Understanding the genetics of complex traits I

Gavin Band gavin.band@well.ox.ac.uk

BA Human Sciences

Friday 27th Feb 2026

Main lecture messages

1. Most human phenotypes are highly heritable
(a large proportion of variation is due to genetics)

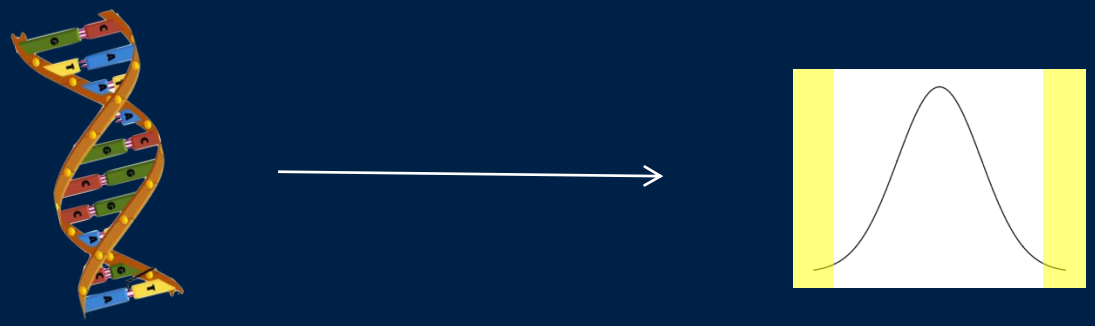
2. But many 'complex' traits are *not* mendelian - they are polygenic - *"common variant, common trait" hypothesis*

3. The discovery of this fact is due to ***genome-wide association studies*** (GWAS), the first of which was conducted in the mid 2000s.
We will go into this in some detail – methodology, population genetics, GWAS in practice

4. But understanding the biology can be hard…

The human genome is ~3.2 billion base pairs long.

About 1 in 100 – 1000 of those bases vary between people.



What proportion of phenotypic variation is due to genetic variation?

Human traits are highly heritable

We don't have to guess!

Idea: if genetics determines a trait, then *more genetically similar individuals should have more similar phenotypes.*

We can estimate how much genetics determines trait variation by comparing trait similarity in more genetically similar and less genetically similar individuals, such as monozygotic and dizygotic twins.

Meta-analysis of the heritability of human traits based on fifty years of twin studies

Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]

(2015)

Large meta-analysis of > 2000 twin studies

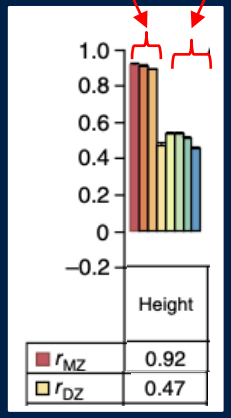
(Browse the results at: https://match.ctglab.nl)

Human traits are highly heritable

Idea: if genetics determines a trait, then *more genetically similar individuals should have more similar phenotypes.*

MZ Twins r~0.92

DZ Twins r~0.47



All studied traits

Compare trait correlations between twins.

(Adult) height is *much* more similar between monozygotic than dizygotic twins. **The heritability** is about 90%.

Heritability is the proportion of trait variation explained by inherited factors (including genetics) . Can be estimated as $h^2 \approx 2 \times (r_{MZ} - r_{DZ})$.

Human traits are highly heritable

If genetics determines a trait, then *more genetically similar individuals should have more similar phenotypes.*

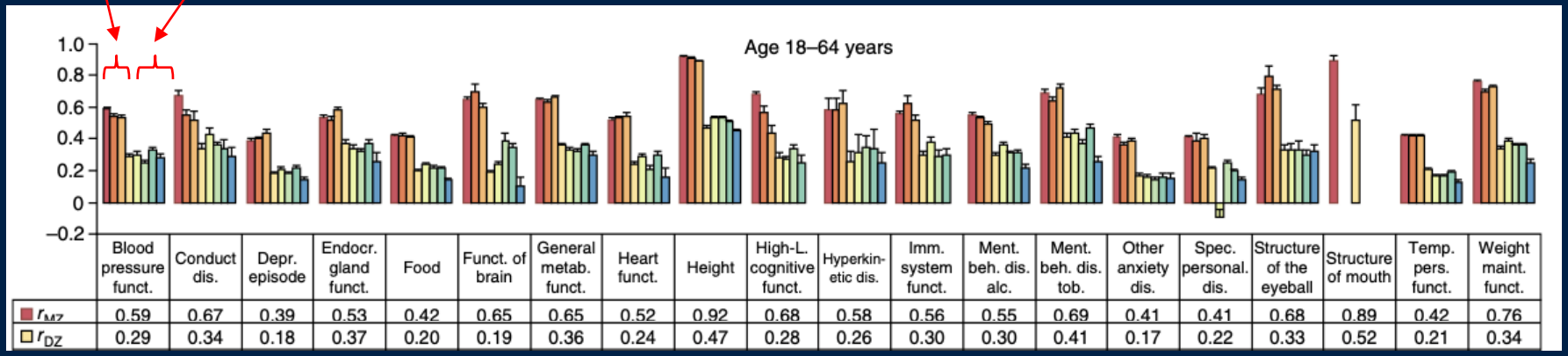
Meta-analysis of the heritability of human traits based on fifty years of twin studies

Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de Leeuw[1,3], Patrick F Sullivan[4–6], Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthuma[1,9,11]

(2015)

Monozygotic

Dizygotic



	Blood pressure funct.	Conduct dis.	Depr. episode	Endocr. gland funct.	Food	Funct. of brain	General metab. funct.	Heart funct.	Height	High-L. cognitive funct.	Hyperkin-etic dis.	Imm. system funct.	Ment. beh. dis. alc.	Ment. beh. dis. tob.	Other anxiety dis.	Spec. personal. dis.	Structure of the eyeball	Structure of mouth	Temp. pers. funct.	Weight maint. funct.
r_{MZ}	0.59	0.67	0.39	0.53	0.42	0.65	0.65	0.52	0.92	0.68	0.58	0.56	0.55	0.69	0.41	0.41	0.68	0.89	0.42	0.76
r_{DZ}	0.29	0.34	0.18	0.37	0.20	0.19	0.36	0.24	0.47	0.28	0.26	0.30	0.30	0.41	0.17	0.22	0.33	0.52	0.21	0.34

Blood pressure
$h^2 \approx 60\%$

Depression
$h^2 \approx 42\%$

Adult height
$h^2 \approx 90\%$

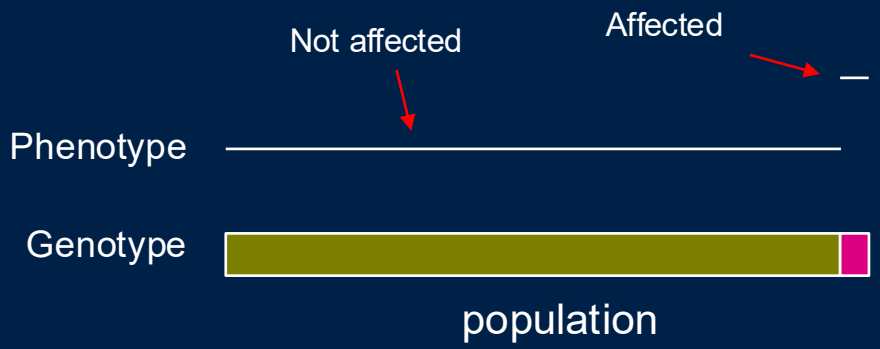
"Higher level cognitive function"
$h^2 \approx 80\%$

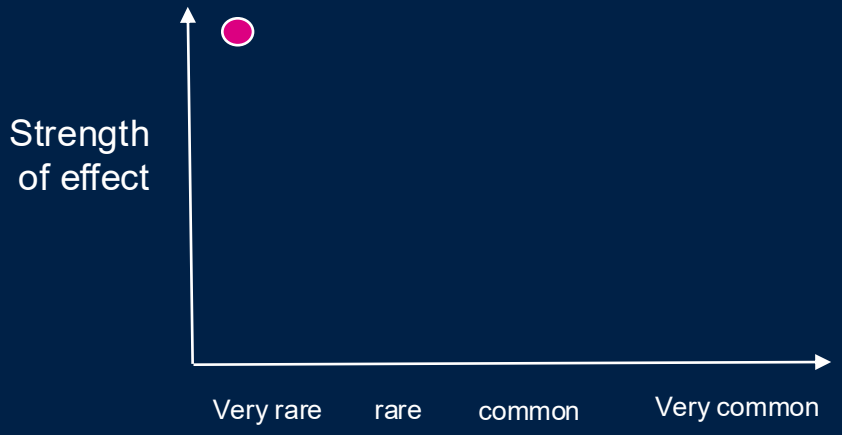
Structure of the eyeball
$h^2 \approx 70\%$

Lots of theoretical caveats might apply here – see Lecture 1. But in general it is true that *a large proportion of variation in most human phenotypes is caused by genetics*.

(Browse the results at: https://match.ctglab.nl)

Two possible extreme genetic architectures





Example: Huntingdon's

Cell, Vol. 72, 971–983, March 26, 1993, Copyright © 1993 by Cell Press

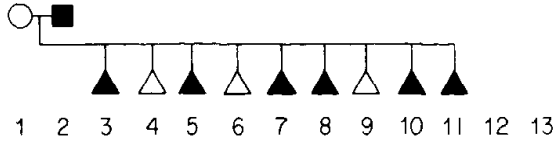
A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes

The Huntington's Disease Collaborative Research Group*

Introduction

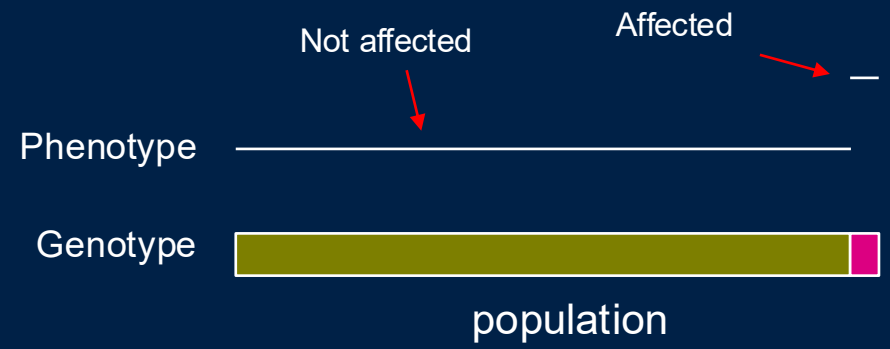
Affects ~1 in 20,000 people of European ancestry
(less in Africa and Asia)

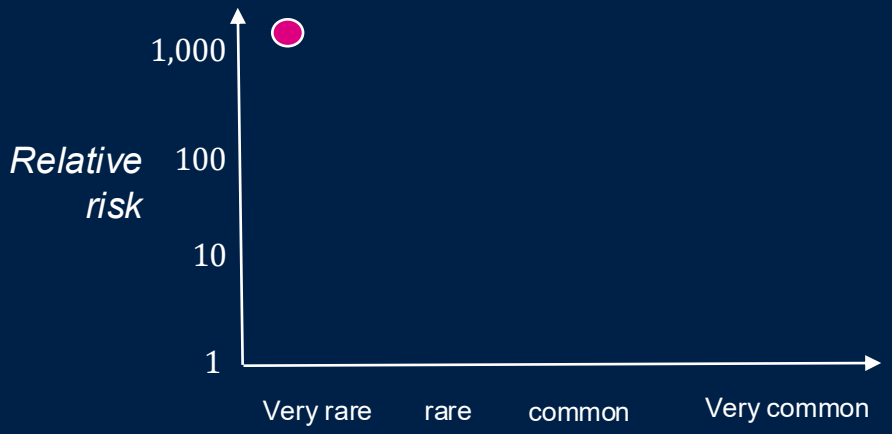
Discovered by looking in families



A **"Mendelian"** trait

Two possible extreme genetic architectures





Example: Huntingdon's

Cell, Vol. 72, 971–983, March 26, 1993, Copyright © 1993 by Cell Press

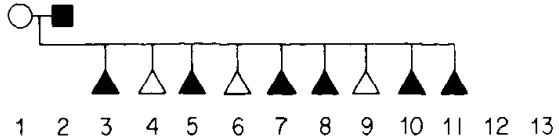
A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes

The Huntington's Disease Collaborative Research Group*

Introduction

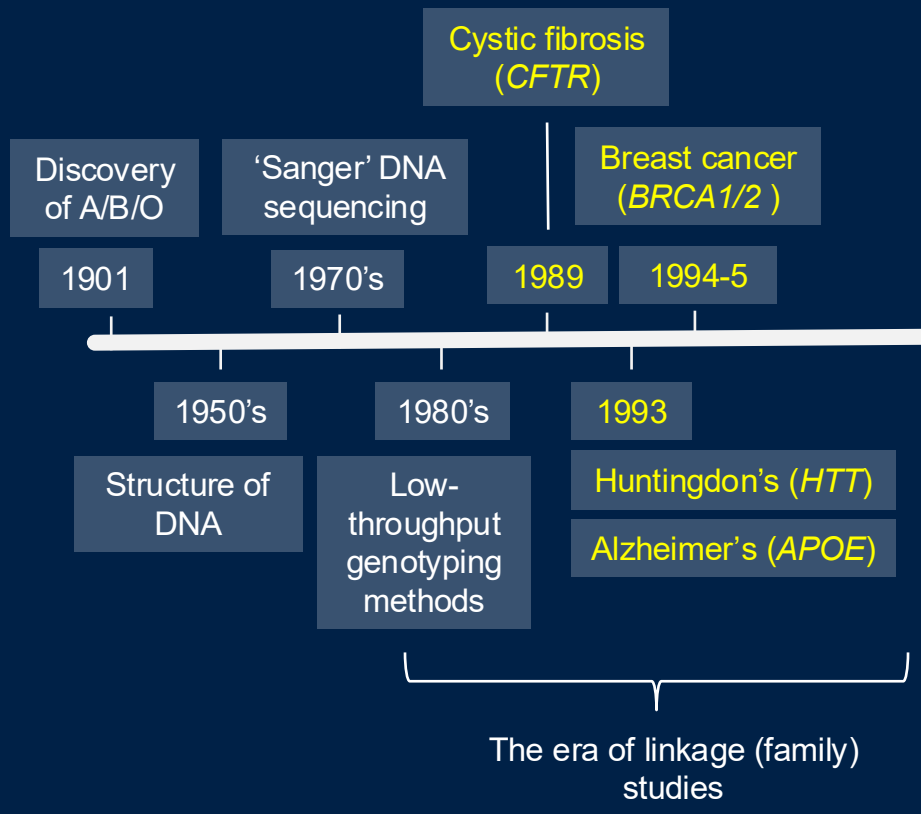
Affects ~1 in 20,000 people of European ancestry
(less in Africa and Asia)

Discovered by looking in families



A **"Mendelian"** trait

End of an era

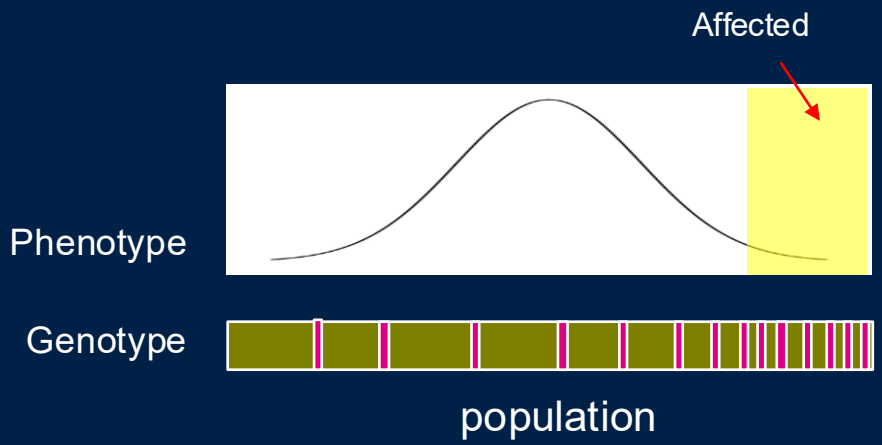


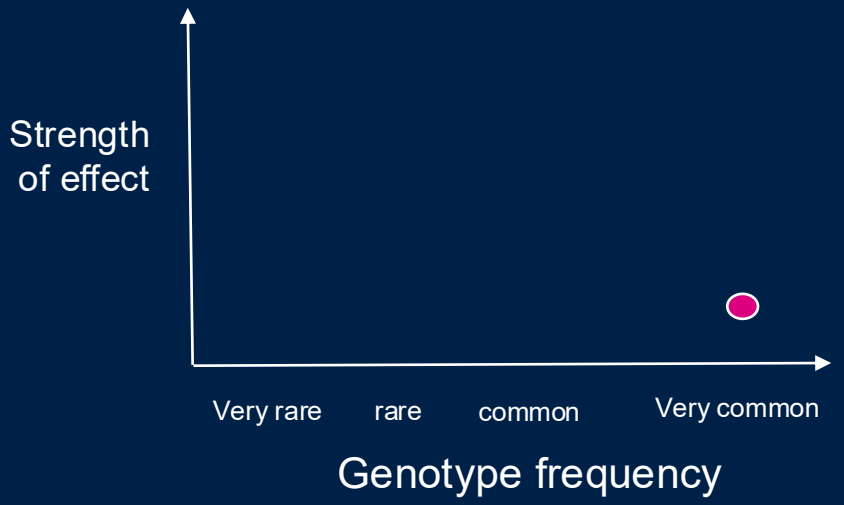
"Linkage Mapping was successful in identifying the genetic basis of many human diseases in which the disease penetrance resembles a simple Mendelian model e.g. Huntington's disease, Cystic Fibrosis, some forms of breast cancer, Alzheimers, …"

"…but the literature is now replete with linkage screens for an array of common 'complex' disorders such as schizophrenia, manic depression, autism, asthma, type I and type II diabetes, Multiple Sclerosis, Lupus. Although many of these studies have reported significant linkage findings, none has lead to convincing replication"

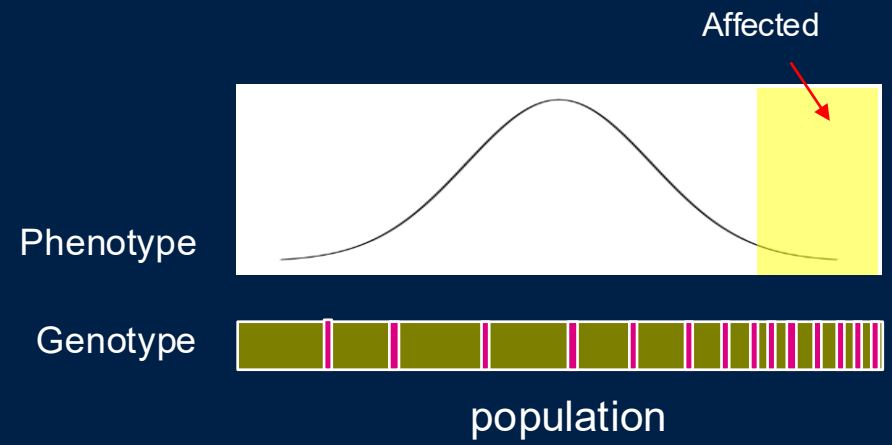
– Risch "*Searching for genetic determinants in the new millennium*" Nature (2000)

