Genome-wide association studies I: Identifying genetic associations with complex traits

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

MSc Global Health Science and Epidemiology

Genetic Epidemiology Module

26th Feb 2024

Learning objectives

Understand a genome-wide association study (GWAS) and the concept of a hypothesis-free approach to studying genetic associations.

Have a working knowledge of the different steps involved in the conduct of GWAS, including study design, quality control and basic analyses.

Be able to interpret and critically appraise evidence from genome-wide association studies.

Understand the relevance of replication, meta-analysis and consortia, and multi-ancestry approaches, in genome-wide association studies.

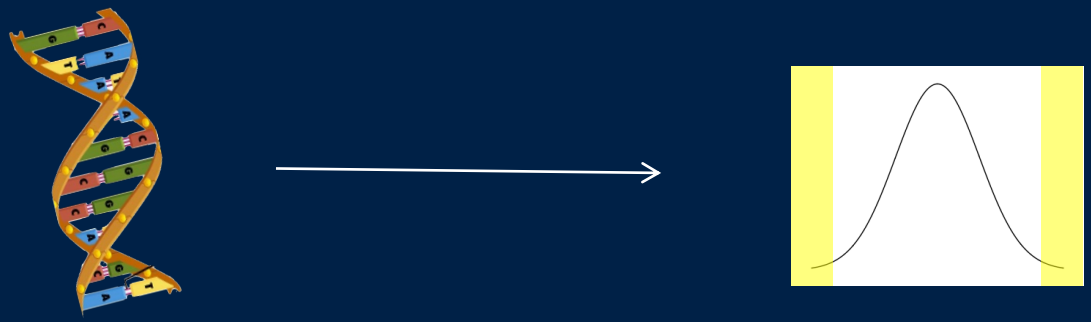
Appreciate the use of post-GWAS analyses including fine mapping, gene and pathway analyses, and the concept of causal variants.

Lecture outline

- Why GWAS? Heritability & genetic architecture
- Testing for association
- What to genotype, and how? LD and the HapMap study
- A real GWAS study
- The challenge of understanding biology

The human genome is ~3.2 billion base pairs long.

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



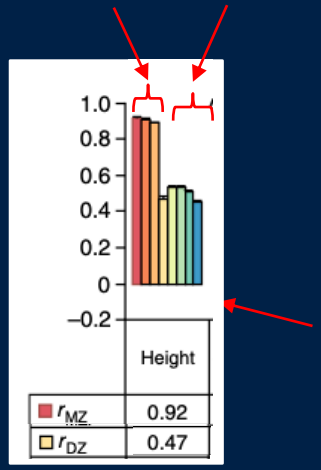
What proportion of phenotypic variation is
due to genetic variation?

Human traits are highly heritable

Idea: if genetics determines a trait, then *more genetically similar individuals should have more similar phenotypes.* Can estimate how much genetics determines trait variation by comparing trait similarity in monozygotic (identical) and dizygotic twins.

MZ Twins r~0.92

DZ Twins r~0.47



All studied traits

(Adult) height is much more highly correlated between monozygotic than dizygotic twins.
Heritability is about 90%.

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)

2748 papers,

Definition: 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

Idea: if genetics determines a trait, then *more genetically similar individuals should have more similar phenotypes.* Can estimate how much genetics determines trait variation by comparing trait similarity in monozygotic (identical) and dizygotic twins.

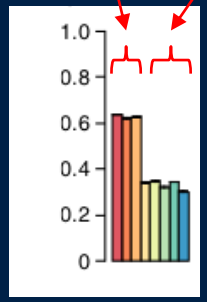
Meta-analysis of the heritability of human traits based on fifty years of twin studies

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)

MZ Twins r~0.64

DZ Twins r~0.34



All studied traits

Across all traits, phenotypes are much more highly correlated between monozygotics than dizygotic twins. Heritability (averaged across traits) is about 60%.

Definition: 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.*

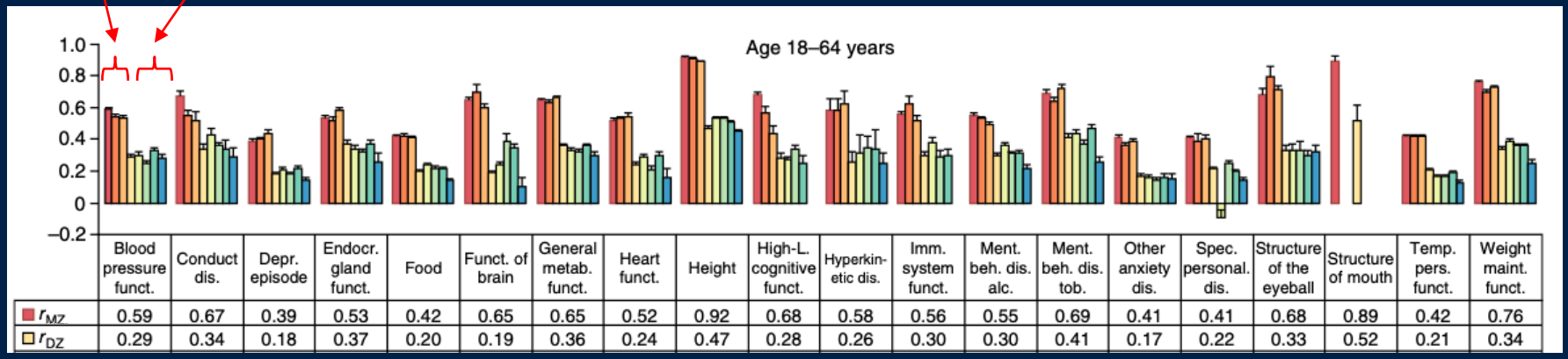
Meta-analysis of the heritability of human traits based on fifty years of twin studies

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 $h^2 \approx 90\%$

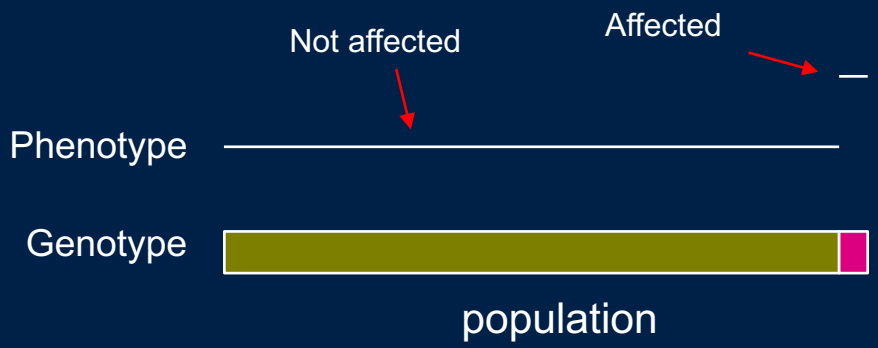
"Higher level cognitive function" $h^2 \approx 80\%$

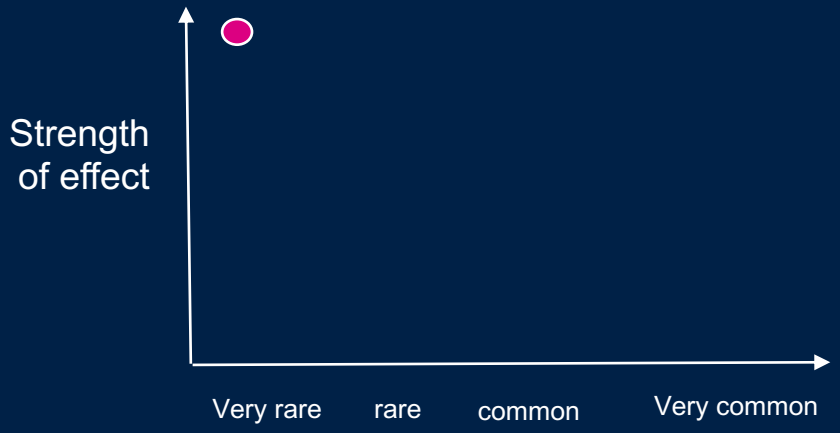
Structure of the eyeball $h^2 \approx 70\%$

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

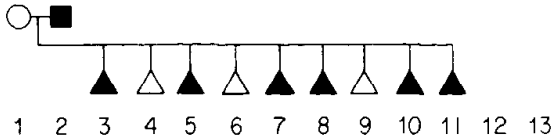
A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes

The Huntington's Disease Collaborative Research Group*

Introduction

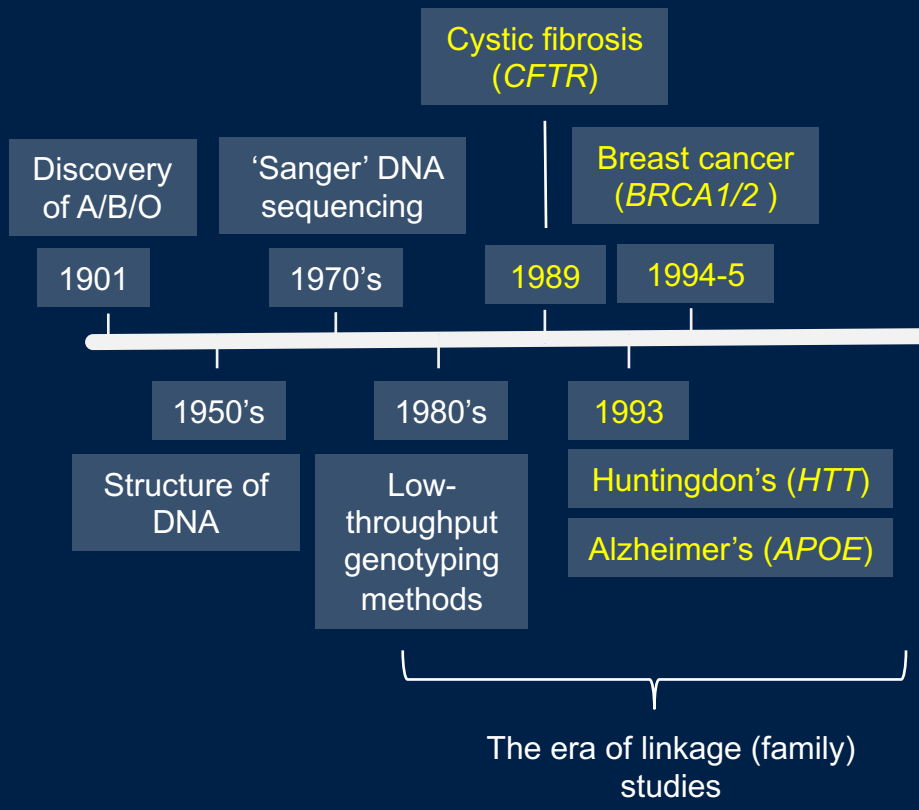
Affects ~1 in 20,000 people of European ancestry
(less in Africa and Asia)

Discovered by looking in families



A **"Mendelian"** trait

End of an era

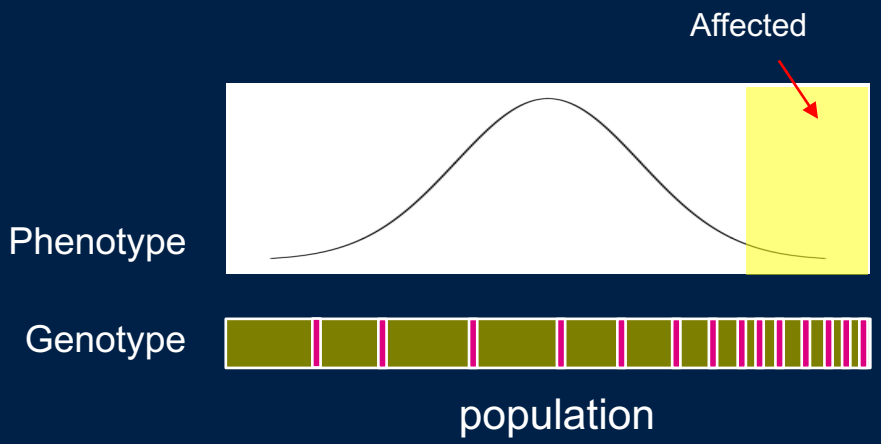


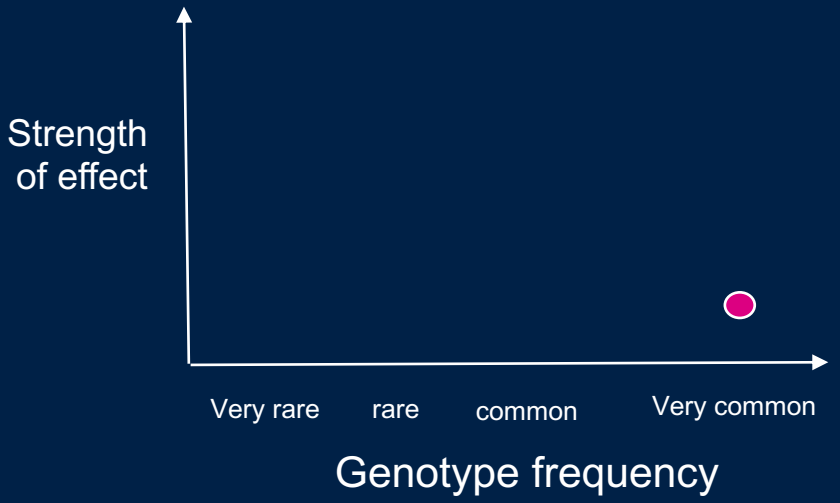
"Linkage Mapping was successful in identifying the genetic basis of many human diseases in which the disease penetrance resembles a simple Mendelian model e.g. Huntington's disease, Cystic Fibrosis, some forms of breast cancer, Alzheimers, …"

"…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"

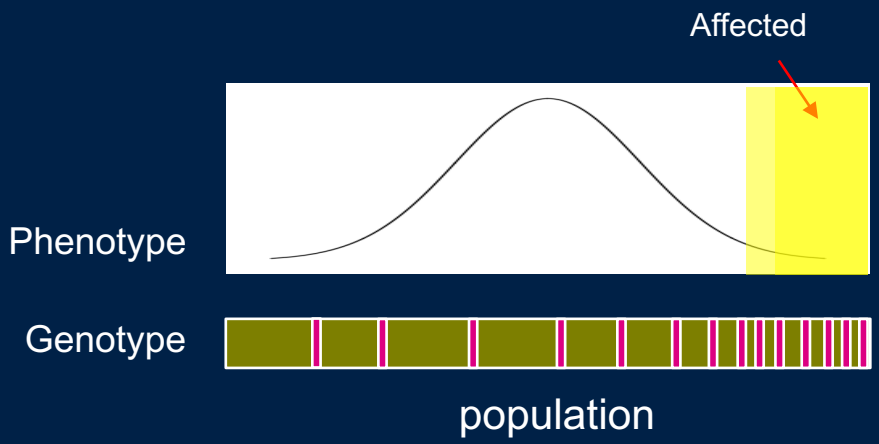
– Risch "*Searching for genetic determinants in the new millennium*" Nature (2000)

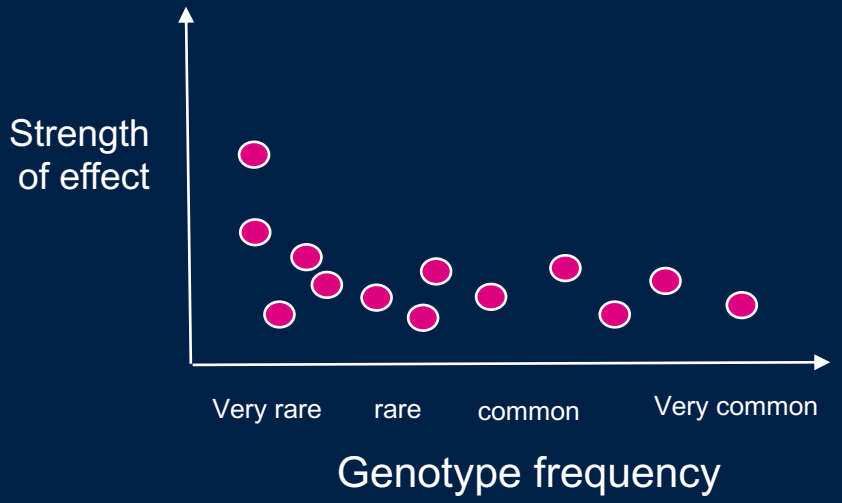
Common variant, common disease hypothesis





Common variant, common disease hypothesis





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.

