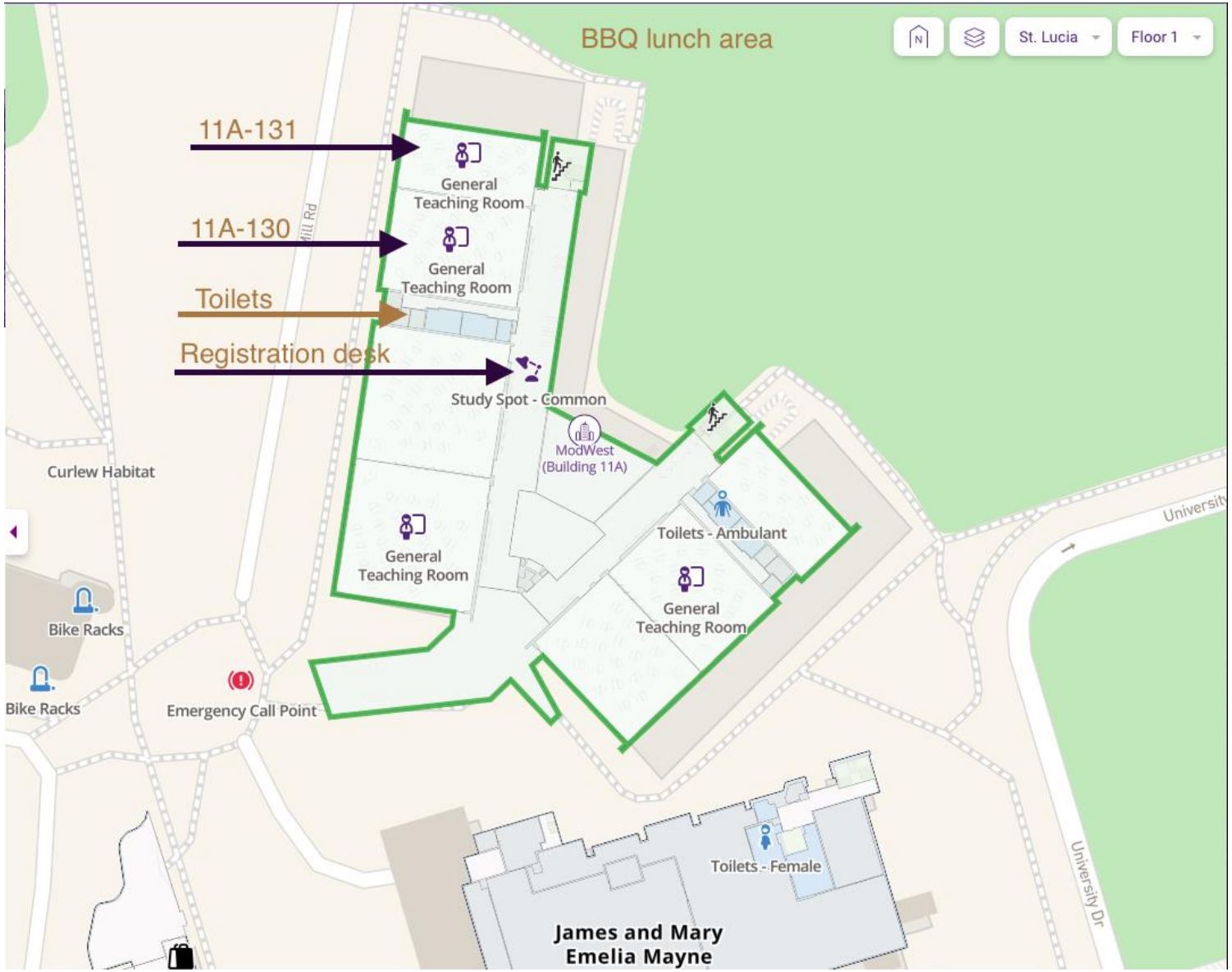
Acknowledgement of Country

- The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.
- We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.
- We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

General Information:

- We are currently located in Building 11A MODWEST
  - Bathrooms
  - Vending machines
- Food court and other bathrooms are located in Building 63 or Building 21B
- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days please ensure you are wearing a mask for the duration of the module



Learning materials

Instructions to access WiFi/desktop/server:

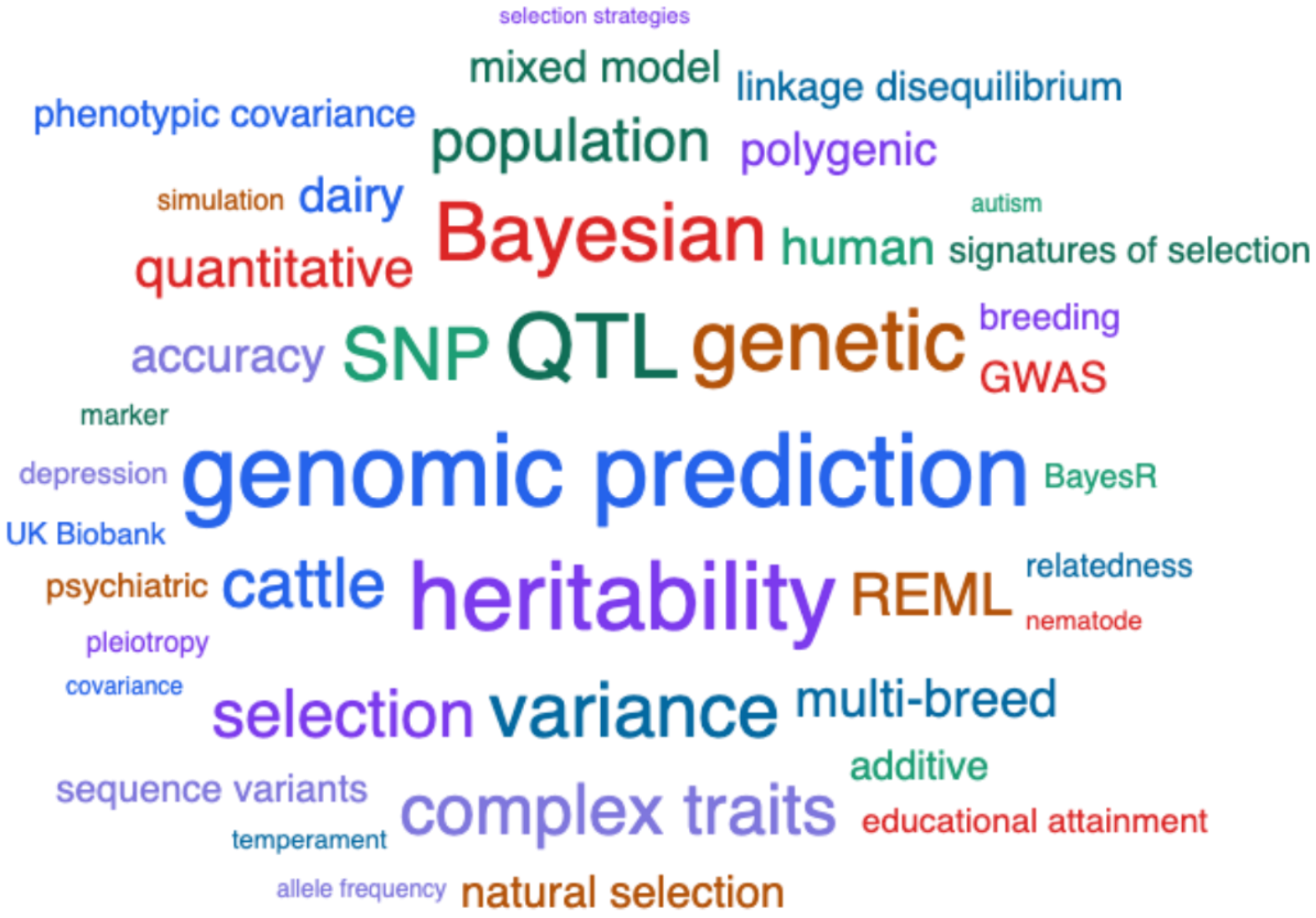
https://cnsgenomics.com/data/teaching/GNGWS26/module0/

The winter school server is available until **24th July 2026** (2 weeks after the course)

Slides and practical notes for this module:

https://cnsgenomics.com/data/teaching/GNGWS26/module1/

A little bit about me...



