



Intro to GWAS

from the
Health Data Science
Sandbox



Alba Refoyo Martinez, PhD

Sandbox Data scientist

Center for Health Data Science (HeaDS)

UNIVERSITY OF
COPENHAGEN



GWAS with the Genomics Sandbox



Intro to GWAS topics

1. Why GWAS?
2. The genome, DNA structure and genetic variations
3. GWAS types and examples
4. GWAS steps overview



What and why?

The genome and the phenome



Genomic make-up of an individual

Variations in the genome between individuals (e.g. SNPs) stays constant through lifetime!

“Genome-wide” studies consider variation in millions of position across the genome



Phenome is a snapshot of your biology comprising all traits/phenotypes

- Measurable traits (blood pressure, BMI)
- Disease status (multiple sclerosis, diabetes)
- Behavioural traits (smoking)



What and why?

The genome and the phenome



association?



What is the genetic contribution to observed phenotypic variation?



Why GWAS?

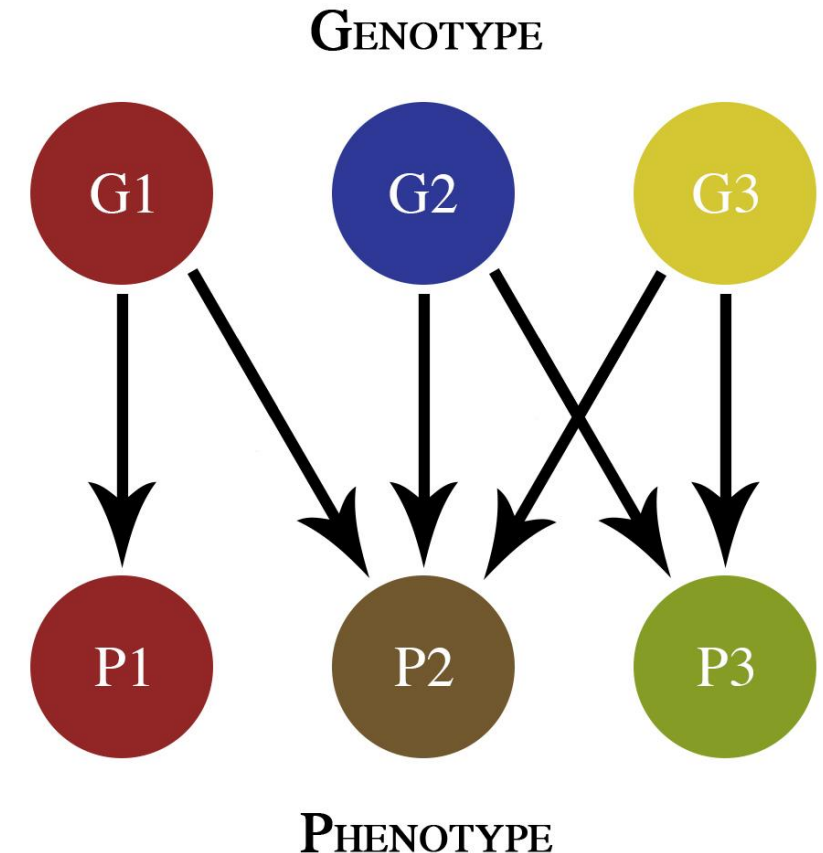
Genome-phenome association

Goal: To identify **specific genetic variants** that influence a specific trait

Why? typically disease related traits, e.g. febrile seizures or the susceptibility to diabetes

Approach: an statistical association test can

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



Why GWAS?

Studying the genome...

It can benefit:

Medicine

Molecular and
environmental
interventions against
harmful phenotypes



Why GWAS?

Studying the genome...

It can benefit:

Medicine

Biotechnology

Improving the ways
we utilize microbes,
plants or animals



Why GWAS?

Studying the genome...

It can benefit:

Medicine

Forensics

Biotechnology

More accurate
identification of
an individual from
a DNA sample



Why GWAS?

Studying the genome...

It can benefit:

Medicine

Biotechnology

Forensics

Ancestry
inference

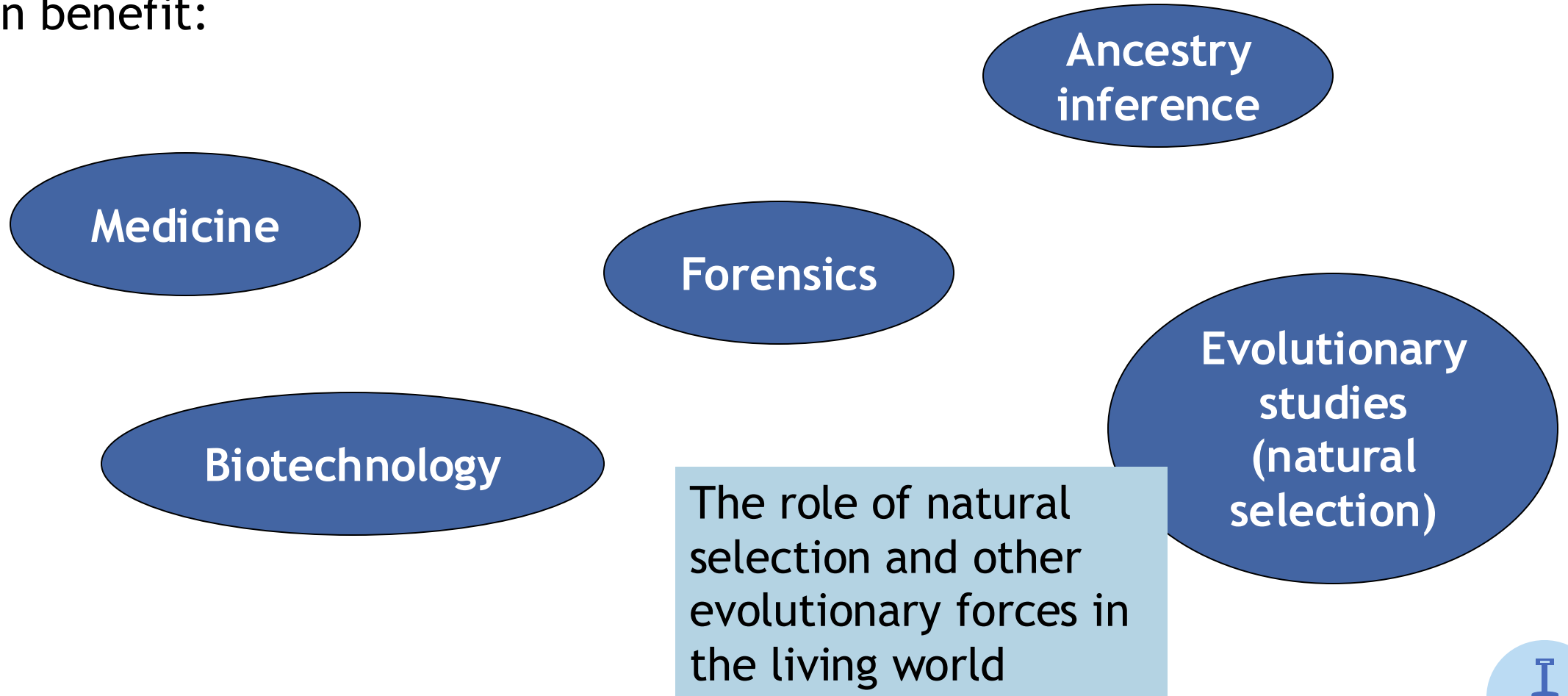
Biogeographic
ancestry
inference of
individuals,
populations &
species



Why GWAS?

Studying the genome...

It can benefit:



Why GWAS?

Studying the genome...

Why pursue this goal?

- Help uncover the underlying **genetic architecture** of the trait
- Hopefully improve the understanding of **diseases mechanisms**
- Ideally lead to better treatments and prevention strategies

It can also be used in evolutionary studies e.g. it helps trace how traits have evolved and uncover adaptations to different environments over time!



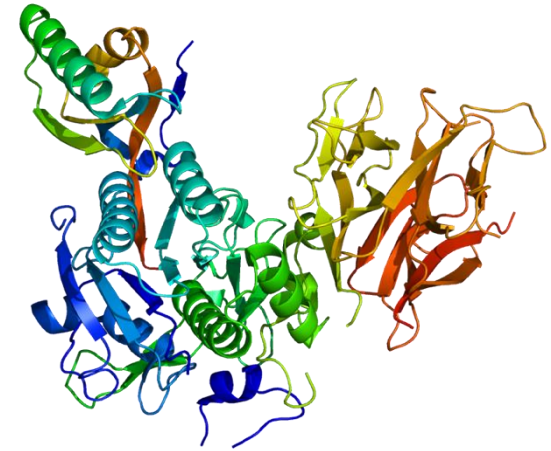
Before GWAS:

Studying the genome...

Single SNP studies - PCSK9

PCSK9 gene on chr1 codes for a protein 692 amino acids long.

The enzyme binds to and degrades the low-density lipoprotein particles (LDL) receptor on liver cells. The receptor initiates the ingestion and destruction of LDL particles.



**Proprotein convertase
subtilisin/kexin type 9**

With fewer LDL receptors, less LDL is cleared from the bloodstream, leading to higher blood cholesterol levels.



Before GWAS:

Studying the genome...

