Summer Institutes of Statistical Genetics, 2023

Module 2: INTRODUCTION TO GENETICS AND GENOMICS

Greg Gibson and Joe Lachance

Georgia Institute of Technology

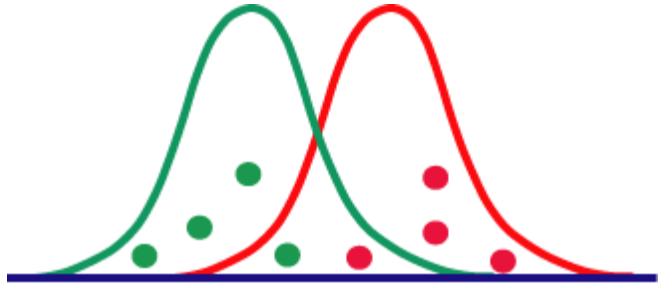
Lecture 5: GENOME-WIDE ASSOCIATION STUDIES

greg.gibson@biology.gatech.edu

http://www.cig.gatech.edu

Principle of Association Studies

Individual	Site	Score
1	A T C C G A	9
2	A C T C G A	8
3	A C C A - G	3
4	T T C A G A	5
5	A T C A G A	2
6	A C C C - G	7
7	T C T A - G	4
8	A T C C G A	8

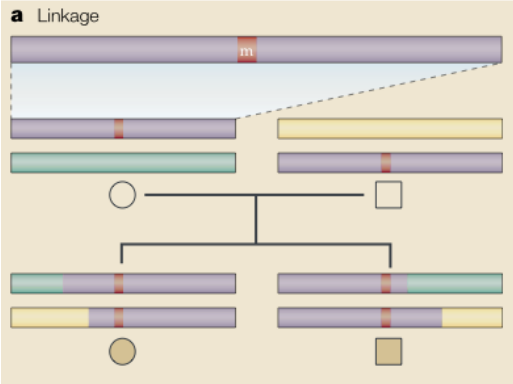


Are the phenotype scores associated with each class of SNP drawn from the same or different distributions ?

Linkage versus Association

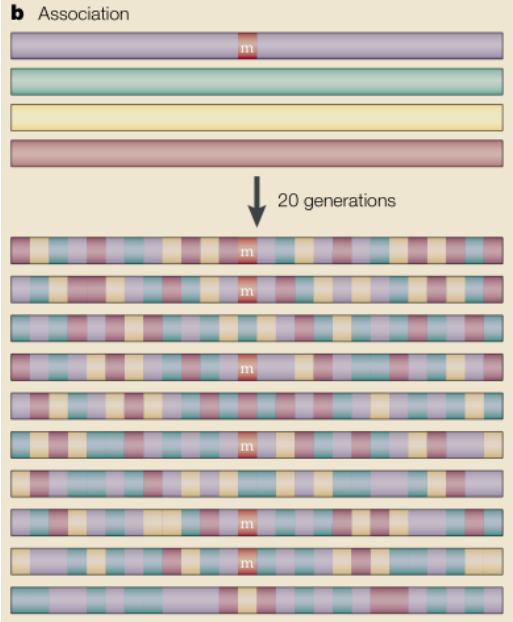
Linkage examines recent recombination events in a pedigree:

- over just several generations
- large chromosomal regions detected
- no information on allele frequency



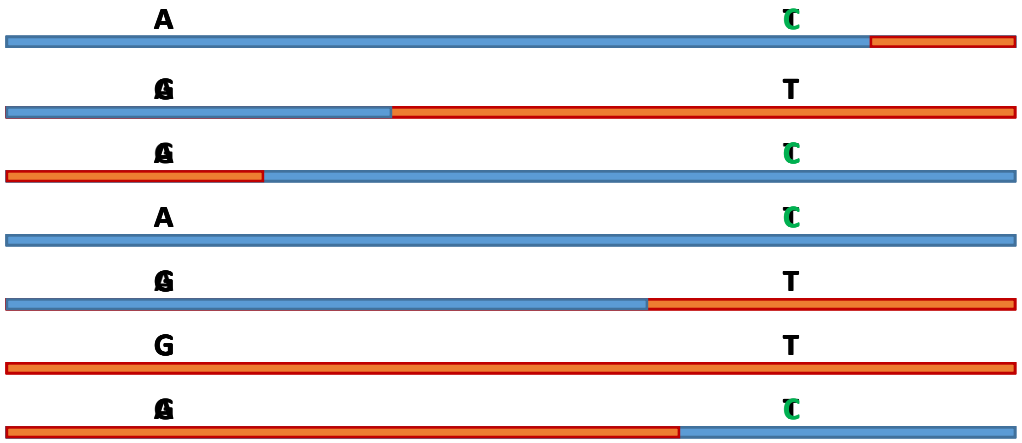
Association examines historical recombination events in a population:

- basically a 10,000 generation pedigree
- resolution to single genes
- estimates effect size and frequency



Cardon & Bell. 2003. *Nat. Rev. Genet*. **2**: 91

Why LD (Linkage Disequlibrium) happens



When a mutation occurs, by definition it is only on one chromosome and hence "associated" with the genotypes elsewhere on that chromosome.

Over time, the mutation increases in frequency and becomes a polymorphism. It remains in LD with the genotypes on the chromosome it appeared on.

Eventually recombination breaks up the LD, in proportion to genetic distance.

Measurement of LD

LD is the non-random association of genotypes.

Expected

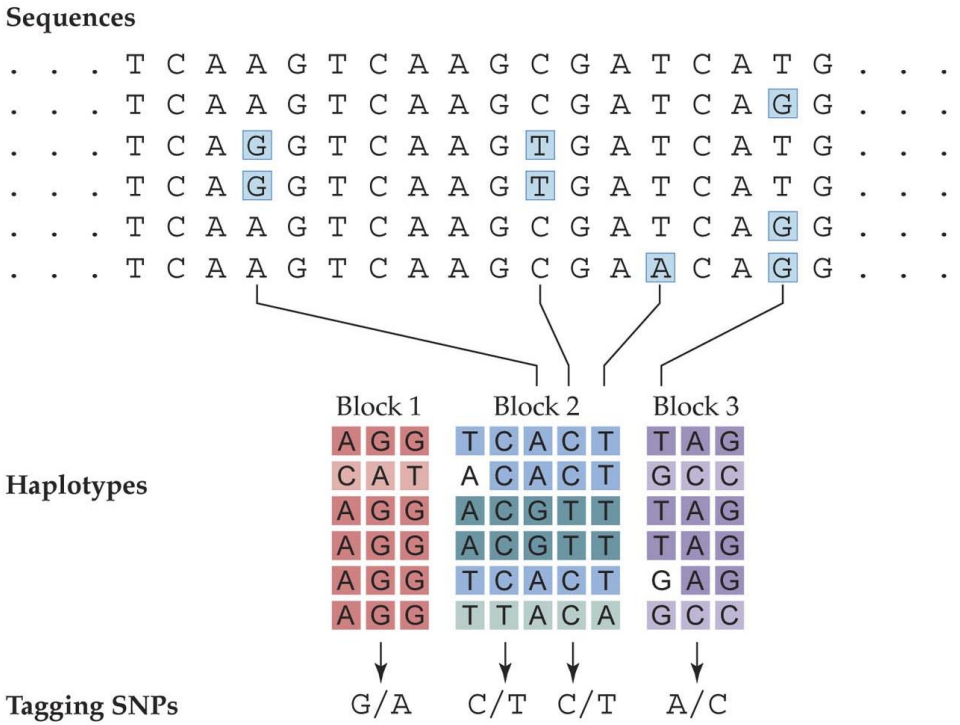
		AA	AG	GG
		24	48	24
TT	24	6	12	6
TC	48	12	24	12
CC	24	6	12	6