Plan for the Module: Genetic Mapping

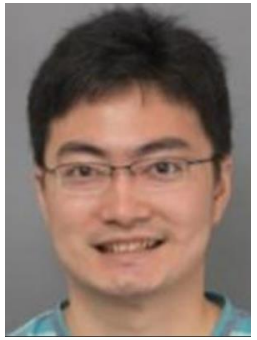
DAY 1 – GWAS, foundations	
1pm	1. Intro to GWAS
2:20 pm	BREAK – afternoon tea
2:40 pm	2. Study Design

DAY 2 – GWAS & Post-GWAS analysis	
9am	3. GWAS in practice I
10:20am	BREAK – morning tea
10:40am	4. GWAS in practice II
12pm	LUNCH
1pm	5. Trouble shooting 6. Summary Stats 7. Meta-analysis
2:20 pm	BREAK – afternoon tea
2:40 pm	8. Independent loci 9. Fine Mapping

Kath
(lecturer)

Alesha
(lecturer)

Tian
(teaching assistant)



JZ
(WS coordinator)

Solal
(cluster admin)

Acknowledgements

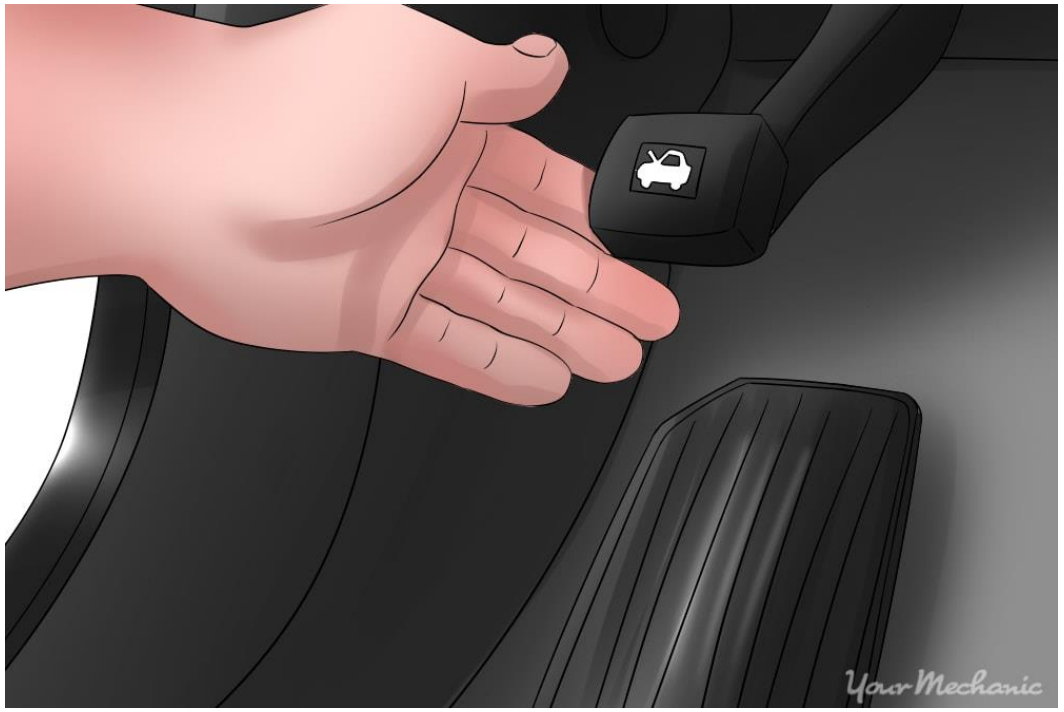
I wish to acknowledge the following people for allowing us to use/adapt their material (slides and practical) in this module:

Allan McRae
Ben Hayes
Joanna Revez
Jian Zeng
Alesha Hatton
Naomi Wray
Fleur Garton

Various internet resources & GWAS protocol papers

Objective for the Module: Genetic Mapping

- during the module you will *not* find recipes and/or pipelines
- We hope you will *<u>understand</u>* what your doing & be able to critique others
- You may not complete all the practicals but hopefully you will have resources to come back to
  - everyone brings different skills



THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

MODULE 1 | GENETIC MAPPING

Session 1. Introduction to GWAS

July 2026

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

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

Introduction to Genome Wide Association Studies (GWAS)

Outcome: Participants are familiar with the pre-genomics motivation for GWAS, terminology and basic outputs of a GWAS study

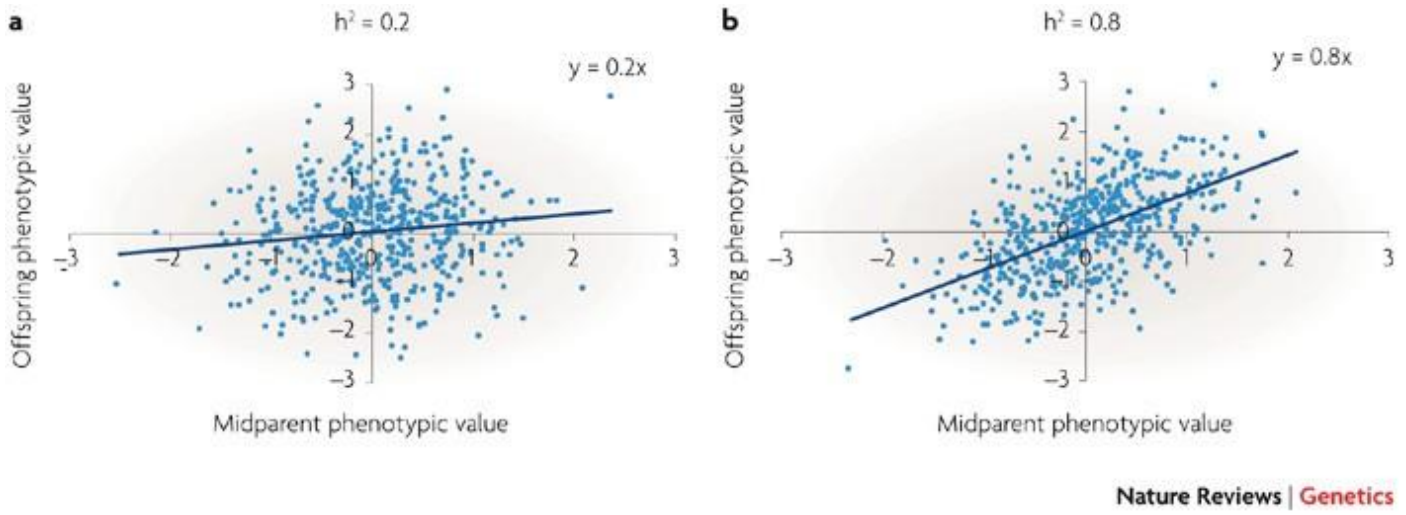
Outline: (1) context & motivation for GWAS

(2) inputs, output and basic approach

What came before GWAS?