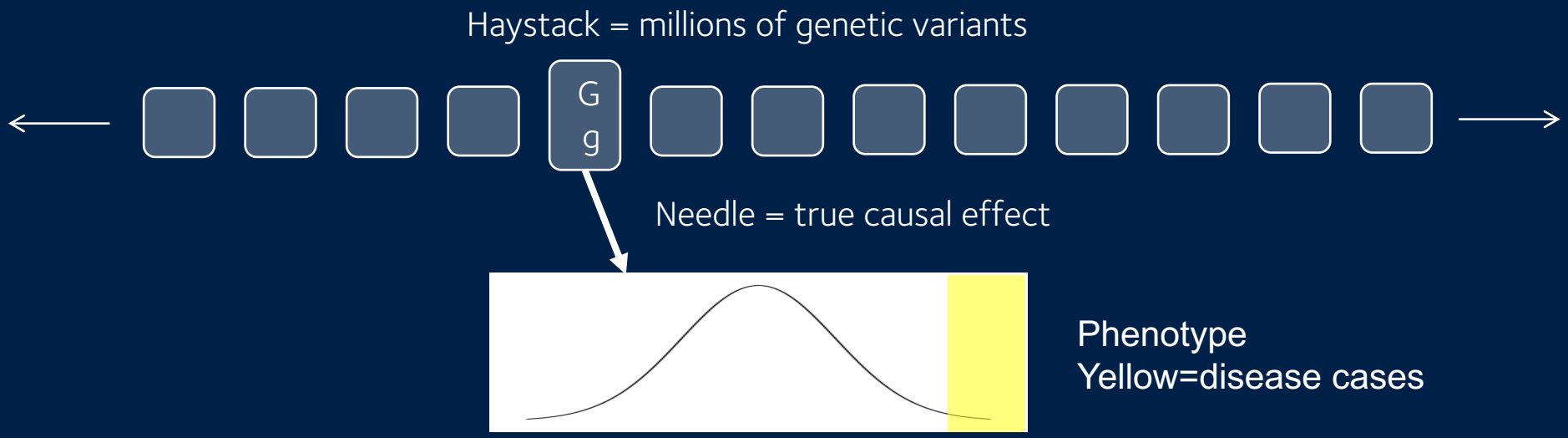
- By the 2000s family studies had identified the causes of several mendelian traits, but had failed to solve the genetics of multiple complex diseases.

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

Lecture outline

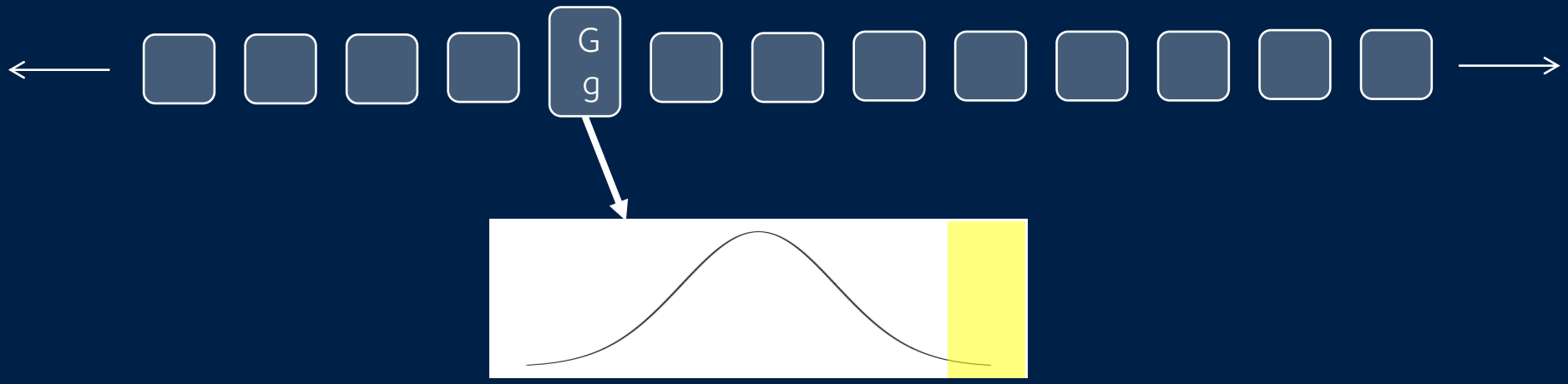
- Why GWAS? Heritability & genetic architecture
- → Testing for association
- What to genotype, and how? LD and the HapMap study
- A real GWAS study
- The challenge of understanding biology

Searching for a needle in a haystack



Aim: find the causal genetic variants

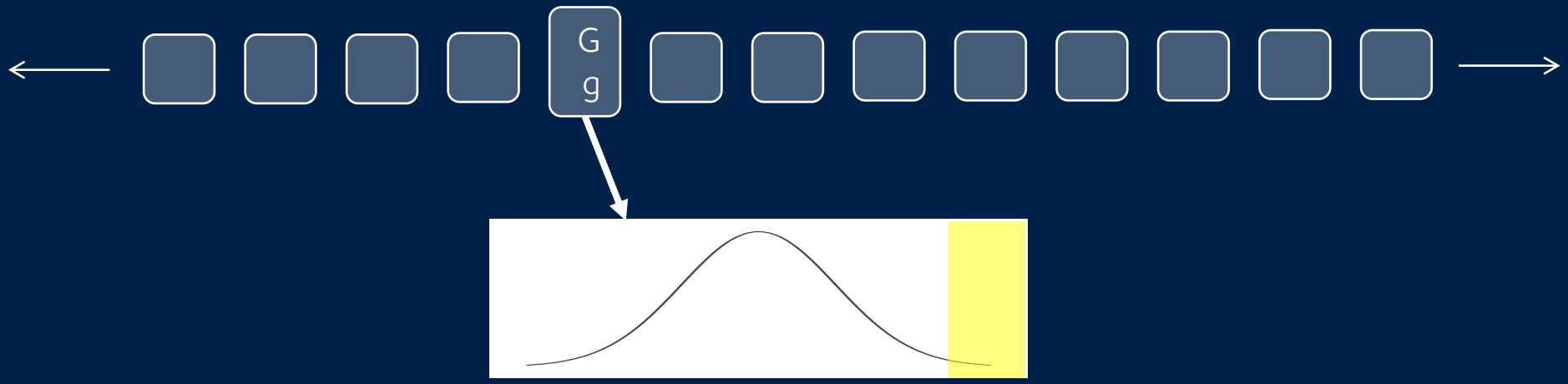
Causal effects generate relative risk ≠ 1



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

$$\frac{\text{“Chance/frequency of disease given genotype } G\text{”}}{\text{“Chance/frequency of disease given genotype } g\text{”}} > 1$$

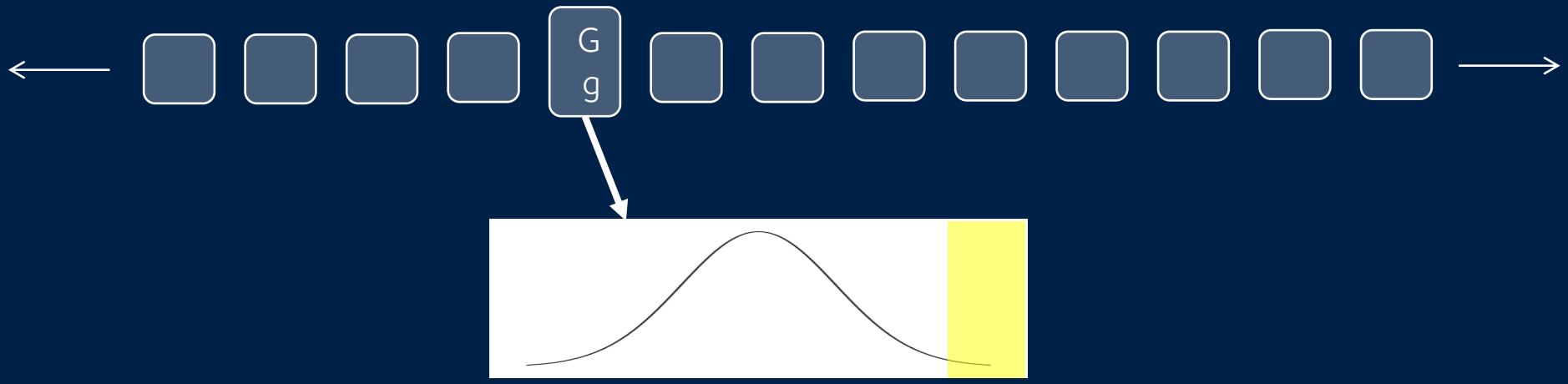
Causal effects generate relative risk ≠ 1



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

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

Causal effects generate relative risk ≠ 1



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

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

=> We can find genetic effects by looking for **RR ≠ 1**

How to run a GWAS, take 1

1. Collect DNA samples from as many cases and controls as possible

2. Genotype genetic variants genome-wide and estimate relative risk

3. If there's enough statistical evidence that RR != 1, bingo!

(+ now try to understand the underlying biology…)

How to run a GWAS, take 1

1. Collect DNA samples from as many cases and controls as possible

Which people? How many?

2. Genotype genetic variants genome-wide and estimate relative risk

What to genotype and how?

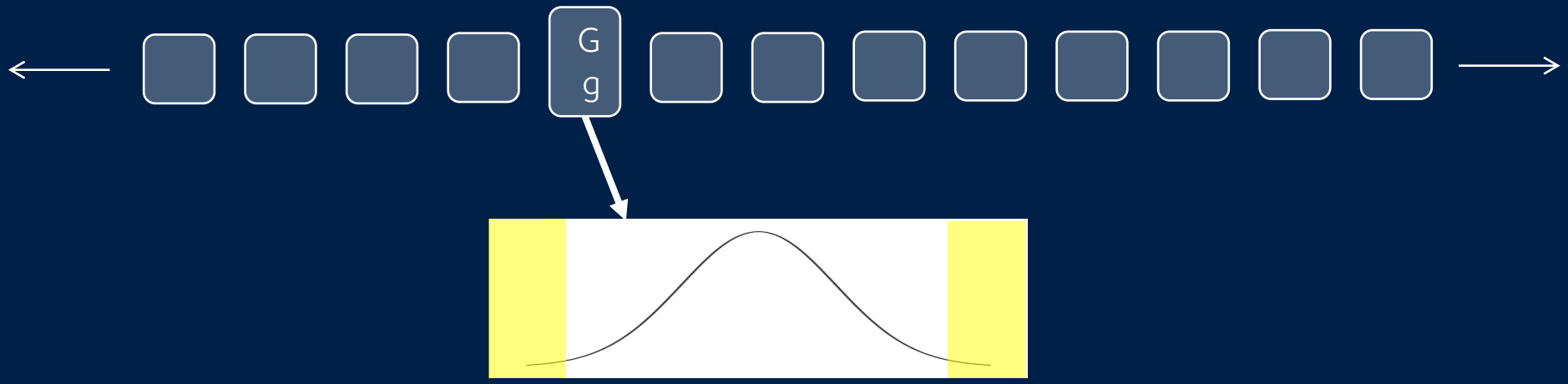
How to estimate?

3. If there's enough statistical evidence that RR != 1, bingo!

How much evidence needed?

(+ now try to understand the underlying biology…)

Causal effects generate relative risk ≠ 1



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

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

Write as "probability" instead of "chance"

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)

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)

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

Note: Also approximately true for 'true' controls, provided the disease is relatively rare.

Example: O blood group and severe malaria

Cases were ascertained as children arriving in hospital with severe symptoms compatible with malaria & parasitaemia

Controls were ascertained from new births in the same hospitals.

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

Data from N=3,068 samples from Kilifi, Kenya
MalariaGEN 2019 doi: 10.1038/s41467-019-13480-z

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

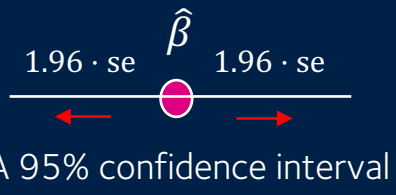
Estimate that O blood group is associated with a ~30% lower chance of severe malaria (all else being equal).

But how accurate is this estimate?
How much evidence that $RR \neq 1$?

Key association test summary statistics

Effect size estimate. i.e. the odds ratio. Typically expressed on the log scale and denoted by beta ($\hat{\beta}$).

The standard error, which reflects uncertainty in the estimate. Also usually expressed on the log(OR) scale and denoted as "se".



A 95% confidence interval

Computed from these is **The p-value** expressing the evidence that $RR \neq 1$.

The P-value is computed in practice by assuming the errors have a normal distribution:

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

Normal distribution function

For a GWAS we typically want P very small. Typical threshold used: $P < 5\times10^{-8}$

i.e. log(OR) estimate is about 5.5 standard errors away from zero.

How accurate is our estimate?

Incredibly useful formula:

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

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

f = frequency of G allele

ϕ = proportion of cases

The accuracy of the estimate (and our ability to distinguish from zero) depends on sample size, variant frequency, and on the proportion of cases in the study.

Note: this formula is appropriate for two alleles G vs. g, e.g. O blood group example. For an additive test GG/Gg/gg, use 2N in place of N.

Example: O blood group is associated with malaria protection

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

Data from *N*=3,068 samples from Kilifi, Kenya
MalariaGEN 2019 doi: 10.1038/s41467-019-13480-z

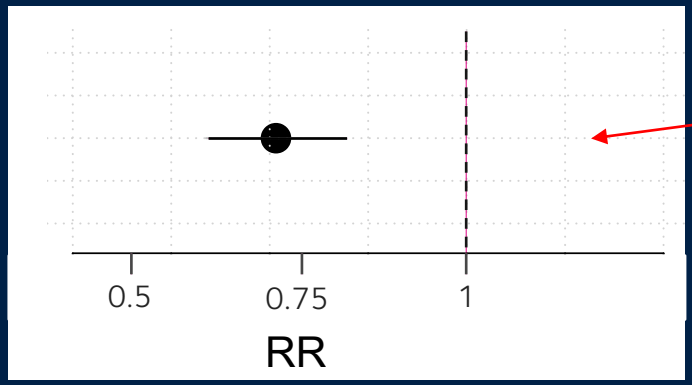
$$OR = \frac{686}{843} \times \frac{700}{839} = 0.68$$

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

$$\text{Standard error}(\log OR) \approx \frac{1}{\sqrt{3068 \times 0.45 \times 0.55 \times 0.25}} \approx 0.073$$ (on log scale)

Estimate is about 5 standard errors from zero

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



Convert back to RR scale:
Estimated relative risk = 0.68
95% CI = 0.59-0.78
(estimate +/- 1.96 standard errors)

Using regression to estimate

In a real study we would typically use a regression method e.g. logistic regression, rather than the above simple 2x2 calculation.
This is more flexible, allowing to control for other variables and/or assessing different models of association e.g. additive/ dominant / recessive and so on:

$$\text{logodds}(\text{disease}|g) = \text{baseline} + \hat{\beta} \times g + \text{other variables} \ldots$$

$\hat{\beta}$: log(OR) estimate

For a continuous trait (e.g. height) might use *linear regression* instead.

The method gives back the estimate, the standard error, and the p-value so don't have to compute by hand as above.

How to run a GWAS

1. Collect DNA samples from as many disease cases and population controls as possible

The se formula can be used to estimate how are many needed for a given effect size and variant frequency

✔ Which people? How many?

2. Genotype at variants genome-wide and estimate relative risk by computing the odds ratio, standard error and P-value.

What to genotype and how?

✔ How to estimate?

3. If P is small enough, bingo!

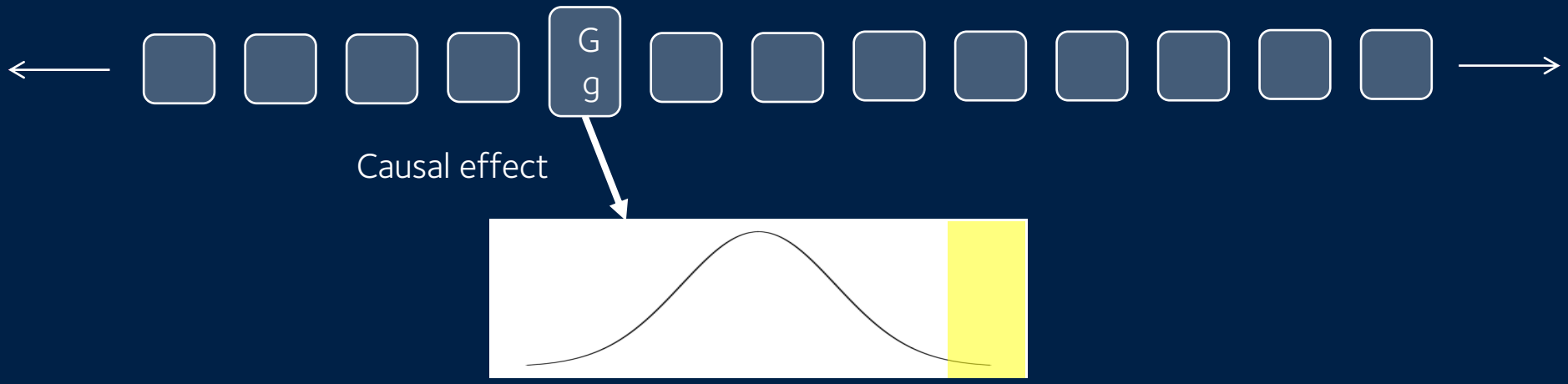
✔ How much evidence needed?