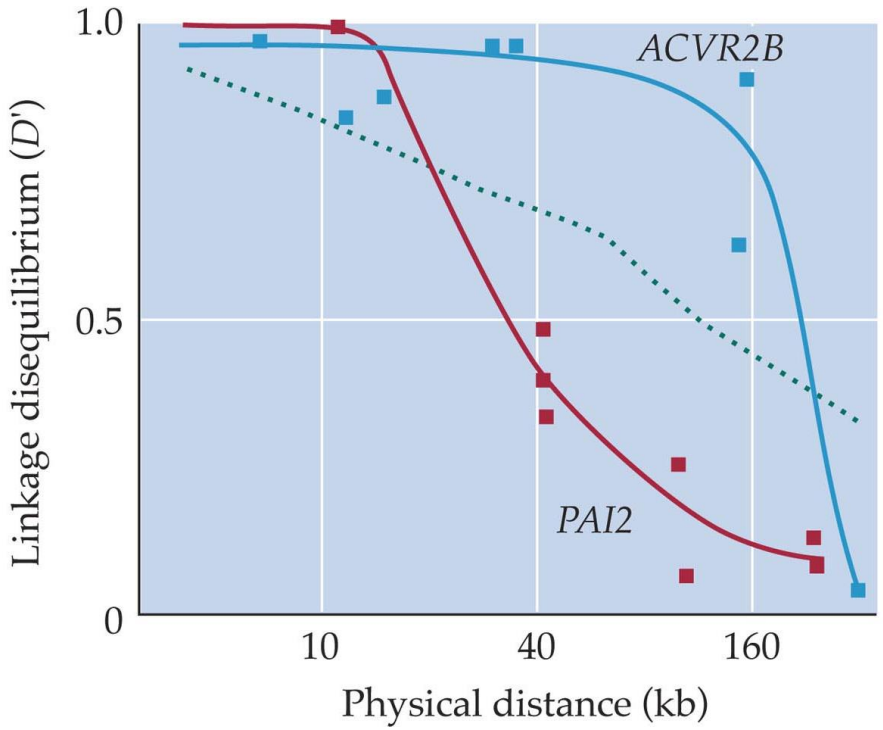
Observed

		AA	AG	GG
		24	48	24
TT	24	24	0	0
TC	48	0	48	0
CC	24	0	0	24

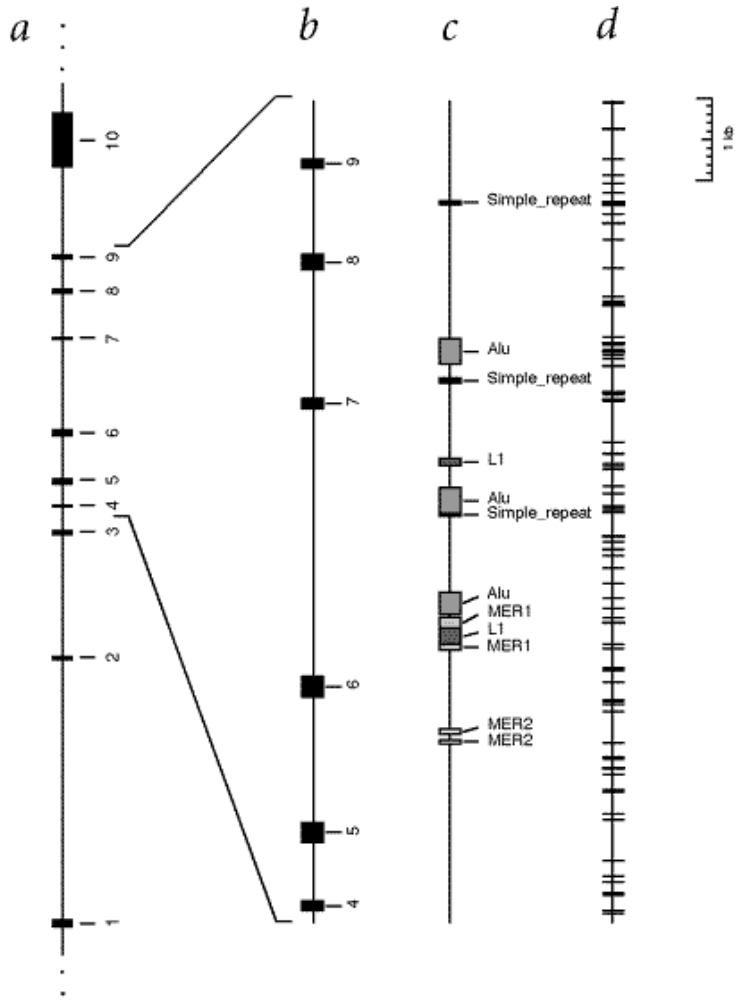
LD can be quantified as a proportion of the maximal possible LD given the allele frequencies (D'),
Or as the squared correlation between allele frequencies (r^2).