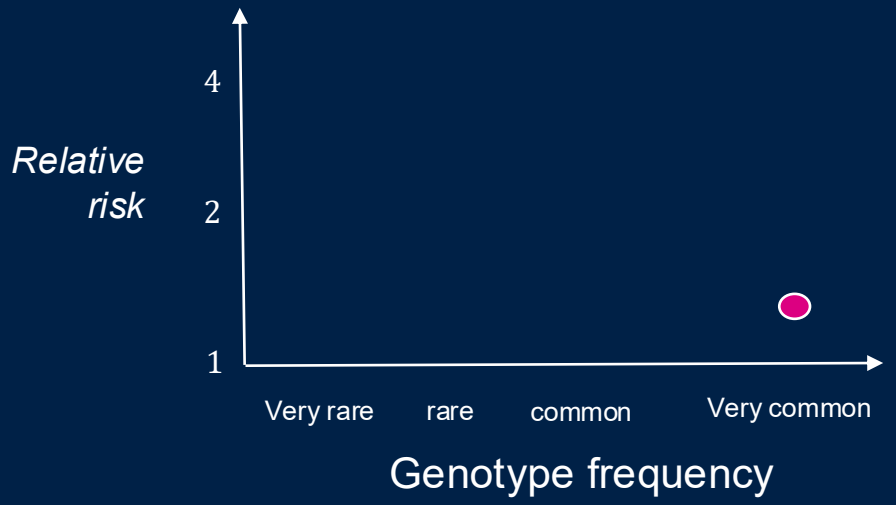
Common variant, common disease hypothesis



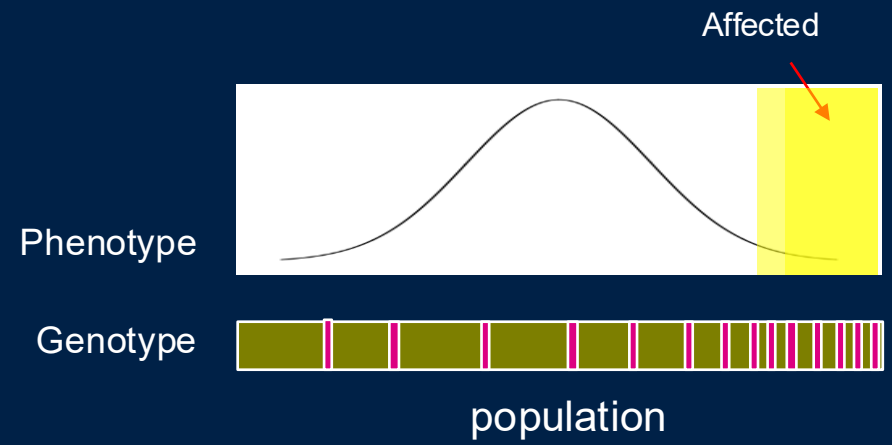


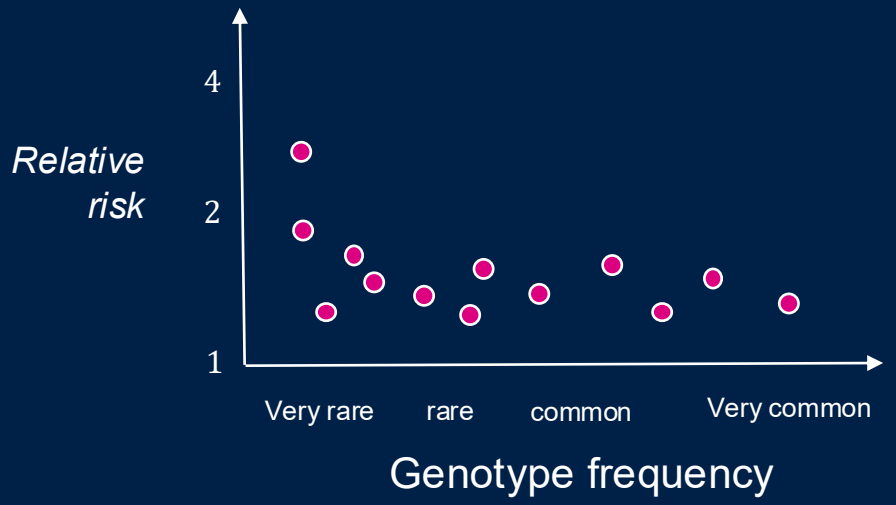
Common variant, common disease hypothesis





Common variant, common disease hypothesis





A **complex trait**.
Caused by many factors, each having a small overall effect. Including

- Many genetic variants, including common ones
- Environmental factors
- Gene-environment or gene-gene interactions
- ...

Summary

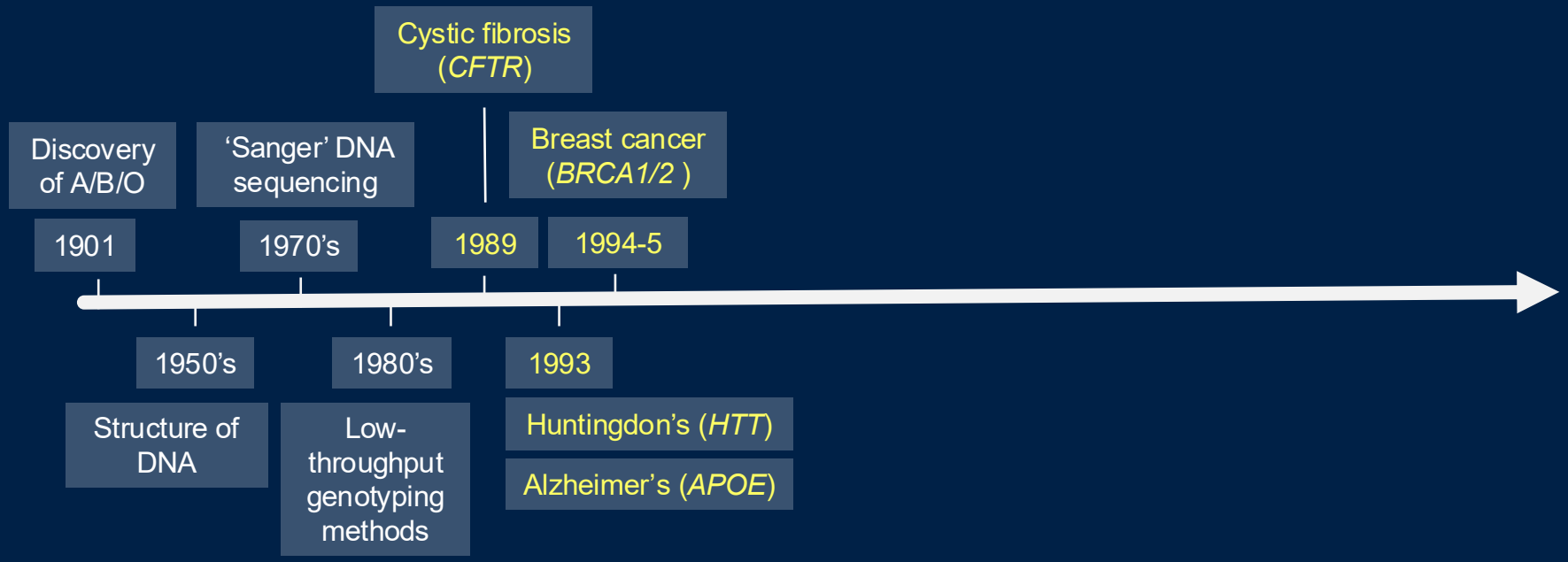
- Most human phenotypes are highly heritable - a large proportion of phenotype variation seems to be caused by genetics. ~60% on average!

- In principle this heritability could occur in different ways – for example through single variants with strong effects, or through multiple variants with small effects.

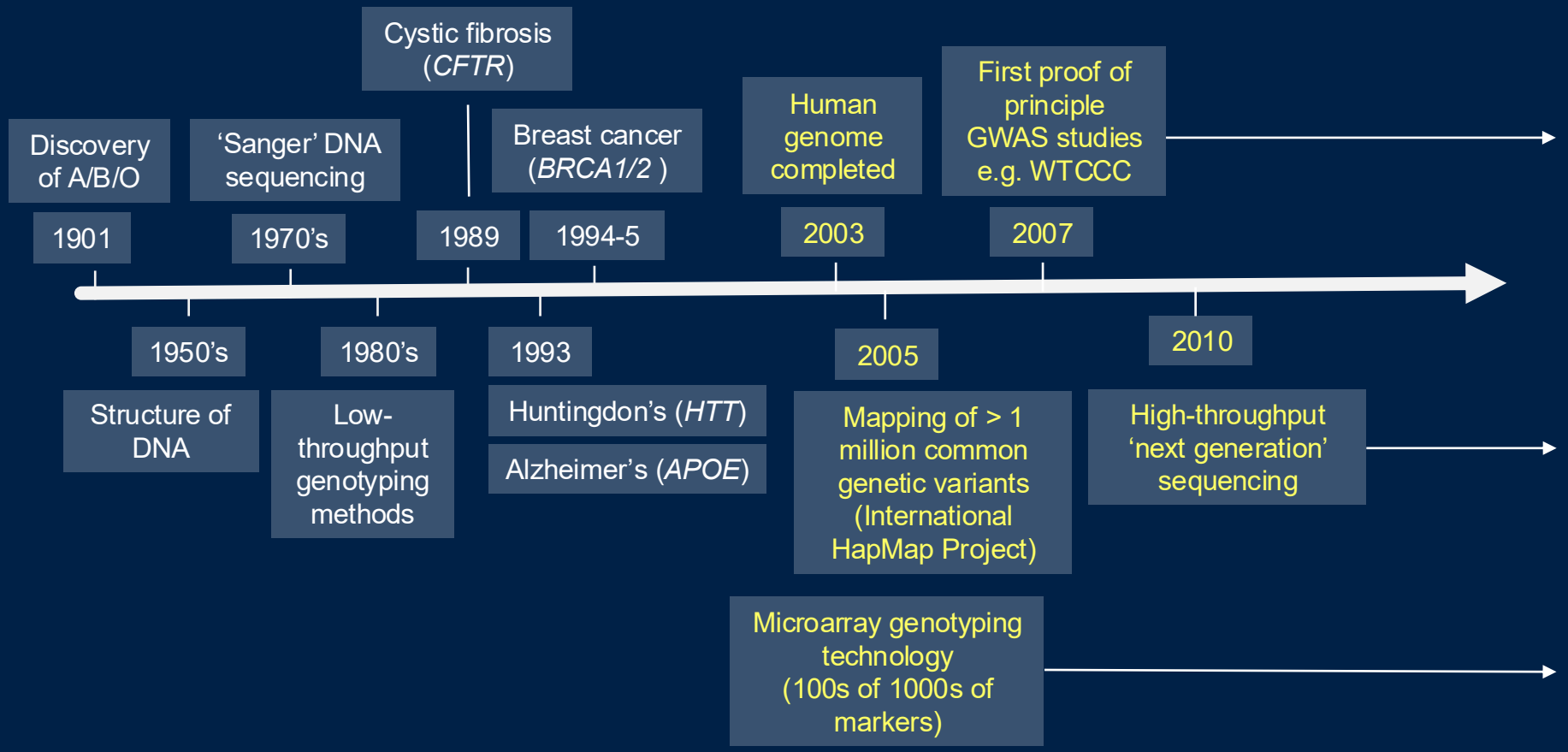
- By the 2000s family studies had identified the causes of several mendelian traits, but had failed to solve the genetics of multiple complex diseases.

Was the “common variant, common disease“ hypothesis true?

End of the linkage era



The birth of GWAS



Rest of lecture - outline

- Testing for association

How to do the statistical tests? How many samples? How to deal with confounders?

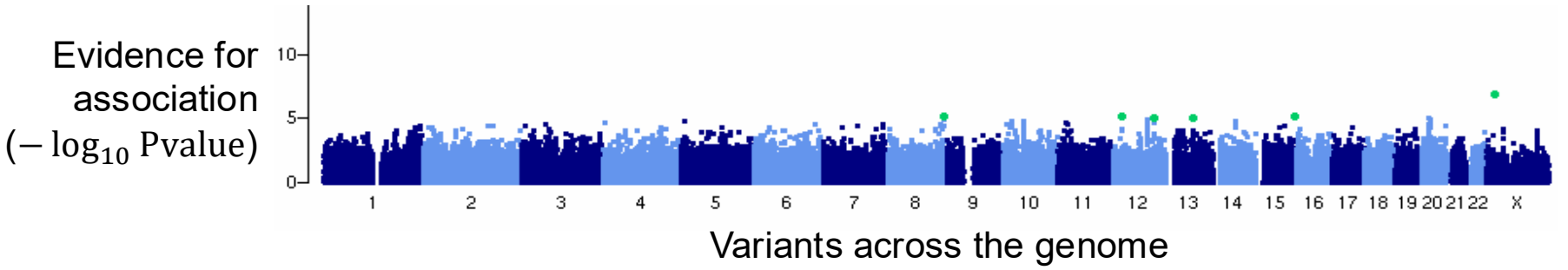
- What variants to genotype, and how?

Patterns of linkage disequilbrium and the HapMap study

- A real GWAS study – WTCCC

The common variant, common disease hypothesis in practice

What we are aiming for:



Results of a statistical test of association between each each genetic variant across the genome and a trait of interest.

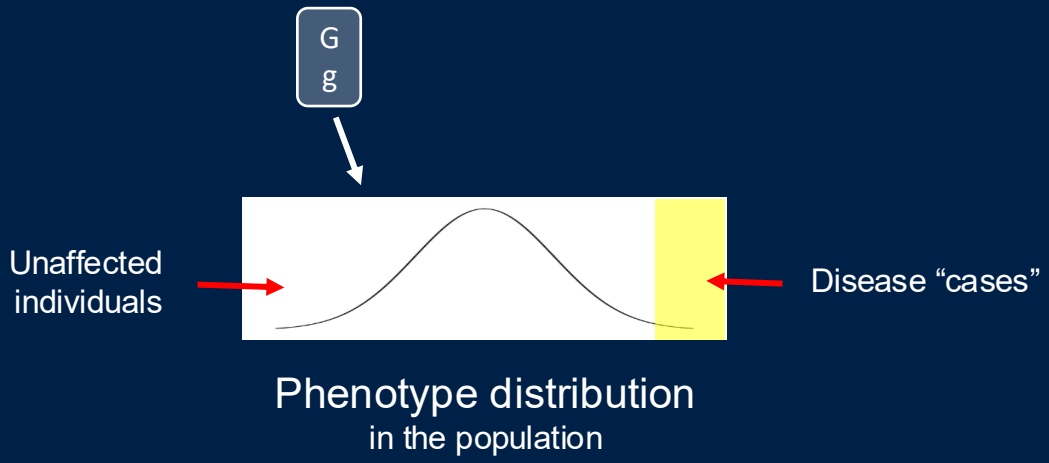
Higher values (lower P-values) are better.

If $P < 5 \times 10^{-8}$ or so, we might get excited

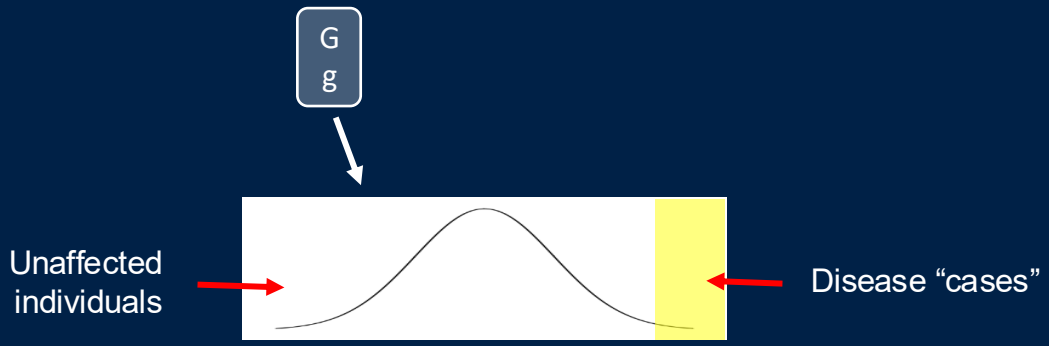
because maybe that variant causes disease?

Testing for association

Imagine a genetic variant that affects risk of disease



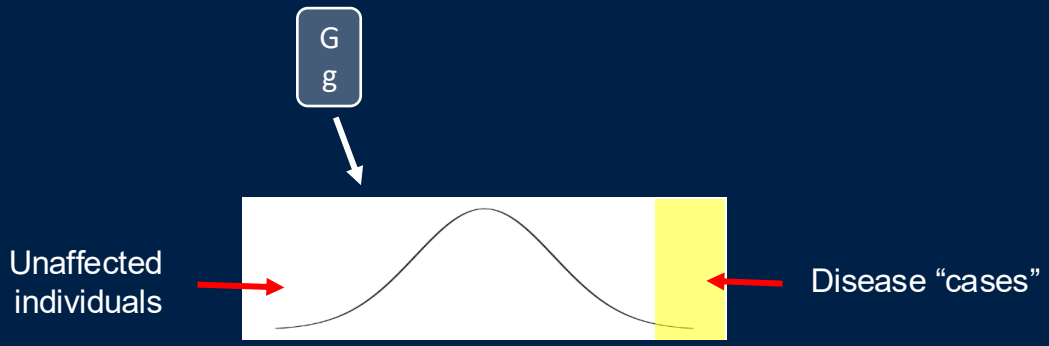
Testing for association



If genotype *G* causes disease, then carrying *G* will make you more likely to have disease.

$$\underline{\textit{Relative risk}} = \frac{\text{“Chance/frequency of disease given genotype } G\text{”}}{\text{“Chance/frequency of disease given genotype } g\text{”}} > 1$$

Testing for association



If genotype *G* causes disease, then carrying *G* will make you more likely to have disease.

$$\underline{\textit{Relative risk}} = \frac{P(\text{disease}|G)}{P(\text{disease}|g)} > 1$$

Using probability notation

If the genotype causes disease, then the relative risk will be different from 1

How to estimate relative risk?

$$RR = \frac{P(\text{disease}|\, G)}{P(\text{disease}|\, g)}$$

(in population)

Disease frequencies given genotype

How to estimate relative risk?

$$RR = \frac{P(\text{disease}|\, G)}{P(\text{disease}|\, g)} = \frac{P(G|\text{disease})}{P(g|\text{disease})} \times \frac{P(g)}{P(G)}$$

(in population)

Disease frequencies given genotype

Genotype frequencies in cases and controls

To estimate the relative risk, we just need to **measure the genotypes in some disease cases and population controls.**

(Note: apply Bayes' theorm)

How to estimate relative risk?

$$RR = \frac{P(\text{disease}|\, G)}{P(\text{disease}|\, g)} = \frac{P(G|\text{disease})}{P(g|\text{disease})} \times \frac{P(g)}{P(G)}$$

(in population)

Disease frequencies given genotype

Genotype frequencies in cases and controls

To estimate the relative risk, we just need to **measure the genotypes in some disease cases and population controls.**

	G	***g***
Disease cases:	a	b
Controls*:	c	d

$$OR = \frac{a}{b} \times \frac{d}{c}$$

(in sample)

Key fact

	G	*g*
Disease cases:	a	b
Population controls*:	c	d

$$OR = \frac{a}{b} \times \frac{d}{c}$$

(in a sample of disease cases and population controls)

The odds ratio **in a sample of cases and controls* estimates the population relative risk.**

Stricyly this applies to 'population controls', but also approximately true for 'true' disease controls, as long as the disease is not too common.

Example: O blood group and severe malaria

Cases were ascertained as children arriving in hospital with severe symptoms compatible with malaria & parasitaemia, in a hospital in Kilifi, eastern Kenya. Controls were ascertained from new births in the same hospitals.

	O	non-O
Severe malaria cases	686	843
Controls:	839	700

N=3,068 samples

MalariaGEN 2019 doi: 10.1038/s41467-019-13480-z

Can you compute the odds ratio?

Example: O blood group and severe malaria

Cases were ascertained as children arriving in hospital with severe symptoms compatible with malaria & parasitaemia, in a hospital in Kilifi, eastern Kenya. Controls were ascertained from new births in the same hospitals.