E.g. $P < 5 \times 10^{-8}$

(+ now try to understand the underlying biology…)

Practical gotchas…

Causal effects generate relative risk ≠ 1



1. If genotype *G* causes disease, then will have $RR \neq 1$

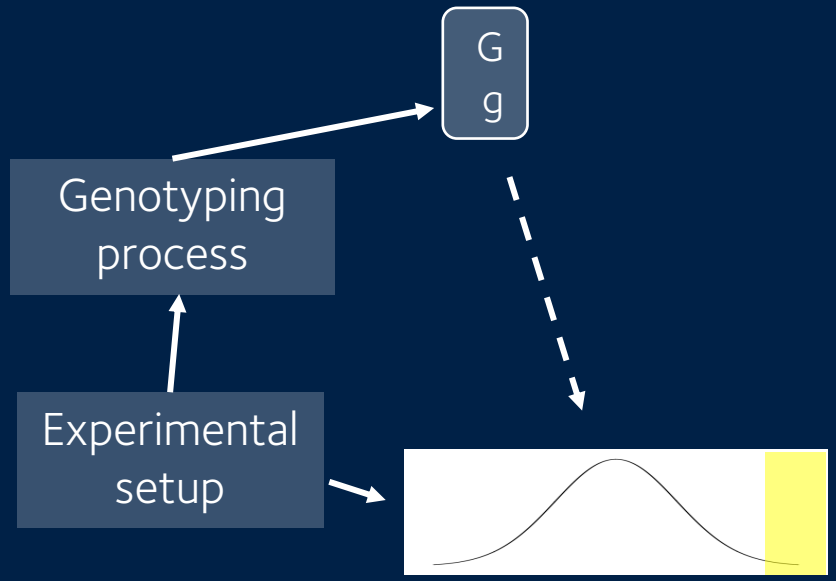
2. We estimate the RR and standard error in a sample of cases and controls. If the estimate is sufficiently far from 1, we declare them associated.

3. If P-value is small enough, they are associated, so start to get interested.

Major confounder 1: poor genotyping

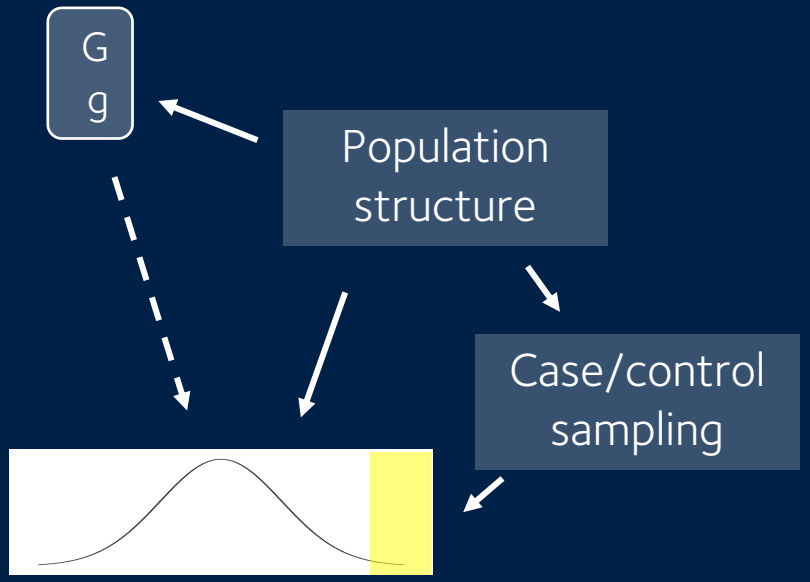
Association tests capture all causal paths from genotype to phenotype – even those that have nothing to do with biology.

E.g. differential genotyping performance between cases and controls



Major confounder 2: population structure

Association tests capture all causal paths from genotype to phenotype – even those that have nothing to do with biology.

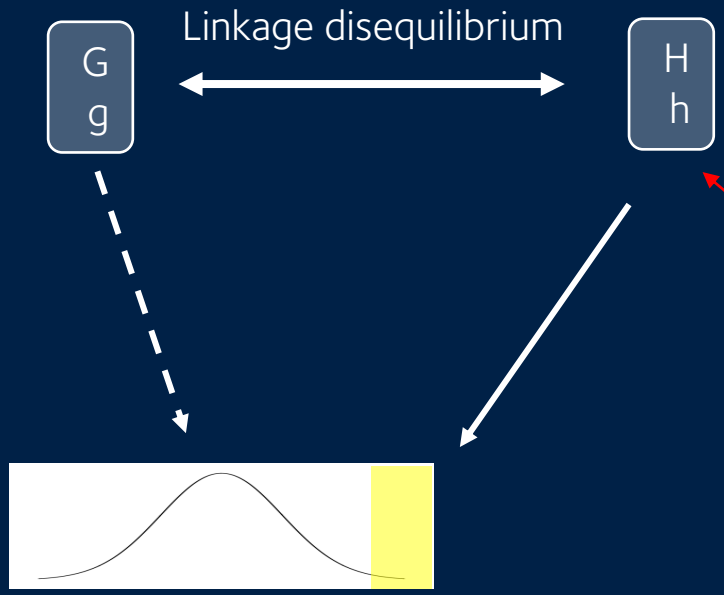


E.g. different rates of sampling cases in different ancestral backgrounds

Major confounder 3: linkage disequilibrium

Association tests capture all causal paths from genotype to phenotype – even those that have nothing to do with biology.

Will also pick up effects from all nearby causal variants that are in LD



If this variant were the causal one, we'd still detect it through G/g

How to run a GWAS

1. Collect DNA samples from as many disease cases and population controls as possible

The se formula can be used to estimate how are many needed for a given effect size and variant frequency

2. Genotype at variants genome-wide and estimate relative risk by computing the odds ratio, standard error and P-value.

3. If P is small enough, bingo!

How to run a GWAS, take 2

1. Collect DNA samples from as many disease cases and population controls as possible

The se formula can be used to estimate how are many needed for a given effect size and variant frequency

2. Genotype at variants genome-wide and perform careful quality control

What to genotype and how?

3. Estimate relative risk by computing the odds ratio, standard error and P-value, controlling for potential confounders

4. If P is small enough, there may be a link between genotype and phenotype

5. Attempt to replicate or find other corroborating evidence

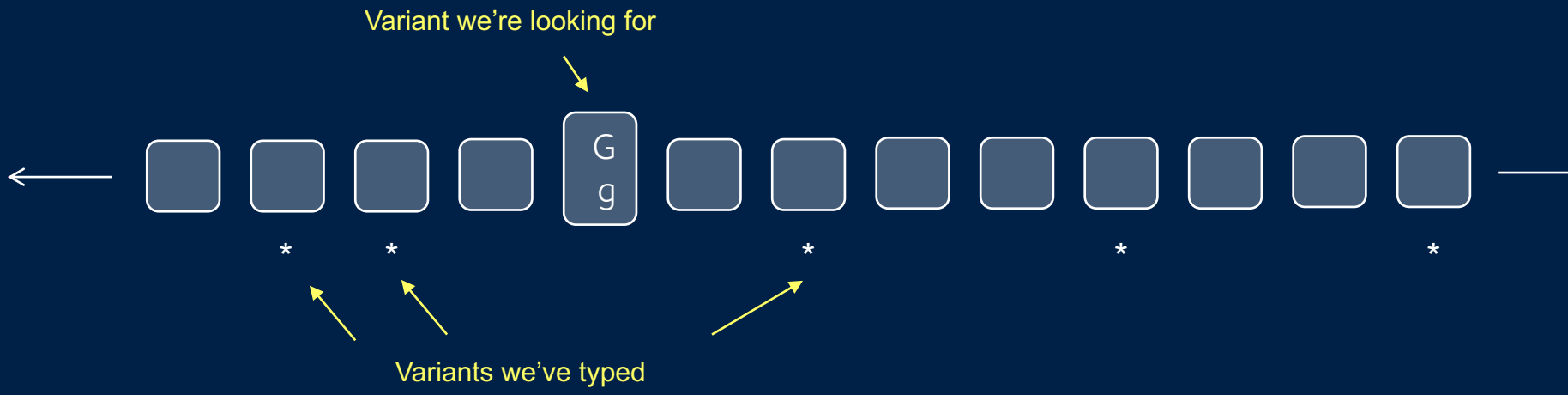
(+ now try to understand the underlying biology…)

Lecture outline

- Why GWAS? Heritability & genetic architecture
- Testing for association
- → What to genotype, and how? LD and the HapMap study
- A real GWAS study
- The challenge of understanding biology

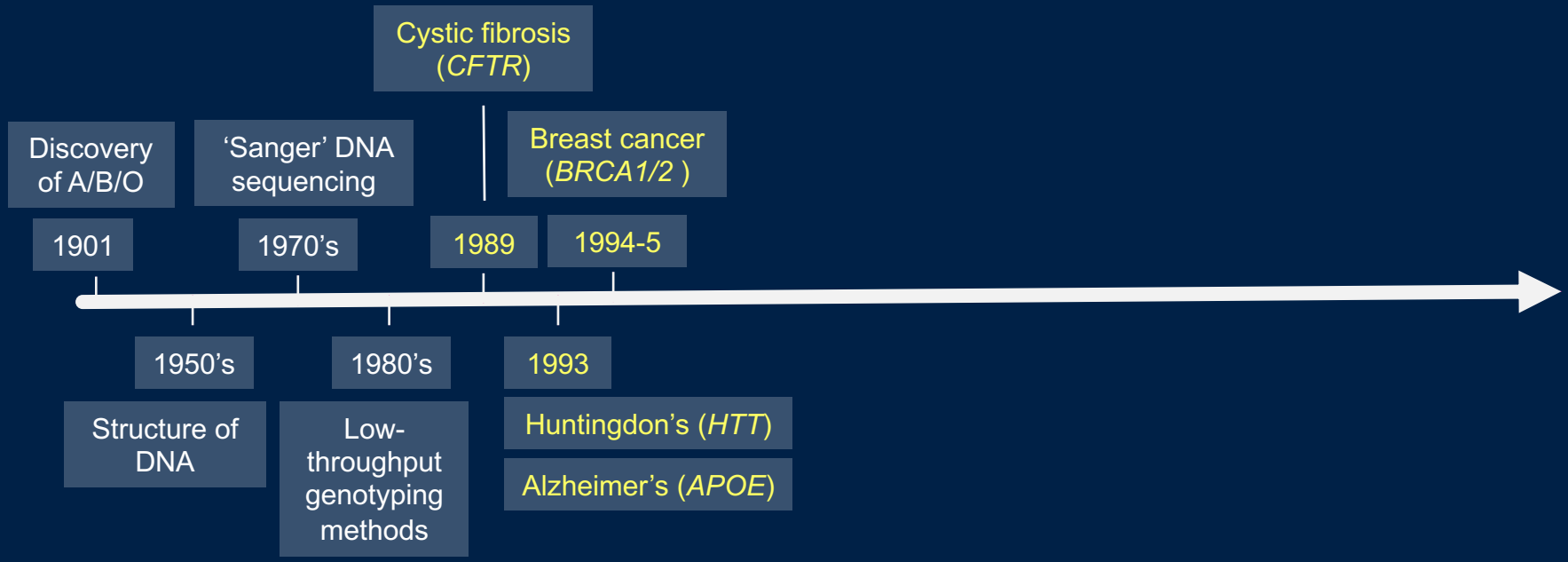
In an ideal world we would just genome sequence everyone and capture all variation.

In practice that is too expensive. We must make do with genotyping a subset of a few million genetic variants instead.

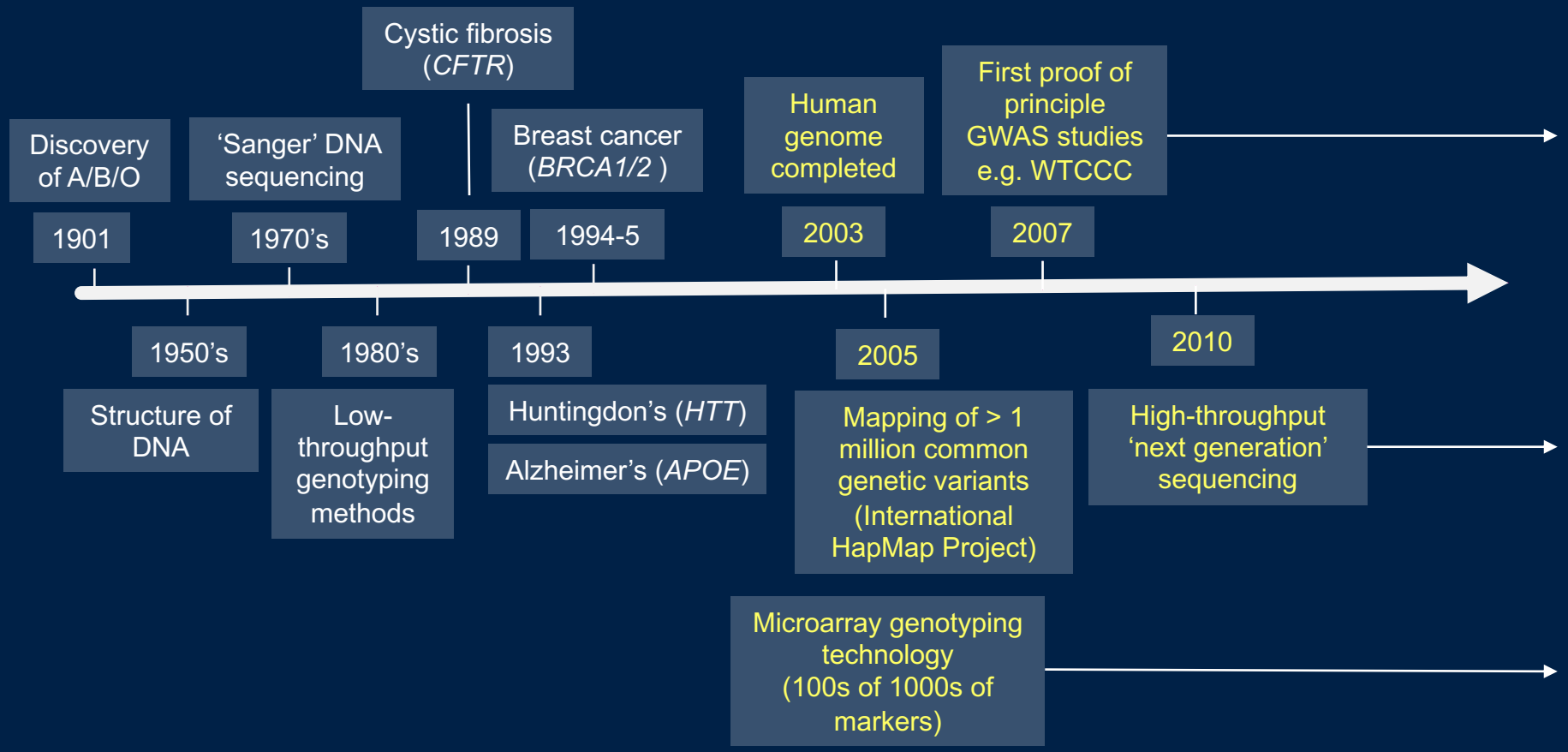


How does this work?