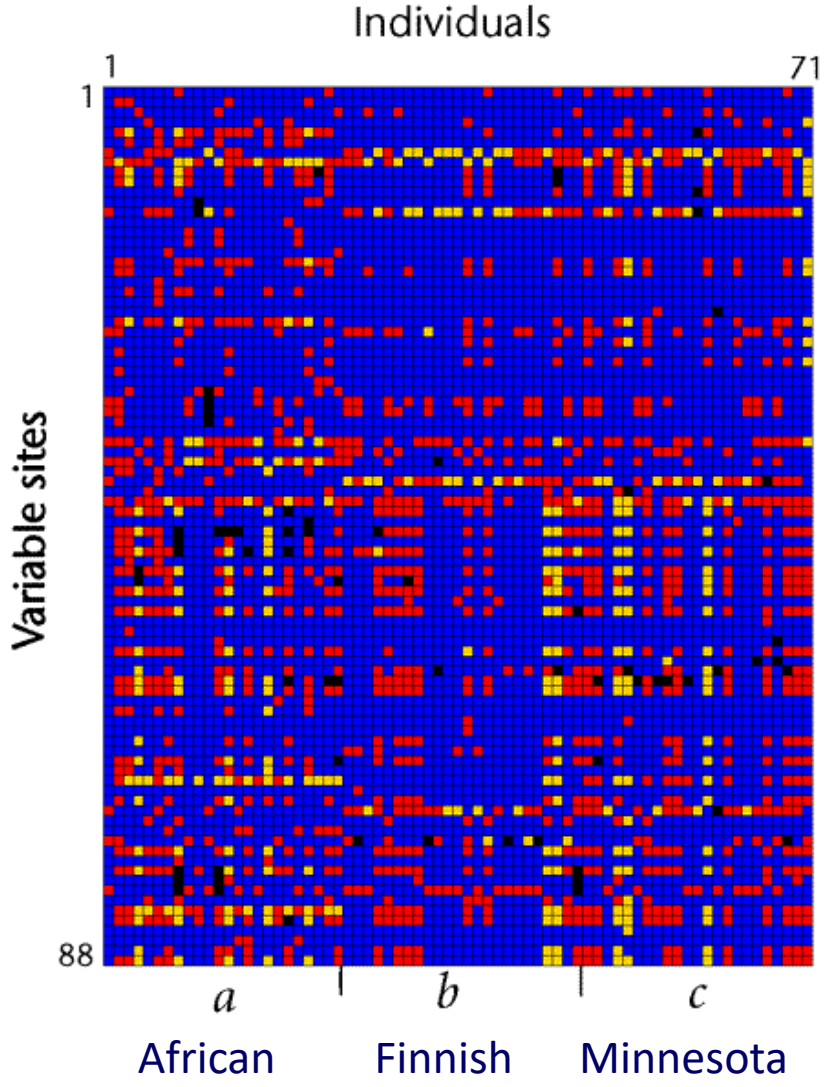
Haplotypes and Tagging SNPs



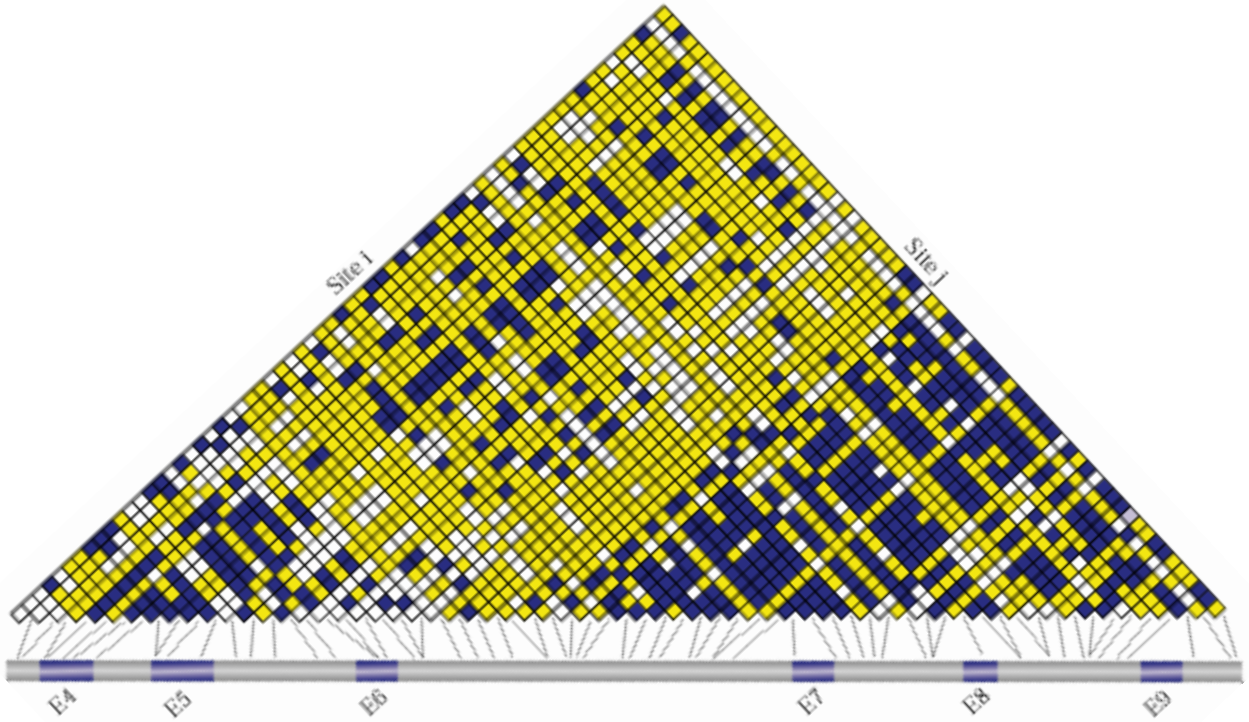


Visualizing LD: The LPL example

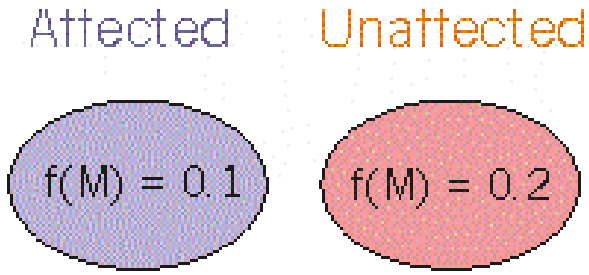




An LD Plot (for the LPL locus)



Case-Control and Family Designs



	Observed		Expected	
	Allele M	Allele m	Allele M	Allele m
Affected	34	278	61	265
Unaffected	69	256	62	269

