Genome-wide association studies and follow-up analyses

Dora Koller, Ph.D.
Postdoctoral Fellow, Cormand Lab
Department of Genetics, Microbiology and Statistics
Universitat de Barcelona
dorakoller@ub.edu, DoraKoller

19/10/2023

About me

Biology BSc
Budapest

2014

2016

Erasmus traineeship
Universitat Pompeu Fabra

Medical Biotechnology MSc
Budapest

2017

MARIE CURIE ACTIONS

2020

Clinical Pharmacology PhD
Madrid, Coimbra

Postdoctoral Fellow
Yale School of Medicine

ongoing

MARIE CURIE ACTIONS

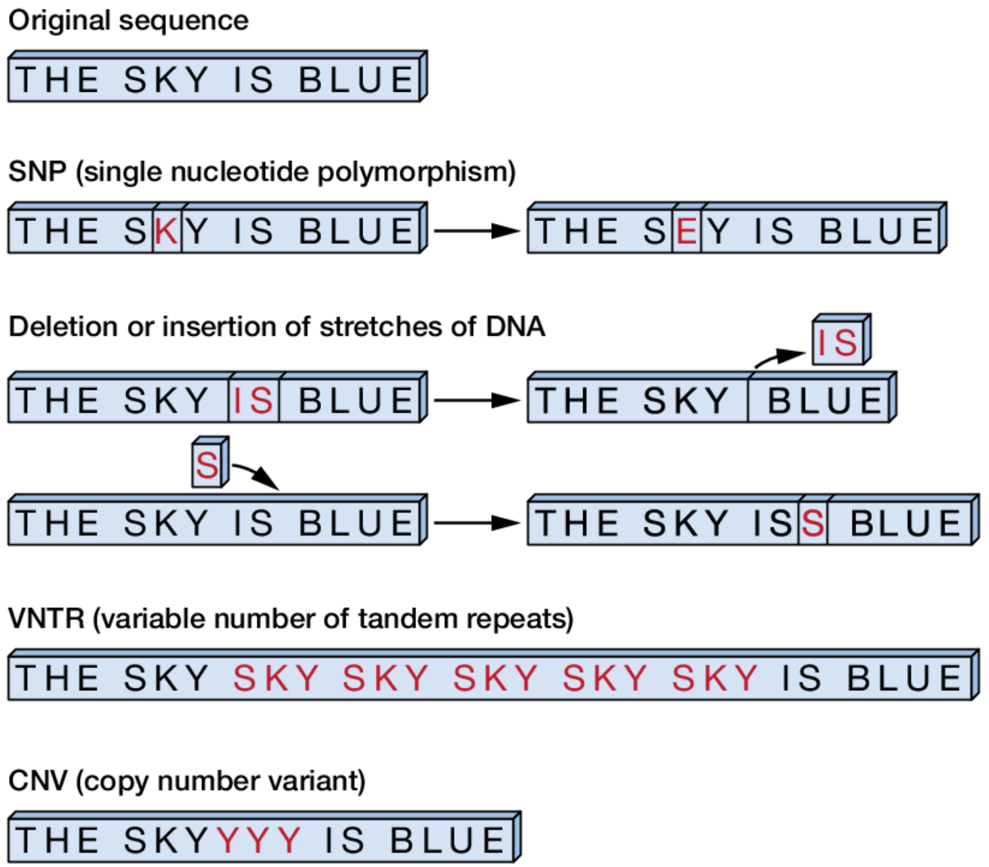
UNIVERSITAT DE BARCELONA

MARIE CURIE ACTIONS

ongoing

Postdoctoral Fellow
Universitat de Barcelona

Genetic variants



Original sequence

THE SKY IS BLUE

SNP (single nucleotide polymorphism)

THE SKY IS BLUE → THE SEY IS BLUE

Deletion or insertion of stretches of DNA

THE SKY IS BLUE → THE SKY BLUE

IS

S

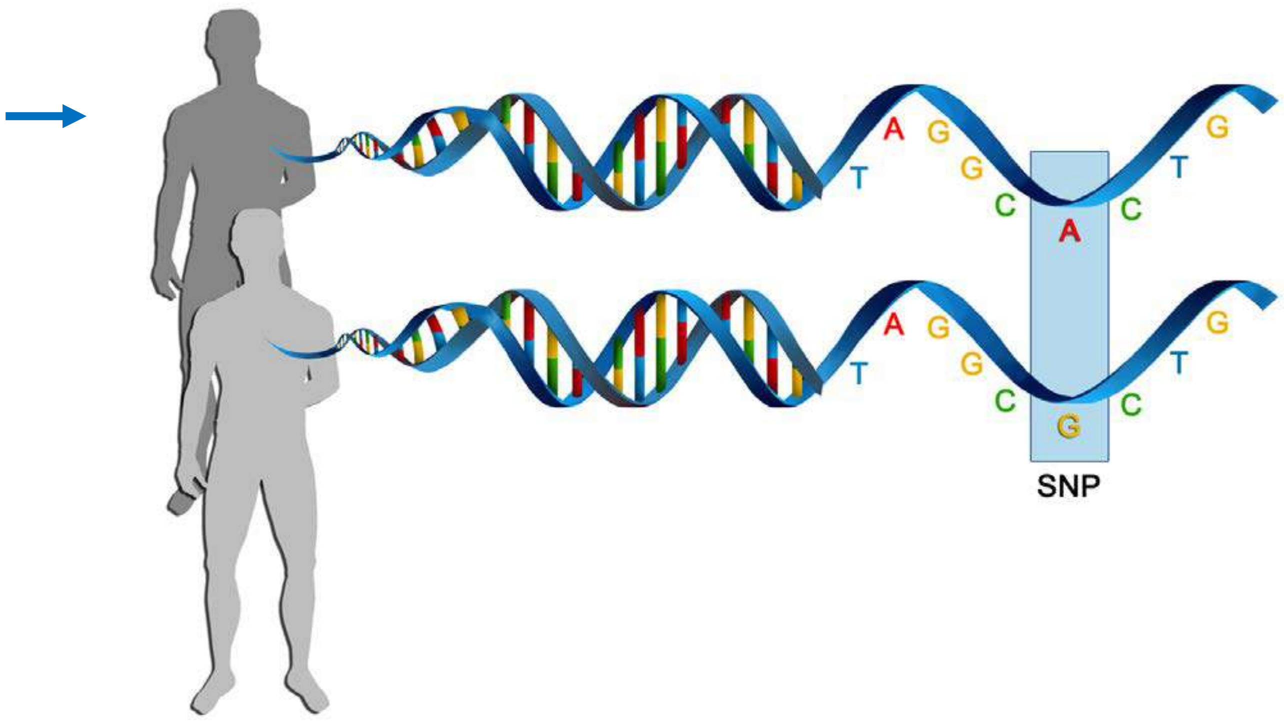
THE SKY IS BLUE → THE SKY ISS BLUE

VNTR (variable number of tandem repeats)

THE SKY SKY SKY SKY SKY SKY IS BLUE

CNV (copy number variant)

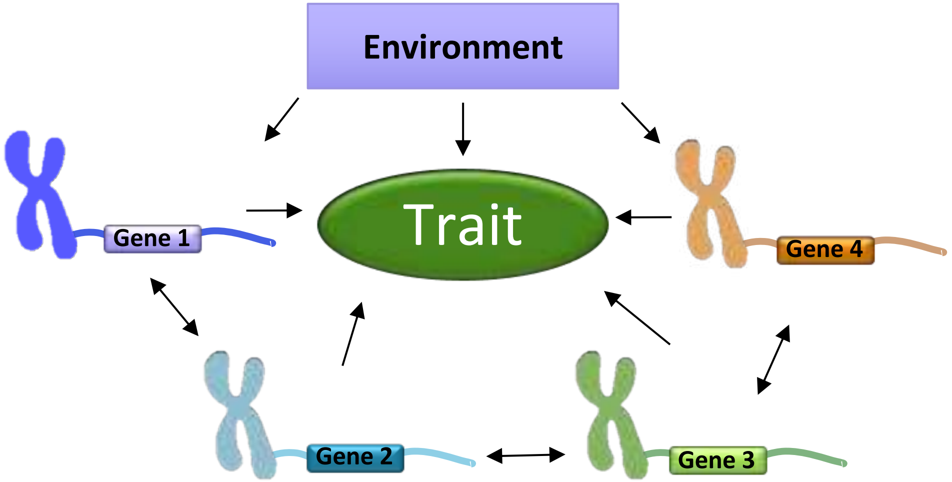
THE SKYYYY IS BLUE



Philibert et. al., Clin Epigenetics. 2014; 6(1): 28.

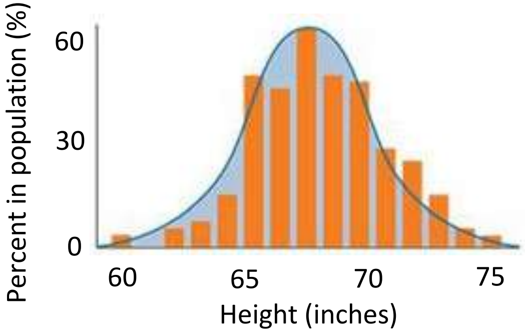
Complex traits

Complex (polygenic) traits

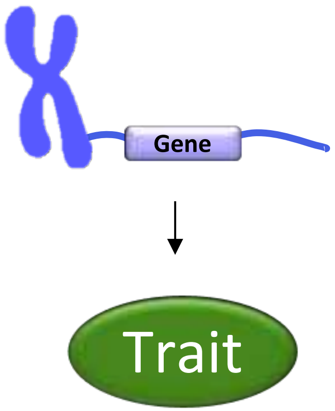


Examples

- Type 2 diabetes - metabolic
- Breast cancer - neoplasm
- LDL cholesterol - biomarker
- Schizophrenia - psychiatric

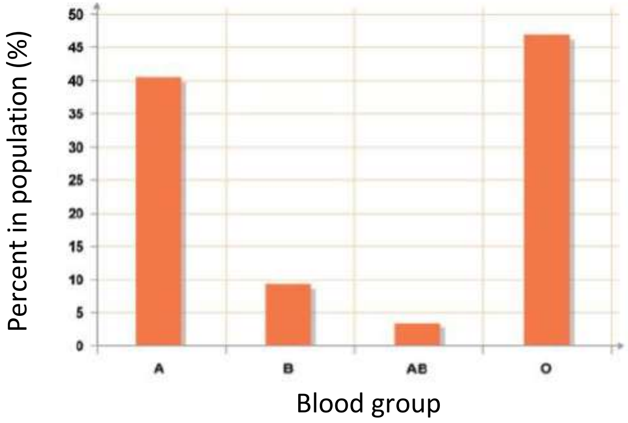


Mendelian (monogenic) traits



Examples

- Sickle cell anemia
- Cystic fibrosis
- Huntington disease
- Duchenne muscular dystrophy



Main approaches to investigate complex traits

Single genes

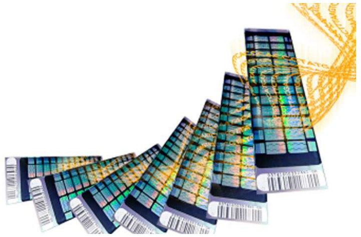
Genotyping of predetermined SNPs

One or a limited number of SNPs are measured in pre-specified genes.

"All" genes

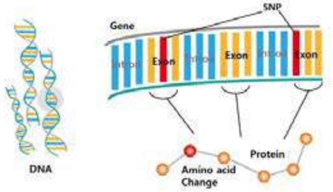
Genome-wide genotyping

A certain number of variants are directly measured, and millions are imputed using a reference panel (e.g., Haplotype Reference Consortium, 1000 Genomes).

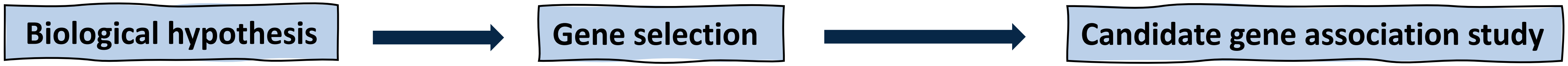


Whole exome sequencing

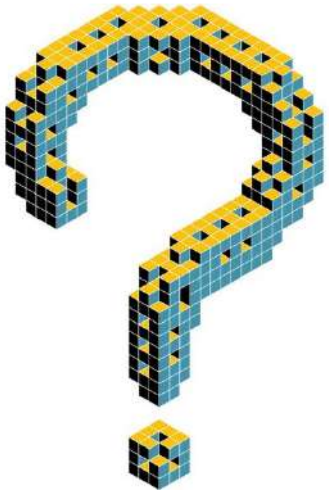
Sequencing regions of the genome (about 2%) that are involved in coding for proteins. Particularly suitable to detect structural variants, i.e., insertions, deletions, and CNVs.



