GWAS I

Matti Pirinen

University of Helsinki

GENOME-PHENOME ASSOCIATION

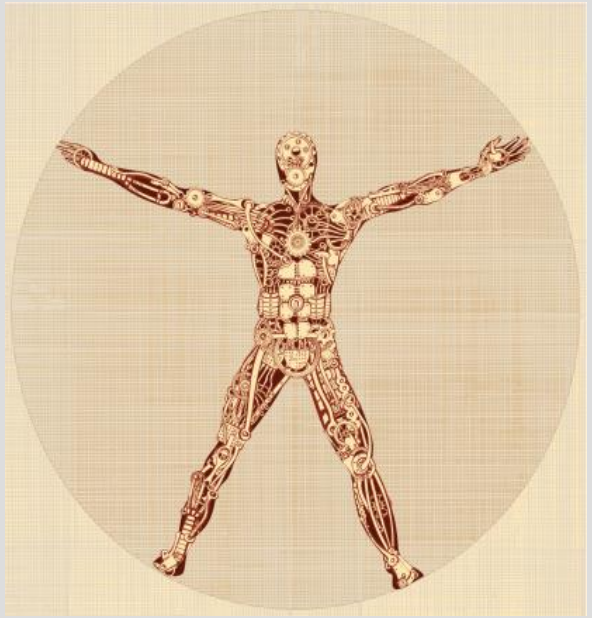
Genome

Phenome = all **phenotypes** combined

Variation in the genome between individuals stays (nearly) constant through lifetime!

"genome-wide" studies consider variation in millions of positions across the genome

association ?



Measurable traits/ Quantitative traits (blood pressure)

Disease traits (MS-disease, diabetes)

Behavioral traits (chronotype, smoking)

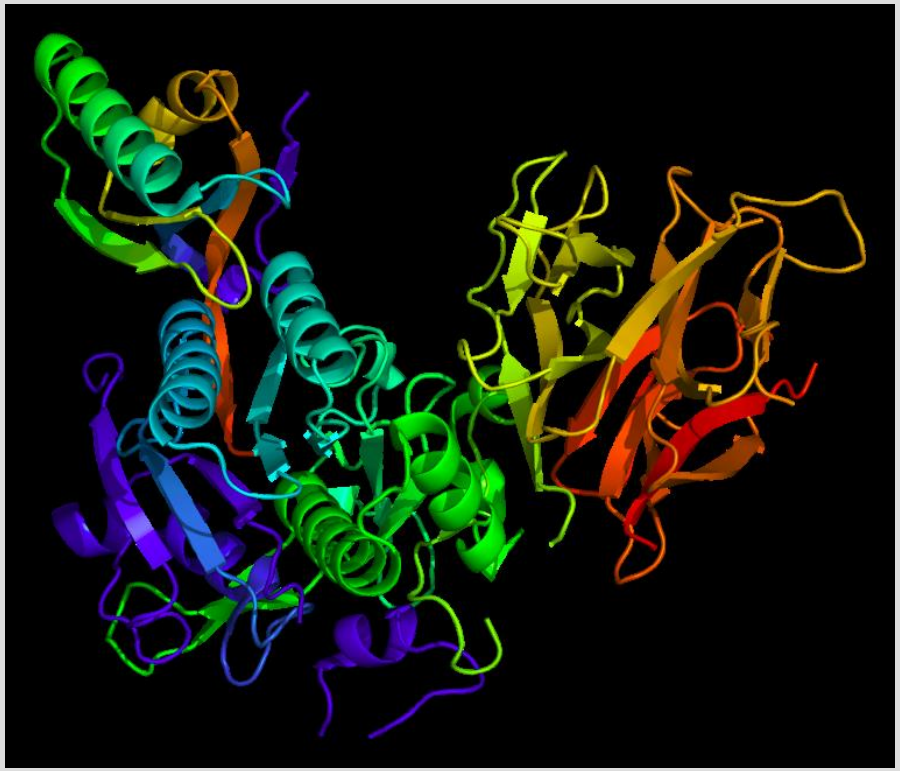
Statistical association can

- allow predicting one from the other
- suggest causal links between the two

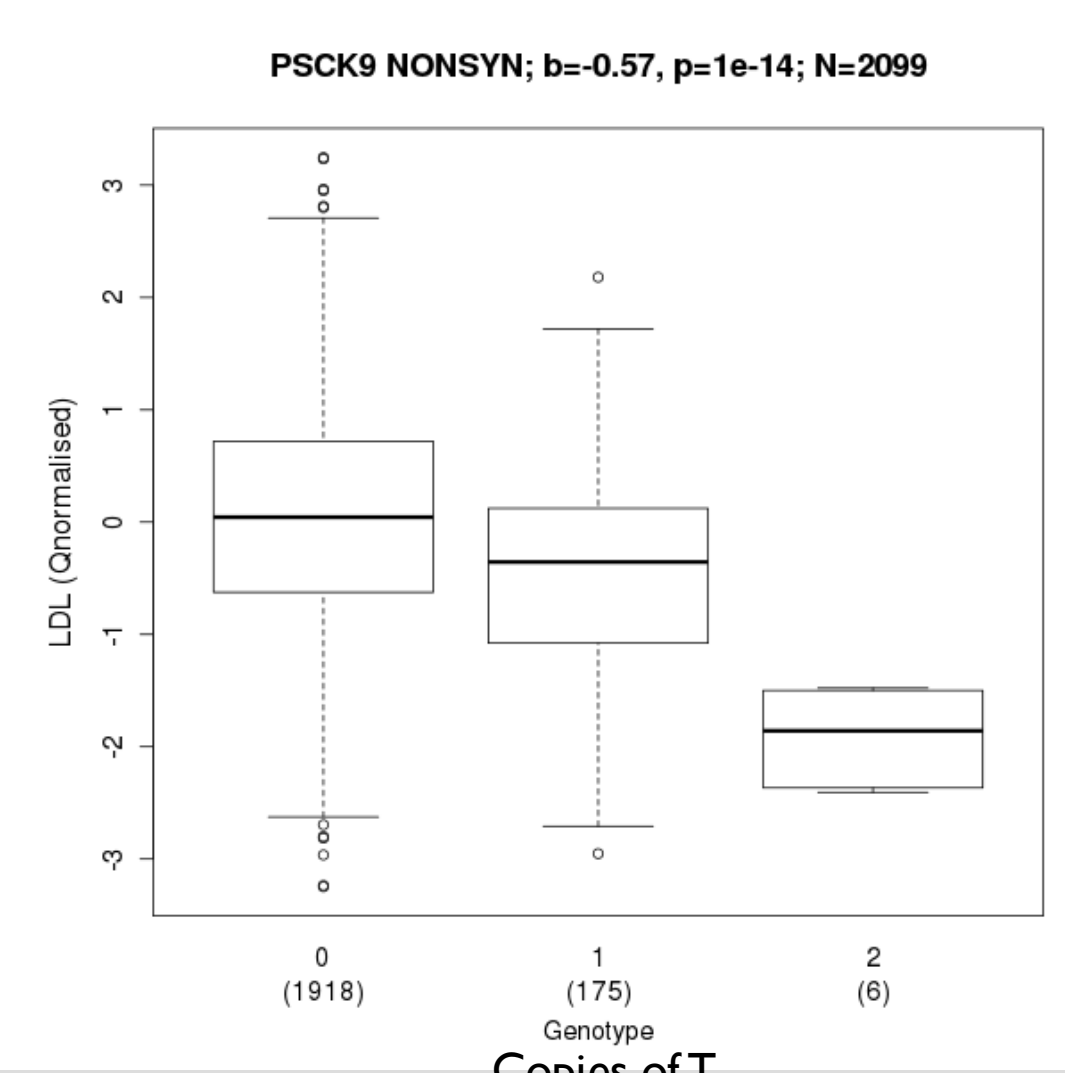
WHY STUDY GENOME?
A STORY ABOUT *PCSK9*

Gene *PCSK9* on chr 1

Codes for protein
692 amino acids long



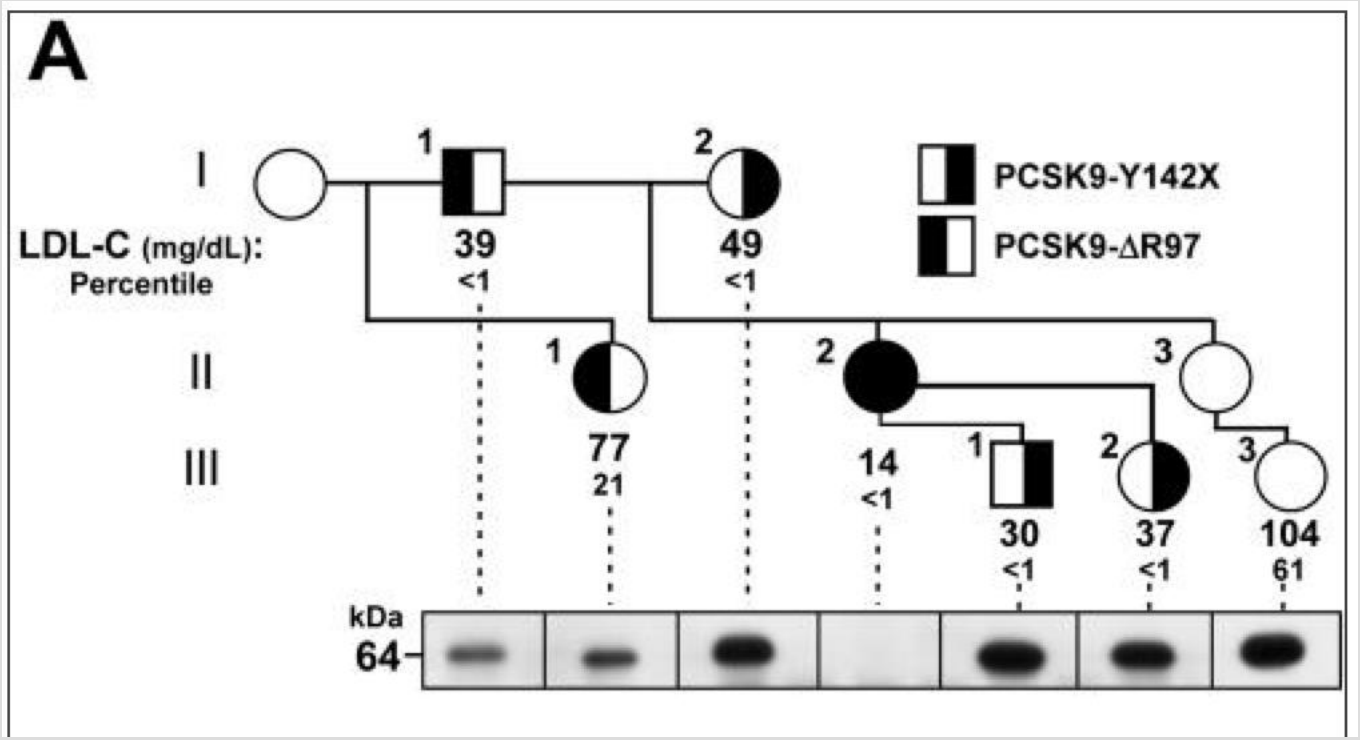
A GENETIC VARIANT IN *PCSK9* IS **ASSOCIATED** WITH CHOLESTEROL LEVELS



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

2099 Finnish individuals

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

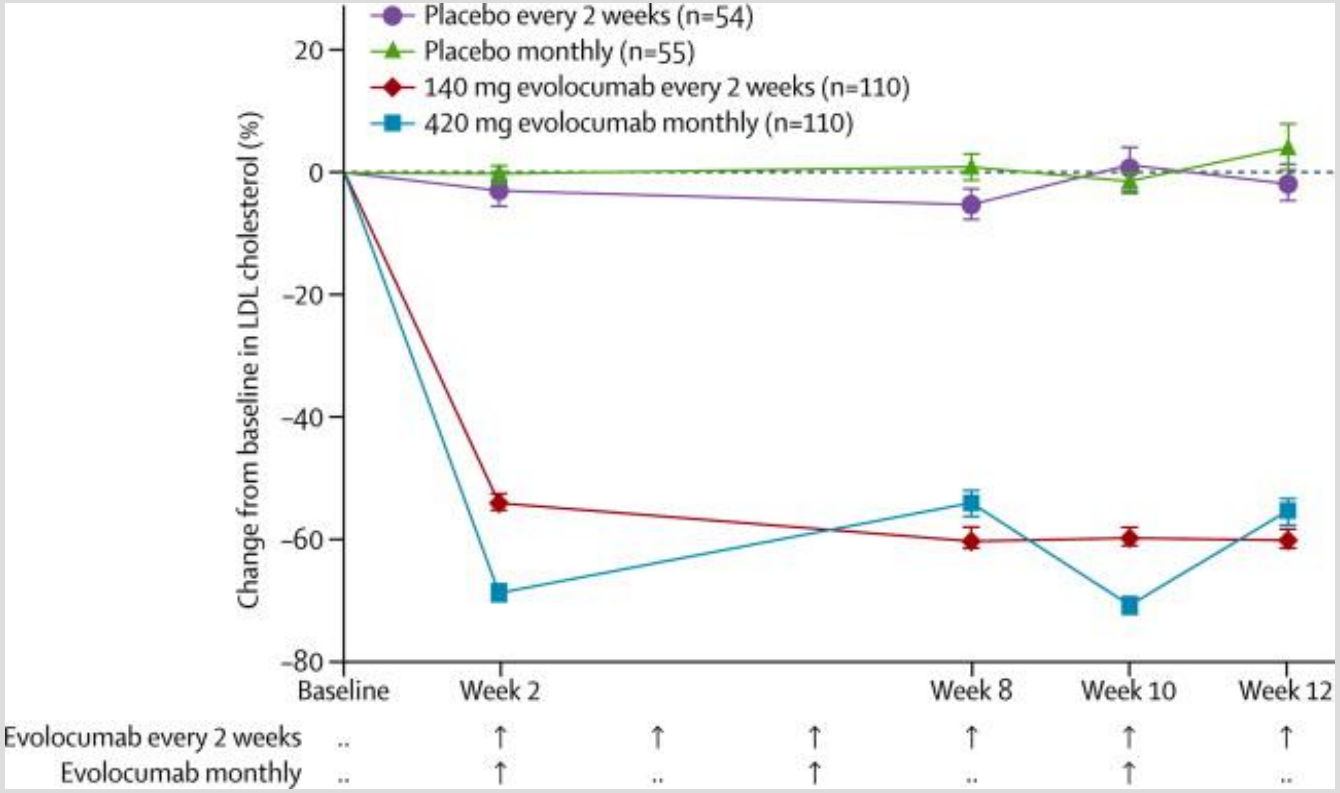


Zhao et al. AJHG 2006

- 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-C

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.

