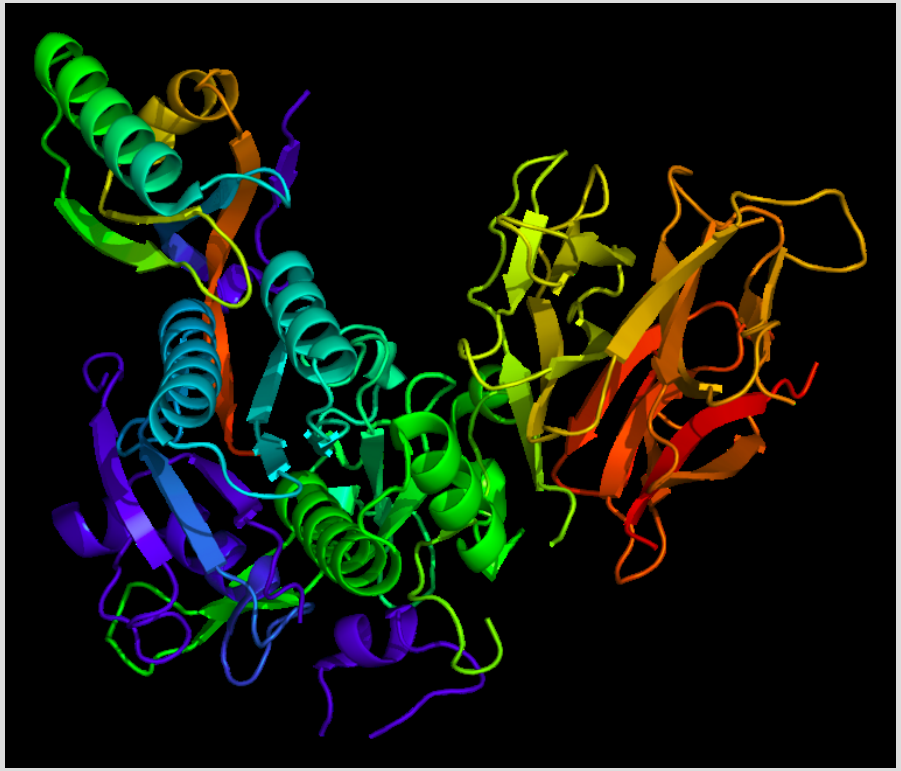
GWAS I

Matti Pirinen
University of Helsinki
28.10.2020

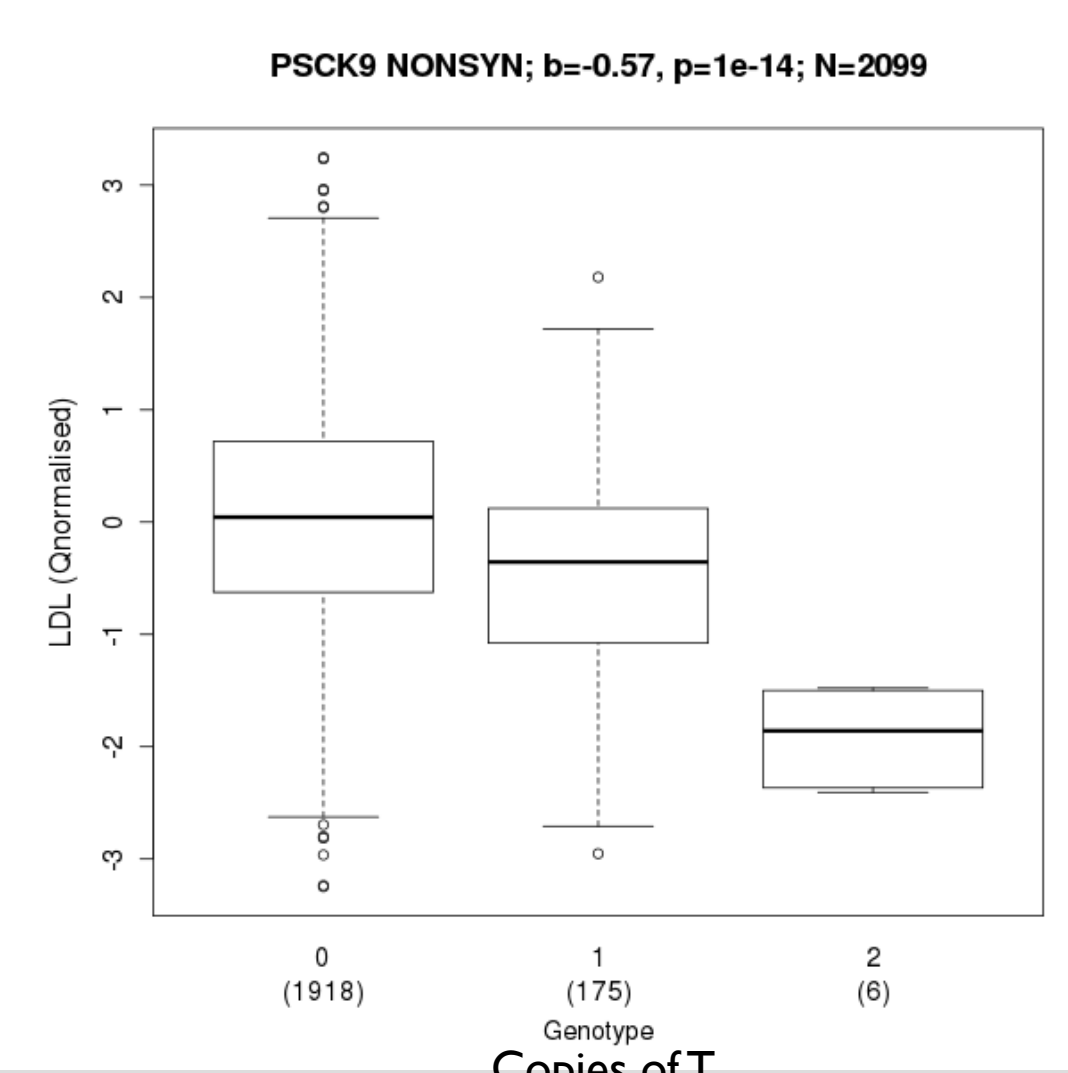
WHY STUDY GENOME?
A STORY ABOUT *PCSK9*

Gene *PCSK9* on chr 1

Codes for protein
692 amino acids long



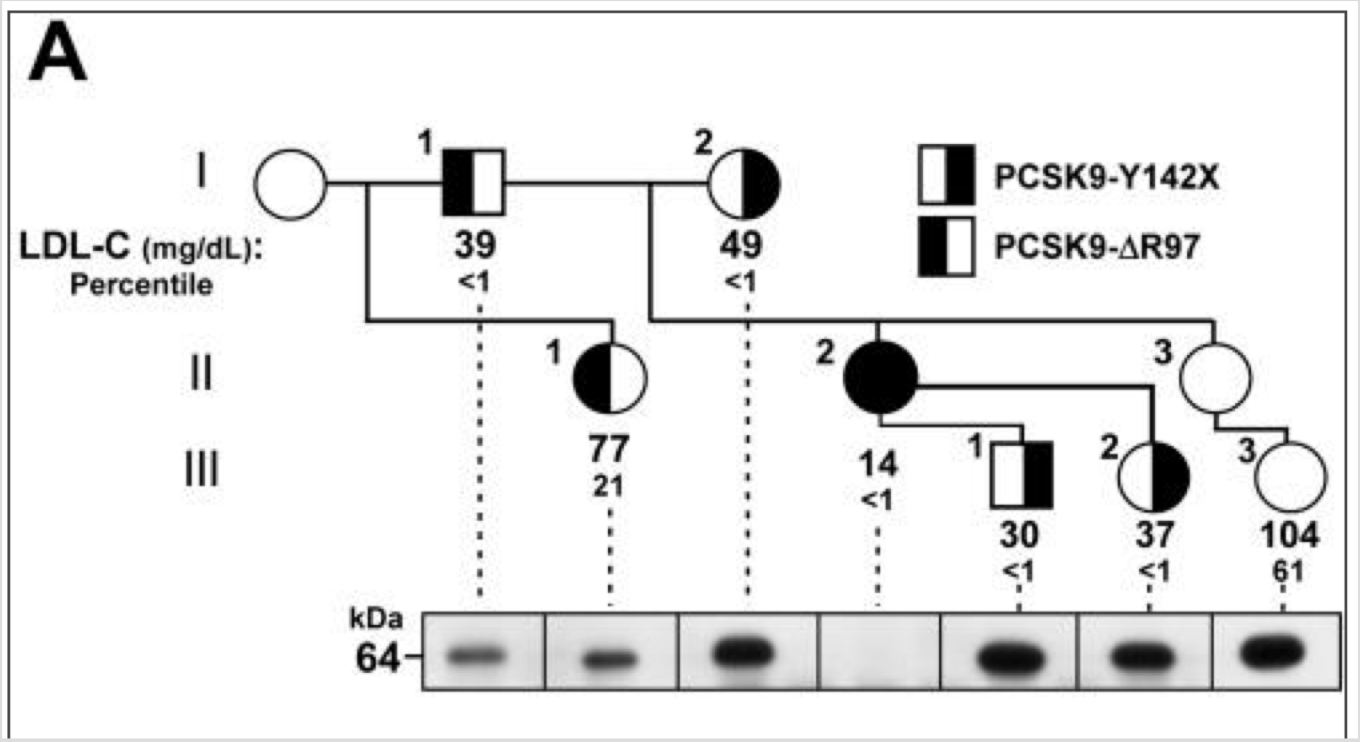
GENETIC VARIANT "RS11591147" IN *PCSK9*



- Carriers of T variant have lower levels of LDL cholesterol than carriers of G variant
- LDL is a strong risk factor for heart disease

2099 Finnish individuals

A HUMAN KNOCK-OUT OF *PCSK9* (2006)

