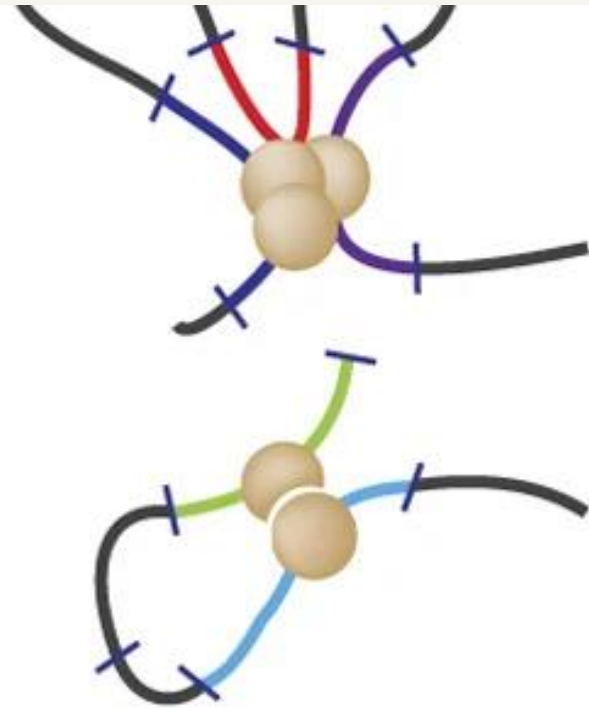
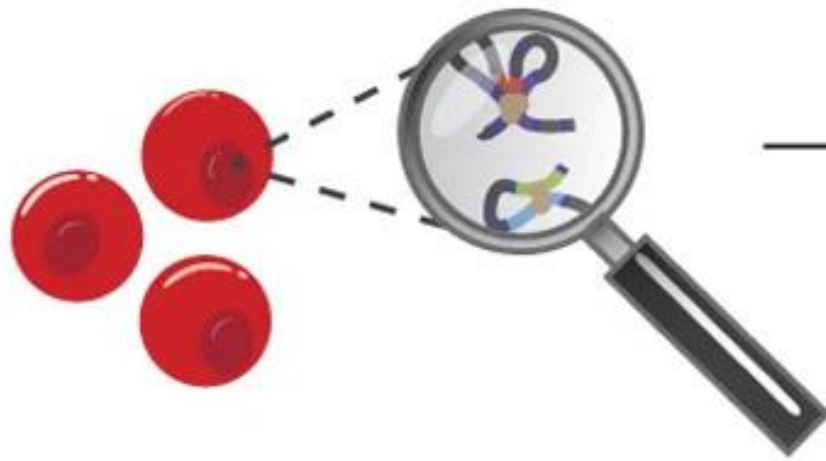


Altered **Protein A** levels, effect on the binding to the transcription factor binding sites of downstream genes

## Trans-eQTL

SNP X has an effect on distant Gene B through an intermediary factor (such as a transcription factor)



