The University Of Queensland AUSTRALIA | CREATE CHANGE

MODULE 1 | GENETIC MAPPING

Session 2. Study design

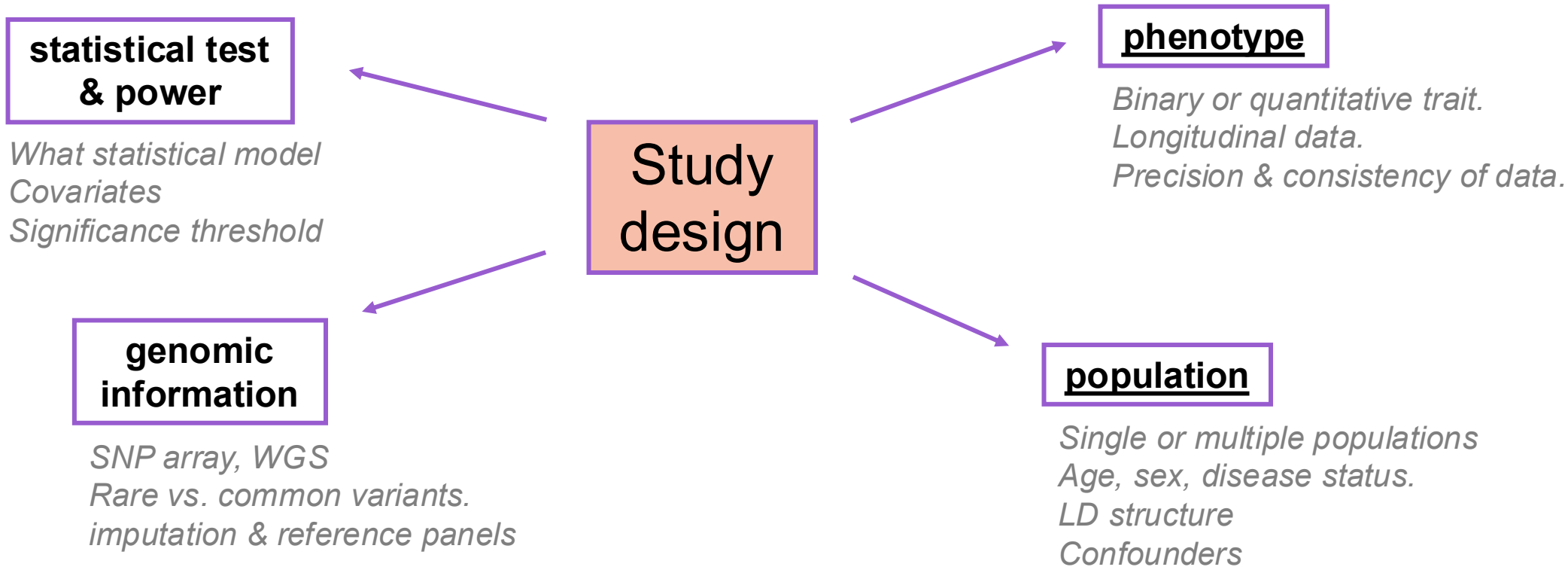
July 2026

Slides, practicals & data can be downloaded from the cluster:
`/data/module1/downloadsMonPM.zip`

Slides & Prac guide can be downloaded from the website:
https://cnsgenomics.com/data/teaching/GNGWS26/module1/

Study design – what is your question!

A set of decisions (made before you start) which determine what question can be answered by your study



Study design

Outcome: Students understand some of the key determinates of experimental power (and confounders) in GWAS

Outline:

Linkage disequilibrium

Factors affecting statistical power

- LD, effect size, sample size, significance threshold

Confounders

- population stratification, close relatives

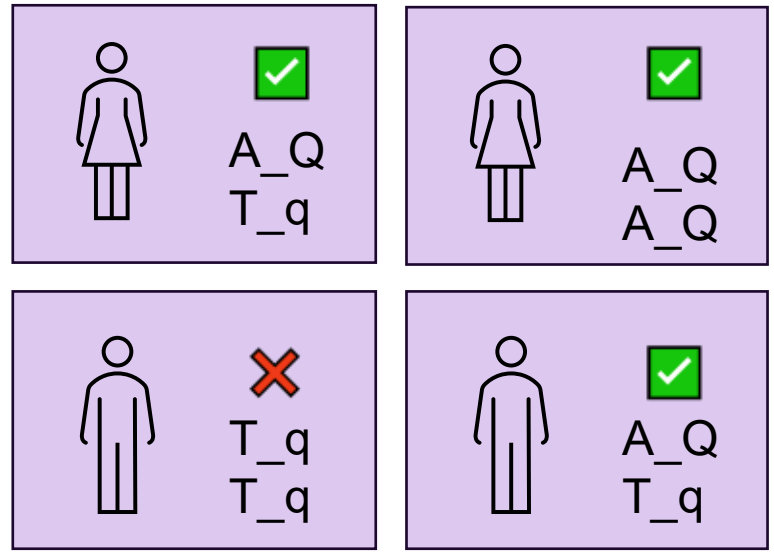
Linkage disequilibrium (LD)

GWAS depend on LD between SNP and causal variant

Hence, we can use a common set of SNP ('SNP chip') for many traits

What is linkage disequilibrium?

Do you like coriander?



Linkage Disequilibrium (LD)

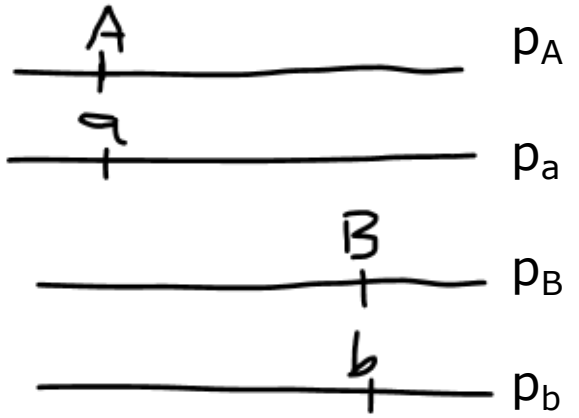
- **Linkage disequilibrium (LD)** is the non-random association between alleles at different locations in the genome.

- GWAS exploit LD between common SNP and ‘causative mutations’
  - the SNP associations in GWAS are (usually) *indirect* associations between the genome and the trait of interest

Definitions of LD

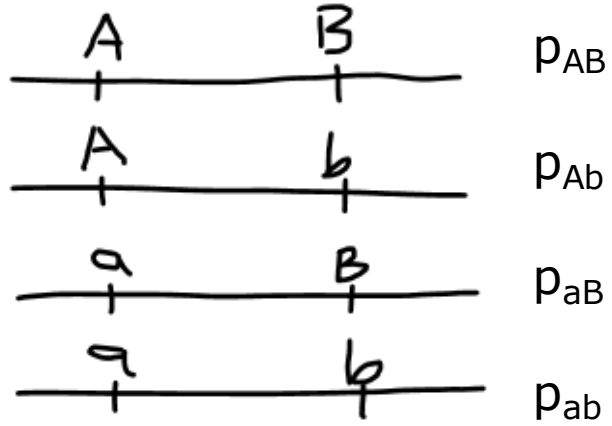
Imagine two bi-allelic SNP loci

- locus 1: alleles are A, a with frequency p_A and p_a
- locus 2: alleles are B, b with frequency p_B and p_b



And there are 4 possible haplotypes:

- A_B with frequency p_{AB}
- A_b with frequency p_{Ab}
- a_B with frequency p_{aB}
- a_b with frequency p_{ab}



https://web.stanford.edu/group/pritchardlab/HGbook.html

Definitions of LD

Under **linkage equilibrium**…….

The alleles are **independent** :- therefore the frequency of the haplotype is predicted by the frequency of the alleles, i.e.

$$p_{AB} = p_A \times p_B$$

$$p_{Ab} = p_A \times p_b$$

$$p_{aB} = p_a \times p_B$$

$$p_{ab} = p_a \times p_b$$

Definitions of LD

Under **linkage <u>dis</u>equilibrium**…….

Allele frequencies **do not** predict haplotype frequencies

$p_{AB} \neq p_A \times p_B$

$p_{Ab} \neq p_A \times p_b$

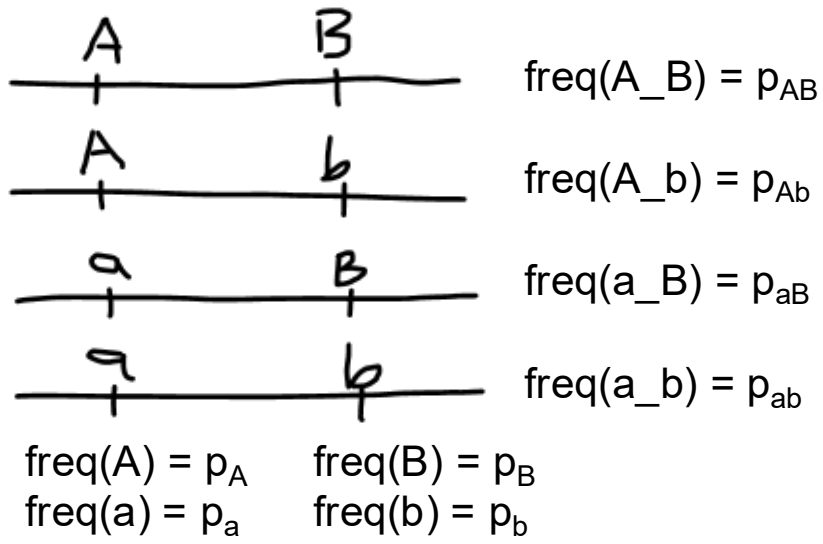
$p_{aB} \neq p_a \times p_B$

$p_{ab} \neq p_a \times p_b$

Definitions of LD

LD is the non-random association of alleles,

i.e. alleles are correlated



LD is measured through D’ (= D prime) and r^2 statistics

- measure difference between observed and expected frequencies

$$D = p_{AB} - p_A p_B$$

$$D' = D / D_{max}$$

$$r^2 = D^2 / p_A p_a p_B p_b$$

Definitions of LD

r^2 ~ **squared correlation co-efficient** between alleles

D’ ranges -1 to 1, and r^2 range from 0-1

D’ & r^2 = 0 means the two loci are independent

|D’| = 1 & r^2 = 1 means the two loci are in perfect LD. Only occurs if two of the possible four haplotypes are observed, i.e. only A_B / a_b or only A_b / a_B

In general,

r^2 & |D’| near 1	often co-inherited	Nearby each other on chromosome
r^2 & |D’| near 0	independent	Far away, maybe on different chromosomes

Definitions of LD

Why have two different measures of LD?