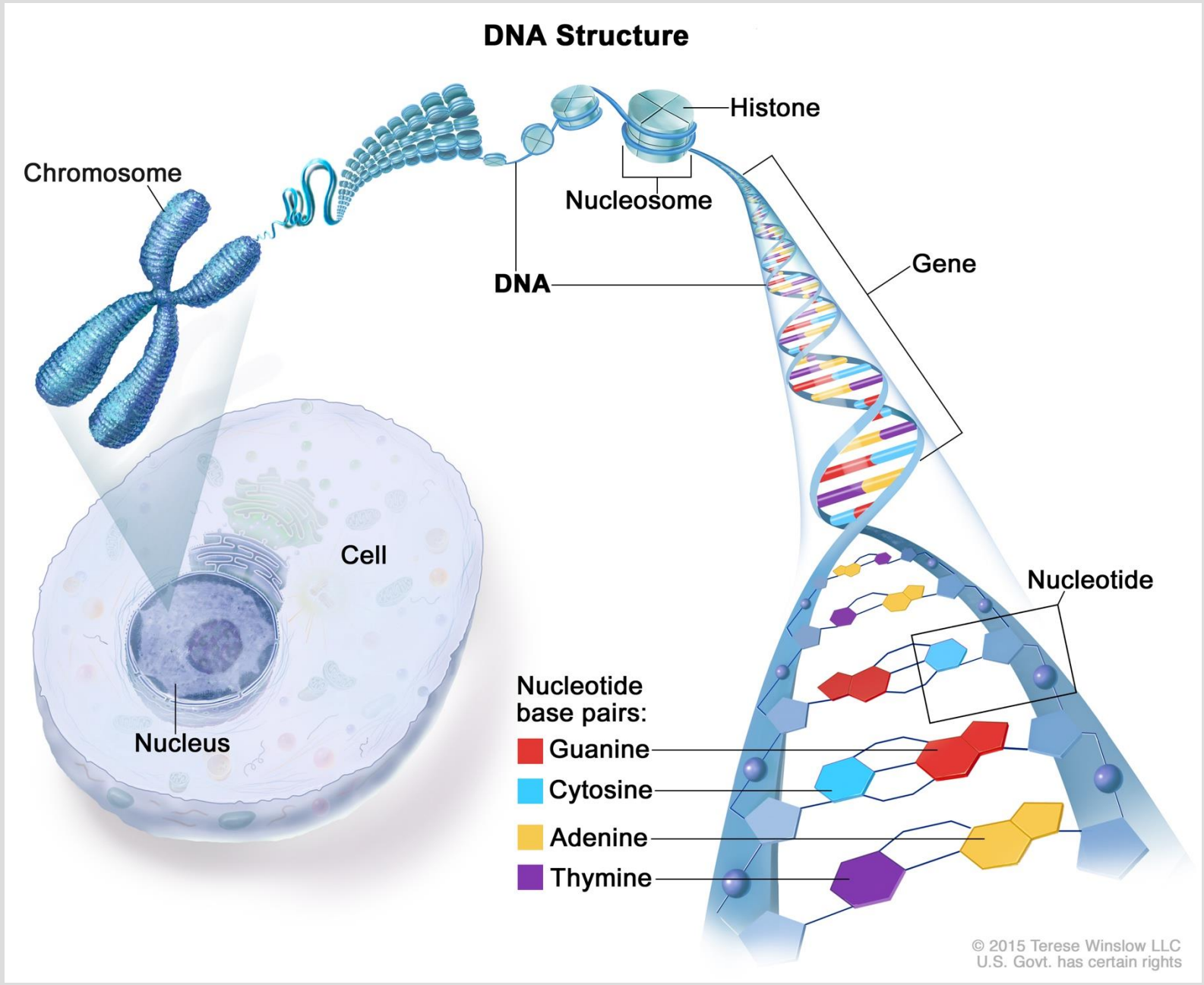
HUMAN GENOME

- Sequence of 3 x 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.

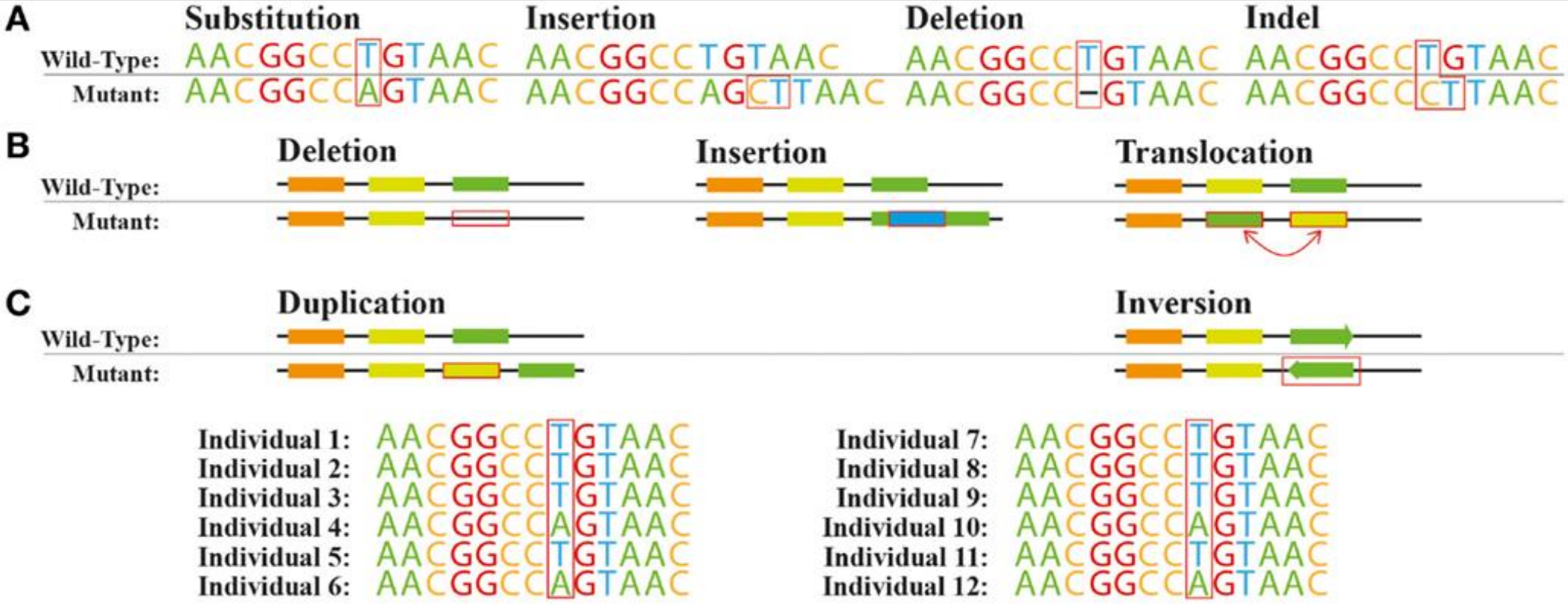
Your genomes are physically divided into 22 pairs of autosomal chromosomes and
1 pair of sex chromosomes (males XY, females: XX)



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%

Only positive strand of genomes is shown here

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

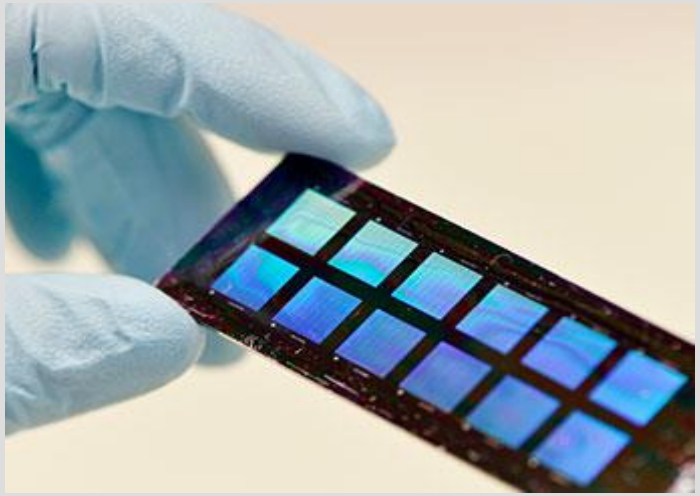
Genotypes at this SNP in population

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

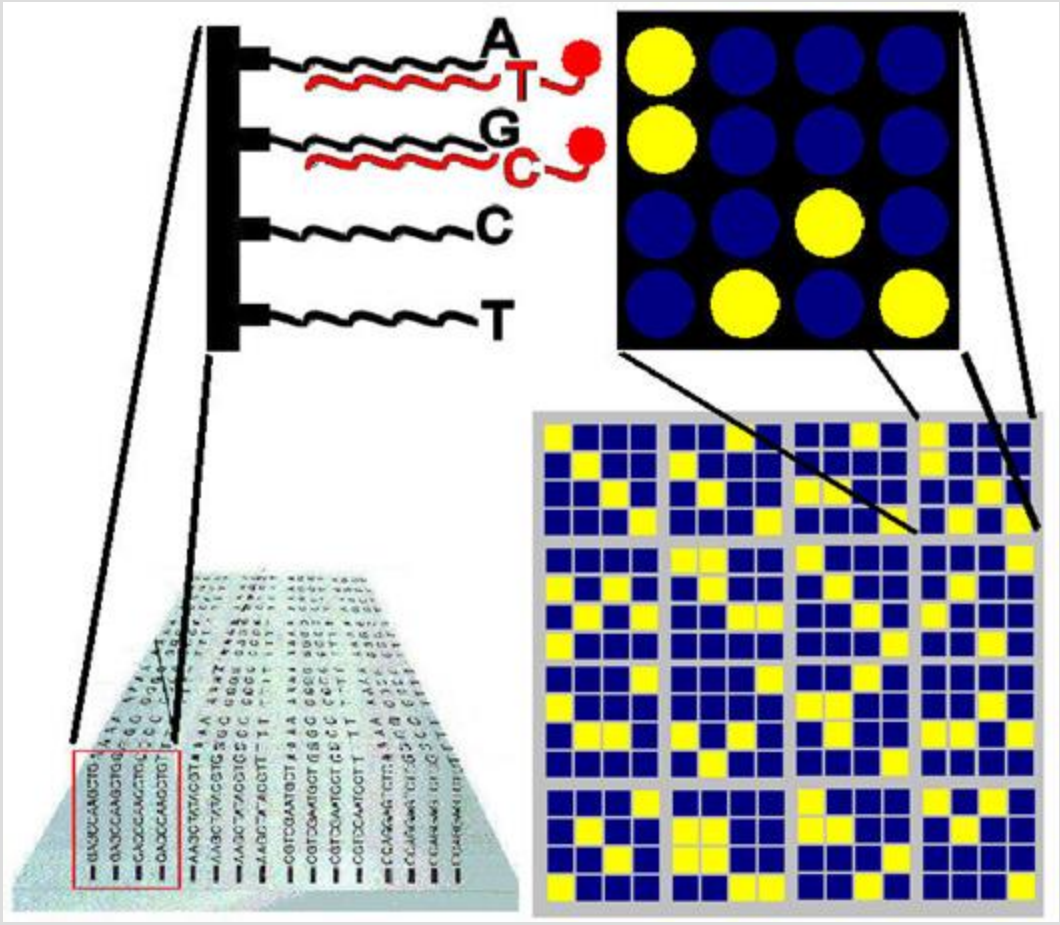
Assuming **Hardy-Weinberg equilibrium frequencies**