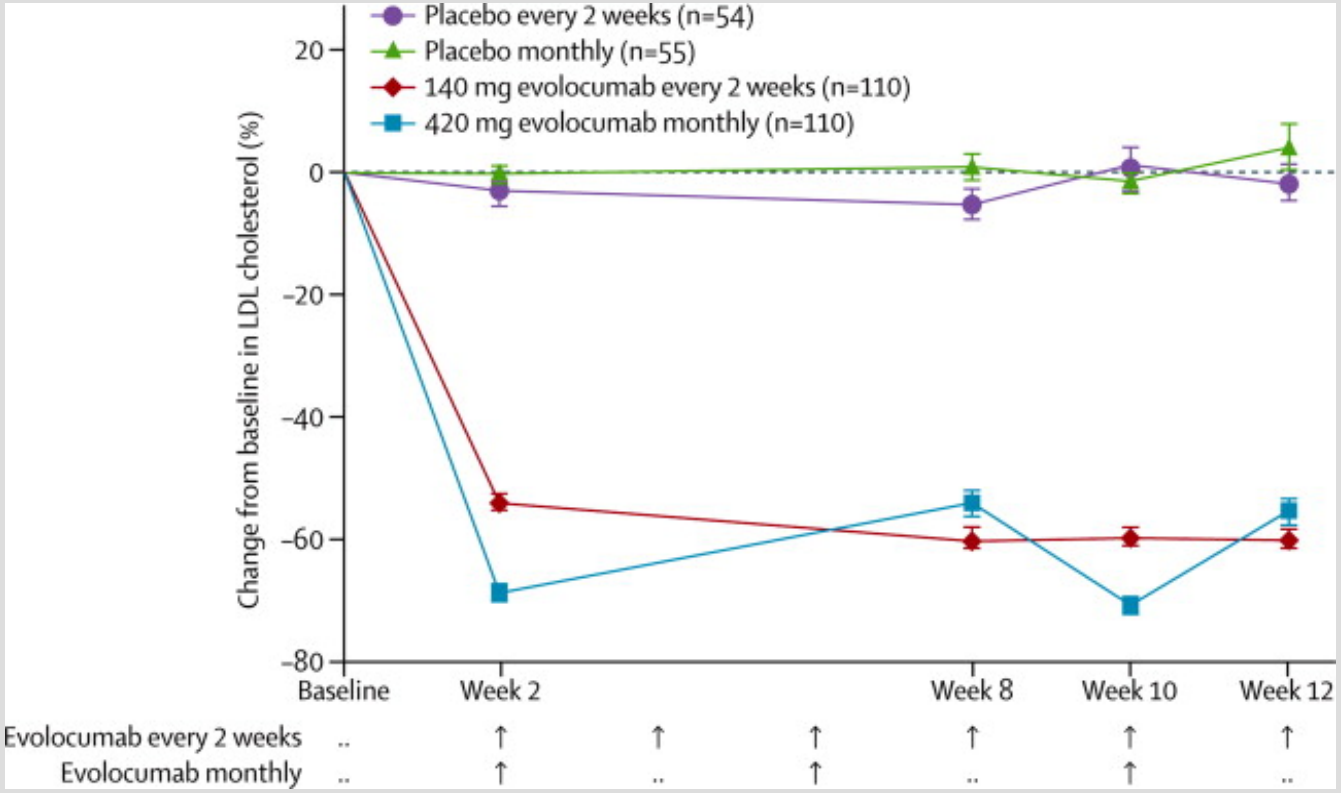


Zhao et al. AJHG 2006

- Individual II.2 has zero working copies of *PCSK9* gene
  - no circulating *PCSK9* and an LDL-C of only 14 mg/dL
  - apparently healthy, fertile, normotensive, college-educated woman with normal liver and renal function tests who works as an aerobics instructor
  - Why is this very interesting observation?
    - Inhibiting *PCSK9* might be a **safe** way to reduce LDL

PCSK9 inhibition with evolocumab (AMG 145) in heterozygous familial hypercholesterolaemia (RUTHERFORD-2): a randomised, double-blind, placebo-controlled trial

Prof Frederick J Raal, PhD, Prof Evan A Stein, PhD, Robert Dufour, MD, Traci Turner, MD, Fernando Civeira, MD, Prof Lesley Burgess, MB, Gisle Langslet, MD, Prof Russell Scott, MD, Prof Anders G Olsson, MD, David Sullivan, MD, G Kees Hovingh, MD, Bertrand Cariou, MD, Ioanna Gouni-Berthold, MD, Ransi Somaratne, MD, Ian Bridges, MSc, Rob Scott, MD, Scott M Wasserman, MD, Prof Daniel Gaudet, MD, for the RUTHERFORD-2 Investigators



Lancet Oct 2014

ORIGINAL ARTICLE

N Engl J Med 2017; 376:1713-1722

Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease

Marc S. Sabatine, M.D., M.P.H., Robert P. Giugliano, M.D., Anthony C. Keech, M.D., Narimon Honarpour, M.D., Ph.D., Stephen D. Wiviott, M.D., Sabina A. Murphy, M.P.H., Julia F. Kuder, M.A., Huei Wang, Ph.D., Thomas Liu, Ph.D., Scott M. Wasserman, M.D., Peter S. Sever, Ph.D., F.R.C.P., and Terje R. Pedersen, M.D. for the FOURIER Steering Committee and Investigators*

FDA Approves Amgen's Repatha (evolocumab) to Prevent Heart Attack and Stroke

Dec 1 2017

In the Repatha cardiovascular outcomes study (FOURIER), Repatha reduced the risk of heart attack by 27 percent, the risk of stroke by 21 percent and the risk of coronary revascularization by 22 percent.

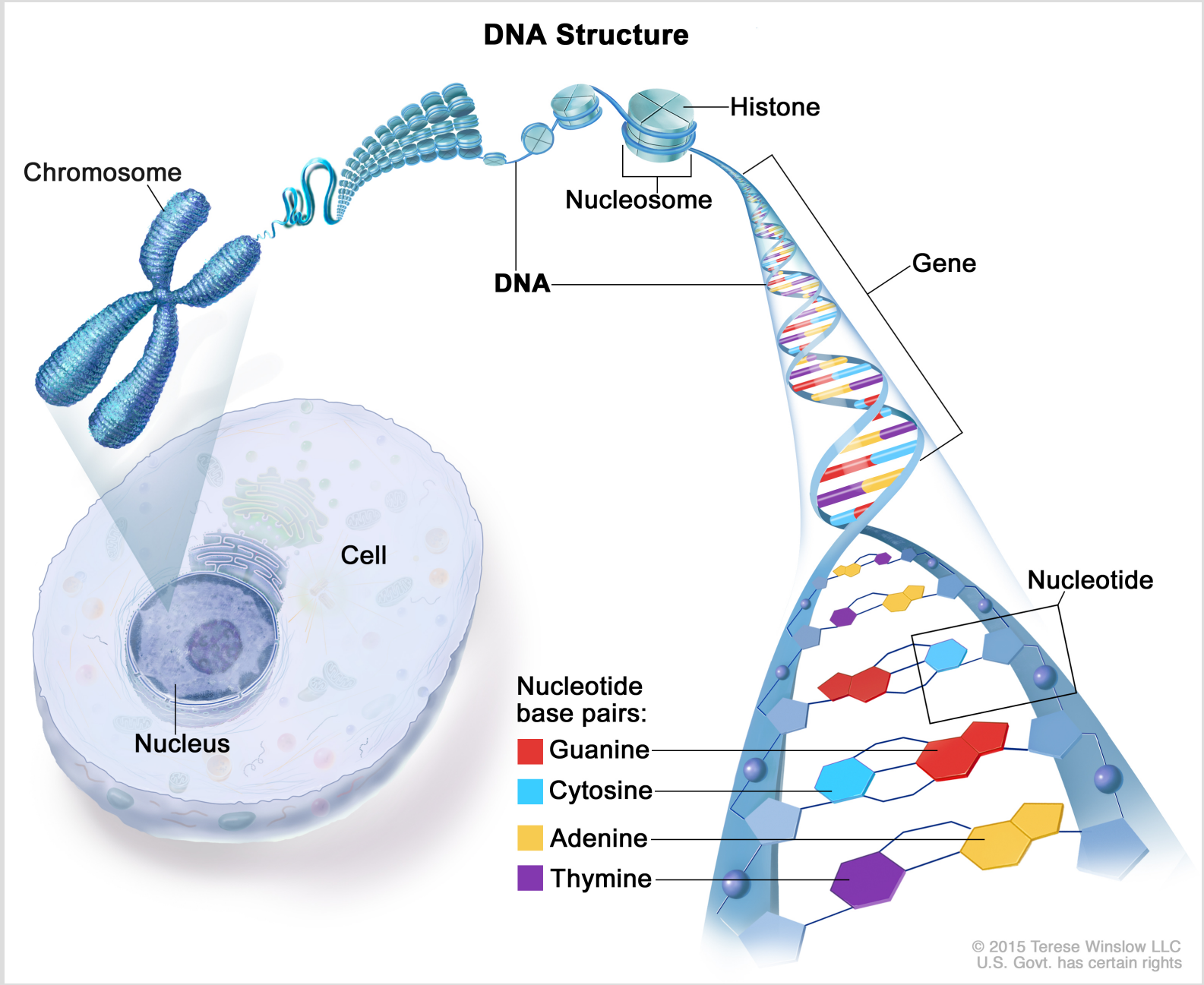
HUMAN GENOME

- Sequence of 3×10^9 letters from alphabet { A, C, G, T }

... G C G T T T A C G ...

You have two genomes: maternal and paternal.