Candidate gene studies



Biological hypothesis → Gene selection → Candidate gene association study

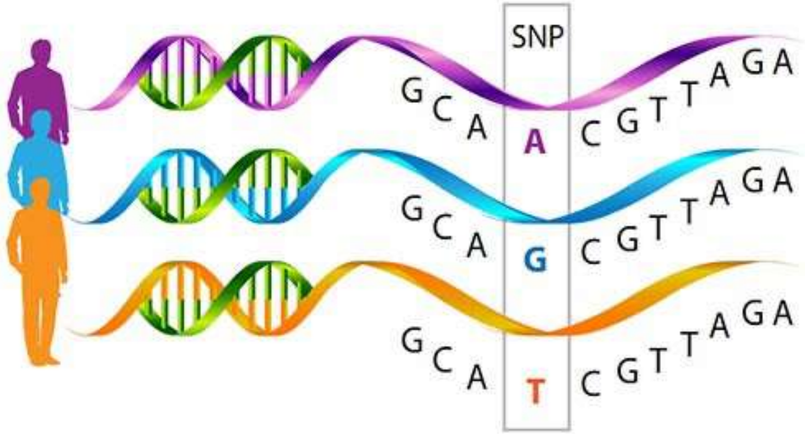


- ◉ Genomic databases

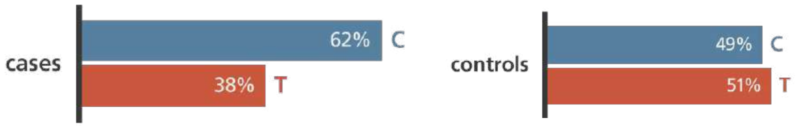
- ◉ Microarray data

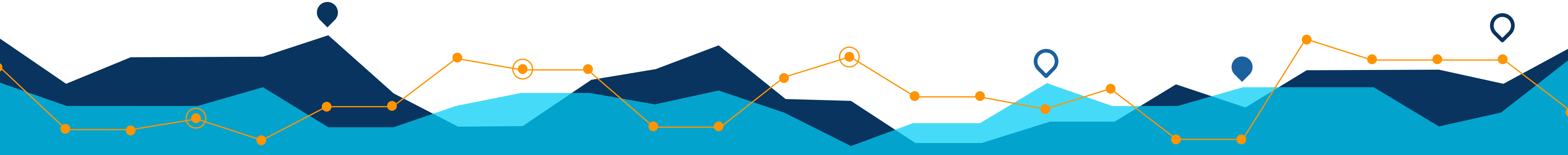
- ◉ Literature search

- ◉ Molecular evidence

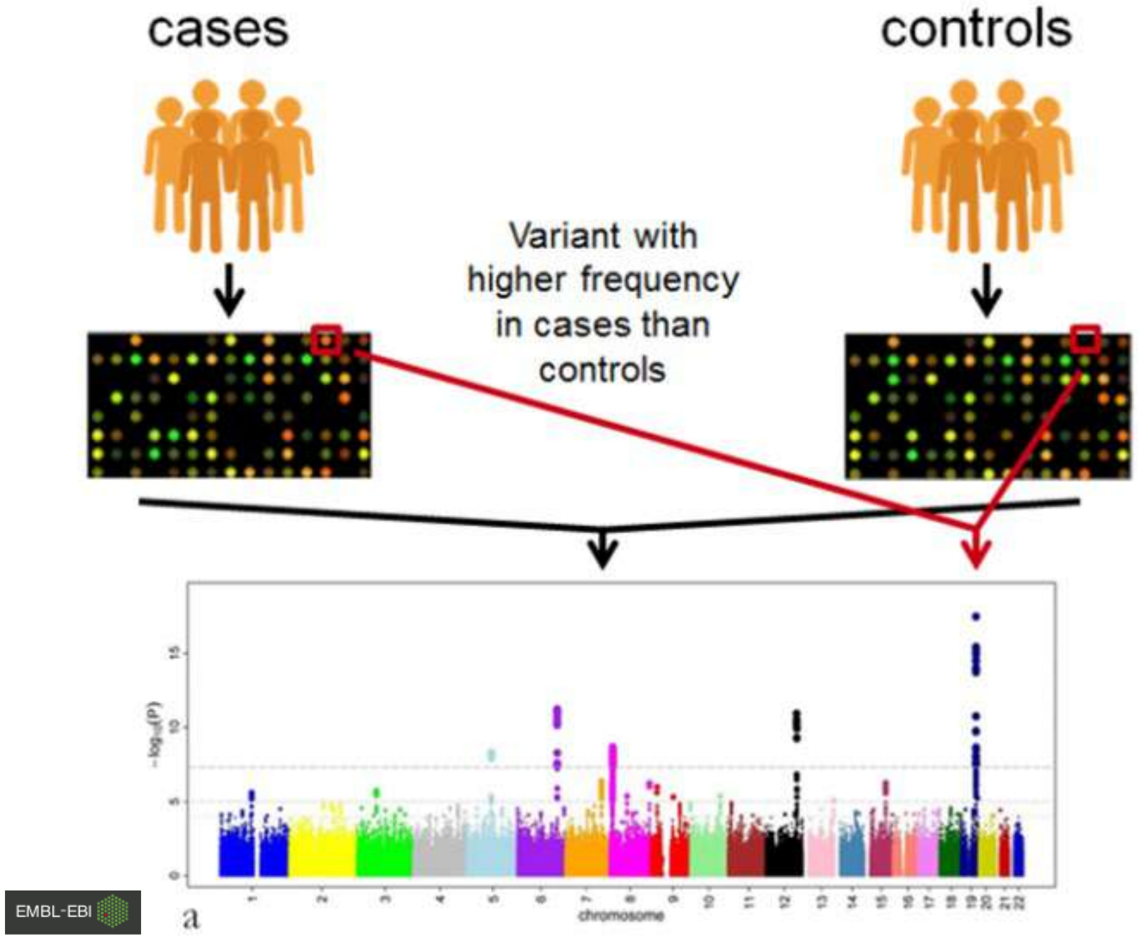


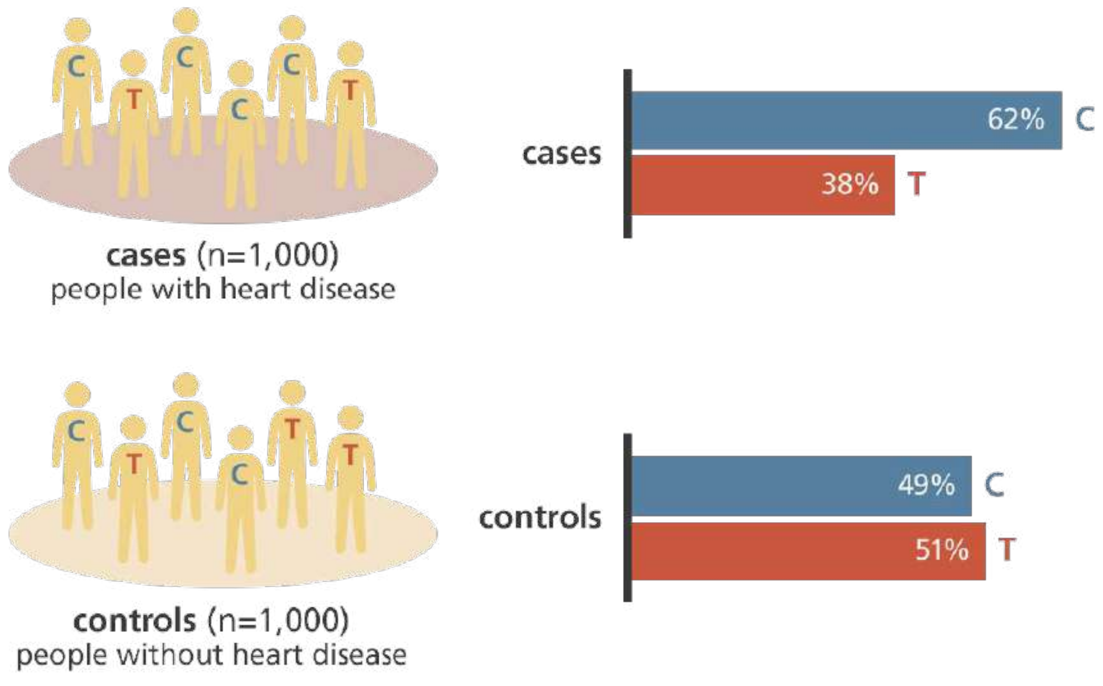
$p < 0.05$





Genome-wide association studies





Genome Research Limited

Types of GWAS

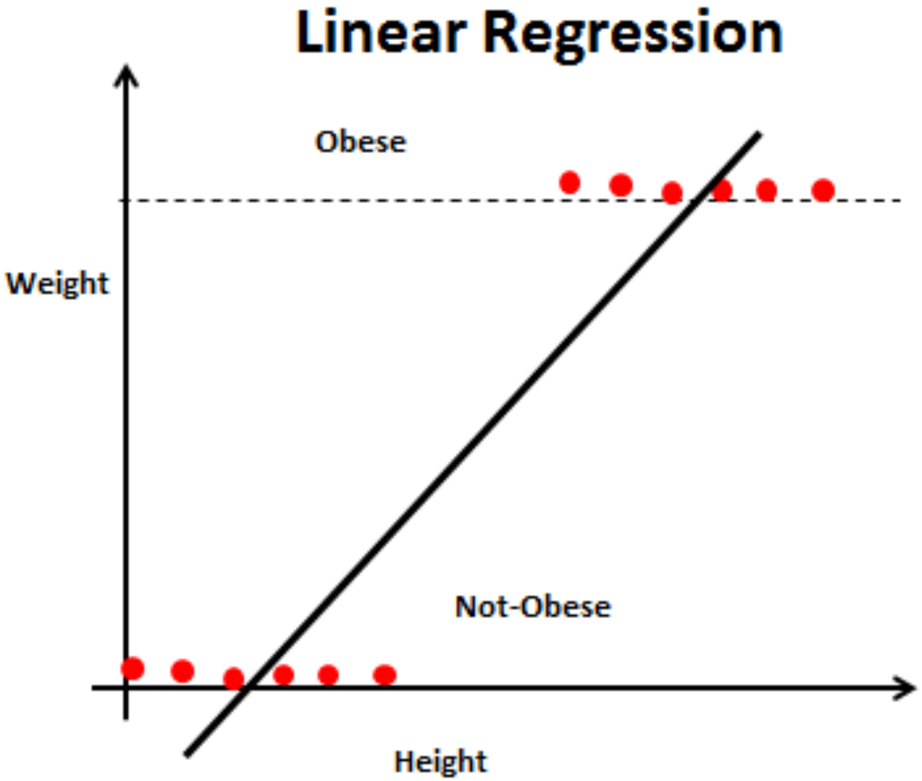
Linear regression

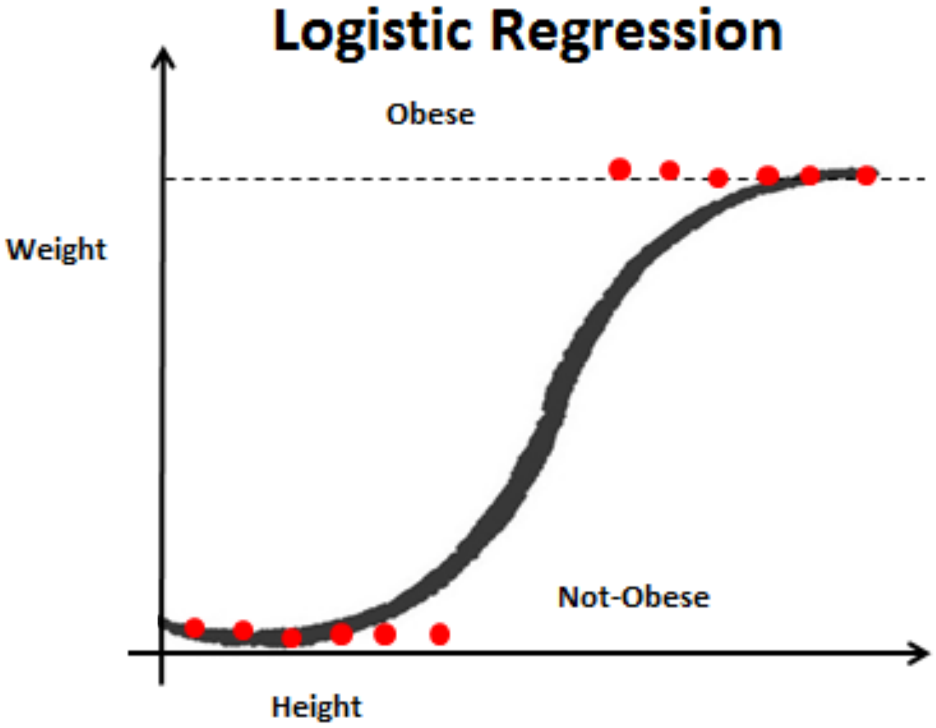
For continuous traits
-height
-BMI
-blood pressure

Covariates such as age, sex and ancestry are included to account for stratification and avoid confounding effects from demographic factors

Logistic regression

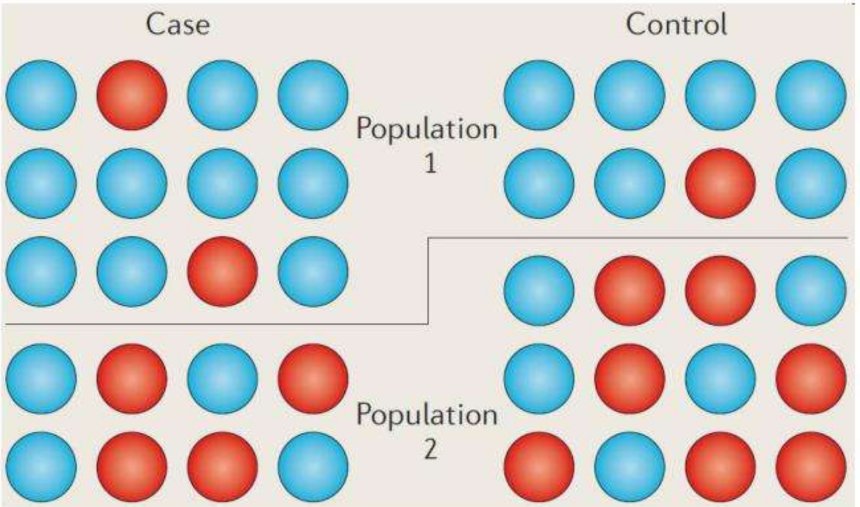
For binary traits
-presence/absence of disease





Confounding factors

Population stratification



Balding, Nature Reviews Genetics 2010

Population stratification arises when cases and controls are sampled from genetically different underlying populations, thus causing any associations found to be due to sampling differences rather than the disease of interest.

Systematic “errors” on the SNP array chips

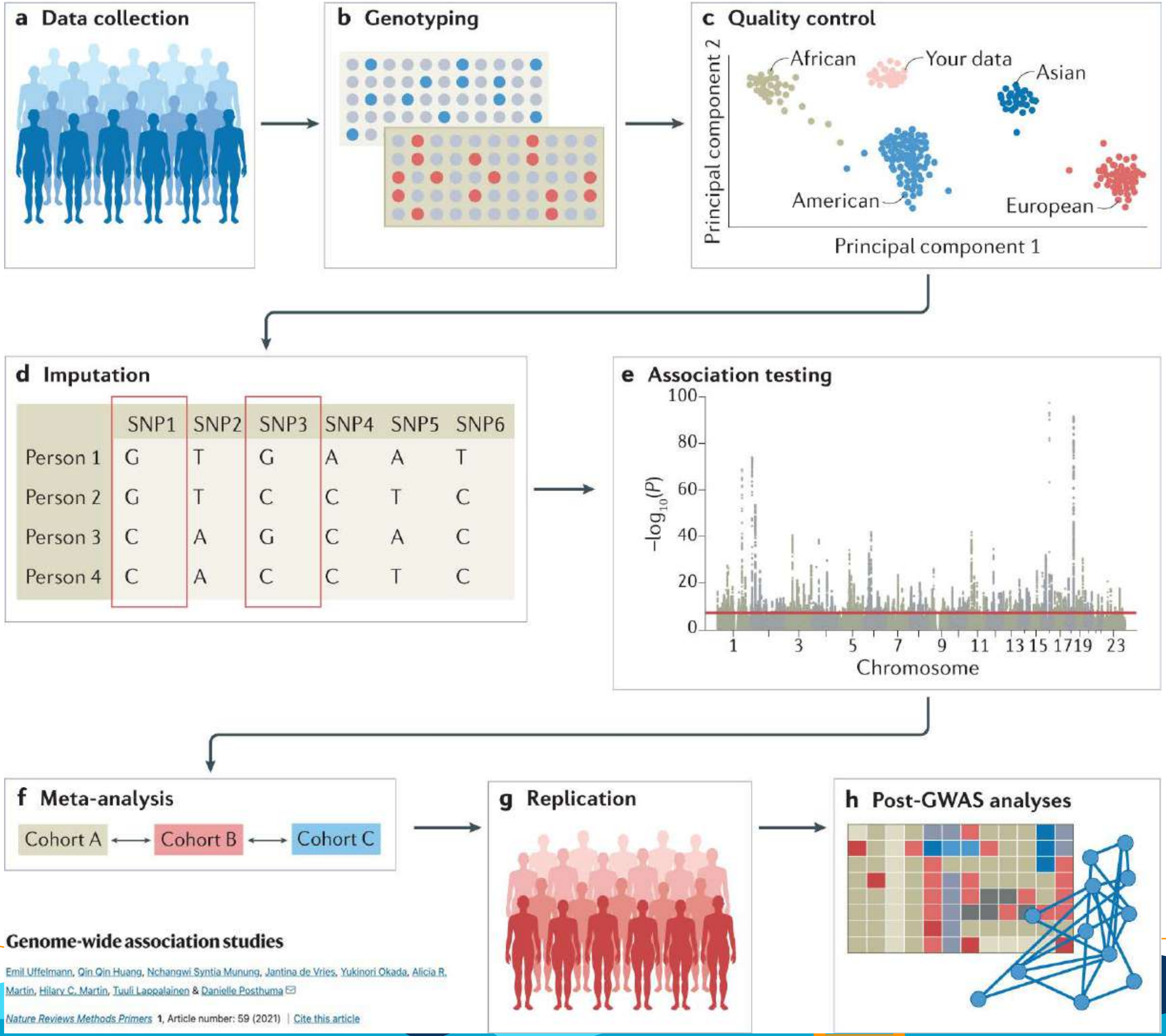
Sex, age

Disease heterogeneity

Appropriate reference panels

Gene annotation

GWAS process



	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6
Person 1	G	T	G	A	A	T
Person 2	G	T	C	C	T	C
Person 3	C	A	G	C	A	C
Person 4	C	A	C	C	T	C

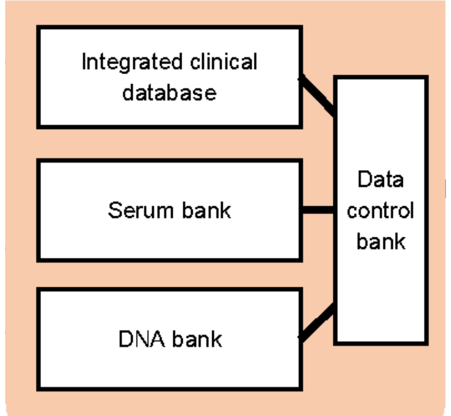
Genome-wide association studies

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen & Danielle Posthuma

Nature Reviews Methods Primers 1, Article number: 59 (2021) | Cite this article

Biobanks for genotypic and phenotypic data

BIOBANK JAPAN



Integrated clinical database

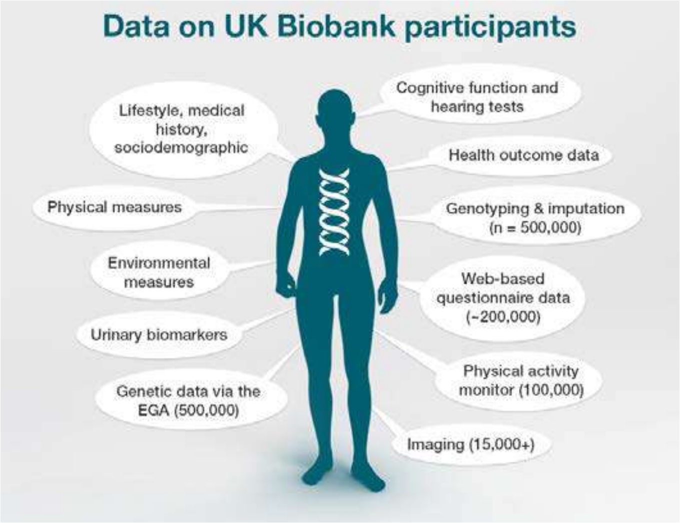
Serum bank

DNA bank

Data control bank

- 200,000 East Asians
- 47 common diseases, 59 quantitative traits
- 12 cooperative medical institutes all over Japan

biobank uk



Data on UK Biobank participants

Lifestyle, medical history, sociodemographic

Physical measures

Environmental measures

Urinary biomarkers

Genetic data via the EGA (500,000)