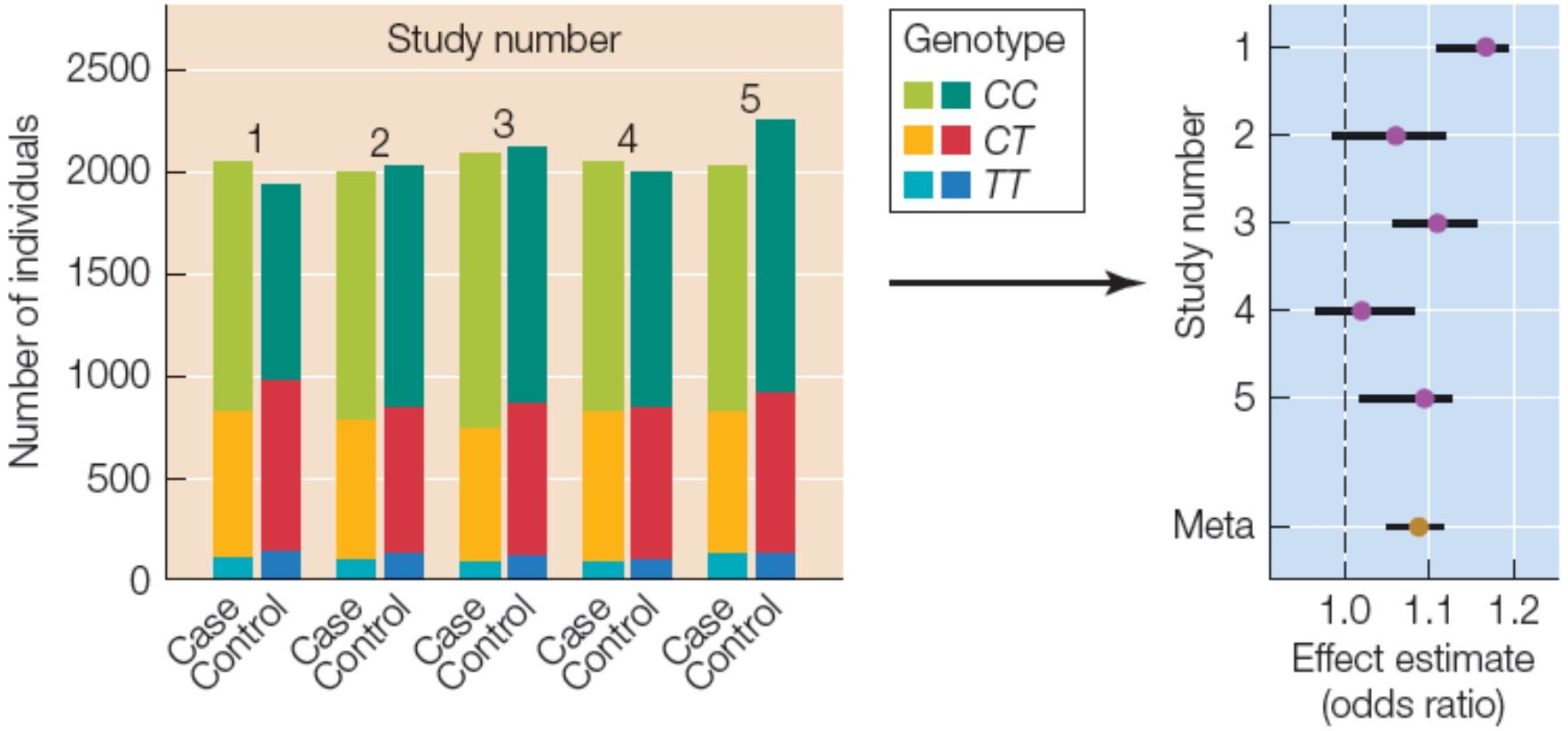
$\chi^2 = 14.0$ $P < 0.001$

$$\frac{M}{m} \times \frac{M}{m} \rightarrow f(M) = 0.63$$

	Transmitted Allele	
	M	m
Observed	78	46
Expected	62	62

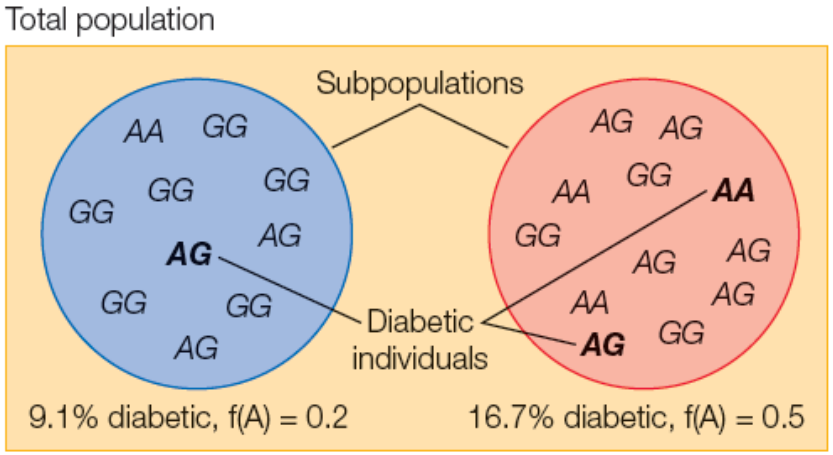
$\chi^2 = 8.2$ $P < 0.001$

Repeatability and Forest Plots



Population Structure

If the allele frequency AND the trait frequency vary among hidden sub-populations, false positives can arise



Blue subpopulation

	AA	*AG*	*GG*
Case	80	640	1280
Control	800	6400	12,800
Case/control	0.1	0.1	0.1

Red subpopulation

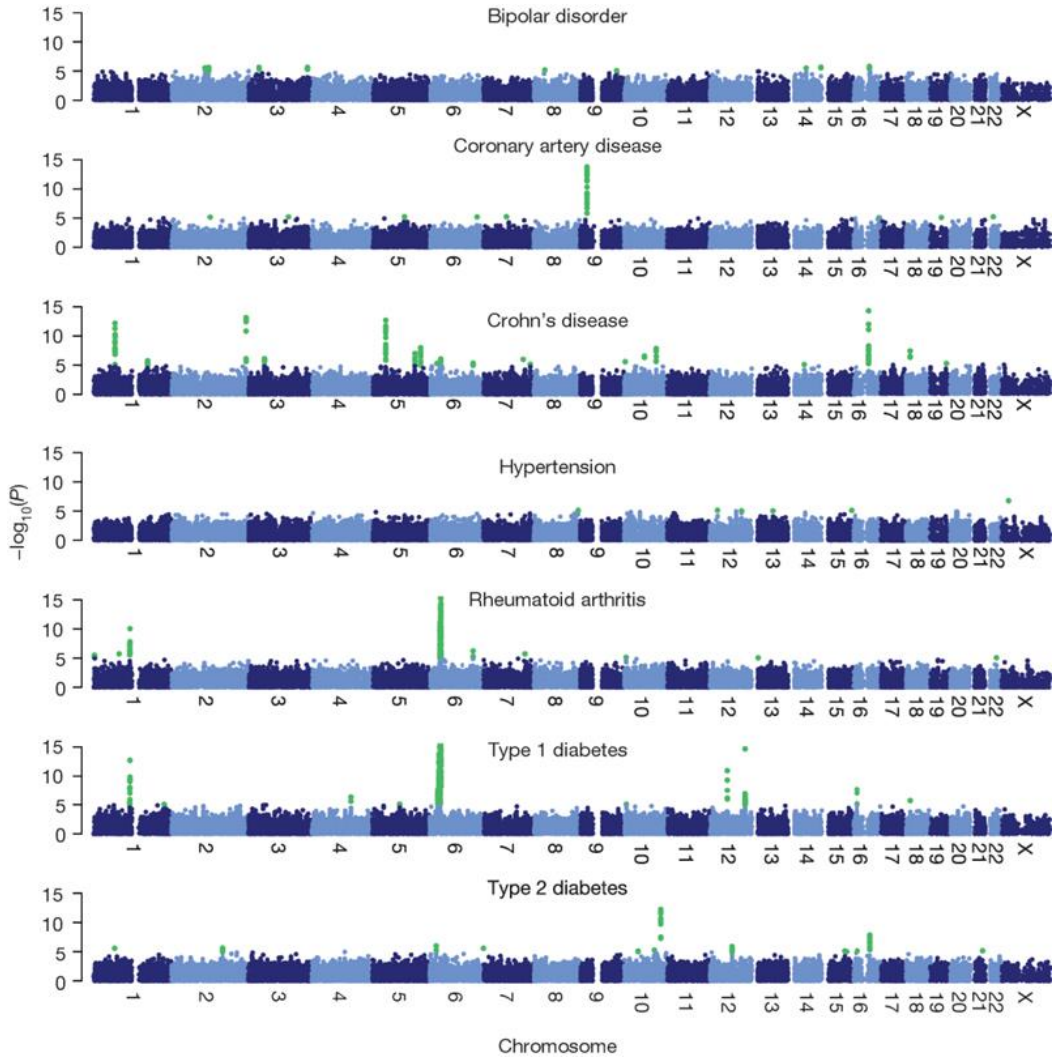
	AA	*AG*	*GG*
Case	200	400	200
Control	1000	2000	1000
Case/control	0.2	0.2	0.2