Pre-genomics:

- good evidence for heritable genetic variation
- Limited knowledge of biological underpinnings



Visscher, Hill & Wray (2008)

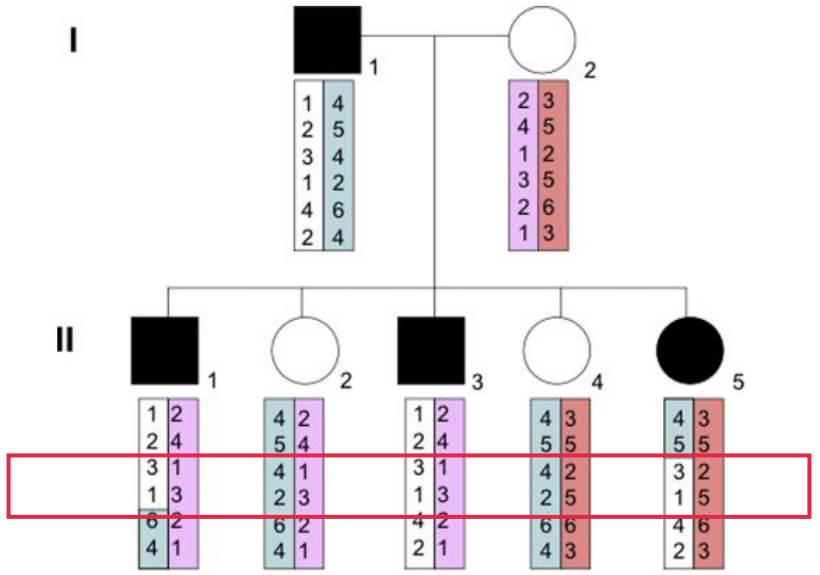
What came before GWAS?

Pre-genomics tools:

- linkage studies for ‘mendelian’ traits
- candidate gene studies

However,

- *limited in scale*
- *not feasible to conduct population-level studies*
- *poor replication (candidate gene studies)*



Korf and Liu (2012)
Principals and Practice of Clinical Research

Human genome project & SNP chips

‘SNP chips’ (developed in late 1990’s & early 2000’s) enabled the GWAS explosion

TIMELINE

1990-2003
Human Genome Project

2002-2005
HapMap Project

2005
First GWAS published

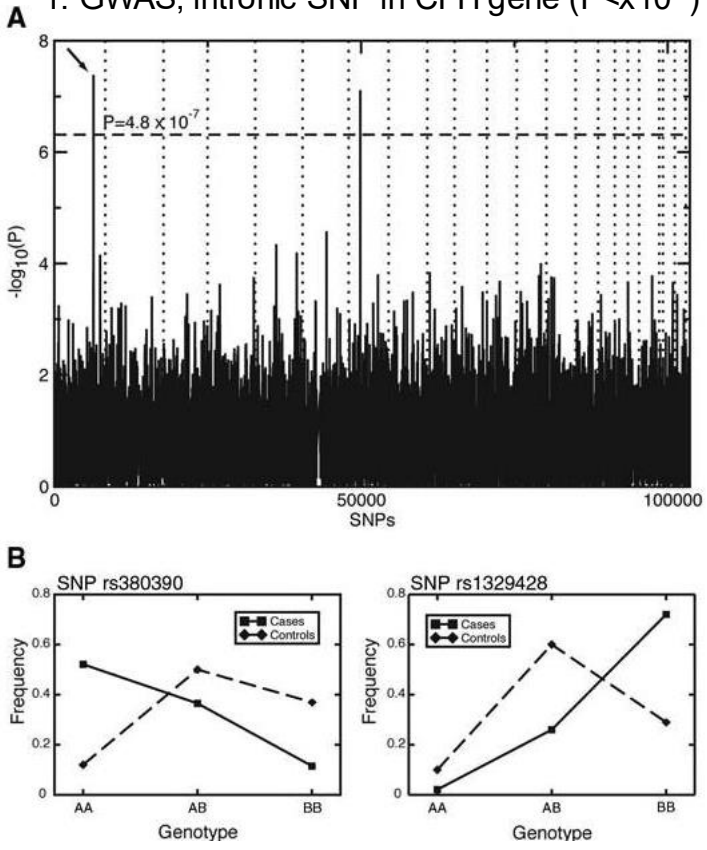
2007
WTCCC published

The first GWAS

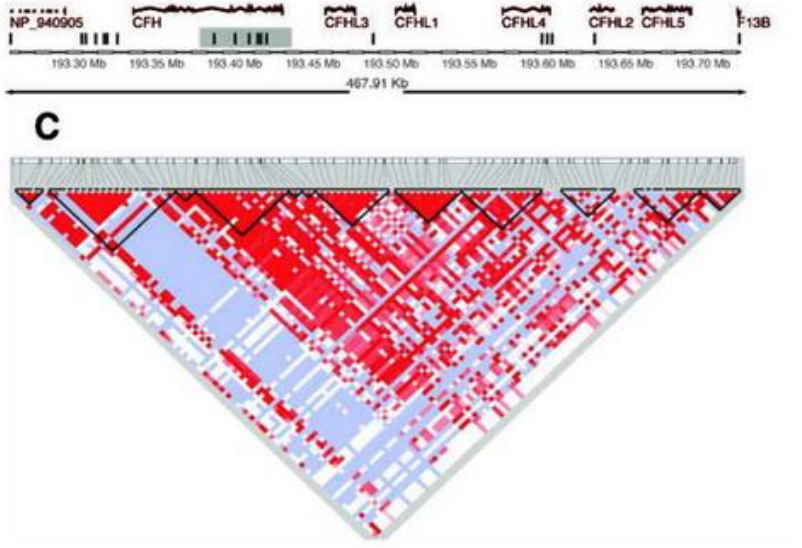
2005: age-related macular degeneration

- Klein et al. 2005 *Science;* 96 cases and 50 controls; 116,204 SNPs

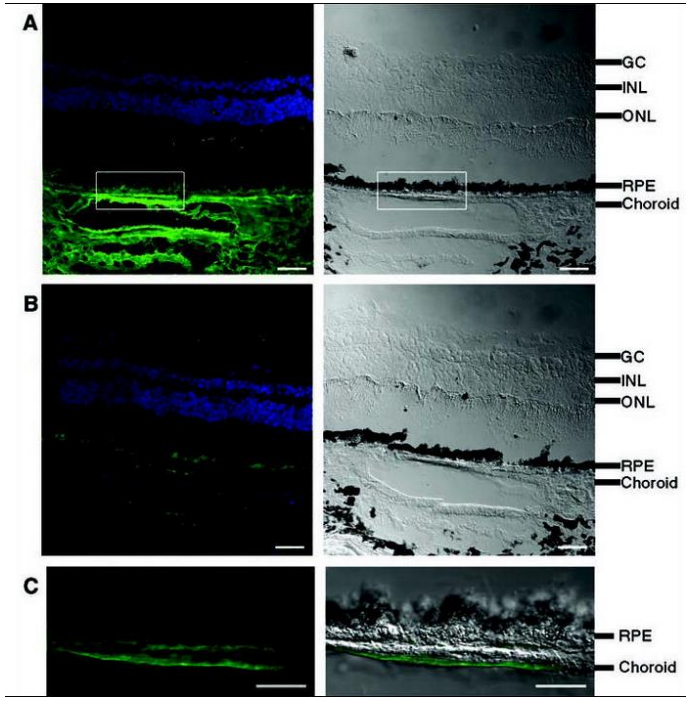
1. GWAS, intronic SNP in CFH gene ($P<x10^{-7}$)