End of the linkage era



The birth of GWAS



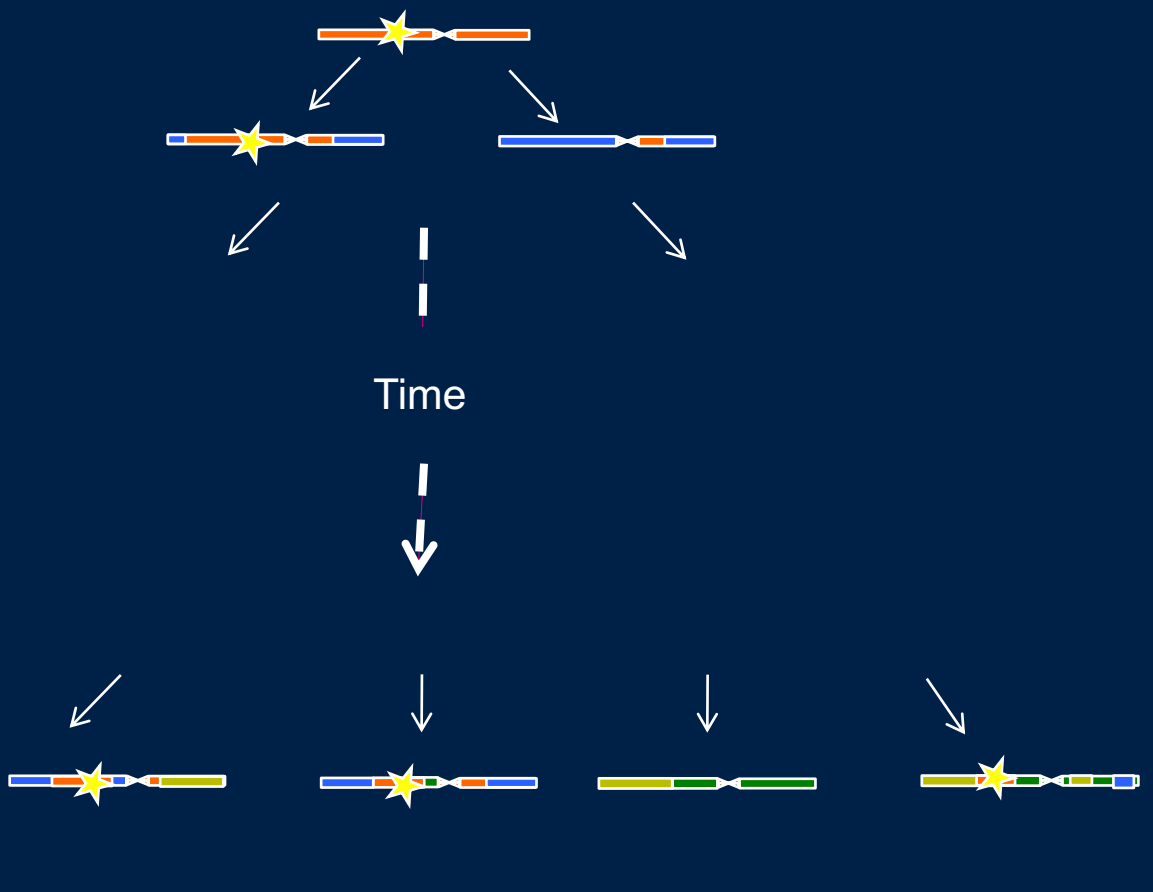
We expect linkage disequilbrium between the causal mutation and nearby variants.

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

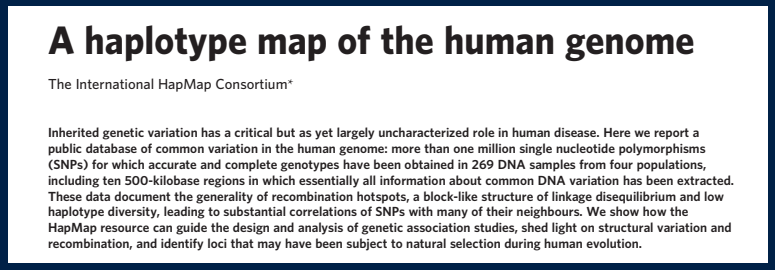


Patterns of LD depend on overall population size.

There are higher levels of LD in smaller populations.

The HapMap project estimated LD

The extent of LD depends on the amount of recombination.



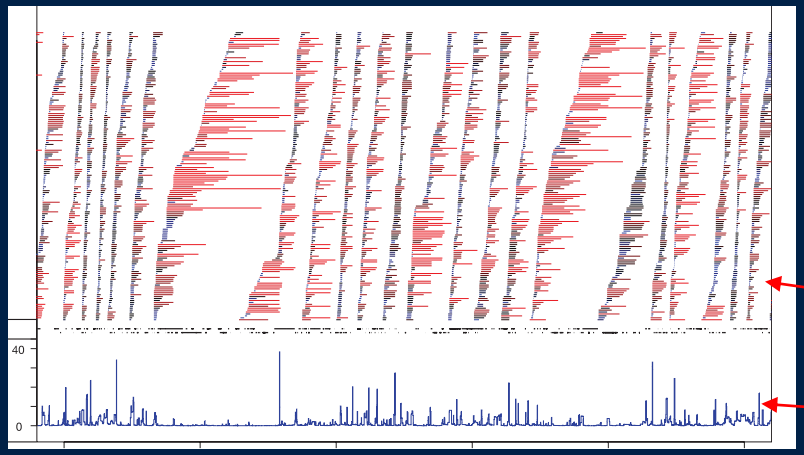
A haplotype map of the human genome

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)

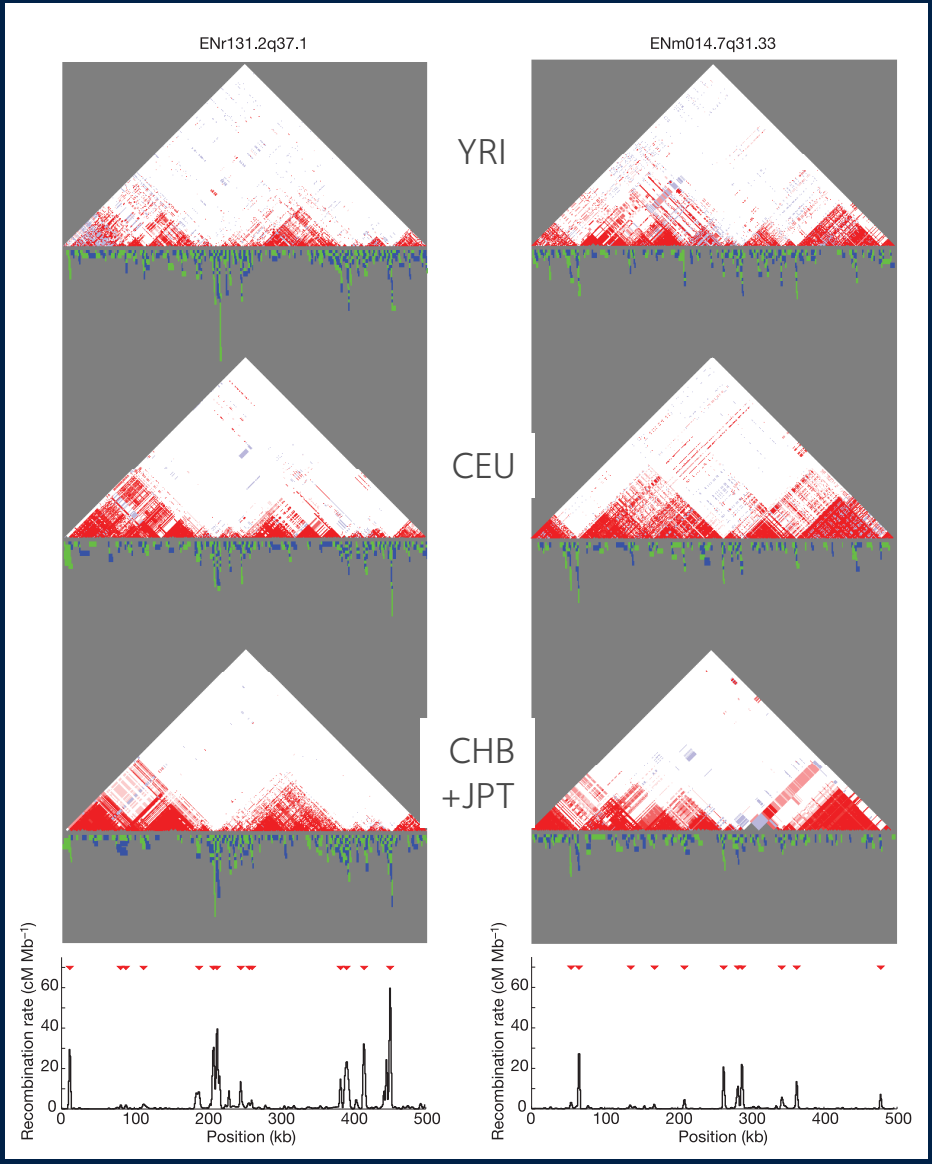
A database of > 1M SNPs found in European, African, and Asian ancestry individuals
(A subset of the samples later used in the 1000 Genomes Project)



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



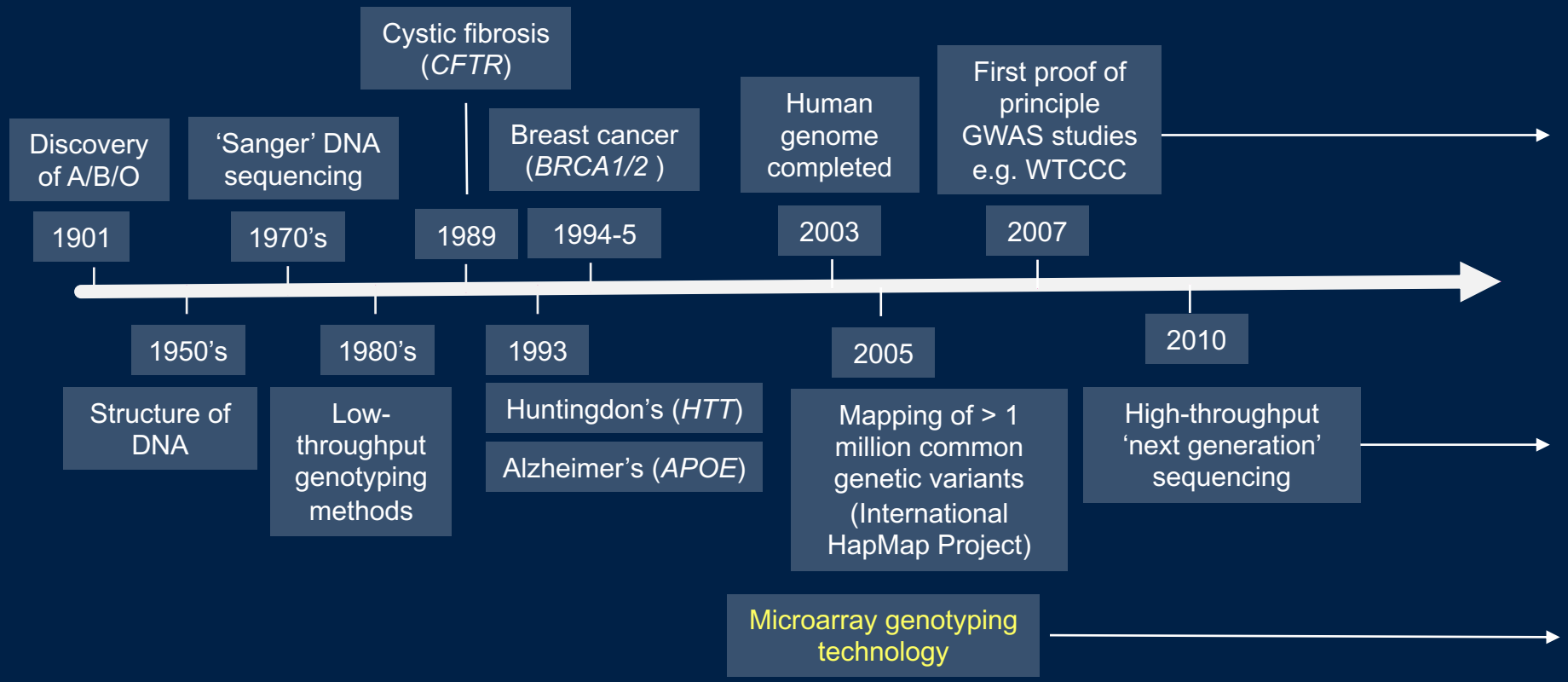
Block-like structure of LD (correlations between SNPs in two different regions)

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 $r^2 \geq 0.8$.

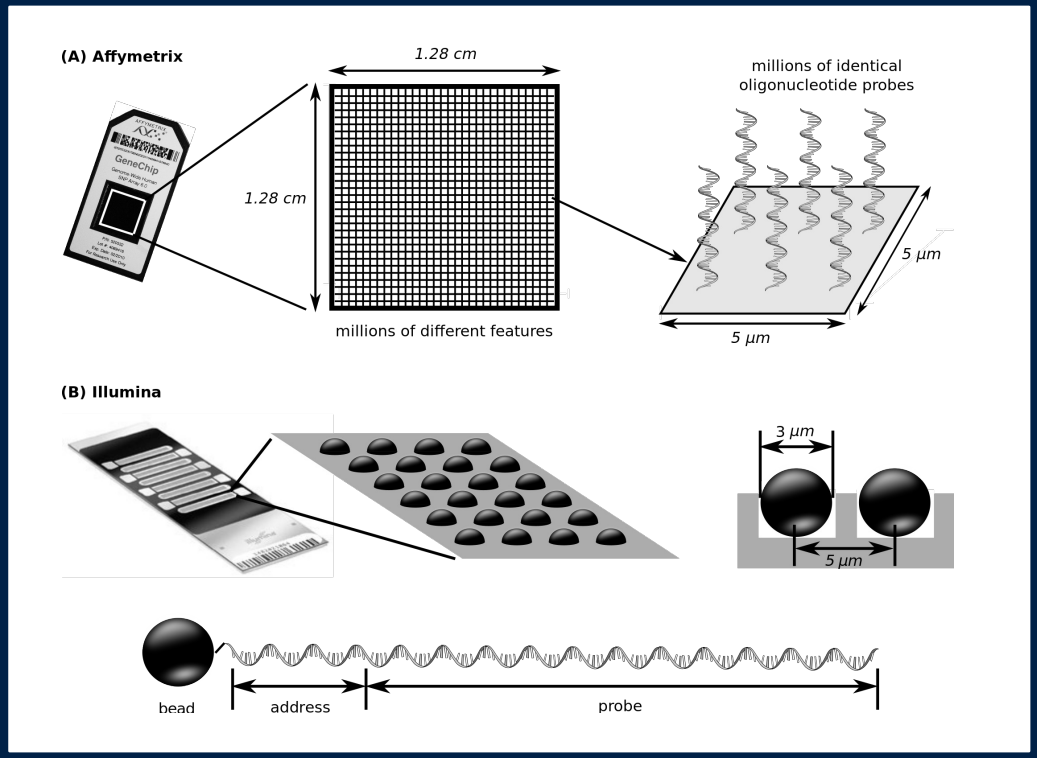
HapMap estimated how many SNPs genome-wide would need to be typed to capture (by LD) most common genetic variants. E.g. 250,000 would capture ~95% of SNPs in European populations.

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

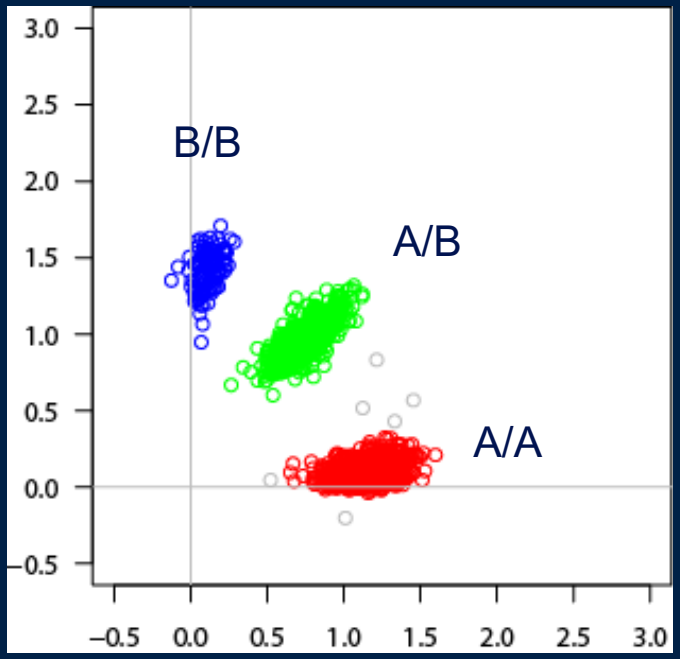


Wash the DNA over and let it hybridise to millions of probes – one for each SNP

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

A microarray gives you intensities, not genotypes

For each SNP, you get back this:



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

An algorithm is needed to turn the intensity values (x/y axis values) into genotype calls (colours).