READING SNPS

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

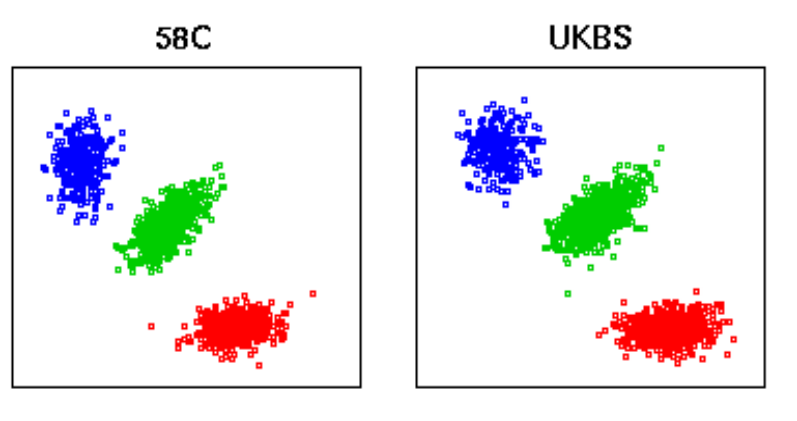


This array can genotype 12 individuals at 10^6 SNPs

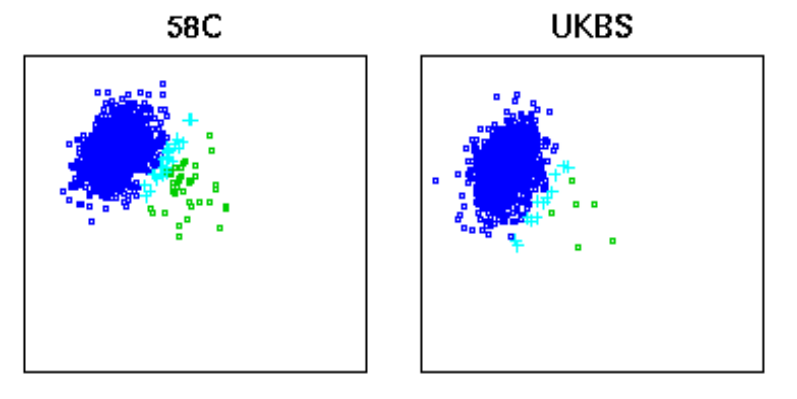


Steven M. Carr
www.mun.ca/biology/scarr/DNA_Chips.html

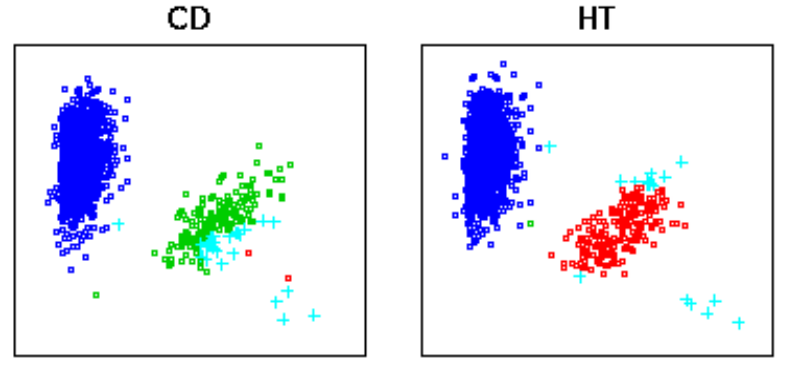
GENOTYPE CALLING FROM SNP ARRAY DATA



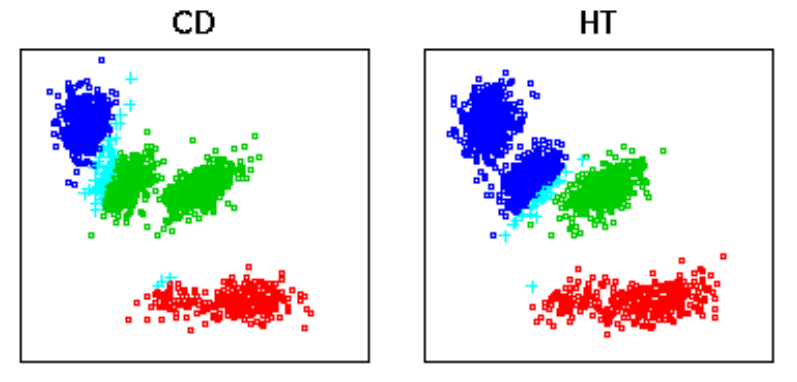
GOOD calling!



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

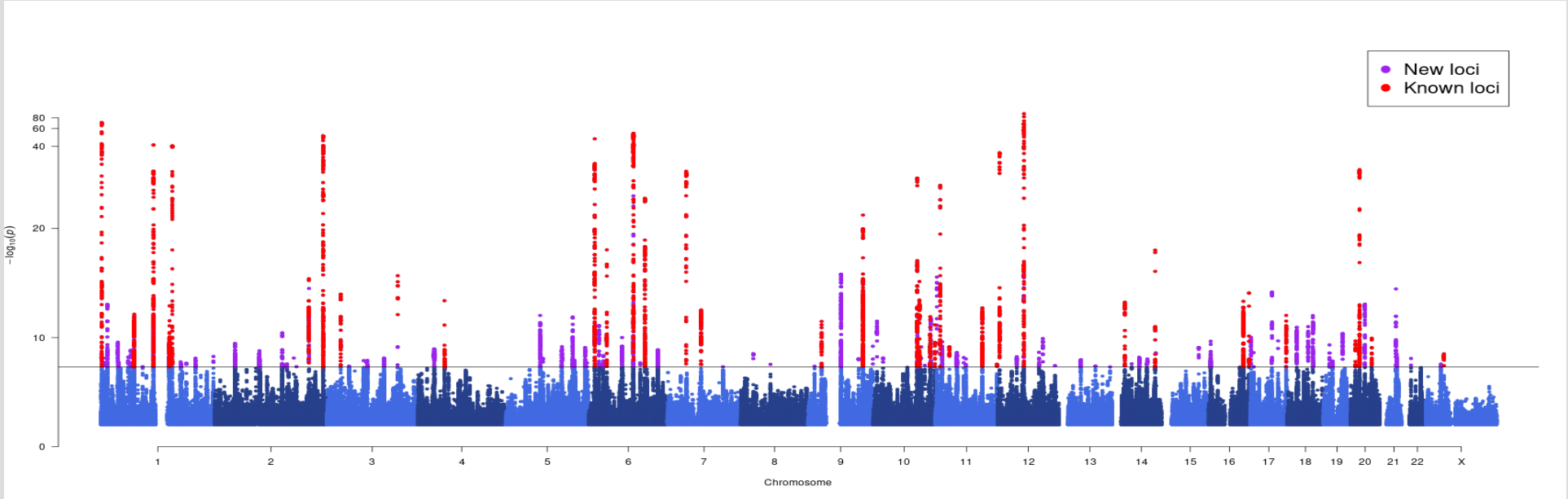
Light blue means algorithm has made no call.

Bottom line errors would likely fail HWE test.

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

GENOME-WIDE ASSOCIATION STUDY

- Statistical problem: Is genetic variation at a particular position associated with observed phenotypic variation?
  - Population cohorts with quantitative trait measurements (lipids, BMI, blood pressure)
  - Case-control studies (diseases such as schizophrenia, breast cancer or migraine)



100,000 migrain cases vs. 750,000 controls analyzed @ Helsinki / FIMM 2022

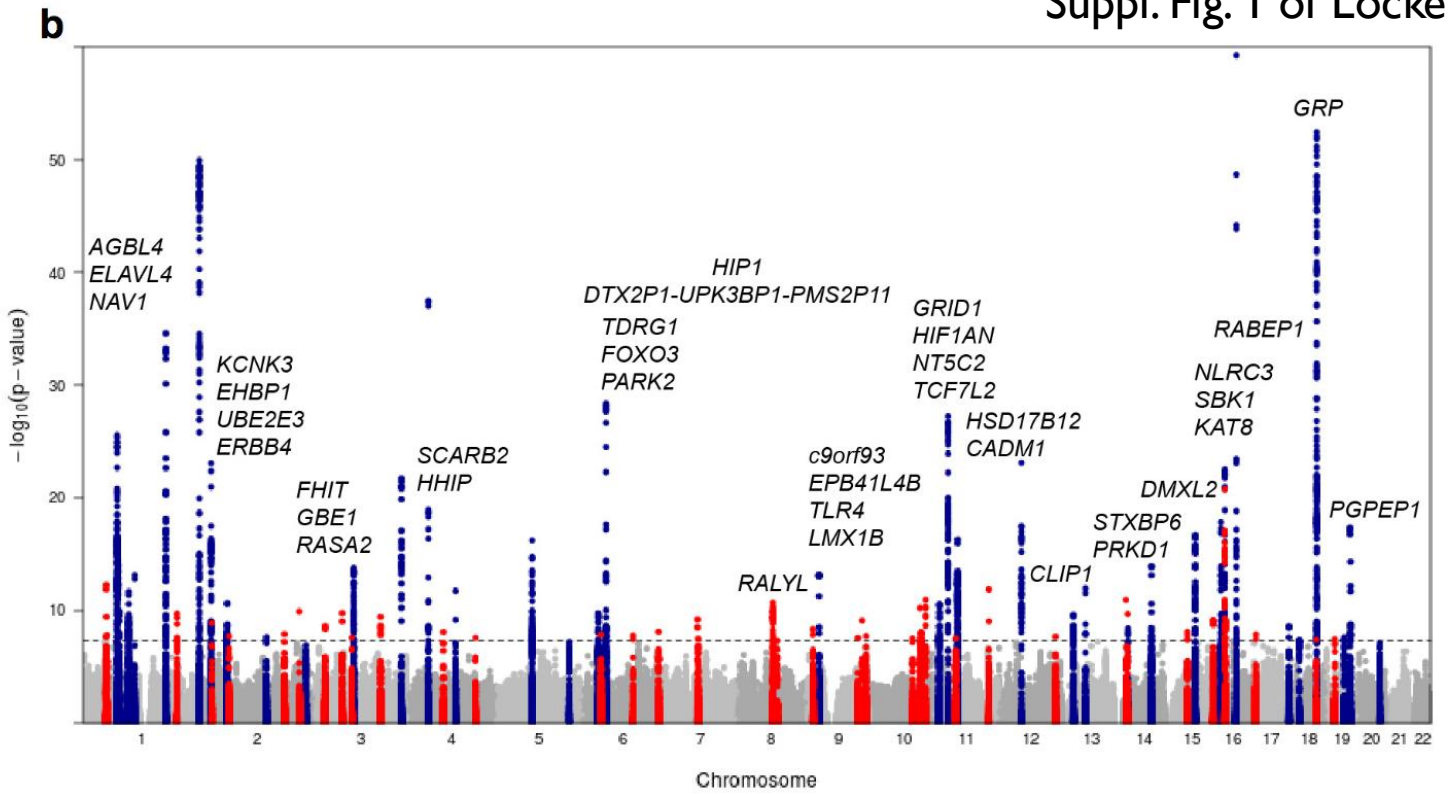
EXAMPLE GWAS

- 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 Hautakangas et al. (Nature Genetics 2022) 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

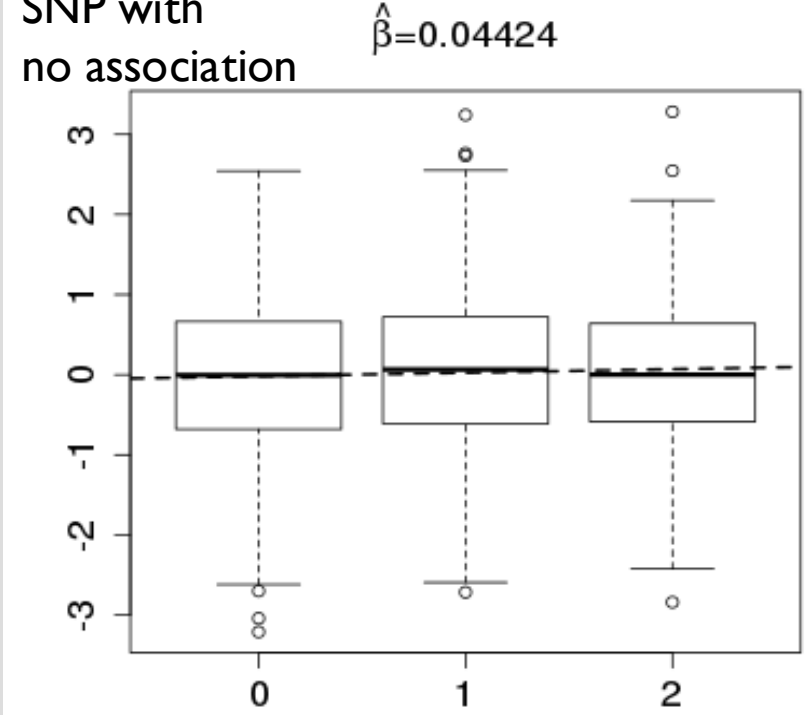
Suppl. Fig. 1 of Locke et al.