	O	non-O
Severe malaria cases	686	843
Controls:	839	700

N=3,068 samples

MalariaGEN 2019 doi: 10.1038/s41467-019-13480-z

$$OR = \frac{686}{843} \times \frac{700}{839} = 0.68$$

Example: O blood group and severe malaria

Cases were ascertained as children arriving in hospital with severe symptoms compatible with malaria & parasitaemia, in a hospital in Kilifi, eastern Kenya. Controls were ascertained from new births in the same hospitals.

	O	non-O
Severe malaria cases	686	843
Controls:	839	700

$$OR = \frac{686}{843} \times \frac{700}{839} = 0.68$$

N=3,068 samples

MalariaGEN 2019 doi: 10.1038/s41467-019-13480-z

People with O blood group get severe malaria at ~70% of the rate of people without.

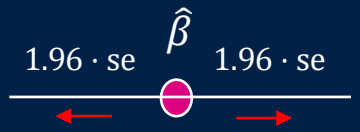
(Could say: *"O blood group is associated with ~30% reduced risk of severe malaria."*)

How much **statistical evidence** is there that this is a real effect?

The key association test summary statistics

Effect size estimate $\hat{\beta} = \log(\text{OR})$

Standard error



95% confidence interval

"How much uncertainty?"

P-value*

*"How **unlikely** was such a big estimate, if actually there was **no true effect**?"*

A **small p-value** means the effect is **unlikely to be zero**

*In practice P-value is computed from the beta and standard error:

$$P = \Phi^{-1}\left(\frac{\log(\text{OR})}{se}\right)$$

Normal distribution function

Incredibly useful formula

Fact: the standard error is largely determined by the study design.

Here is a very useful formula which approximates it in the 2x2 table example:

$$\text{Standard error}(\log OR) \approx \frac{1}{\sqrt{N \times f(1-f) \times \phi(1-\phi)}}$$

N = sample size = $a + b + c + d$

f = frequency of G allele

ϕ = proportion of cases

Note: this example is a recessive effect of O blood group. Use $2N$ instead if testing an additive effect

The standard error depends on **sample size**, **frequency**, and **case/control ratio.**

It gets smaller (at rate $\frac{1}{\sqrt{N}}$) as the sample size increases.

Example: O blood group is associated with malaria protection

	O	non-O
Severe malaria cases	686	843
Controls:	839	700

$N = 3{,}068$
$f \approx 0.55$
$\phi \approx 0.5$

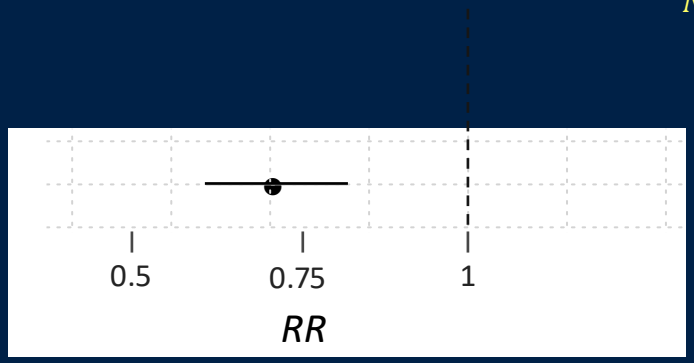
$$OR = \frac{686}{843} \times \frac{700}{839} = 0.68$$

i.e. $\log(OR) \approx -0.386$

$$\text{Standard error}(\log OR) \approx \frac{1}{\sqrt{3068 \times \underbrace{0.45 \times 0.55}_{f(1-f)} \times \underbrace{0.5^2}_{\phi(1-\phi)}}} \approx 0.073$$

(N)

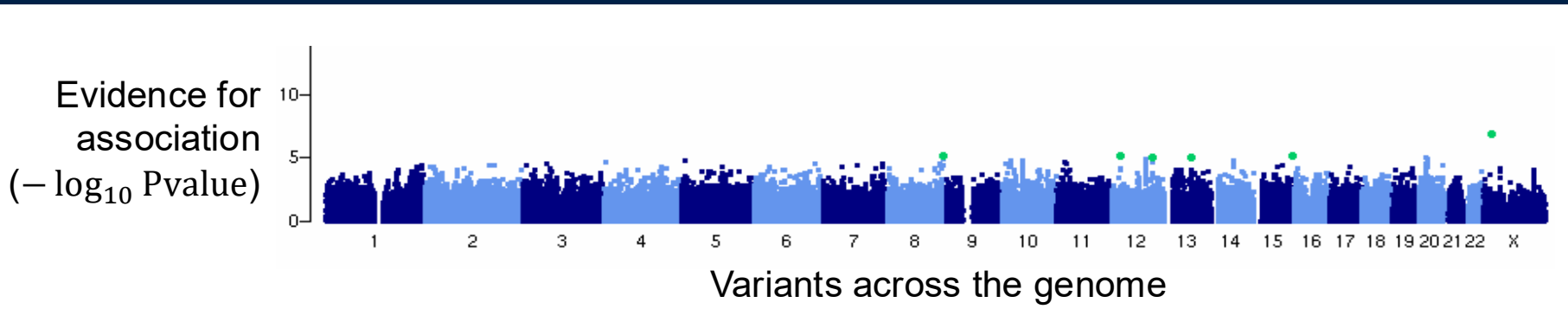
(on log scale)



Estimated relative risk = **0.68**
se = **0.07**
95% CI = 0.59-0.78 (estimate +/- 1.96 standard errors)

Estimate is about 5 standard errors from zero

$$P = 9.6 \times 10^{-8}$$



However, remember we are going to do this for **millions** of variants

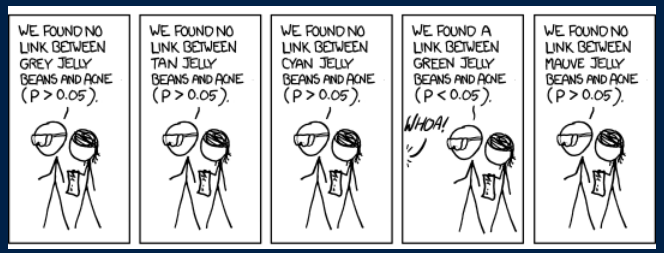
Most of them not well-known variants like O blood group!

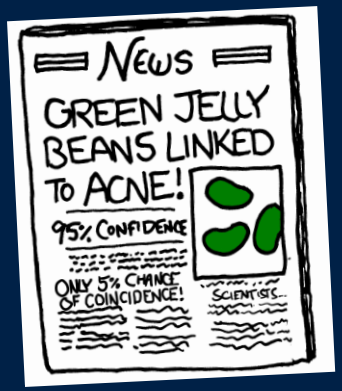
Each one has very little 'prior' chance of being associated

Lots of statistical tests so to get excited we need strong evidence

$P < 5 \times 10^{-8}$ is a commonly-used threshold.

https://xkcd.com/882/





How many samples did we need anyway?

…to have a chance of such small P-values?

$$\text{Standard error}(\log OR) \approx \frac{1}{\sqrt{N \times f(1-f) \times \phi(1-\phi)}}$$

N = sample size = $a + b + c + d$

f = frequency of G allele

ϕ = proportion of cases

The standard error depends on **sample size**, **frequency**, and **case/control ratio.**

It gets smaller (at rate $\frac{1}{\sqrt{N}}$) as the sample size increases.

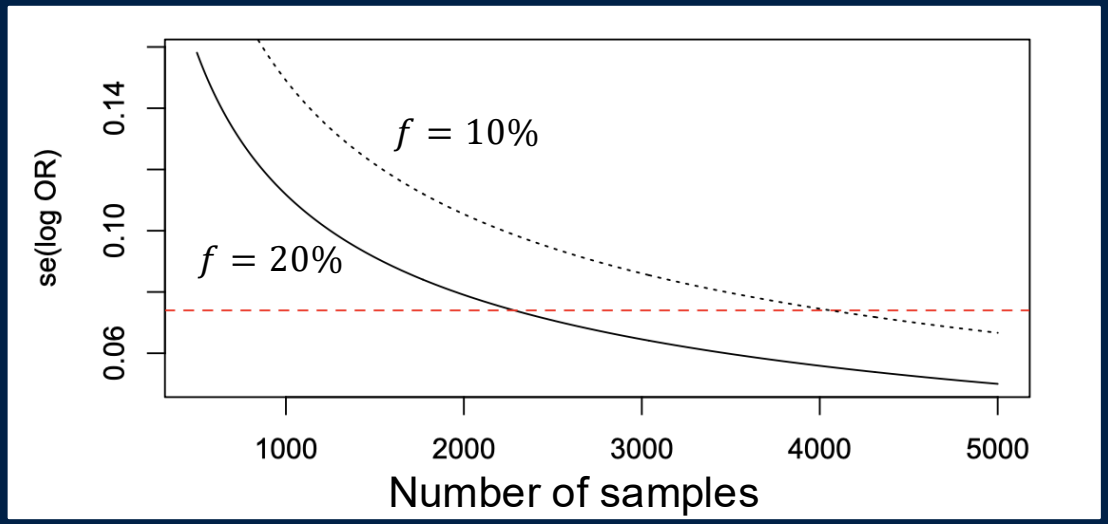
How many samples did we need anyway?

E.g. suppose the variant we're looking for has frequency $f = 20\%$ and the effect size is $RR = 1.5$.

$P = 5 \times 10^{-8}$ corresponds to an effect about 5.5 standard errors from zero, so very roughly we need a standard error at least as small as

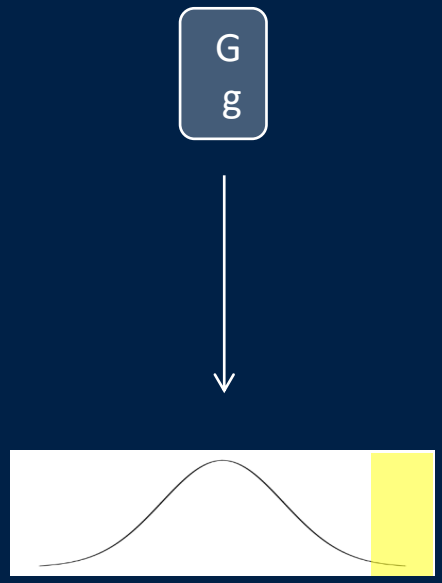
$$\frac{log(1.5)}{5.5} \approx 0.07$$

$$\text{se}(\log OR) \approx \frac{1}{\sqrt{2N \times f(1-f) \times 0.5^2}}$$



Answer: we need thousands!

Major gotcha! Confounders

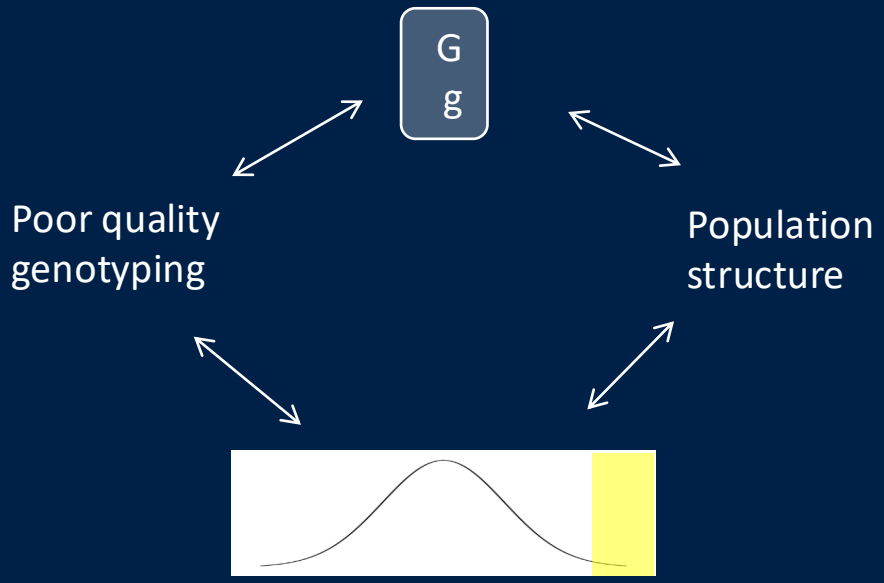


$P < 5 \times 10^{-8}$, yay!

So you've found an association

Can you really conclude it is the genetic variant causing the disease?

Major gotcha! Confounders



So you've found an association with a small P-value

Can you really conclude it is the genetic variant causing the disease?

Answer: not if it is confounded.

$\Rightarrow$ All good GWAS studies pay careful attention to possible confoudners including **population structure** and **genotyping quality**

Association testing in practice

In practice would use a **regression method*** (instead of the simple 2x2 table) to make these estimates:

Can control for possible confounders like measures of population structure

Also more flexible - can also model different modes of inheritance – e.g. additive, dominant or recessive effects.

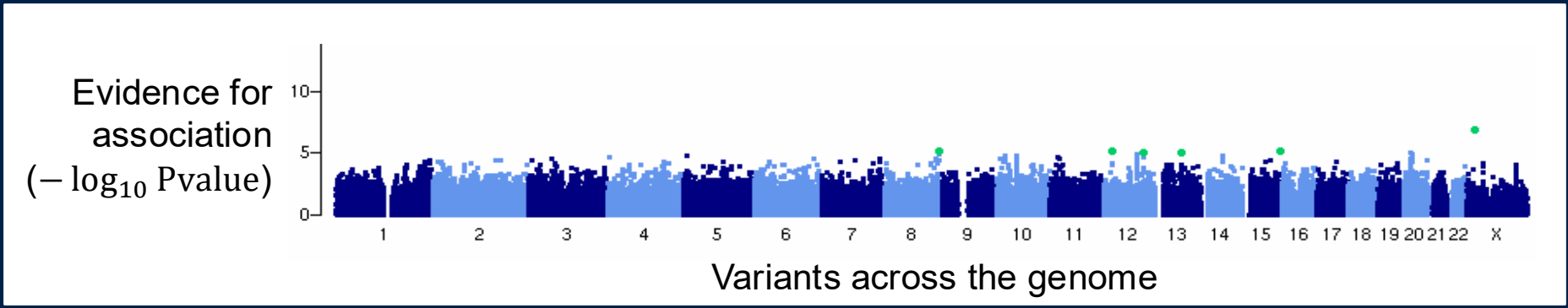
*E.g logistic regression (for case/control traits) or linear regression for continuous traits.

GWAS association testing – putting it together

1. Collect as many samples as possible

2. Genotype at as many variants across the genome as possible

3. Run a statistical test for genotype-phenotype association



If strong evidence e.g. $P < 5 \times 10^{-8}$ => maybe we've found our disease-causing variant

GWAS association testing – put together

1. Collect as many samples as possible

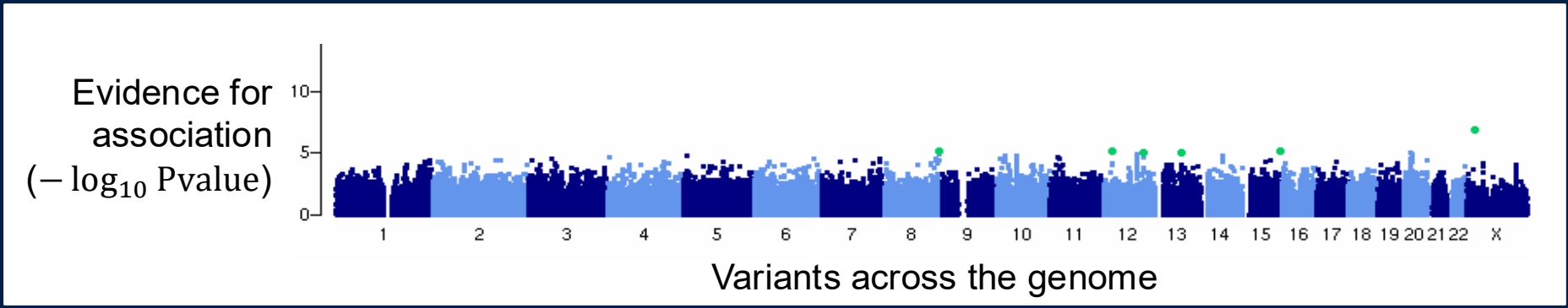
How many samples?

2. Genotype at as many variants across the genome as possible

How to deal with confounders?
Genotype quality control!

3. Run a statistical test for genotype-phenotype association

How to do the stats test?
Can we deal with confounders?



If strong evidence e.g. $P < 5 \times 10^{-8}$ => maybe we've found our disease-causing variant

Rest of lecture - outline

- Testing for association

How to do the statistical tests? How many samples? How to deal with confounders?

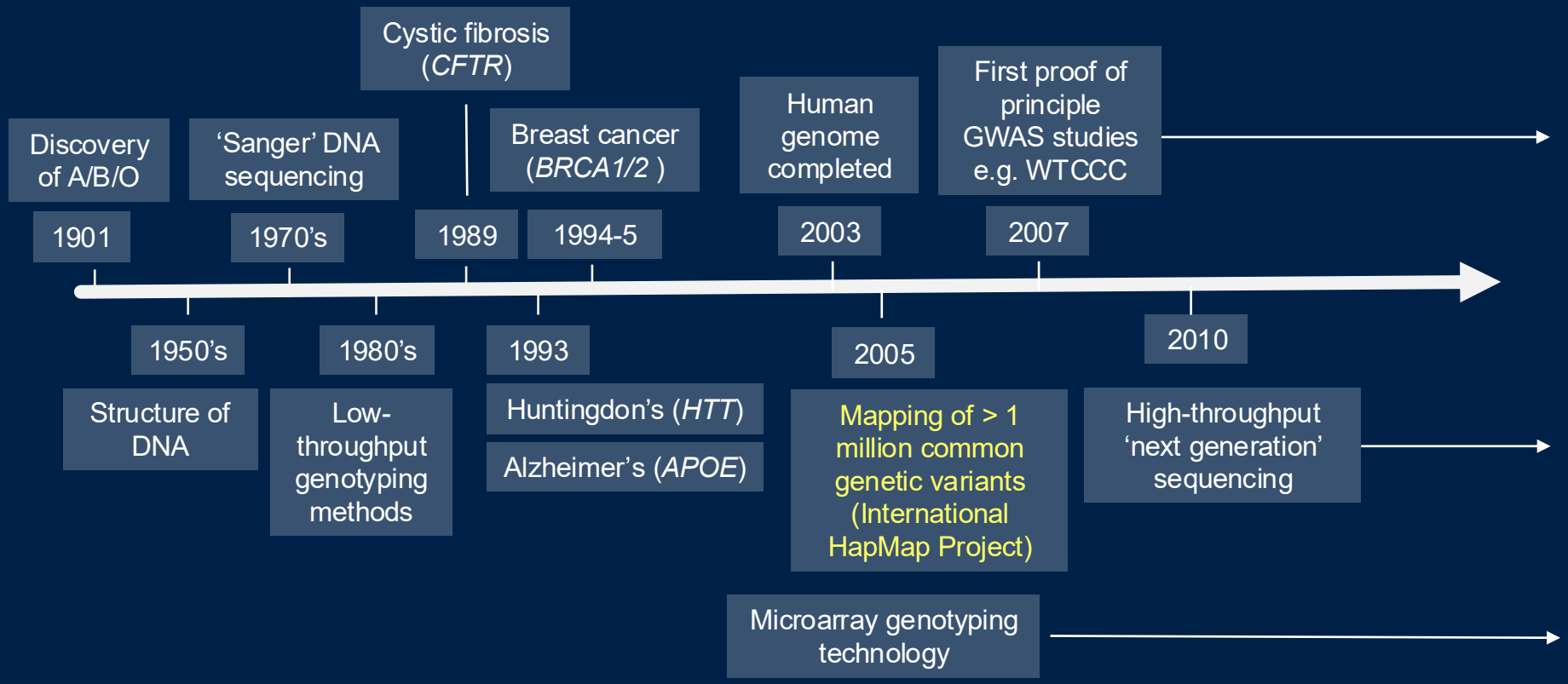
- What variants to genotype, and how?