Cognitive function and hearing tests

Health outcome data

Genotyping & imputation (n = 500,000)

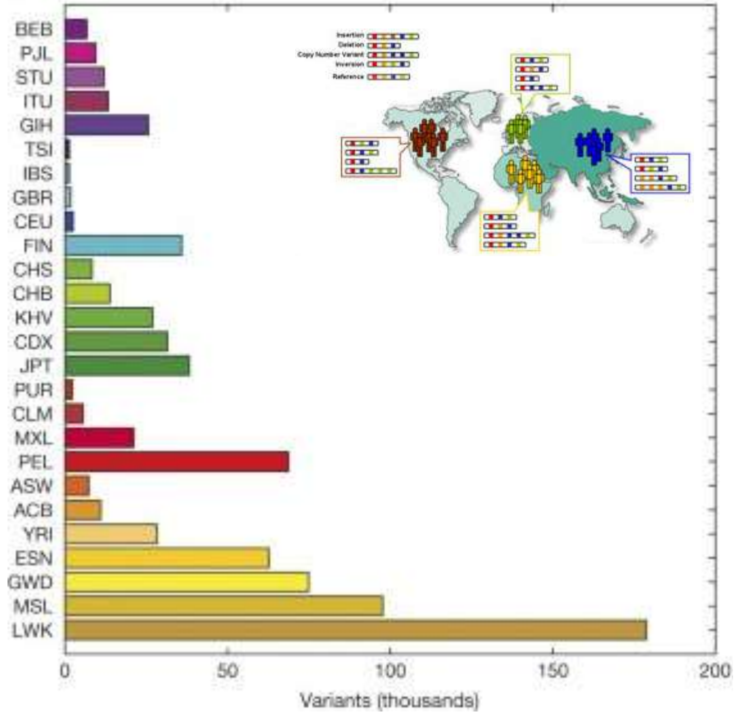
Web-based questionnaire data (~200,000)

Physical activity monitor (100,000)

Imaging (15,000+)

- 500,000 participants
- six ancestry groups
- ≥ 7000 phenotypes
- 23 cooperative medical institutes all over the UK

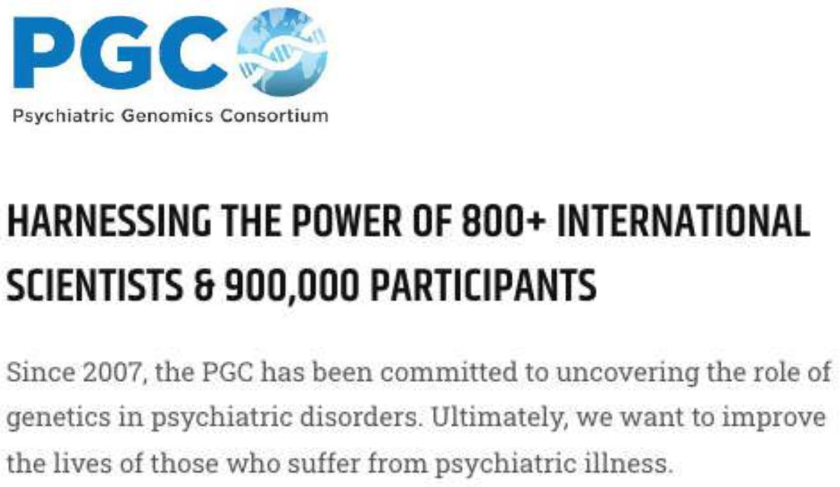
Why do we need biobanks for different populations?



The number of variants more common in the population compared to global population

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015)

Genomic data science - collaboration is the way

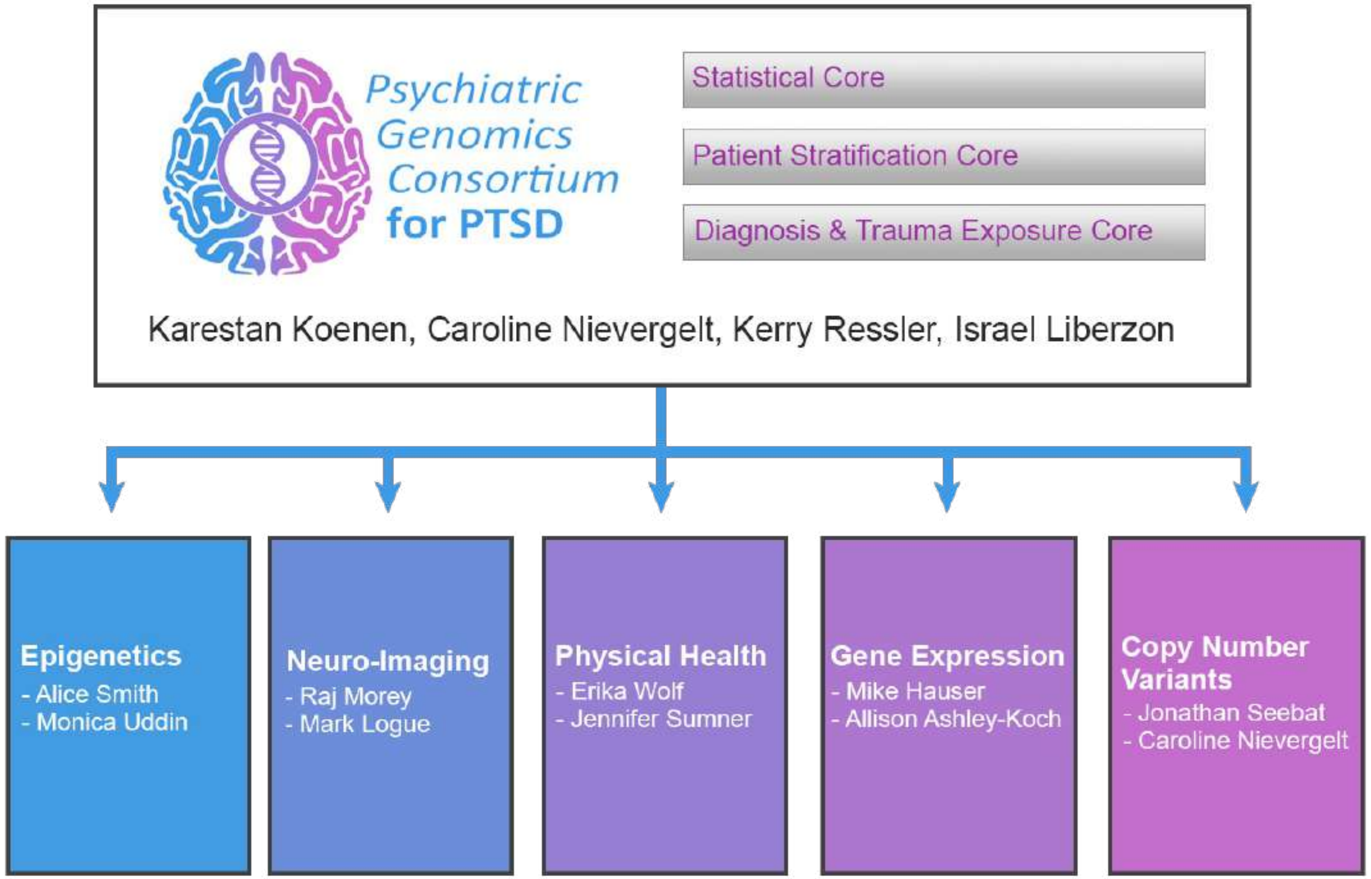


PGC

Psychiatric Genomics Consortium

HARNESSING THE POWER OF 800+ INTERNATIONAL SCIENTISTS & 900,000 PARTICIPANTS

Since 2007, the PGC has been committed to uncovering the role of genetics in psychiatric disorders. Ultimately, we want to improve the lives of those who suffer from psychiatric illness.



Psychiatric Genomics Consortium for PTSD

- Statistical Core
- Patient Stratification Core
- Diagnosis & Trauma Exposure Core

Karestan Koenen, Caroline Nievergelt, Kerry Ressler, Israel Liberzon

Epigenetics
- Alice Smith
- Monica Uddin

Neuro-Imaging
- Raj Morey
- Mark Logue

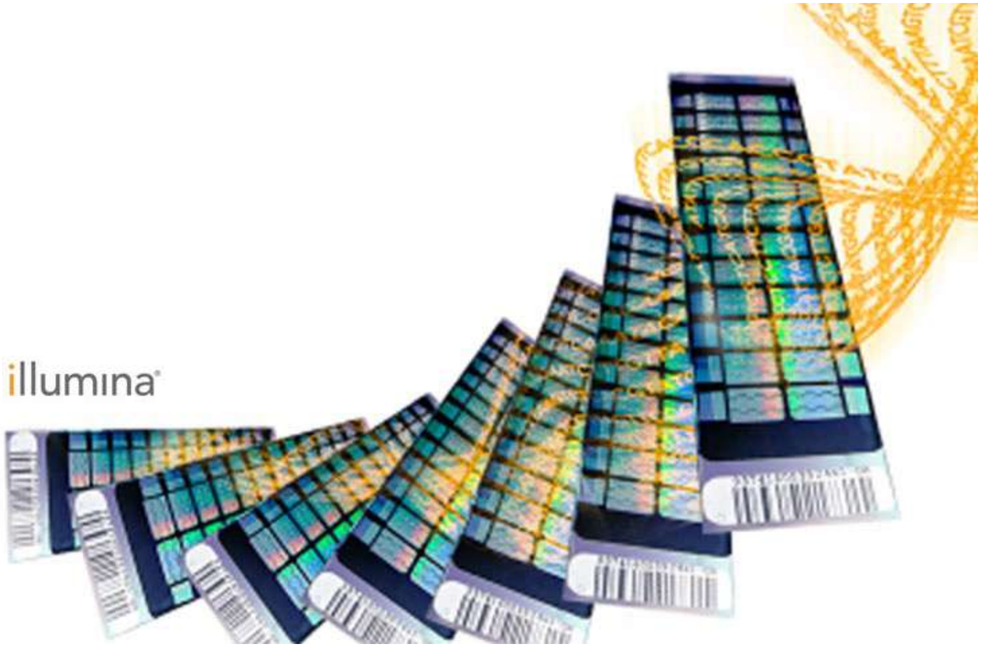
Physical Health
- Erika Wolf
- Jennifer Sumner

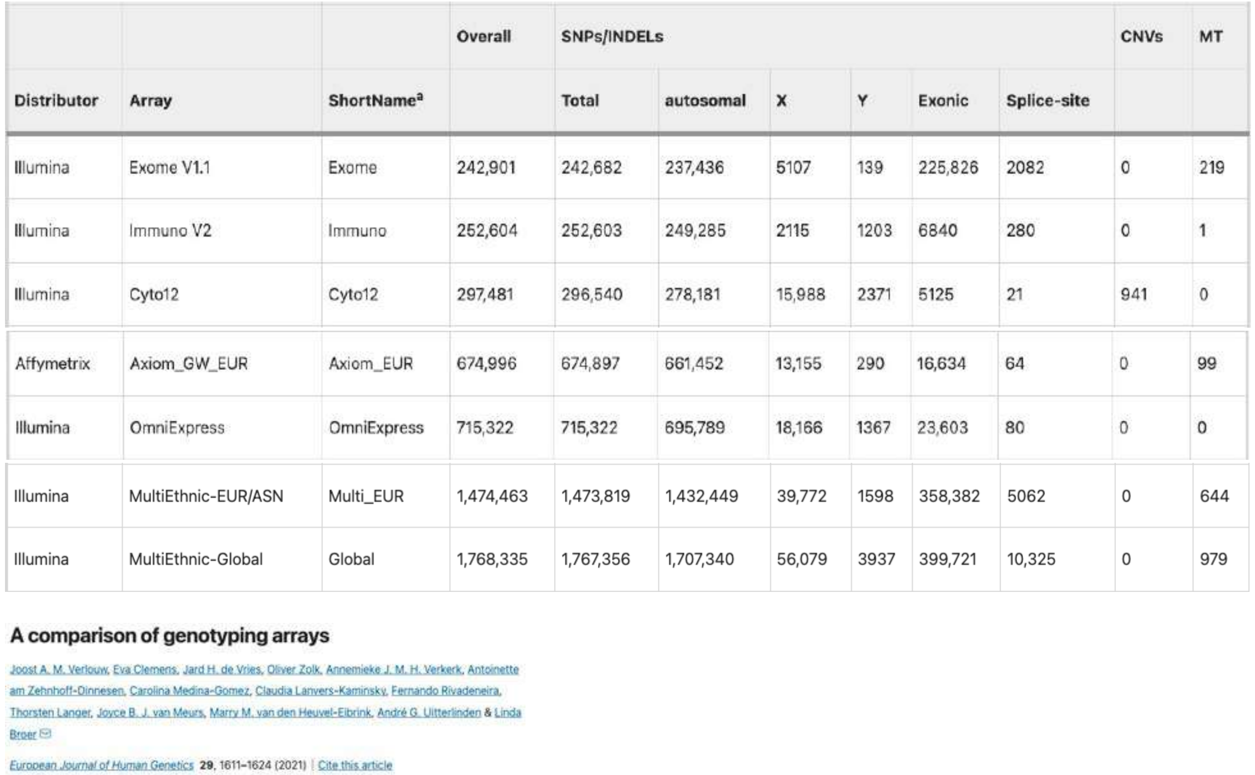
Gene Expression
- Mike Hauser
- Allison Ashley-Koch

Copy Number Variants
- Jonathan Seebat
- Caroline Nievergelt

Genome-wide genotyping

A certain number of variants (e.g., 850,000 for the UK Biobank) are directly measured, and millions (e.g., >90million for the UK Biobank) are imputed using a reference panel (e.g., Haplotype Reference Consortium, 1000 Genomes).





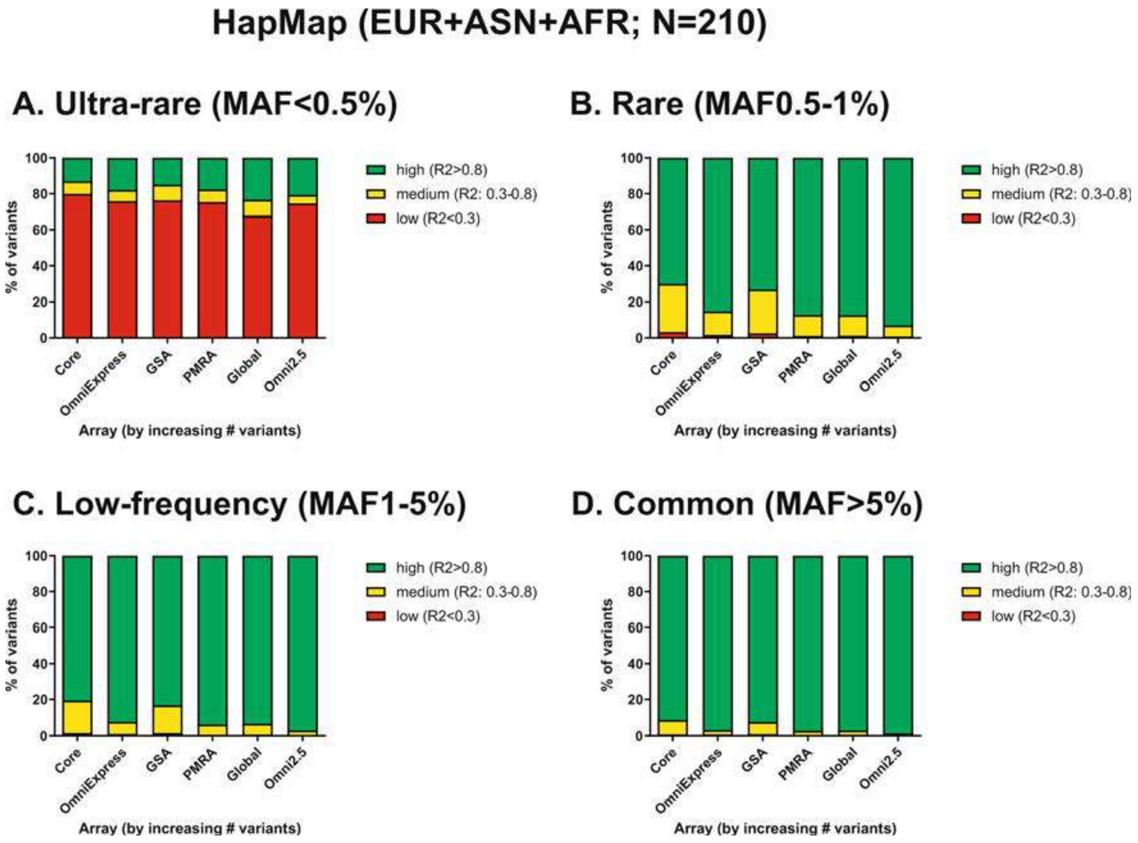
			Overall	SNPs/INDELs						CNVs	MT
Distributor	Array	ShortName[a]		Total	autosomal	X	Y	Exonic	Splice-site		
Illumina	Exome V1.1	Exome	242,901	242,682	237,436	5107	139	225,826	2082	0	219
Illumina	Immuno V2	Immuno	252,604	252,603	249,285	2115	1203	6840	280	0	1
Illumina	Cyto12	Cyto12	297,481	296,540	278,181	15,988	2371	5125	21	941	0
Affymetrix	Axiom_GW_EUR	Axiom_EUR	674,996	674,897	661,452	13,155	290	16,634	64	0	99
Illumina	OmniExpress	OmniExpress	715,322	715,322	695,789	18,166	1367	23,603	80	0	0
Illumina	MultiEthnic-EUR/ASN	Multi_EUR	1,474,463	1,473,819	1,432,449	39,772	1598	358,382	5062	0	644
Illumina	MultiEthnic-Global	Global	1,768,335	1,767,356	1,707,340	56,079	3937	399,721	10,325	0	979