D’ tells you about ancestral recombination

r^2 tells you if knowing alleles at locus 1 is informative about locus 2

Sometimes they are different

For example, APOE alleles

Haplotype	Allele	Isoform	Observed frequency	Expected frequency*
CT	$\varepsilon 3r$	Arg^{112}-Cys^{158}	0.0001	0.0125
TT	$\varepsilon 2$	Cys^{112}-Cys^{158}	0.0804	0.0680
CC	$\varepsilon 4$	Arg^{112}-Arg^{158}	0.1554	0.1430
TC	$\varepsilon 3$	Cys^{112}-Arg^{158}	0.7641	0.7765

AD risk →

*frequency rs429358 C allele = 0.1547 and frequency rs7412 C allele = 0.9195; expected haplotype frequency assumes independence between the SNP.

e.g. missing haplotypes or when allele frequencies are very different

D_{CT} = freq(obs) – freq(exp) = - 0.0123

Note: $D_{CT} = -D_{TT} = -D_{CC} = D_{TC}$

$|D'_{CT}| = D_{CT}/max(D_{CT}) = 0.182$

$|D'_{TT}| = D_{TT}/max(D_{TT}) = 0.988$

$|D'_{CC}| = |D'_{TT}|$

$|D'_{TC}| = |D'_{CT}|$

$r^2 = D^2/p(C1)p(T1)\ p(C2)p(T2) = 0.016$

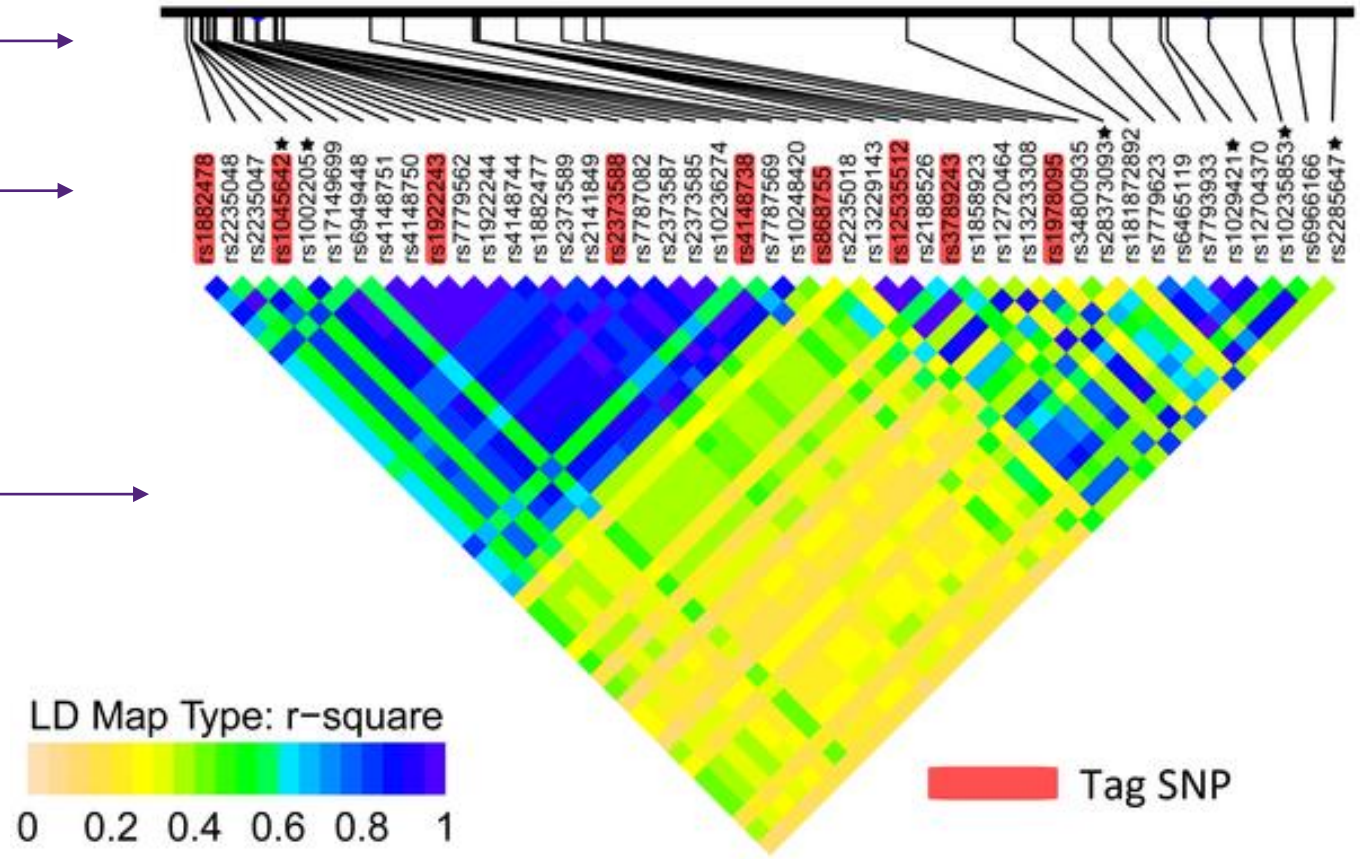
Visualising LD – pairwise LD

Many tools, e.g.

physical location →

marker name →

Pair-wise LD →



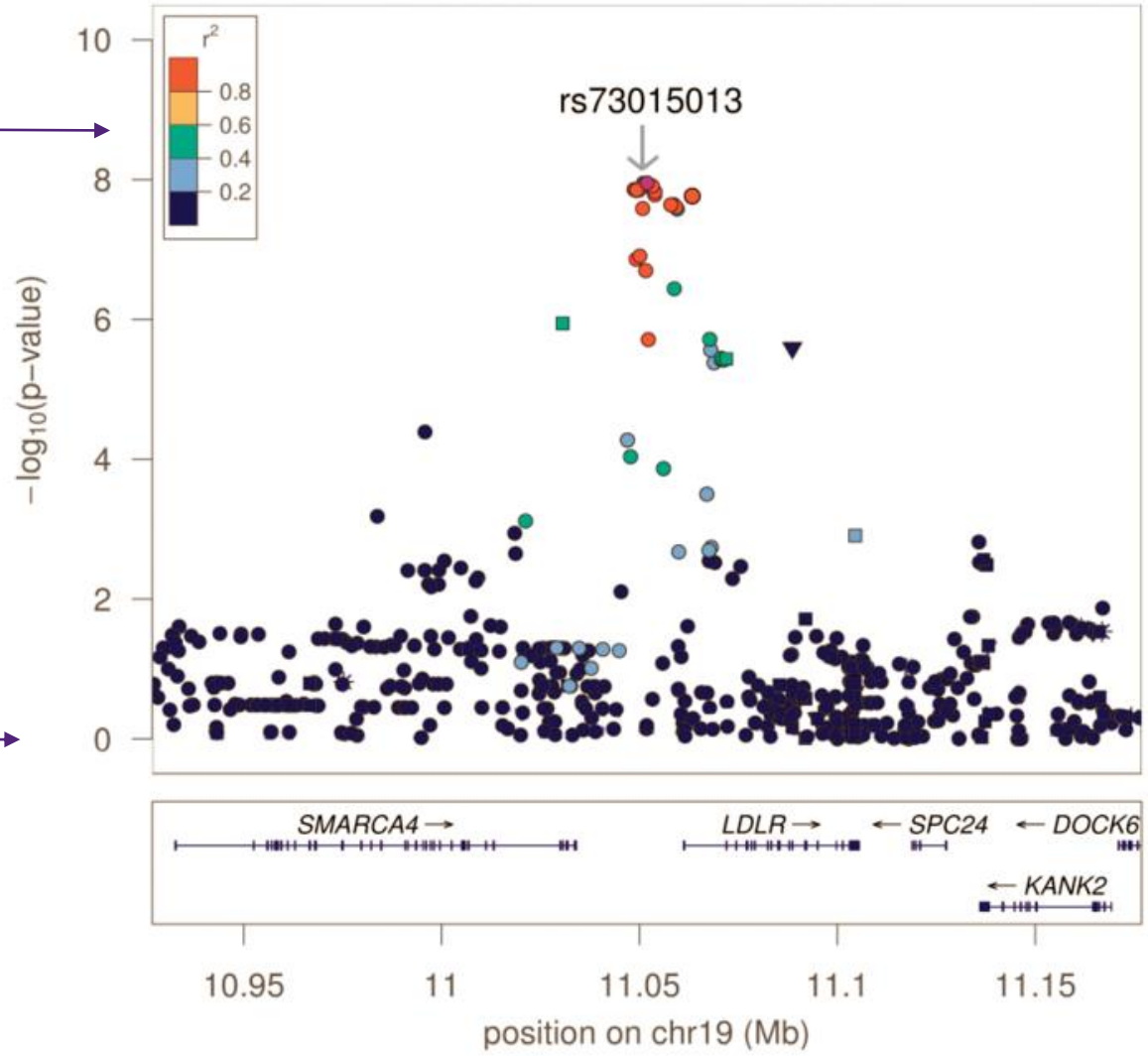
Shou et al. (2012) *PLOS ONE*

Visualising LD – regional association plots

Many tools, e.g.