Typical studies use microarrays to genotype hundreds of thousands to millions of genetic markers genome-wide.

They rely on patterns of LD to ‘access’ all the remaining genome-wide variation.

Imputatation may also be used to effectively extend the number of variants accessed.

Lecture outline

- Why GWAS? Heritability & genetic architecture
- Testing for association
- What to genotype, and how? LD and the HapMap study
- → A real GWAS study: WTCCC
- The challenge of understanding biology

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?

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)

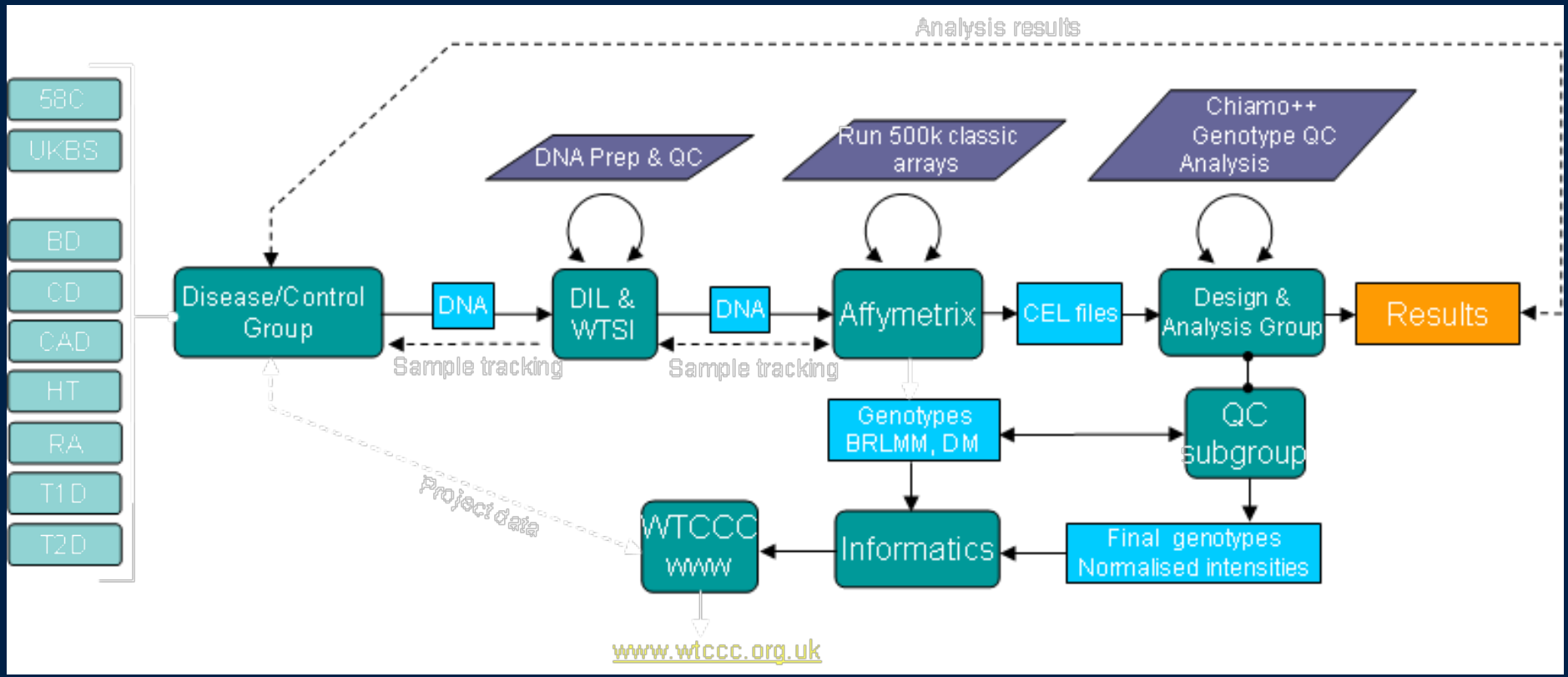
Studied seven common diseases in the UK

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.

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 strong evidence?

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

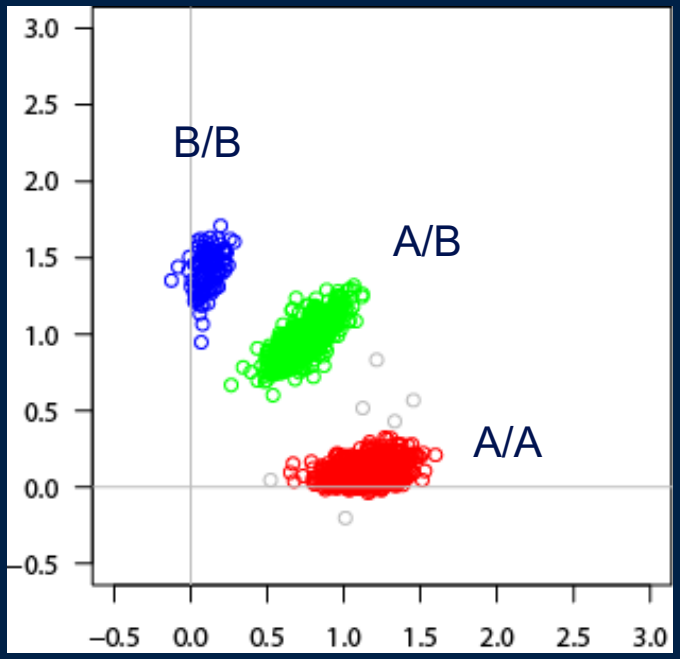
Is it convincing?

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

What about biology?

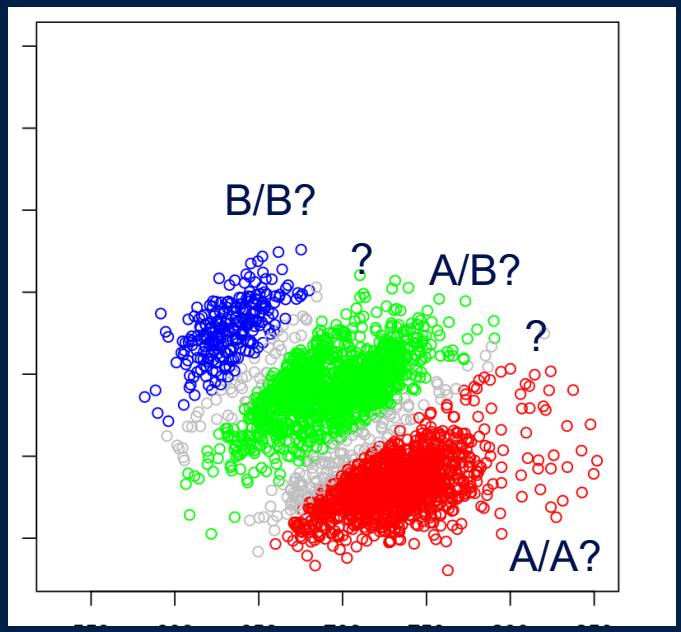
A microarray gives you intensities, not genotypes

For each SNP, you get back this:



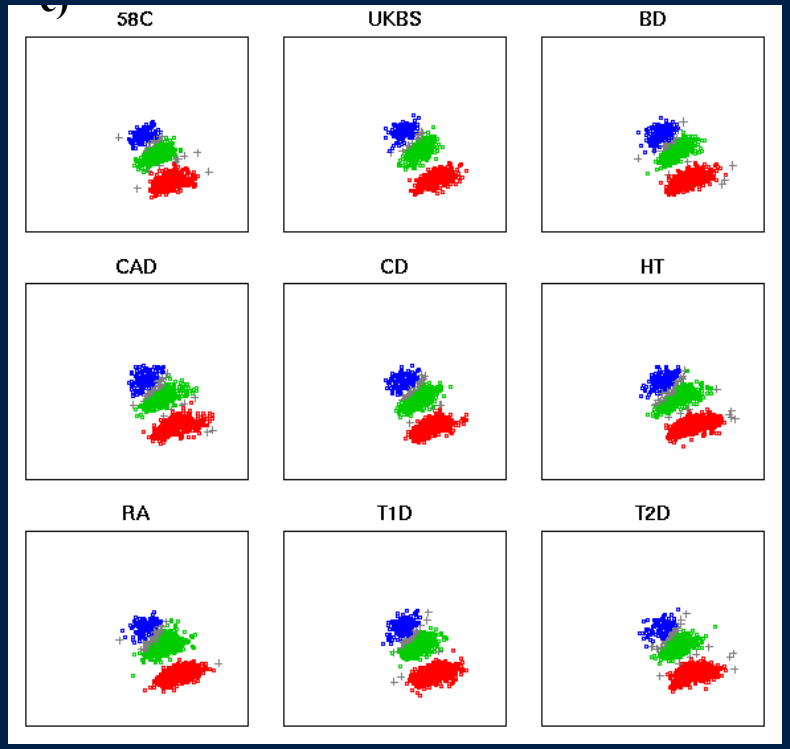
Each dot represents DNA from one individual.
X axis = image intensity for 1st allele probe
Y axis = image intensity for 2nd allele probe

Or this if you're less lucky:



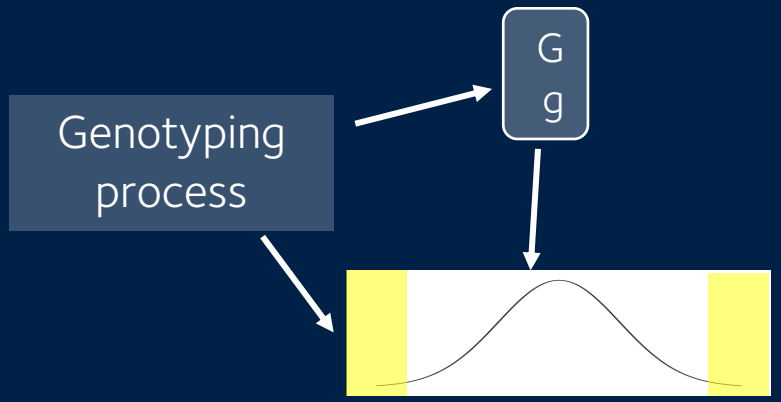
Small genotyping errors in cases or controls could easily confound the study

An algorithm is needed to call genotypes



The authors developed a genotype calling algorithm to turn these data (intensities, X and Y axis) into genotype calls (colours). Samples lying outside clusters, or in overlapping clusters, would be called as missing. (NB. Nowadays most studies use off-the-shelf algorithms for this.)

In particular cases and controls were jointly called.



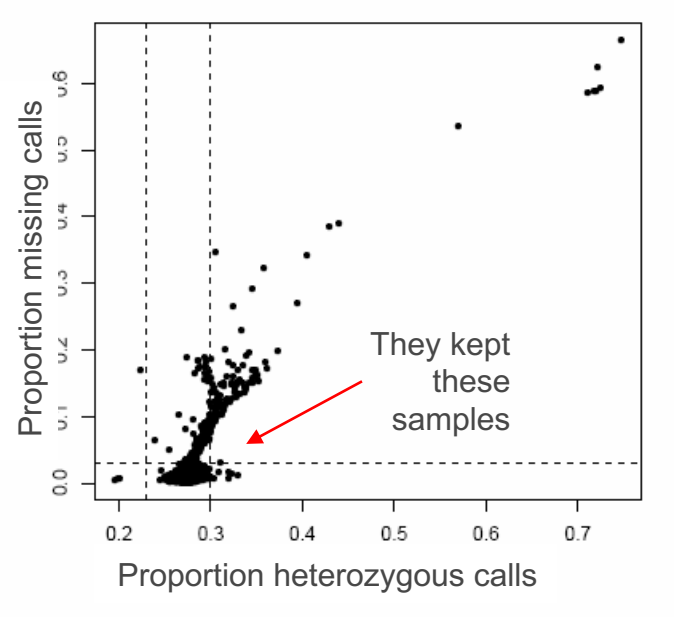
Collection	Missingness	Heterozygosity	External discordance	Non-European ancestry	Duplicate	Relative	Total
58C	9	0	4	6	4	1	**24**
UKBS	8	0	5	14	0	15	**42**
BD	30	0	0	9	77	13	**129**
CAD	41	1	0	13	2	5	**62**
CD	43	4	6	54	131	18	**256**
HT	29	0	0	2	6	11	**48**
RA	47	1	0	26	53	9	**136**
T1D	7	2	1	18	6	3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

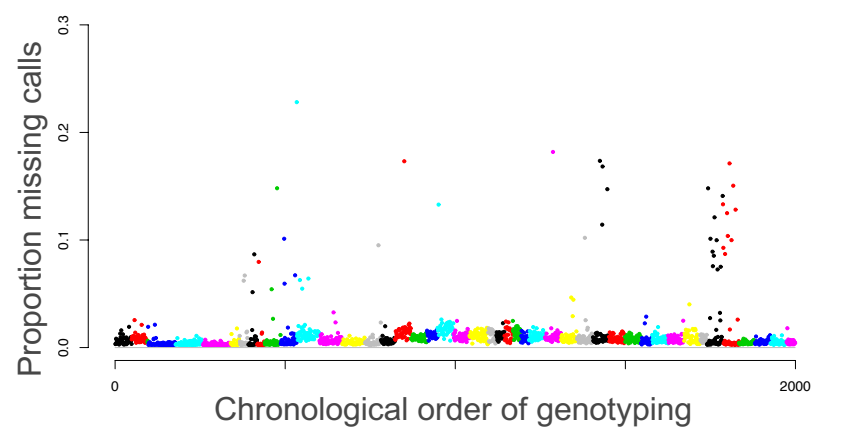
Supplementary Table 4 | Exclusion summary by collection. Six filters were applied for sample exclusion: 1. SNP call rate < 97% (missingness). 2. Heterozygosity > 30% or < 23% across all SNPs. 3. External discordance with genotype or phenotype data. 4. Individuals identified as having recent non-European 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!

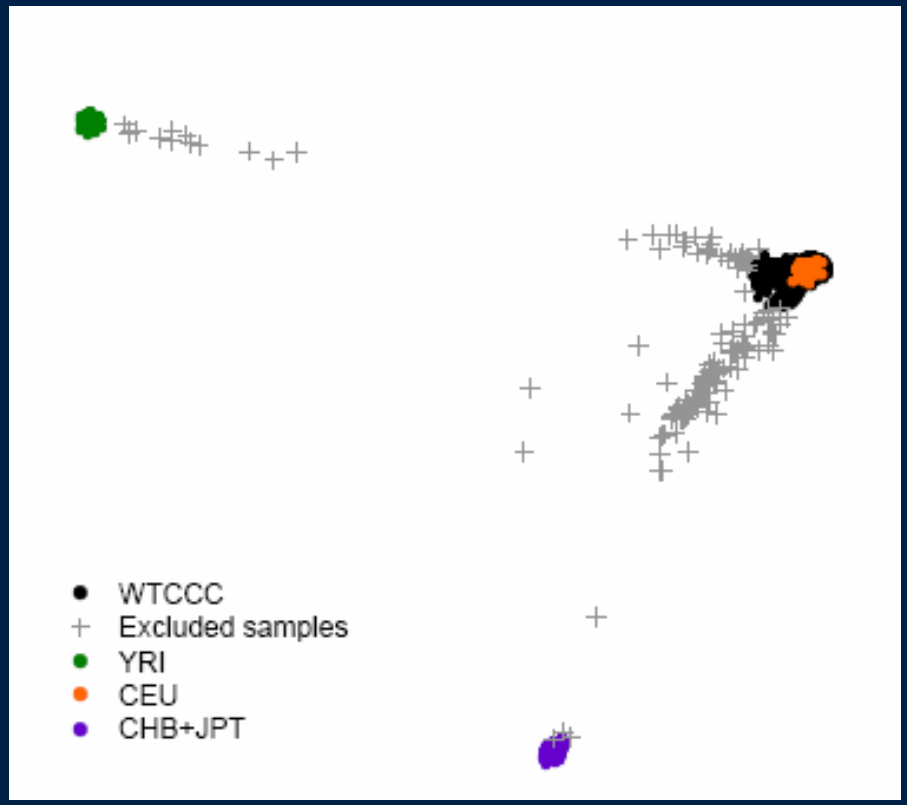
Due to:

- 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.