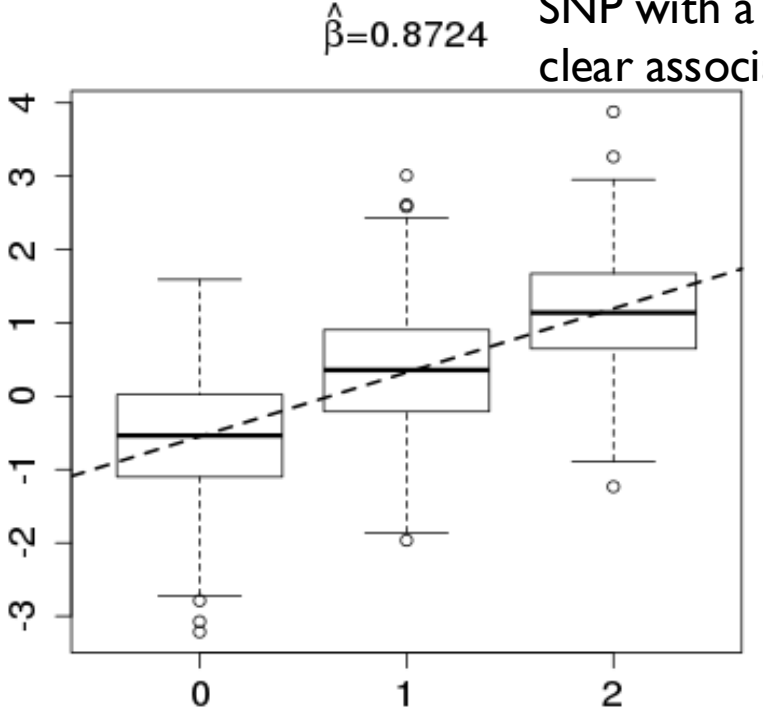
Manhattan plot shows –log10 P-value of each SNP tested in GWAS.
Genome-wide significance level at P=5e-8 or –log10(P) = 7.3.
Previously known loci are in blue, new findings are in red.
Each locus in named by a nearby gene (but that gene is not necessarily causal.)

LOCKE ET AL. DID ASSOCIATION TESTS AT 2.5 MILLION SNPS

SNP with no association

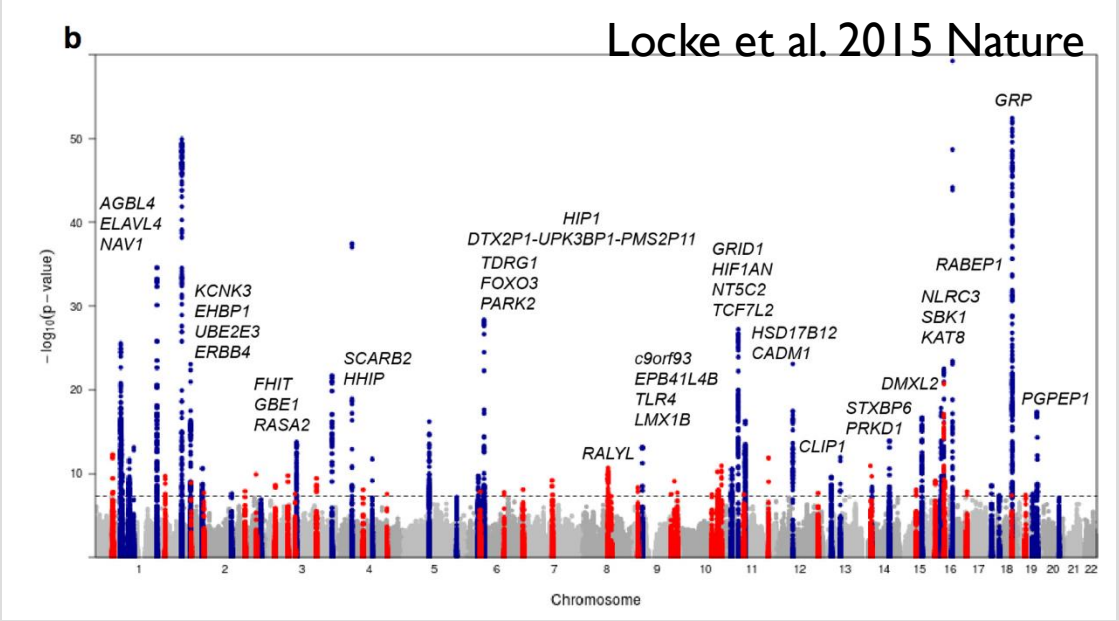


SNP with a clear association



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

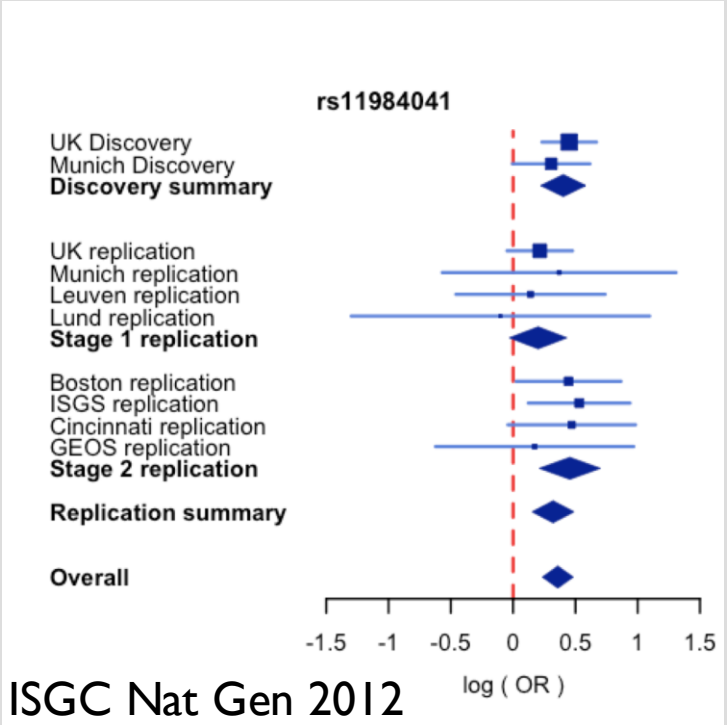
- ”339,000 individuals with genotypes and BMI available”
- ”125 studies around the world participated”



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

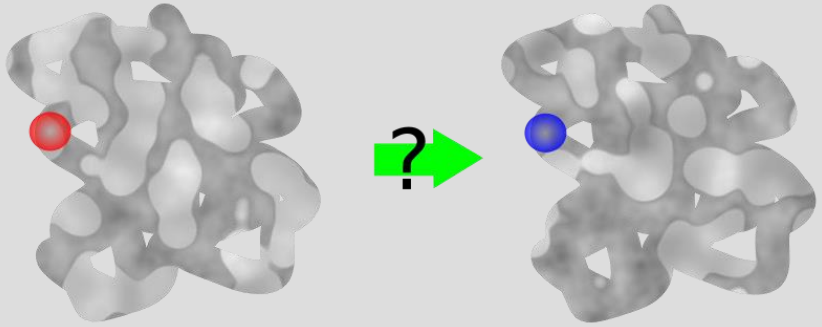
While no-one has access to all original genotype-phenotype data, everyone can access the meta-analyzed GWAS results as they are (often) publicly available.

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

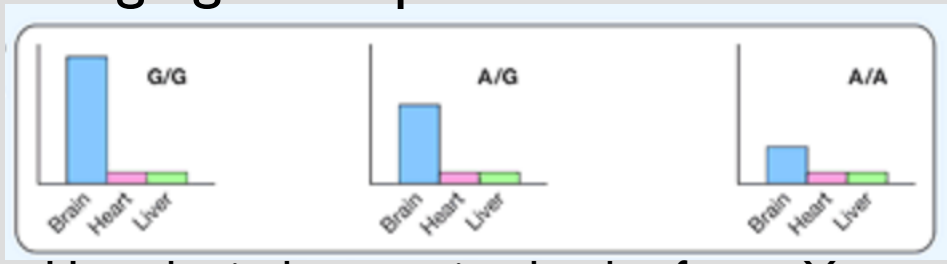
- ”97 loci (regions in the genome) convincingly associated”
- ”Each locus is a hint to biology of BMI”

What does each variant do?

Change protein? (only few GWAS hits do)

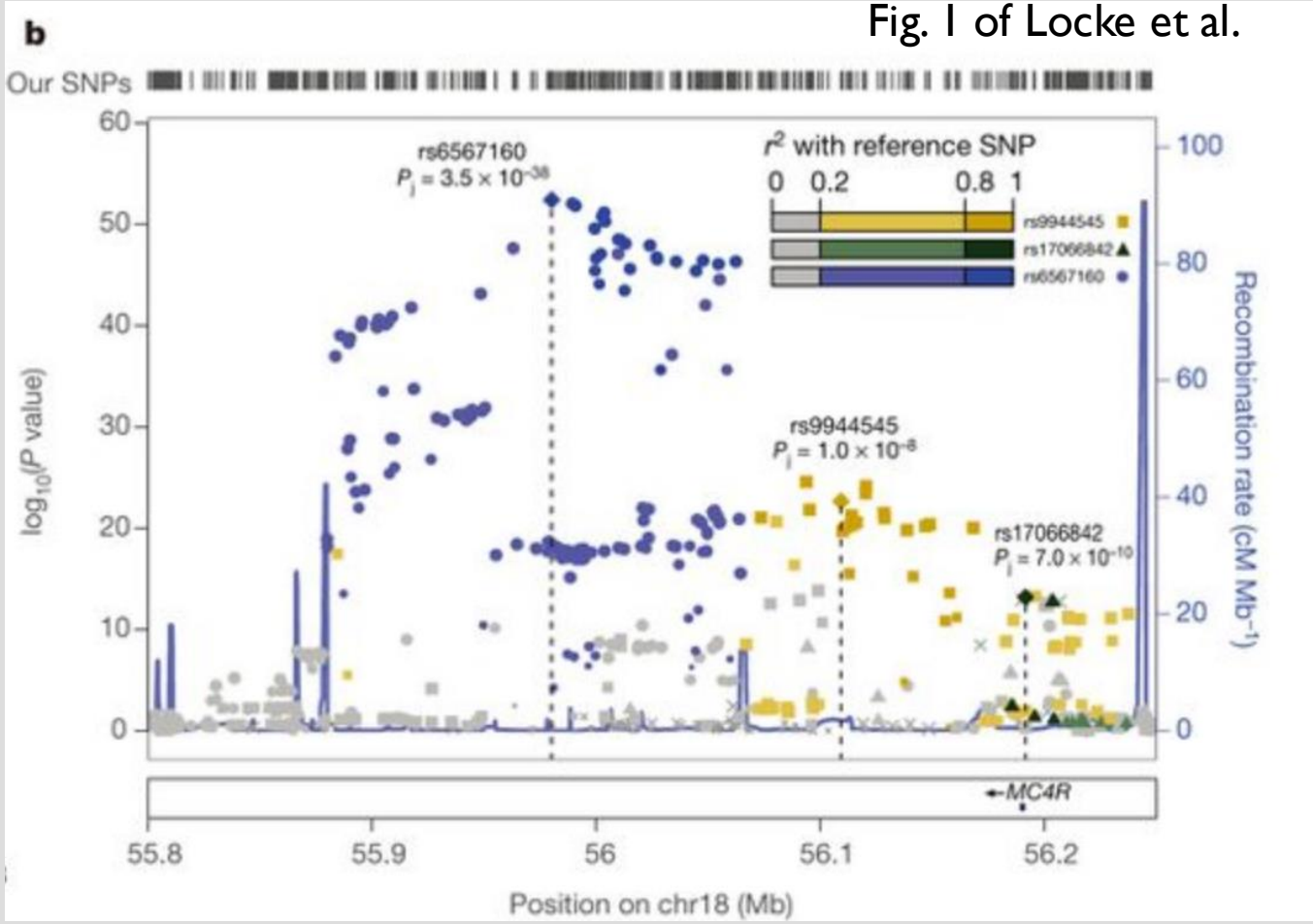


Change gene expression in certain context?



Hypothetical expression levels of gene X

Fig. 1 of Locke et al.