A comparison of genotyping arrays

Joost A. M. Verlouw, Eva Clemens, Jard H. de Vries, Oliver Zolk, Annemieke J. M. H. Verkerk, Antoinette am Zehnhoff-Dinnesen, Carolina Medina-Gomez, Claudia Lanvers-Kaminsky, Fernando Rivadeneira, Thorsten Langer, Joyce B. J. van Meurs, Marry M. van den Heuvel-Eibrink, André G. Uitterlinden & Linda Broer ✉

European Journal of Human Genetics **29**, 1611–1624 (2021) | Cite this article

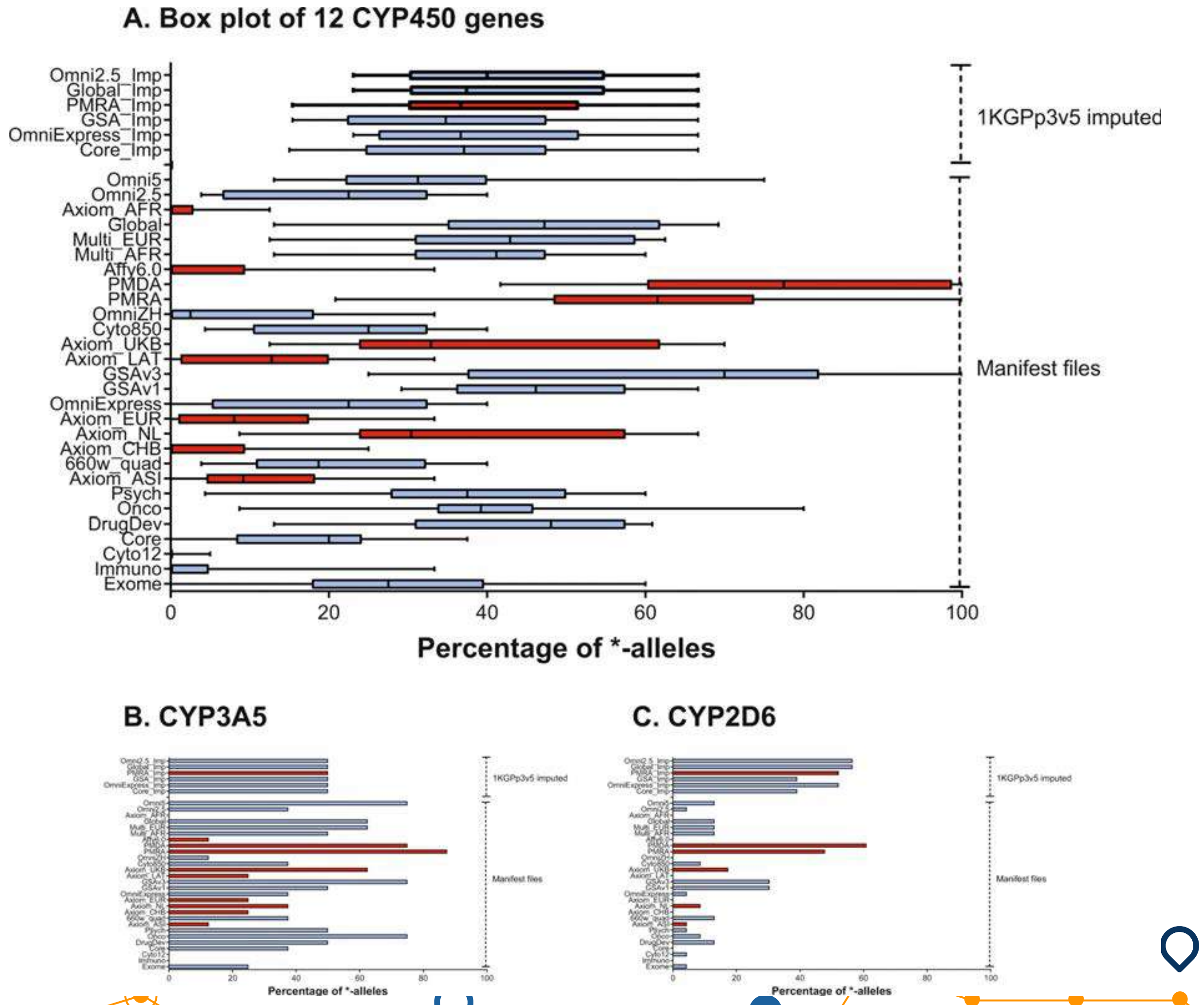
Genome-wide genotyping



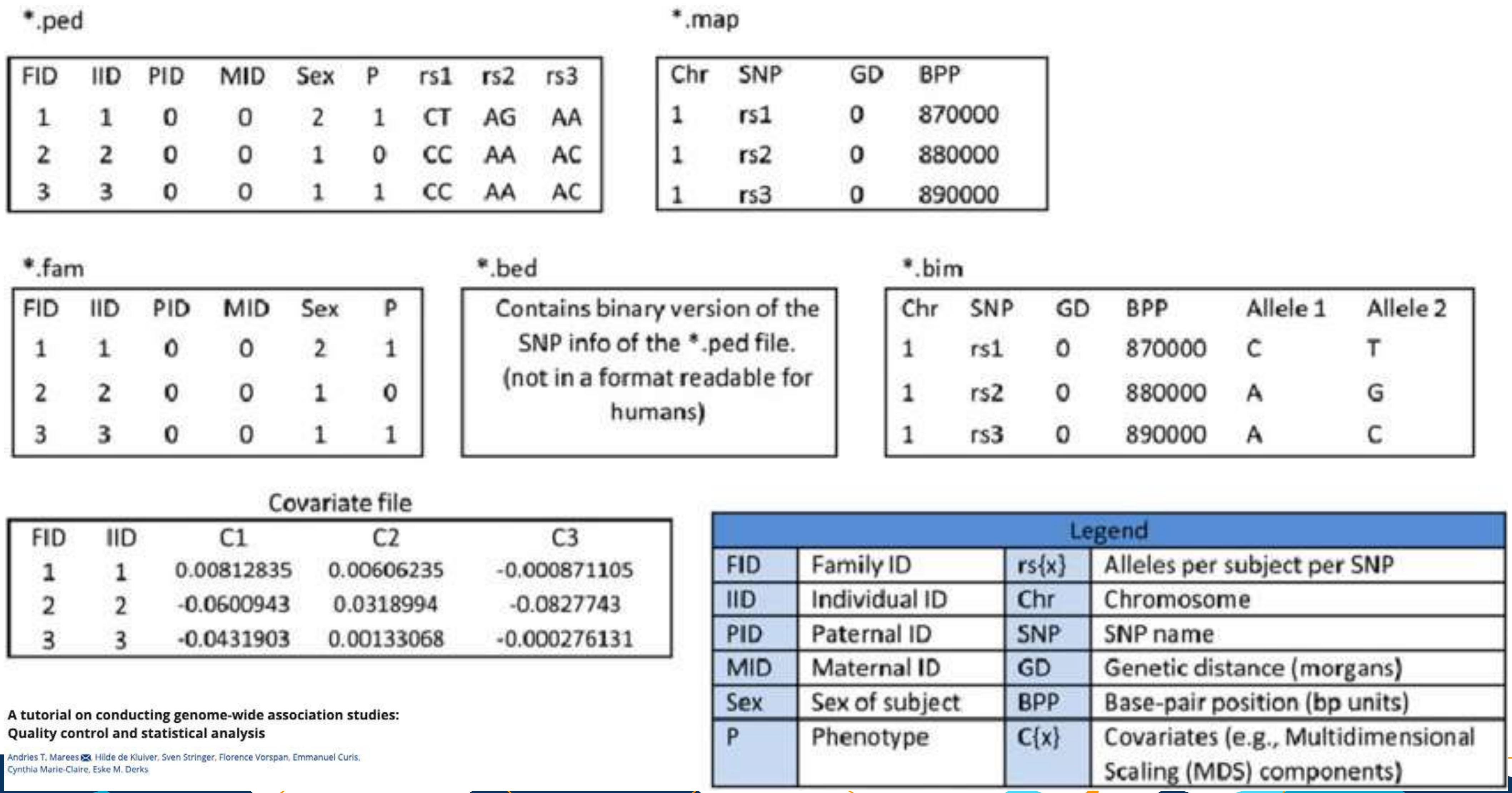
A comparison of genotyping arrays

Joost A. M. Verlouw, Eva Clemens, Jard H. de Vries, Oliver Zolk, Annemieke J. M. H. Verkerk, Antoinette am Zehnhoff-Dinnesen, Carolina Medina-Gomez, Claudia Lanvers-Kaminsky, Fernando Rivadeneira, Thorsten Langer, Joyce B. J. van Meurs, Marry M. van den Heuvel-Eibrink, André G. Uitterlinden & Linda Broer

European Journal of Human Genetics **29**, 1611–1624 (2021) | Cite this article



GWAS data - plink



*.ped

FID	IID	PID	MID	Sex	P	rs1	rs2	rs3
1	1	0	0	2	1	CT	AG	AA
2	2	0	0	1	0	CC	AA	AC
3	3	0	0	1	1	CC	AA	AC

*.map

Chr	SNP	GD	BPP
1	rs1	0	870000
1	rs2	0	880000
1	rs3	0	890000

*.fam

FID	IID	PID	MID	Sex	P
1	1	0	0	2	1
2	2	0	0	1	0
3	3	0	0	1	1

*.bed

Contains binary version of the SNP info of the *.ped file. (not in a format readable for humans)

*.bim

Chr	SNP	GD	BPP	Allele 1	Allele 2
1	rs1	0	870000	C	T
1	rs2	0	880000	A	G
1	rs3	0	890000	A	C

Covariate file

FID	IID	C1	C2	C3
1	1	0.00812835	0.00606235	-0.000871105
2	2	-0.0600943	0.0318994	-0.0827743
3	3	-0.0431903	0.00133068	-0.000276131

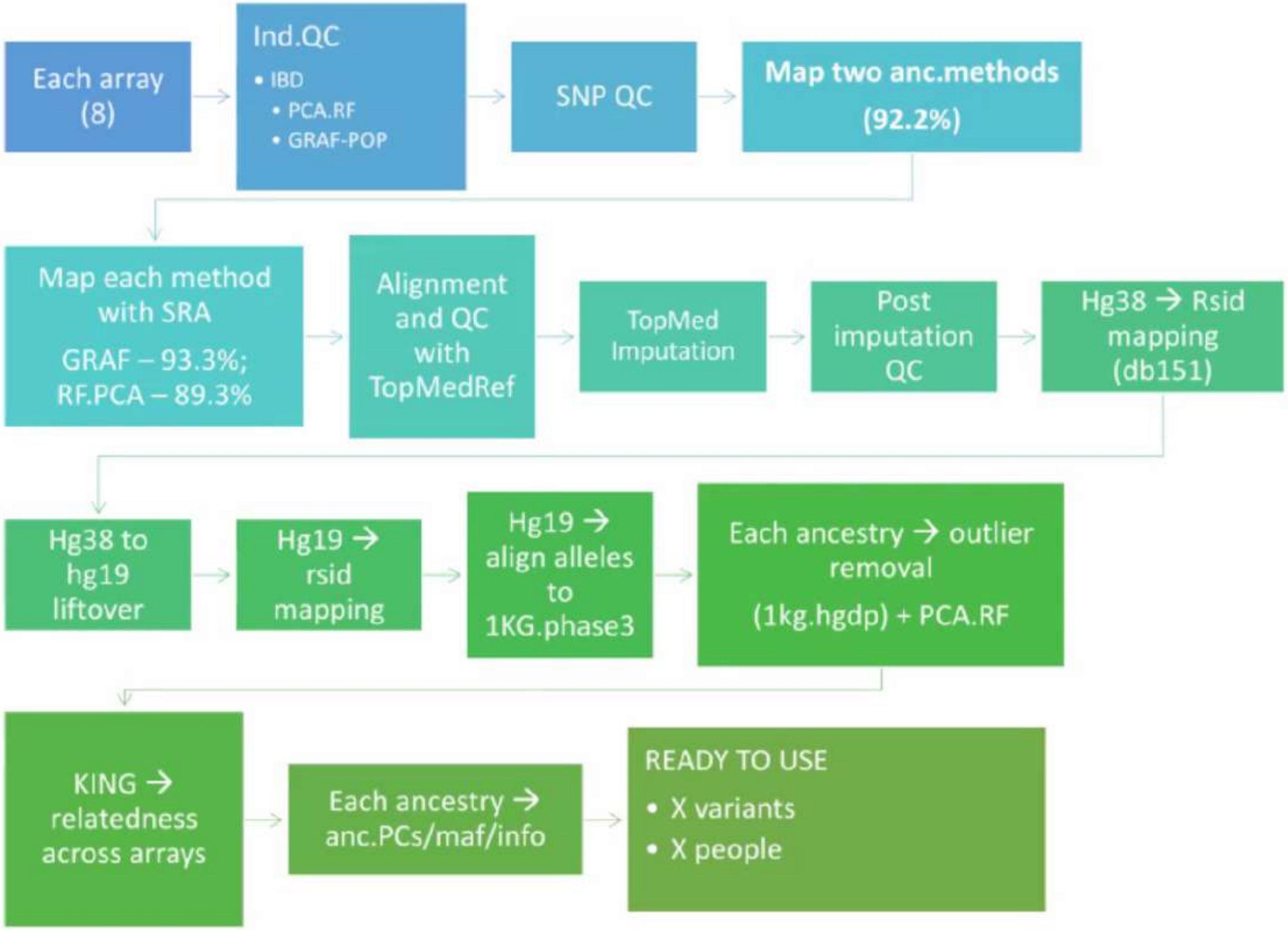
Legend

FID	Family ID	rs{x}	Alleles per subject per SNP
IID	Individual ID	Chr	Chromosome
PID	Paternal ID	SNP	SNP name
MID	Maternal ID	GD	Genetic distance (morgans)
Sex	Sex of subject	BPP	Base-pair position (bp units)
P	Phenotype	C{x}	Covariates (e.g., Multidimensional Scaling (MDS) components)

A tutorial on conducting genome-wide association studies: Quality control and statistical analysis

Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, Eske M. Derks

Quality control



Each array (8)

Ind.QC

- IBD
  - PCA.RF
  - GRAF-POP

SNP QC

Map two anc.methods (92.2%)

Map each method with SRA

GRAF – 93.3%;
RF.PCA – 89.3%

Alignment and QC with TopMedRef

TopMed Imputation

Post imputation QC

Hg38 → Rsid mapping (db151)