2. re-sequencing region to identify missense variant



3. functional follow-up of CFH gene in retina

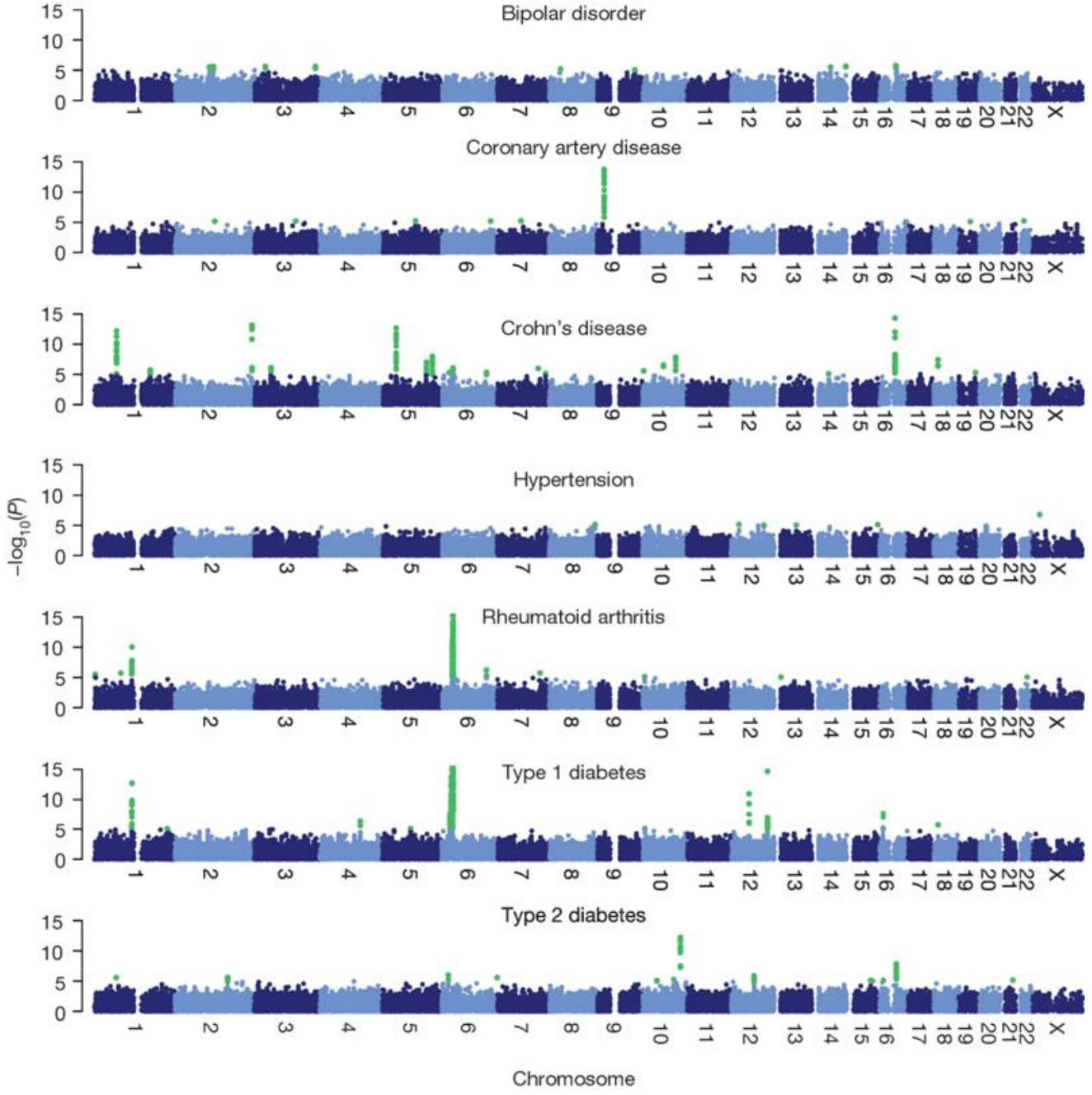


From concept to reality

- Landmark paper,

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), 661–678.

- First large scale GWAS
- 14,000 cases over 7 diseases
- 3,000 shared controls
- 500K Affymetrix GeneChip
- Consortium model



Initial empirical results

Interesting biological insights

However, even a genome wide scan did not explain all the heritable variation

- effects were small
- few loci significant, given the multiple testing burden

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Genome-wide association studies have identified hundreds of genetic variants associated with complex human diseases and traits, and have provided valuable insights into their genetic architecture. Most variants identified so far confer relatively small increments in risk, and explain only a small proportion of familial clustering, leading many to question how the remaining, 'missing' heritability can be explained. Here we examine potential sources of missing heritability and propose research strategies, including and extending beyond current genome-wide association approaches, to illuminate the genetics of complex diseases and enhance its potential to enable effective disease prevention or treatment.

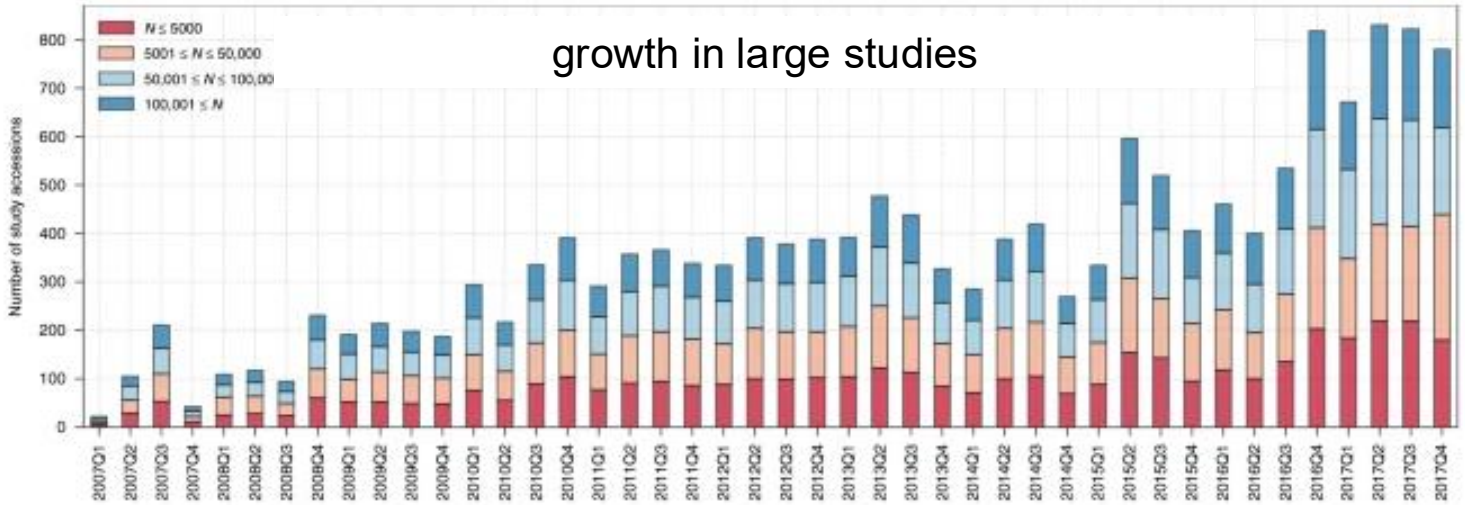
Table 1 | Estimates of heritability and number of loci for several complex traits

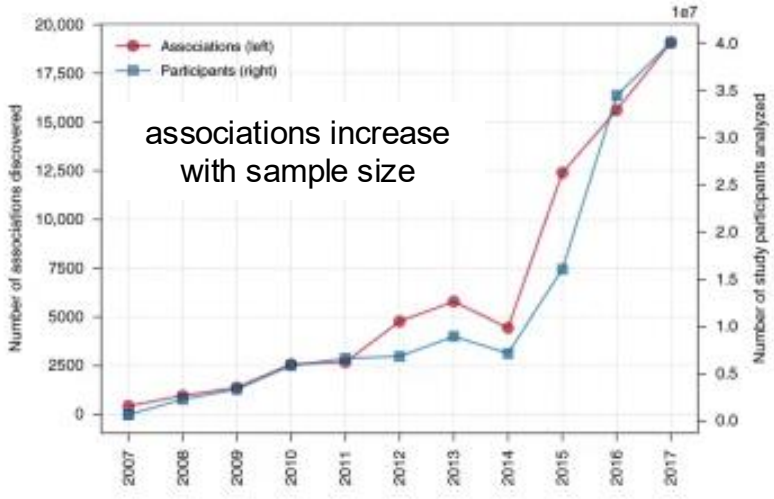
Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual* phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