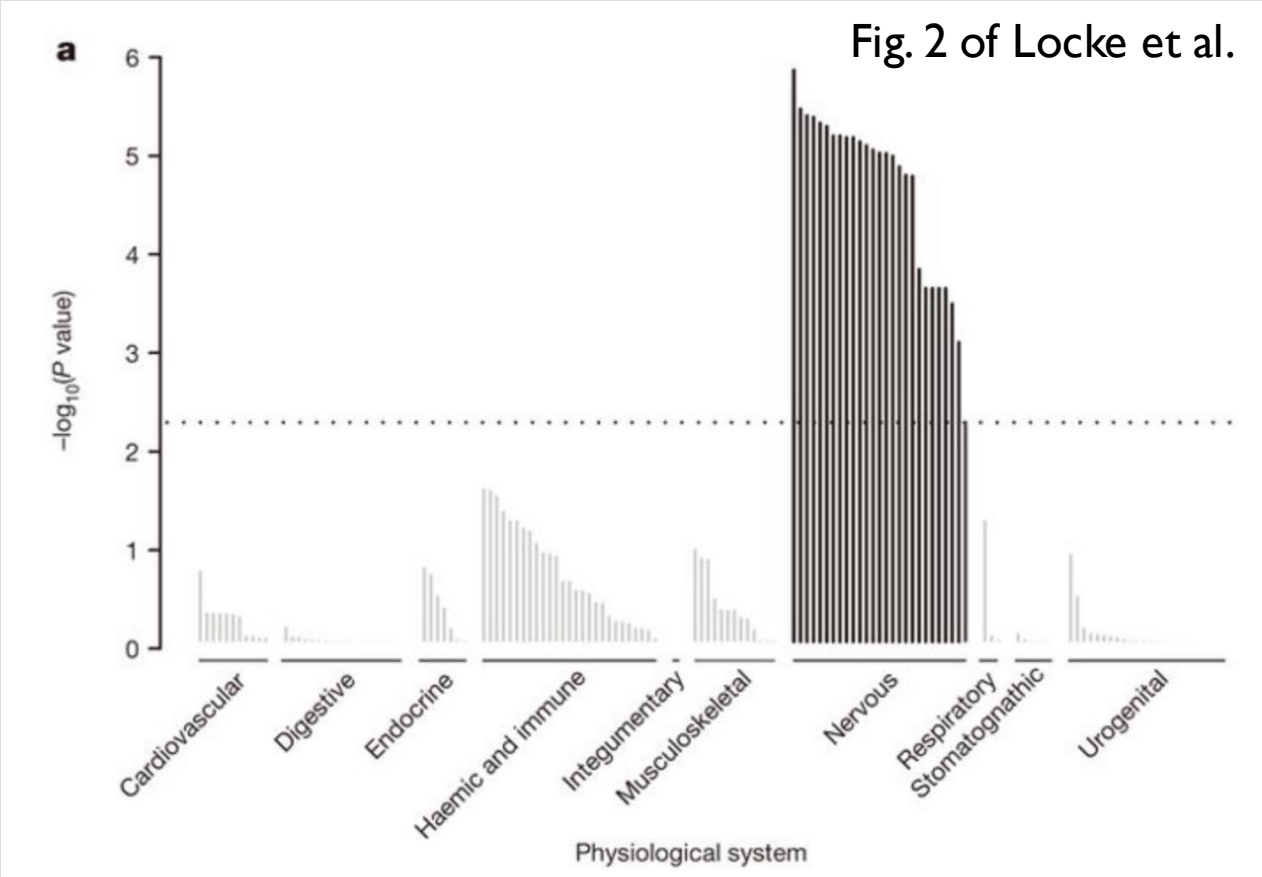


Zooming into one associated region (MC4R) on chr 19.
Many SNPs show strong association; not clear which are causal ones.
Three SNPs are highlighted as possible independent signals.

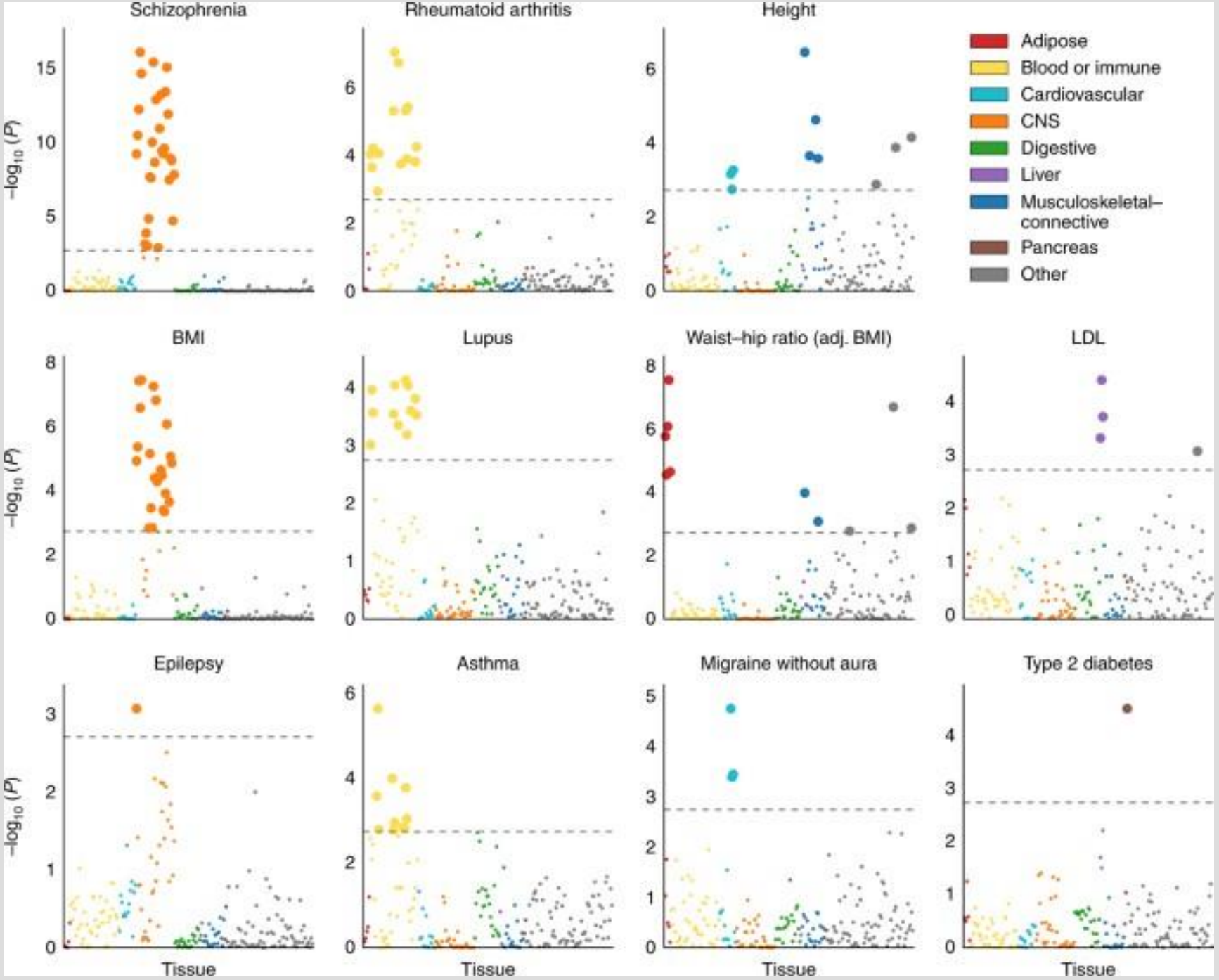
- ”Results highlight role of central nervous system in BMI”

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.



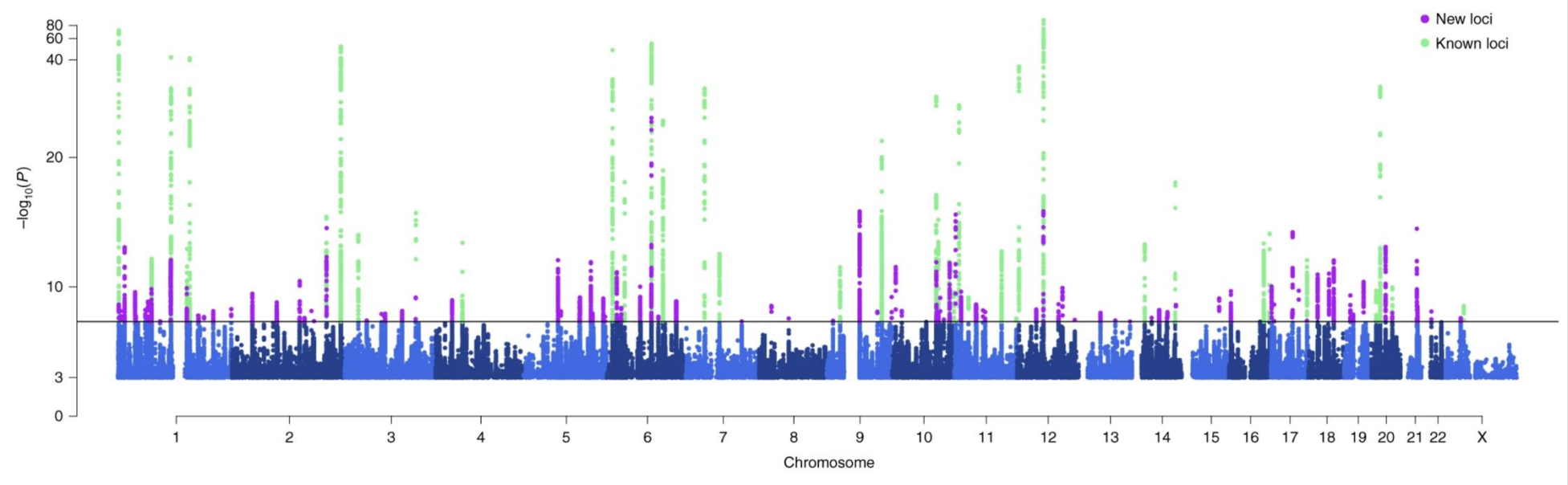
Are GWAS signals enriched in/near genes specifically expressed in certain tissue/cell type(s).

-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

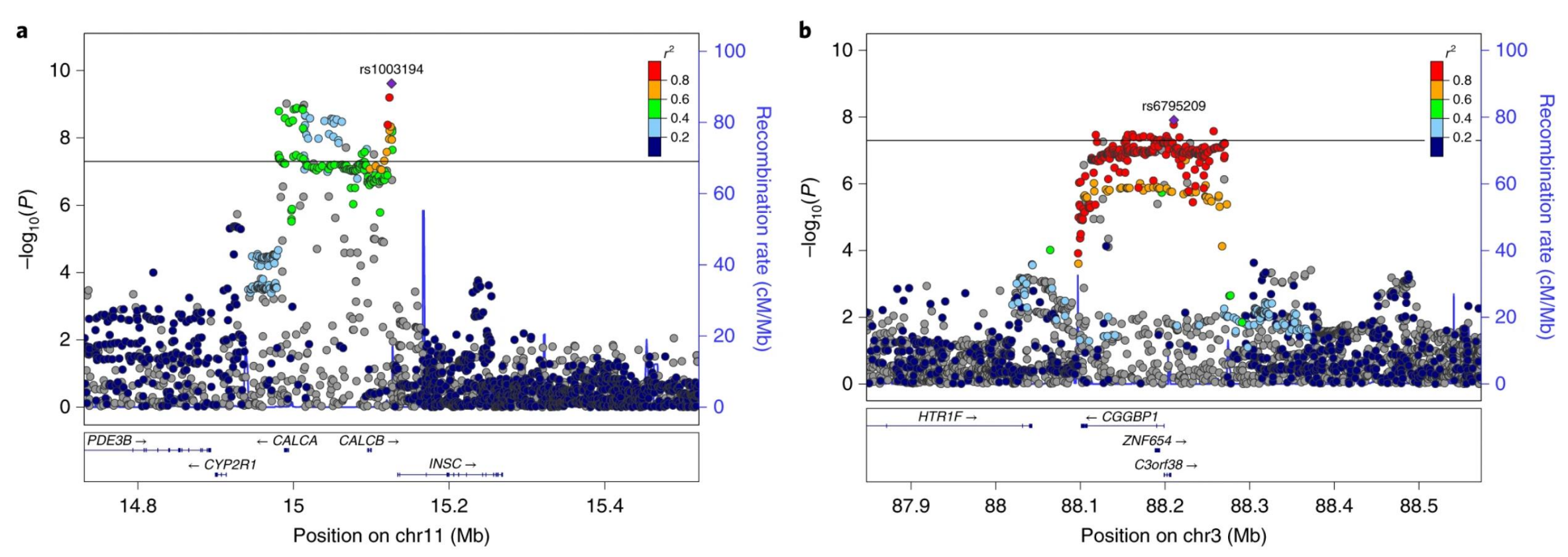
MIGRAINE GWAS (HAUTAKANGAS ET AL. 2022) (1/3)

- 102,084 cases and 771,257 controls from 25 studies
- 123 loci with convincing association



Manhattan plot of results. On the *x*-axis, variants are plotted along the 22 autosomes and the X chromosome. The *y*-axis shows the statistical strength of the association (negative $\log_{10}$ of *P*-value). The horizontal line is the genome-wide significance threshold ($P = 5 \times 10^{-8}$). The 123 risk loci passing the threshold are divided into 86 new loci (purple) and 37 previously known loci (green). Adjacent chromosomes are colored in different shades of blue.