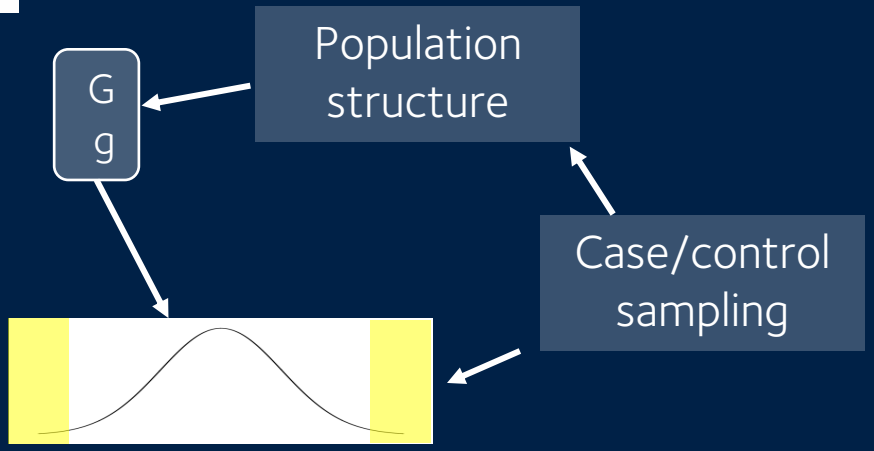


PCA computes genome-wide relationships between samples and then looks for directions of greatest variation. Since relatedness typically decreases with geographic distance, principal components typically reflect geography.

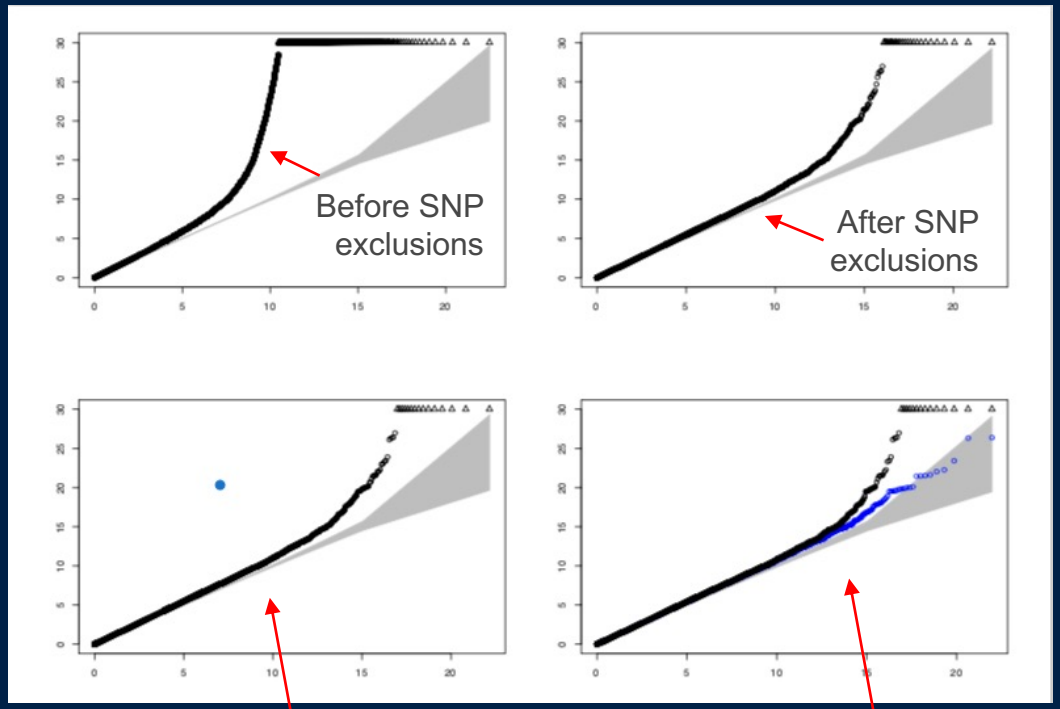
To avoid confounding by population structure, the samples were all supposed to be from the United Kingdom, and with European ancestry.

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



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

After visually inspecting cluster plots for remaining associated SNPs

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

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.

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.

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.

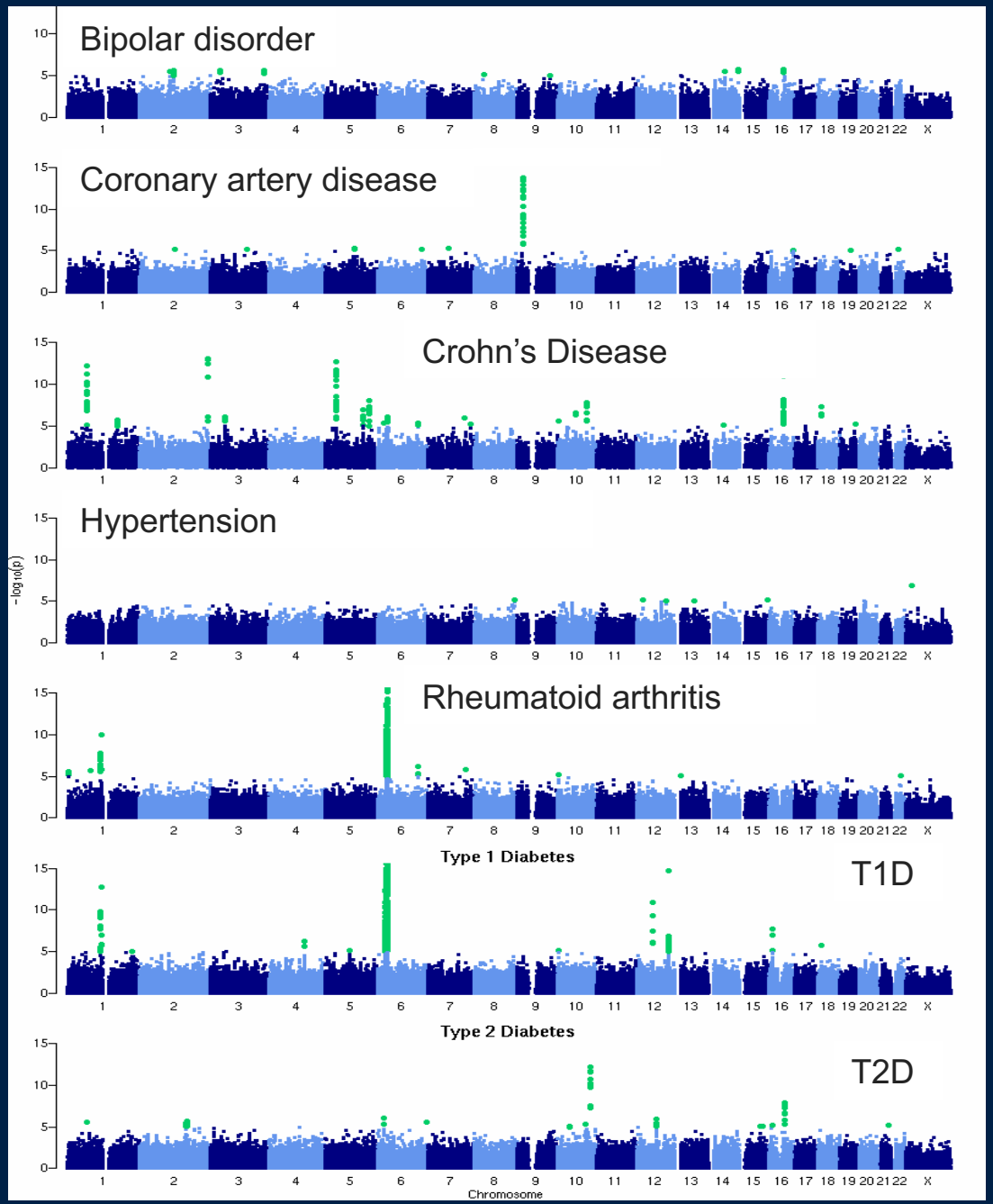
Did they find anything with strong evidence?

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

Is it convincing?

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

What about biology?



Number of associations with strong evidence

1

1

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

9

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

0

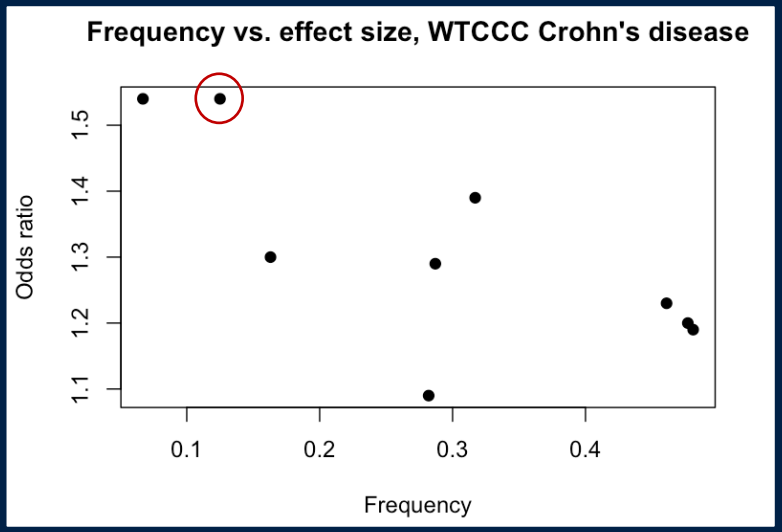
3

The traits clearly differ in their genetic architecture

7

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

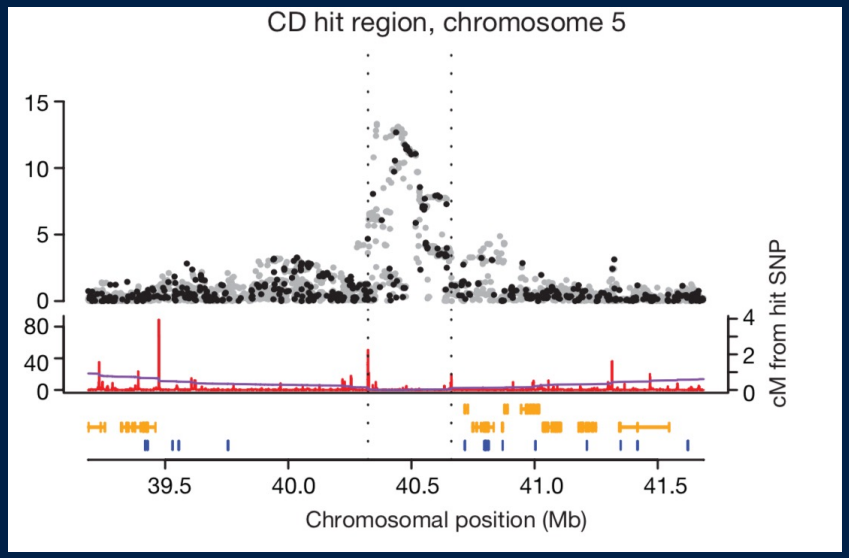
3



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 observed in the MHC locus)



Zooming into these associations vies us a more detailed picture of the regional association – here shown for the strong association on chromosome 5.