A genetic variant in PCSK9 associated with cholesterol levels

2099 individuals in Finland

Carriers of T variant have lower LDL cholesterol levels than carriers of G

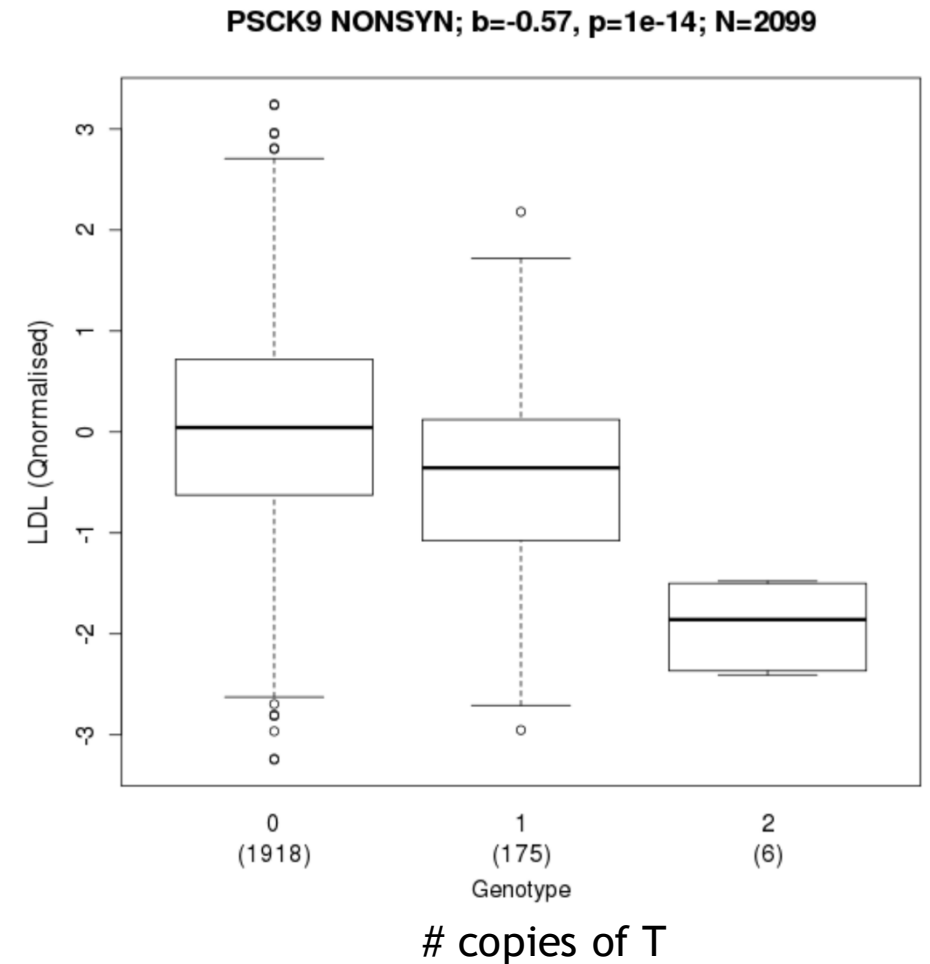
Importance? LDL is a strong risk factor of heart disease

Zhao et al. AJHG 2006 discovered that **knocking out PCSK9** could reduce LDL.

ARTICLE

Molecular Characterization of Loss-of-Function Mutations in *PCSK9* and Identification of a Compound Heterozygote

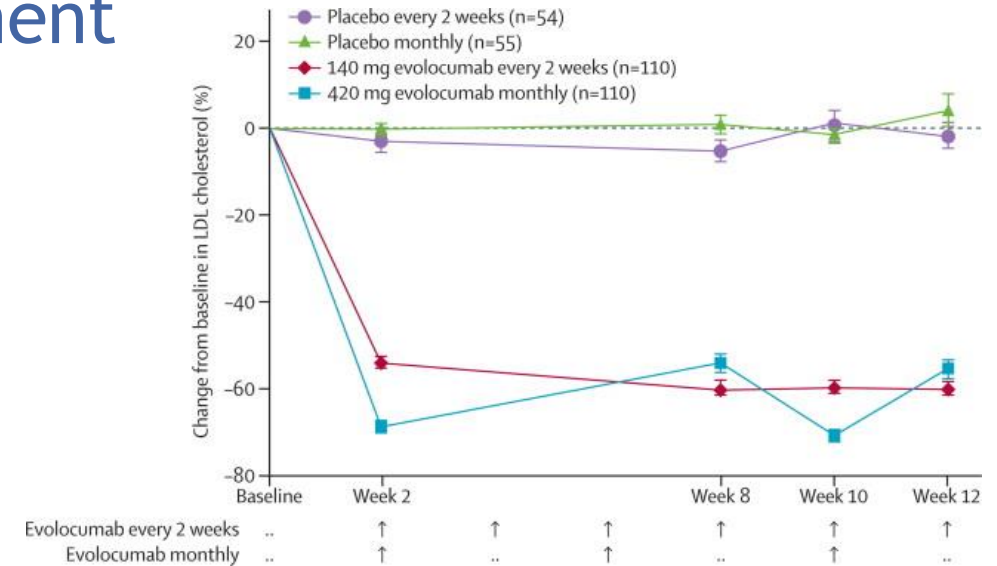
Zhenze Zhao,* Yetsa Tuakli-Wosornu,* Thomas A. Lagace, Lisa Kinch, Nicholas V. Grishin, Jay D. Horton, Jonathan C. Cohen, and Helen H. Hobbs



Applications: From genomics to treatment

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

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



ORIGINAL ARTICLE



Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease

Authors: 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., [+5](#), for the FOURIER Steering Committee and Investigators* [Author Info & Affiliations](#)

Published May 4, 2017 | N Engl J Med 2017;376:1713-1722

DOI: 10.1056/NEJMoa1615664 | VOL. 376 NO. 18 | Copyright © 2017

FDA Approves Amgen's Repatha® (evolocumab) To Prevent Heart Attack And Stroke

Evolocumab 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



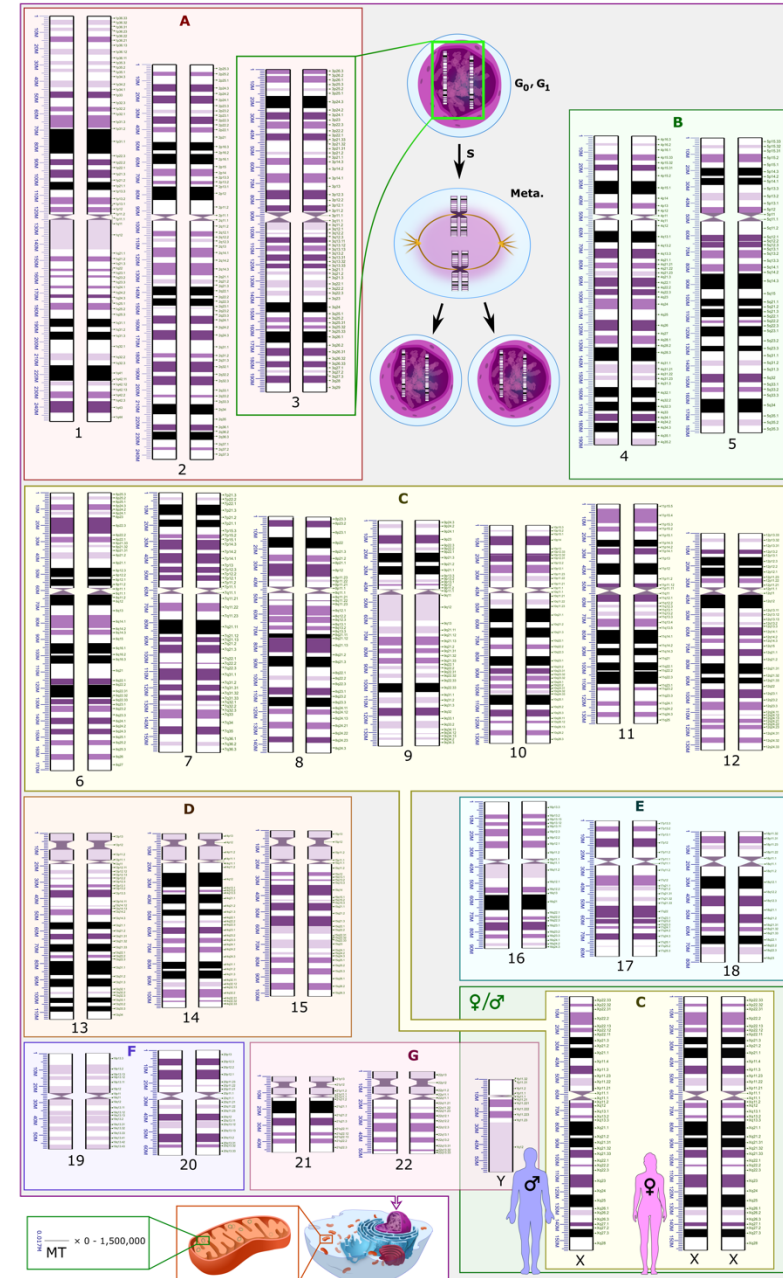
The human genome

Sequence of 3.2 billion nucleotides
(or base pairs: { A, C, G, T })