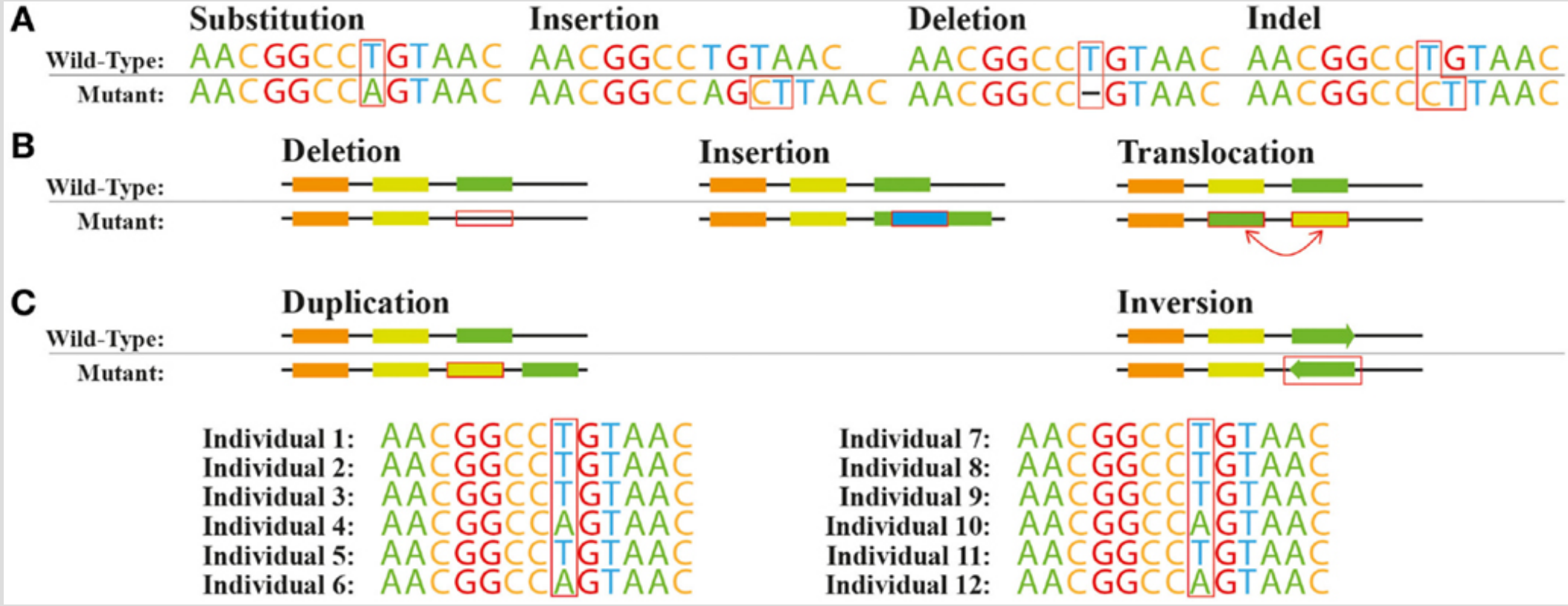
Your genomes are physically divided into 22 pairs of autosomal chromosomes and
1 pair of sex chromosomes (males XY, females: XX)



Most DNA is found inside the nucleus of a cell, where it forms the chromosomes. Chromosomes have proteins called histones that bind to DNA. DNA has two strands that twist into the shape of a spiral ladder called a helix. DNA is made up of four building blocks called nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). The nucleotides attach to each other (A with T, and G with C) to form chemical bonds called base pairs, which connect the two DNA strands. Genes are short pieces of DNA that carry information for creating proteins.

https://siteman.wustl.edu/glossary/cdr0000046470/

TYPES OF VARIATION



Nucleotide level variation

Structural variation

SNP A/T in population

Cardoso et al. 2015
Front. Bioeng. Biotechnol., 16 February 2015 | https://doi.org/10.3389/fbioe.2015.00013

SINGLE NUCLEOTIDE POLYMORPHISM (SNP)

On average, 1:300 positions in genome has common (MAF>1%) variation in population; these are called "SNPs"

Genomes in population

... G C G T T ... 96%

... G C T T T ... 4%

This is a SNP, with alleles: G / T, minor allele frequency (MAF) = 4%

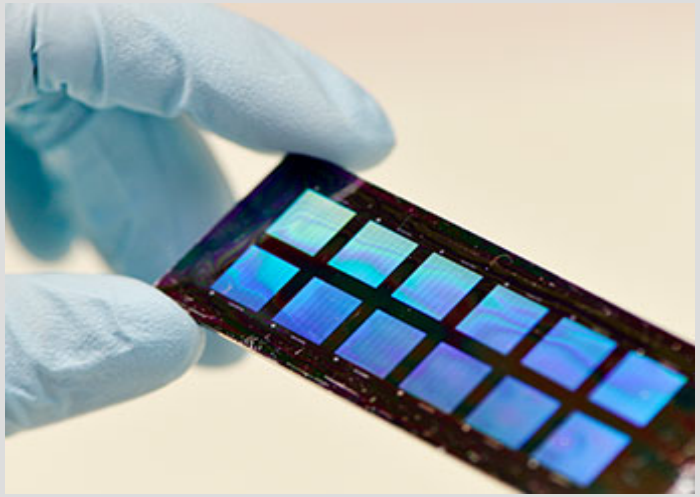
Only forward strand of genomes is shown here

Genotypes at this SNP in population

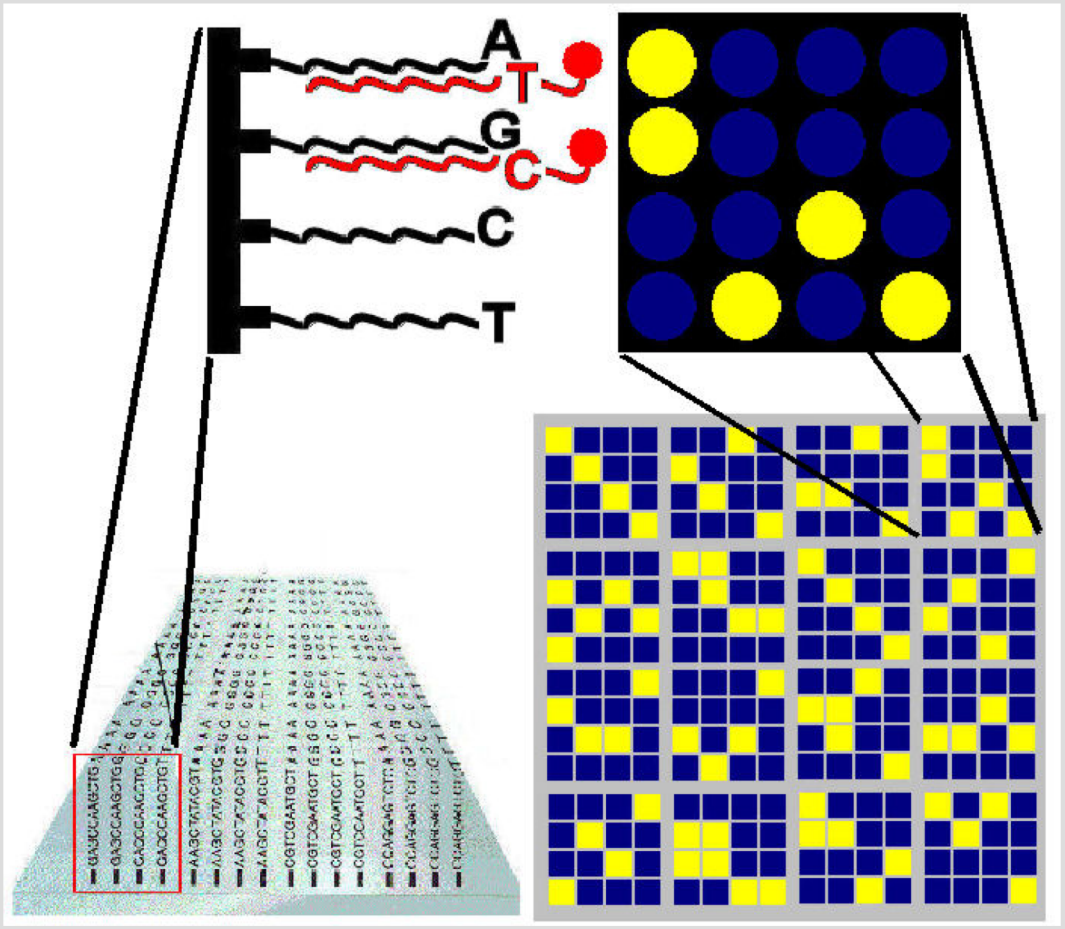
0: GG ~ 92.1%
1: GT ~ 7.7 %
2: TT ~ 0.2 %

READING SNPS

- Human SNP array can measure 10^6 SNPs
- Cost per individual ~30 euros

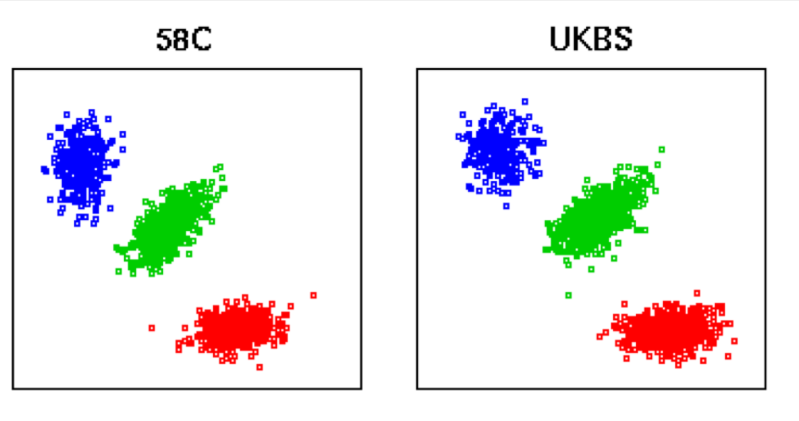


This array can genotype 12 individuals at 10^6 SNPs

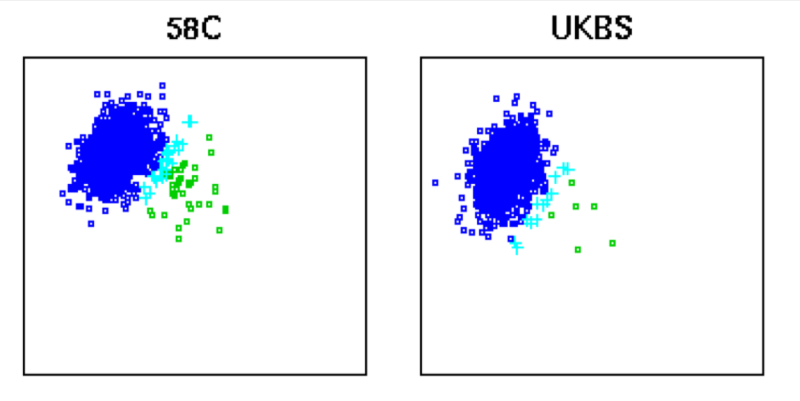


Steven M. Carr
www.mun.ca/biology/scarr/DNA_Chips.html

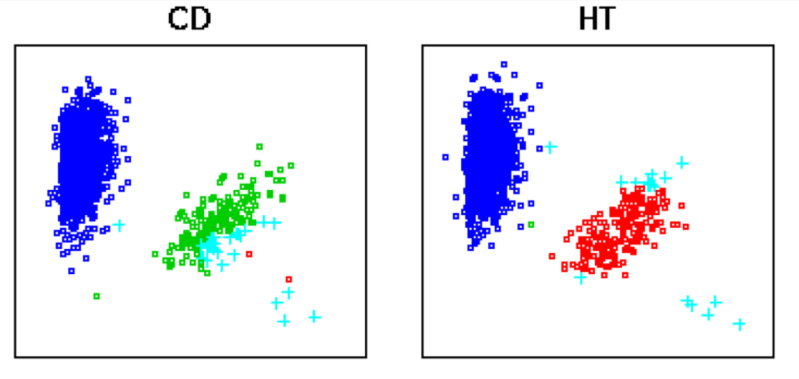
GENOTYPE CALLING FROM SNP ARRAY DATA



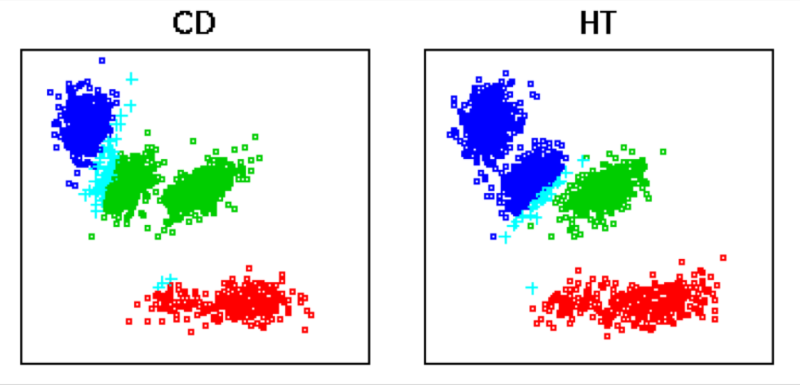
GOOD calling!