LD with ‘key’ variant

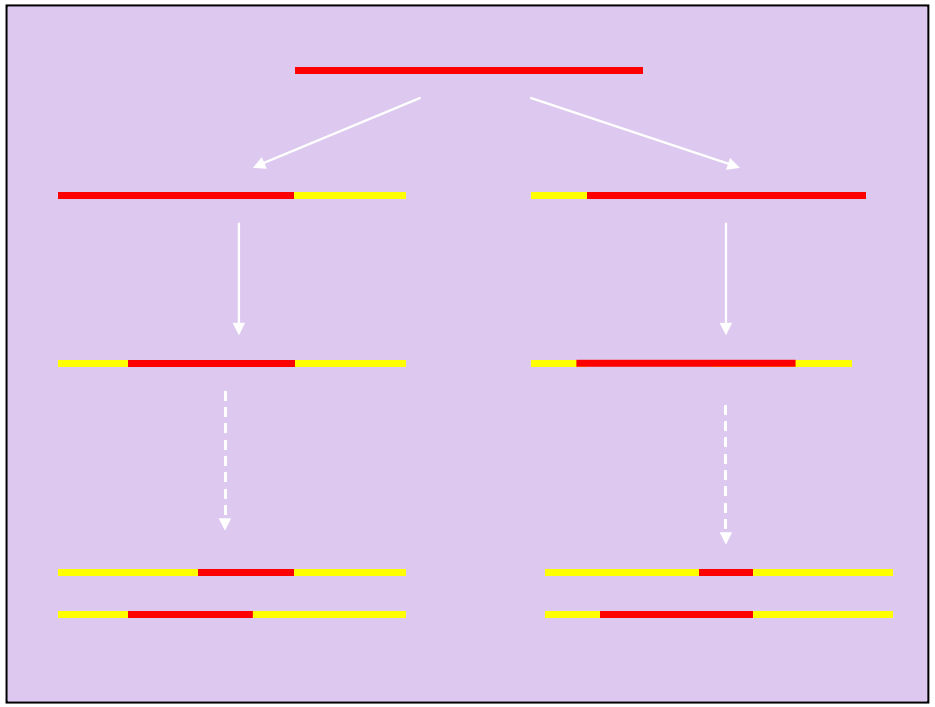
physical location



Factors influencing LD

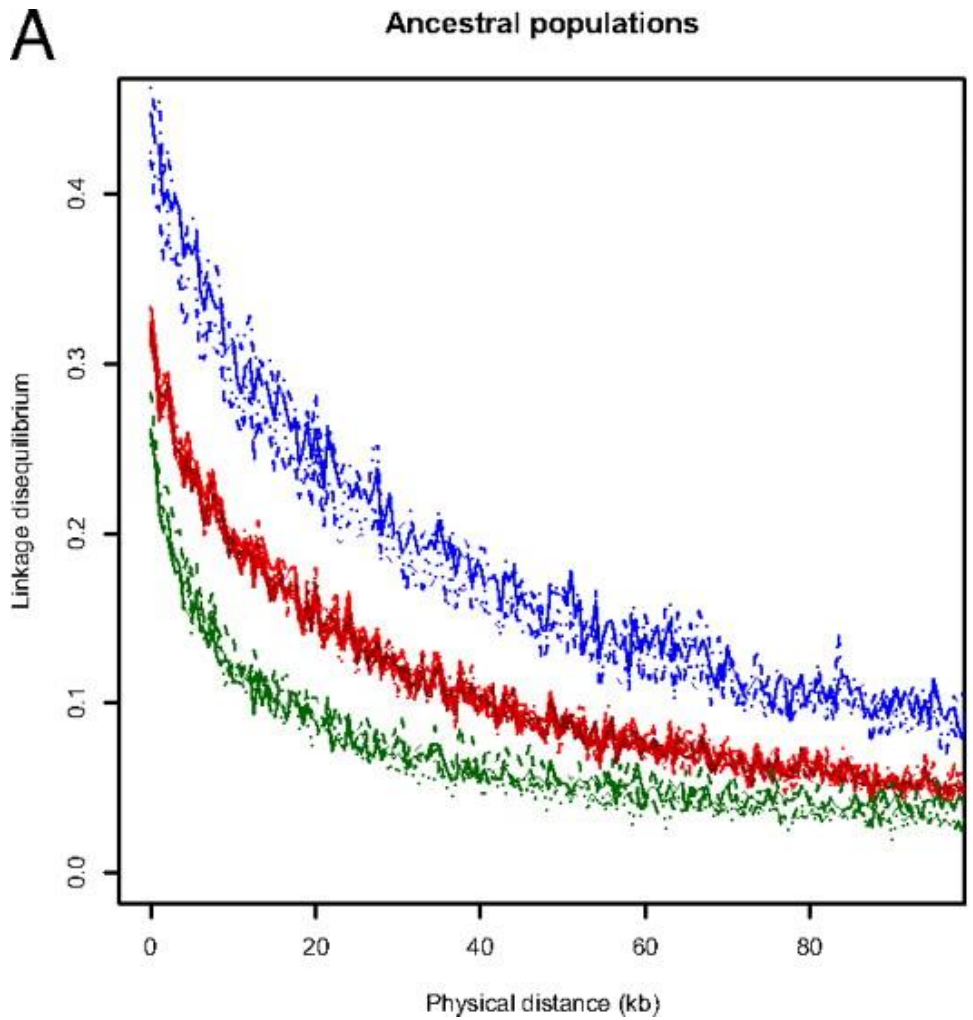
Recombination breaks up LD. Thus factors influencing LD include:

- **Physical distance**
  - Markers far away from each other more likely to be broken up by recombination
- **Genome location**
  - Recombination hotspots
- **Age of the mutation**
  - Older mutations have lower surrounding LD compared to younger mutations
- **Demographic events**
  - Population size / N_e
  - Population bottlenecks
  - Admixture
- **Selection**
  - positive or negative



Factors influencing LD

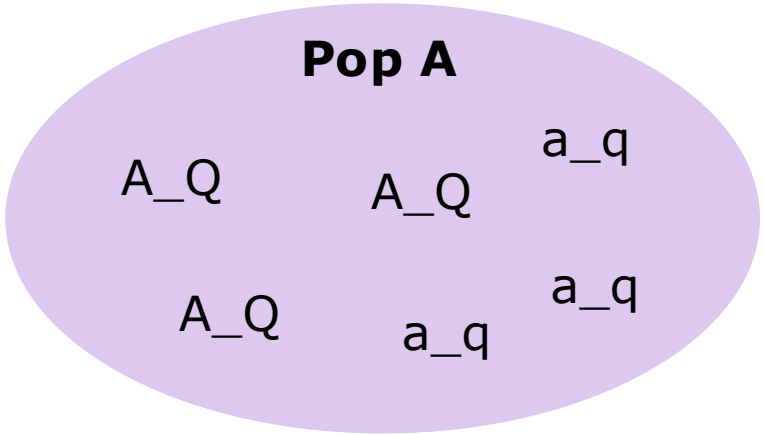
The level of LD is dependent on population history, and distance between loci



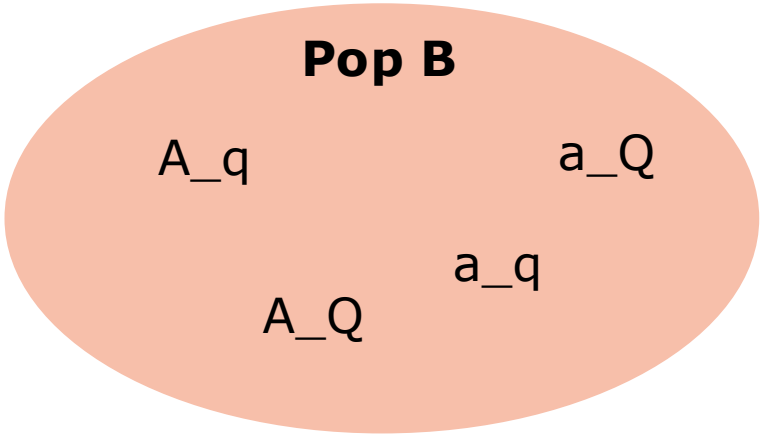
Linkage disequilibrium, genotype r2 estimated by PLINK, by population as a function of physical distance (kb). Native America, European and African populations. https://www.pnas.org/doi/full/10.1073/pnas.0914618107

Factors influencing LD

The association between a marker and a ‘causative mutation’ may be population dependent



‘A’ in LD with (unobserved) mutation ‘Q’



‘A’ is uncorrelated with mutation ‘Q’

Manipulating & exploiting LD in study design

GWAS rely on LD between SNP and unobserved causal variants

- Higher LD means greater power to detect associations

LD is a property of a population, but it can be

- **Manipulated via recruitment or experiments**. e.g. recruiting populations with different histories (e.g. Icelandic vs British), recruiting multiple populations (e.g. South Asian + European), targeting groups within a population (e.g. full sib families) or experimentally creating LD (e.g. F1 crosses)

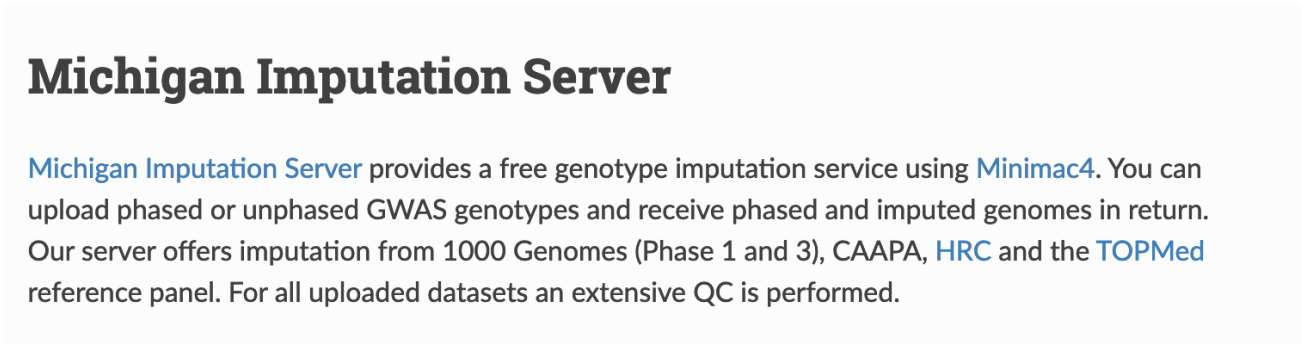
AND

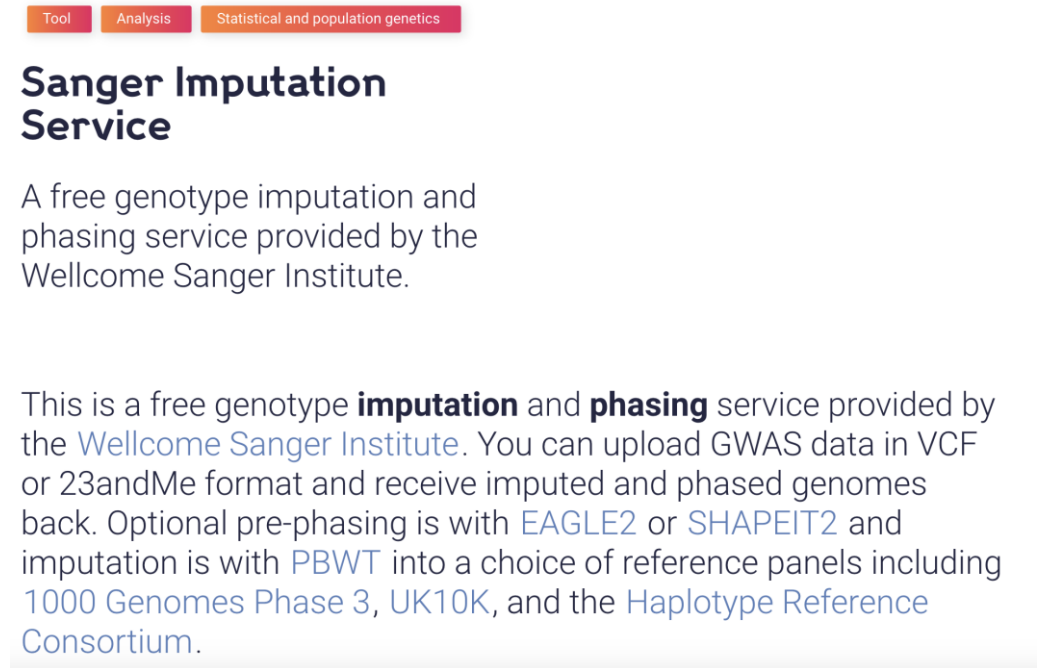
- **Leveraged** via imputation to increase SNP density.

Exploiting LD via Imputation

SNP-chip data is typically imputed to full sequence. *Why?*

- Imputation requires a relevant reference dataset & phased genotypes
- In human genetics, can be done for *free* using online imputation servers
  - e.g. Michigan or Sanger imputation servers





Imputation – brief overview

We can use LD to fill in missing genotypes via imputation

a Genotype data with missing data at untyped SNPs (grey question marks)