MIGRAINE GWAS (HAUTAKANGAS ET AL. 2022) (2/3)

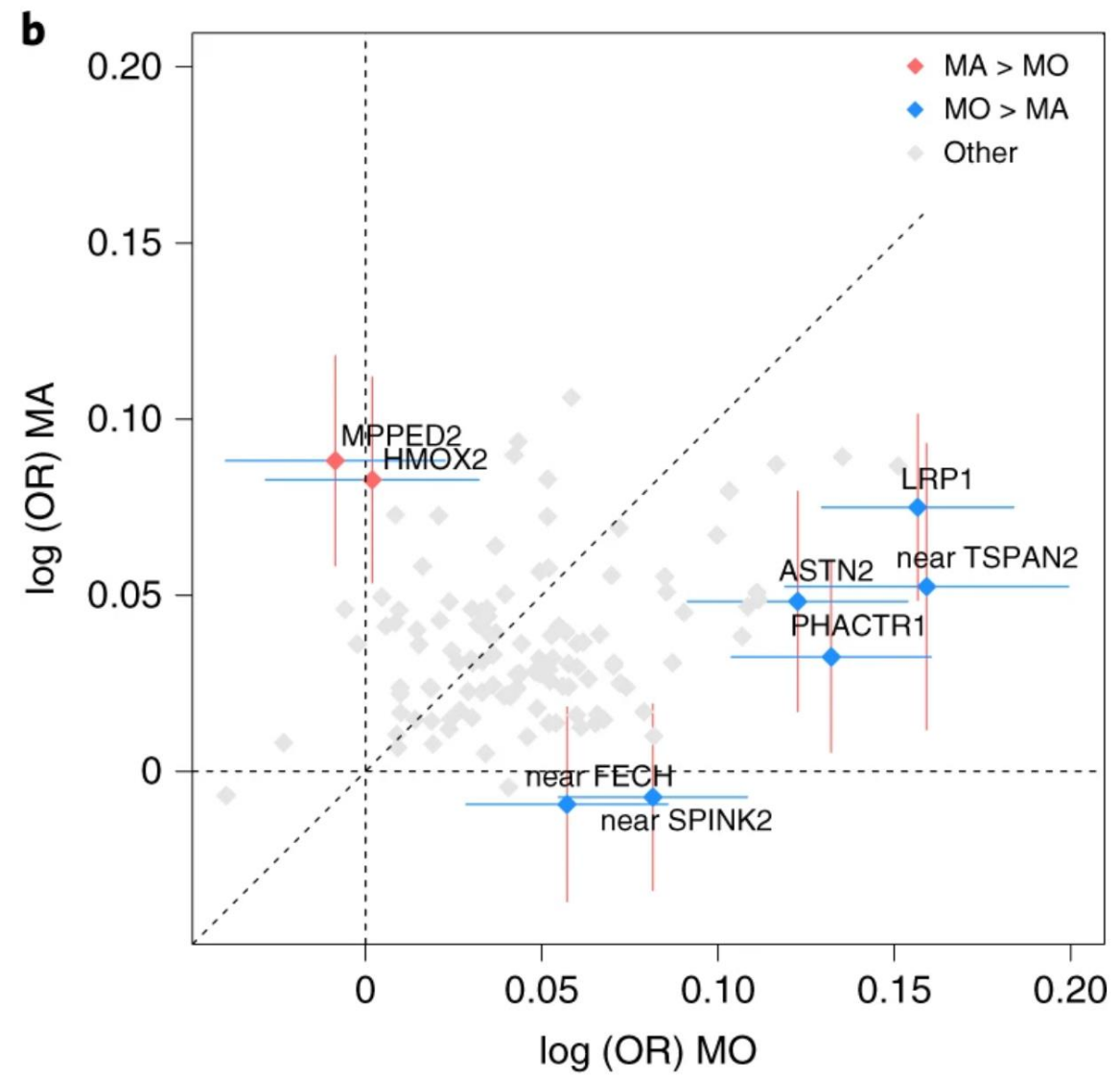


a, Locus containing *CALCA* and *CALCB* genes, encoding CGRP, which is the target of preventive and acute therapies via monoclonal antibodies and gepants.

b, Locus containing the *HTR1F* gene, which encodes a serotonin 5-HT_{1F} receptor that is the target of acute therapies via ditans.

Maybe there are promising candidates for drug therapies among the other 121 genetic regions highlighted by this GWAS?

MIGRAINE GWAS (HAUTAKANGAS ET AL. 2022) (3/3)



Is biology behind different migraine subtypes same or different?
Here comparing:

- Migraine with aura (MA) and
- Migraine without aura (MO).

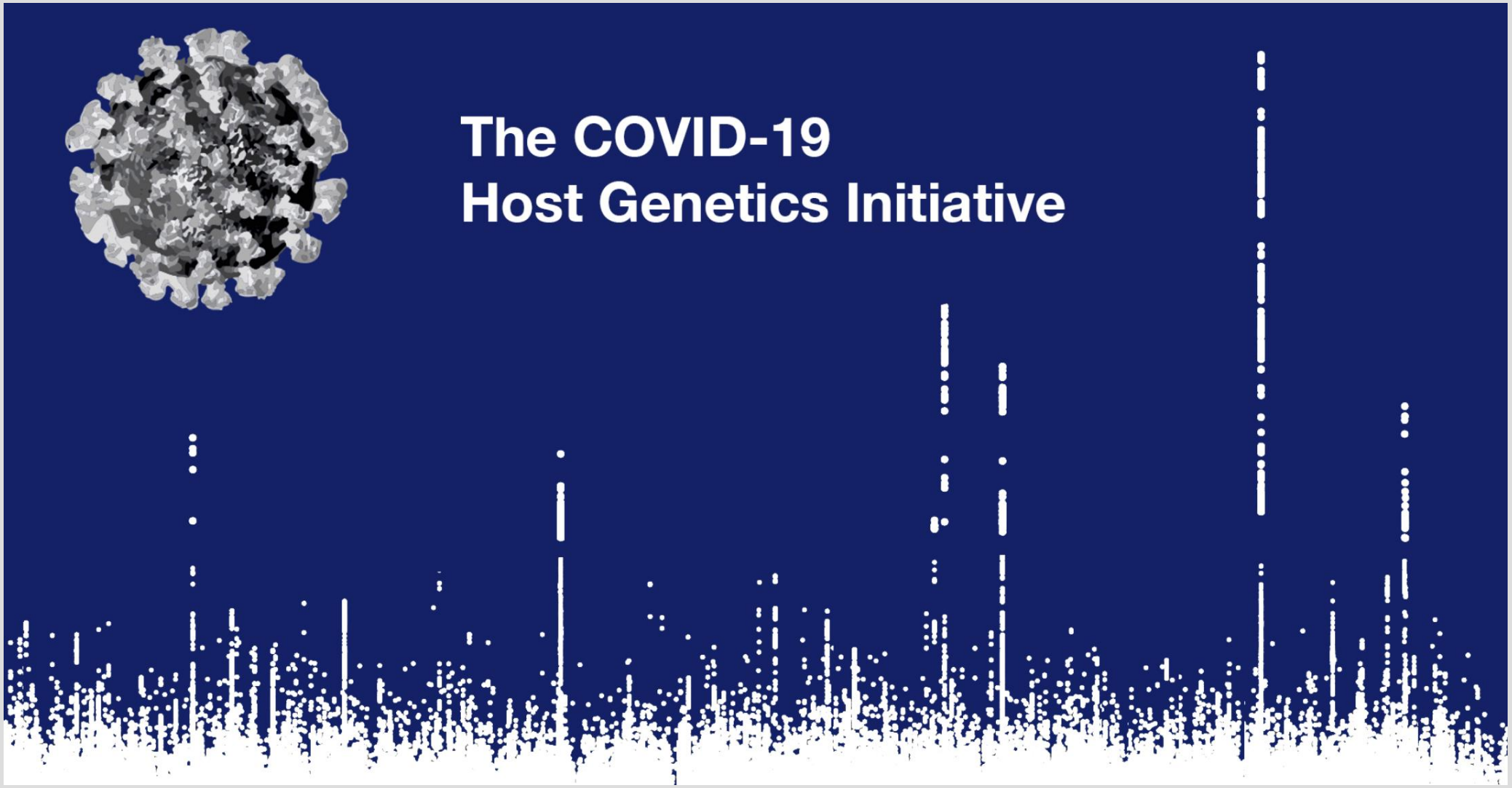
Axes show effect sizes of migraine risk alleles
as logarithm of odds ratios (OR) for
MO (*x* axis; 15,055 MO cases and 682,301 controls) and
MA (*y* axis;14,624 MA cases and 703,852 controls).

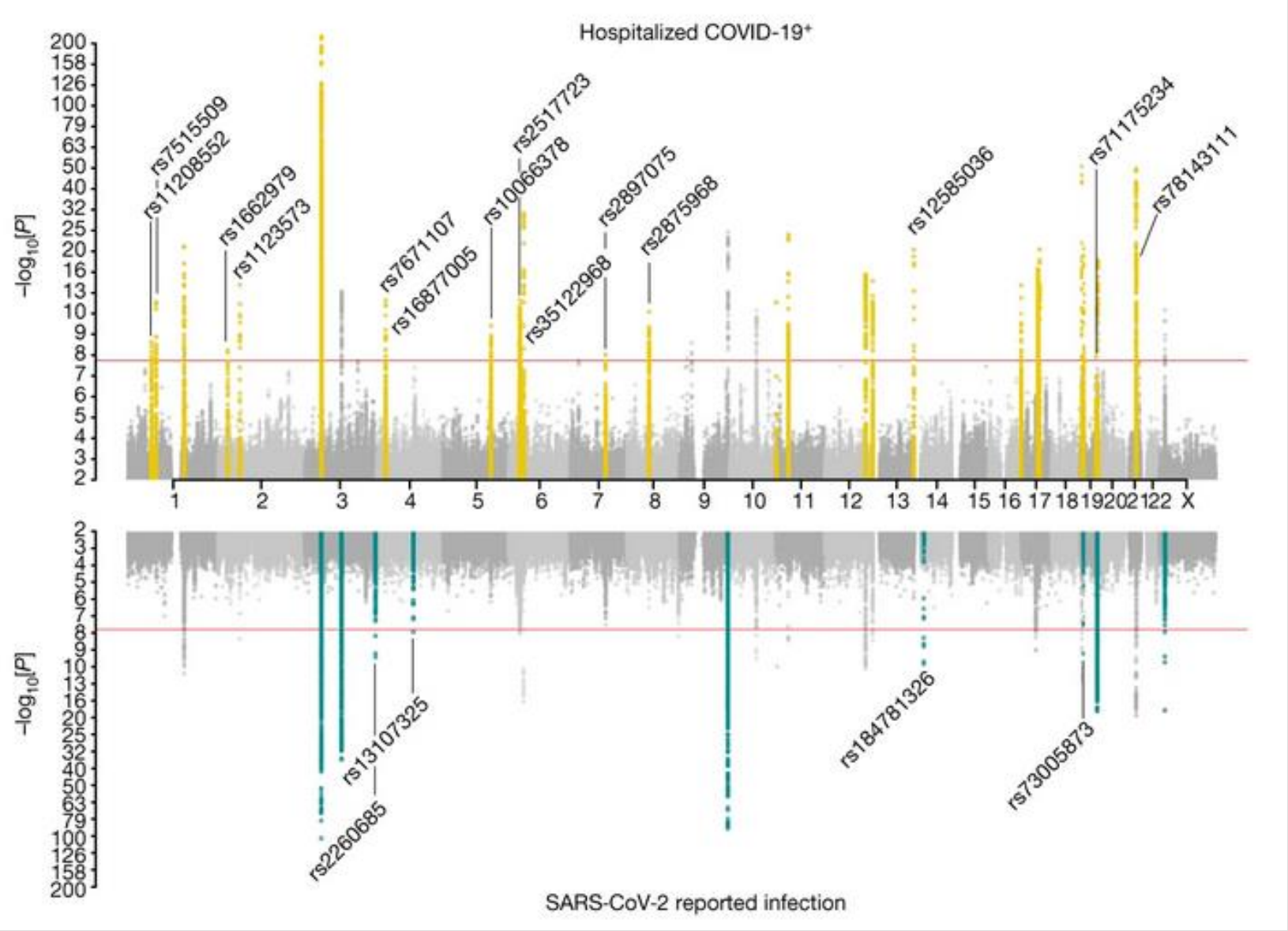
Colored variants show statistically different effects size between the two subtypes. The variants are named by their nearest genes.