It is an organism's **complete set of
genome instructions**

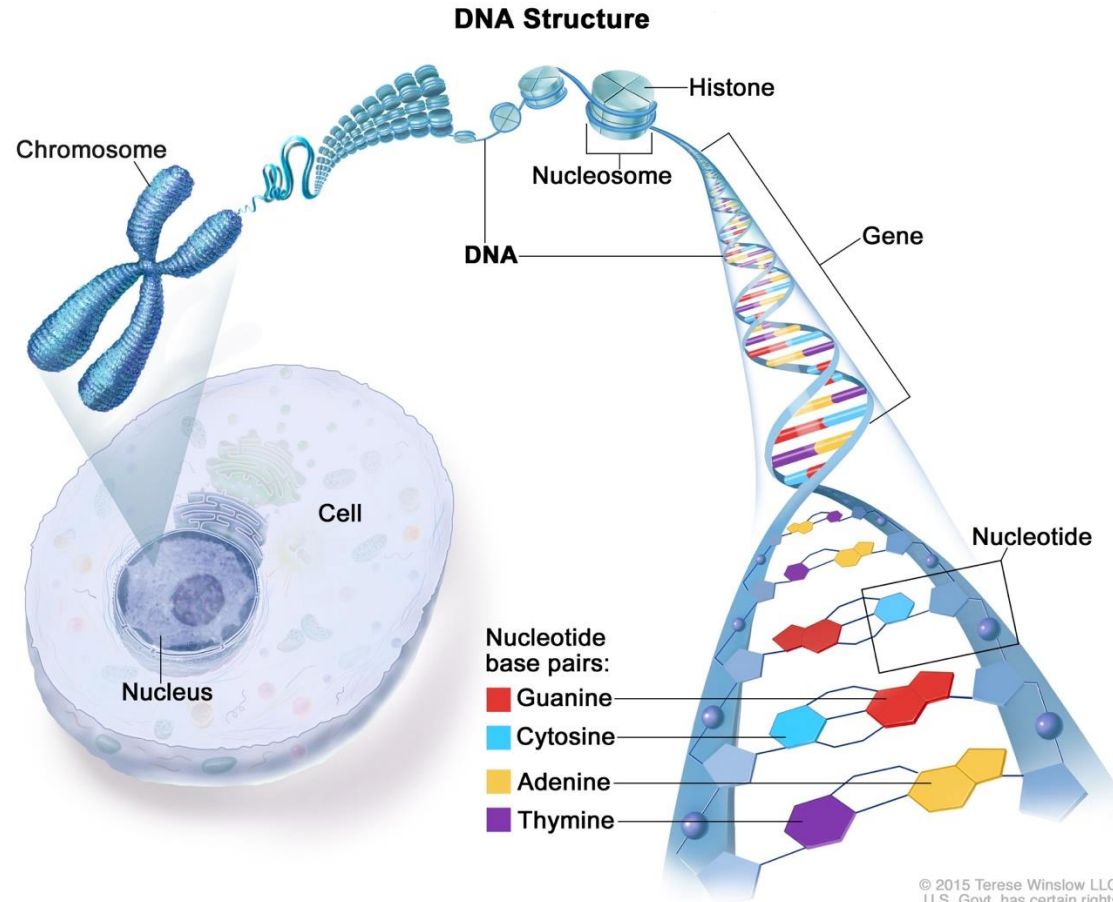
Two copies: maternal and paternal

Genome is physically divided into 22
pairs of autosomal chromosomes and
1 pair of sex chromosomes (males XY,
female XX)



General concepts

DNA structure



- **DNA is in the nucleus, forming chromosomes.**
- **Chromosomes have histones that bind to DNA.**
- **DNA is a double helix (spiral ladder shape).**
- **Made up of four bases: A, T, G, C (A-T, G-C pairing).**
- **Genes = DNA segments that code for proteins.**



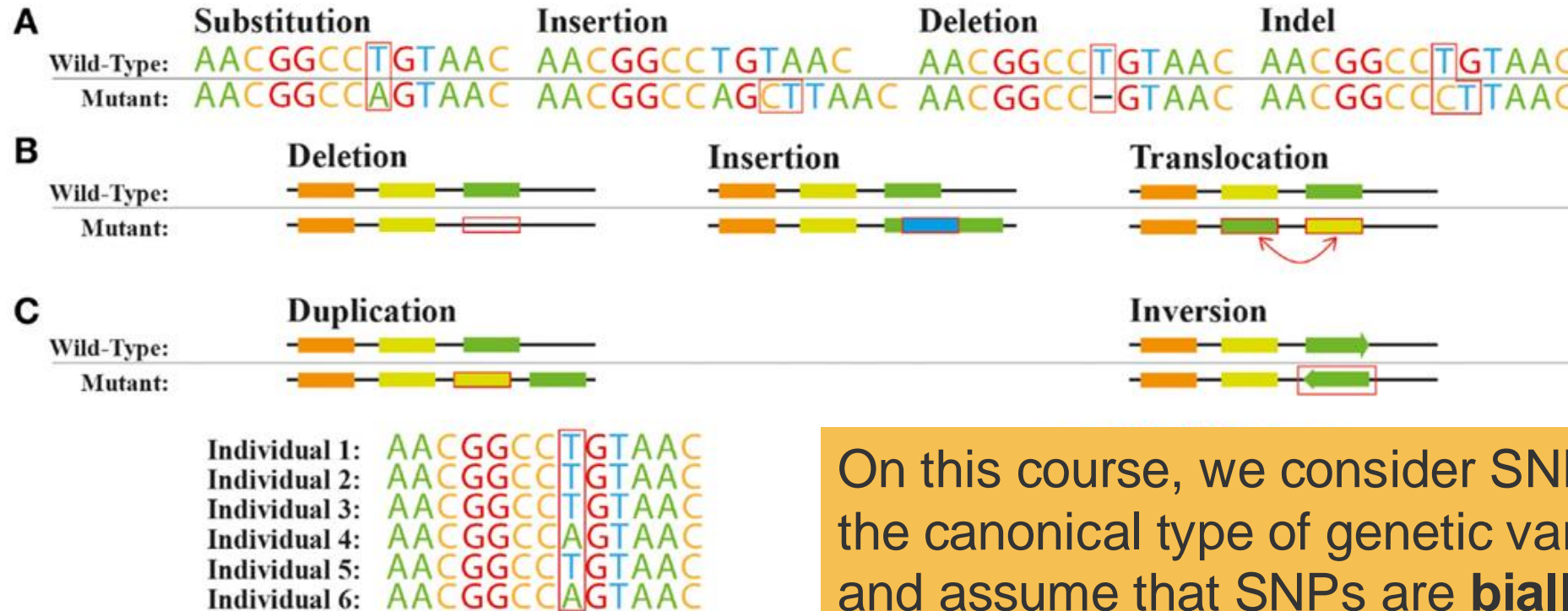
General concepts

Genetic variation



General concepts

Types of variation



On this course, we consider SNPs as the canonical type of genetic variation and assume that SNPs are **biallelic** (only 2 alleles present in the population)



General concepts

Single nucleotide polymorphism (SNP)

On average, 1 in 300 positions in genome exhibit **common variation** (MAF>5% or 1%) in the population; these are called “SNPs”.

Rare variant if MAF< 0.01%

Genomes in population

... G C G T T ...	96%
... G C T T T ...	4 %

Genotype at this SNP in population

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

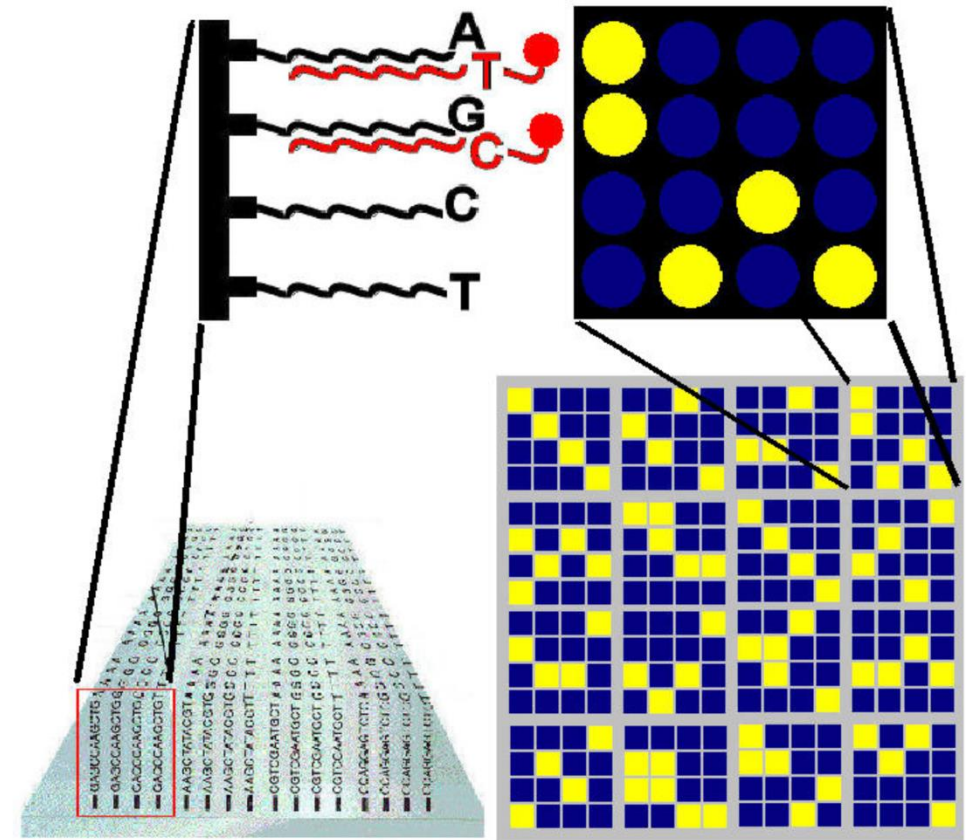
Only forward
strand of
genomes is
shown here

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



Reading SNPs

- Human SNP array can measure 10^6 SNPs reliably (predefined set)
- Fairly cheap: cost per individual ~30 euros -> key to making GWAS possible



Principle of a DNA microarray chip: use as Variant Detector Arrays (VDAs)

(after SM Carr *et al.* 2008. Comp Biochem Physiol D, 3:11)

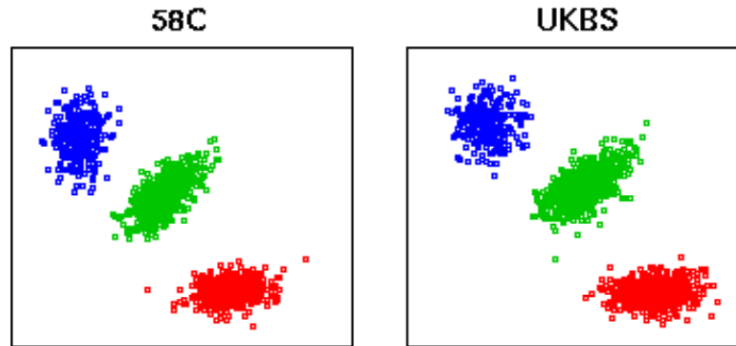
https://www.mun.ca/biology/scarr/DNA_Chips.html



General concepts

Genotype calling from SNP array data

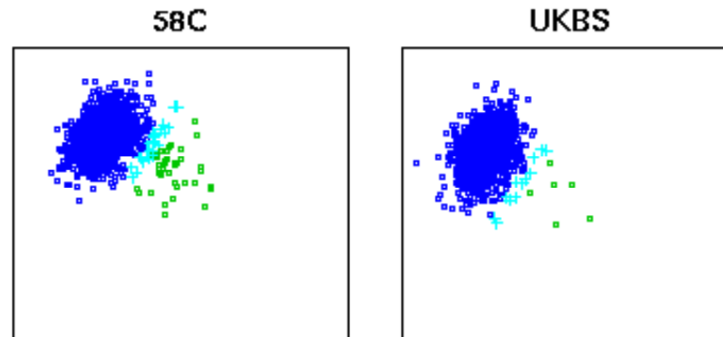
Example: variant rs6540301



Dark blue and red
are homozygotes
for one allele
Green is
heterozygote

The calling algorithm tries to
find the three genotype
clusters.
<- Good calling

Example: variant RS727641



Light blue means algorithm has
made no call. ERROR, rare
variant has less than 3 clusters.