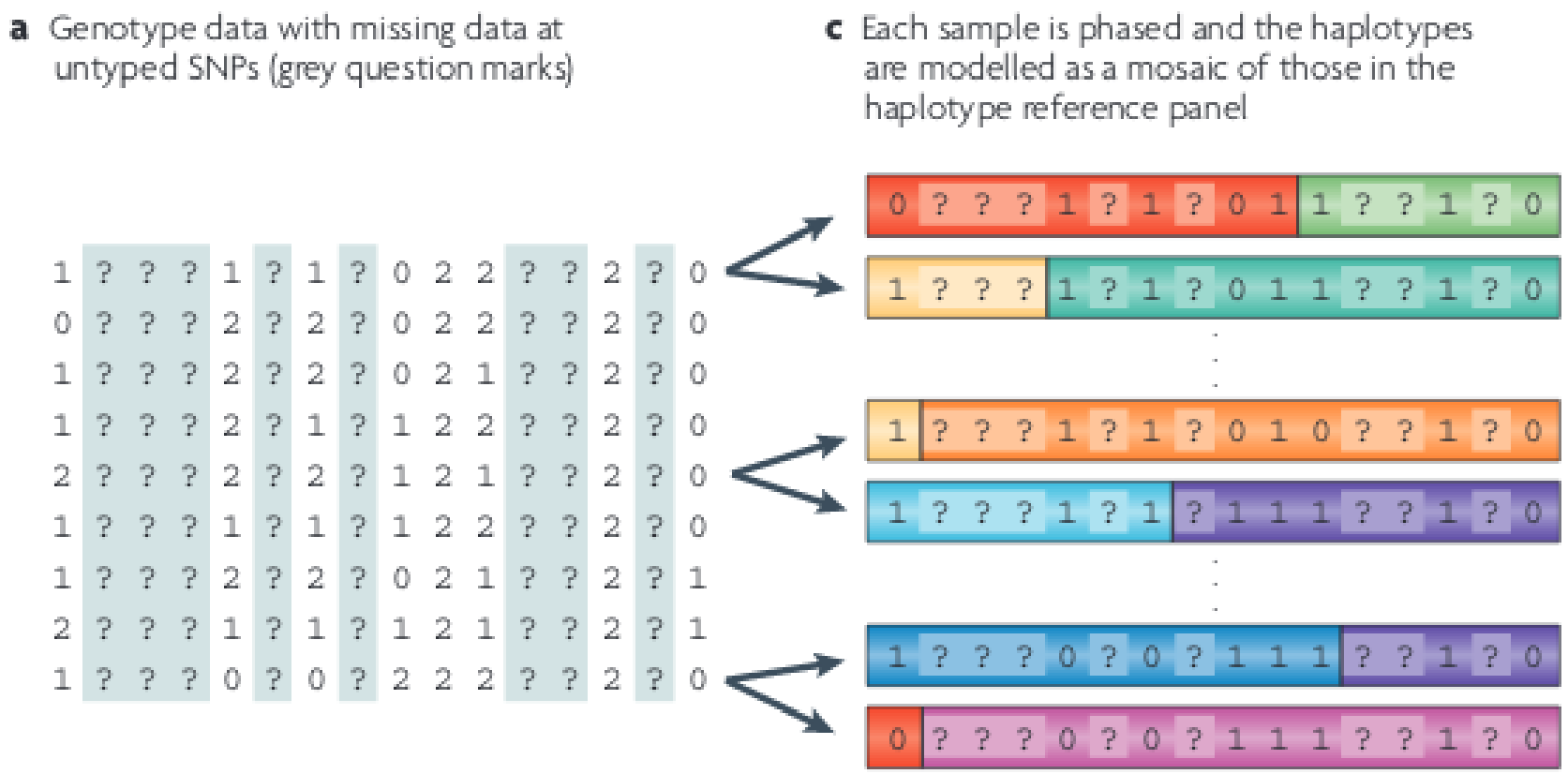
1	?	?	?	1	?	1	?	0	2	2	?	?	2	?	0
0	?	?	?	2	?	2	?	0	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	0
1	?	?	?	2	?	1	?	1	2	2	?	?	2	?	0
2	?	?	?	2	?	2	?	1	2	1	?	?	2	?	0
1	?	?	?	1	?	1	?	1	2	2	?	?	2	?	0
1	?	?	?	2	?	2	?	0	2	1	?	?	2	?	1
2	?	?	?	1	?	1	?	1	2	1	?	?	2	?	1
1	?	?	?	0	?	0	?	2	2	2	?	?	2	?	0

d Reference set of haplotypes, for example, HapMap

0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	1	1	0	0	1	1	1	0	1	1	1	0
0	0	1	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	1	1	0	1	0	0	1	0	0	0	1	0	1
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0
0	0	0	0	1	1	1	0	0	1	1	1	1	1	1	0
1	1	1	0	0	1	0	0	1	1	1	0	1	1	1	0

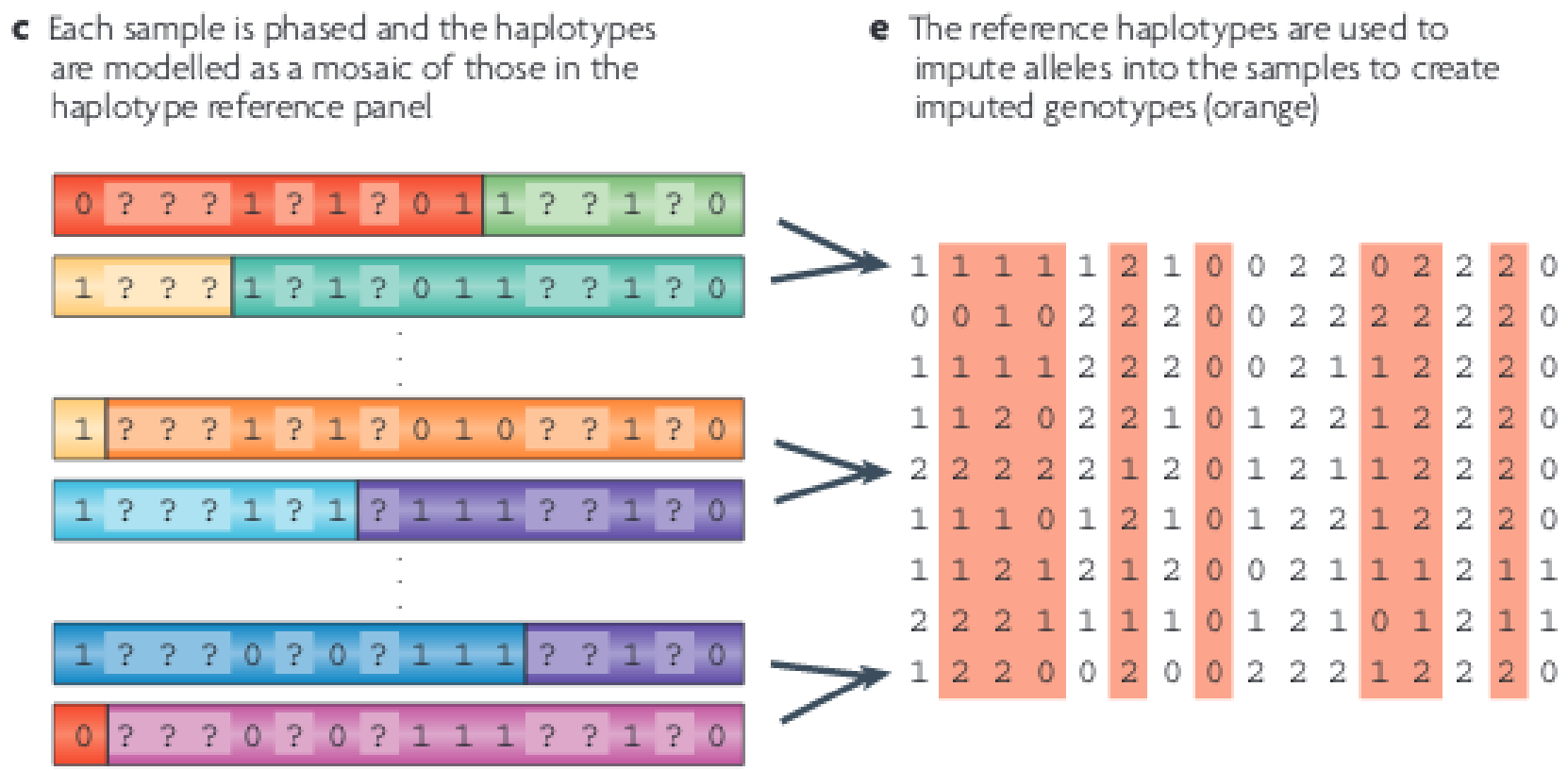
Imputation – brief overview

We can use LD to fill in missing genotypes via imputation



Imputation – brief overview

We can use LD to fill in missing genotypes via imputation



Exploiting LD via Imputation

Imputation is used to:

- fill in missing data, i.e. SNP removed during QC or poorly genotyped in some samples
- completely impute (unobserved) SNP in genotyped individuals from the reference panel

Imputed SNPs can be used in GWAS like genotyped SNPs

- increases the power to detect associations

Outcome of a statistical test

In GWAS a P-value tells you the probability of observing the data **given** that null hypothesis of no association is true, i.e. P(data|H_0). We accept or reject the null at a given level (α)

- Four possible outcomes

the truth

your decision

	H_0 true	H_A true
Reject H_0	False-positive (α)	Correct
Fail to reject H_0	Correct	False-negative

Power to detect loci

In GWAS experimental **power** (i.e. correctly rejecting H_0) is a function of:

- LD between SNP and causal variant
- Effect size of variant
- Sample size
- Significance threshold (α)

Power – Variance explained by SNP or locus

So far, we have talked about the effect size β

i.e. a regression co-efficient in trait units

To determine power, we need to know how much <u>variance</u> a marker explains

$$Var(\beta x) = \hat{\beta}^2 Var(x)$$

$$= \hat{\beta}^2 2p(1-p)$$

where p is the allele frequency

Often expressed as the proportion or % of 'phenotypic variance' (σ_P^2)

Power – (1) LD between SNP and causal variant

We don't typically expect the most significant GWAS variant in a region to be causal/functional

- i.e. tested SNP in LD with an unobserved 'causal variant'
- this reduces statistical power

Sample size must increase by $1/r^2$ to detect an unobserved variant, compared to sample size required for testing causal variant itself

- Hence increased SNP density (i.e. imputation, WGS) to maximise LD between causal variants & genotyped SNP

Power – (1) LD between SNP and causal variant

Example:

The variance explained by a ‘causal variant’ is 1% of σ_P^2

How much phenotypic variance does a genotyped SNP explain when the LD between the causal variant and SNP is 0.2 or 0.8 ?

- r^2 = 0.8 ; variance explained by SNP = 0.8 x 0.01 = 0.008 σ_P^2
- r^2 = 0.2 ; variance explained by SNP = 0.2 x 0.01 = 0.002 σ_P^2

The r^2 between a SNP and a ‘causal variant’ is the proportion of the phenotypic variance which can be observed at the SNP

Power – (2) effect size

How much of σ_P^2 is a marker expected to explain?

For human height, the first detected (i.e. largest) effect explained 0.3% σ_P^2

LETTERS

nature genetics

A common variant of *HMGA2* is associated with adult and childhood height in the general population

Michael N Weedon[1,2,21], Guillaume Lettre[3,4,21], Rachel M Freathy[1,2,21], Cecilia M Lindgren[5,6,21], Benjamin F Voight[3,7], John R B Perry[1,2], Katherine S Elliott[5], Rachel Hackett[3], Candace Guiducci[3], Beverley Shields[2], Eleftheria Zeggini[5], Hana Lango[1,2], Valeriya Lyssenko[8,9], Nicholas J Timpson[5,10], Noel P Burtt[3], Nigel W Rayner[6], Richa Saxena[3,7,11], Kristin Ardlie[3], Jonathan H Tobias[12], Andrew R Ness[13], Susan M Ring[14], Colin N A Palmer[15], Andrew D Morris[16], Leena Peltonen[3,17,18], Veikko Salomaa[19], The Diabetes Genetics Initiative, The Wellcome Trust Case Control Consortium, George Davey Smith[10], Leif C Groop[8,9], Andrew T Hattersley[1,2], Mark I McCarthy[5,6,21], Joel N Hirschhorn[3,4,20,21] & Timothy M Frayling[1,2,21]

Human height is a classic, highly heritable quantitative trait. To begin to identify genetic variants influencing height, we examined genome-wide association data from 4,921 individuals. Common variants in the *HMGA2* oncogene, exemplified by rs1042725, were associated with height ($P = 4 \times 10^{-8}$). *HMGA2* is also a strong biological candidate for height, as rare, severe mutations in this gene alter body size in mice and humans, so we tested rs1042725 in additional samples. We confirmed the association in 19,064 adults from four further studies ($P = 3 \times 10^{-11}$, overall $P = 4 \times 10^{-16}$, including the genome-wide association data). We also observed the association in children ($P = 1 \times 10^{-6}$, $N = 6,827$) and a tall/short case-control study ($P = 4 \times 10^{-6}$, $N = 3,207$). We estimate that rs1042725 explains ~0.3% of population variation in height (~0.4 cm increased adult height per C allele). There are few examples of common genetic variants reproducibly associated with human quantitative traits; these results represent, to our knowledge, the first consistently replicated association with adult and childhood height.