Patterns of linkage disequilbrium and the HapMap study

- A real GWAS study – WTCCC

The common variant, common disease hypothesis in practice

The birth of GWAS



Microarrays developed in the late 90's / early 2000's.
For the first time was possible to rapidly type hundreds of thousands or millions of SNPs

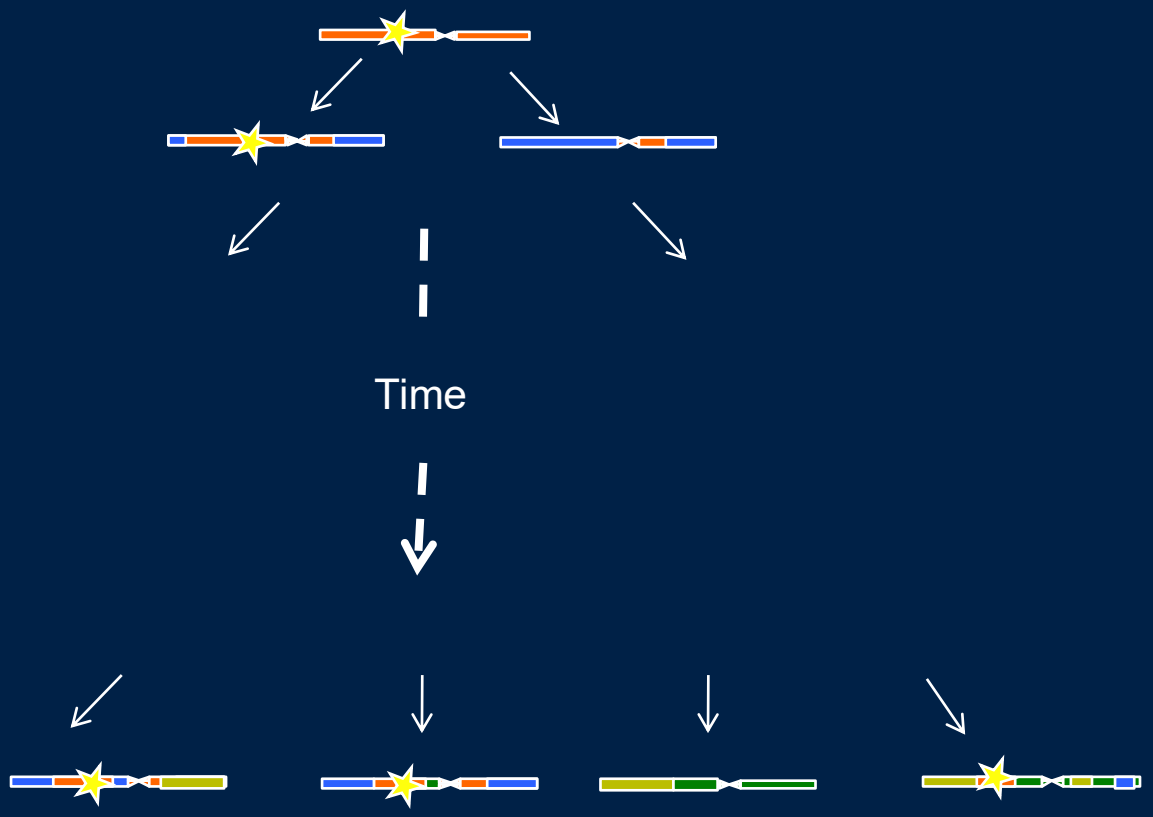
Patterns of inheritance generate linkage disequilibrium*

Mutation arises

Gets passed on through many generations

Changes in frequency cause variants to become correlated (LD)

Recombination breaks this down leading to local patterns



** LD = correlations between genotypes at nearby genetic variants along a chromosome*

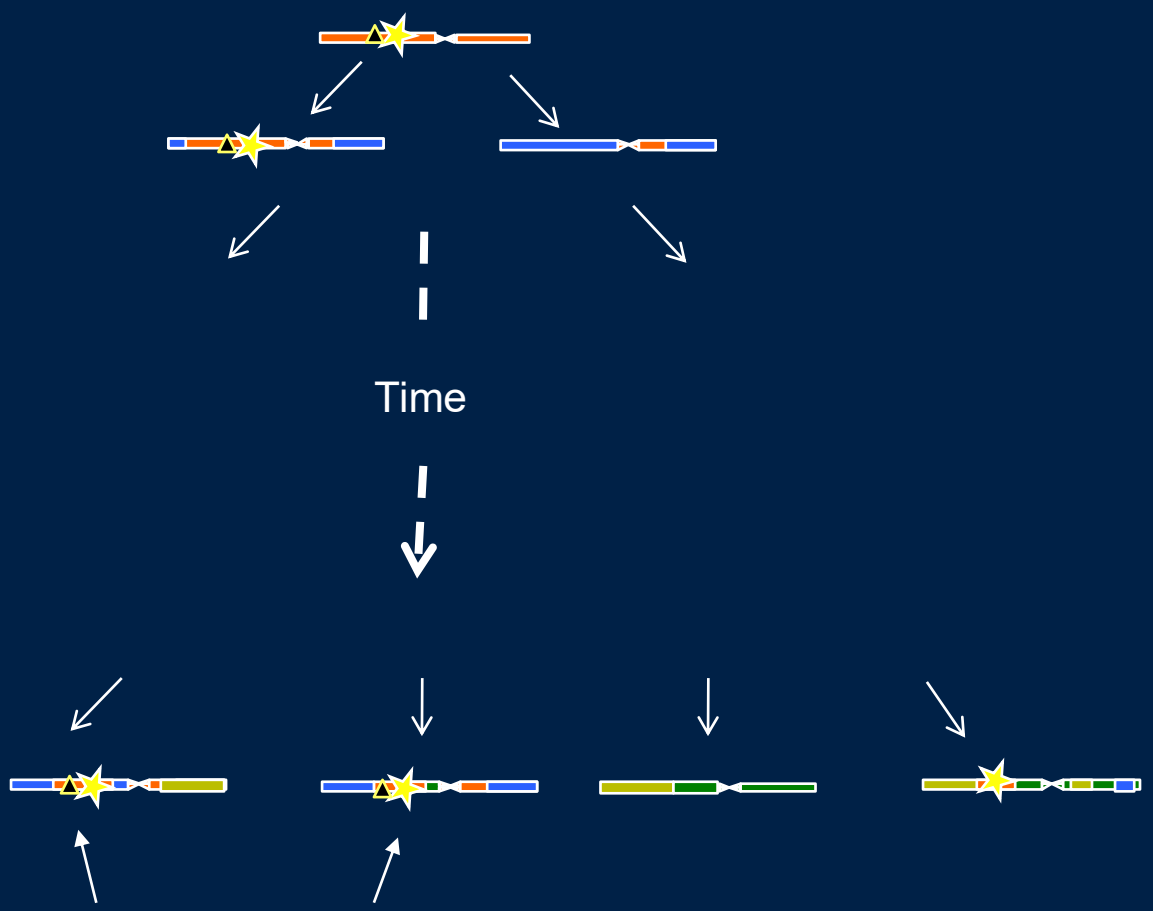
Patterns of inheritance generate linkage disequilibrium*

Mutation arises

Gets passed on through many generations

Changes in frequency cause variants to become correlated (LD)

Recombination breaks this down leading to local patterns



These variants are correlated (tend to be seen together).

** LD = correlations between genotypes at nearby genetic variants along a chromosome*

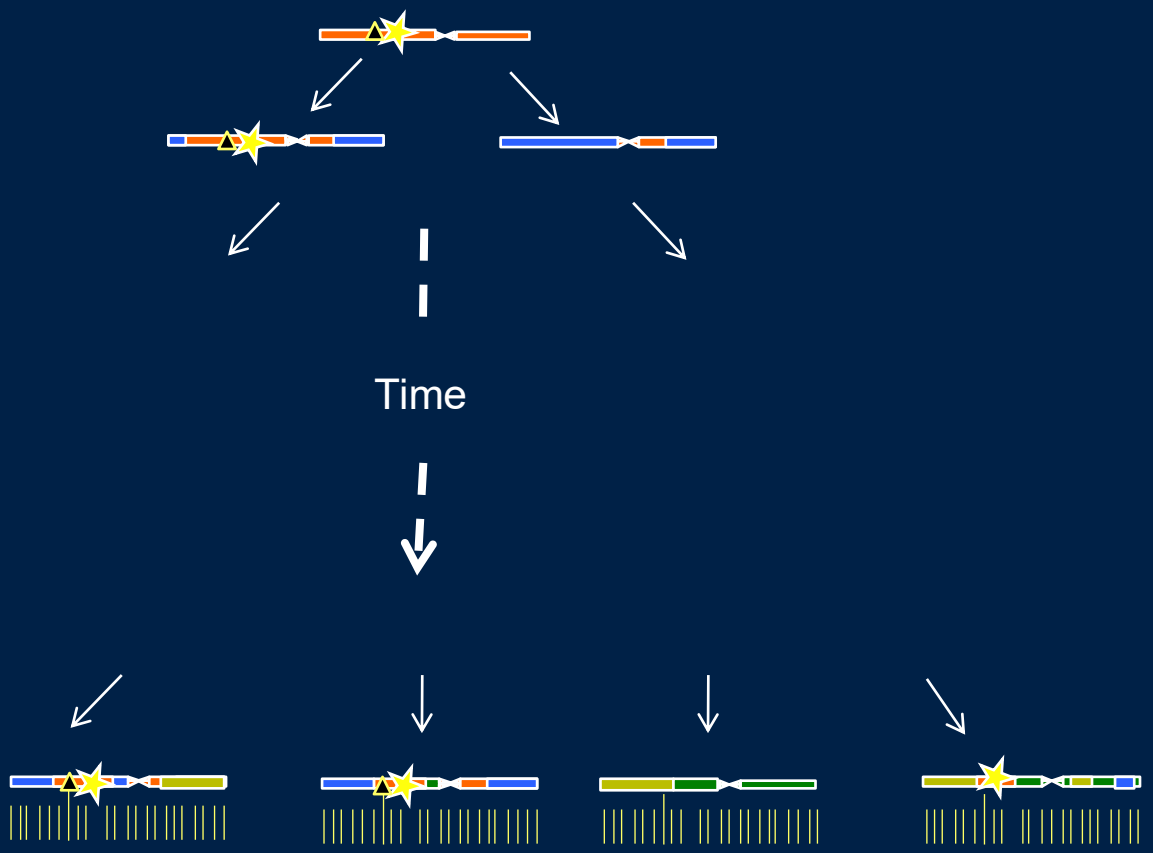
Patterns of inheritance generate linkage disequilbrium

Mutation arises

Gets passed on through many generations

Changes in frequency cause variants to become correlated (LD)

Recombination breaks this down leading to local patterns

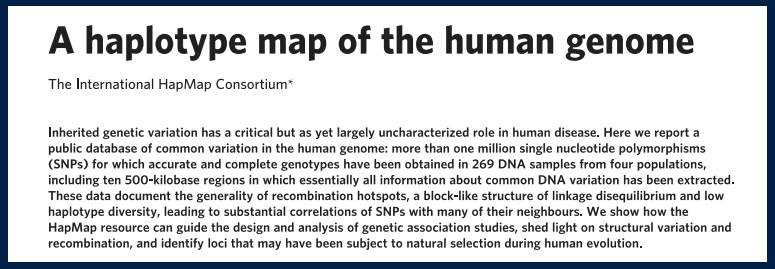


Idea: maybe we can just genotype a dense set of marker genotypes

If we happened to include △ , we might pick up the true signal at ✶

The HapMap project estimated LD

The extent of LD depends on the amount of recombination.



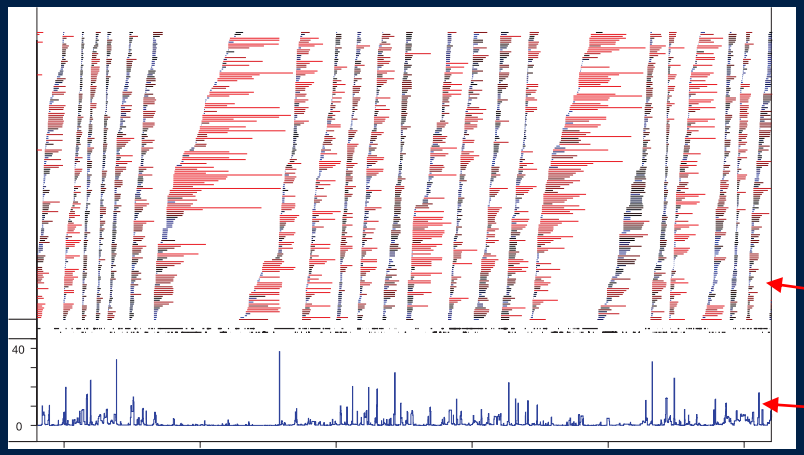
A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

International HapMap Project
doi:10.1038/nature0422 (2005)

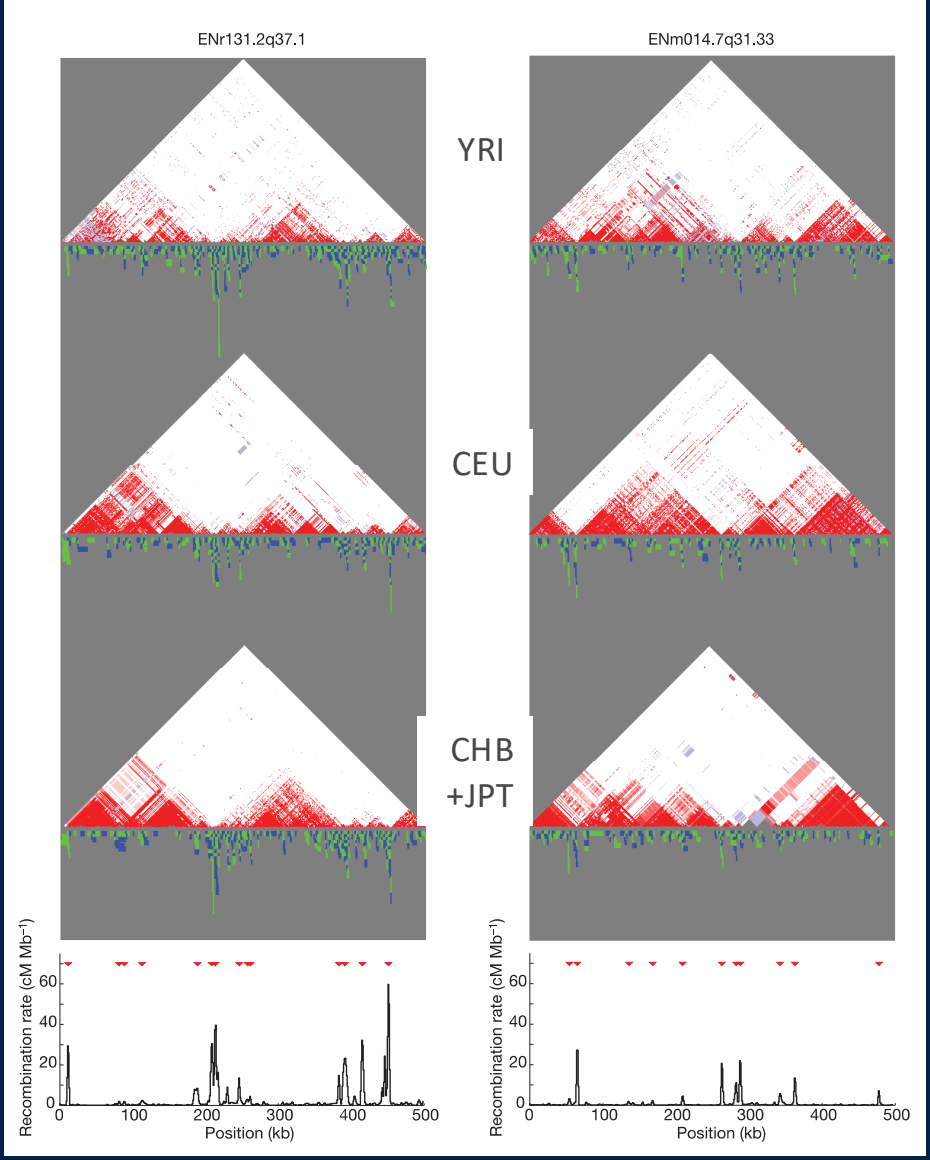
A database of > 1M SNPs found in European, African, and Asian ancestry individuals
(A subset of the samples later used in the 1000 Genomes Project)



Recombination turns out to be highly nonuniform. It is concentrated in *recombination hotspots*. So mutations are carried on longer haplotypes than had been expected.

Shared haplotype lengths

Map of recombination rate



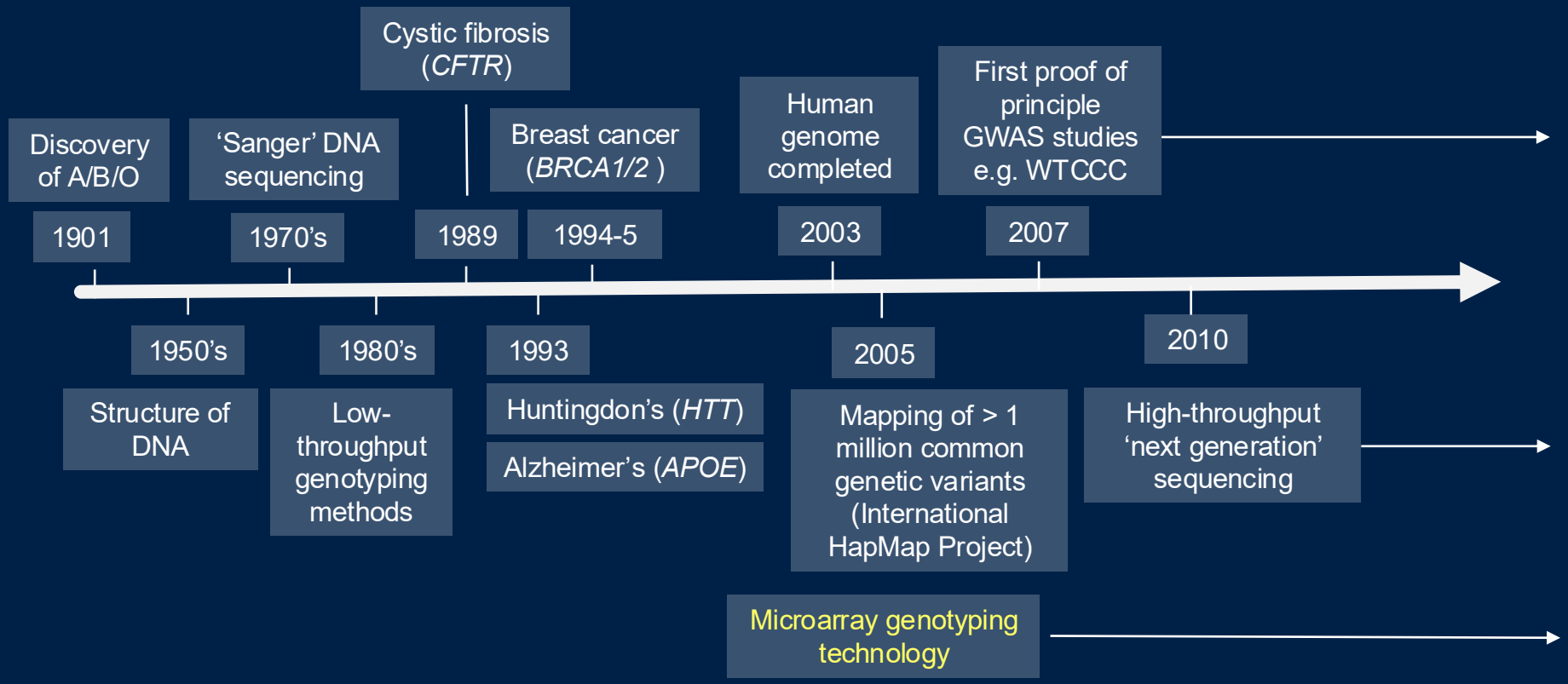
Block-like structure of LD (correlations between SNPs in two different regions)

Tag SNP set size	Common SNPs captured (%)		
	YRI	CEU	CHB + JPT
10,000	12.3	20.4	21.9
20,000	19.1	30.9	33.2
50,000	32.7	50.4	53.6
100,000	47.2	68.5	72.2
250,000	70.1	94.1	98.5

As in Table 7, tag SNPs were picked to capture common SNPs in HapMap release 16c1 using Haploview, selecting SNPs in order of the fraction of sites captured. Common SNPs were captured by fixed-size sets of pairwise tags at $r^2 \geq 0.8$.