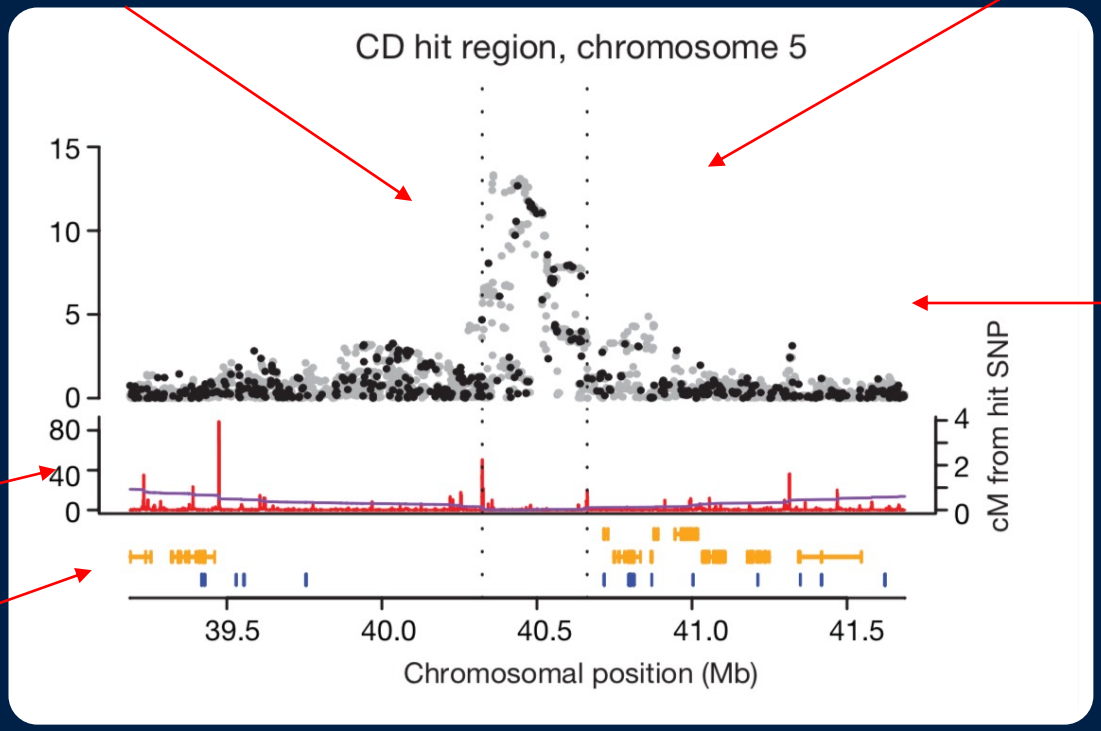
Zooming in to a GWAS 'hit' plot

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

Table 4 | Regions of the genome showing moderate evidence of association

Collection	Chromosome	Region (Mb)	SNP	Trend *P* value	Genotypic *P* value	$\log_{10}$(BF) additive	$\log_{10}$(BF) general	Risk allele	Minor allele	Heterozygote odds ratio	Homozygote odds ratio	Control MAF	Case MAF
BD	2p25	11.94–12.00	rs4027132	1.31×10^{-05}	9.68×10^{-06}	3.07	2.84	A	G	1.39 (1.19–1.64)	1.51 (1.27–1.79)	0.459	0.414
BD	2q12	104.41–104.58	rs7570682	3.11×10^{-06}	1.64×10^{-05}	3.68	3.23	A	A	1.23 (1.09–1.40)	1.64 (1.28–2.12)	0.214	0.255
BD	2q14	115.63–116.11	rs1375144	2.43×10^{-06}	1.31×10^{-05}	3.80	2.92	A	G	1.32 (1.07–1.63)	1.59 (1.29–1.96)	0.337	0.291
BD	2q37	241.23–241.28	rs2953145	1.11×10^{-05}	6.57×10^{-06}	3.22	3.50	C	G	1.84 (1.31–2.58)	2.14 (1.53–2.98)	0.226	0.189
BD	3p23	32.26–32.33	rs4276227	4.57×10^{-06}	2.62×10^{-05}	3.52	3.04	C	T	1.20 (0.99–1.46)	1.49 (1.23–1.81)	0.371	0.326
BD	3q27	184.29–184.40	rs683395	2.30×10^{-06}	5.11×10^{-06}	3.87	3.73	G	G	1.47 (1.26–1.71)	1.30 (0.69–2.46)	0.080	0.109
BD	6p21	42.82–42.86	rs6458307	3.43×10^{-01}	4.35×10^{-06}	−0.80	2.84	T	T	0.84 (0.75–0.96)	1.39 (1.13–1.69)	0.312	0.321
BD	8p12	34.22–34.61	rs2609653	6.86×10^{-06}	-	3.44	3.21	C	C	1.43 (1.19–1.71)	3.62 (1.26–10.44)	0.052	0.074
BD	9q32	114.31–114.39	rs10982256	8.80×10^{-06}	4.41×10^{-05}	3.23	2.37	T	C	1.26 (1.08–1.47)	1.47 (1.24–1.74)	0.471	0.425
BD	14q22	57.17–57.24	rs10134944	3.21×10^{-06}	6.89×10^{-06}	3.73	3.59	T	T	1.45 (1.24–1.68)	1.32 (0.74–2.33)	0.086	0.115
BD	14q32	103.43–103.62	rs11622475	2.10×10^{-06}	8.14×10^{-06}	3.87	3.24	C	T	1.13 (0.89–1.44)	1.47 (1.17–1.86)	0.300	0.256
BD	16q12	51.36–51.50	rs1344484	1.64×10^{-06}	1.03×10^{-05}	3.94	3.41	T	C	1.24 (1.03–1.48)	1.52 (1.27–1.82)	0.402	0.353
BD	20p13	3.70–3.73	rs3761218	4.43×10^{-05}	6.71×10^{-06}	2.58	3.18	T	C	0.97 (0.81–1.15)	1.31 (1.09–1.57)	0.397	0.356
CAD	1q43	236.77–236.85	rs17672135	1.04×10^{-04}	2.35×10^{-06}	2.36	3.88	T	C	0.70 (0.61–0.81)	1.32 (0.79–2.22)	0.134	0.108
CAD	5q21	99.98–100.11	rs383830	5.72×10^{-06}	1.34×10^{-05}	3.49	3.26	T	A	1.60 (1.16–2.21)	1.92 (1.40–2.63)	0.220	0.182
CAD	6q25	151.34–151.42	rs6922269	6.33×10^{-06}	1.50×10^{-05}	3.38	3.14	A	A	1.17 (1.04–1.32)	1.65 (1.32–2.06)	0.253	0.294
CAD	16q23	81.72–81.79	rs8055236	9.73×10^{-06}	5.60×10^{-06}	3.28	3.59	G	T	1.91 (1.33–2.74)	2.23 (1.56–3.17)	0.198	0.162
CAD	19q12	34.74–34.78	rs7250581	9.12×10^{-06}	2.50×10^{-05}	3.30	2.87	G	A	1.06 (0.79–1.43)	1.40 (1.05–1.86)	0.220	0.182
CAD	22q12	25.01–25.06	rs688034	6.90×10^{-06}	3.75×10^{-06}	3.33	3.15	T	T	1.11 (0.98–1.25)	1.62 (1.34–1.95)	0.310	0.355
CD	1q24	169.53–169.67	rs12037606	1.79×10^{-06}	1.09×10^{-05}	3.89	3.35	A	A	1.22 (1.07–1.40)	1.52 (1.28–1.82)	0.388	0.438
CD	5q23	131.40–131.90	rs6596075	5.40×10^{-07}	3.19×10^{-06}	4.54	4.01	C	G	1.55 (1.00–2.39)	2.06 (1.35–3.14)	0.166	0.127
CD	6p22	20.83–20.85	rs6908425	5.13×10^{-06}	1.10×10^{-05}	3.55	3.38	C	T	1.63 (1.18–2.25)	1.95 (1.43–2.67)	0.230	0.190
CD	6p21	32.79–32.91	rs9469220	8.65×10^{-07}	2.28×10^{-06}	4.19	3.92	A	A	1.14 (0.98–1.32)	1.52 (1.28–1.79)	0.481	0.534
CD	6q23	138.06–138.17	rs7753394	4.42×10^{-06}	2.59×10^{-05}	3.52	2.99	C	C	1.21 (1.04–1.40)	1.48 (1.25–1.76)	0.482	0.531
CD	7q36	147.62–147.70	rs7807268	6.89×10^{-06}	4.42×10^{-06}	3.33	3.58	G	G	1.38 (1.20–1.60)	1.47 (1.24–1.74)	0.462	0.509
CD	10p15	38.52–38.57	rs6601764	2.56×10^{-06}	8.95×10^{-06}	3.74	3.01	C	C	1.16 (1.01–1.33)	1.52 (1.28–1.80)	0.408	0.458
CD	19q13	50.89–51.07	rs8111071	6.14×10^{-06}	1.75×10^{-05}	3.48	3.29	G	G	1.47 (1.25–1.73)	1.28 (0.56–2.88)	0.070	0.096
HT	1q43	235.67–235.79	rs2820037	5.76×10^{-05}	7.66×10^{-07}	2.54	3.99	T	T	1.54 (1.03–2.31)	1.09 (0.74–1.62)	0.141	0.171
HT	8q24	140.17–140.35	rs6997709	7.88×10^{-06}	4.36×10^{-05}	3.32	2.60	G	T	1.20 (0.94–1.52)	1.49 (1.18–1.89)	0.285	0.244
HT	12p12	24.86–24.95	rs7961152	7.39×10^{-06}	3.03×10^{-05}	3.29	2.51	A	A	1.16 (1.01–1.32)	1.47 (1.25–1.74)	0.415	0.461
HT	12q23	100.52–100.58	rs11110912	9.18×10^{-06}	1.94×10^{-05}	3.27	3.11	G	G	1.33 (1.18–1.51)	1.34 (0.96–1.86)	0.165	0.200
HT	13q21	66.90–67.04	rs1937506	9.23×10^{-06}	4.53×10^{-05}	3.25	2.85	G	A	1.33 (1.04–1.69)	1.60 (1.26–2.02)	0.289	0.248
HT	15q26	94.60–94.67	rs2398162	7.85×10^{-06}	5.67×10^{-06}	3.33	3.40	A	G	0.97 (0.76–1.25)	1.31 (1.03–1.67)	0.258	0.218
RA	1p36	2.44–2.77	rs6684865	5.37×10^{-06}	3.14×10^{-05}	3.47	2.97	G	A	1.27 (1.02–1.56)	1.54 (1.25–1.90)	0.338	0.294
RA	1p31	80.16–80.36	rs11162922	1.80×10^{-06}	-	4.11	3.80	A	G	1.27 (0.41–4.01)	2.00 (0.64–6.20)	0.072	0.048
RA	4p15	24.99–25.13	rs3816587	7.65×10^{-03}	9.25×10^{-06}	0.50	2.64	C	C	0.91 (0.80–1.04)	1.35 (1.14–1.59)	0.406	0.434
RA	6q23	138.00–138.06	rs6920220	4.99×10^{-06}	1.58×10^{-05}	3.49	3.17	A	A	1.20 (1.06–1.36)	1.72 (1.33–2.22)	0.223	0.263
RA	7q32	130.80–130.84	rs11761231	1.74×10^{-06}	2.65×10^{-06}	3.92	3.42	C	T	1.44 (1.19–1.75)	1.64 (1.35–1.99)	0.375	0.327
RA	10p15	6.07–6.16	rs2104286	7.02×10^{-06}	2.52×10^{-05}	3.37	2.57	T	C	1.41 (1.10–1.81)	1.68 (1.31–2.14)	0.286	0.244
RA	13q12	19.845–19.855	rs9550642	8.44×10^{-06}	3.90×10^{-05}	3.35	3.02	A	A	1.34 (1.15–1.56)	2.23 (1.21–4.13)	0.084	0.112
RA	21q22	41.430–41.465	rs2837960	3.45×10^{-02}	1.68×10^{-06}	0.05	2.70	G	G	0.95 (0.83–1.08)	2.30 (1.64–3.23)	0.171	0.188
RA	22q13	35.870–35.885	rs743777	7.92×10^{-06}	1.15×10^{-06}	3.29	3.52	G	G	1.09 (0.97–1.24)	1.72 (1.40–2.11)	0.292	0.336
T1D	1q42	221.92–222.17	rs2639703	8.46×10^{-06}	1.74×10^{-05}	3.25	3.06	C	C	1.15 (1.02–1.30)	1.61 (1.31–1.99)	0.276	0.318
T1D	4q27	123.02–123.92	rs17388568	5.01×10^{-07}	3.27×10^{-06}	4.42	3.89	A	A	1.26 (1.11–1.42)	1.58 (1.27–1.95)	0.260	0.307
T1D	5q14	86.20–86.50	rs2544677	8.23×10^{-06}	4.43×10^{-05}	3.32	2.70	C	G	1.34 (1.00–1.79)	1.65 (1.24–2.18)	0.242	0.204
T1D	5q31	132.64–132.67	rs17166496	6.06×10^{-01}	5.20×10^{-06}	−0.97	3.25	C	G	0.77 (0.68–0.87)	1.09 (0.92–1.29)	0.391	0.386
T1D	10p15	6.07–6.18	rs2104286	7.96×10^{-06}	4.32×10^{-05}	3.31	2.88	T	C	1.30 (1.02–1.65)	1.57 (1.25–1.99)	0.286	0.245
T1D	12p13	9.71–9.80	rs11052552	1.02×10^{-04}	7.24×10^{-07}	2.22	3.80	G	T	1.49 (1.28–1.73)	1.43 (1.21–1.69)	0.486	0.446
T1D	18p11	12.76–12.91	rs2542151	1.89×10^{-06}	1.16×10^{-05}	3.91	3.52	G	G	1.30 (1.15–1.47)	1.62 (1.17–2.24)	0.163	0.201
T2D	1p31	66.04–66.36	rs4655595	2.68×10^{-06}	1.33×10^{-05}	3.81	3.47	G	G	1.37 (1.17–1.59)	2.33 (1.23–4.42)	0.080	0.108
T2D	2q24	160.90–161.17	rs6718526	2.40×10^{-06}	1.16×10^{-05}	3.86	3.35	C	T	1.49 (1.05–2.11)	1.86 (1.32–2.63)	0.209	0.171
T2D	3p14	55.24–55.32	rs358806	4.77×10^{-01}	3.05×10^{-06}	−0.83	2.72	A	A	0.86 (0.75–0.97)	1.78 (1.34–2.36)	0.198	0.204
T2D	4q27	122.92–123.02	rs7659604	2.1×10^{-02}	9.42×10^{-06}	0.13	2.74	T	T	1.35 (1.19–1.54)	1.09 (0.91–1.30)	0.380	0.403
T2D	10q11	43.43–43.63	rs9326506	7.78×10^{-06}	2.99×10^{-05}	3.27	2.92	C	C	1.28 (1.11–1.48)	1.46 (1.24–1.72)	0.492	0.538
T2D	12q13	49.50–49.87	rs12304921	5.37×10^{-02}	7.07×10^{-06}	−0.09	2.68	G	G	2.50 (1.53–4.09)	1.94 (1.20–3.15)	0.145	0.159
T2D	12q15	69.58–69.96	rs1495377	1.31×10^{-06}	6.52×10^{-06}	4.01	3.15	G	G	1.28 (1.11–1.49)	1.51 (1.28–1.78)	0.497	0.547
T2D	15q24	72.24–72.50	rs2930291	7.72×10^{-06}	4.40×10^{-05}	3.30	2.42	G	A	1.25 (1.04–1.51)	1.50 (1.24–1.82)	0.377	0.332
T2D	15q25	78.12–78.36	rs2903265	9.57×10^{-06}	4.98×10^{-05}	3.24	2.53	G	A	1.18 (0.93–1.49)	1.47 (1.17–1.86)	0.284	0.243

Regions with at least one SNP with a *P* value of greater than 5×10^{-7} and less than 1×10^{-5} for either the trend or the genotypic test. Columns as for Table 3. Cluster plots for each SNP have been inspected visually. Positions are in NCBI build-35 coordinates. Genotypic *P* values were not calculated for SNPs with the lowest MAFs owing to low numbers of rare-allele homozygotes and sensitivity to genotype calling errors.

The results above used a P-value threshold of $P < 5 \times 10^{-7}$

They also reported a longer list of association at lesser levels of evidence ($P < 5 \times 10^{-7}$). Many of these must be real as well.

How much statistical evidence do we really need? How did they choose a good threshold?

How to choose a P-value threshold

They reasoned like this: Based on what we know from HapMap, there are maybe 1 million 'LD blocks' in the human genome. Suppose maybe 10 of them, or so, are associated with the trait. Then the prior chance of association for a randomly chosen region (i.e. chosen 'hypothesis free') will be 10 in a million, i.e. plausibly

Prior odds = 1×10^{-5} **before** we see any data.

For a P-value threshold α it works out that:

$$\text{odds}(\text{associated}|\text{P} < \alpha) = \frac{\text{statistical power}}{\alpha} \times \text{prior odds}$$

=> If the statistical power is 50%, say, then setting $\alpha = \mathbf{5\times10^{-7}}$ will give a posterior odds of 10 to 1.

This was a good choice! All of their associations have subsequently replicated in larger studies.

Many GWAS use a more stringent $\alpha = 5\times10^{-8}$ threshold, while still others attempt to directly estimate the above (c.f. 'False discovery rate' methods).

Statistical power

The statistical power says "how likely are we to detect a true effect". It is essentially determined by:

- The true effect size β (which of course we don't know beforehand)

- The standard error, which we do know approximately $\text{se} \approx \frac{1}{\sqrt{N \times f(1-f) \times \phi(1-\phi)}}$

- And also the threshold α, which says 'how many standard errors away from zero do we need?

$$\text{power} \approx N \times f(1-f) \times \phi(1-\phi) \times \beta^2$$

The true, causal effect size

Sample size

Genotype frequency

Ratio of cases to controls in study

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.

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 strong evidence?

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

Is it convincing?

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

What about biology?

Summary

- GWAS is a very simple study design in principle - just genotyping a lot of cases and controls, and test for association. The hard parts are in the implementation details

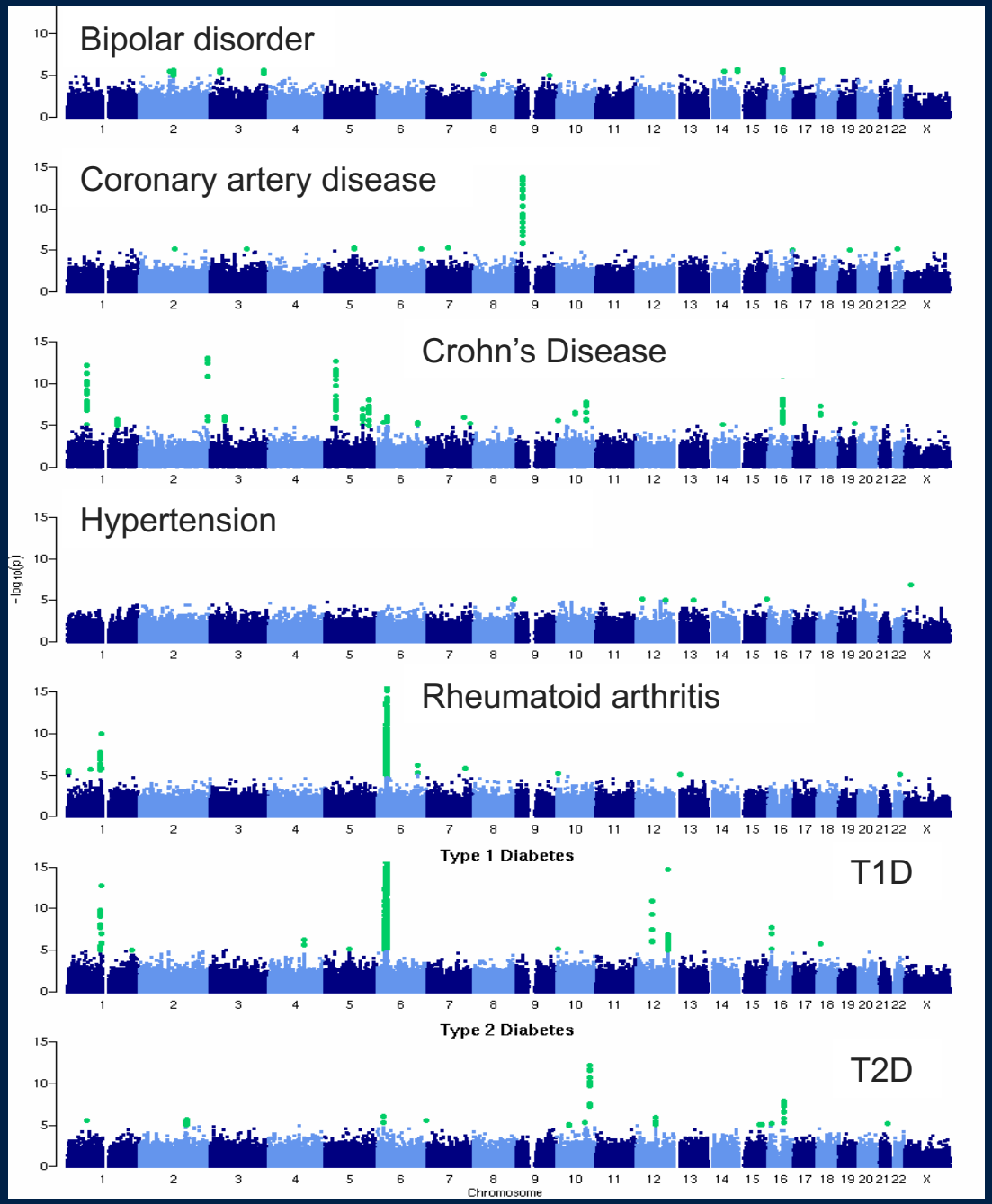
- In the early 2000's, The HapMap and other projects enabled the first GWAS by mapping SNPs genome-wide, and describing human haplotype variation.and patterns of LD. High-throughput genotyping microarray technology was developed to type these SNPs.