Hg38 to hg19 liftover

Hg19 → rsid mapping

Hg19 → align alleles to 1KG.phase3

Each ancestry → outlier removal

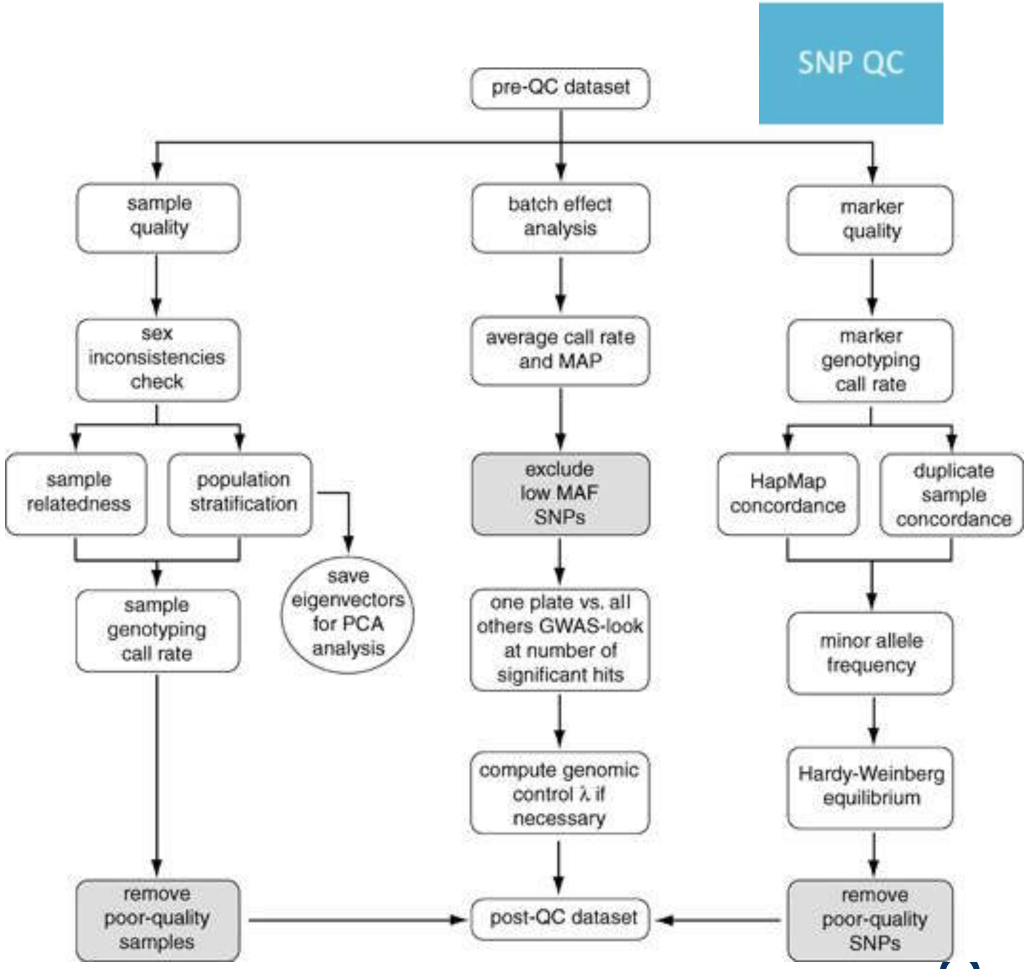
(1kg.hgdp) + PCA.RF

KING → relatedness across arrays

Each ancestry → anc.PCs/maf/info

READY TO USE

- X variants
- X people



SNP QC

pre-QC dataset

sample quality

batch effect analysis

marker quality

sex inconsistencies check

average call rate and MAP

marker genotyping call rate

sample relatedness

population stratification

exclude low MAF SNPs

HapMap concordance

duplicate sample concordance

save eigenvectors for PCA analysis

sample genotyping call rate

one plate vs. all others GWAS-look at number of significant hits

minor allele frequency

compute genomic control λ if necessary

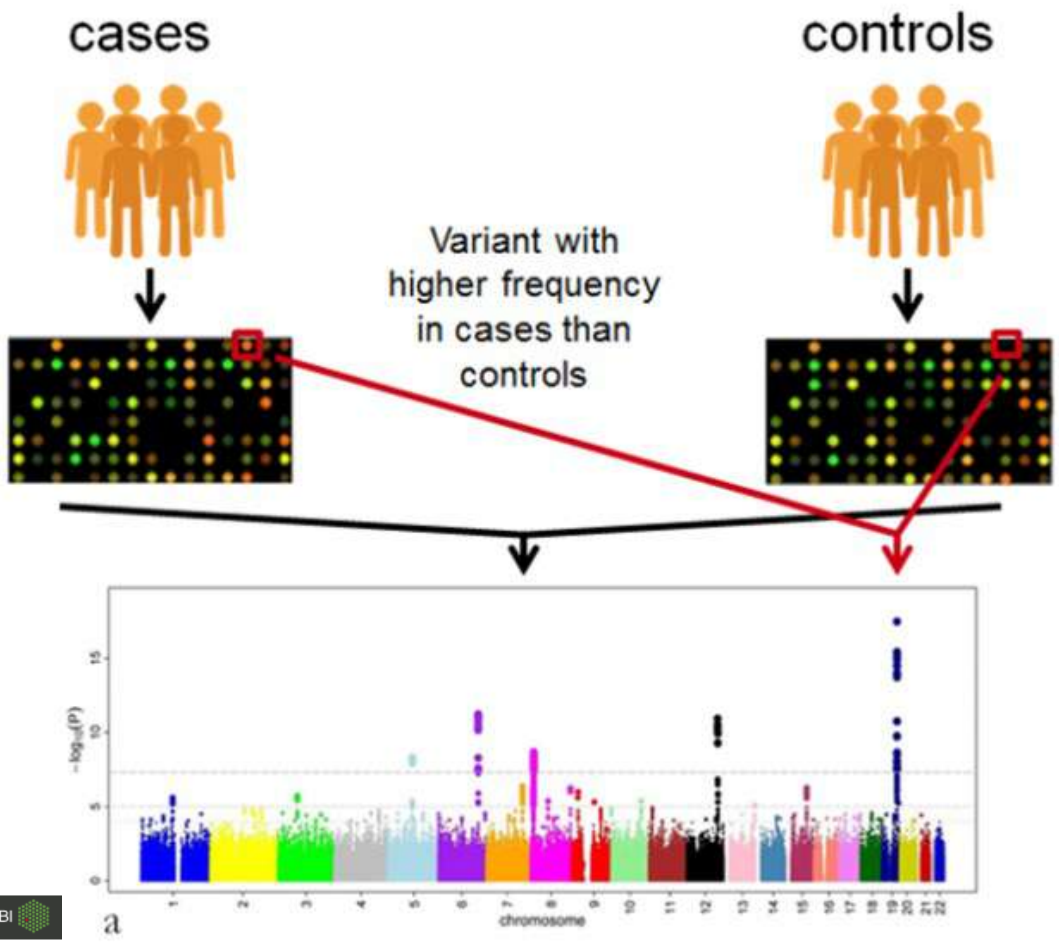
Hardy-Weinberg equilibrium

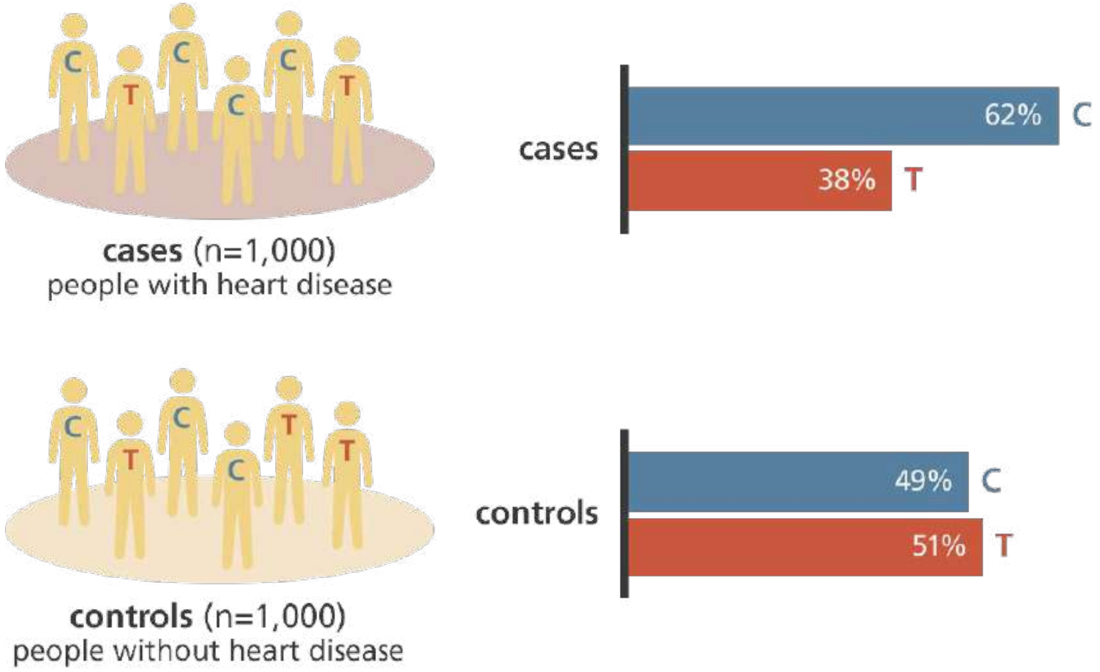
remove poor-quality samples

post-QC dataset

remove poor-quality SNPs

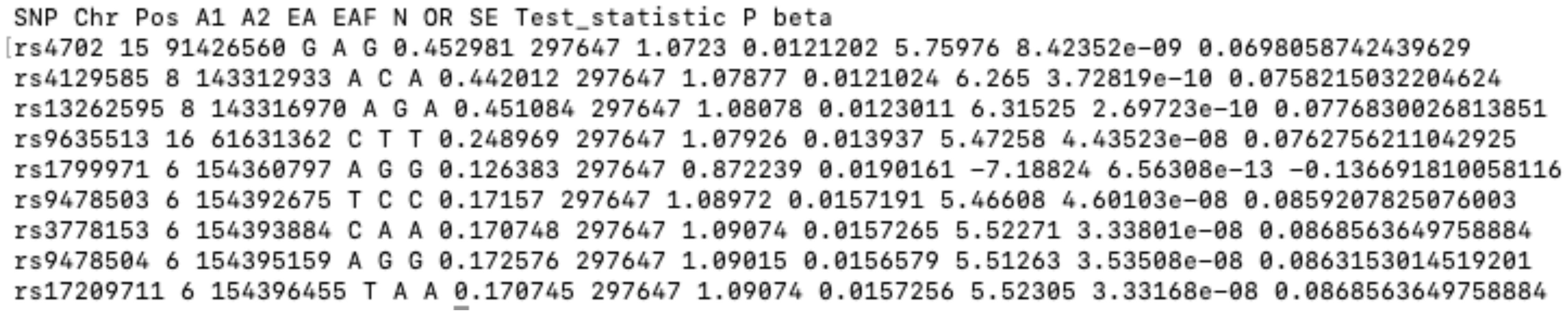
Association testing





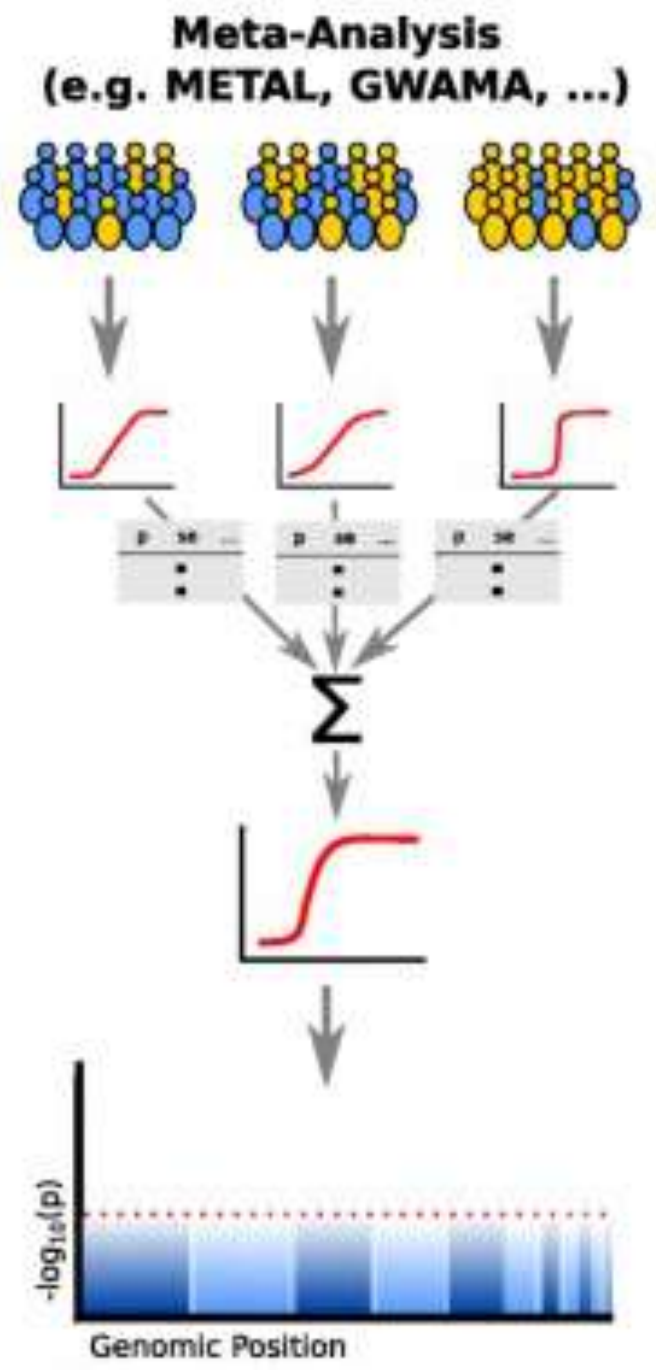
Genome Research Limited

GWAS summary data



```
SNP Chr Pos A1 A2 EA EAF N OR SE Test_statistic P beta
rs4702 15 91426560 G A G 0.452981 297647 1.0723 0.0121202 5.75976 8.42352e-09 0.0698058742439629
rs4129585 8 143312933 A C A 0.442012 297647 1.07877 0.0121024 6.265 3.72819e-10 0.0758215032204624
rs13262595 8 143316970 A G A 0.451084 297647 1.08078 0.0123011 6.31525 2.69723e-10 0.0776830026813851
rs9635513 16 61631362 C T T 0.248969 297647 1.07926 0.013937 5.47258 4.43523e-08 0.0762756211042925
rs1799971 6 154360797 A G G 0.126383 297647 0.872239 0.0190161 -7.18824 6.56308e-13 -0.136691810058116
rs9478503 6 154392675 T C C 0.17157 297647 1.08972 0.0157191 5.46608 4.60103e-08 0.0859207825076003
rs3778153 6 154393884 C A A 0.170748 297647 1.09074 0.0157265 5.52271 3.33801e-08 0.0868563649758884
rs9478504 6 154395159 A G G 0.172576 297647 1.09015 0.0156579 5.51263 3.53508e-08 0.0863153014519201
rs17209711 6 154396455 T A A 0.170745 297647 1.09074 0.0157256 5.52305 3.33168e-08 0.0868563649758884
```

Meta-analysis



ADVANTAGES

- ➢Greater statistical power.
- ➢Confirmatory data analysis.
- ➢Meta-analyses are used by researchers to review large and sometimes complex research.
- ➢Greater ability to extrapolate to general population affected.
- ➢Considered an evidence-based resource.