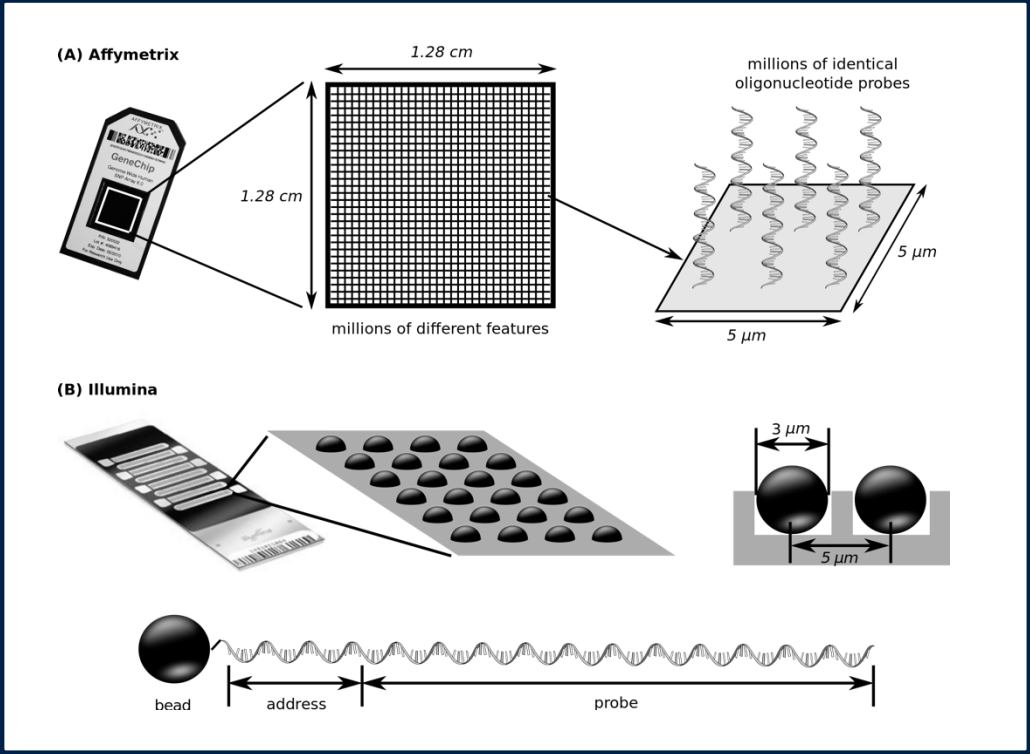
HapMap estimated how many SNPs genome-wide would need to be typed to capture (by LD) most common genetic variants. E.g. 250,000 would capture ~95% of SNPs in European populations.

The birth of GWAS



Microarrays developed in the late 90's / early 2000's.
For the first time was possible to rapidly type hundreds of thousands or millions of SNPs

How a microarray works

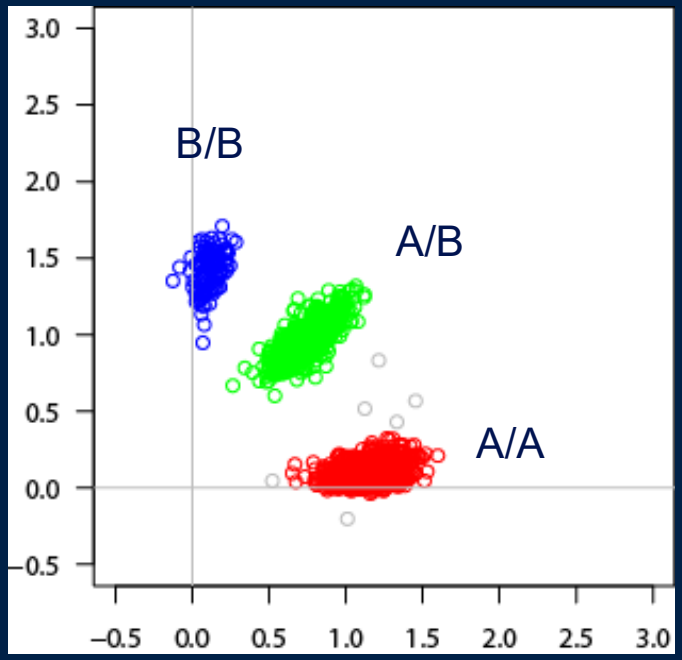


Wash the DNA over and let it hybridise to millions of probes – one for each SNP

Flourescent markers are then attached. A picture is taken of the array.

A microarray gives you intensities, not genotypes

For each (well-genotyped) SNP, you get back this:

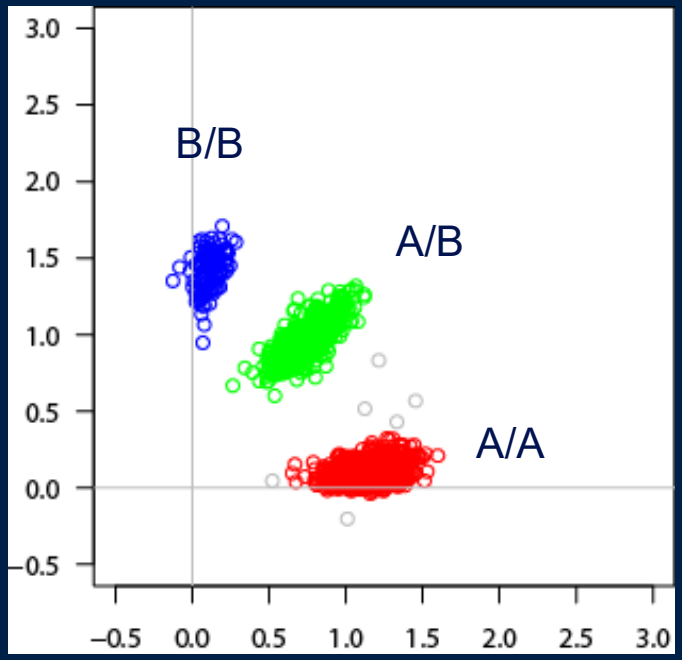


A clustering algorithm has been used to turn the intensity values (x/y axis values) into genotype calls (colours).

Each dot represents DNA from one individual.
X axis = image intensity for 1st allele
Y axis = image intensity for 2nd allele

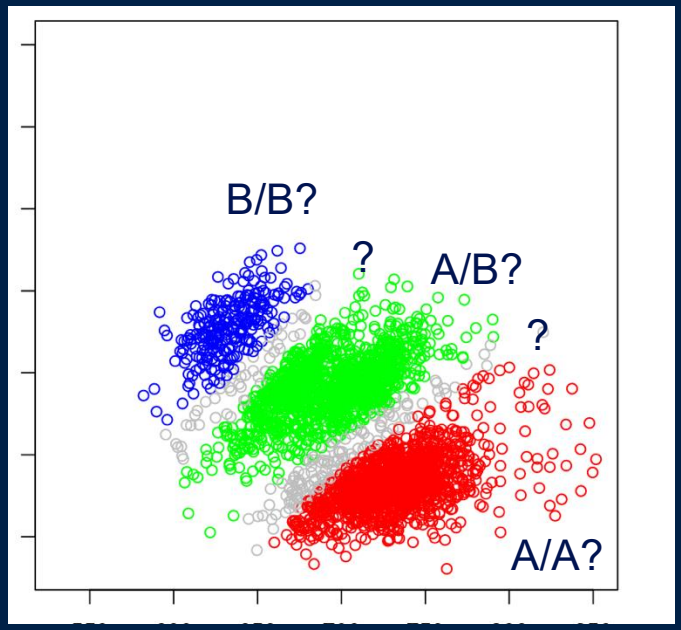
A microarray gives you intensities, not genotypes

For each SNP, you get back this:



Each dot represents DNA from one individual.
X axis = image intensity for 1st allele
Y axis = image intensity for 2nd allele

Or this if you're less lucky:



Small genotyping errors in cases or controls could easily confound the study

Careful quality control needed with these technologies

Rest of lecture - outline

- Testing for association

How to do the statistical tests? How many samples? How to deal with confounders?

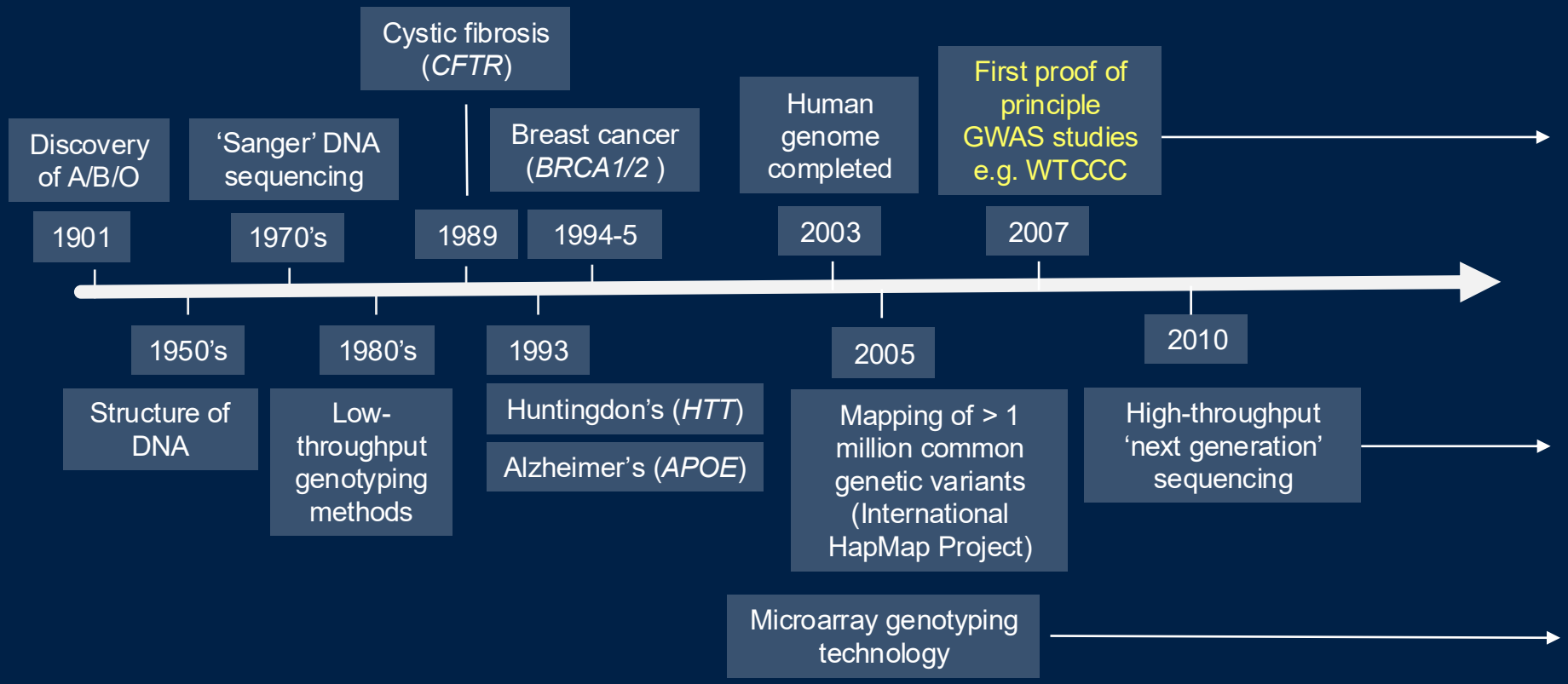
- What variants to genotype, and how?

Patterns of linkage disequilbrium and the HapMap study

→ • A real GWAS study – WTCCC

The common variant, common disease hypothesis in practice

The birth of GWAS



Microarrays developed in the late 90's / early 2000's.
For the first time was possible to rapidly type hundreds of thousands or millions of SNPs

Anatomy of a GWAS – what to look for

1. Collect as many cases and controls as possible

What samples How many?

2. Genotype (or impute) them at as many variants across the genome as possible

How many?

3. Deal with potential confounders – careful data quality control and handle population structure.

Have they done adequate data quality control? Have they dealt with possible confounders?

4. Estimate relative risks, and look for statistical evidence that of $RR \neq 1$

5. If estimate is many standard deviations from zero, bingo! We may have found a true causal effect.

Did they find anything with enough evidence?

6.Replicate in other studies, or find other corroborating evidence?

Is it corroborated by other evidence?

7. (Now try to understand the underlying biology.)

Can they understand the biology?

A real GWAS study - WTCCC

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The Wellcome Trust Case Control Consortium*

Nature (2007)

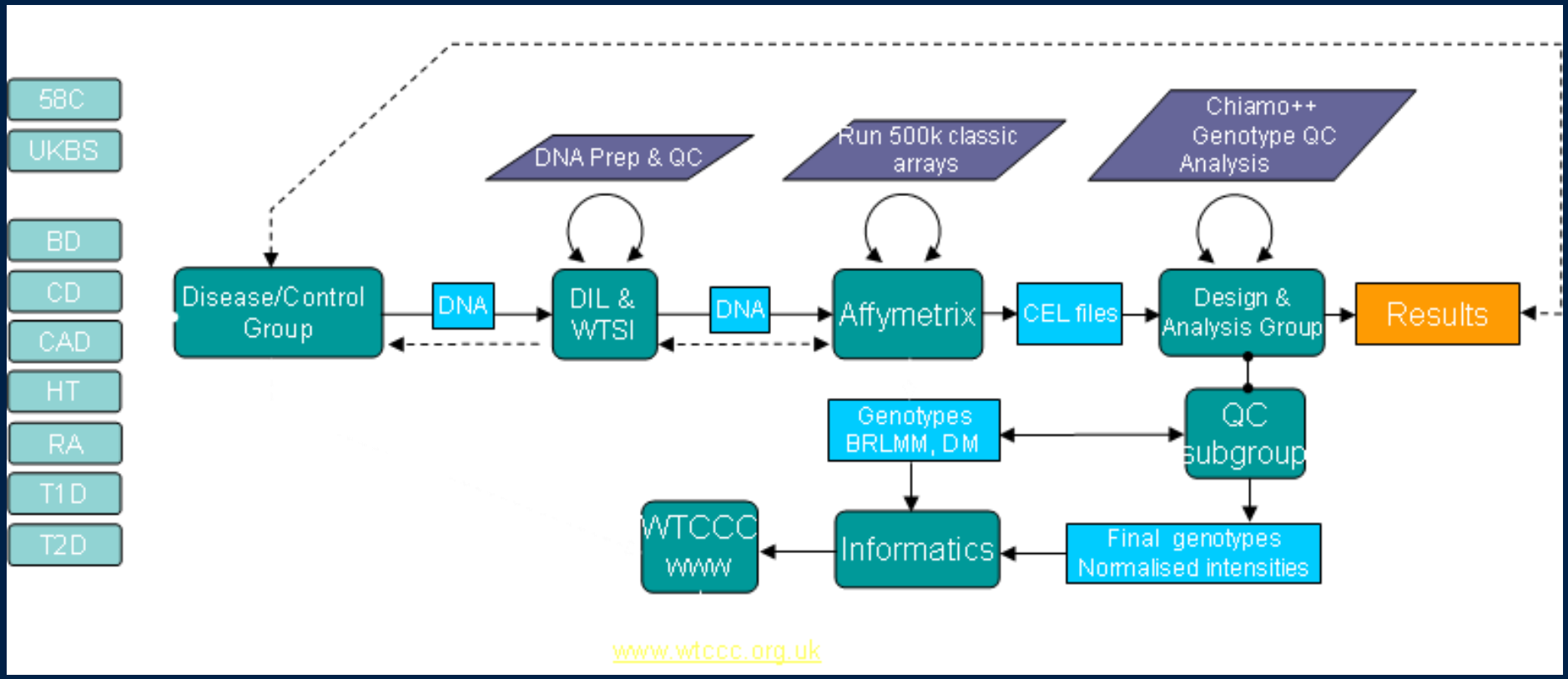
Studied seven common diseases in the UK

Bipolar disorder, Coronary Artery Disease, Crohn's disease, Hypertension, Rheumatoid arthritis, Type 1 and Type 2 Diabetes

Genotyped at 500,000 SNPs across the genome

doi:10.1038/nature05911

A real study - WTCCC



Anatomy of a GWAS – what to look for

1. Collect as many cases and controls as possible

N=2,000 cases and 3,000 controls

2. Genotype (or impute) them at as many variants across the genome as possible

Genotyped at 500k SNPs

3. Deal with potential confounders – careful data quality control and handle population structure.

4. Estimate relative risks, and look for statistical evidence that of $RR \neq 1$

5. If estimate is many standard deviations from zero, bingo! We may have found a true causal effect.

6. Does it replicate in other studies, or have other corroborating evidence?

7. (Now try to understand the underlying biology.)

Anatomy of a GWAS – what to look for

1. Collect as many cases and controls as possible

2. Genotype (or impute) them at as many variants across the genome as possible

3. Deal with potential confounders – careful data quality control and handle population structure.

4. Estimate relative risks, and look for statistical evidence that of $RR \neq 1$

5. If estimate is many standard deviations from zero, bingo! We may have found a true causal effect.

6. Does it replicate in other studies, or have other corroborating evidence?

7. (Now try to understand the underlying biology.)

Have they done adequate data quality control? Have they dealt with possible confounders?