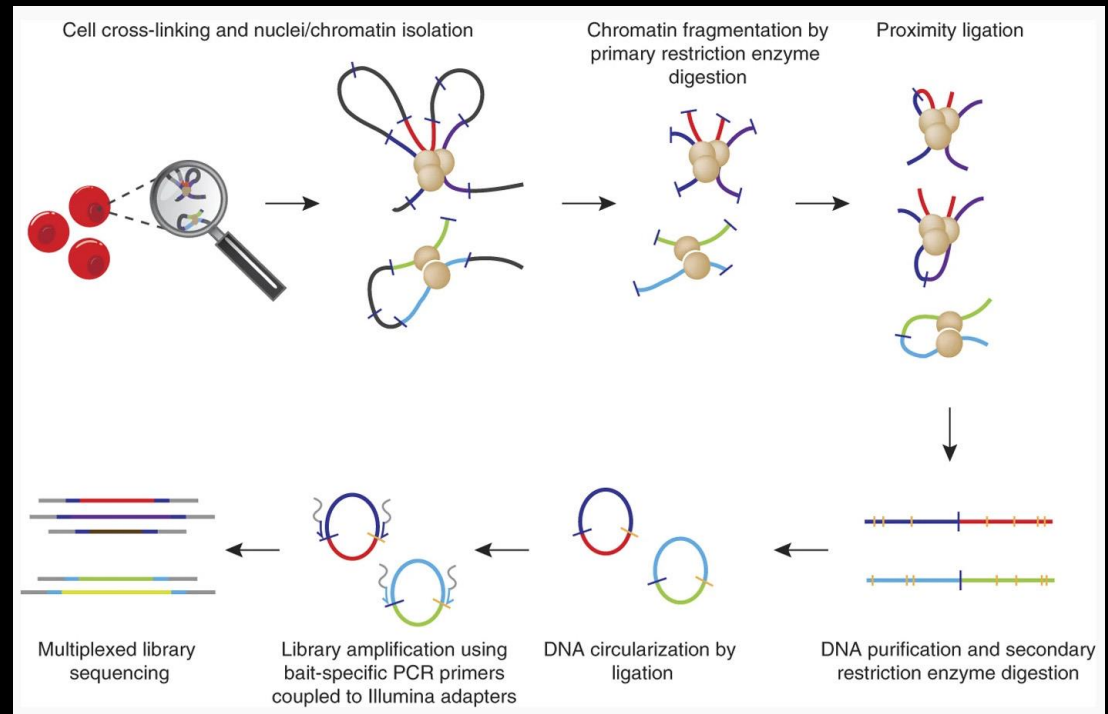


# Chromatin Conformation Assays

- A SNP can affect a gene that it's far away from.
- DNA folds up and puts distant regions in contact.
- Chromatin Conformation Assays allow us to search for such interactions with high-throughput sequencing.