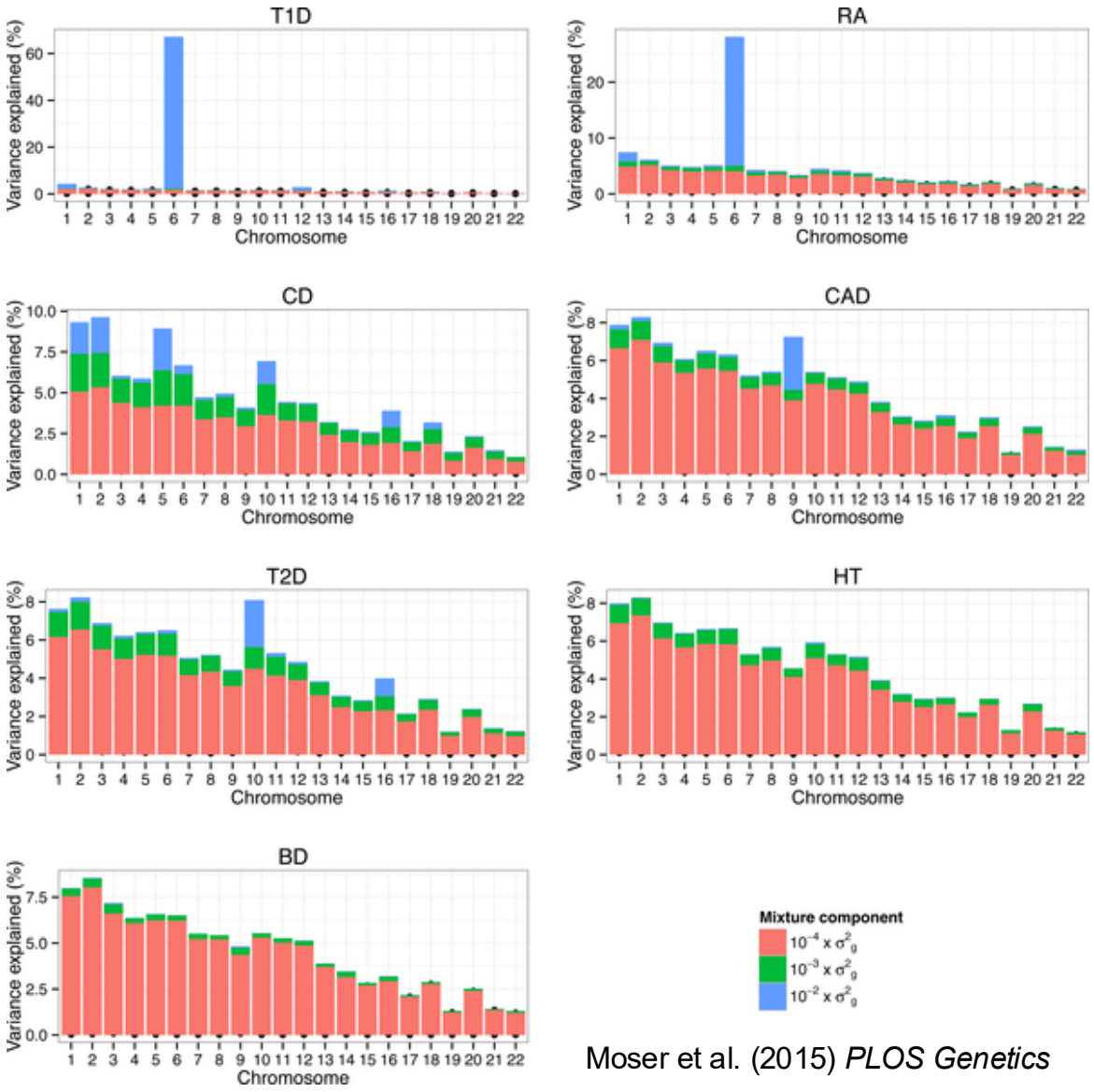
Adult height is a classic polygenic trait. The genetics of height were central to the mendelian versus biometrician debate in the early part of the twentieth century that was resolved by Fisher, who proposed that height and other human phenotypes showed multifactorial inheritance[1]. Twin, family and adoption studies suggest that up to 90% of normal variation in human height within populations is due to genetic variation[2–6]. Severe mutations in several genes cause rare syndromes with extreme stature; however, these cannot explain normal population height variation[7]. Many regions of the genome have been linked with height based on numerous genome-wide linkage scans, with some overlap between studies[6], but thus far there have not been any examples of gene variants that are reproducibly associated with height variation in the general population.

The recent flood of data from many genome-wide association (GWA) studies offers new opportunities to identify genes influencing adult height. The identification of such genes will probably provide important insights into how best to dissect the genetics of polygenic quantitative traits. The identification of genes influencing growth may also have important medical implications. Height is associated with several common disorders, including a number of cancers[8,9].

Power – (2) effect size

How much of σ_P^2 is a marker expected to explain?

It is trait dependent



Moser et al. (2015) *PLOS Genetics*

Power – (2) effect size

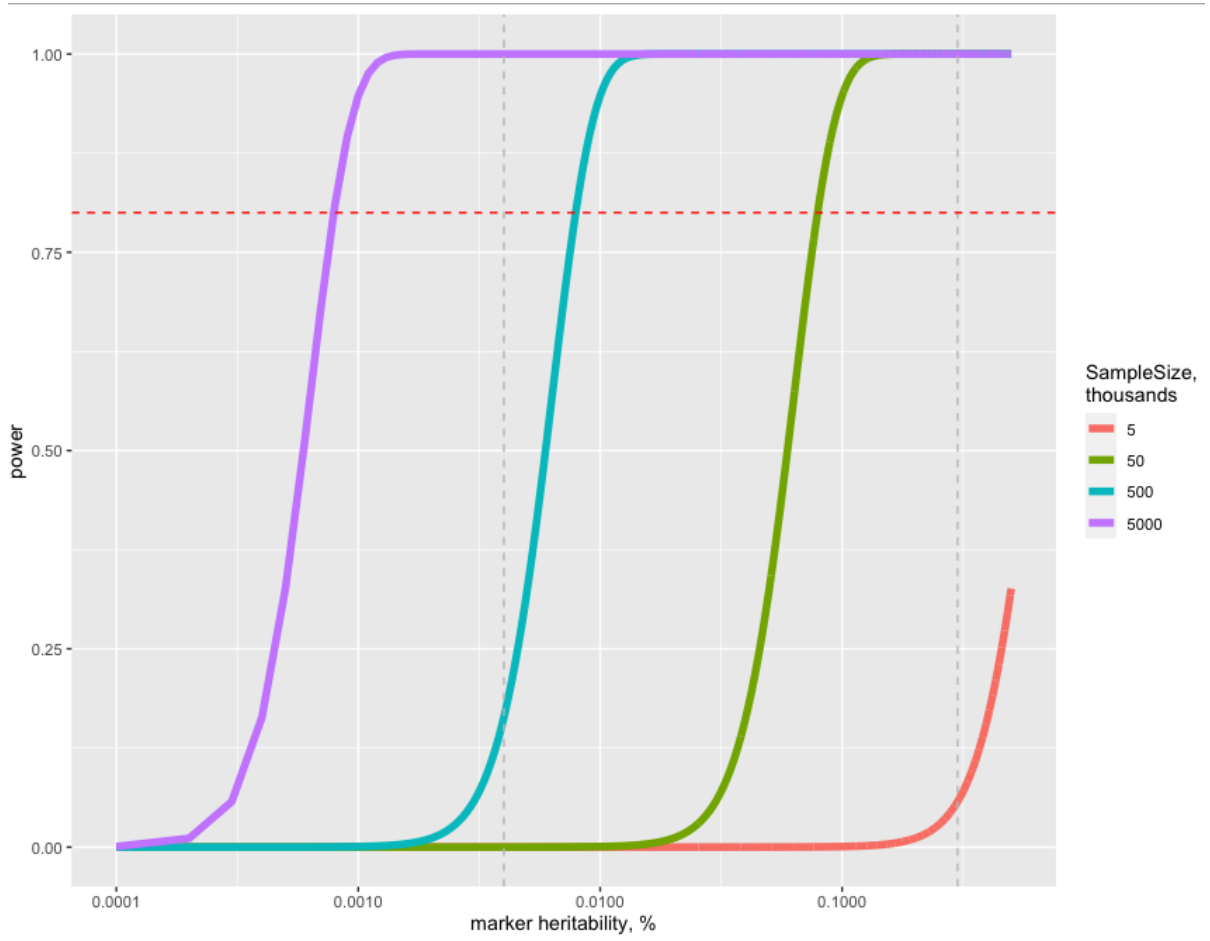
How much of σ_P^2 is a marker expected to explain?