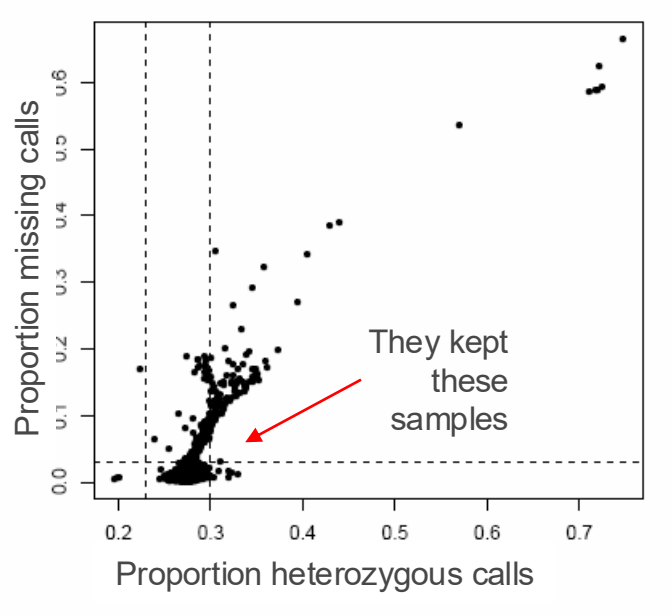
Collection	Missingness	Heterozygosity	External discordance	Non-European ancestry	Duplicate	Relative	Total
58C	9	0	4	6	4	1	**24**
UKBS	8	0	5	14	0	15	**42**
BD	30	0	0	9	77	13	**129**
CAD	41	1	0	13	2	5	**62**
CD	43	4	6	54	131	18	**256**
HT	29	0	0	2	6	11	**48**
RA	47	1	0	26	53	9	**136**
T1D	7	2	1	18	6	3	**37**
T2D	36	1	0	11	16	11	**75**
Total	**250**	**9**	**16**	**153**	**295**	**86**	**809**

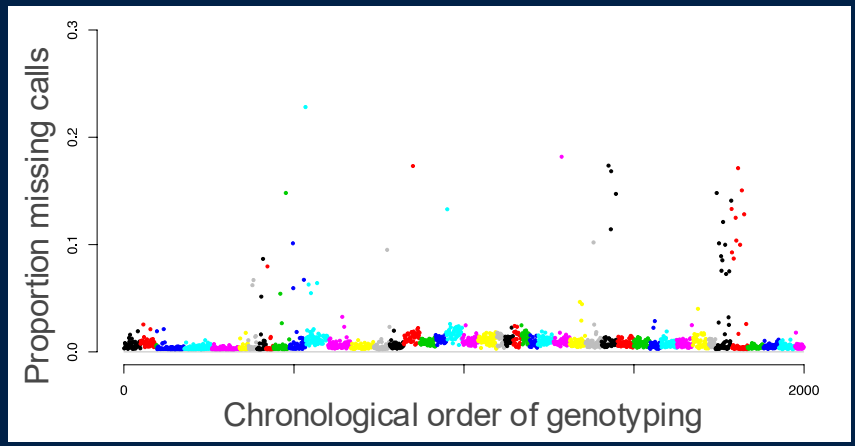
Supplementary Table 4 | Exclusion summary by collection. Six filters were applied for sample exclusion: 1. SNP call rate < 97% (missingness). 2. Heterozygosity > 30% or < 23% across all SNPs. 3. External discordance with genotype or phenotype data. 4. Individuals identified as having recent non-European ancestry by the Multidimensional Scaling analysis (see Methods). 5. Duplicates (the copy with more missing data was removed) 6. Individuals with too much IBS sharing (>86%); likely relatives. Where individuals could be excluded for more than one reason, they appear in the leftmost such column.

They then threw away 809 samples!

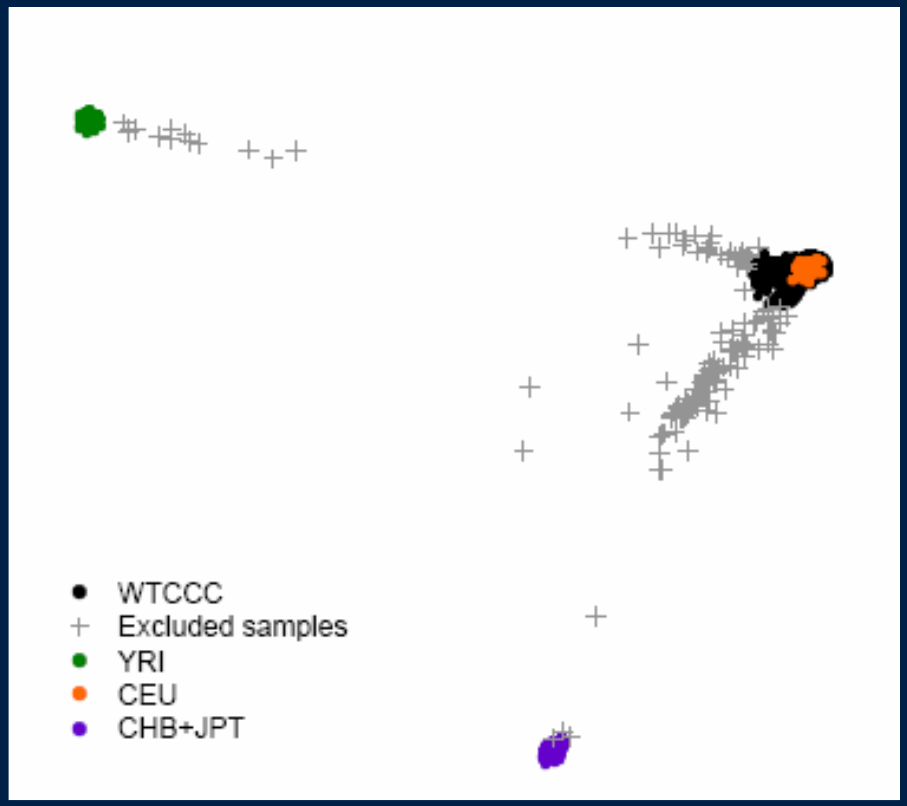
Due to:

- Poor genotyping rates
- Evidence of contamination (too many heterozygous genotypes)
- Evidence of being not of European ancestry
- A duplicate, or close relative of another sample





Some of the poor quality data was apparently due to batch effects.

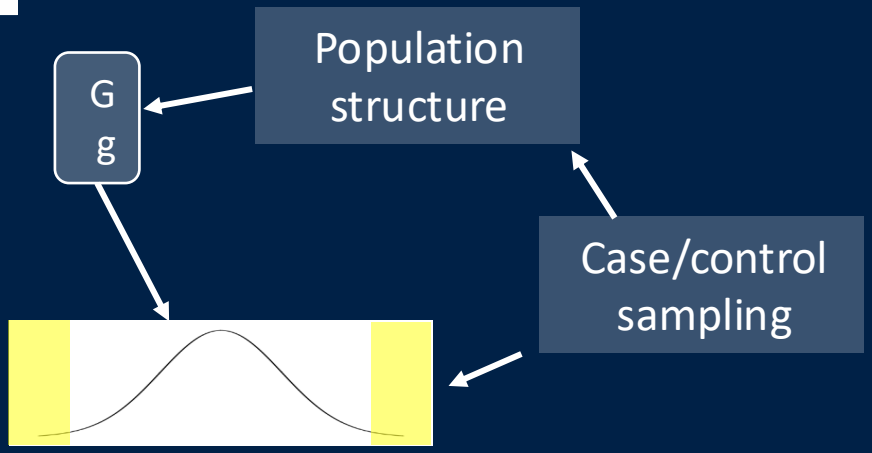


PCA computes genome-wide relationships between samples and then looks for directions of greatest variation. Since relatedness typically decreases with geographic distance, principal components typically reflect geography.

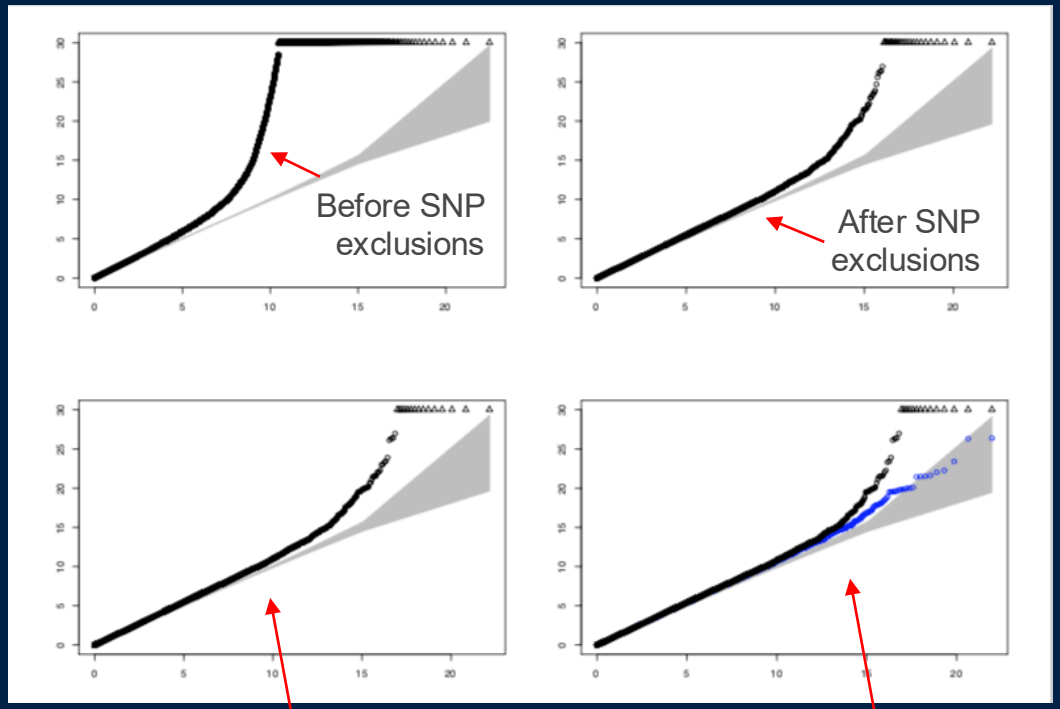
To avoid confounding by population structure, the samples were all supposed to be from the United Kingdom, and with European ancestry.

They used a method called *principal components analysis* to detect ancestry against the HapMap project samples. Some non-European ancestry individuals had been typed.

153 individuals were excluded on this basis.



Using quantile-quantile plots to assess residual confounding



After visually inspecting cluster plots for remaining associated SNPs

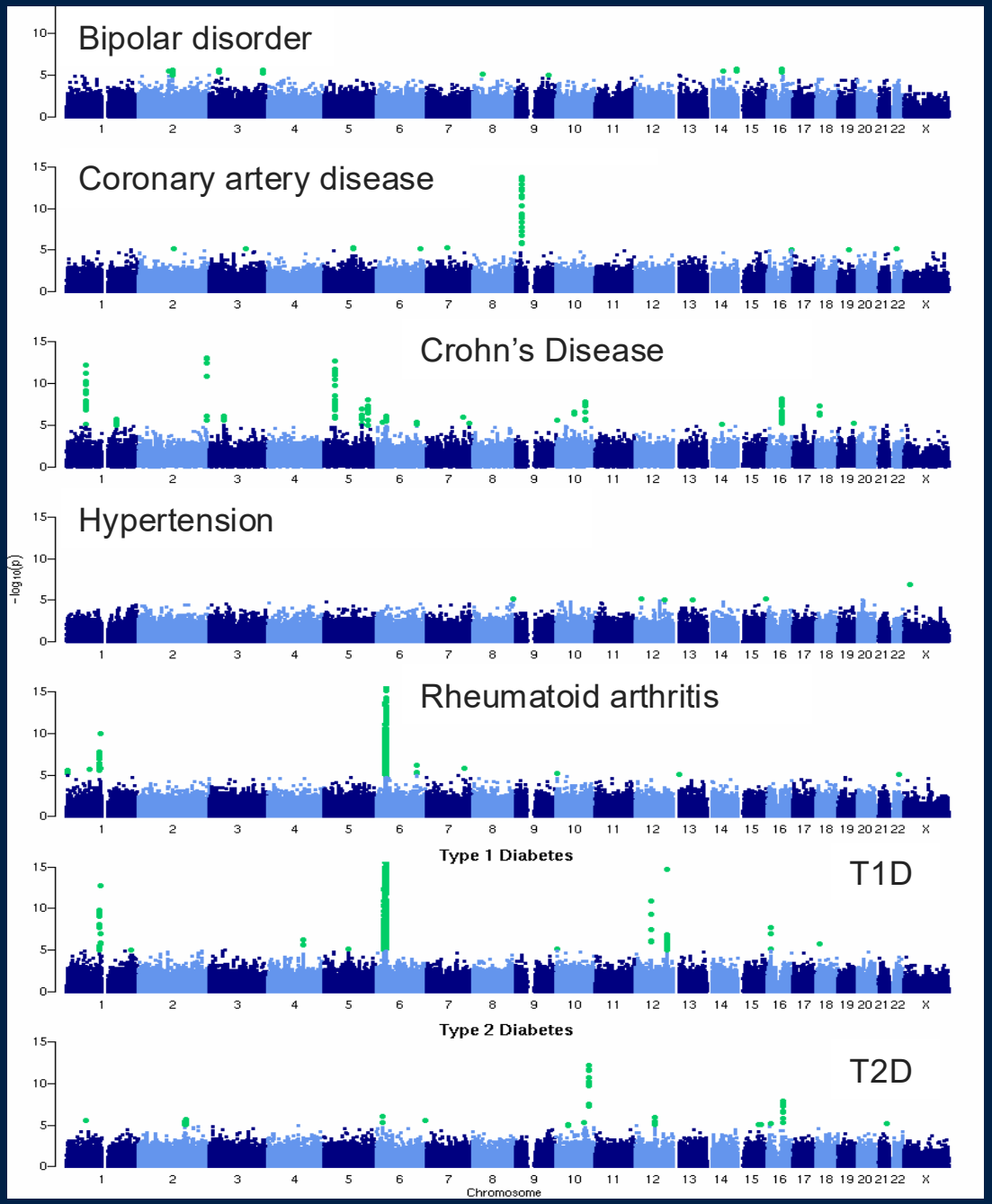
(Blue dots)… and after removing remaining strongly-associated regions that they claim to be real

They also excluded 25,567 SNPs from the study for

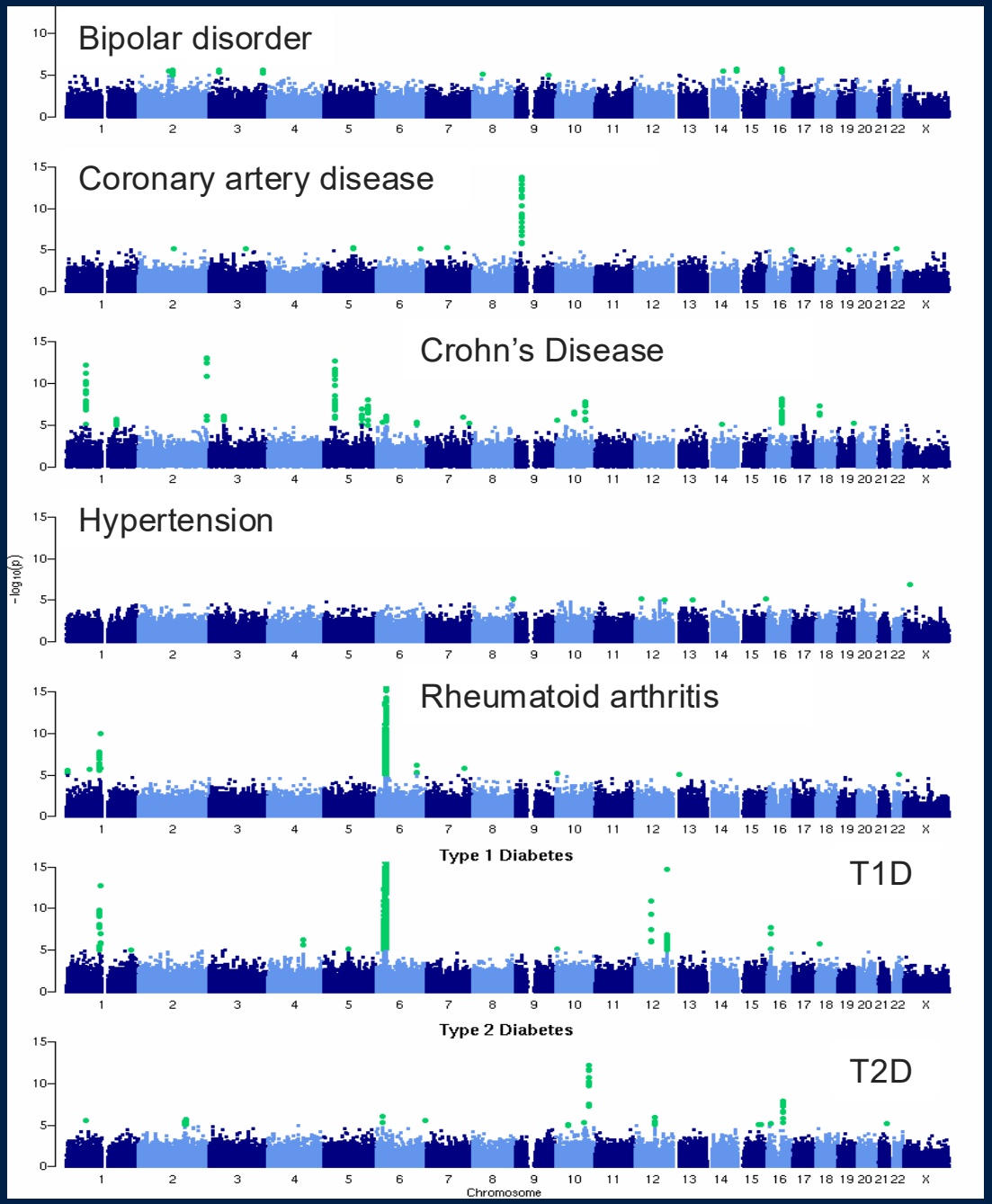
- High missing data rates
- Deviation from Hardy-Weinberg equilibrium (lecture 1) in controls
- Frequency differences between the two control groups
- And they visually inspected cluster plots for remaining SNPs

If there are few true signals, and if we have removed confounders – then P-values should largely come from a uniform distribution - they should lie on the diagonal.

Phew!



The main result of the study



Number of associations with strong evidence

1

1

9

0

3

7

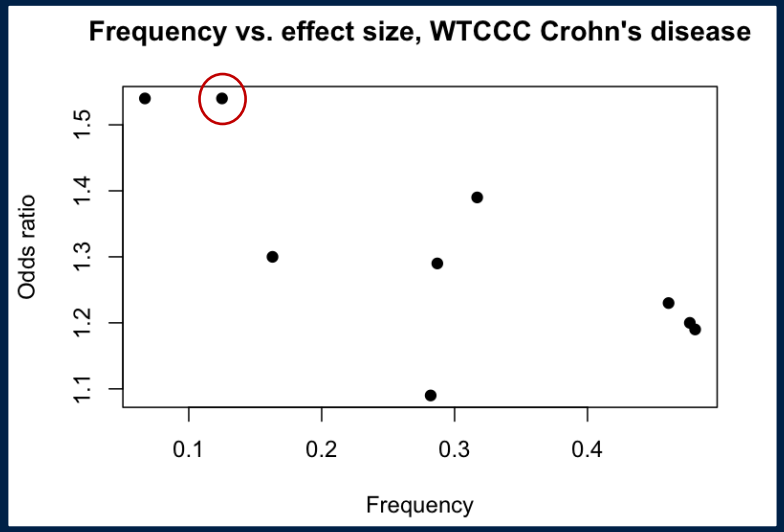
3

The study found 25 associations at their nominal P-value threshold.

Twelve of these provided replication of previously implicated variants. Thirteen were new associations.

The traits clearly differ in their genetic architecture

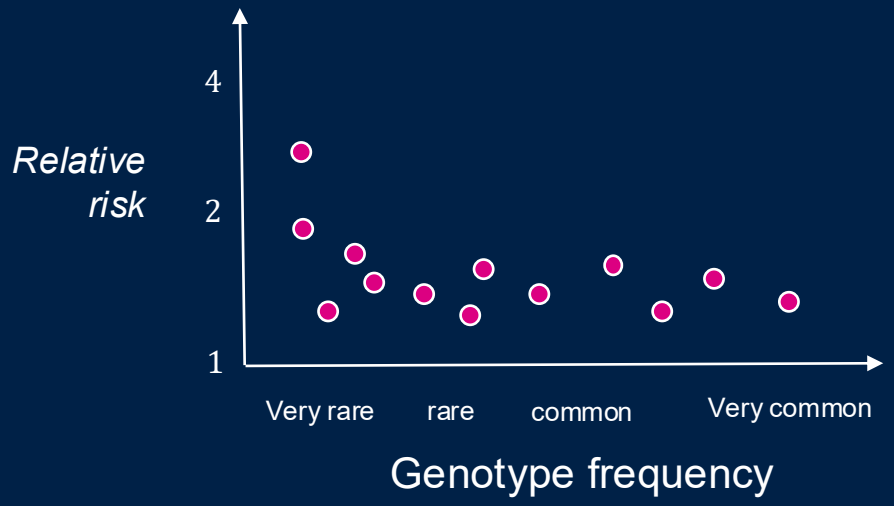
Some SNPs were associated with some evidence with multiple traits (mainly for the autoimmune diseases).