- The WTCCC was one of the first large GWAS studies. It provided compelling evidence that the 'common variant, common disease' hypothesis really holds.

- Although the overall design is simple, we are looking for small differences in risk between cases and controls (often RR = 1.5 or smaller). Consequently a lot of careful work is needed to ensure there is no subtle confounding – e.g. from sample collection, genotyping and data quality issues, or environmental covariates.

Lecture outline

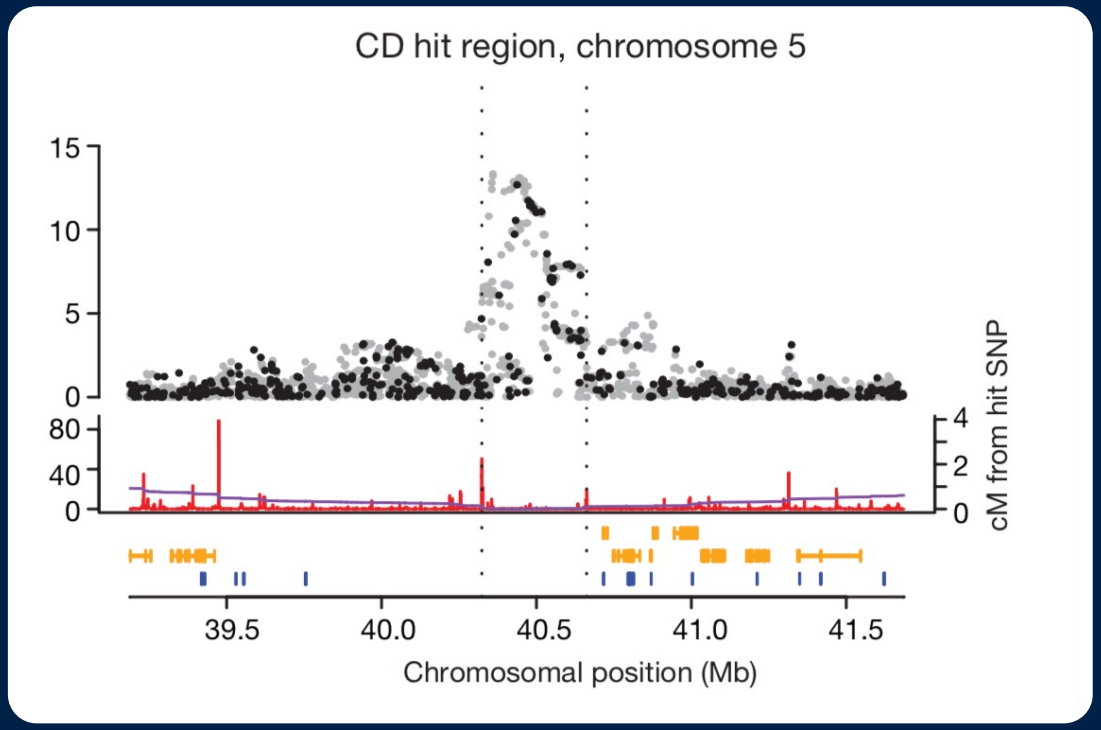
- Why GWAS? Heritability & genetic architecture
- Testing for association
- What to genotype, and how? LD and the HapMap study
- A real GWAS study: WTCCC
- → The challenge of understanding biology



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

What about the underlying causal variants?

The challenge of understanding biology



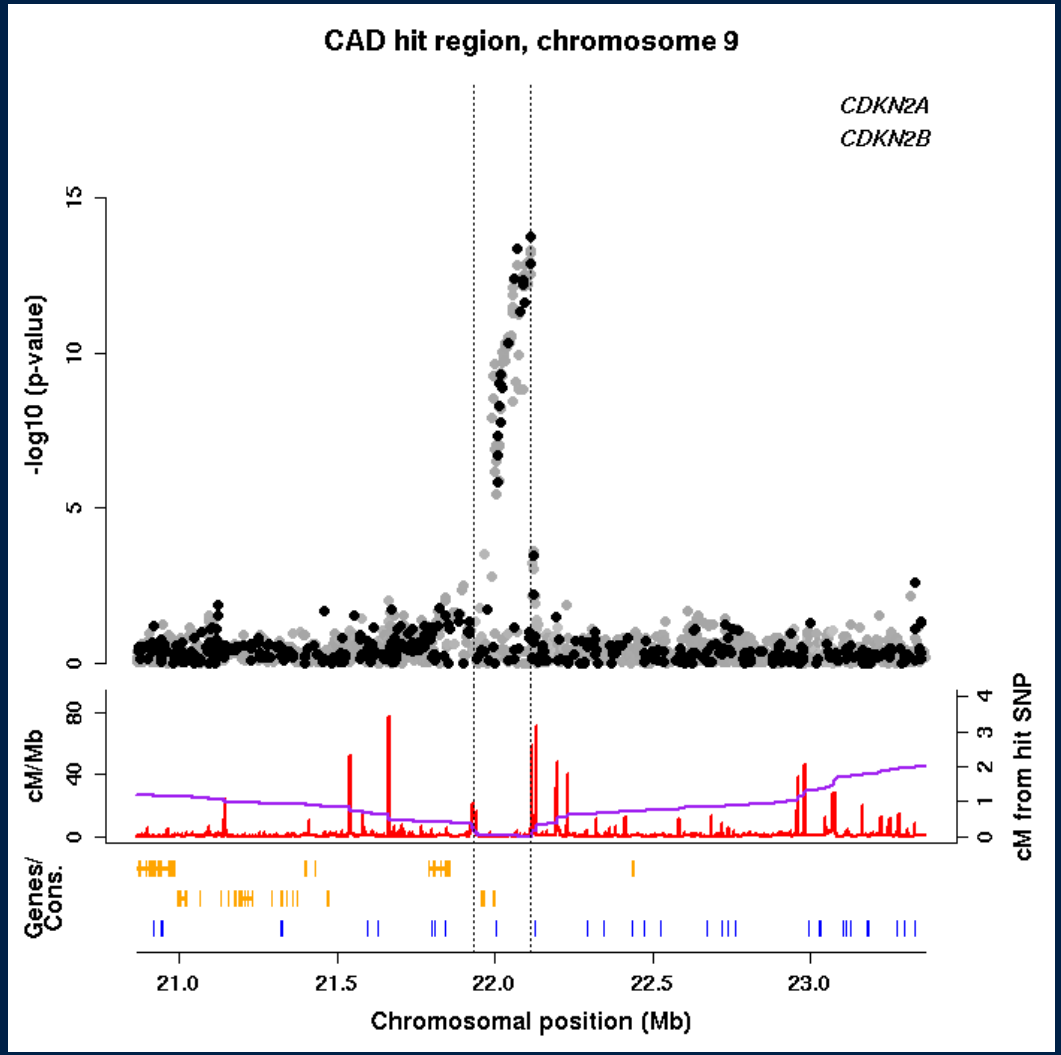
No genes under the main association signal!

Biology is complicated

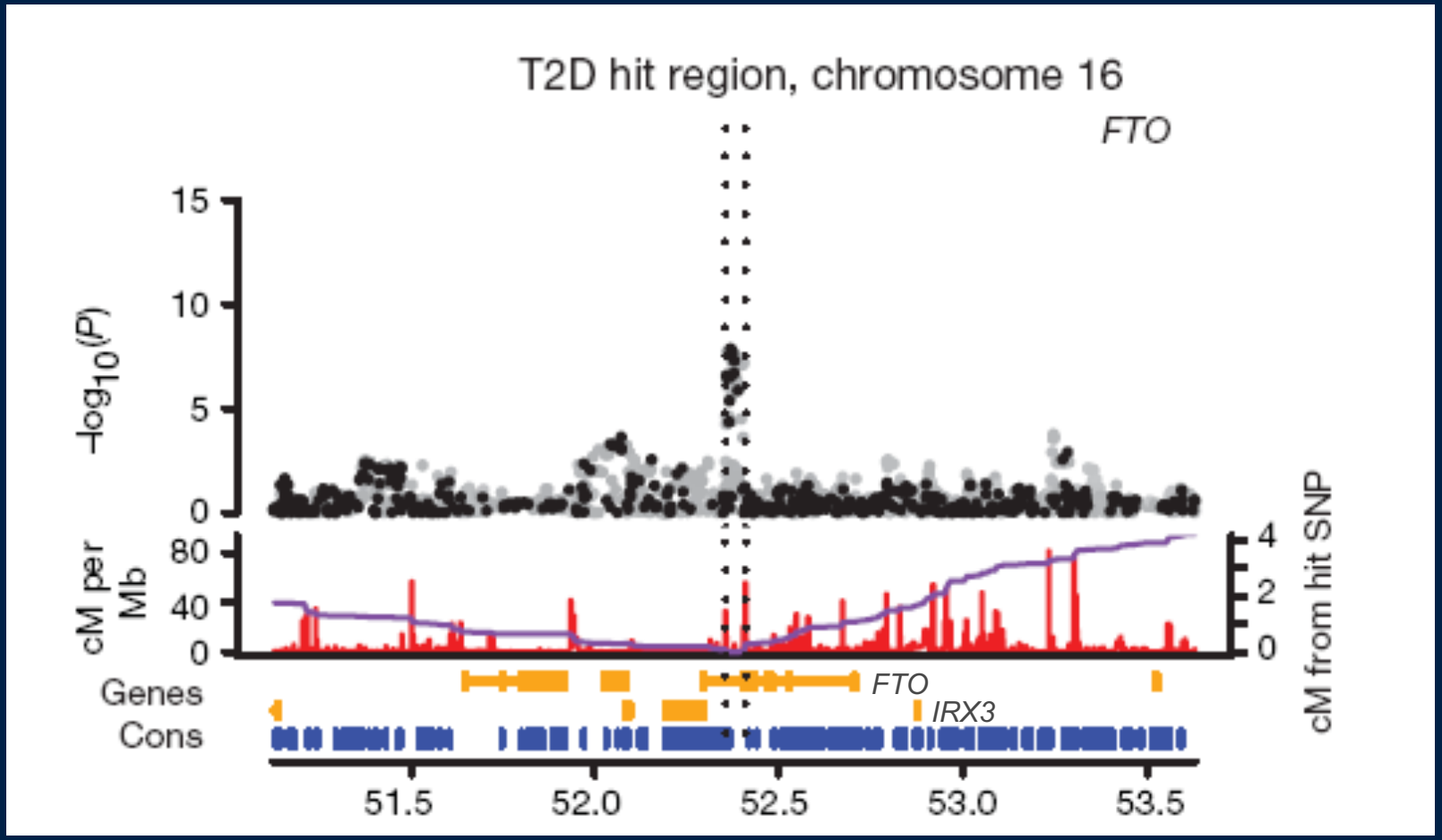
Association observed with CAD over a ~100kb region of chromosome 9. This is unquestionably a real association (it has been replicated in several independent studies).

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.

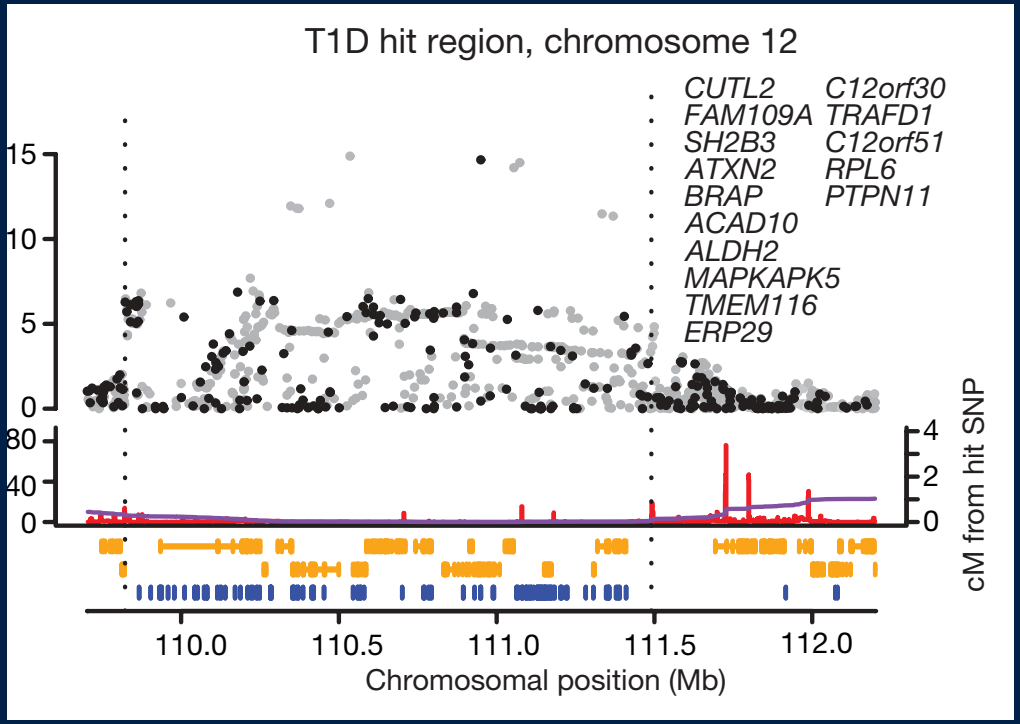


Harismendy et al Nature 2011; Almontishari, et al JACC 2013; Almonitishari et al, Circulation 2015



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.

How to read a GWAS – checklist

What is the sample size?

How are the samples genotyped? Are cases and controls typed in the same way?

What have the authors done to deal with potential confounders – good data quality control? Population structure? Is it convincing?

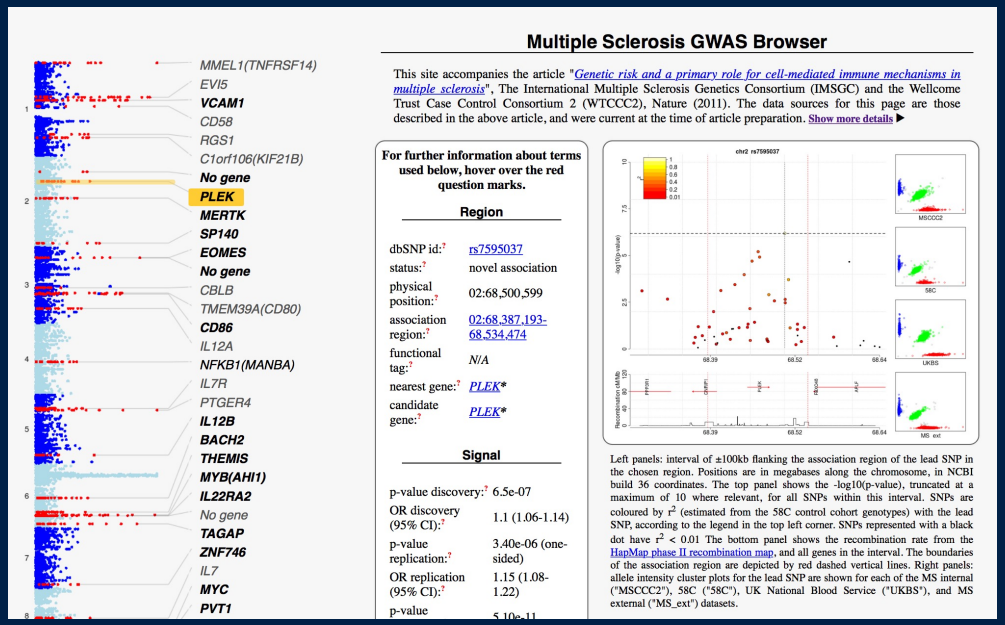
Do the results look sensible? Are the effect sizes reasonable? How strong is the evidence?

Does the signal replicate?

Does the association follow patterns of LD?

If all the above seem fine – what genes are nearby? Can you figure out biology?

Consolidation question



GWAS of multiple sclerosis (2011)

9772 cases, 17,376 controls from across Europe

www.well.ox.ac.uk/wtccc2/ms/
(I think this 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: Wednesay 28th Feb @11am

Genome-wide association studies II: Identifying genetic associations with complex traits

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

MSc Global Health Science and Epidemiology

Genetic Epidemiology Module

26th Feb 2024

CENTRE for HUMAN GENETICS