It is trait dependent

For human height, the first detected (i.e. largest) effect explained 0.3% σ_P^2

Yengo et al. (2022) detected 12,111 SNP collectively explaining ~ 0.5 σ_P^2

i.e. 0.004 % σ_P^2 per SNP



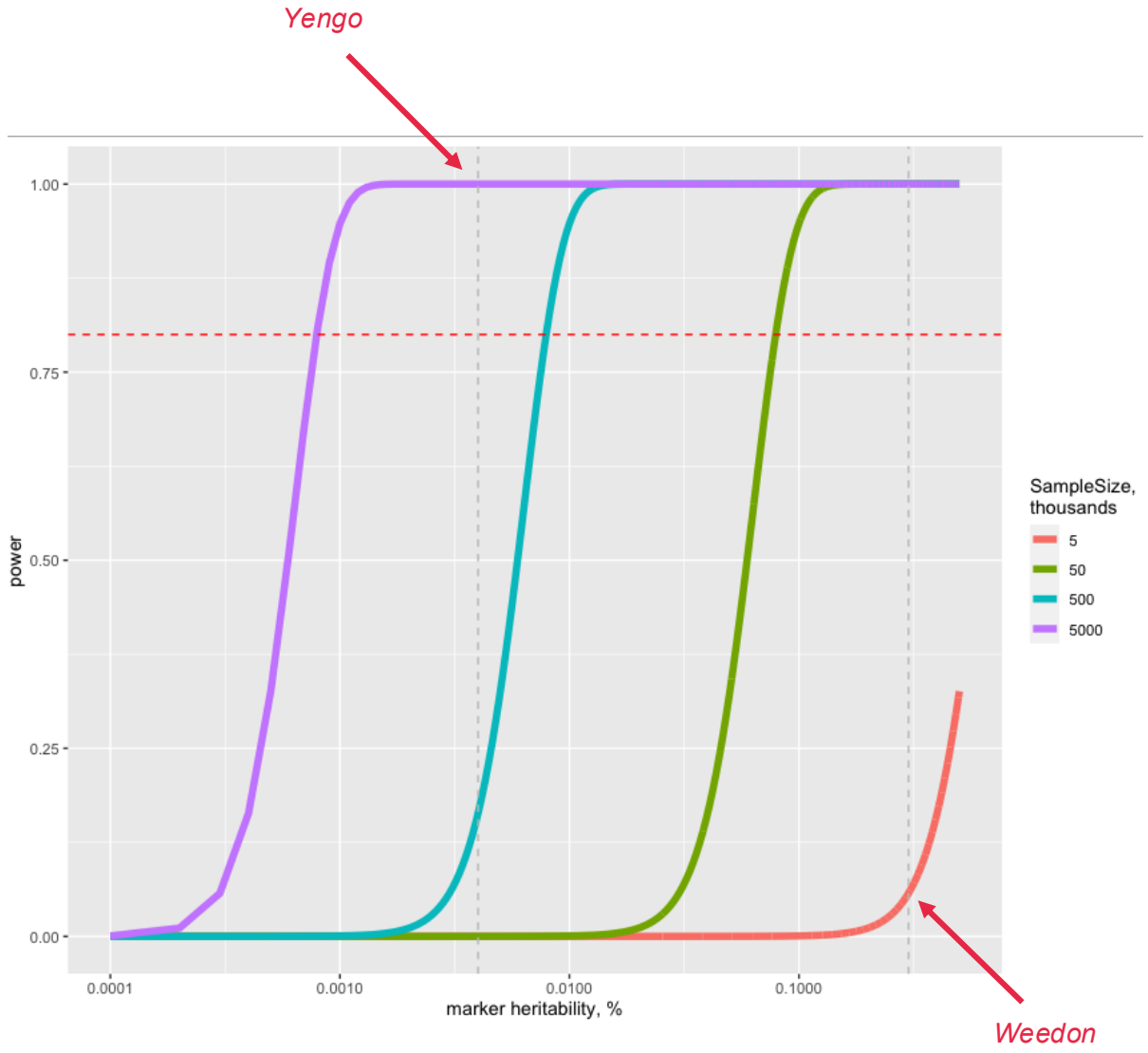
Power – (3) sample size

How big do sample sizes need to be?

For human height,

5K individuals to detect loci 0.3% σ_P^2

5M to detect loci explaining ~ 0.004 % σ_P^2



Power – (4) significance threshold

- GWAS performs millions of tests... many will be 'significant' ($P < 0.05$) by chance
- Easiest way to account for all these tests is to correct the significance threshold (α) for number of independent tests
  - correcting for the total number of tests is overly conservative due to the LD

- LD varies between populations, thus
  - EUR: 1 million independent tests ($0.05/1x10^6$) → sig. threshold p = $5x10^{-8}$
  - AFR: 2 million independent tests ($0.05/2x10^6$) → sig. threshold p = $2.5x10^{-8}$

Pe'er et al. (2008) *Genetic Epidemiology*

Power (4) – significance threshold

Permutation testing is ‘gold standard’

In non-human GWAS, an experiment-wise FDR (**F**alse **D**iscovery **R**ate) of 5% may be preferable to blind acceptance of $5x10^{-8}$

FDR = # expected ‘significant’ SNP / # observed ‘significant’ SNP

Two ways, (1) calculate FDR at a given nominal p-value or
(2) determine which p-value will give FDR of 5%

e.g. If we test 1M loci with α = 0.0001, we expect $1x10^6$ X 0.0001 = 100 sig. loci by chance

Say we observe 150 sig. loci at α = 0.0001

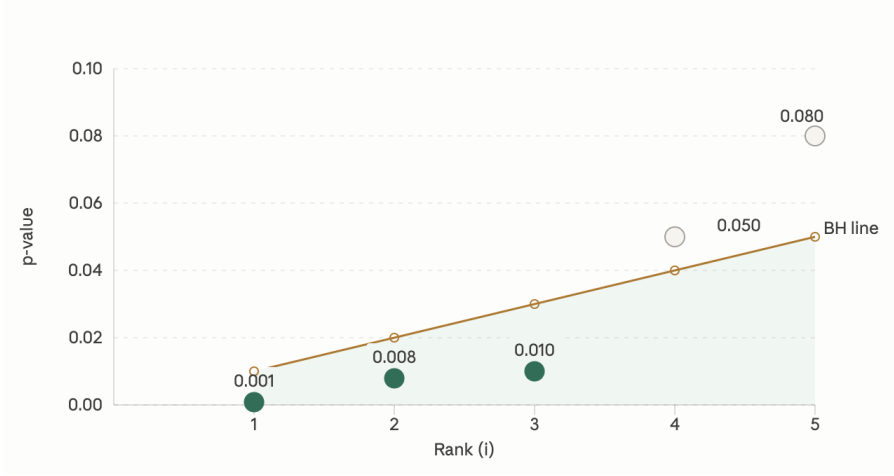
FDR = expected/observed = 100/150 = 0.67

Power (4) – significance threshold

Benjamini-Hochberg Procedure

1. Sort p-values from smallest to largest
2. Assign ranks, 1 to m (m = total number of tests)
3. Choose FDR rate (α = 0.05)
4. Calculate critical value for each marker, $q = \alpha(\frac{rank}{m})$
5. Starting from smallest p-value (rank = 1), accept if q < p-value

Rank	P-value	q	p < q?	decision
1	0.001	(0.05x1)/5 = 0.01	Y	reject H_0
2	0.008	(0.05x2)/5 = 0.02	Y	reject H_0
3	0.01	(0.05x3)/5 = 0.03	Y	reject H_0
4	0.05	(0.05x4)/5 = 0.04	N	accept H_0
5	0.08	(0.05x5)/5 = 0.05	N	accept H_0



A final word on replication

GWAS may have many false-positives

- Replication in an independent cohort typically required before follow-up
- Be mindful of sample size (is there enough power to replicate?)
- Replicate size and direction of effect
- *'Winner's curse'*

Meta-analysis of CHARGE and Global BPgen of Top 10 Loci for Systolic and Diastolic Blood Pressure and Hypertension in CHARGE

SNP identifier	Chr	Position	Nearest Gene	Alleles (coded / other)	Freq. of coded allele	discovery			replication		
						Beta	SE	p-value	Beta	SE	p-value
Systolic blood pressure											
rs12046278	1	10,722,164	*CASZ1*	T/C	0.64	-0.84	0.18	1.84E-06	-0.29	0.15	5.71E-02
rs7571613	2	190,513,907	*PMS1*	A/G	0.82	-0.96	0.19	7.28E-07	-0.23	0.16	1.59E-01
rs448378	3	170,583,593	*MDS1*	A/G	0.52	-0.71	0.15	1.28E-06	-0.36	0.13	4.76E-03
rs2736376	8	11,155,175	*MTMR9*	C/G	0.13	-1.08	0.23	1.90E-06	-0.06	0.19	7.36E-01
rs1910252	8	49,569,915	*EFCAB1*	T/C	0.18	-0.93	0.19	1.70E-06	-0.07	0.17	6.80E-01
rs11014166	10	18,748,804	*CACNB2*	A/T	0.66	0.74	0.16	2.11E-06	0.33	0.13	1.31E-02
rs1004467	**10**	**104,584,497**	***CYP17A1***	**A/G**	**0.90**	**1.20**	**0.25**	**1.99E-06**	**0.94**	**0.21**	**1.08E-05**
rs381815	**11**	**16,858,844**	***PLEKHA7***	**T/C**	**0.26**	**0.84**	**0.17**	**5.76E-07**	**0.52**	**0.14**	**2.72E-04**
rs2681492	**12**	**88,537,220**	***ATP2B1***	**T/C**	**0.80**	**1.26**	**0.19**	**3.01E-11**	**0.50**	**0.17**	**4.07E-03**
rs3184504	**12**	**110,368,991**	***SH2B3***	**T/C**	**0.48**	**0.75**	**0.15**	**5.73E-07**	**0.45**	**0.13**	**6.36E-04**

Levy et al. (2009) *Nature Genetics*

Power to detect loci

Power is a function of:

- LD between SNP and causal variant (dense SNPs to maximise LD)
- Proportion of phenotypic variance explained by SNP
  - Typically: < 0.005 σ_P^2 for quantitative traits, OR 1.1-1.2 binary traits
  - Can't change genetic architecture
- Sample size (bigger is more powerful)
- Significance threshold (α) – 5% FDR experiment wise

Outcome of a statistical test

In GWAS a P-value tells you the probability of observing the data **given** that null hypothesis of no association is true, i.e. P(data|H_0). We accept or reject the null at a given level (α)

- Four possible outcomes

the truth

your decision

	H_0 true	H_A true
Reject H_0	False-positive (α)	Correct
Fail to reject H_0	Correct	False-negative

Confounders in GWAS

Confounders are unmeasured or uncontrolled factors that can result in spurious false-positive associations

A variable correlated with genotype and phenotype but not through the SNP that you are testing

3 overlapping sources

(1) Technical confounding : batch effects, array differences, ascertainment

(2) Population structure, genetic confounding i.e. stratification & relatives

(3) Environmental confounding : socioeconomic and environmental factors

Technical confounding

Many different sources of technical confounders, i.e. something with non-biological origin that causes a systematic difference between groups

QC QC QC : Many different approaches at different levels,

e.g. PC coloured by batch, missingness, heterozygosity, GWAS across platforms within controls

-> More on this tomorrow

Technical confounding - HWE

Hardy Weinberg equilibrium is the probabilistic relationship between allele and genotype frequencies at a single locus,

Consider an A/a bi-allelic locus:

i.e. alleles are A and a

Frequency of **A** is p

Frequency of **a** is q (thus $p = 1 - q$)

Three possible genotypes:

AA has expected frequency p^2

Aa has expected frequency $2pq$

aa has expected frequency q^2

MODIFIED PUNNET SQUARE

Allele (freq)	**A** *(p)*	**a** *(q)*
A *(p)*	**AA** *(p^2)*	**Aa** *(pq)*
a *(q)*	**aA** *(qp)*	**aa** *(q^2)*

Technical confounding - HWE

Hardy Weinberg equilibrium is the probabilistic relationship between allele and genotype frequencies at a single locus,

We can test for HWE using a chi-squared test with 1df

There are legitimate reasons for a SNP to fail this test, e.g. selection or demographic events, population structure, non-random mating, etc. but they are assumed to be rare

In GWAS, SNP failing HWE are more likely to be genotyping errors, e.g. poor cluster separation, probe failure or DNA quality issues

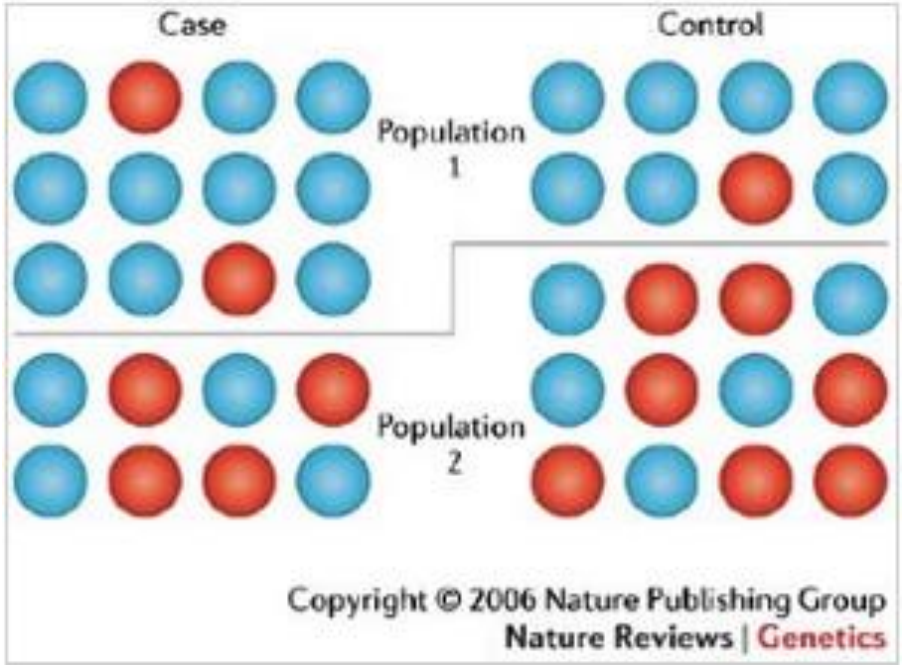
Population structure - stratification

Population stratification is a major source of bias

- it creates spurious associations

Occurs when there are unknown subpopulations within the study sample which have *systematic differences in both ancestry (allele frequencies) and phenotypes*