There are other cases when the algorithm
performs the wrong call (violations of HWE)

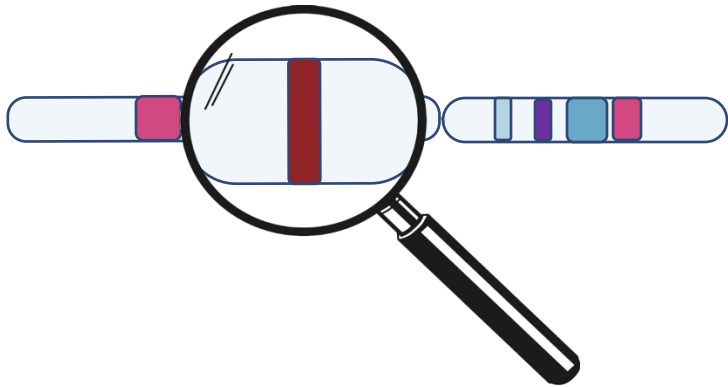


General concepts

Traits of interest

Monogenic trait

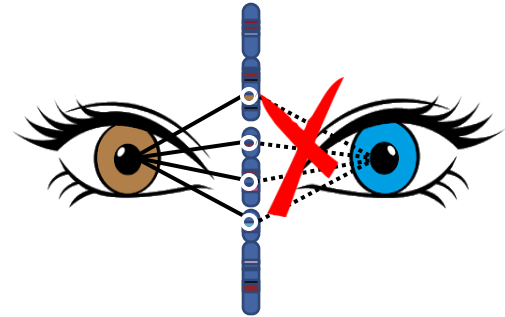
- Single variant
- 1 base change == disease
- Relatively easy to detect



Polygenic trait

- Thousands of variants with small effects
- Probability of the disease or having the trait (high/low)
- Hard(er) to detect

↓
Complex traits
(environmental factors)

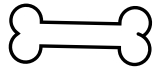


Many complex traits are highly polygenic

Continuous traits

Anthropometric traits

BMI
Height
Hip circumference
Waist circumference
Waist-to-hip ratio
Age at menarche



Cholesterol traits

HDL cholesterol
LDL cholesterol
Total cholesterol
Triglycerides



Early growth traits

Child birth length
Child birth weight
Childhood obesity
Infant head circumference

Intelligence, cognitive ability, and educational attainment traits

Childhood IQ
Cognitive performance
Intelligence
Years of schooling

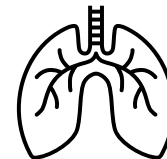
Diseases

Common chronic diseases

Asthma
Alzheimer's disease
Type 2 diabetes
Coronary artery disease

Inflammatory bowel diseases

Crohn's disease
Inflammatory bowel disease
Ulcerative colitis

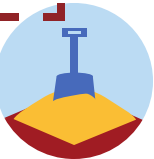
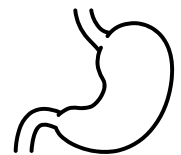


Psychiatric diseases and brain conditions

Autism spectrum disorder
Bipolar disorder
Major depressive disorder
Schizophrenia

Other

College completion
Rheumatoid arthritis



Genome-wide association study

Statistical problem: is a genetic variation at a particular position associated with observed phenotypic variation?



- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body measurement
- Cardiovascular measurement

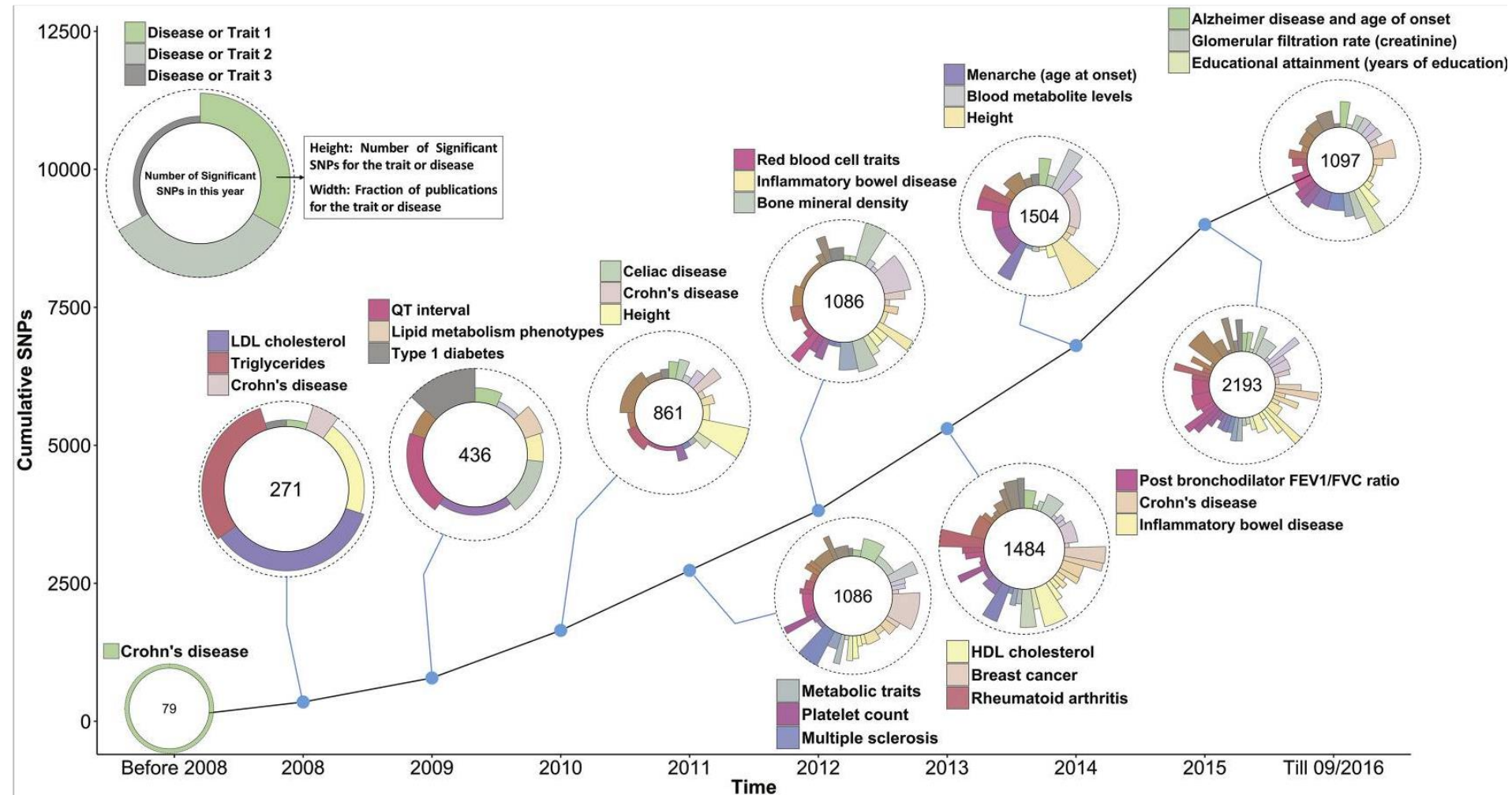


- Immune system disease
- Inflammatory marker measurement
- Cholesterol levels
- ...



GWAS

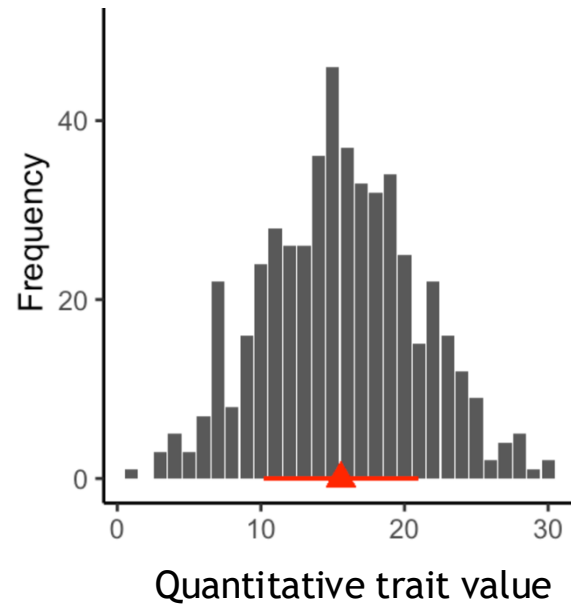
SNP-Trait Discovery Timeline



GWAS

Primary types of GWAS

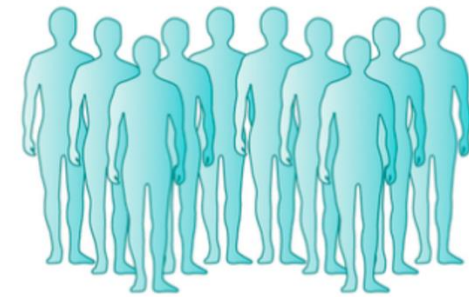
Quantitative trait-based GWAS



Disease trait-based GWAS



Cases



Controls

Let's next look at two examples GWAS



Example 1: Quantitative GWAS

Body-mass index (BMI)