Effect sizes were generally modest

E.g. across the 9 associations with Crohn's disease, the maximum estimated odds ratio was 1.54, (similar to the O blood group example)

(A strong effect with Type 1 Diabetes was also observed in the MHC locus)



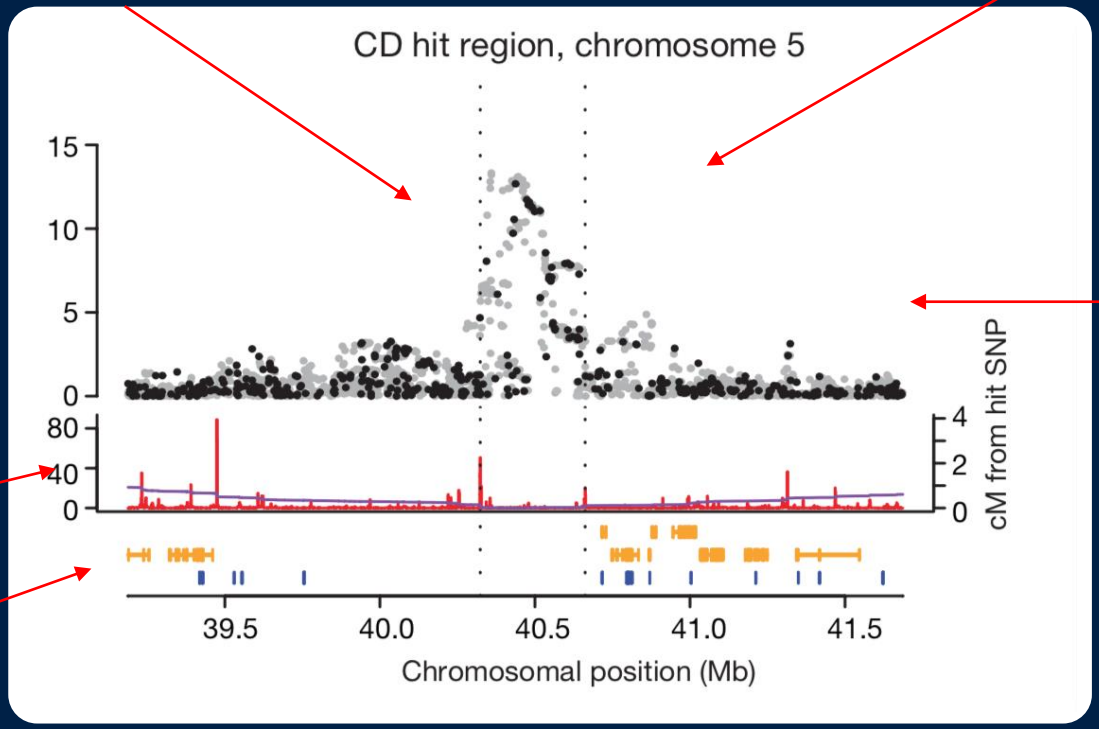
Zooming in to a GWAS 'hit' plot

Sometimes called a 'locus zoom' plot. Here are some things to look for:

Evidence for association with each SNP (-log10 P-value or log10 Bayes factor)

Delineation of association region boundaries (usually based on heuristics)

Black points were typed, grey points were imputed from HapMap



The recombination rate, here as estimated by HapMap

Signal ought to follow LD patterns. In particular ought to drop off near recombination hotspots

Regional genes

Position of SNPs in the reference genome assembly

Main lecture messages

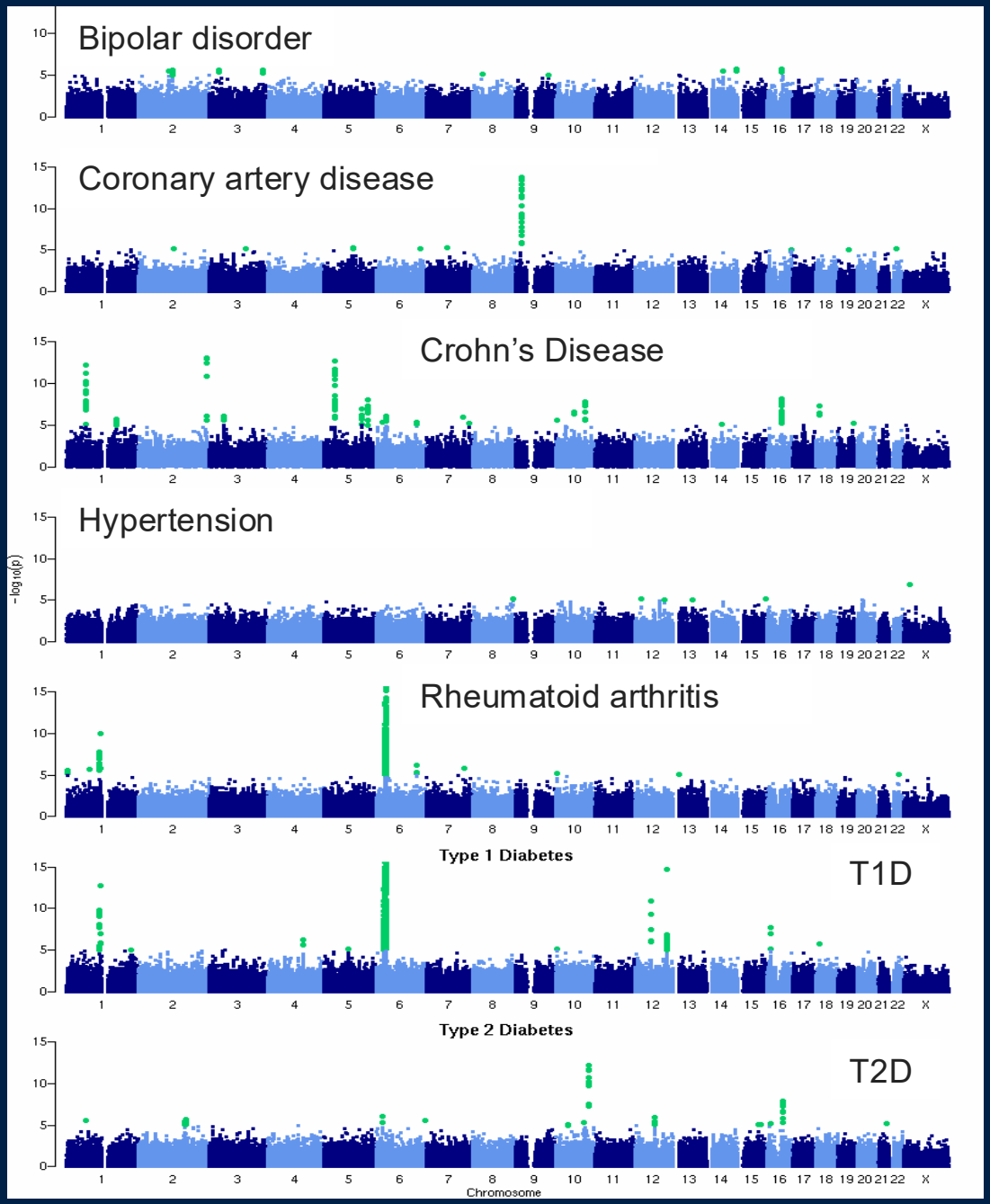
1. Most human phenotypes are highly heritable
(a large proportion of variation is due to genetics)

2. But many 'complex' traits are *not* mendelian - they are polygenic - *"common variant, common trait" hypothesis*

3. The discovery of this fact is due to ***genome-wide association studies*** (GWAS)
Collect **lots** of samples – type **lots** of variants – careful quality control – conduct statistical test of association – use stringent statistical thresholds.

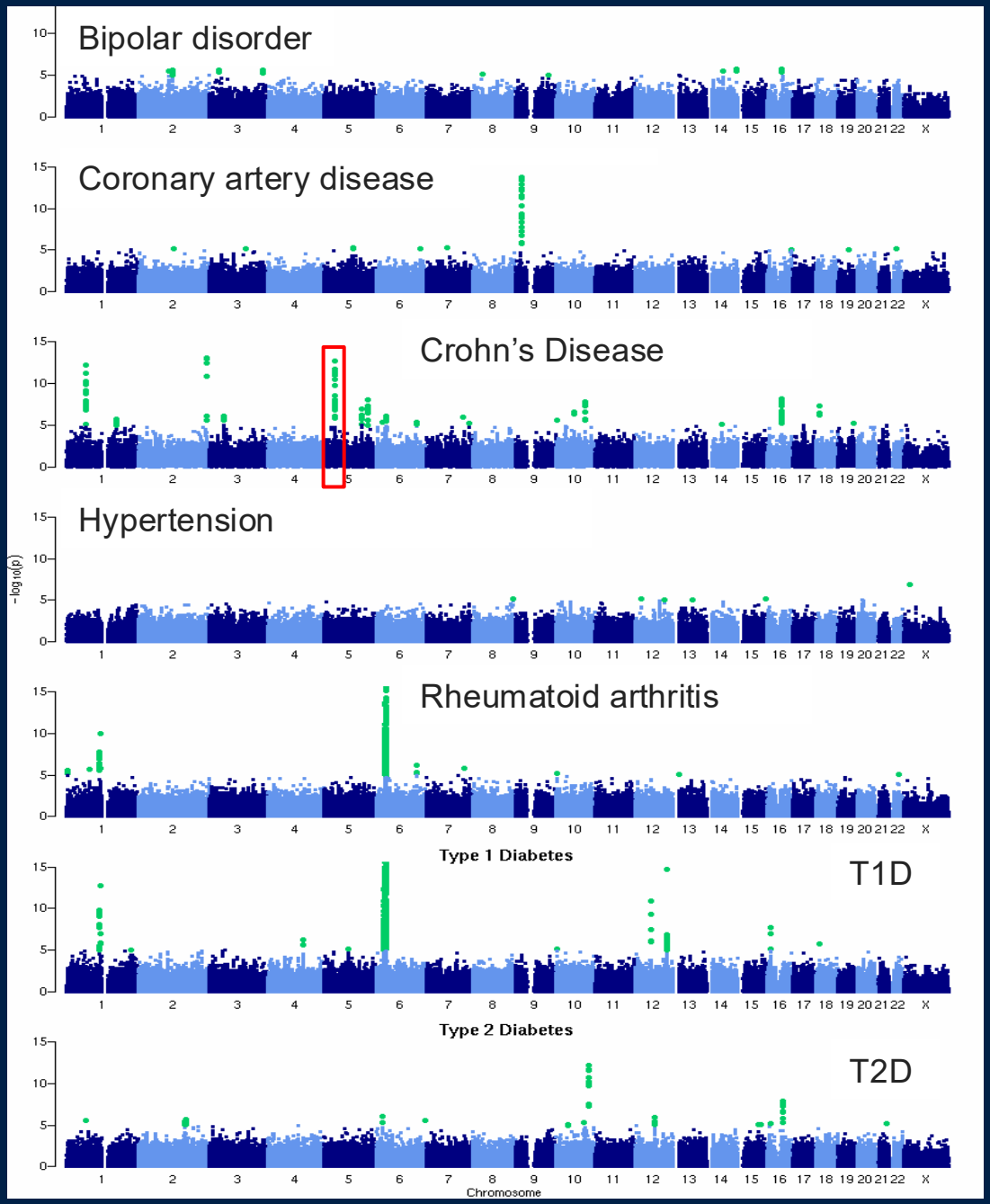
The first of these studies was conducted in the mid 2000s and showed the hypothesis is **definitely true**.

What about the underlying biology?



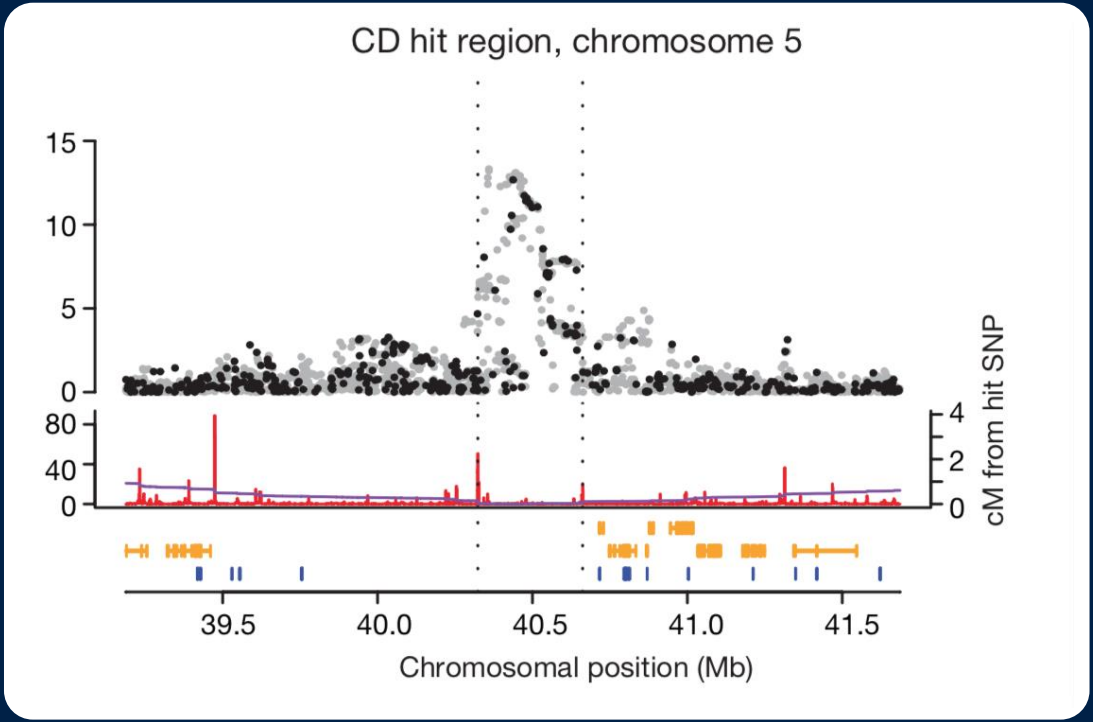
We have clearly learned something about the biology of these traits.

…but what have we learned?

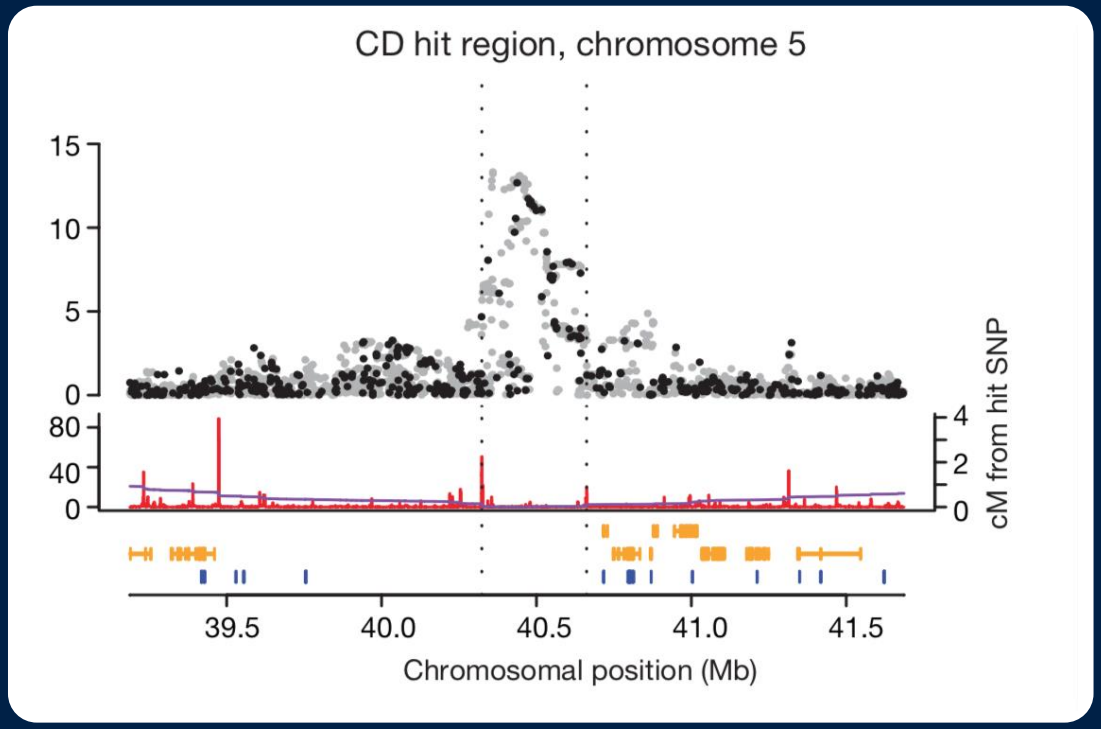

Bipolar disorder
Coronary artery disease
Crohn's Disease
Hypertension
Rheumatoid arthritis
Type 1 Diabetes
T1D
Type 2 Diabetes
T2D
Chromosome

Let's zoom in

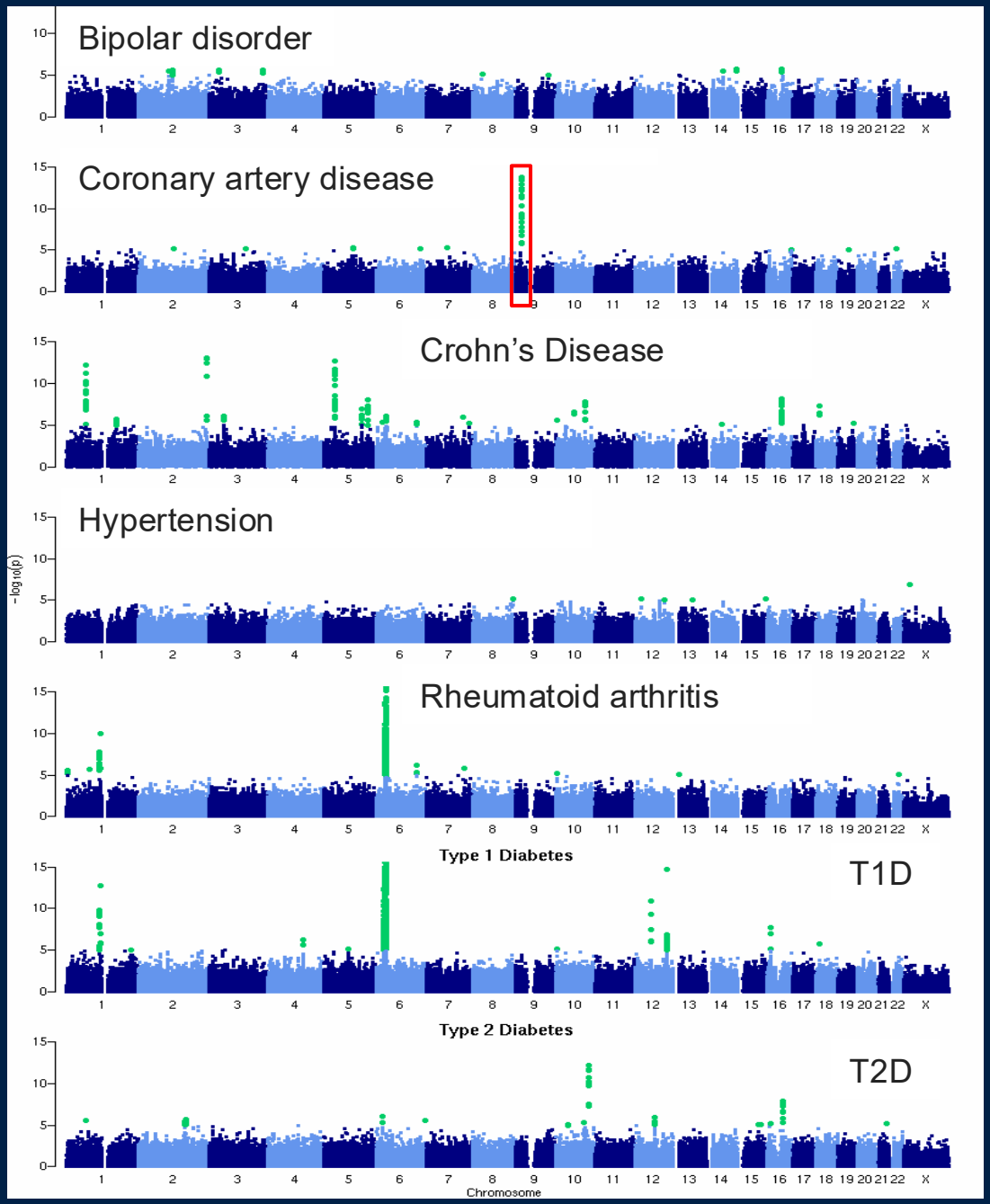
Biology is hard



Biology is hard



No genes under the main association signal!