- e.g. when one subpopulation contributes more cases to a case-control GWAS



	Case	Control
ALL	14/20 = 0.7	12/20 = 0.6

Balding (2006) *Nature Rev. Genetics*

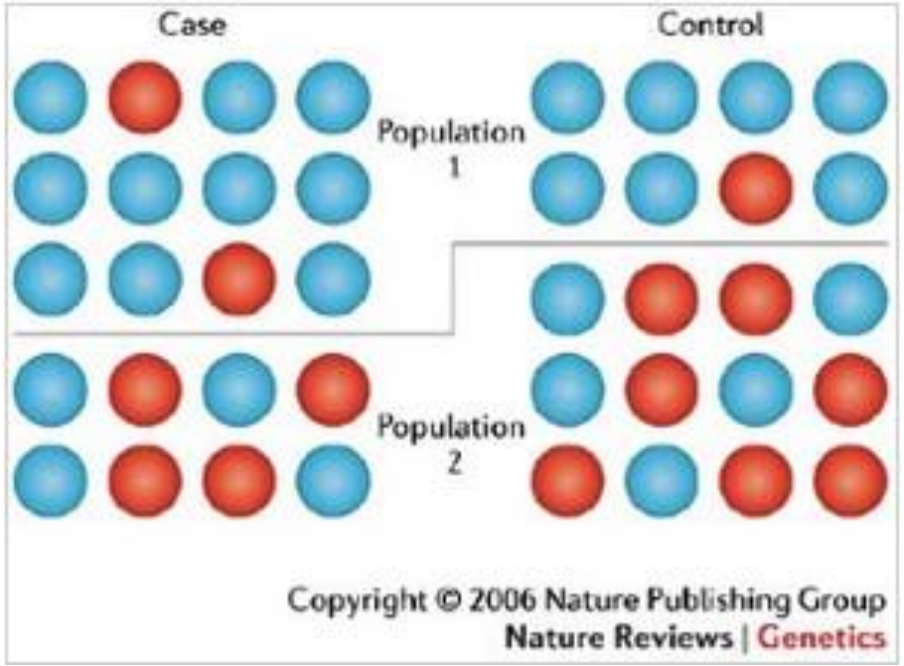
Population structure - stratification

Population stratification is a major source of bias

- it creates spurious associations

Occurs when there are unknown subpopulations within the study sample which have *systematic differences in both ancestry (allele frequencies) and phenotypes*

- e.g. when one subpopulation contributes more cases to a case-control GWAS



	Case	Control
Pop 1	10/12 = 0.83	7/8 = 0.87
Pop 2	4/8 = 0.5	5/12 = 0.41
ALL	14/20 = 0.7	12/20 = 0.6

Balding (2006) *Nature Rev. Genetics*

Population structure - stratification

also occurs for quantitative/continuous traits

e.g. Campbell et al. performed a GWAS on two groups of individuals of European descent that were discordant for height and identified an association with the LCT (lactase) locus

	Height (Adult men)	Lactose Tolerance
Northern (Sweden)	5 ft 11 1/2 in	98%
Southern (Italy)	5 ft 9 1/2 in	$\sim$ 50%

Campbell et al. (2005) *Nature Genetics*

Population structure - close relatives

Sometimes we might want to include relatives to increase sample size.
However, close relatives have correlated genomes! They are not independent.

➔Inflated tests statistics

They might also share environmental effects which might bias effects

Population structure - implications

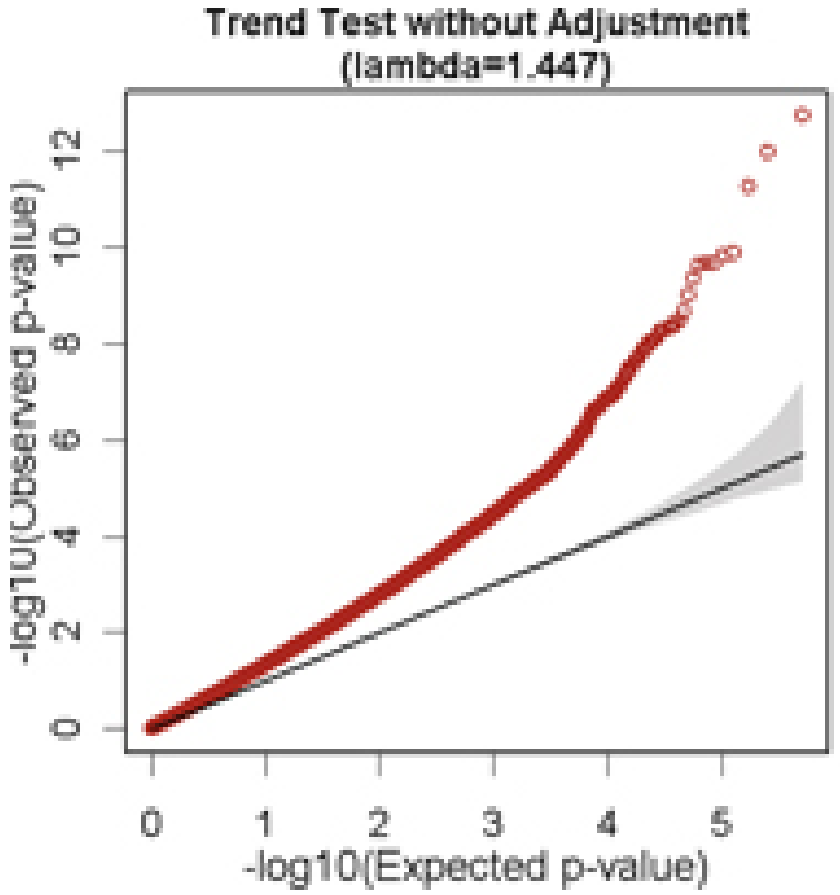
Both types of population structure can inflate test statistics.

Population structure is sometimes difficult to avoid. Common approaches to deal with structure include:

(1) removing individuals

(2) fit PCs as covariates (ancestry)

(3) mixed model approaches (relatedness)



The genomic relationship matrix

Many approach to detect & account for population structure rely on a genomic relationship matrix or GRM

What is a GRM?

$$\begin{pmatrix} 1.1 & 0.22 & 0.12 & -0.01 \\ 0.22 & 0.95 & 0.12 & 0.01 \\ 0.12 & 0.12 & 1.05 & 0.52 \\ -0.01 & 0.01 & 0.52 & 1.00 \end{pmatrix}$$

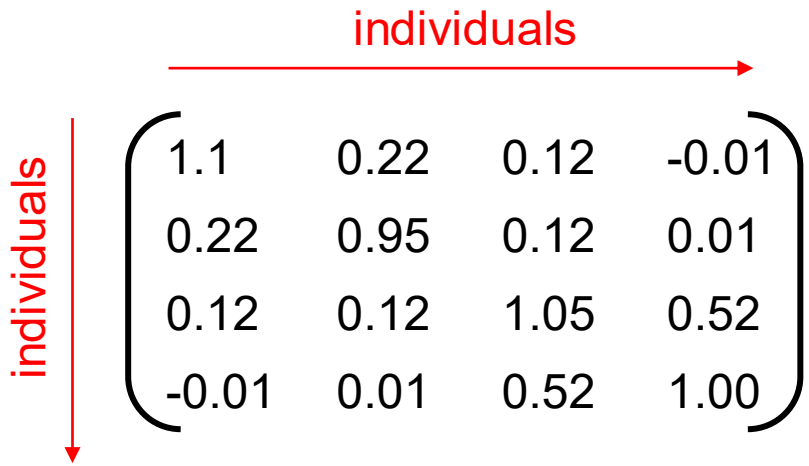
off-diagonal elements of **A** estimate the genomic relationship (π) between pairs [i.e. average allele sharing]

The genomic relationship matrix

Many approach to detect & account for population structure rely on a genomic relationship matrix or GRM

What is a GRM?

Square symmetric matrix



off-diagonal elements of **A** estimate the genomic relationship (π) between pairs [i.e. average allele sharing]

The genomic relationship matrix

Many approach to detect & account for population structure rely on a genomic relationship matrix or GRM

What is a GRM?

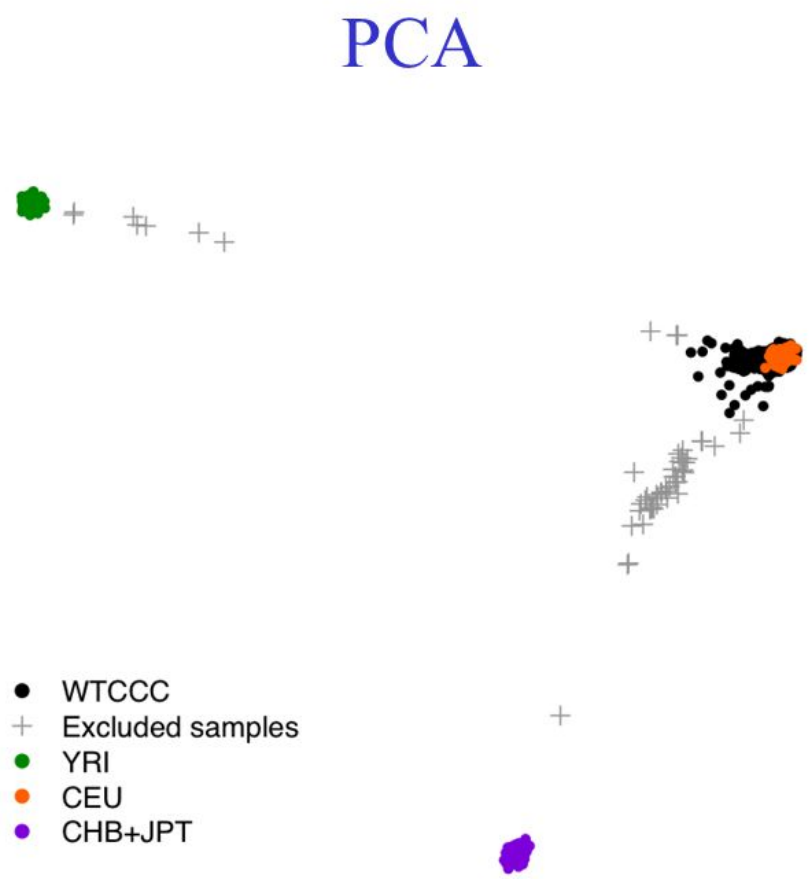
Square symmetric matrix

off-diagonal elements of **A** estimate the genomic relationship (π) between pairs [i.e. average allele sharing]