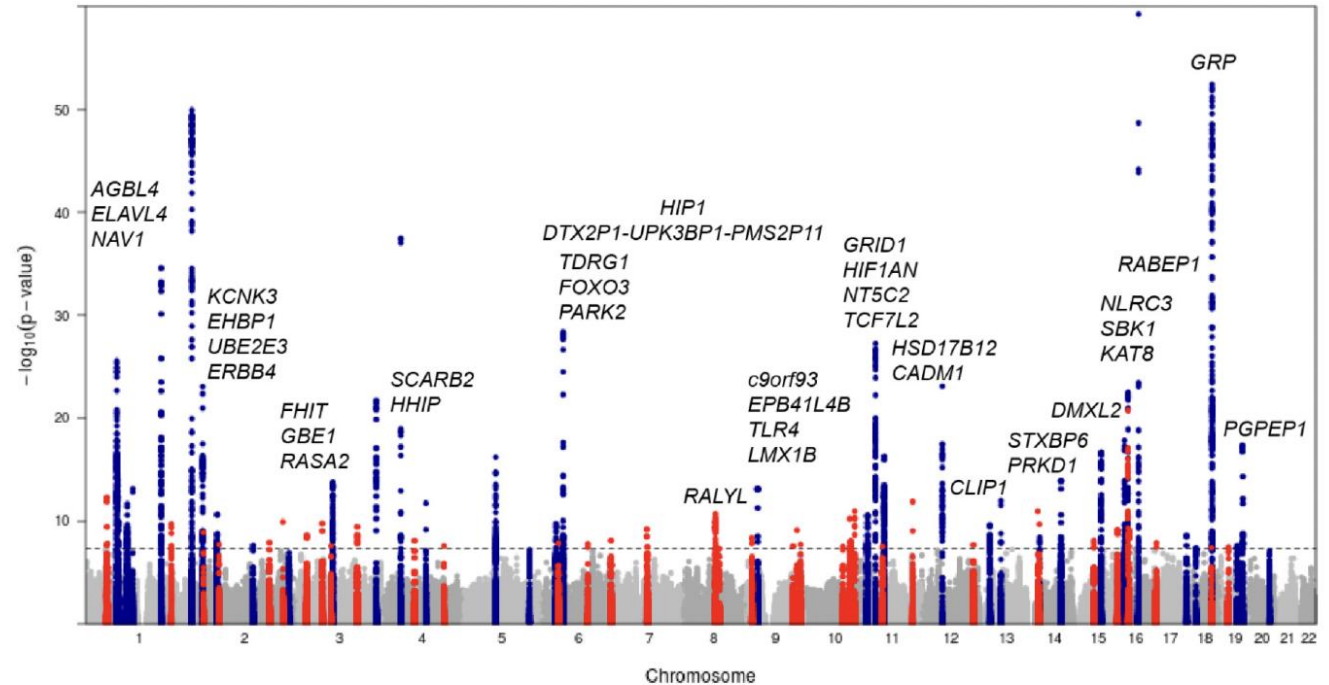
> 300,000 individuals from 125 cohorts (studies)

97 loci associated (genomic regions)

Each locus is a hint to biology of BMI

Genes involved in energy metabolism, lipid biology, insulin secretion and adipogenesis

Results highlight role of central nervous system in BMI



Manhattan plot shows $-\log_{10}$ P-value of each SNP tested in GWAS. Previously known loci are in blue, new findings are in red.

Suppl. Fig 1, Locke et al. 2015



Example 1: Quantitative GWAS

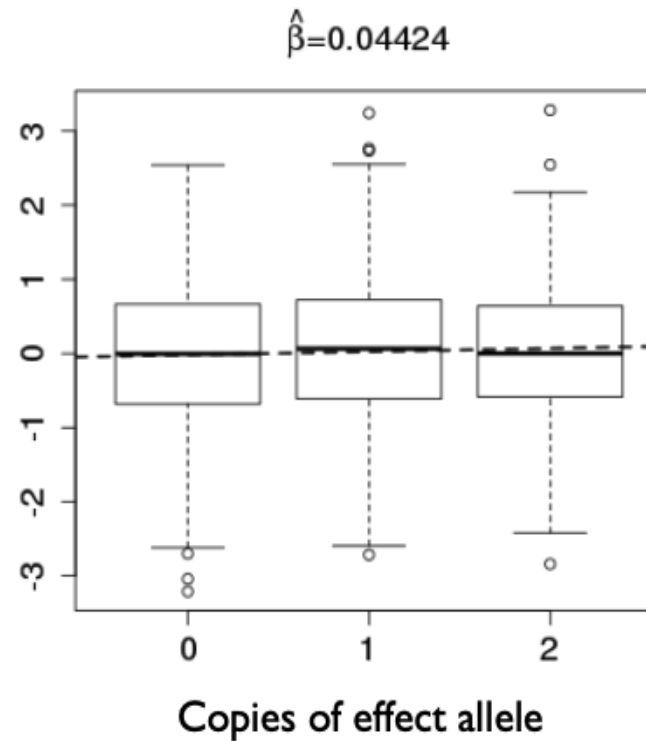
Body-mass index (BMI)

Association tests at **2.5 million SNPs**

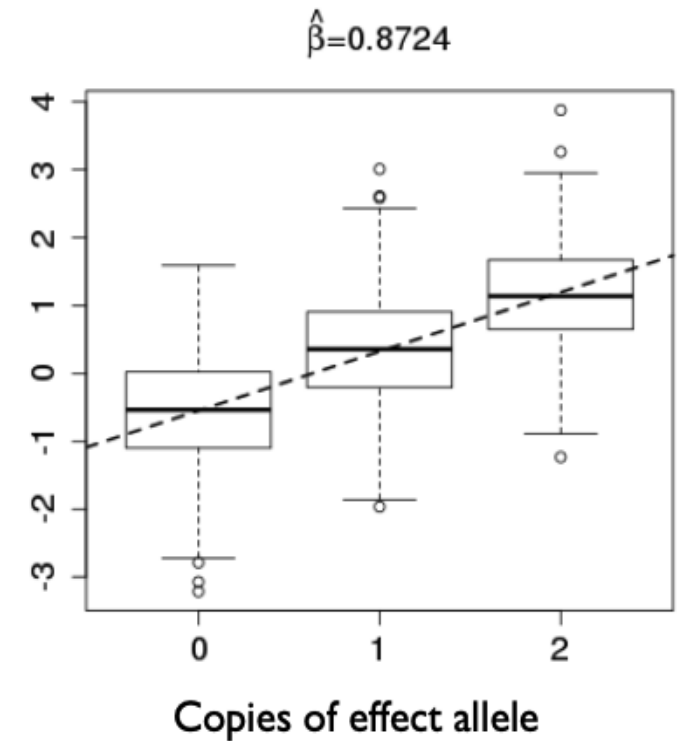
For each test, does the BMI differ between genotype groups?

Linear regression slope, β
(SE and P-value)

SNP with no association



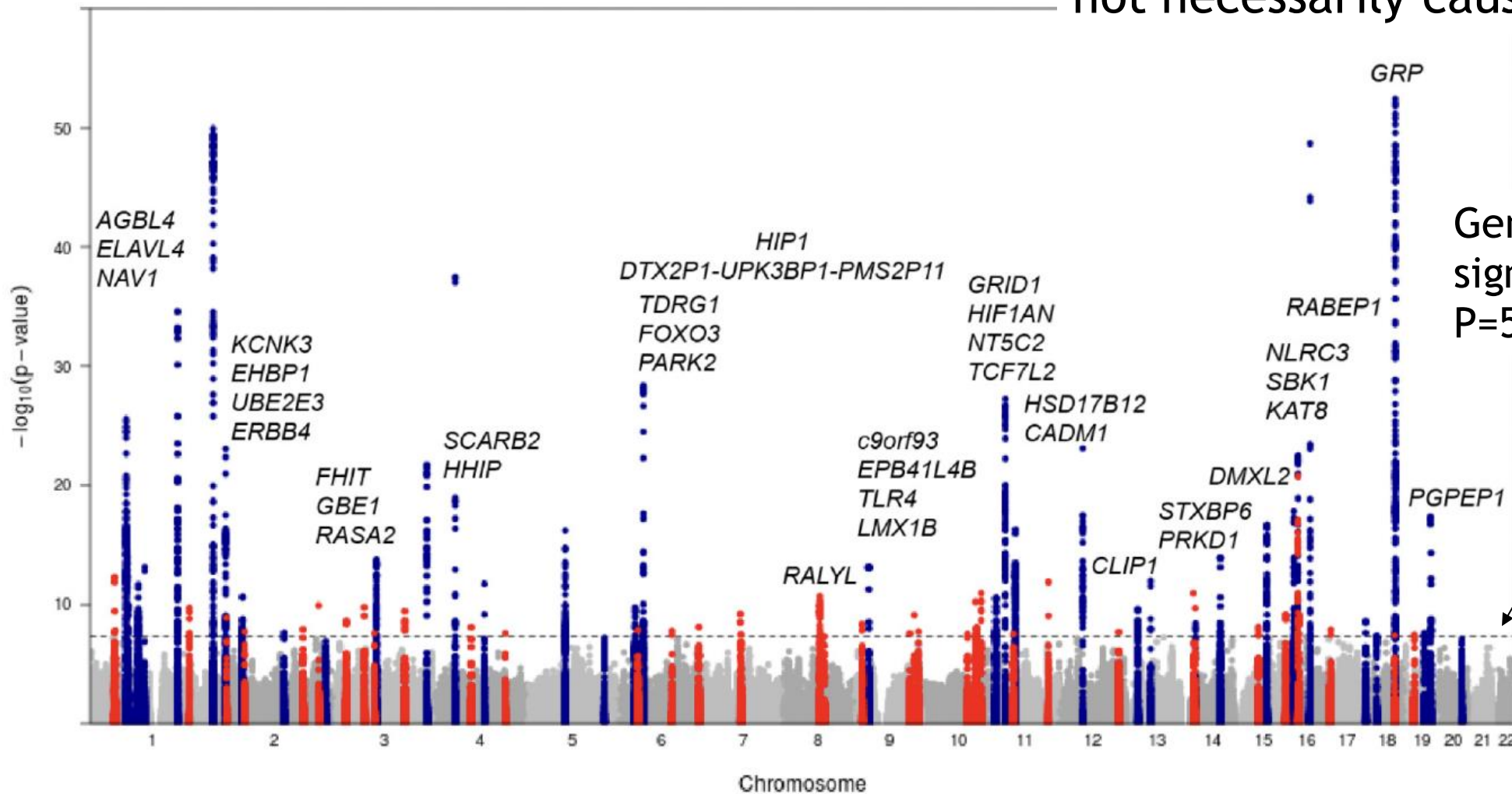
SNP with clear association



Example 1: Quantitative GWAS

Body-mass index (BMI)

Each locus is labelled by a nearby gene (but that gene is not necessarily causal.)