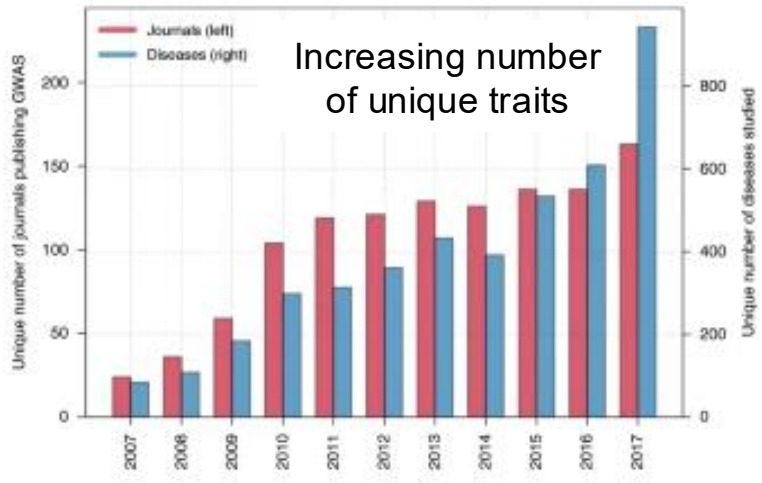
*Residual is after adjustment for age, gender, diabetes.

Trends in GWAS

- More markers
- Bigger sample sizes
- New traits & diseases







Mills & Rahal (2019) *Communications Biology*

Largest published GWAS has 5.4M people!

Article

A saturated map of common genetic variants associated with human height

https://doi.org/10.1038/s41586-022-05275-y

Received: 19 December 2021

Accepted: 24 August 2022

Published online: 12 October 2022

Open access

Check for updates

Common single-nucleotide polymorphisms (SNPs) are predicted to collectively explain 40–50% of phenotypic variation in human height, but identifying the specific variants and associated regions requires huge sample sizes[1]. Here, using data from a genome-wide association study of 5.4 million individuals of diverse ancestries, we show that 12,111 independent SNPs that are significantly associated with height account for nearly all of the common SNP-based heritability. These SNPs are clustered within 7,209 non-overlapping genomic segments with a mean size of around 90 kb, covering about 21% of the genome. The density of independent associations varies across the genome and the regions of increased density are enriched for biologically relevant genes. In out-of-sample estimation and prediction, the 12,111 SNPs (or all SNPs in the HapMap 3 panel[2]) account for 40% (45%) of phenotypic variance in

Yengo et al. (2022) *Nature*

What have we learned?

- smaller effect sizes for variants
- most ‘hits’ in non-coding regions
- thousands of variants, e.g. > 10,000 for height
- pleiotropy is pervasive
- association ≠ causation
- QC is critical due to multiple testing burden

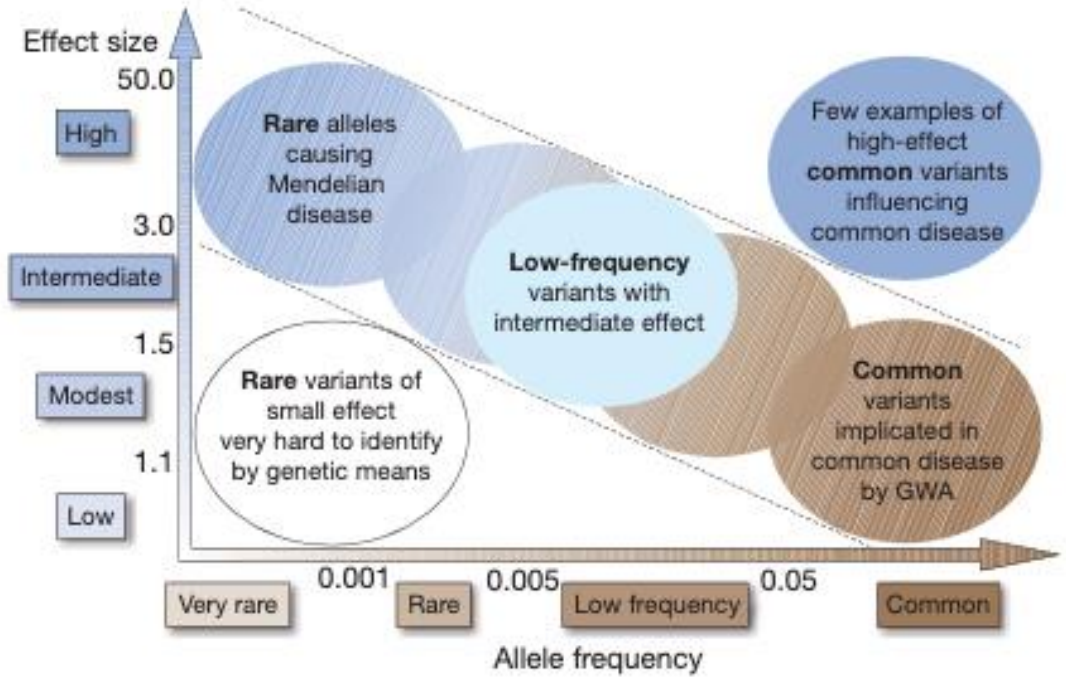


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

Manolio et al. (2009)

The GWAS approach - methodology

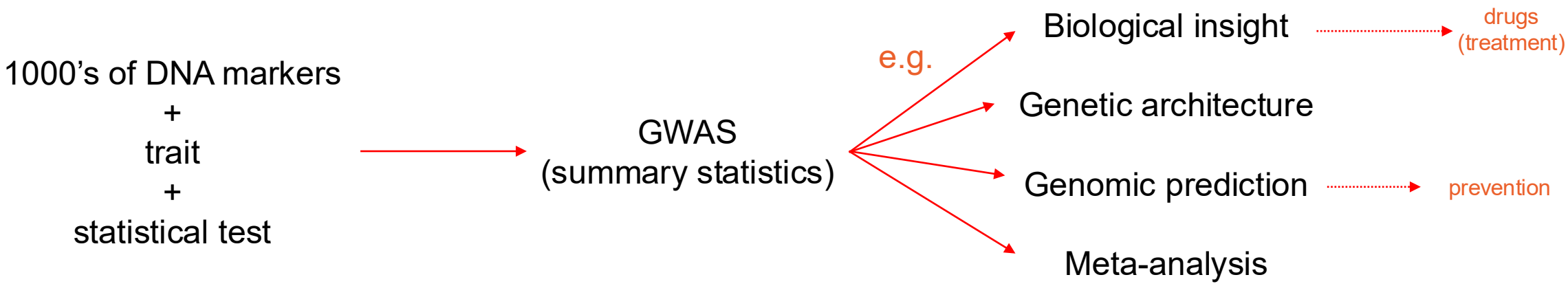
- A **G**enome **W**ide **A**ssociation **S**tudy is a method for identifying associations between locations in the genome and a trait of interest

- Three key parts to a GWAS:
  - A trait of interest or phenotype
  - Genetic markers measured across the genome
  - Statistical test of association between markers & phenotype

Why conduct a GWAS?