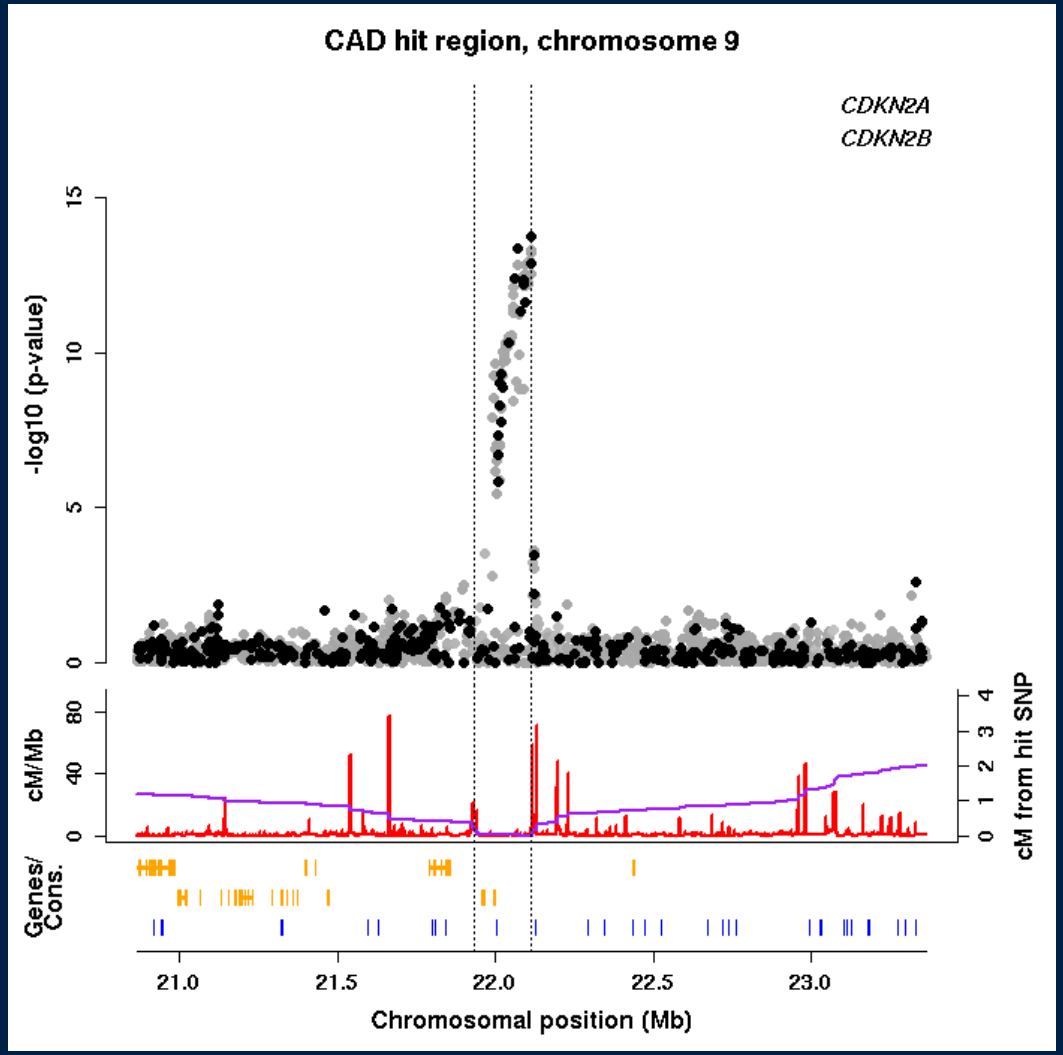

Bipolar disorder
Coronary artery disease
Crohn's Disease
Hypertension
Rheumatoid arthritis
Type 1 Diabetes
T1D
Type 2 Diabetes
T2D
−log10(p)
Chromosome

Biology is hard

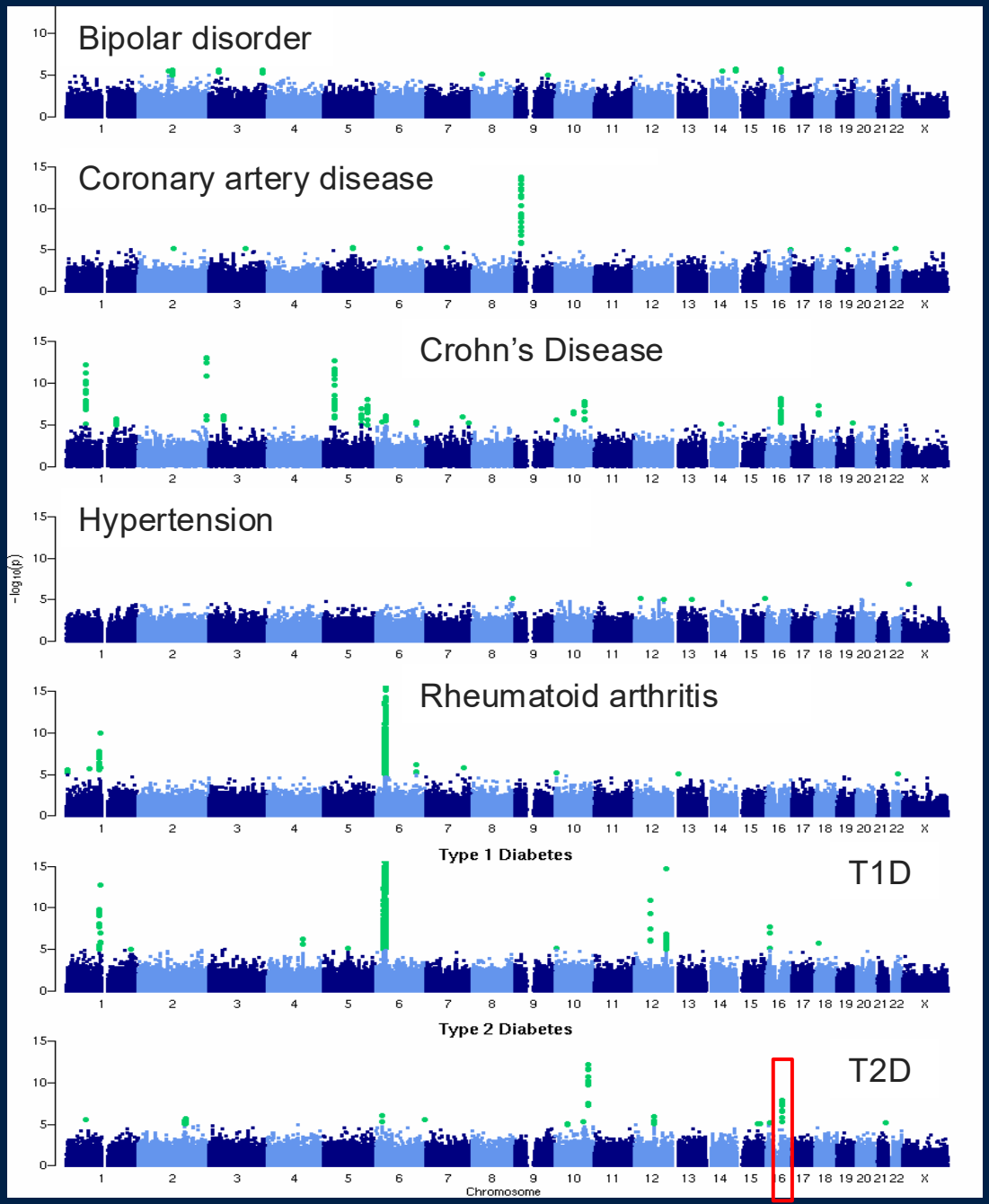
Association observed with CAD over a ~100kb region of chromosome 9. This is unquestionably a real association (it has been replicated in several independent studies).

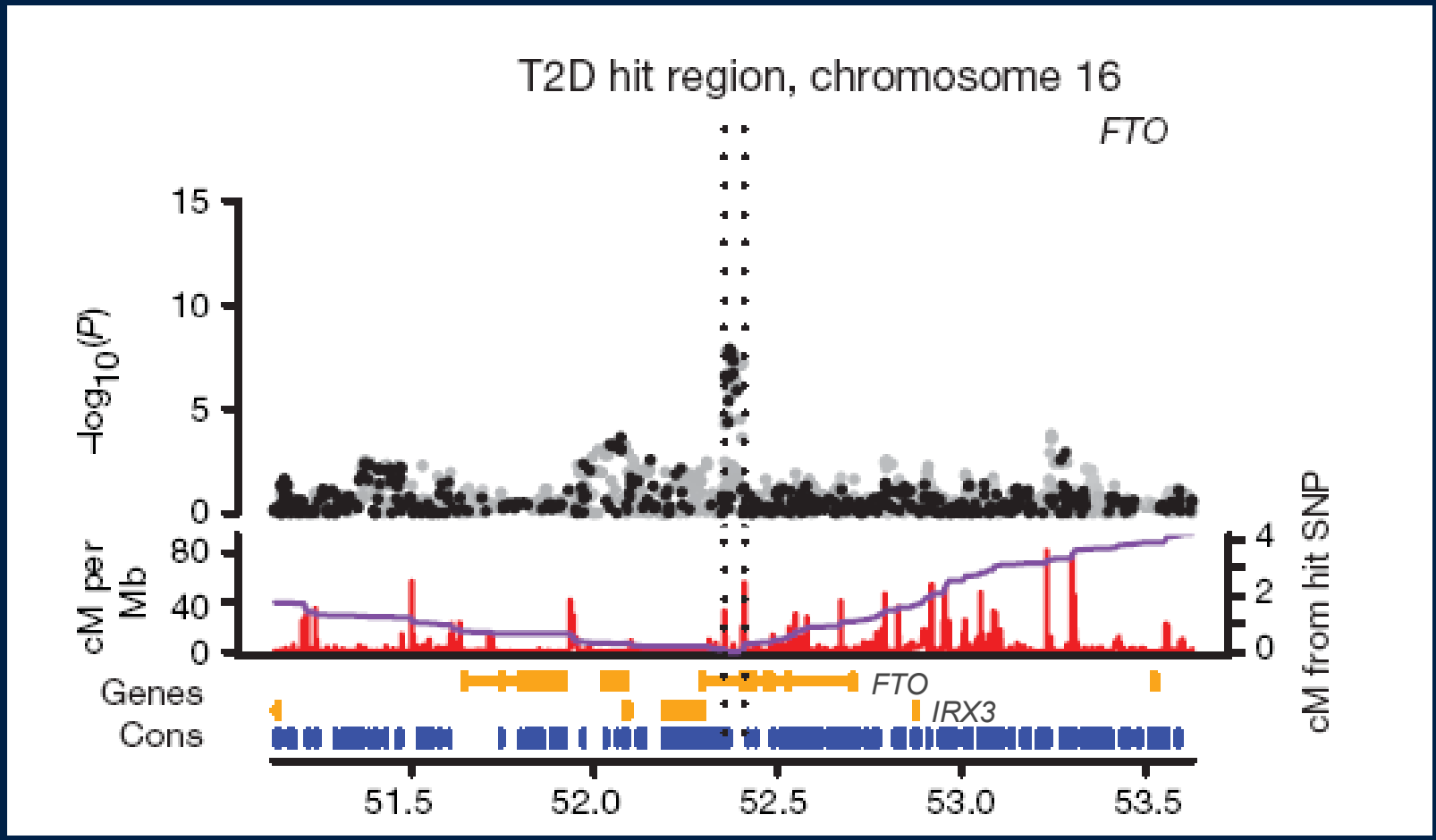
The functional mechanism of this association is not fully solved; it probably involves regulation of expression of the two nearby genes *CDKN2A/B*.

Neither gene was an obvious candidate beforehand - thus, this association does point to novel biology.



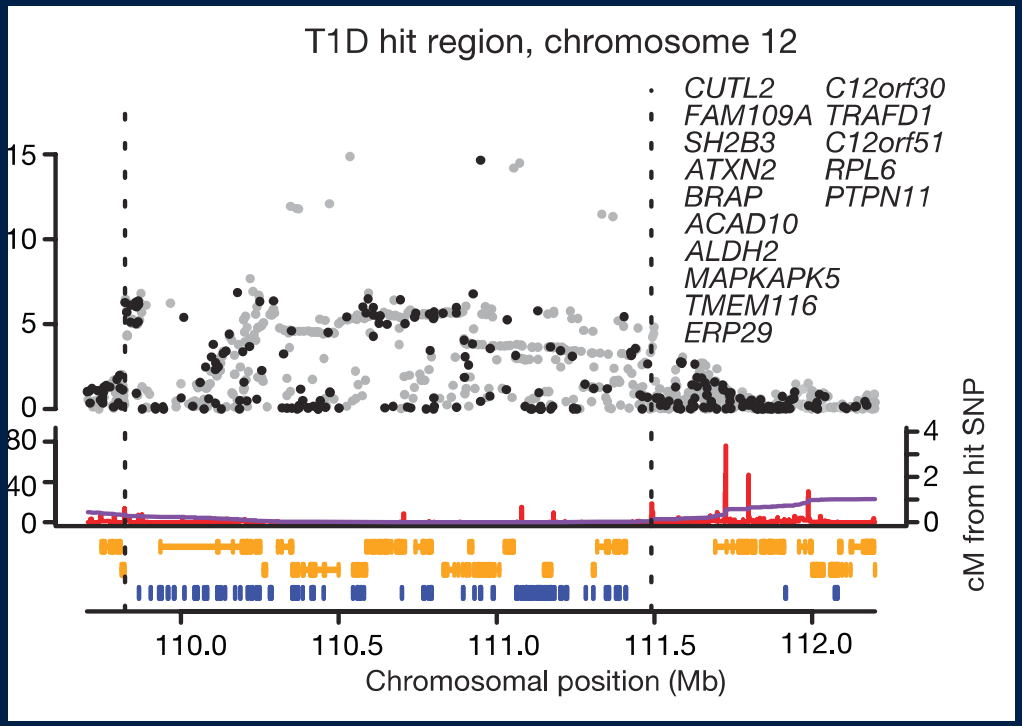
Harismendy et al Nature 2011; Almontishari, et al JACC 2013; Almonitishari et al, Circulation 2015


Bipolar disorder
Coronary artery disease
Crohn's Disease
Hypertension
Rheumatoid arthritis
Type 1 Diabetes
T1D
Type 2 Diabetes
T2D
−log10(p)
Chromosome



This association with Type 2 Diabetes turned out to be through a second, related trait (obesity), again unquestionably a real effect. But as of 2018 the functional mechanism remains unclear. Expression of *FTO* is known to affect obesity, but the SNPs may also affect expression of another gene, IRX3, 200kb away.

Smemo et al, Nature 2014



This pattern has turned out to be typical. It has generally proven extremely hard to narrow down GWAS associations to underlying ‘causal’ variants.

LD is a double-edged sword.

Next lecture: we will look at this.

Anatomy of a GWAS – what to look for

1. Collect as many cases and controls as possible

What samples How many?

2. Genotype (or impute) them at as many variants across the genome as possible

How many?

3. Deal with potential confounders – careful data quality control and handle population structure.

How did they do quality control – is it adequate?

4. Estimate relative risks, and look for statistical evidence that of $RR \neq 1$

5. If estimate is many standard deviations from zero, bingo! We may have found a true causal effect.

Did they find anything with enough evidence?

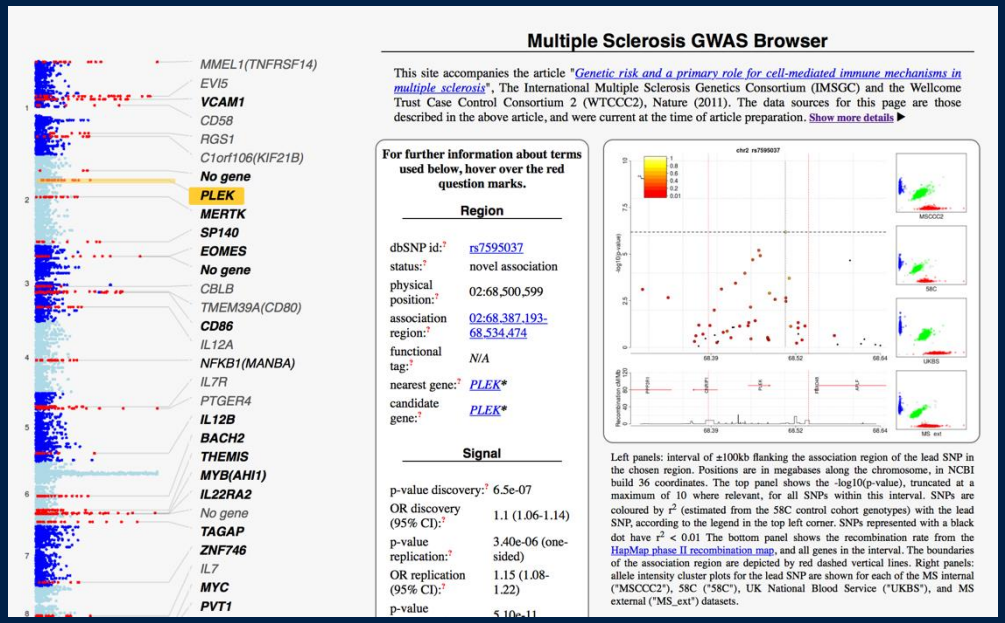
6.Replicate in other studies, or find other corroborating evidence?

Is it convincing?

7. (Now try to understand the underlying biology.)

Can they understand the biology?

Consolidation question



GWAS of multiple sclerosis (2011)

9772 cases, 17,376 controls from across Europe

www.chg.ox.ac.uk/wtccc2/ms/
(I think this url requires the trailing /)

Visit the above site and make sure you understand what is shown. Pick a signal and try to work out

- What is the estimated effect size?
- How strong was the evidence?
- Did it replicate?
- Does the association signal look sensible – does it follow LD patterns, and do the cluster plots look sensible?
- Can you figure out what the nearby genes do? (**Warning**: this can be a time sink!)

Bonus question: read the paper and try to figure out the questions on the checklist.

Next lecture: Friday 6th Mar @14:00

Understanding the genetics of complex traits II

Gavin Band gavin.band@well.ox.ac.uk

BA Human Sciences

Friday 6th March 2026