Population structure – ancestry outliers

remove ancestry outliers

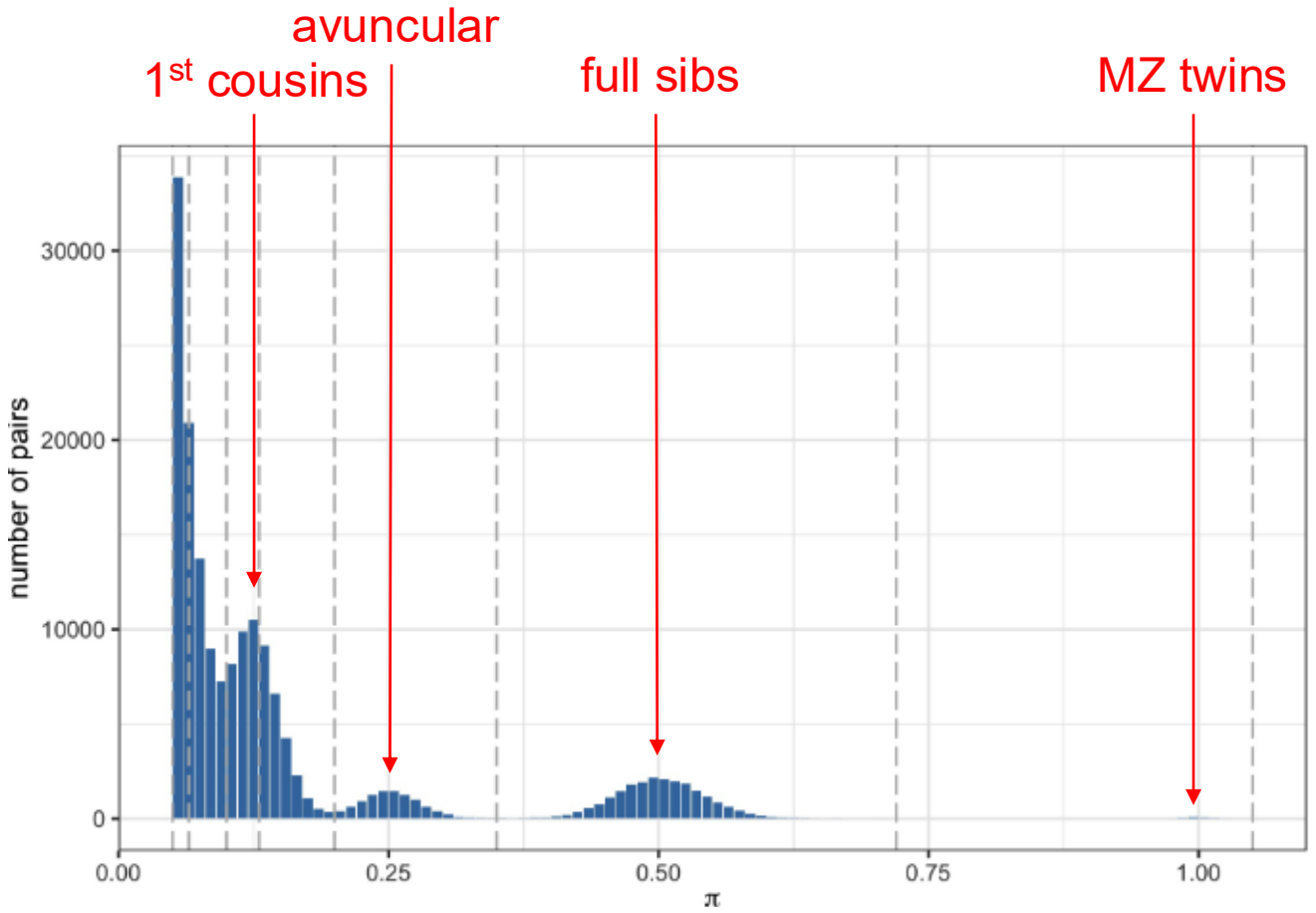
1. Perform PCA on GRM of diverse individuals with known ancestry, e.g. 1000 Genomes
2. Project your samples onto PCs
3. Exclude ‘outliers’



Population structure – close relatives

Large GRM off-diagonal elements (π > 0.05)

remove one member from each pair



Genomic relationship among each pair in UK Biobank (π)

Kemper et al. (2021) *Nature Communications*

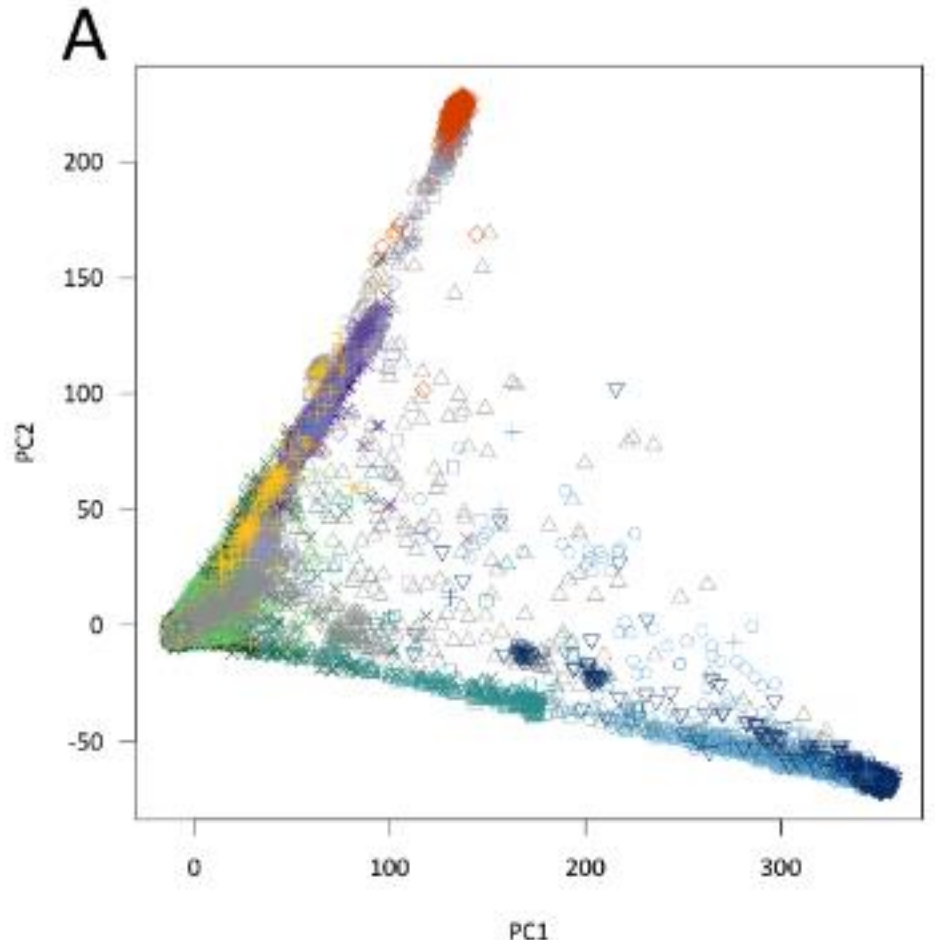
Population structure – within a population

fitting PCs as covariates

1. Perform PCA on your GRM
2. Fit PCs as covariates in GWAS

e.g. file of covariates:

ID	PC1	PC2
456859	-10	0
456185	150	-10
523014	323	-47
…	…	…



1st and 2nd principal components in UK Biobank, coloured by self-reported ancestry

Bycroft et al. (2021) *Nature Genetics*

Population structure – within a population

Fit the GRM using mixed model approach

Alternative (or sometimes in addition to!) PCs for highly structured populations

$$\mathbf{y} = \mathbf{1}\alpha + \mathbf{x}\beta + \mathbf{W}\mathbf{g} + \mathbf{e}$$

phenotype

intercept

genotype

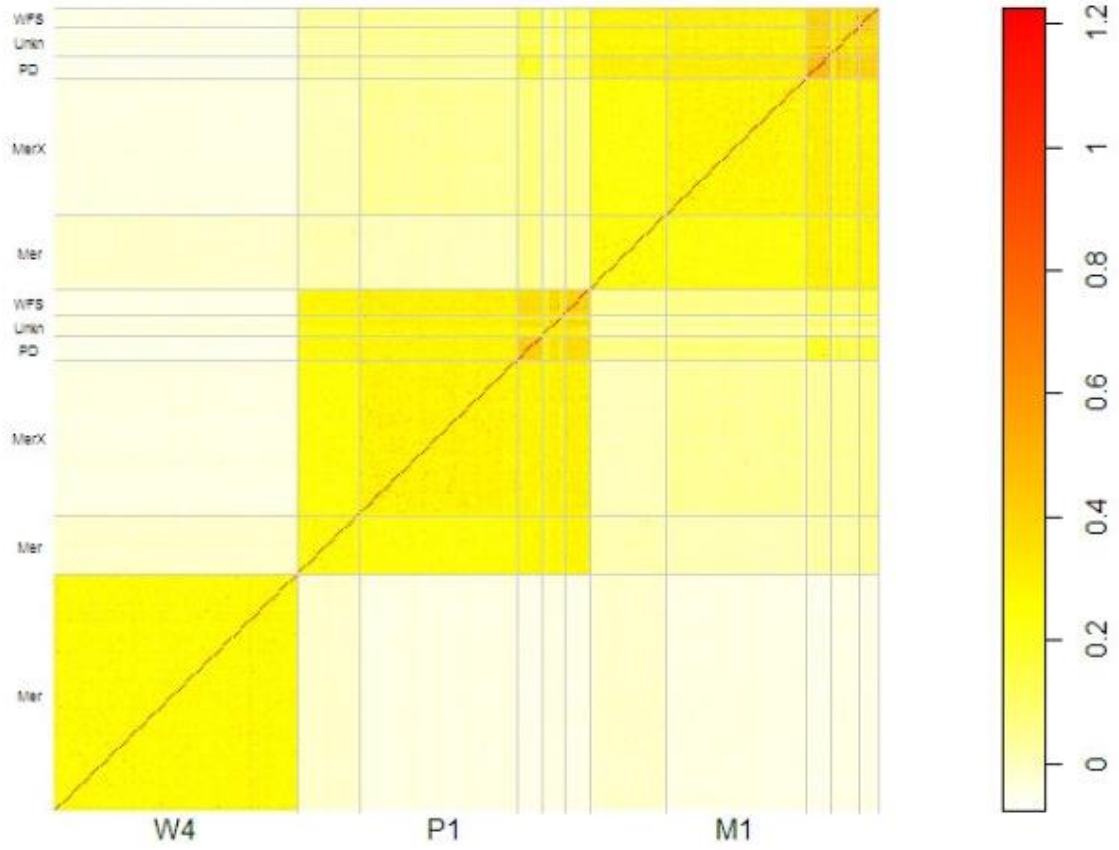
SNP effect

design matrix

(random) additive genetic effect

error

Example GRM from sheep



Kemper et al. (2011) *Genetics Research*

Summary – study design

GWAS require careful thought – what question am I asking?

Important considerations to maximise experimental power & avoid confounding

- Study population and sample size
- Size of effects
- Multiple testing & significance thresholds

Practical Session

Choose Part 1, 2, 3 or 4 (!)

Part 1: LD between loci
Part 2: power to detect loci
Part 3: construct a small GRM
Part 4: simple PCA in R

- download practical notes & slides from
  https://cnsgenomics.com/data/teaching/GNGWS26/module1/

- On the cluster please work in your own folder, `/scratch/username/`

- data can be downloaded for the cluster from:
  `/data/module1/downloadsDataMonPM.zip`