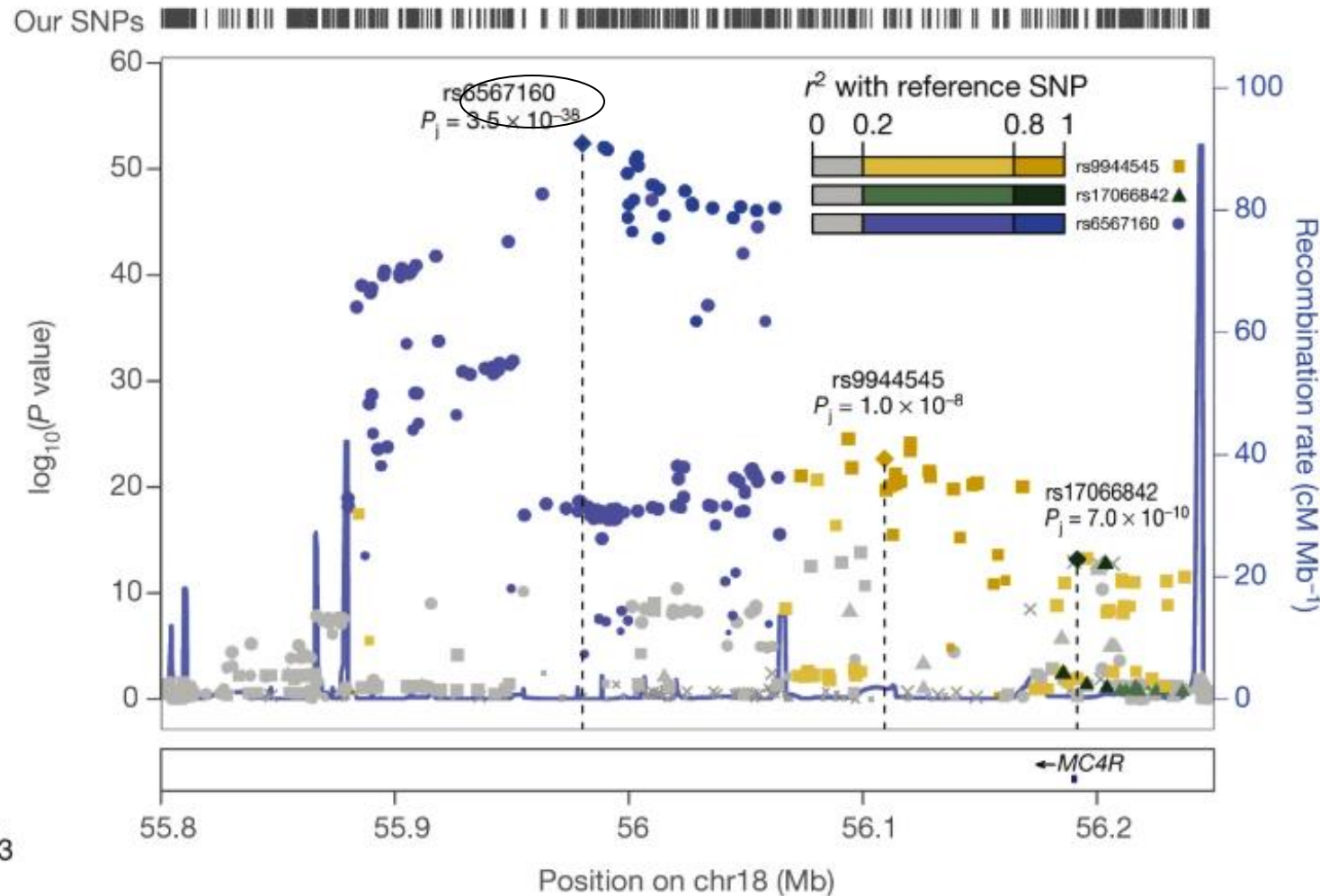


Genome-wide significance level at $P=5e^{-8}$ or $-\log_{10}(P) = 7.3$.



Example 1: Quantitative GWAS

Zooming into one associated region



Many SNPs show strong association; not clear which are causal ones.

What does each variant do? Change a protein? The gene expression in certain condition?



Example 1: Quantitative GWAS

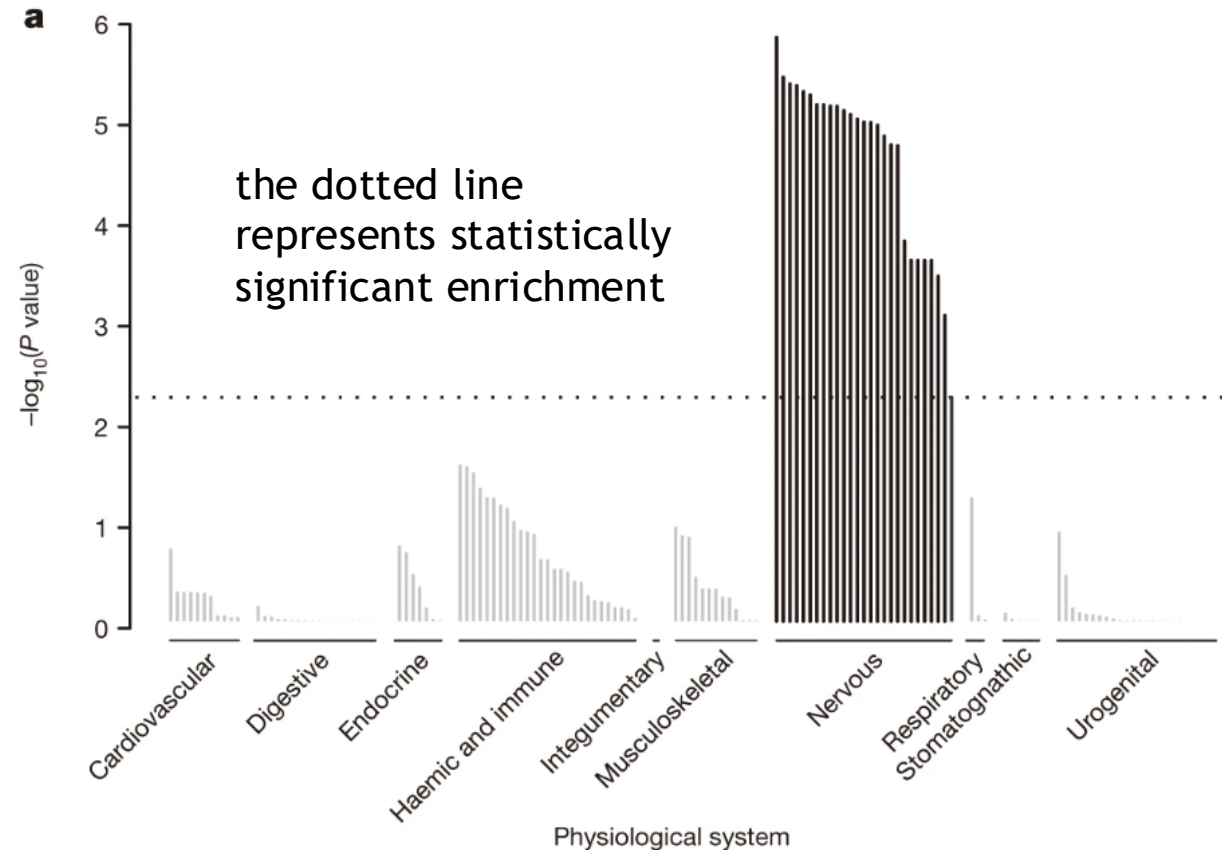
Looking for patterns

Results highlight role of central nervous system in BMI

How? By combining signals across the genome.

Does the significantly associated variation tend to be near genes? Regulatory regions?

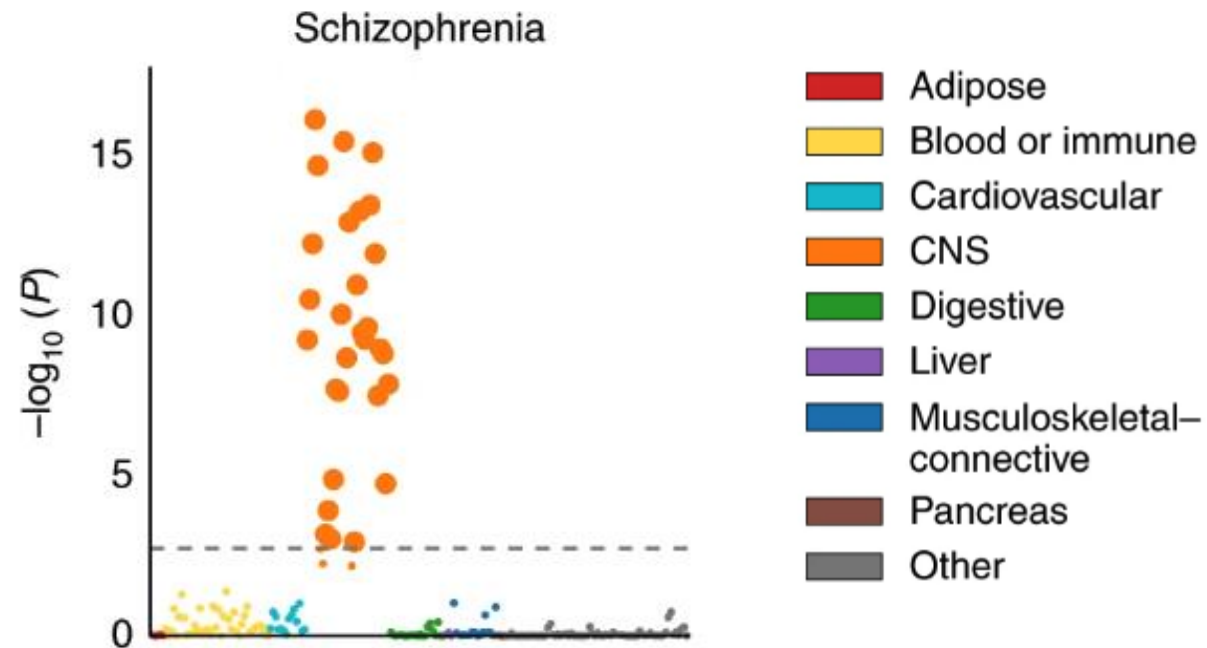
Are the GWAS signals in/near genes enriched for expression in a particular tissue/cell