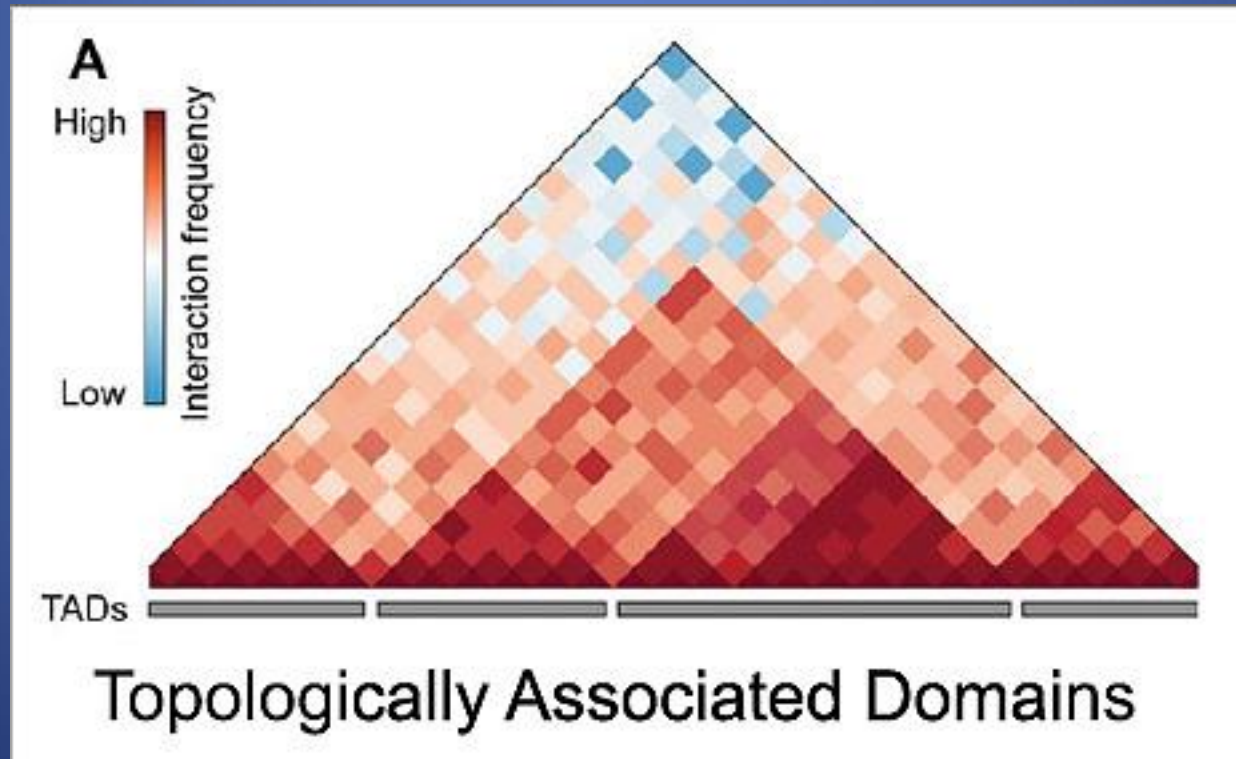
# Hi-C

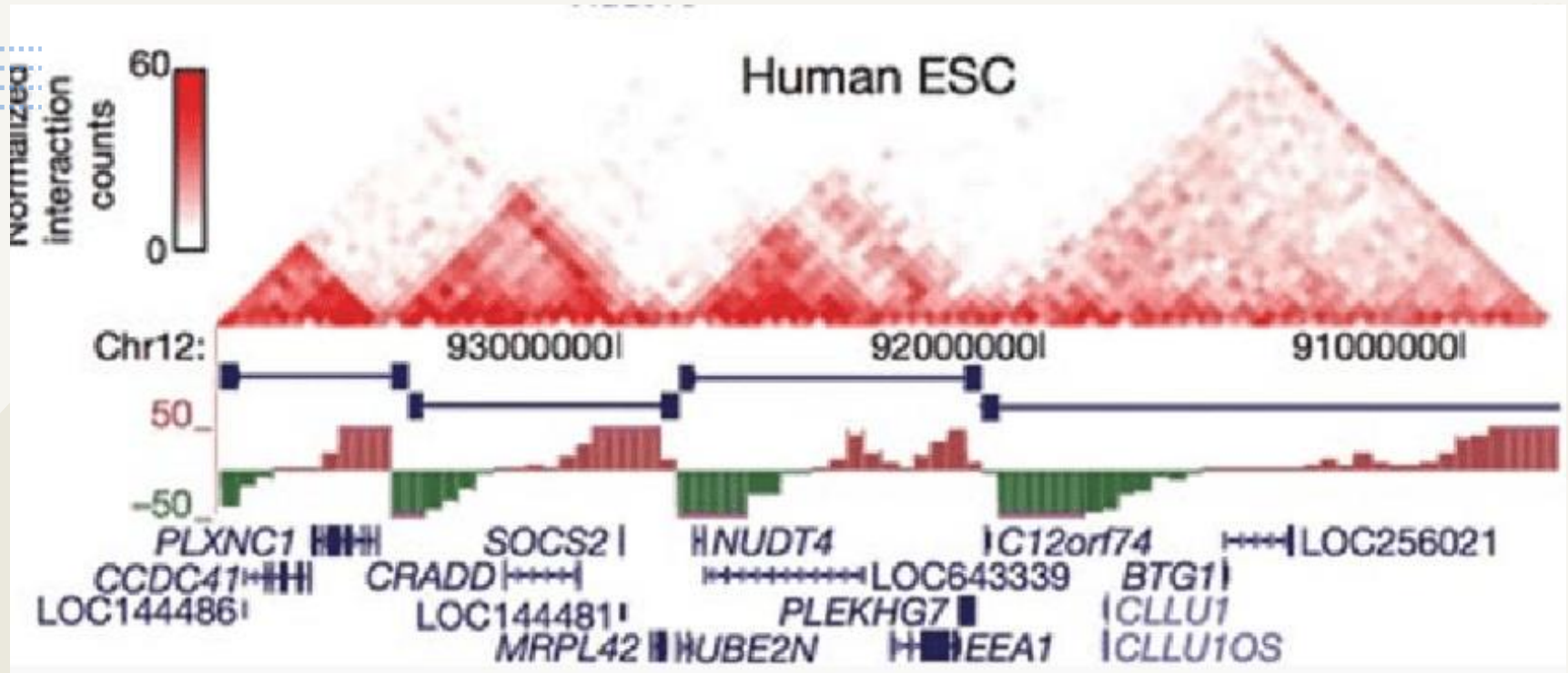
- There are many assays depending on whether you're investigating one SNP versus one gene, one SNP versus all genes or all SNPs versus all genes.



# Interaction Diagram

- Interactions are depicted by interaction diagram.





# TADs

- Associations tend to separate into regions where a lot of interactions happen, and few happen across them.
- These are called Topologically Associated Domains.

# Fine Mapping

- Chromatin Conformation Assays and eQTLs are just two of many tools used to get at this problem.
- CHIP-Seq assays also play a major role.
- Most every tool in the toolbox comes to bear on this problem.
- We won't have time to go into any more detail on this, but it's another highly marketable skill.