DISADVANTAGES

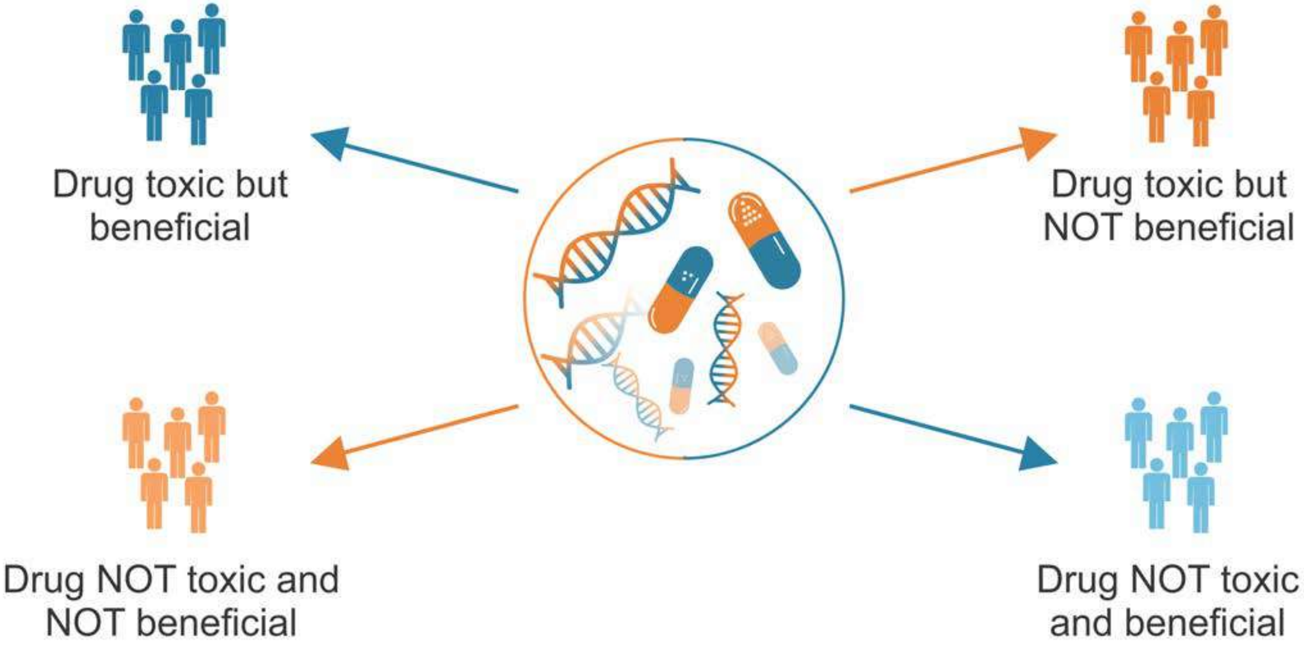
- ➢Difficult and time consuming to identify appropriate studies.
- ➢Not all studies provide adequate data for inclusion and analysis.
- ➢Requires advanced statistical techniques.
- ➢Heterogeneity of study populations.

sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies

Reza Nasirigerdeh, Reihaneh Torkzadehmahani, Julian Matschinske, Tobias Frisch, Markus List, Julian Spath, Stefan Weiss, Uwe Völker, Esa Pitkänen, Dominik Heider, Nina Kerstin Wenke, Georgios Kaissis, Daniel Rueckert, Tim Kacprowski & Jan Baumbach

Genome Biology **23**, Article number: 32 (2022) | Cite this article

Pharmacogenomics

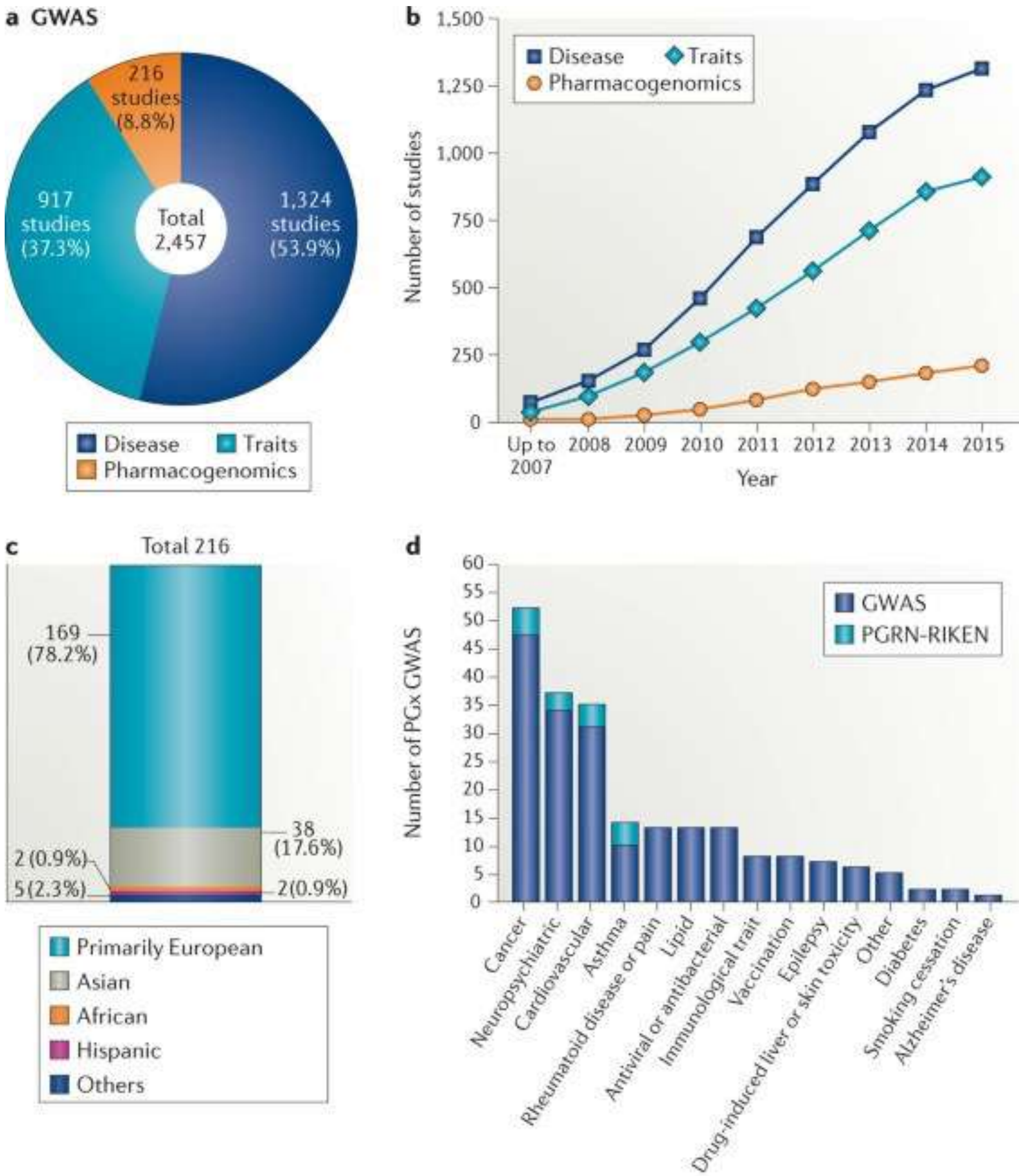


Drug toxic but beneficial

Drug toxic but NOT beneficial

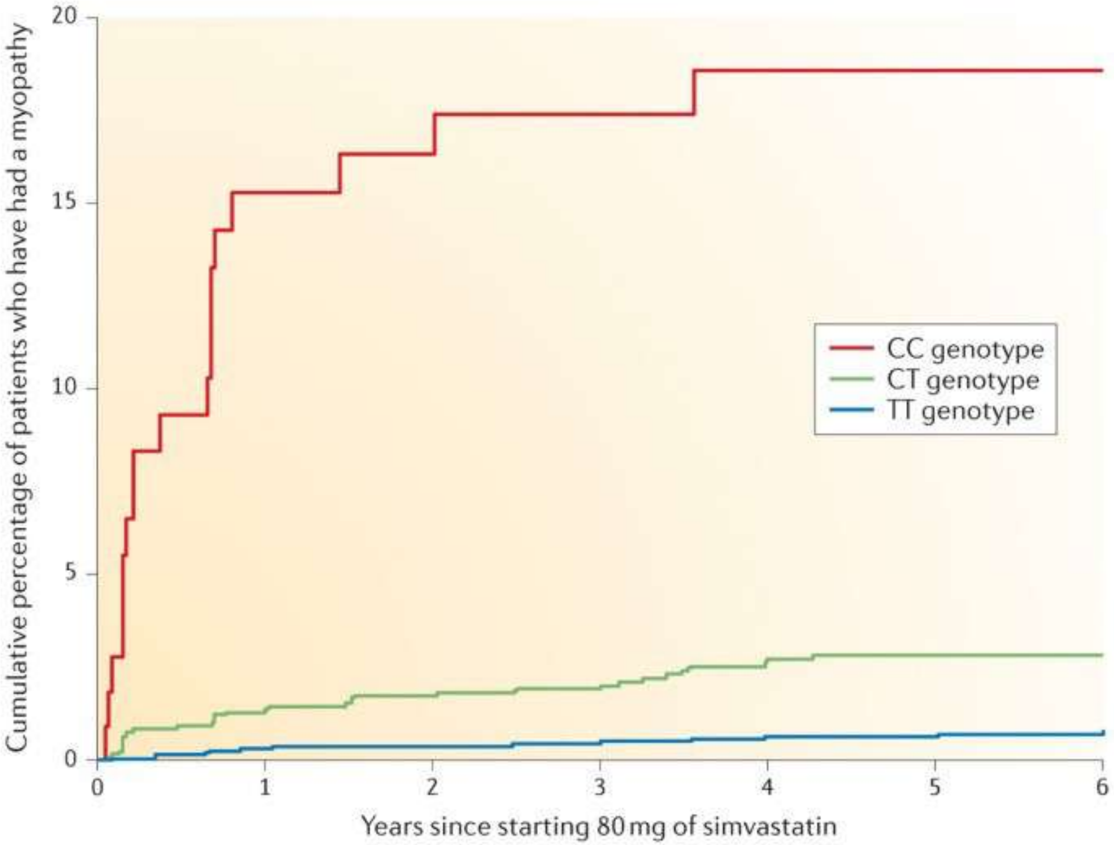
Drug NOT toxic and NOT beneficial

Drug NOT toxic and beneficial

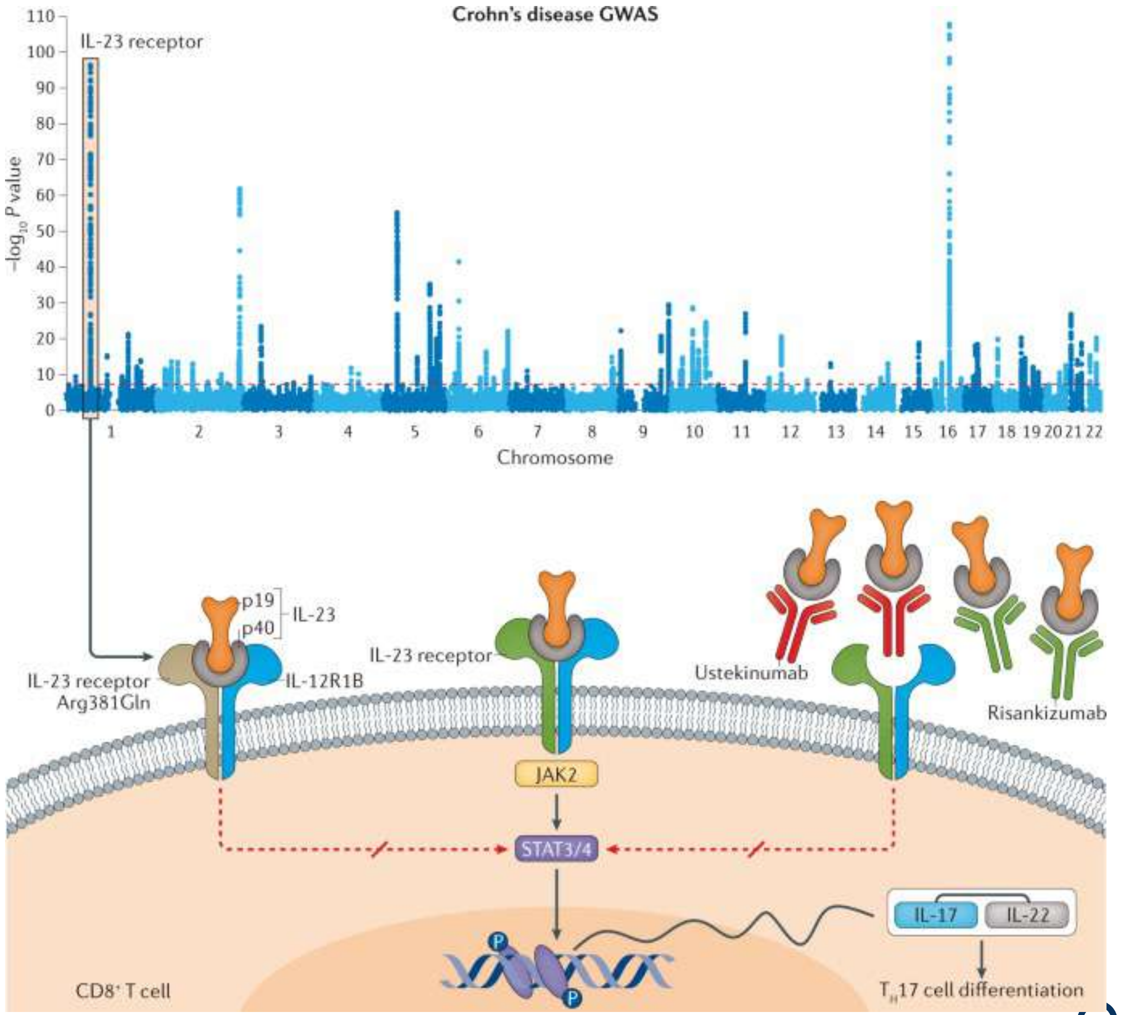


Nature Reviews | Drug Discovery

Clinical application of GWAS



Nature Reviews | Genetics

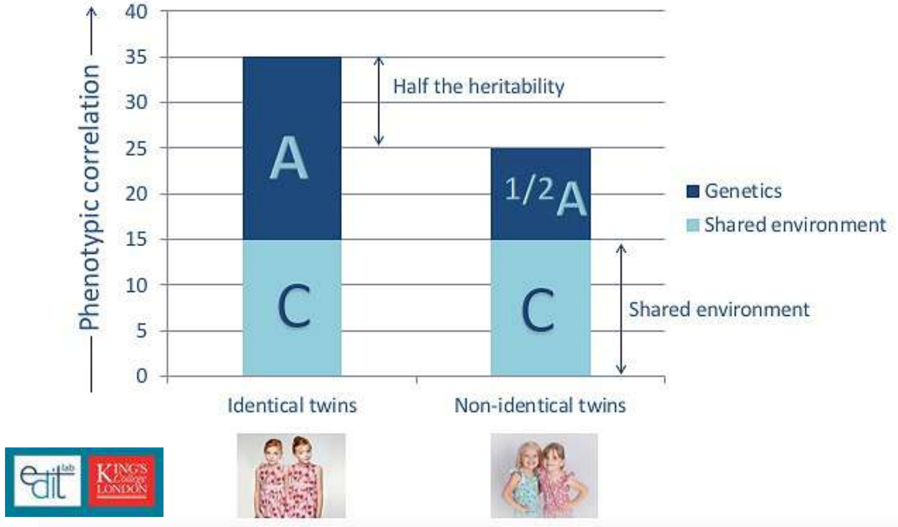


Post-GWAS analyses

Heritability

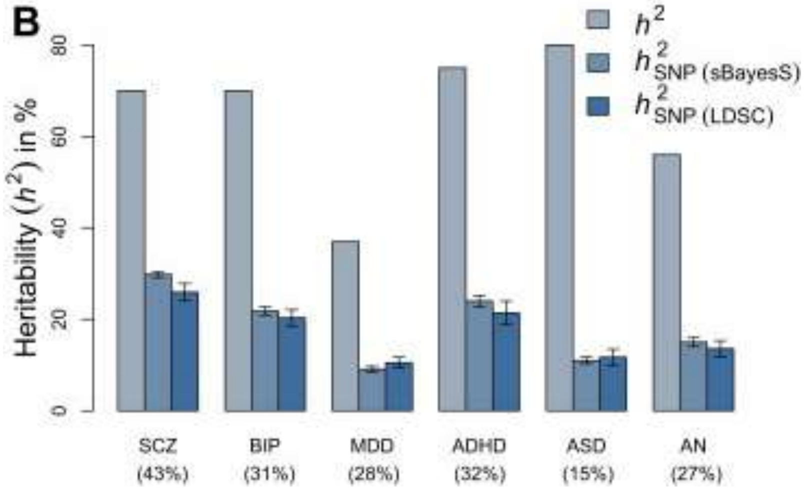
Heritability

Estimates the degree of variation in a phenotypic trait in a population that is due to genetic variation between individuals in that population. With other words, how well genetic differences among individuals account for differences in their complex traits.



SNP heritability

The fraction of the phenotypic variance explained by additive effects of a given set of genetic variants.

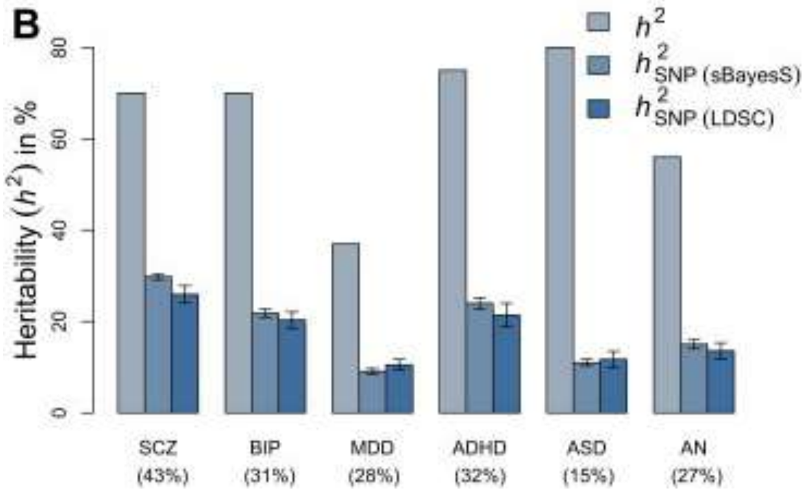


Baselmans et al., Biol. Psy., 2020.