GWAS has established itself as a foundation of genomic analysis (mapping genotype – to – phenotype) in human genetics, *Why?*

- identify genes involved in disease, i.e. biological insight
- starting point for many downstream analyses
- easy data sharing, increased experimental power



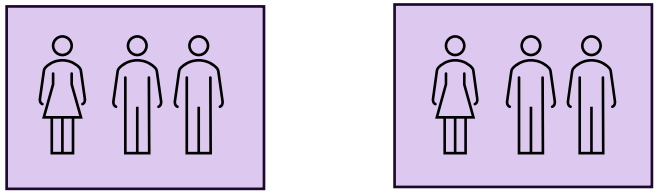
GWAS methodology - overview

phenotype of interest → genotypes → Association between genotype & phenotype

QC

Binary trait:

case: control:



Binary trait: e.g. chi-square test

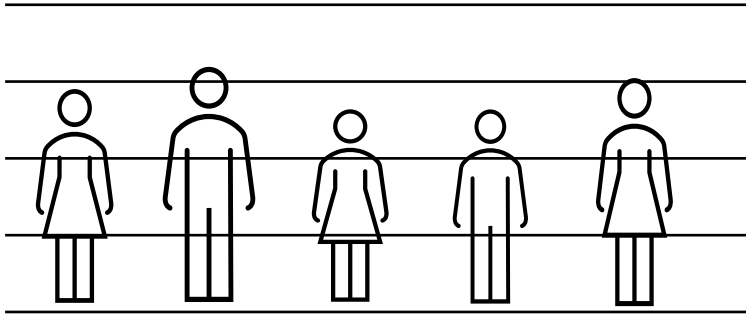
Alleles

	1	2	Total
Case	n_1	n_2	2N
Ctrl	m_1	m_2	2M
Total	T_1	T_2	2(N+M)

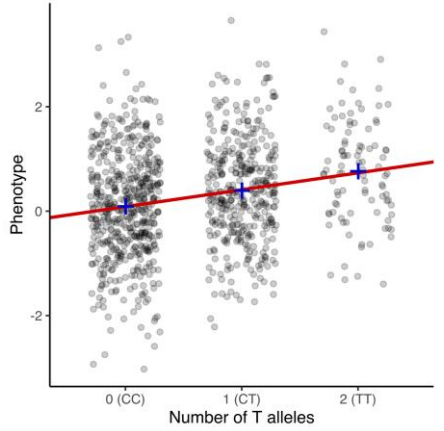
2x2 contingency table

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Quantitative trait:

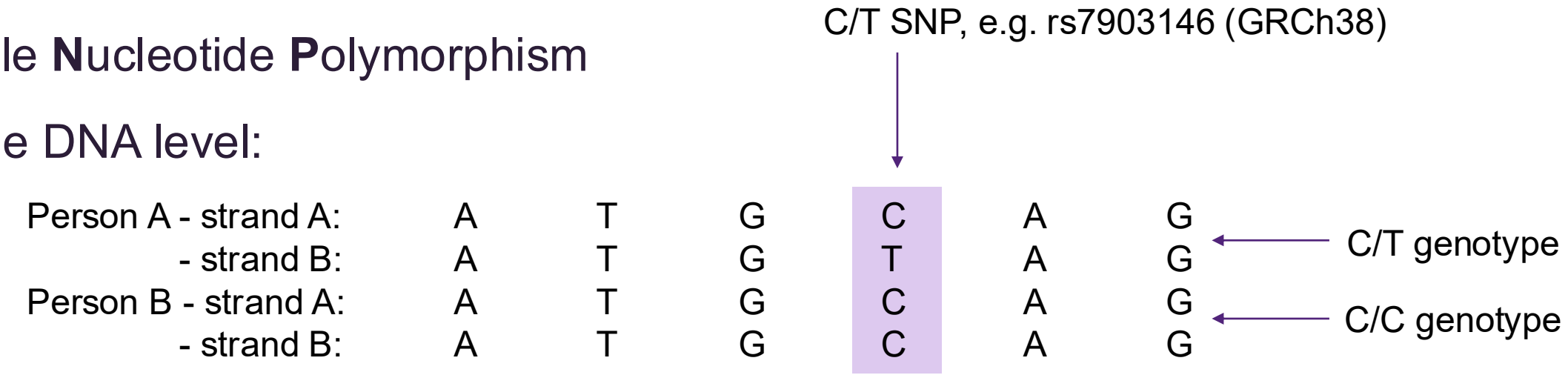


Quantitative trait: linear regression



GWAS methodology – what's a SNP?

- **S**ingle **N**ucleotide **P**olymorphism
- At the DNA level:



- At the population level:

C/C genotype homozygous 20/100 people	C/T genotype heterozygous 55/100 people	T/T genotype homozygous 25/100 people

frequency (T allele) = (55x1 + 25x2)/(100x2) = 0.525
frequency (C allele) = (20x2 + 55x1)/(100x2) = 0.475
'C' is the minor allele

GWAS methodology – what is a SNP?

Minor allele = allele with lowest frequency

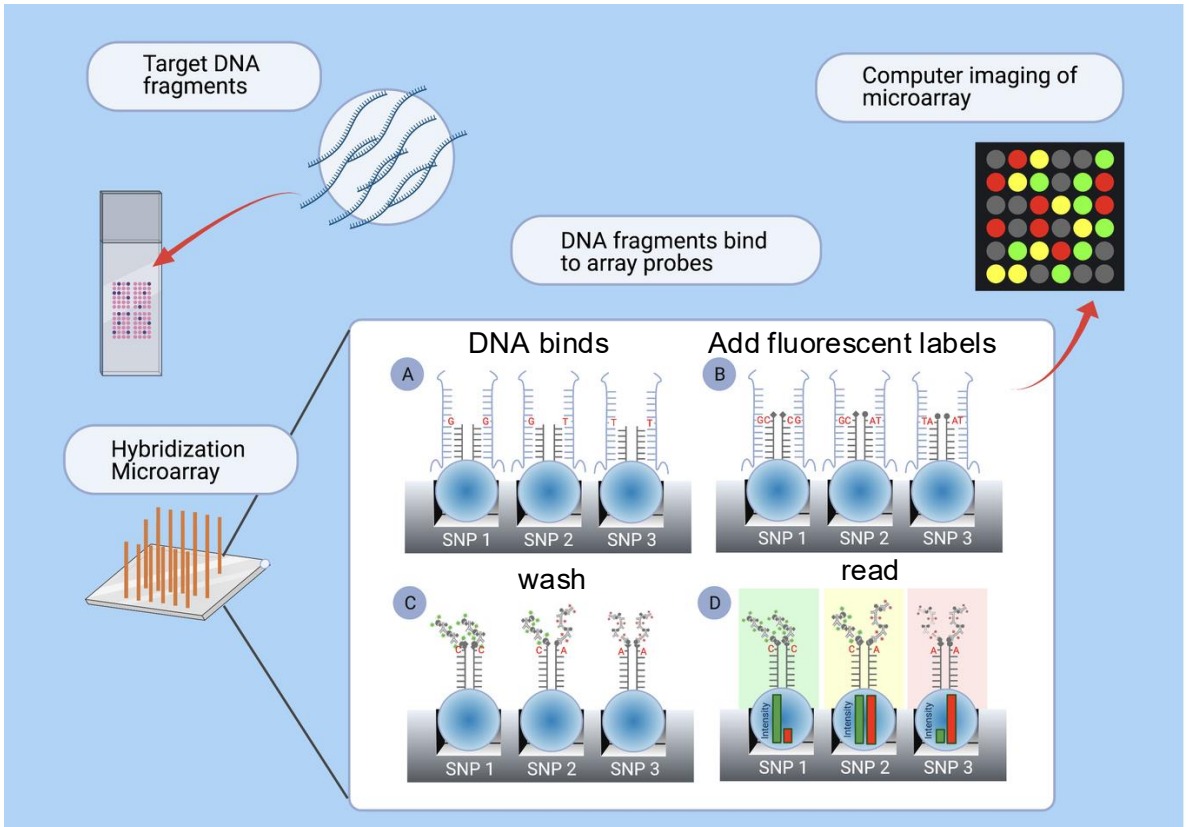
choose a reference allele, count non-reference (alternate) alleles

e.g. ‘C’ allele as the reference, count ‘T’ alleles

C/C genotype homozygous reference 0 alternate alleles	C/T genotype heterozygous 1 alternate allele	T/T genotype homozygous alternate 2 alternate alleles