Total population

	AA	*AG*	*GG*
Case	280	1040	1480
Control	1800	8400	13,800
Case/control	0.155	0.124	0.107

Odds ratio (*A*:*G*) = 1.2 $p = 10^{-8}$

GWAS in 2009: The WTCCC

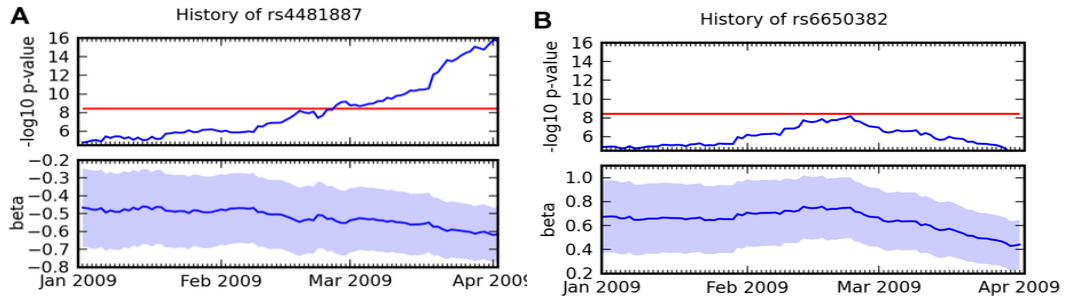


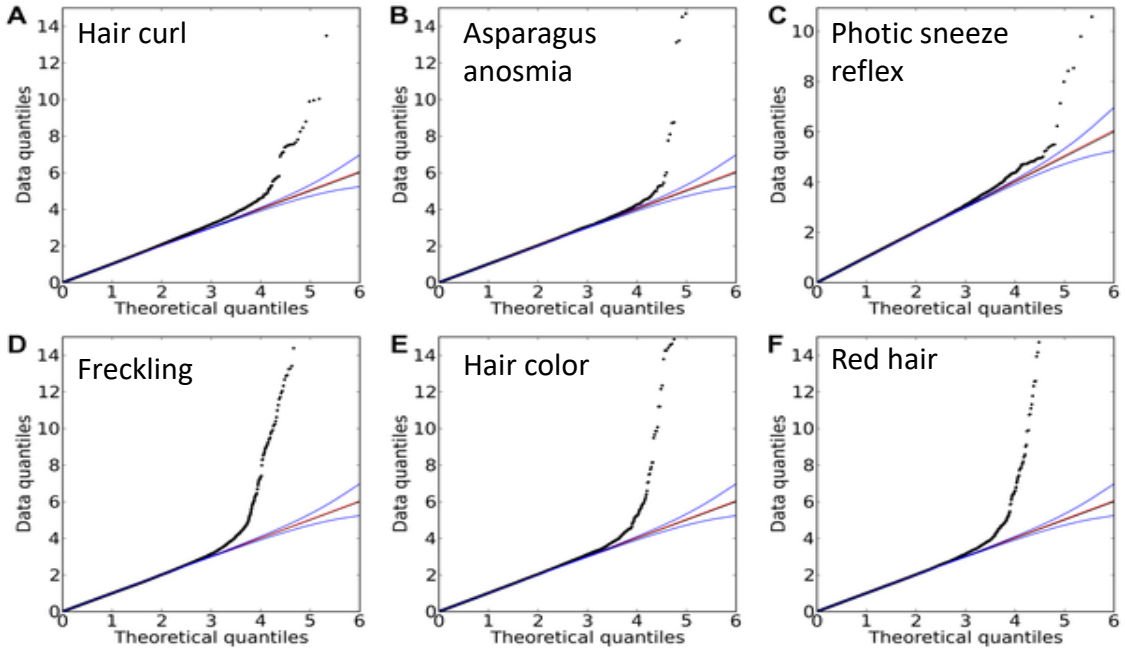
GWAS first appeared 10 years ago, now several new diseases each month

Inflammatory diseases show multiple associations, with some common variants (notably the MHC)

Depression and Hypertension show nothing: likely no variants with a relative risk greater than 1.5

Q-Q Plots in 23andme studies





Other interesting traits:

Endurance Runner vs Sprinter (30% of people change their answer if they know their ACTN3)

Left vs Right Handedness (nothing striking)

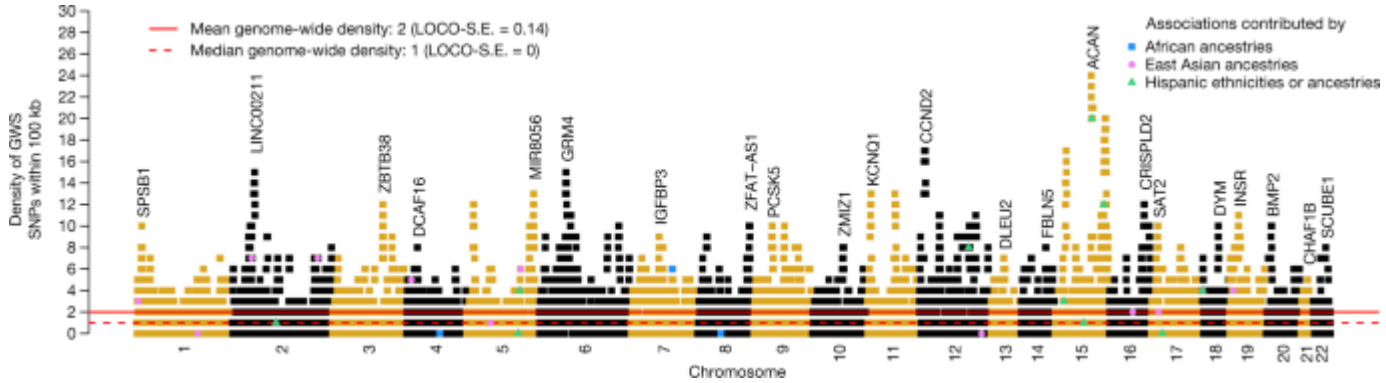
Have you ever needed braces or wisdom teeth surgery?