ERROR, rare variant has less than 3 clusters



ERROR, clustering algorithm performs differently in two cohorts



ERROR, structural variant has more than three genotypes

The calling algorithm tries to find the three genotype clusters.

Figures shows how an algorithm has clustered individuals into three groups

Light blue means algorithm has made no call.

Bottom line errors would likely fail HWE test.

From D.Phil thesis of Damjan Vukcevic, Oxford, 2009

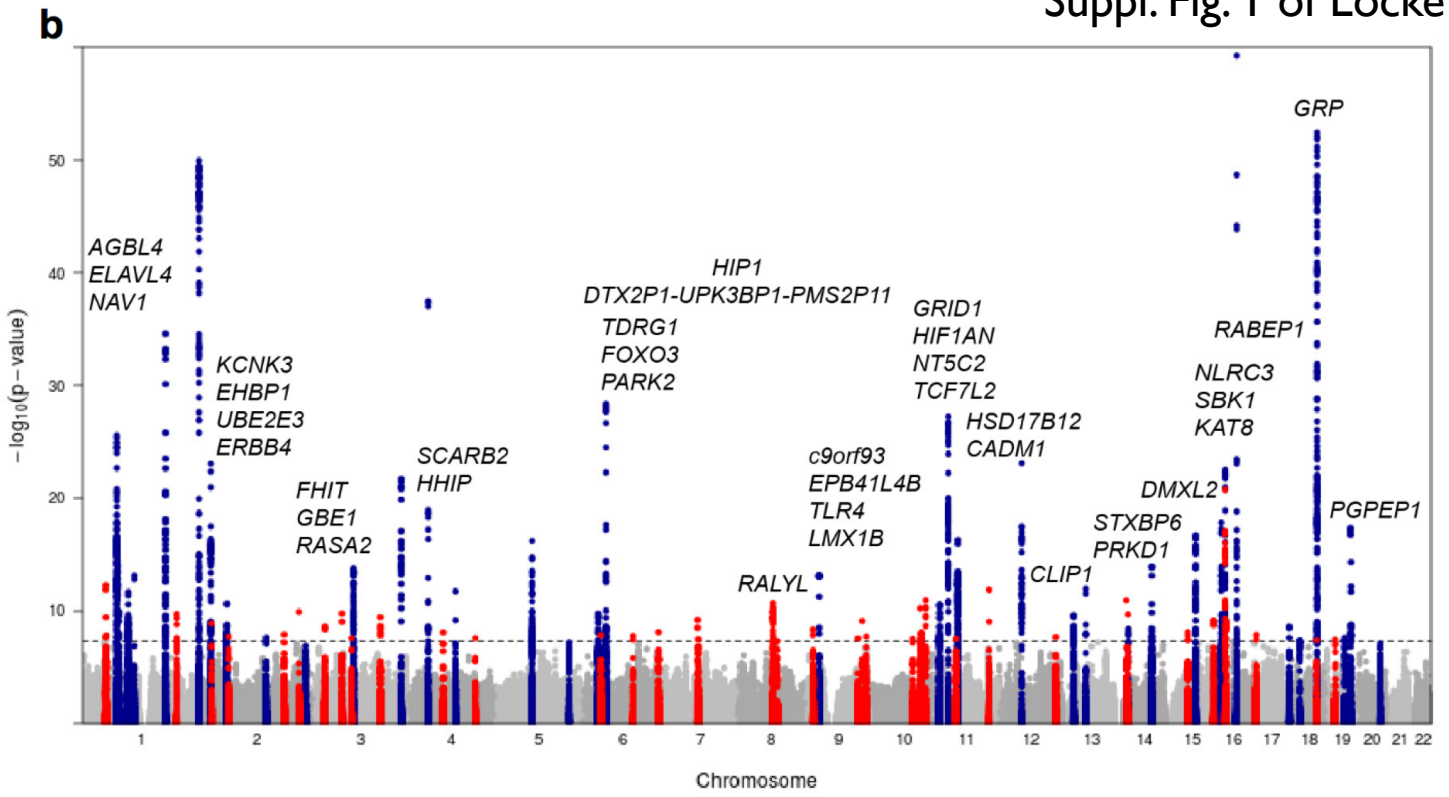
EXAMPLE GWAS

- Let's next look at two examples GWAS
- Body-mass index GWAS by Locke et al. (Nature 2015) as an example of a quantitative trait analysis
- Migraine GWAS by Gormley et al. (Nature Genetics 2016) as an example of case-control analysis.

GWAS ON BODY MASS INDEX (BMI) (LOCKE ET AL. 2015, NATURE)

- 339,000 individuals with genotypes and BMI available
- 125 cohorts around the world participated
- 97 loci (regions in the genome) convincingly associated
- Each locus is a hint to biology of BMI
- Results highlight role of central nervous system in BMI

Suppl. Fig. 1 of Locke et al.

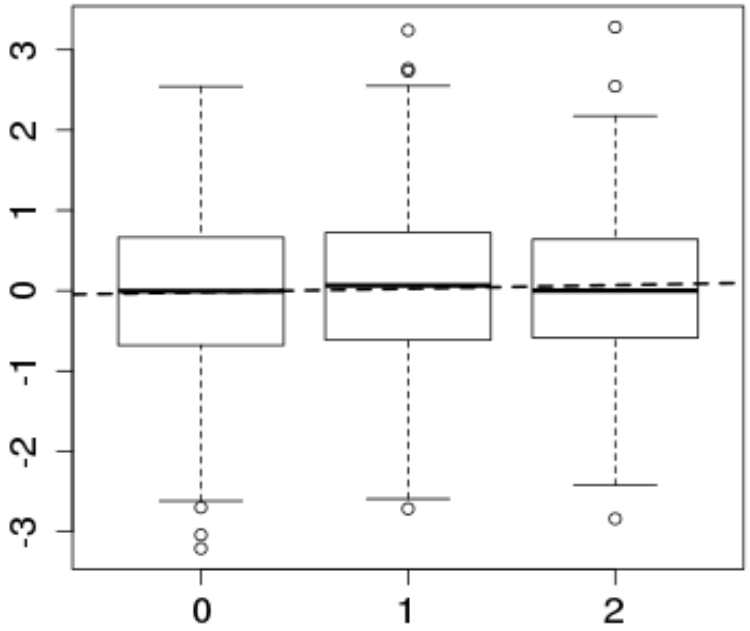


Manhattan plot shows –log10 P-value of each SNP tested in GWAS.
Genome-wide significance level at P=5e-8 or –log10(P) = 7.3.
Previously known loci are in blue, new findings are in red.
Each locus in named by a nearby gene (but that gene is not necessarily causal.)

LOCKE ET AL. DID ASSOCIATION TESTS AT 2.5 MILLION SNPS

SNP with no association

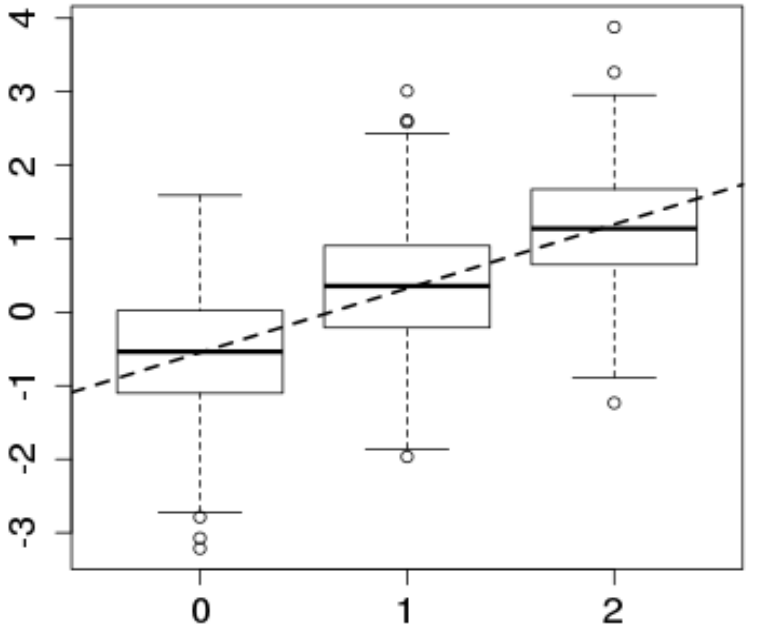
$\hat{\beta}$=0.04424



Copies of effect allele

$\hat{\beta}$=0.8724

SNP with a clear association



Copies of effect allele

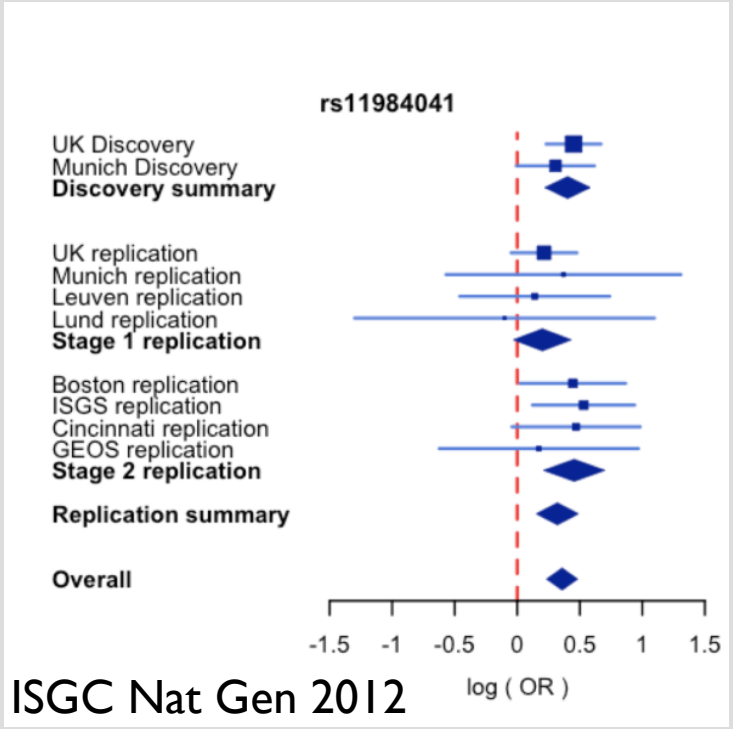
Association test: "Does the mean BMI differ between genotype groups?"
(output are linear regression slope $\hat{\beta}$, its standard error SE and P-value)