(significance level is Bonferroni corrected $P = 0.05$).

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

https://www.covid19hg.org/





Studied 219,692 cases and over 3 million controls across 82 studies from 35 countries.

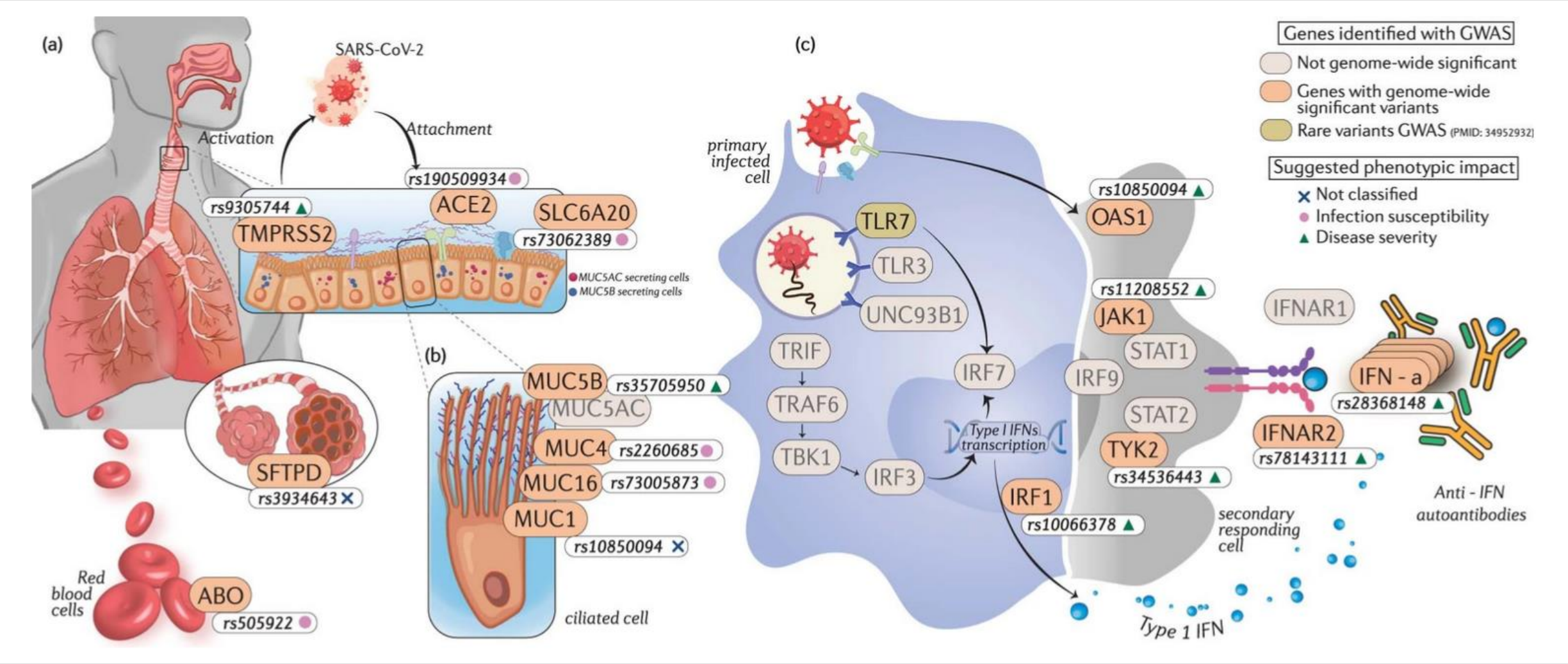
Data collected before vaccination.

Results for 3 phenotypes:

- critical illness (30 loci)
- hospitalization for COVID (40 loci)
- Infection with SARS-CoV-2 (21 loci)

https://www.nature.com/articles/s41586-023-06355-3

https://www.nature.com/articles/s41586-023-06355-3



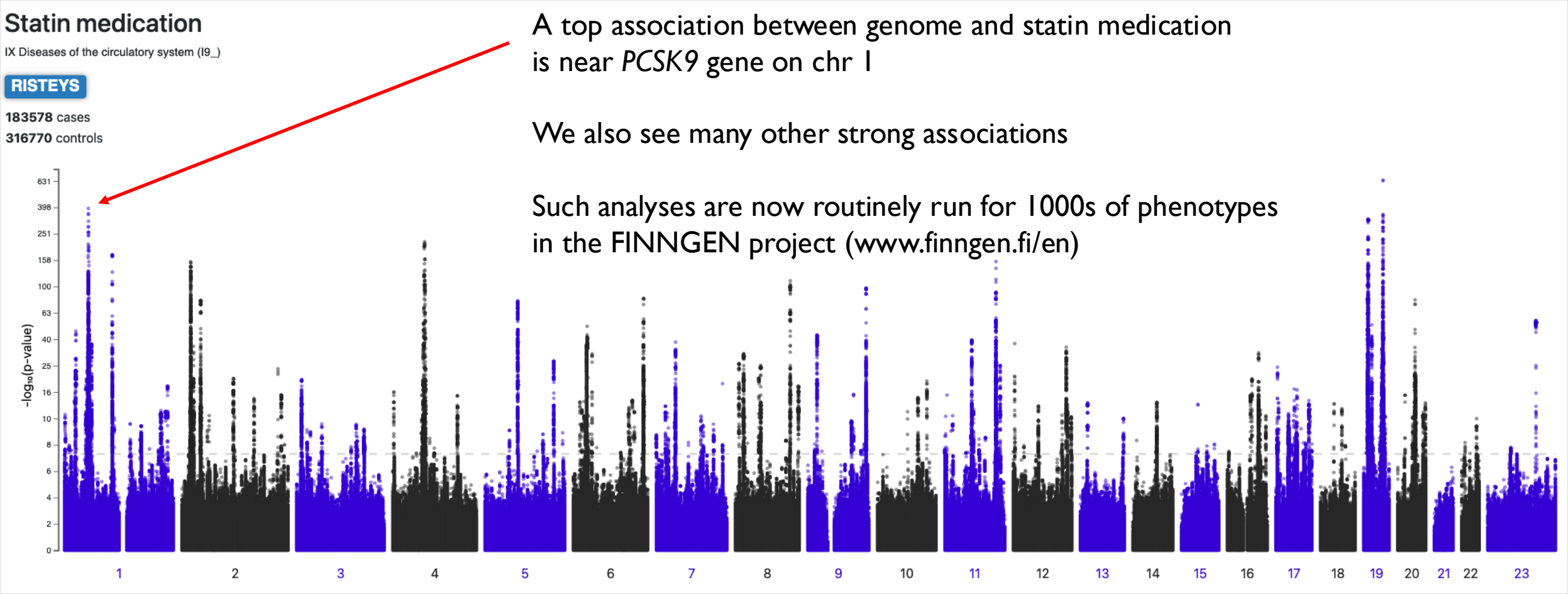
Genome-wide significant variants associated with COVID-19 (white boxes) and the annotated genes (peach boxes) are mapped on to pathways known to be involved in (**a**) viral entry and innate immunity, (**b**) entry defence in airway mucus, and (**c**) type I interferon. The suggested phenotypic impact of the significant variants using a Bayesian approach are denoted with shapes; COVID-19 susceptibility (pink circles), COVID-19 disease severity (green triangles), and unclassified variants (blue cross). Other genes known to be involved in the aforementioned pathways are shown using grey boxes.

ASSOCIATIONS AT SCALE

- Biobanks with 100,000s of samples and 1000s of phenotypes are now being analyzed
- FINNGEN project collected genetic data of 500,000 Finnish samples and combines it with health care records
  - GWAS for 1000s of diseases
    - Variant X reduces risk of disease D but what else does it do?
  - Possibility to recontact individuals and gather more information on them

www.finngen.fi/en

https://r12.finngen.fi/pheno/RX_STATIN



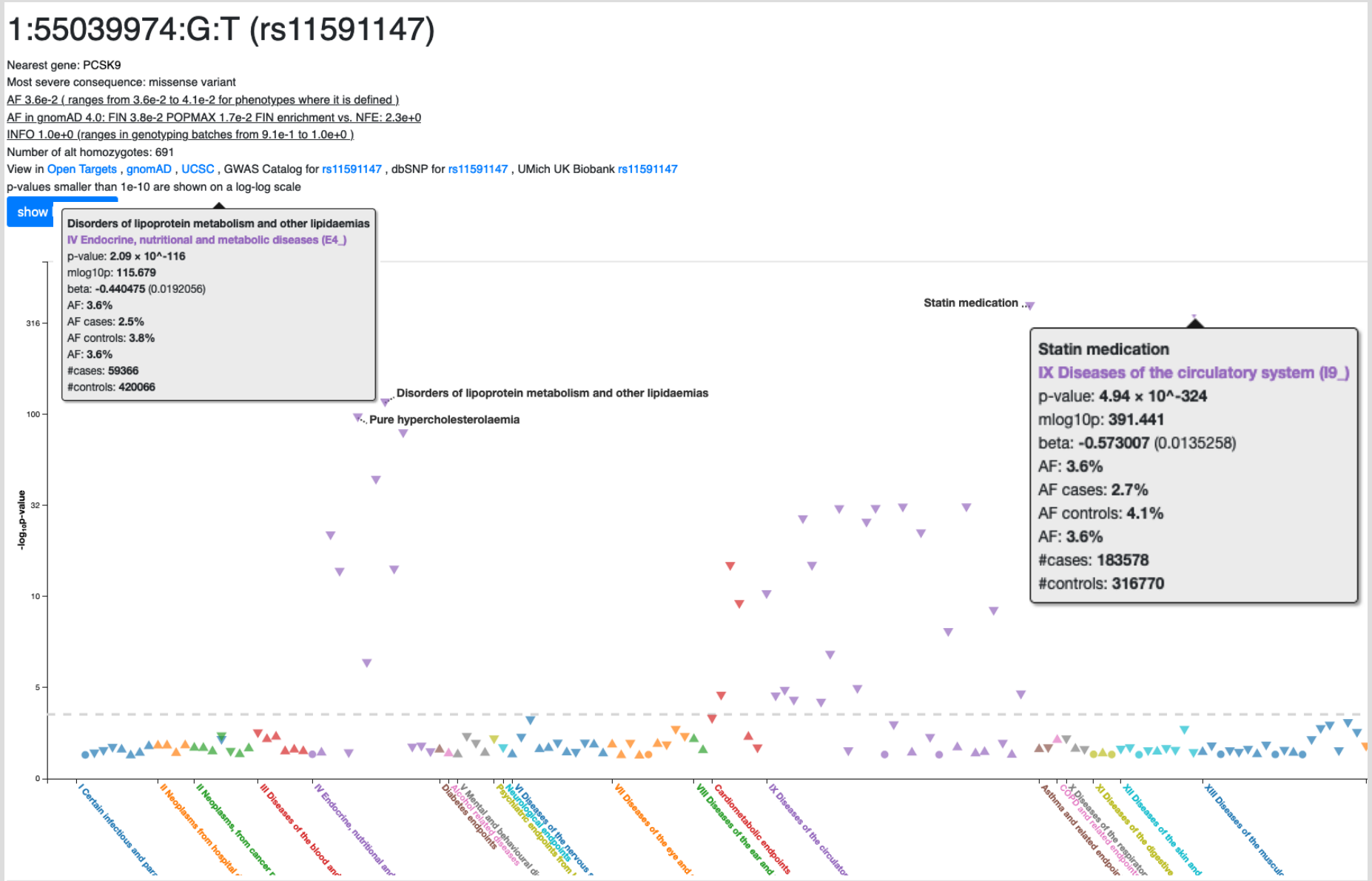
A top association between genome and statin medication is near *PCSK9* gene on chr 1

We also see many other strong associations

Such analyses are now routinely run for 1000s of phenotypes in the FINNGEN project (www.finngen.fi/en)

Genome from chromosomes 1 to 22 + X chr (labelled as 23).