Breast size
(finds breast cancer risk loci)

Hand-clasp dominance …

Eriksson et al, (2010) *PLoS Genetics* **6**: e1000993

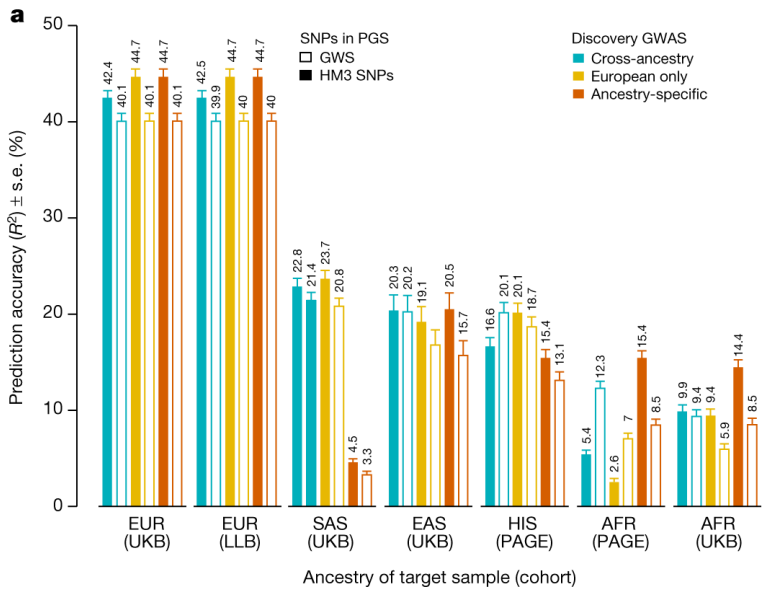
Genetics of Height



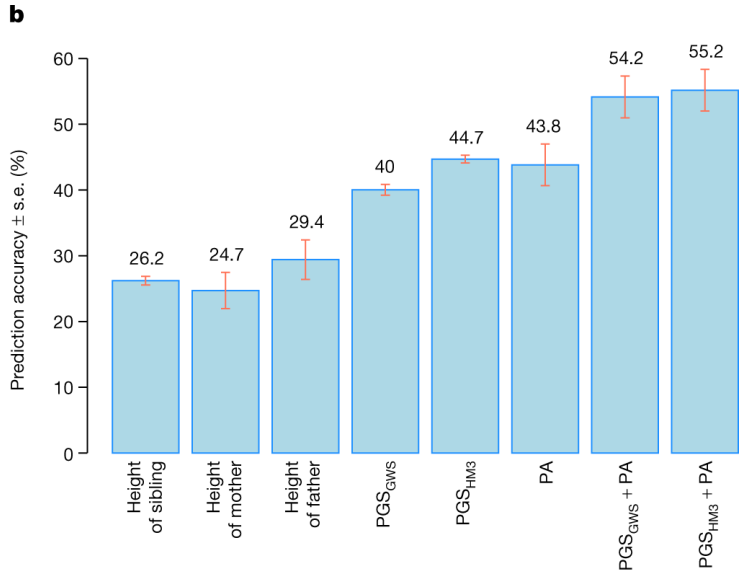
"Brisbane Plot" shows that the saturated map of 12,111 SNPs explaining the common variant contribution to height in EUR is concentrated in ~7200 bins, each 90kb covering 21% of the genome.

90% of the h^2 in Asian and African Continental Ancestry Groups is contained within loci tagged by the European variants

Prediction accuracy drops by ancestry

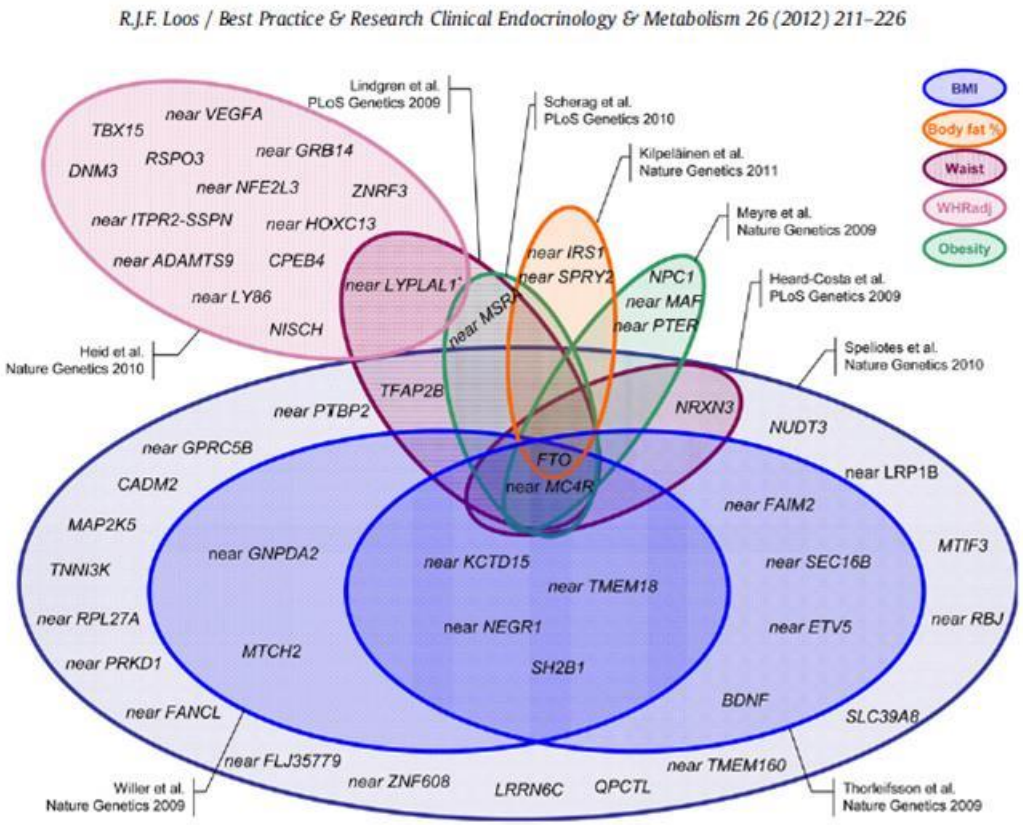


Prediction accuracy adds to parent avg.



Yengo et al (2022) *Nature* **610**: 704-712

Genetics of Obesity



Heritability of obesity ~ 60%

2/3 Americans BMI > 25

One gene, FTO, is repeatedly associated with BMI, hip circumference and weight, in most human populations