- ”339,000 individuals with genotypes and BMI available”
- ”125 studies (”cohorts”) around the world participated”

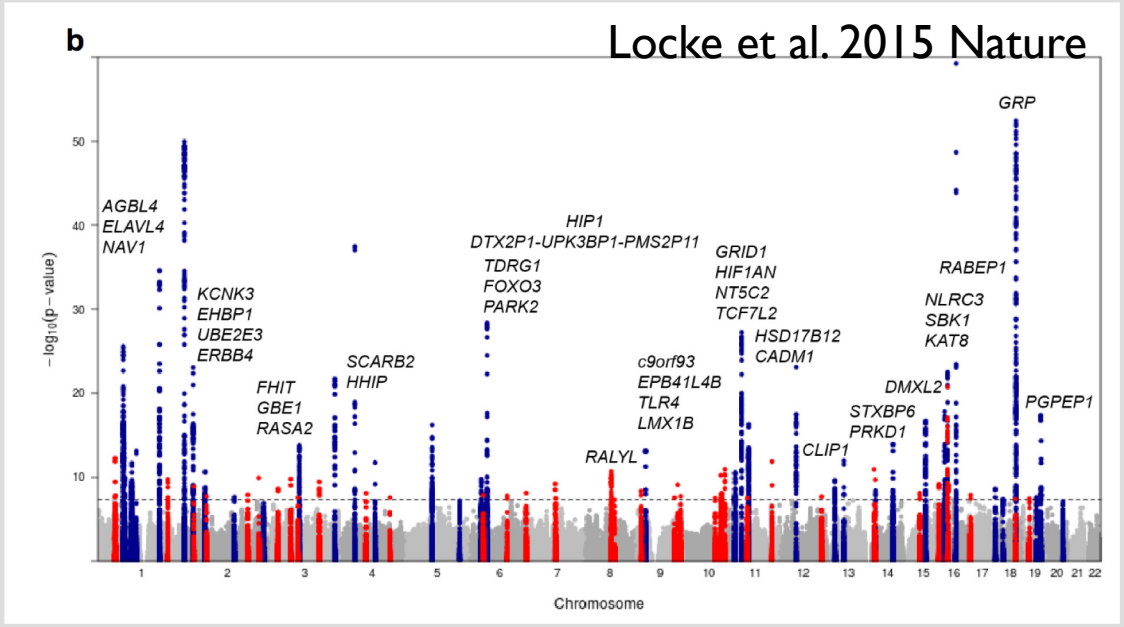
This means that **meta-analysis** is done across the studies.

A **meta-analysis** is a statistical analysis that combines the results of multiple scientific studies on the same question.

Here it works on GWAS results, not requiring original genotype-phenotype data.



ISGC Nat Gen 2012



SNP	A1	A2	Freq1.Hapmap	b	se	p	N
rs1000000	G	A	0.6333	1e-04	0.0044	0.9819	231410
rs10000010	T	C	0.575	-0.0029	0.003	0.3374	322079
rs10000012	G	C	0.1917	-0.0095	0.0054	0.07853	233933
rs10000013	A	C	0.8333	-0.0095	0.0044	0.03084	233886
rs10000017	C	T	0.7667	-0.0034	0.0046	0.4598	233146
rs10000023	G	T	0.4083	0.0024	0.0038	0.5277	233860

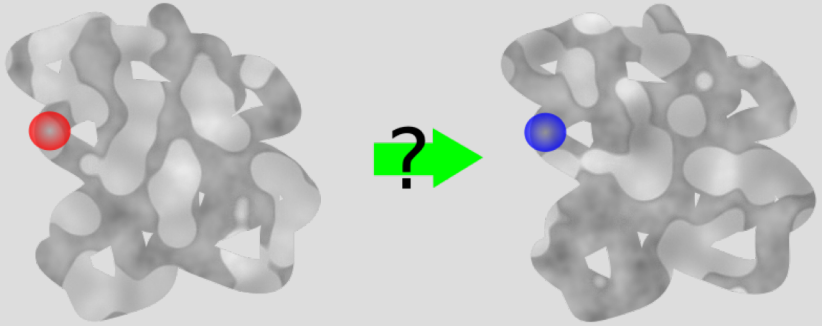
While no-one has access to all original genotype-phenotype data, everyone can access the meta-analyzed GWAS results as they are (often) publicly available.

For this BMI analysis, results are here
https://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

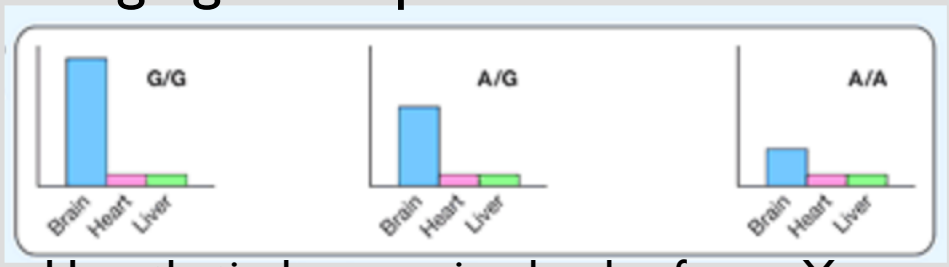
- ”97 loci (regions in the genome) convincingly associated”
- ”Each locus is a hint to biology of BMI”

What does each variant do?

Change protein? (only few GWAS hits do)

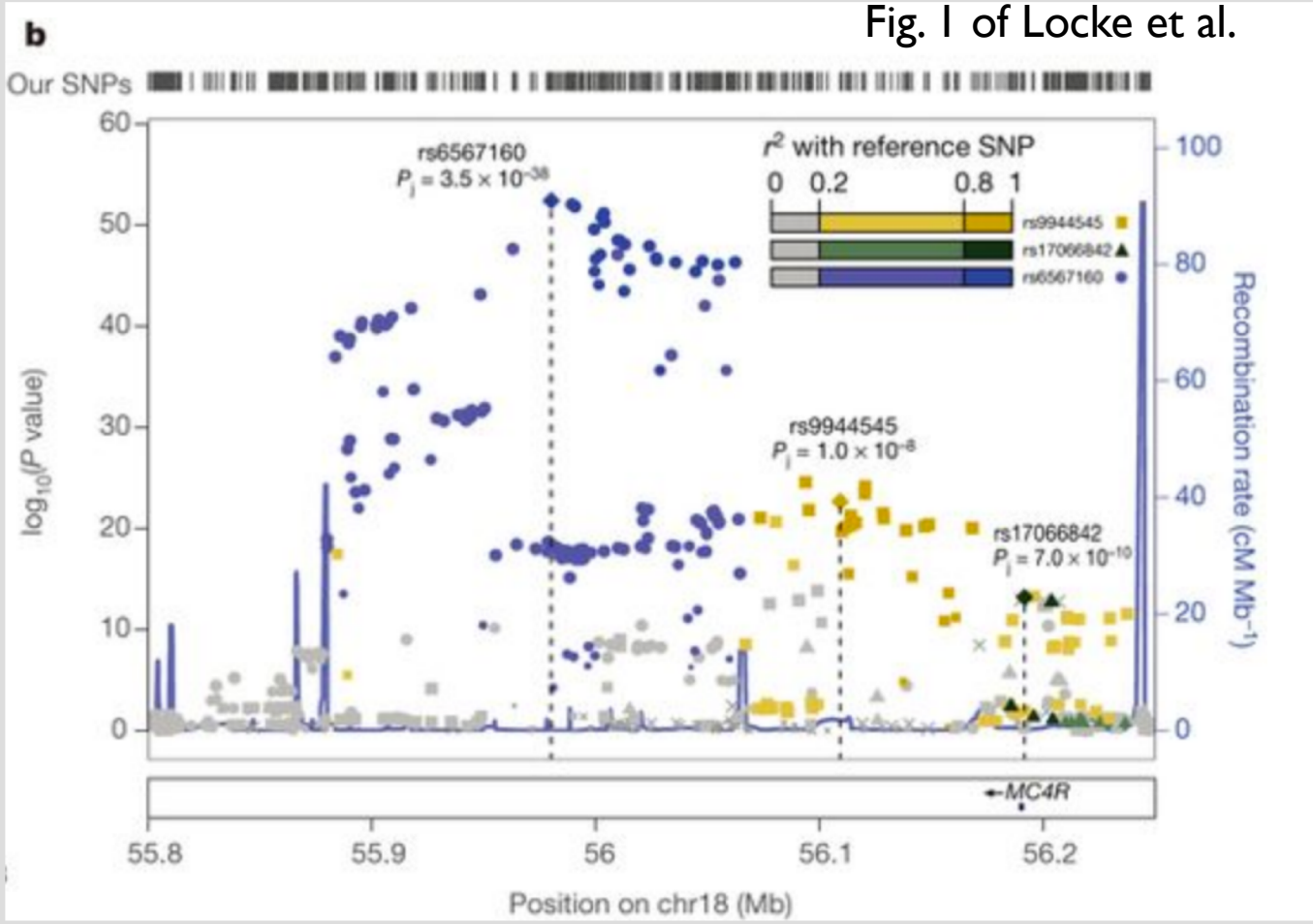


Change gene expression in certain context?