Heritability of complex traits

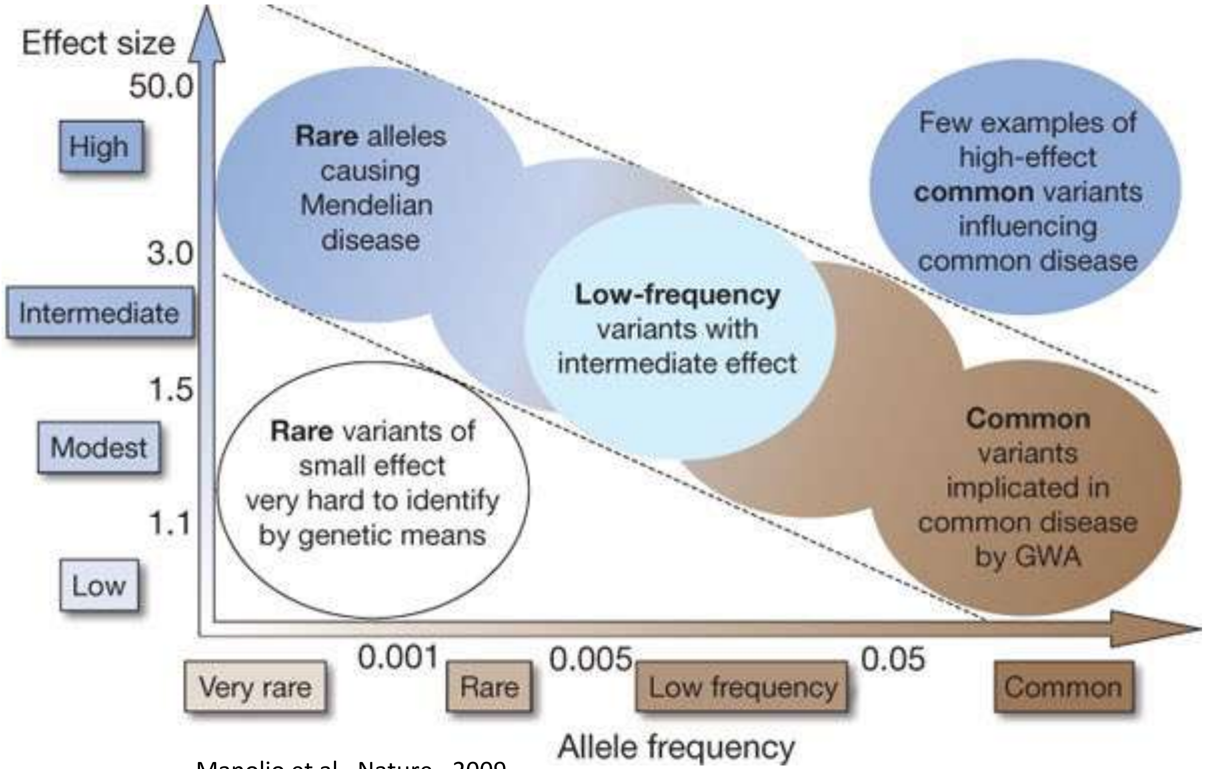
SNP heritability

The fraction of the phenotypic variance explained by additive effects of a given set of genetic variants.



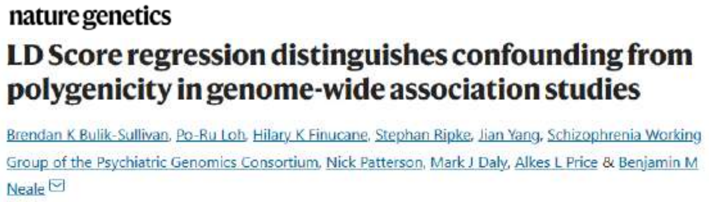
Baselmans et al., Biol. Psy., 2020.

Missing heritability


Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rare alleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
0.001
0.005
0.05
Very rare
Rare
Low frequency
Common
Allele frequency

Manolio et al., Nature., 2009.

Linkage disequilibrium score regression



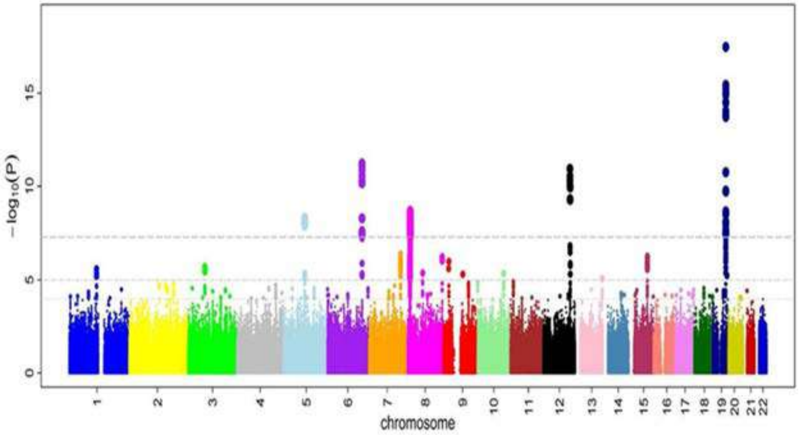
nature genetics

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

The approach involves using regression analysis to examine the relationship between LD scores and the test statistics of SNPs from the GWAS. The lowest LD Score of a SNP is one, which is obtained when a SNP is in perfect linkage equilibrium with all other SNPs.

GWAS summary statistics



Aggregate p-values and association data for every variant analyzed in a GWAS

+

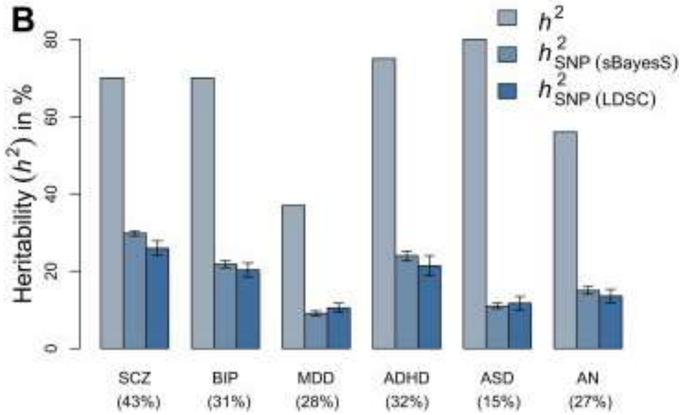
LD scores

Sum of LD r2 between a variant and all the variants in a region

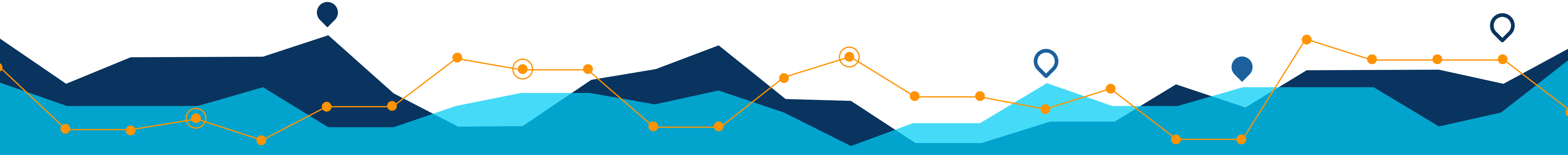
=

Estimating SNP heritability (h2)

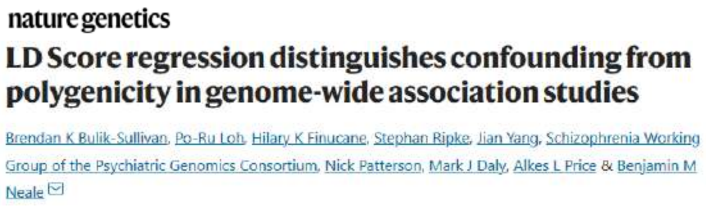
The fraction of the phenotypic variance explained by additive effects of a given set of genetic variants



Baselmans et al., Biol. Psy., 2020.



Linkage disequilibrium score regression



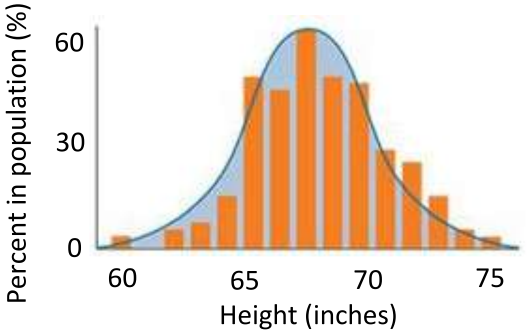
nature genetics

LD Score regression distinguishes confounding from polygenicity in genome-wide association studies

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price & Benjamin M Neale

Aims to quantify the separate contributions of polygenic effects and various confounding factors, such as population stratification, based on summary statistics from genome-wide association studies (GWASs).

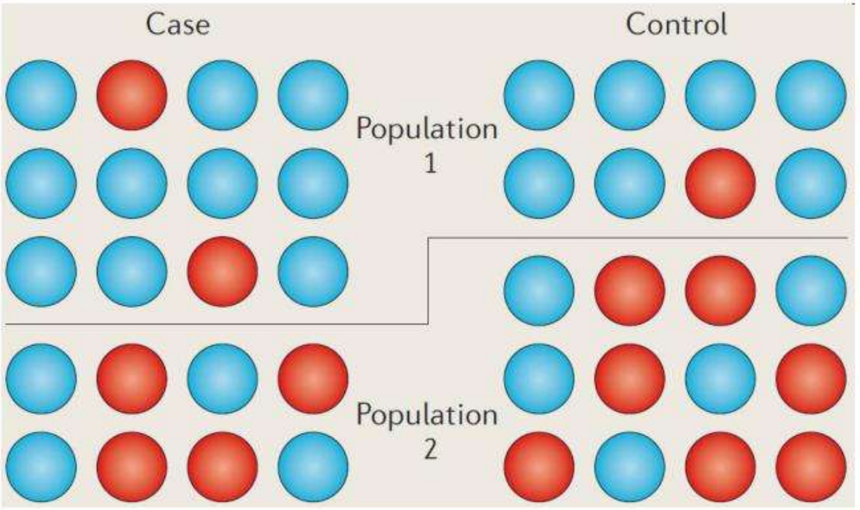
Polygenicity



one characteristic is controlled by two or more genes

Population stratification

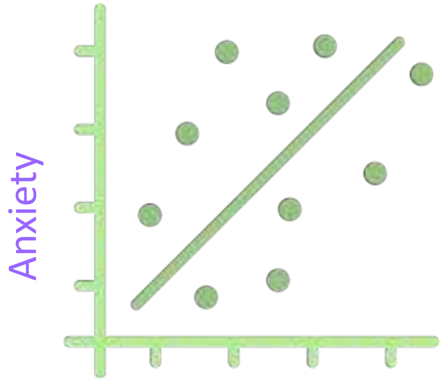
Population stratification arises when cases and controls are sampled from genetically different underlying populations, thus causing any associations found to be due to sampling differences rather than the disease of interest.



Balding, Nature Reviews Genetics 2010

Genetic correlation

- The proportion of variance that two traits share due to genetic causes
- The correlation between the genetic influences on a trait and the genetic influences on a different trait
- Estimates the degree of pleiotropy



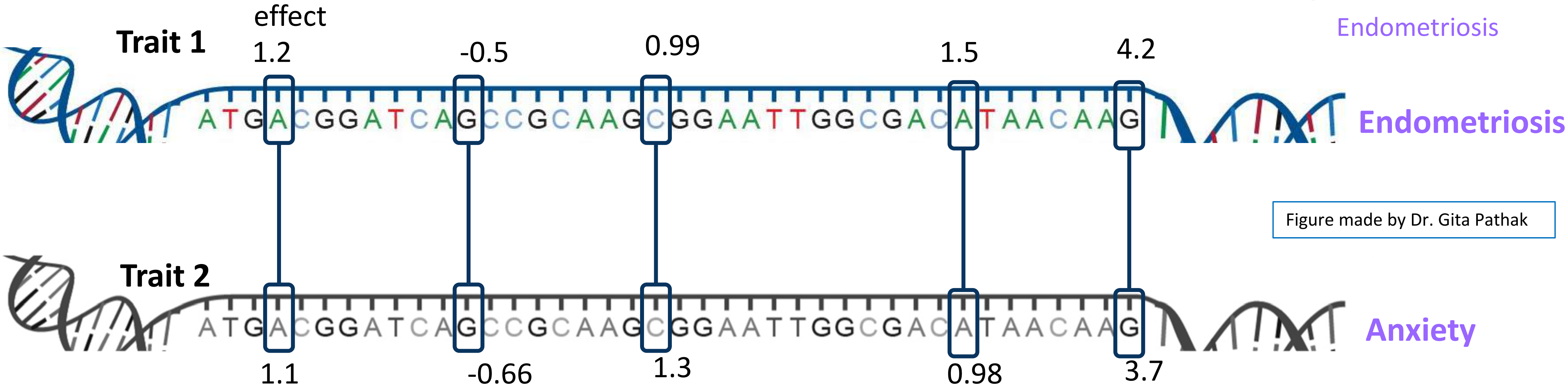
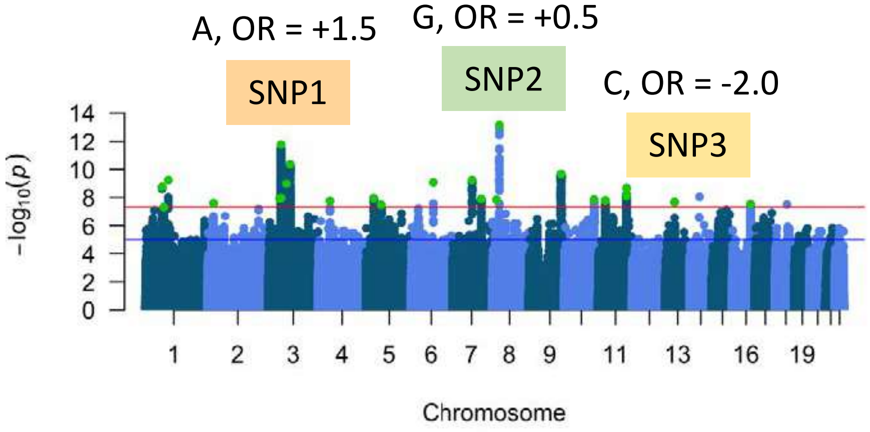

effect
Trait 1
1.2
-0.5
0.99
1.5
4.2
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
Endometriosis
Trait 2
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAACAAG
Anxiety
1.1
-0.66
1.3
0.98
3.7

Figure made by Dr. Gita Pathak

Polygenic risk scoring

Trait 1 – GWAS summary statistics



Base data

- Summary statistics
- Betas/ORs as weights in PRS calculation

+

Trait 1 – Individual level data

Ind	Pheno	SNP	A1	A2
P1	1	SNP1	A	T
P2	0	SNP2	G	C
P3	1	SNP3	C	A

Target data

- Individual-level genotype and phenotype data
- Often small sample size

PGS of trait 1

AT (+1.5x1) + GC (+0.5x1) + CC (-2.0x2)

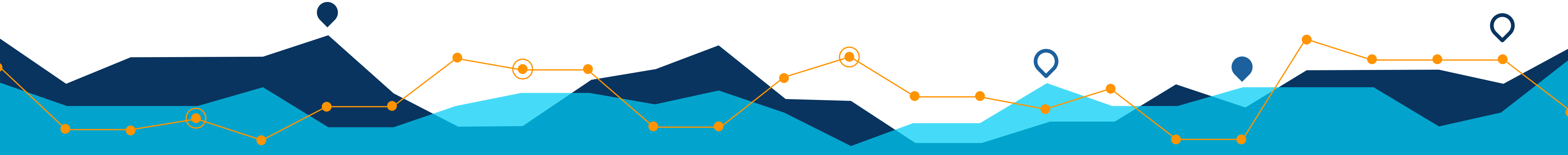
AA (+1.5x2) + GC (+0.5x1) + CA (-2.0x1)

TT (+1.5x0) + GG (+0.5x2) + CC (-2.0x2)