Homozygote classes differ in weight by up to 2 kg

Study of 230,000 people →

49 loci for WHR, many linked to adipose, insulin biology 20 loci only in women

Study of 340,000 people →

97 loci for BMI, many linked to neuronal function

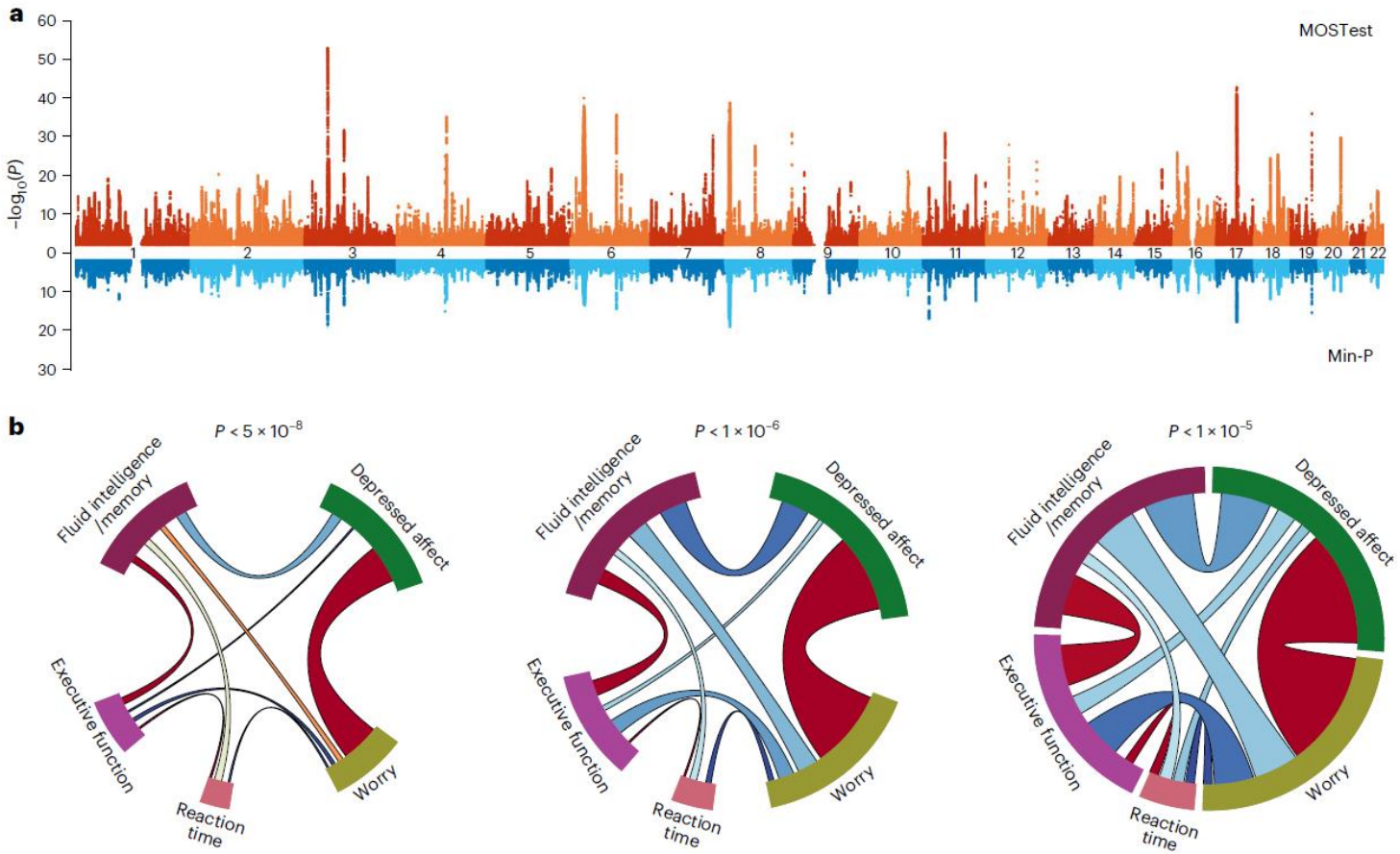
Little overlap with WHR

Genetics of Cognition and Personality

Measures	Cluster	Measure	Abbreviation	Sample size
Neuroticism	Depressed affect	Neuroticism sum-score	NEUR sum-score	274,056
		Are you an irritable person?	Irritable	322,599
		Do you often feel lonely?	Lonely	332,193
		Do you often feel fed-up?	Fed-up	330,478
		Do you ever feel just miserable for no reason?	Miserable	331,782
		Does your mood often go up and down?	Mood swings	329,358
	Worry	Do you suffer from nerves?	Nerves	325,181
		Are you often troubled by feelings of guilt?	Guilt	328,700
		Would you call yourself tense or highly strung?	Tense	327,162
		Are your feelings easily hurt?	Feelings hurt	327,762
		Would you call yourself a nervous person?	Nervous	328,653
		Do you worry too long after an embarrassing experience?	Embarrass	323,698
		Are you a worrier?	Worrier	328,647
Cognition	Fluid intelligence/memory	Fluid intelligence sum-score	FI sum-score	163,375
		Word interpolation	Word interp.	162,937
		Positional arithmetic	Pos. math.	161,768
		Family relationship calculation	Fam. rel. calc.	158,977
		Conditional arithmetic	Cond. math.	144,648
		Synonym	Synonym	120,891
		Chained arithmetic	Chained math.	109,731
		Concept interpolation	Concept interp.	50,331
		Arithmetic sequence recognition	Seq. recog. 2	34,286
		Square sequence recognition	Seq. recog. 1	11,679
		Numeric memory	Num. memory	104,319
		Prospective memory	Prosp. memory	111,079
		Matrix pattern completion	Matrix pattern	22,335
	Executive function	Pair matching:		
		Full game	Pair match. 1	336,993
		Time of full game	Pair match. 2	330,143
		Basic game	Pair match. 3	336,993
		Time of basic game	Pair match. 4	330,777
		Symbol digit substitution	Symb. dig. subs.	94,153
		Tower rearranging	Tower rearrang.	22,159
		Trail making:		
		Part A	Trail making 1	85,595
		Part B	Trail making 2	85,597
	Reaction time	Reaction time	Reaction time	335,066
	Non-heritable	Fluid intelligence:		
		Numeric addition test	–	162,846
		Identify largest number	–	162,989
		Antonym	–	17,417
		Subset inclusion logic	–	3,627

Clusters are derived from genetic correlation-based hierarchical clustering (Fig. 1). Further details are provided in Supplementary Table 1.