DEPICT predicts genes within BMI-associated loci ($P < 5 \times 10^{-4}$) are enriched for expression in the brain and central nervous system



Testing multiple phenotypes for tissue-specific enrichment across many tissues



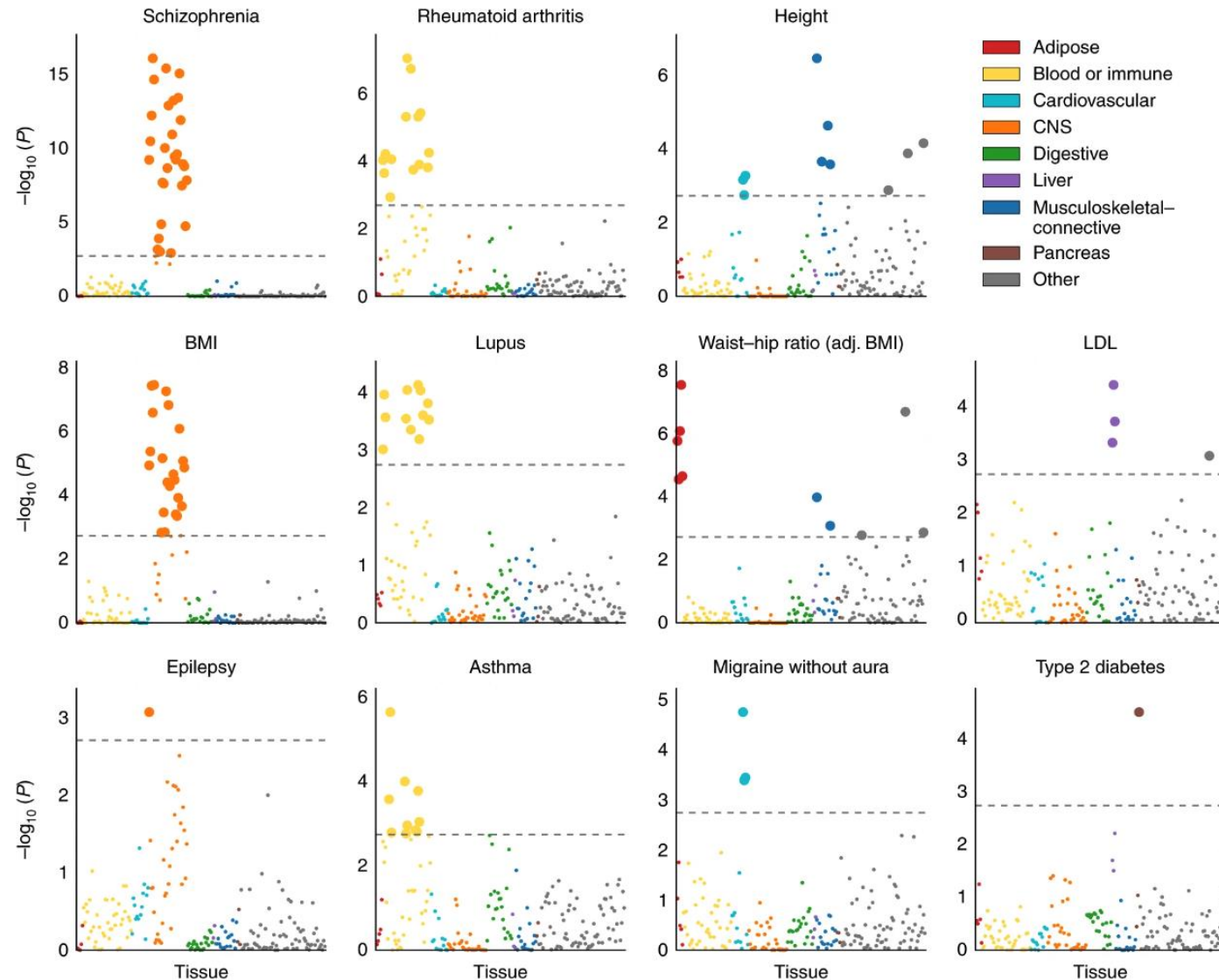
$-\log_{10}(P)$ is of the association between the trait (title) and the tissue/cell type (see legend)

Large circles pass the cutoff of $FDR < 5\%$ at $-\log_{10}(P) = 2.75$

Finucane et al. 2018 Nat Gen



Testing multiple phenotypes for tissue-specific enrichment across many tissues



$-\log_{10}(P)$ is of the association between the trait (title) and the tissue/cell type (see legend)

Large circles pass the cutoff of FDR < 5% at $-\log_{10}(P) = 2.75$

Finucane et al. 2018 Nat Gen



Example 1: Quantitative GWAS

Meta-analysis

- Locke et al. gathered data from 125 cohorts!
- A meta-analysis is a statistical analysis that combines the results of multiple scientific studies on the same question.
- It works on GWAS results, not requiring original genotype-phenotype data.
- While no-one has access to all original genotype-phenotype data, everyone can access the meta-analysed GWAS results as they are (often) publicly available.

GWAS output

Summary statistics

Variant identifier "chr:pos:ref:alt", where "ref" is aligned to the forward strand of GRCh37 and "alt" is the effect allele.

variant	minor_allele	minor_AF	low_confidence _variant	n_complete_sa mples	AC	ytx	beta	se	tstat	pval
1:69487:G:A	A	2,15E+00	TRUE	23483	1,01E+05	-9,18E+04	-8,33E+04	9,97E+04	-8,35E+04	4,04E+04
1:69569:T:C	C	1,36E+01	TRUE	23483	6,38E+05	-2,01E+05	-2,99E+04	4,10E+04	-7,29E+04	4,66E+04
1:139853:C:T	T	2,13E+00	TRUE	23483	1,00E+05	-9,13E+04	-8,28E+04	9,97E+04	-8,30E+04	4,06E+04
1:692794:CA:C	C	1,13E+04	FALSE	23483	5,33E+08	-1,69E+06	-2,59E+02	1,60E+03	-1,62E+04	8,71E+04
1:693731:A:G	G	1,17E+04	FALSE	23483	5,48E+08	-1,79E+05	1,89E+02	1,53E+03	1,24E+04	9,02E+04
1:707522:G:C	C	9,87E+03	FALSE	23483	4,64E+08	3,58E+06	1,24E+03	1,71E+03	7,22E+04	4,70E+04
1:717587:G:A	A	1,58E+03	FALSE	23483	7,40E+07	2,33E+06	4,34E+03	4,10E+03	1,06E+05	2,90E+04
1:723329:A:T	T	2,01E+02	FALSE	23483	9,43E+06	-8,32E+05	-1,03E+04	1,11E+04	-9,34E+04	3,50E+04



Example 2: Disease GWAS

Migraine

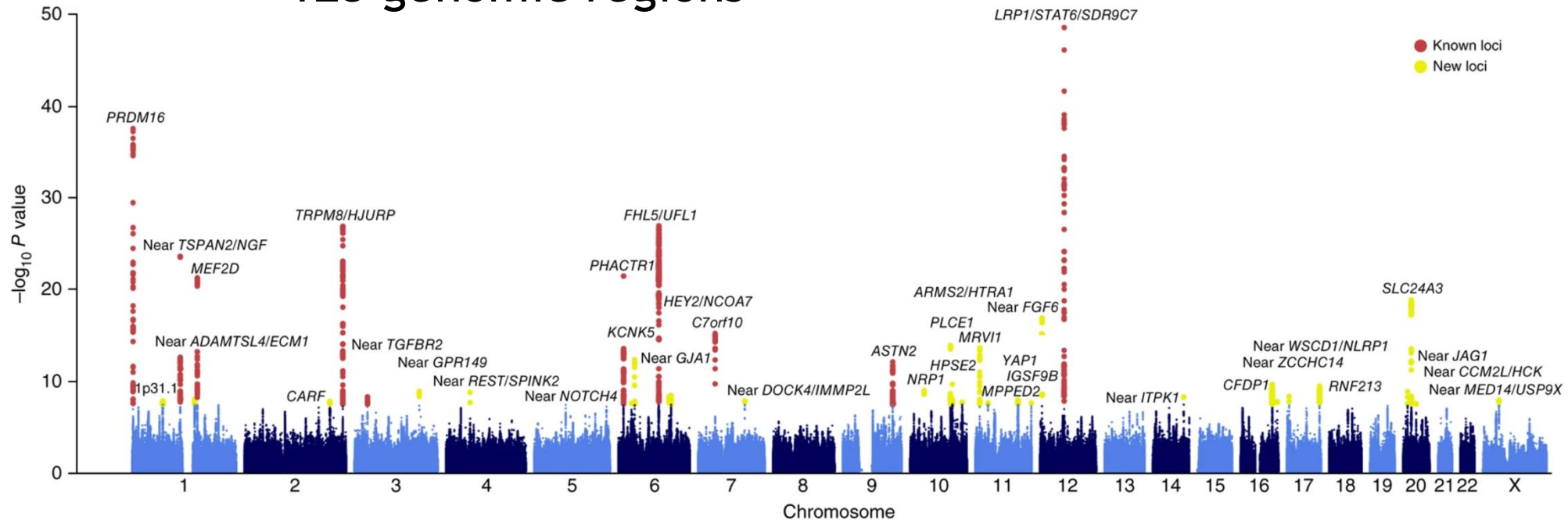
38 susceptibility loci

Gormely et al. 2016

102,000 cases and 771,000 controls

25 cohorts (meta-analysis)