Hypothetical expression levels of gene X

Fig. 1 of Locke et al.



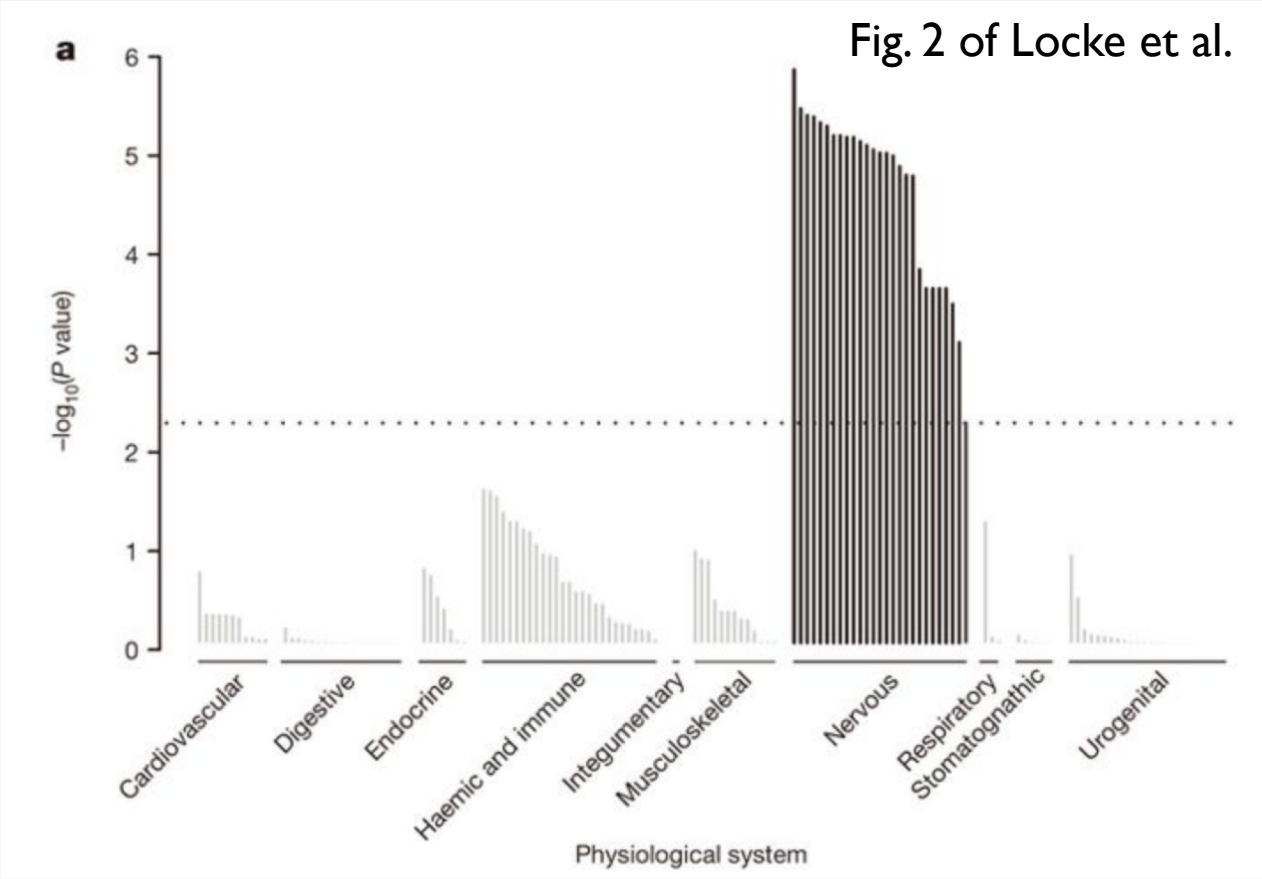
Zooming into one associated region (MC4R) on chr 19.
Many SNPs show strong association; not clear which are causal ones.
Three SNPs are highlighted as possible independent signals.

- ”Results highlight role of central nervous system in BMI”

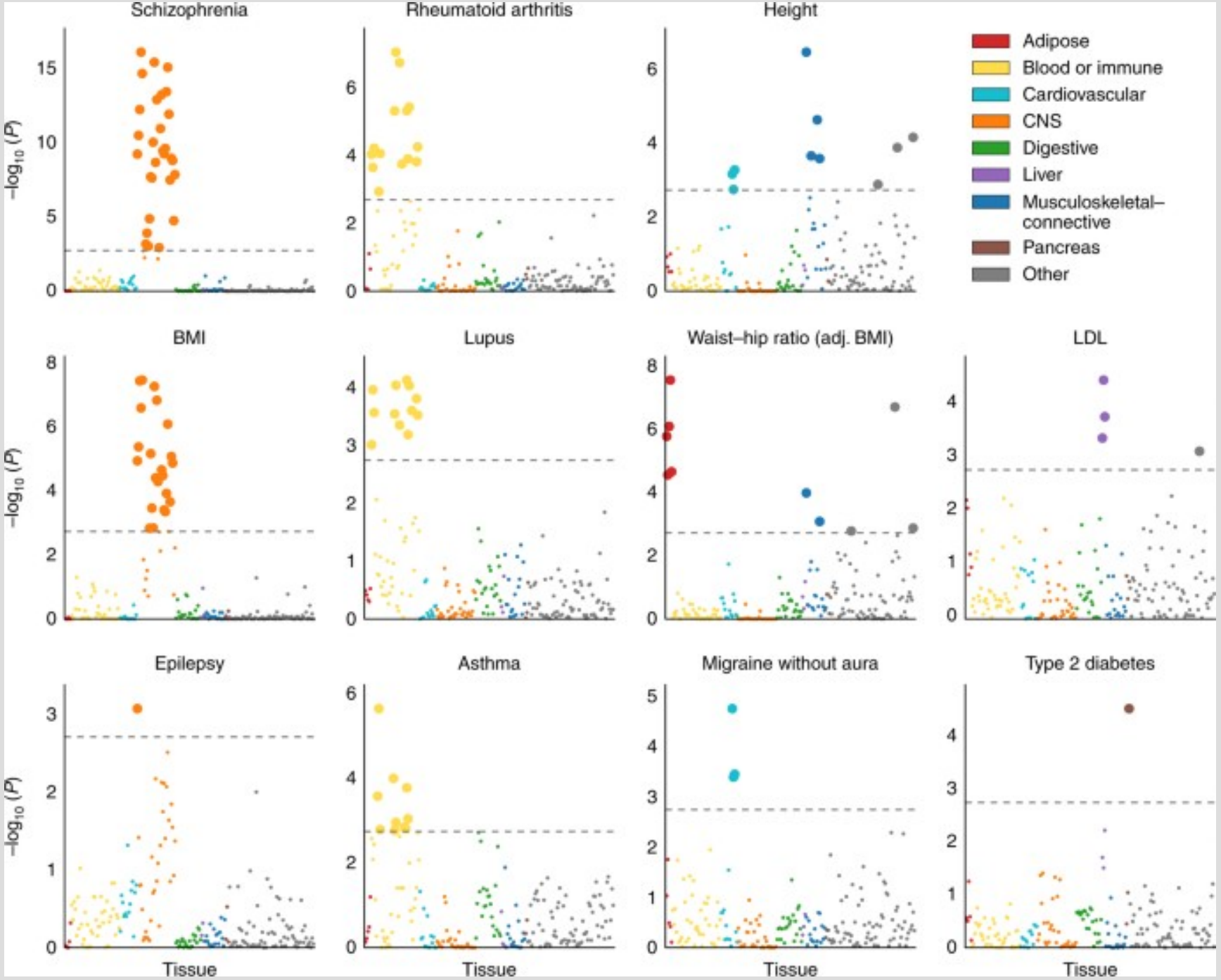
Combining signals across the genome:
Does the significantly associated variation tend to be near certain types of

- Genes?
- Or their regulatory regions?
- Or are there other patterns?

Fig. 2 of Locke et al.



DEPICT predicts genes within BMI-associated loci ($P < 5 \times 10^{-4}$) are enriched for expression in the brain and central nervous system. Tissues are sorted by physiological system and significantly enriched tissues are in black; the dotted line represents statistically significant enrichment.



Are GWAS signals enriched in/near genes specifically expressed in certain tissue/cell type(s).

-log10 P-value is of the association between the trait in the title and the tissue/cell type listed in the legend.

Finucane et al. 2018 Nat Gen

GWAS ON MIGRAINE (GORMLEY ET AL. 2016) (1/3)

- 60,000 cases and 315,000 controls from 22 studies
- 38 loci with convincing association
- Highlights vascular system

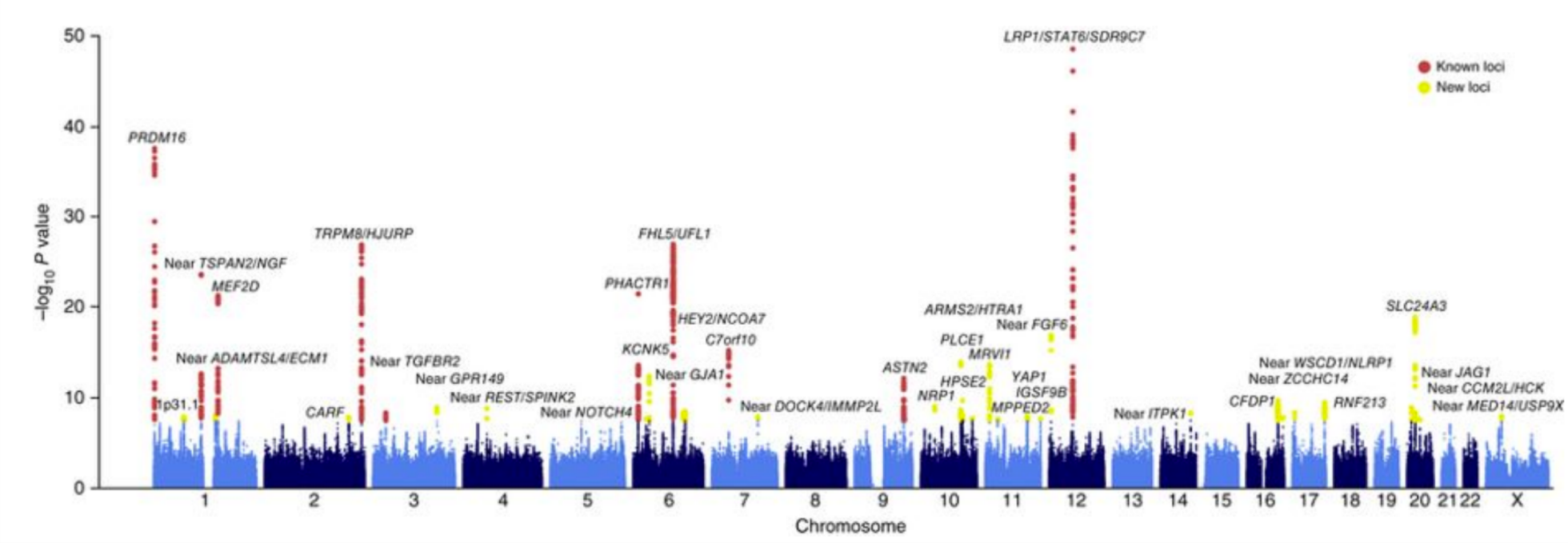


Fig. 1 of Gormley et al.
Manhattan plot of results.