Multivariable Analysis Greatly Boosts Power

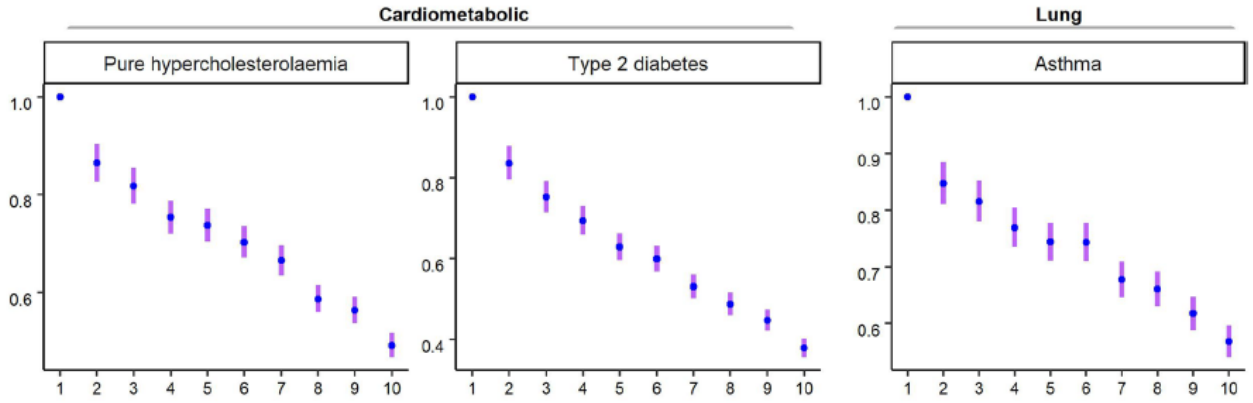


Extensive Pleiotropy within and across cognitive and psychological domains

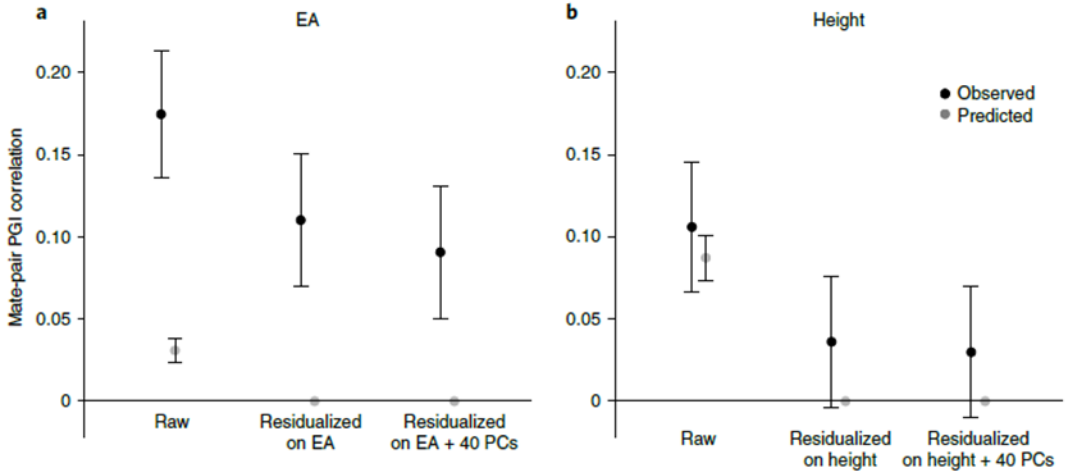
Hindley et al (2023) *Nature Human Behavior* **PMID:** 37365406

Genetics of Educational Attainment (on 3M people)

A PGI for Educational Attainment is also predictive of a wide rang of health outcomes

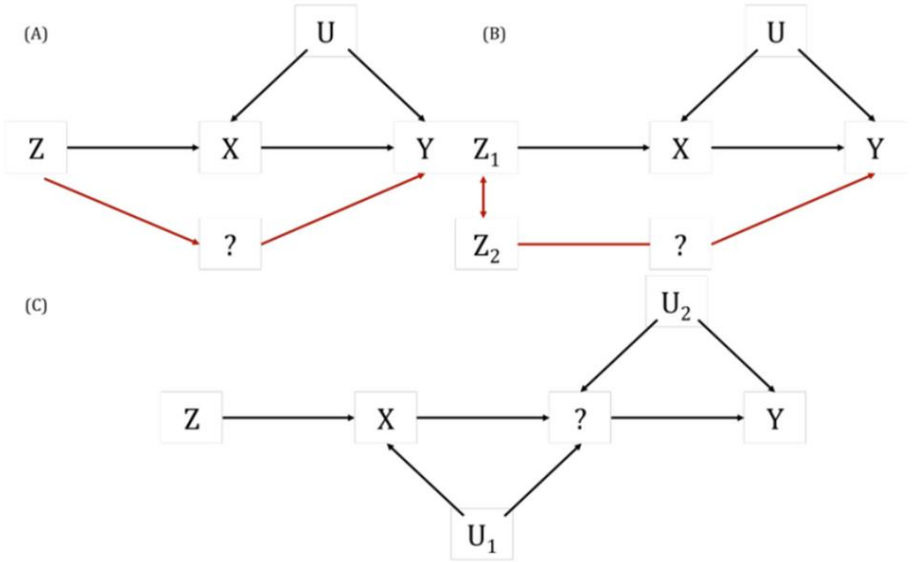


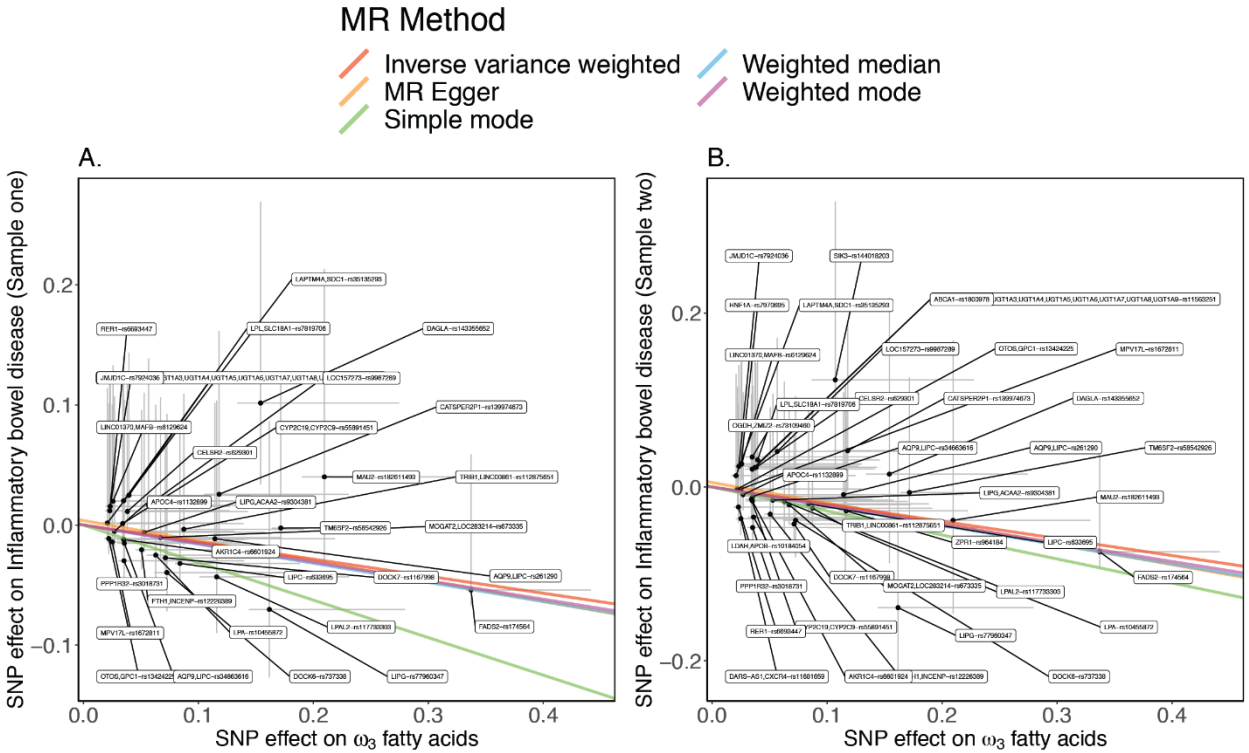
Couples are much more genetically similar for EA (but not height) than expected given their phenotypes



Okbay et al (2022) *Nature Genetics* **54**: 437-449

Mendelian Randomization establishes Causality





Hemani et al (2018) *eLife* **7**: e34408; Astore and Gibson, submitted