Phenome-wide view on the top variant for usage of statins taken from:
https://r12.finngen.fi/variant/1-55039974-G-T

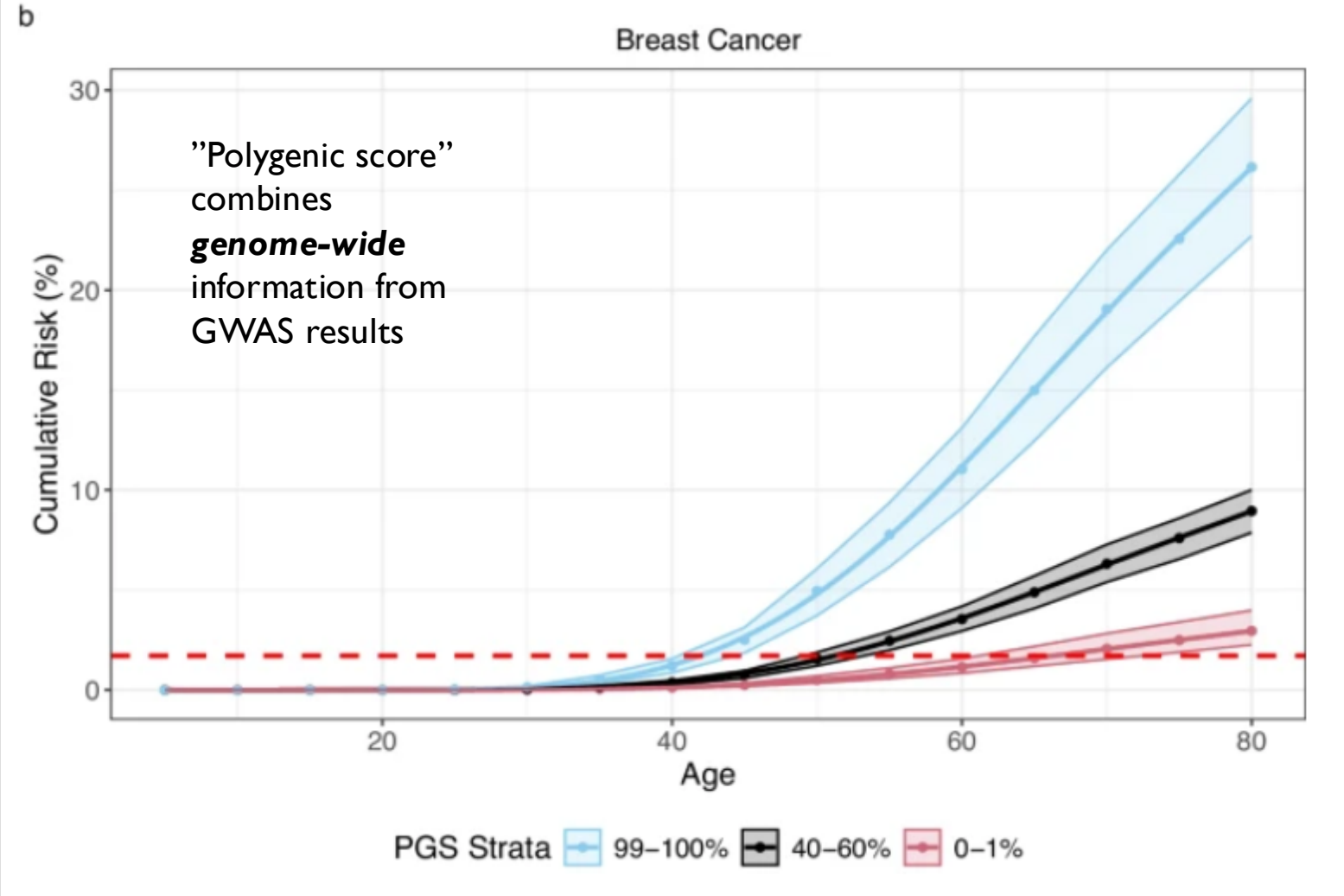


What else does a variant associate with?

Is the same allele that reduces risk of high cholesterol also increasing risk of some other disorder?

This is a crucial question for designing safe therapeutics.

GENETIC PREDICTION



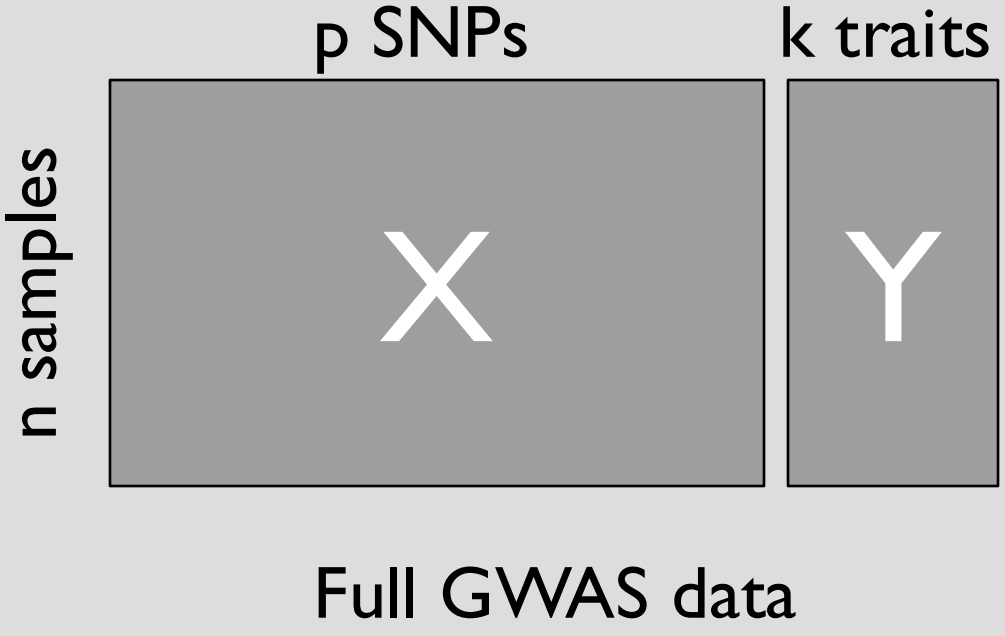
The three curves show the cumulative risk for breast cancer (BC) for three groups of women depending on their genetically predicted BC risk:
Lowest 1% of population (red)
Middle 20% of population (black)
Highest 1% of population (blue)

The red dashed line shows 2% risk that is observed at the 50 year-old female population, where the screening is currently started.

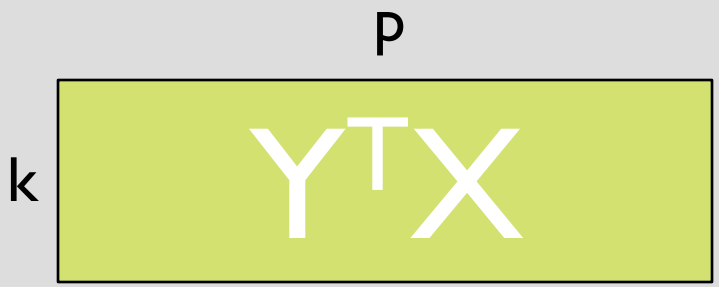
For high-risk population, the 2% risk is present almost 10 years earlier than in the general population.

https://www.nature.com/articles/s41467-024-48938-2

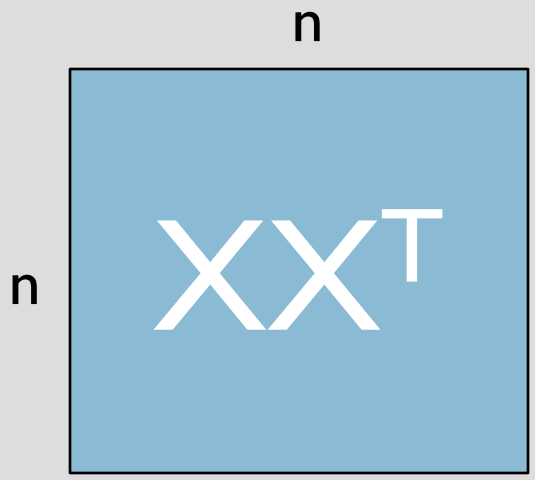
GWAS IN MATRICES



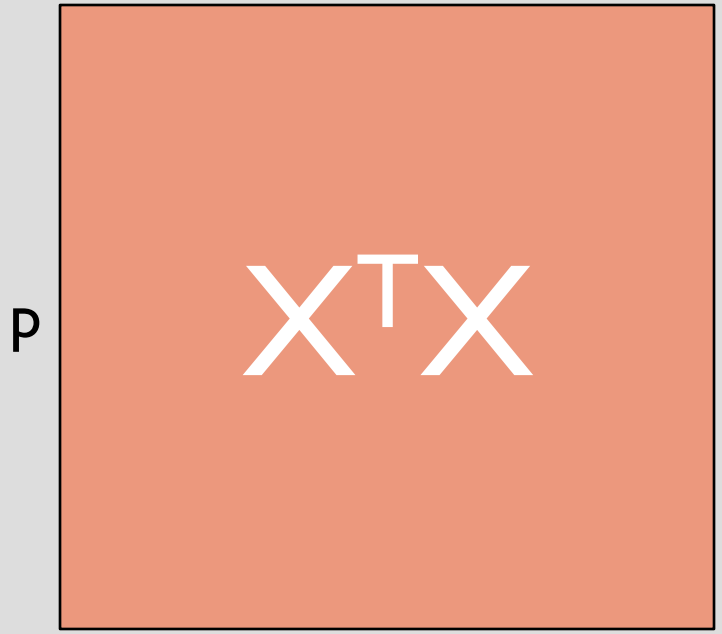
Full GWAS data



GWAS results (betas, SE, P-value)

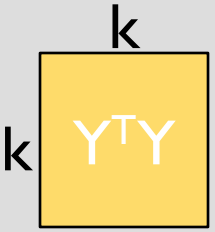


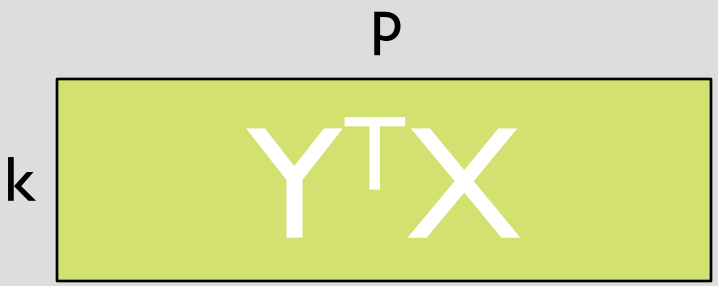
sample relatedness



LD matrix (correlation of variants)

trait correlation

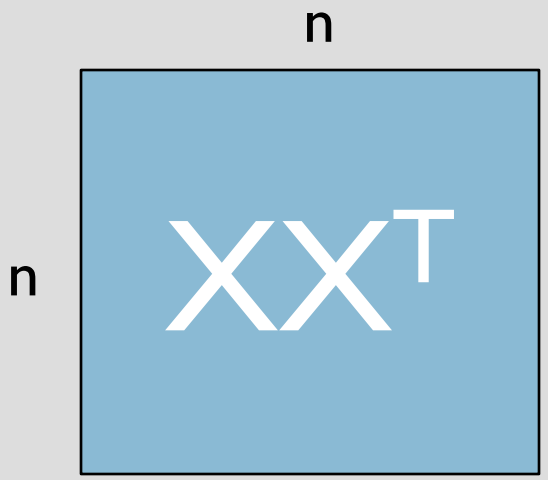




GWAS results (beta, SE, P-value)

Weeks 1-7:

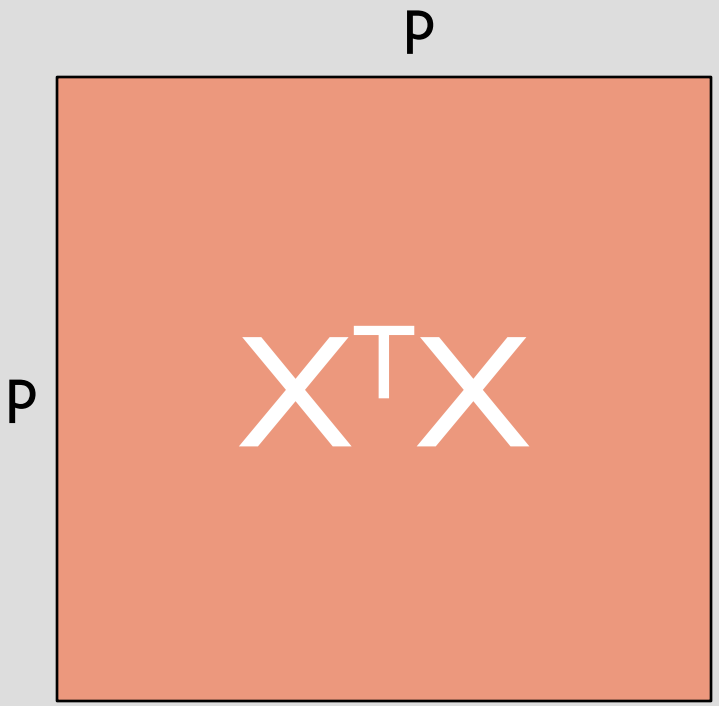
statistical inference,
statistical power,
confounders,
covariates,
summary statistics,
meta-analysis
polygenic scores



sample relatedness

Weeks 3, 5:

Relatedness & population structure
Heritability & mixed models



LD matrix (correlation of variants)

Weeks 4,5:

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