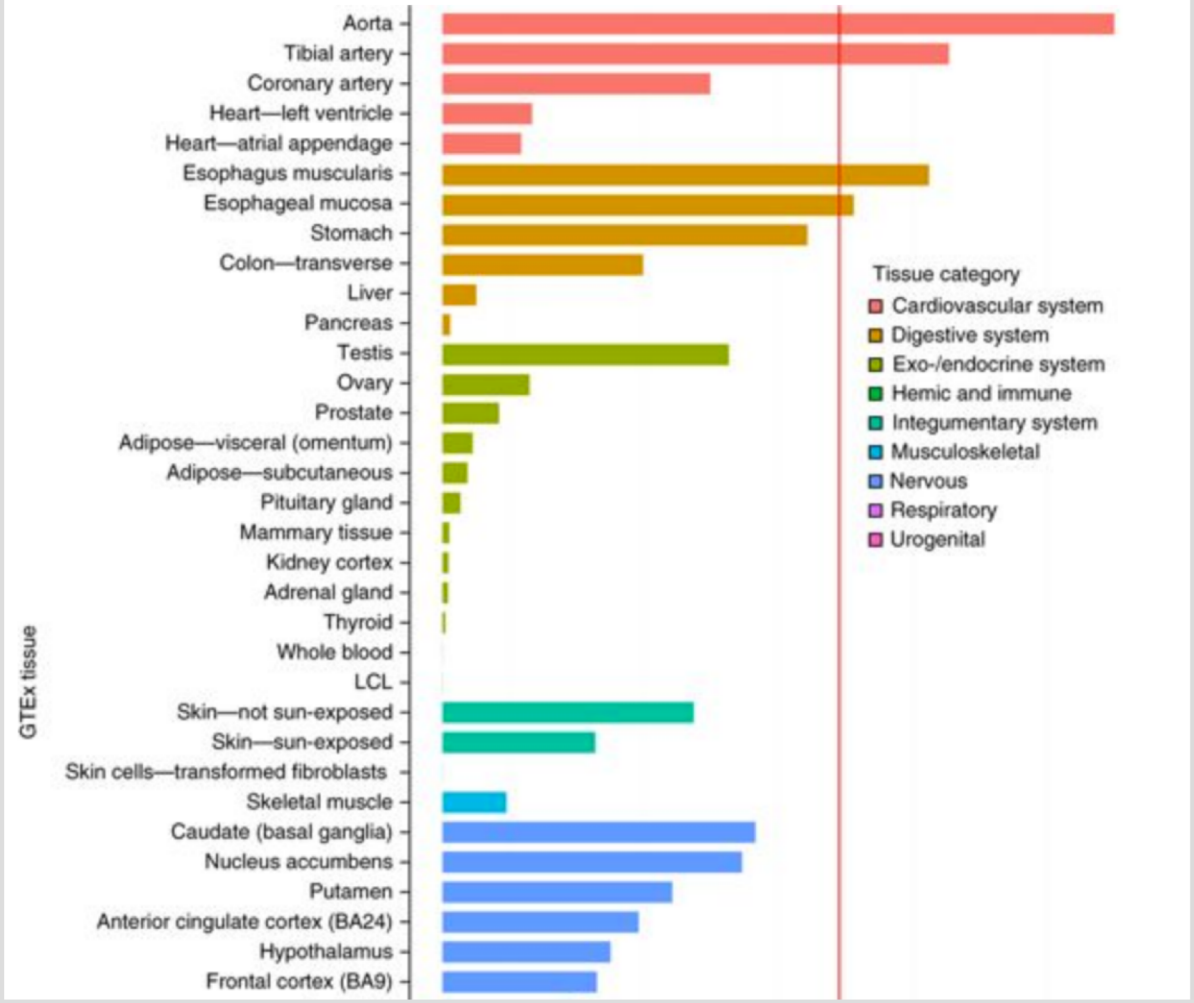
GWAS ON MIGRAINE (GORMLEY ET AL. 2016) (2/3)

Locus rank	Locus	Chr	Index SNP	Minor allele	MAF	All migraine		Secondary signal		Migraine without aura		Previous publication PMID
						OR (95% CI)	*P*	Index SNP	*P*	Index SNP	*P*	
1	*LRP1–STAT6–SDR9C7*	12	rs11172113	C	0.42	0.90 (0.89–0.91)	5.6×10^{-49}	rs11172055	1.3×10^{-9}	rs11172113	4.3×10^{-16}	21666692
2	*PRDM16*	1	rs10218452	G	0.22	1.11 (1.10–1.13)	5.3×10^{-38}	rs12135062	3.7×10^{-10}	-	-	21666692
3	*FHL5–UFL1*	6	rs67338227	T	0.23	1.09 (1.08–1.11)	2.0×10^{-27}	rs4839827	5.7×10^{-10}	rs7775721	1.1×10^{-12}	23793025
4	Near *TSPAN2–NGF*	1	rs2078371	C	0.12	1.11 (1.09–1.13)	4.1×10^{-24}	rs7544256	8.7×10^{-9}	rs2078371	7.4×10^{-9}	23793025
5	*TRPM8–HJURP*	2	rs10166942	C	0.20	0.94 (0.89–0.99)	1.0×10^{-23}	rs566529	2.5×10^{-9}	rs6724624	1.1×10^{-9}	21666692
6	*PHACTR1*	6	rs9349379	G	0.41	0.93 (0.92–0.95)	5.8×10^{-22}	-	-	rs9349379	2.1×10^{-9}	22683712
7	*MEF2D*	1	rs1925950	G	0.35	1.07 (1.06–1.09)	9.1×10^{-22}	-	-	-	-	22683712
8	*SLC24A3*	20	rs4814864	C	0.26	1.07 (1.06–1.09)	2.2×10^{-19}	-	-	-	-	-

Table 1 shows summary statistics for each locus.
Index SNP is the one with lowest P-value and 2nd signal is conditionally independent signal in the same locus.

GWAS ON MIGRAINE (GORMLEY ET AL. 2016) (3/3)

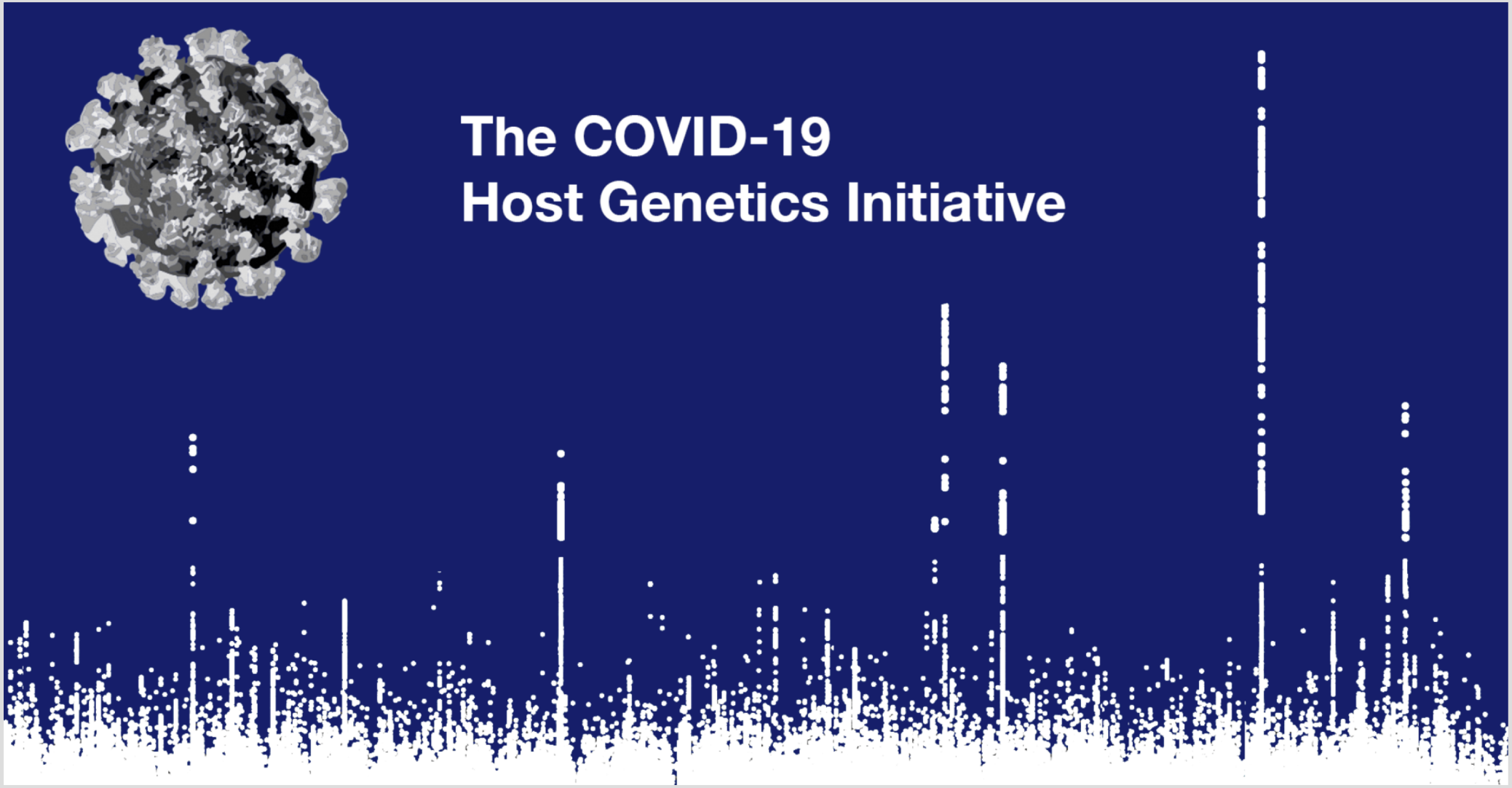
Figure 2: Expression enrichment of genes from the migraine loci in GTEx tissue samples.