123 genomic regions



Haukatangas et al 2022



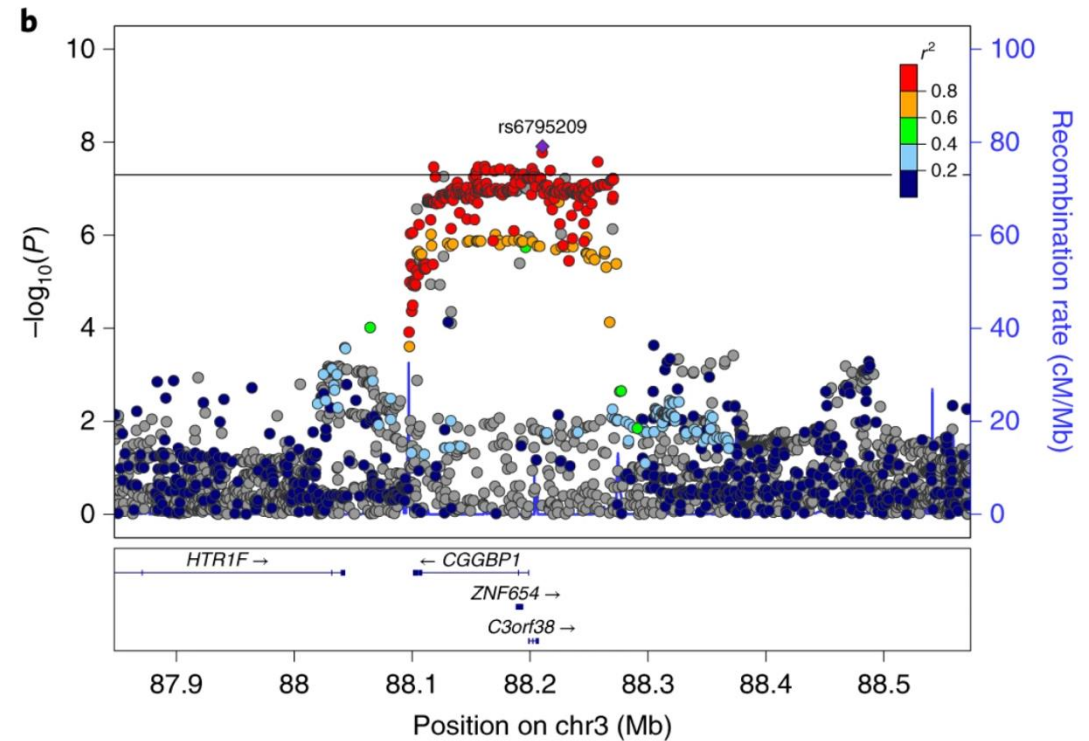
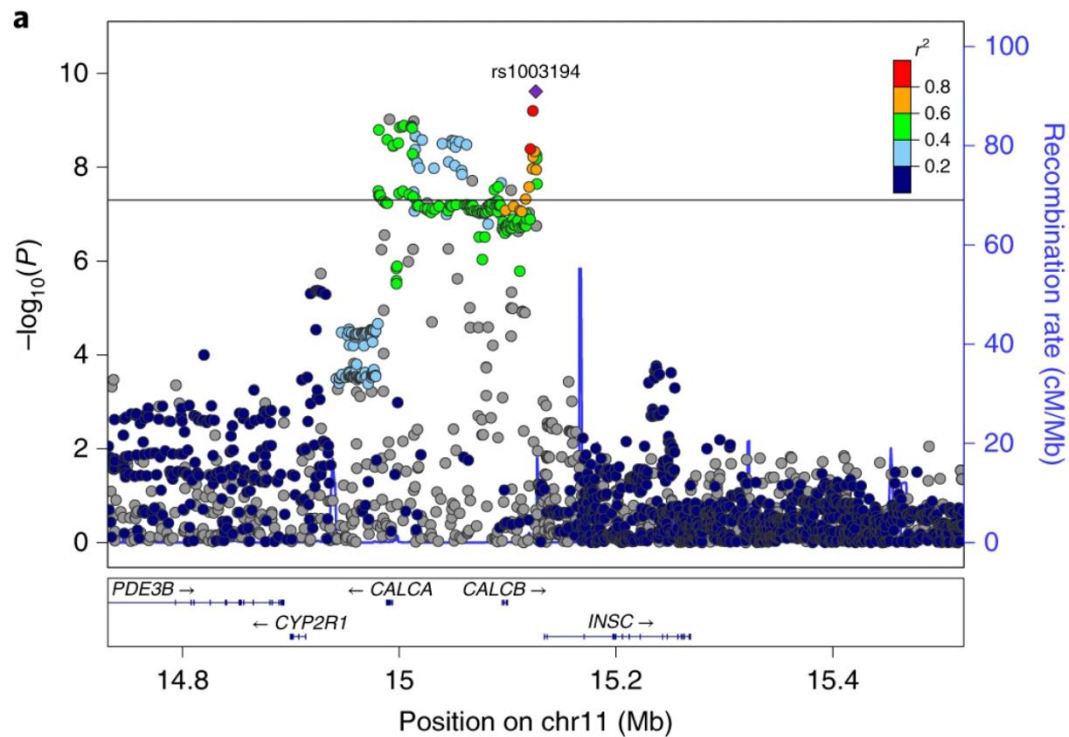
Example 2: Disease GWAS

Migraine, zoom-in regions

Haukatangas et al 2022

2 of the regions w. genes that are target of molecular therapies.

Genes involved in vascular system & central nervous system



Could other hits potentially become promising candidates for drug therapies?

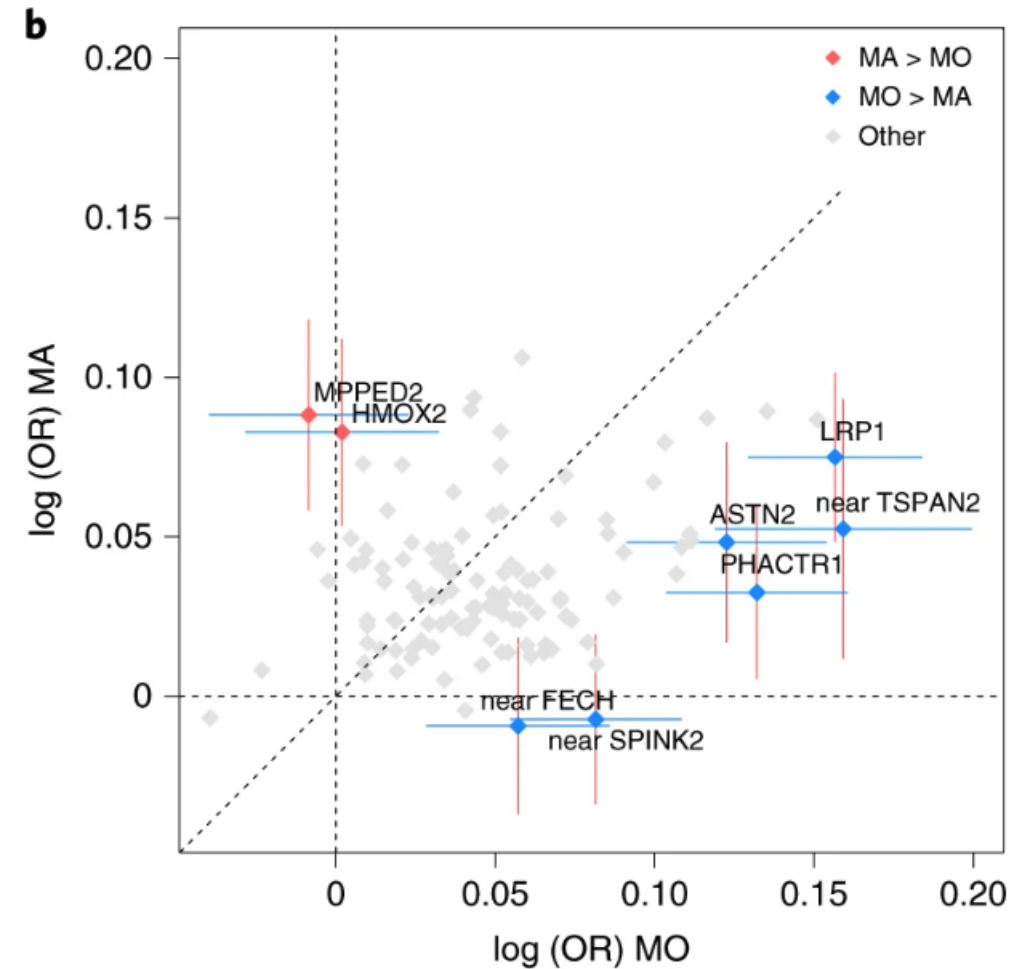


Example 2: Disease GWAS

Migraine, OR

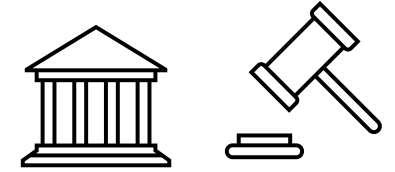
Effect sizes of migraine risk alleles (OR)

- MA: migraine with aura
- MO: migraine without aura
- Is the biology of different migraine subtypes same/different?
- Could GWAS help finding better treatments?
- Could the same allele that reduces the risk of one disease increase the risk of some other disorder?



Ethical aspects

Considerations for genetic data



Access: Who can access genetic data (e.g., individuals, researchers, medical professionals)?

Information to Return: Should individuals/relatives receive actionable health risks, sensitive traits (e.g., IQ), or unexpected ancestry findings?

Gene Editing: Is it ethical for curing diseases, preventing severe mutations, or designing favourable traits?

Data Responsibility: Genetic data contain deeply personal information and must be handled with care, requiring clear agreements on usage and purpose.

GWAS overview

GWAS types

Genotypic data:

- microarrays
- WGS
- WES

GWAS (genetic + phenotypic data) vs. meta-analysis (study cohorts)