GWAS methodology – SNP chips

‘SNP chips’ measure 1000’s of markers all over the genome at low cost



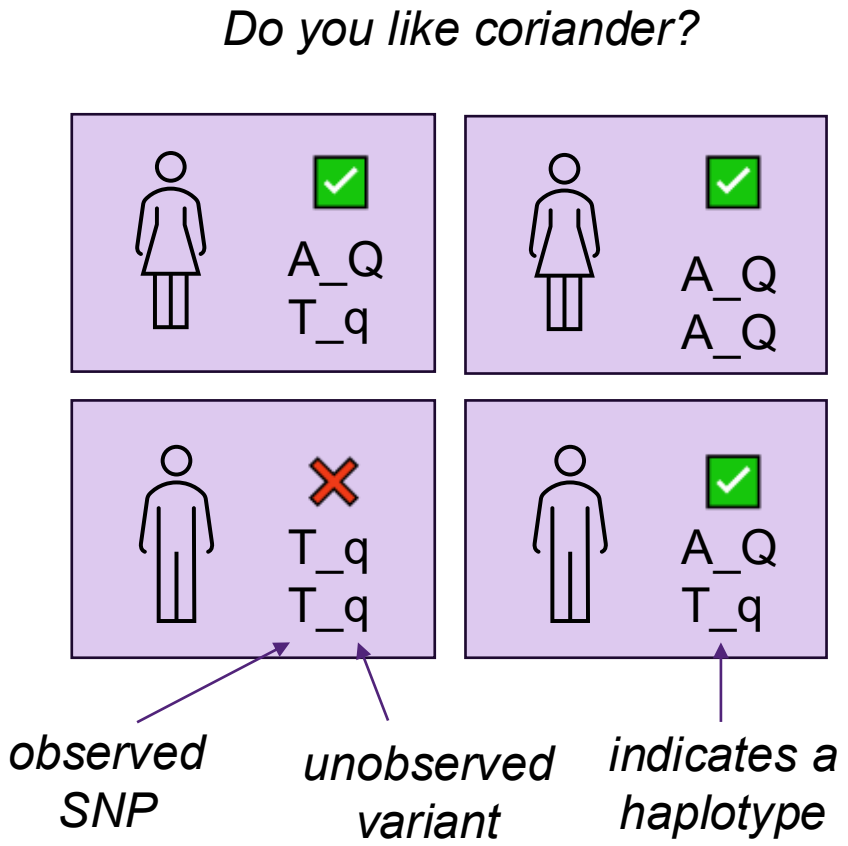
https://www.fjc.gov/content/361256/dna-genotyping-how-it-differs-sequencing-and-relevant-methods

GWAS methodology - basic principle

‘SNP chips’ they allow us to test for associations between DNA and phenotype

- exploit population-level LD (linkage disequilibrium) between SNP and a causal variant
- 1000’s of SNP lets you to scan the whole genome

e.g. the ‘A’ allele of a SNP is associated with coriander preference (y/n). It is in LD with an unobserved variant ‘Q’ which causes variation in the trait.



GWAS methodology – single SNP

Tests of association:

for a quantitative (continuous) trait, use linear regression,

$$y = a + bx$$

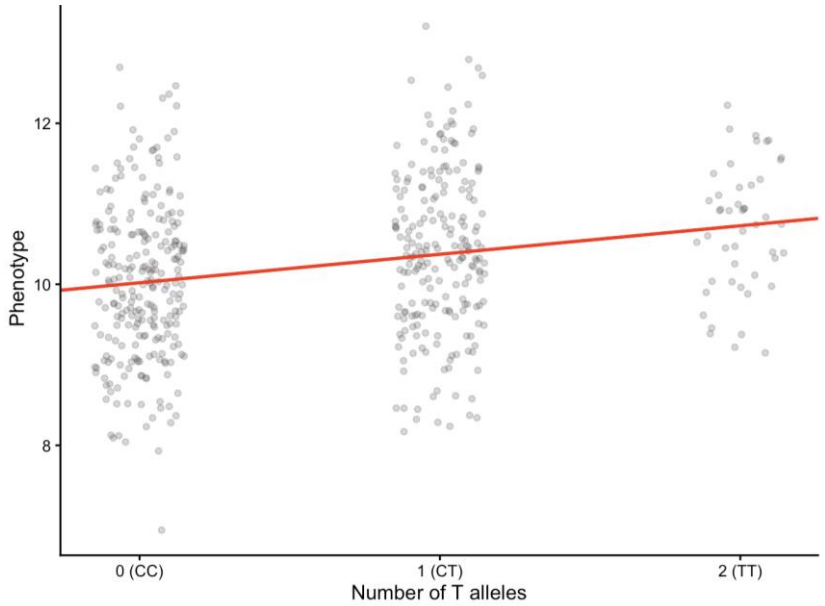
phenotype
e.g. height

intercept

number of T alleles

regression slope

Is the slope different from 0?



GWAS methodology – single SNP

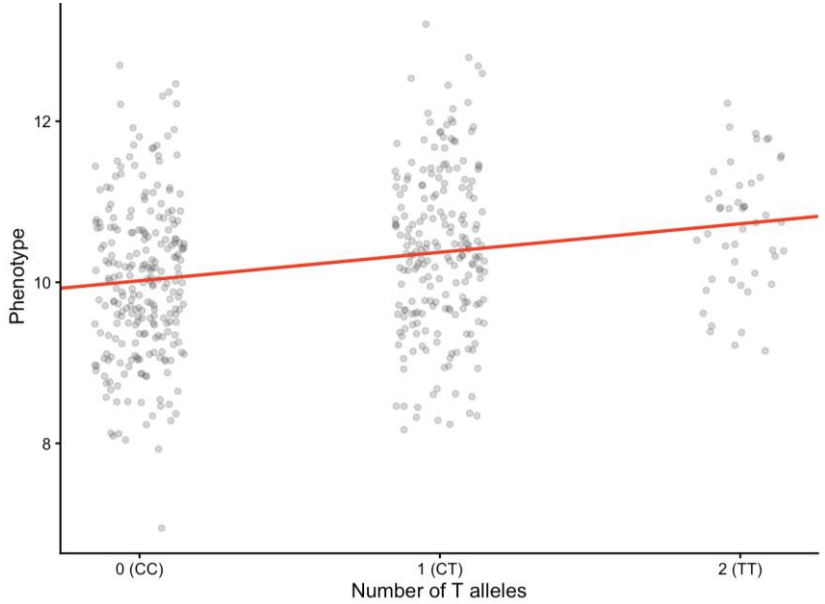
Tests of association:

for a quantitative (continuous) trait, use linear regression,

```
Call:
lm(formula = pheno ~ snp)

Residuals:
     Min       1Q   Median       3Q      Max
-3.07229 -0.68485 -0.00349  0.65494  2.83166

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.01924    0.05903 169.743  < 2e-16 ***
snp          0.35477    0.06624   5.356  1.3e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



GWAS methodology – single SNP, binary trait

Tests of association:

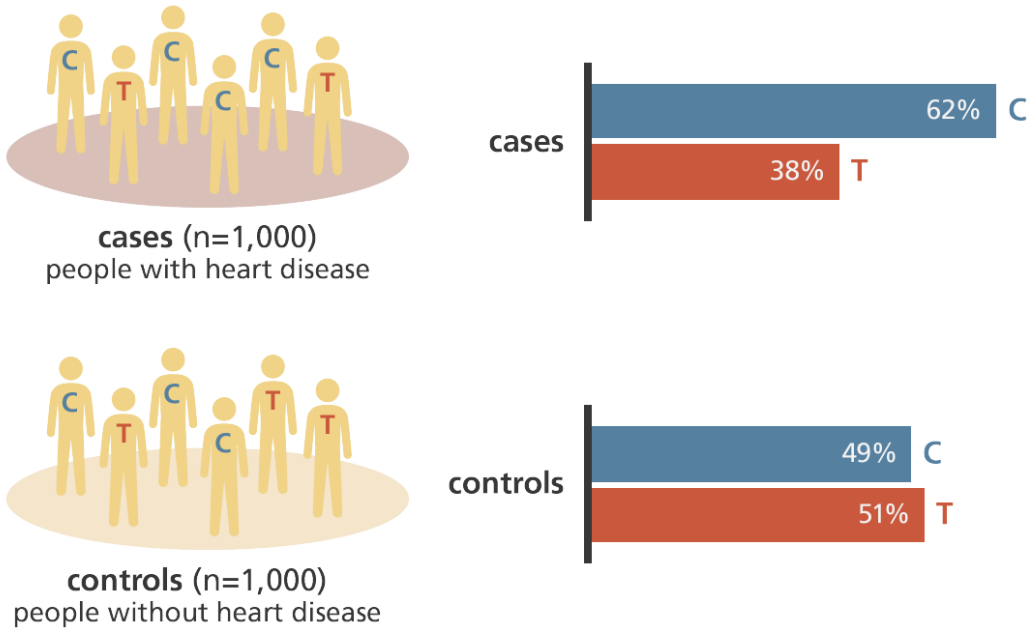
for a binary (0/1) trait,

	C allele	T allele	Total
Case	1240	760	2000
Control	980	1020	2000
Total	2220	1780	4000

Chi-squared test of association:

"Is the frequency of the 'C' allele different in cases vs. controls"

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$



GWAS methodology – single SNP, binary trait

For a binary (0/1) trait,

	C allele	T allele	Total
Case	1240	760	2000
Control	980	1020	2000
Total	2220	1780	4000

OPTION A:

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  mat
X-squared = 67.903, df = 1, p-value < 2.2e-16
```

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

OPTION B:

```
Call:
glm(formula = case ~ snp, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.23531    0.04274   5.505 3.68e-08 ***
snp         -0.52955    0.06421  -8.247  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

GWAS methodology – variations

Many variations on the basic ‘y = a + bx’ model:

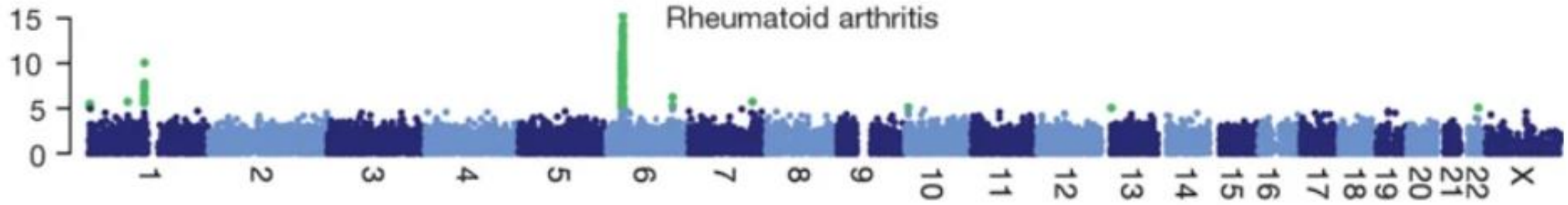
1. change ‘x’
   - different genetic models
   - different genomic data e.g. WGS
2. change ‘y’
   - quantitative vs. binary
   - gene expression, methylation, etc.
3. change statistical model
   - linear mixed models, generalised linear mixed model
   - rare burden tests

genotype	C/C	C/T	T/T	
additive	0	1	2	linear increment
dominant	0	1	0	heterozygote vs. homozygote
recessive	0	0	1	homozygote alt. vs. other genotypes

GWAS methodology - output

GWAS results are typically visualised as a ‘Manhattan plot’

y-axis: $-\log_{10}P$



x-axis: genome location of markers

- SNPs/markers with the strongest associations will have the greatest negative logarithms, and will tower over the background of unassociated SNPs
  - like skyscrapers in Manhattan →

P-value	$-\log_{10}(P)$
0.5	0.30
0.01	2
0.001	3
0.0001	4
... etc.	

GWAS methodology - output

Usually, GWAS ‘summary statistics’ are required (by journals) to be made public

```
SNP A1 A2 freq b se p N
rs1001 A G 0.8493 0.0024 0.0055 0.6653 129850
rs1002 C G 0.0306 0.0034 0.0115 0.7659 129799
rs1003 A C 0.5128 0.0045 0.0038 0.2319 129830
```

Cell Genomics

CellPress
OPEN ACCESS

Perspective

Workshop proceedings: GWAS summary statistics standards and sharing

Jacqueline A.L. MacArthur,[1,2,*] Annalisa Buniello,[1] Laura W. Harris,[1] James Hayhurst,[1] Aoife McMahon,[1] Elliot Sollis,[1] Maria Cerezo,[1] Peggy Hall,[3] Elizabeth Lewis,[1] Patricia L. Whetzel,[1] Orli G. Bahcall,[4] Inês Barroso,[5] Robert J. Carroll,[6] Michael Inouye,[7,8,9] Teri A. Manolio,[3] Stephen S. Rich,[10] Lucia A. Hindorff,[3] Ken Wiley,[3] and Helen Parkinson[1,*]

Table 1. Recommended standard reporting elements for GWAS SumStats

Data element	Column header	Mandatory/Optional
variant id	variant_id	One form of variant ID is mandatory, either rsID or chromosome, base pair location, and genome build[a]
chromosome	chromosome	
base pair location	base_pair_location	
p value	p_value	Mandatory
effect allele	effect_allele	Mandatory
other allele	other_allele	Mandatory
effect allele frequency	effect_allele_frequency	Mandatory
effect (odds ratio or beta)	odds_ratio or beta	Mandatory
standard error	standard_error	Mandatory
upper confidence interval	ci_upper	Optional
lower confidence interval	ci_lower	Optional

GWAS methodology – multiple testing burden

In GWAS we are particularly concerned with the false-positive rate. *Why?*

Every statistical test has a false-positive rate 'α' - set by the researcher. This is the probability of rejecting the null hypothesis (of no association) when it's true. α is the threshold where you declare your test 'significant', i.e. if $p < \alpha$ then we reject H_0

If you run one test at $\alpha = 0.05$, you expect to reject H_0 when it's true 5% of the time. But in GWAS there are $1x10^6$ tests,

If $\alpha = 0.05$ you expect $0.05x1x10^6 = 50{,}000$ significant tests (!)

Thus, we need to adjust α to keep the <u>experiment-wise</u> false-positive rate at 5%

- are the tests independent?

Yes

Bonferroni correction, $\alpha^* = \alpha/n$

(n = number of tests)

(typical for GWAS)

No – because of LD

What to do? – P = $5x10^{-8}$

(more on this next session)

Summary

- GWAS scan the genome for associations between DNA markers (e.g. SNPs) and a trait of interest (e.g. height or heart disease) to identify genomic regions associated with the trait
  - Rely on linkage disequilibrium between markers & causal variant
  - Understanding biological underpinning of associations remains a difficult
  - GWAS are starting point for many downstream analyses

- Issues around multiple testing are the main statistical challenge for GWAS

Practical Session

Part 1: conduct a small GWAS in R
Part 2: make a QQ-plot

- 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 & run locally for the pracs from:
  `/data/module1/downloadsDataMonPM.zip`