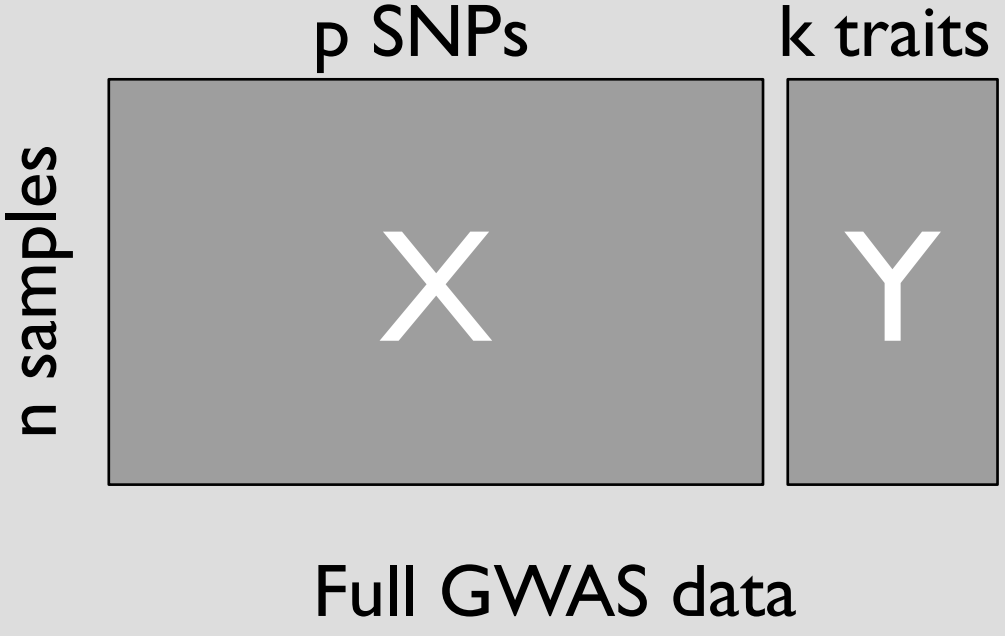
We @FIMM, UH
are currently expanding
this migraine GWAS to over
100,000 cases and 750,000
controls and will report over
nearly 100 new migraine loci.

GWAS on COVID-19 severity and susceptibility is being coordinated from FIMM, UHelsinki

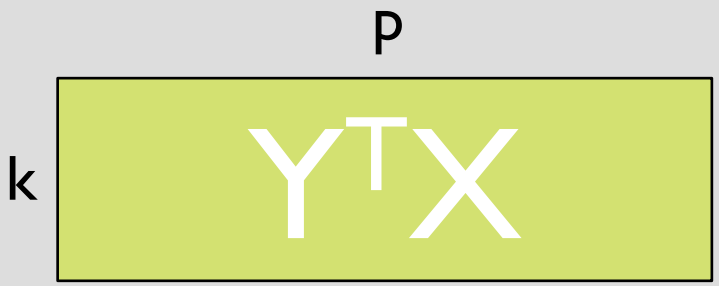
https://www.covid19hg.org/



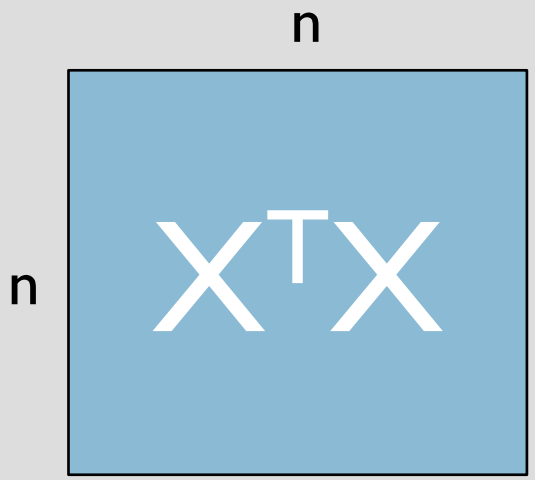
GWAS IN MATRICES



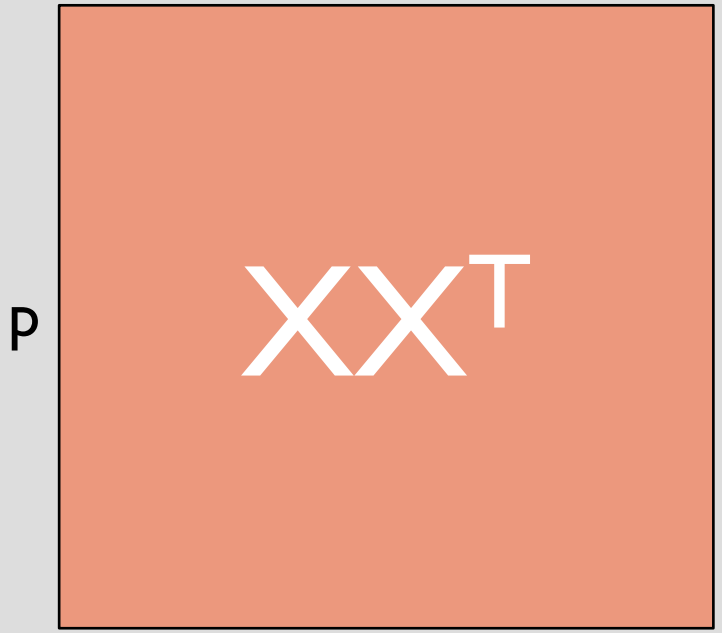
Full GWAS data



GWAS results (betas, SE, P-value)

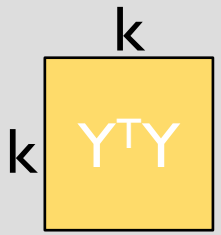


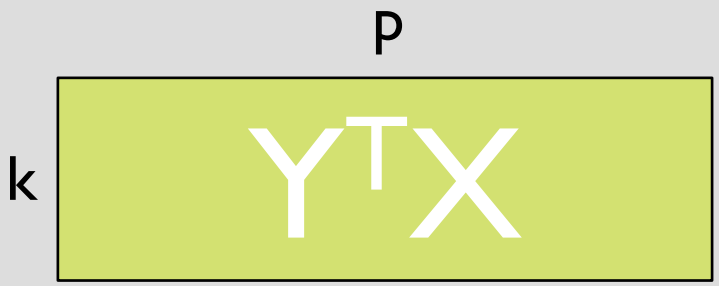
sample relatedness



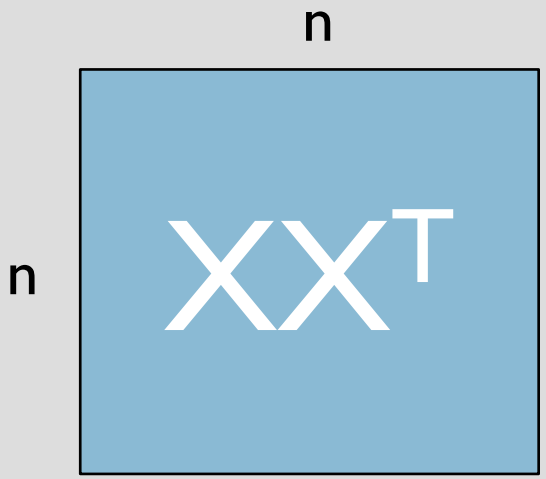
LD matrix (correlation of variants)

trait correlation

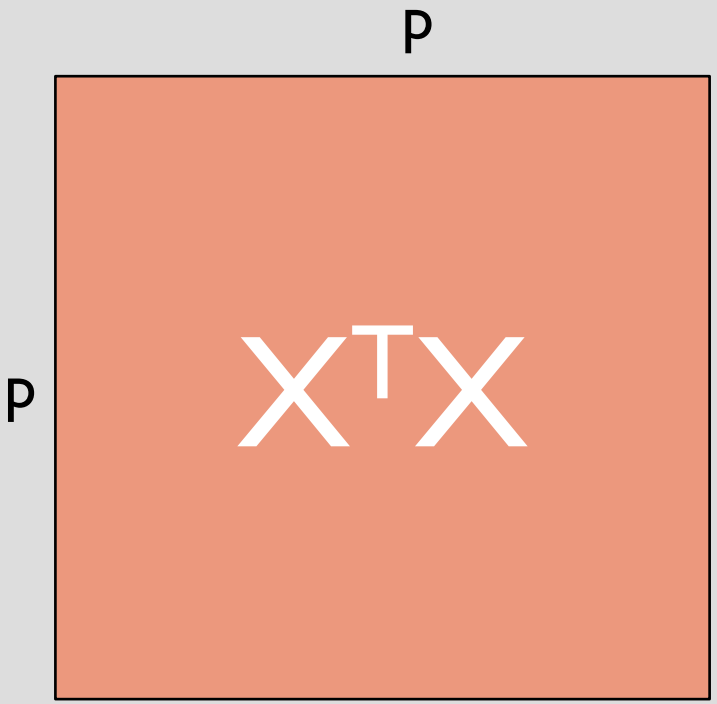




GWAS results (beta, SE, P-value)



sample relatedness



LD matrix (correlation of variants)

Weeks 1-7:

statistical inference,
statistical power,
confounders,
covariates,
summary statistics,
meta-analysis

Weeks 3, 5:

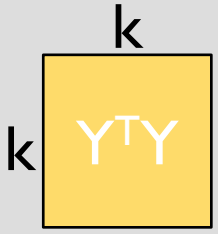
Relatedness & population structure
Heritability & mixed models

Weeks 4,5:

Haplotypes & linkage disequilibrium
Imputation & fine-mapping
LD-score regression

Week 6:

Multi-trait GWAS
Genetic correlation
Mendelian Randomisation



trait correlation