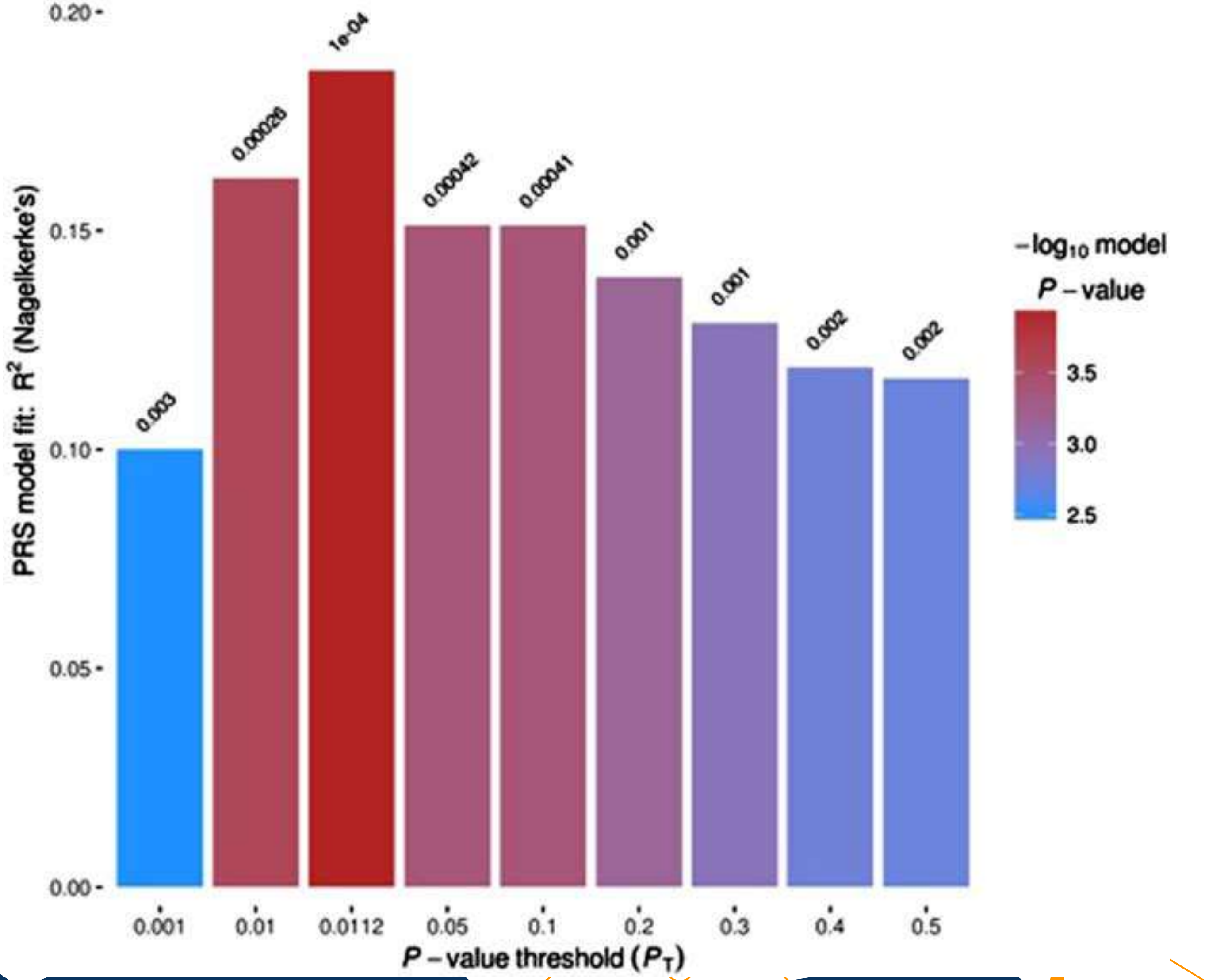
AT (+1.5x1) + CC (+0.5x0) + AA (-2.0x0)

$$\sum_{p\text{-val}}\sum_{\text{SNPs}}(\text{allele count})_{\text{test}} \times (\text{effect size})_{\text{training}}$$

Trait 2, 3, 4… – GWAS summary statistics



Polygenic risk scoring

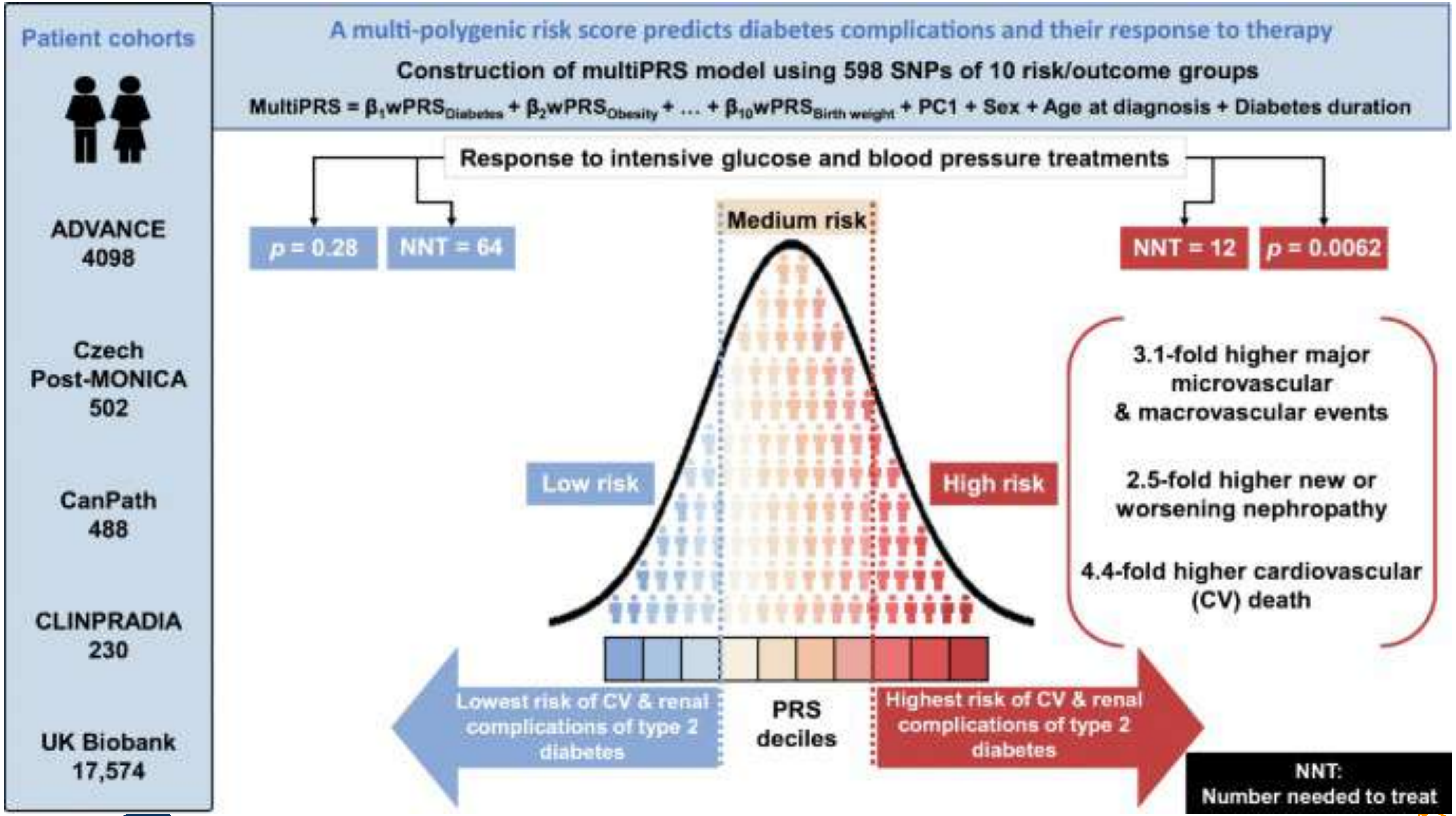


Only SNPs with a GWAS association *P*-value below a certain threshold (e.g. $P < 0.01$) are included in the calculation of the PRS, while all other SNPs are excluded

R²: how the PRS at a given threshold explains the difference between cases and controls

Optimal threshold:
Number of SNPs are not too large
Subset of SNPs that are predictive of the target trait

Polygenic risk scoring



Patient cohorts

ADVANCE 4098

Czech Post-MONICA 502

CanPath 488

CLINPRADIA 230

UK Biobank 17,574

A multi-polygenic risk score predicts diabetes complications and their response to therapy

Construction of multiPRS model using 598 SNPs of 10 risk/outcome groups

MultiPRS = $\beta_1 wPRS_{Diabetes} + \beta_2 wPRS_{Obesity} + \ldots + \beta_{10} wPRS_{Birth\ weight}$ + PC1 + Sex + Age at diagnosis + Diabetes duration

Response to intensive glucose and blood pressure treatments

3.1-fold higher major microvascular & macrovascular events

2.5-fold higher new or worsening nephropathy

4.4-fold higher cardiovascular (CV) death

Lowest risk of CV & renal complications of type 2 diabetes

Highest risk of CV & renal complications of type 2 diabetes

NNT: Number needed to treat

Polygenic risk scoring

Schizophrenia, Nature, 2022

PRS analysis explained a median of 0.073 of variance in liability (SNPs with GWAS $P < 0.05$), and 0.024 when restricted to genome-wide significant SNPs

7.3%

Depression, Nature Neuro, 2019

PRS analysis explained a median of 0.015 of variance in liability (SNPs with GWAS $P < 0.05$)

1.5%

ADHD, Nature Genetics, 2019

PRS analysis explained a median of 0.055 of variance in liability (SNPs with GWAS $P < 0.05$)

5.5%

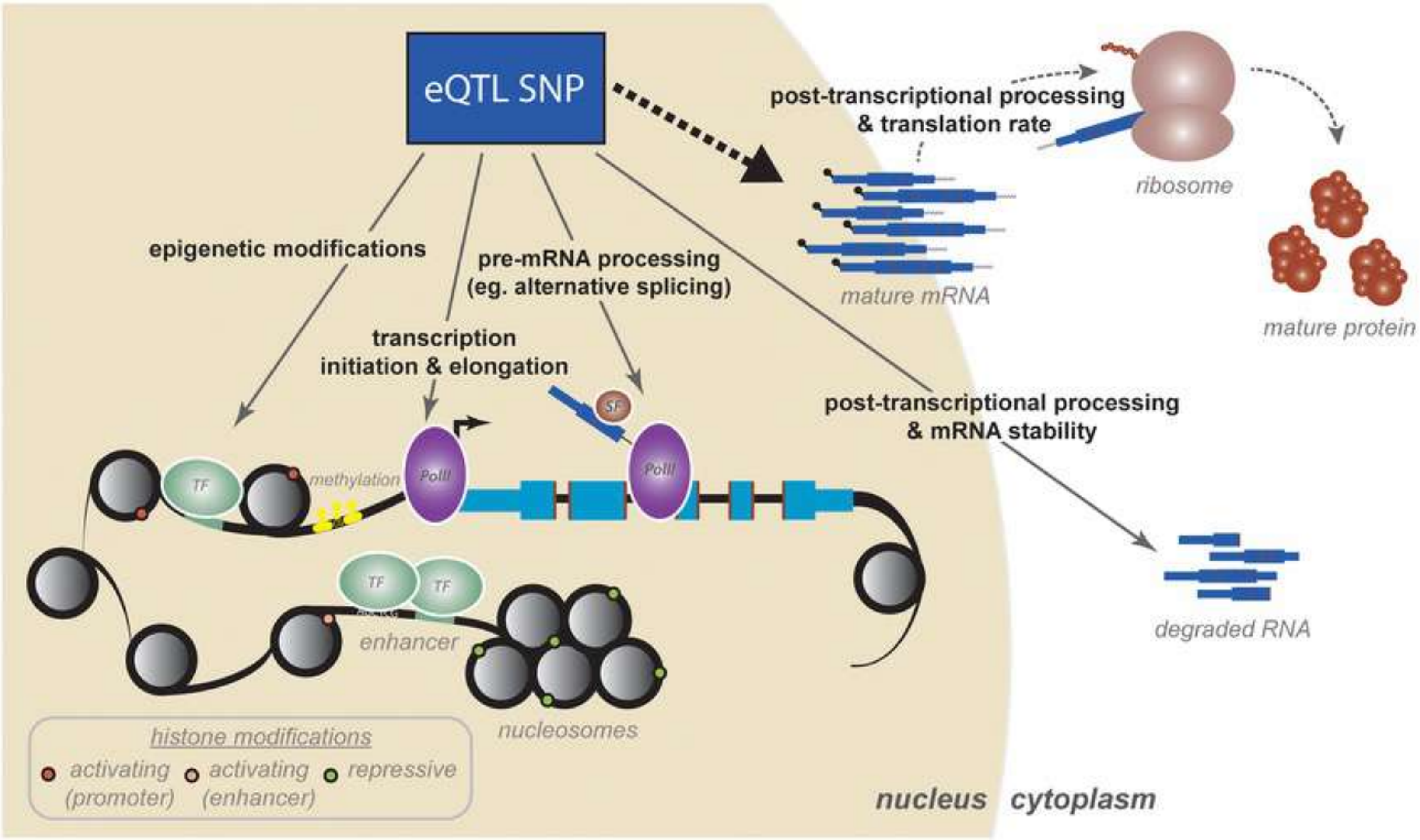
Annotation of SNPs to genes - positional mapping

Gene	Current result	Previous GWAS
ZFHX3	Arrhythmia	Atrial fibrillation
KCNQ1 *LINC01153*	Type 2 diabetes	Type 2 diabetes
IP6K3	Rheumatoid arthritis	Platelet crit Testicular carcinoma
GLB1	Atopic dermatitis	Atopic dermatitis

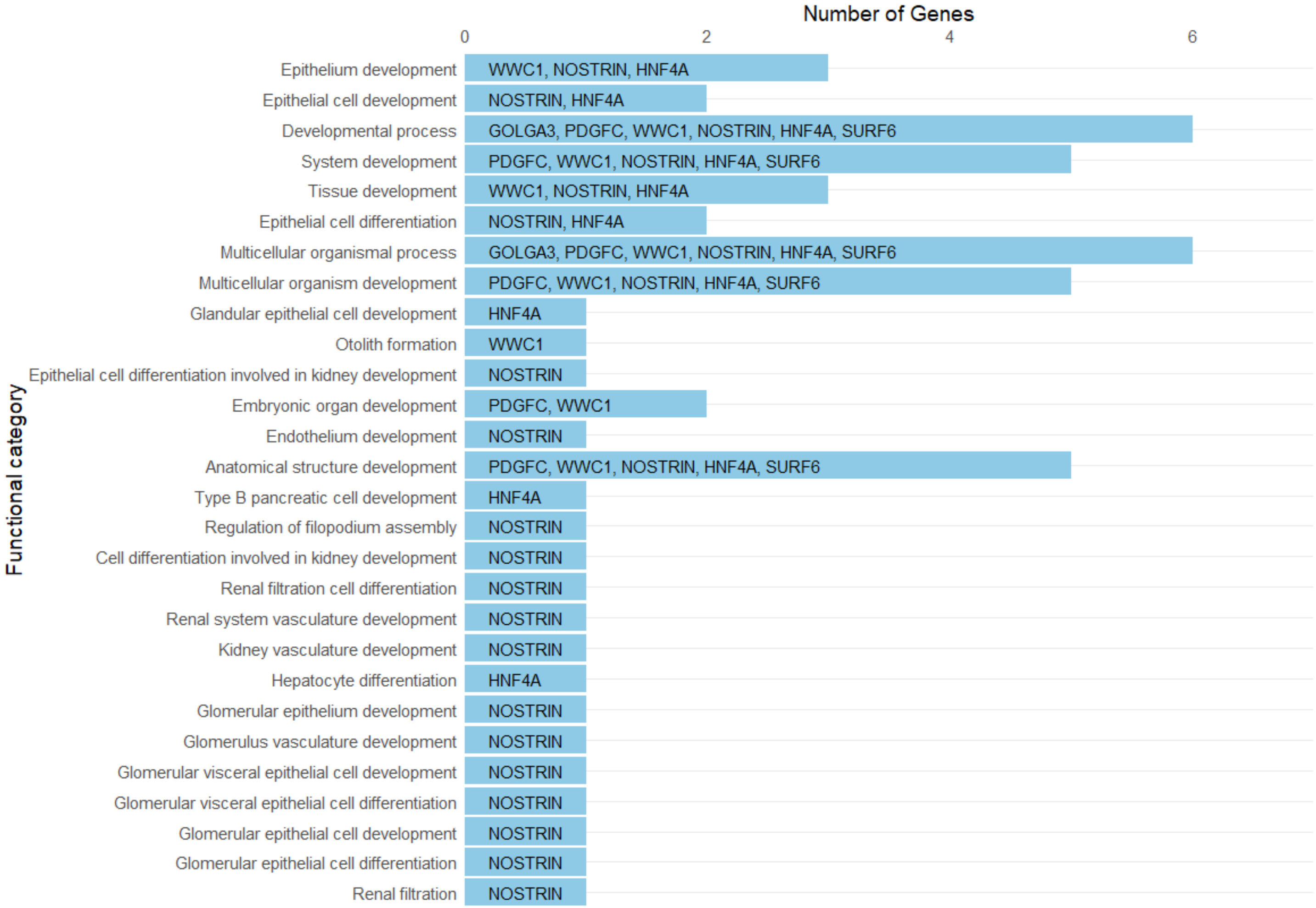
SNPnexus

e!Ensembl

Annotation of SNPs to genes – eQTLs

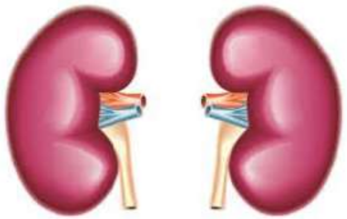


Enrichment for biological processes, cellular components, molecular functions



ShinyGO v0.741

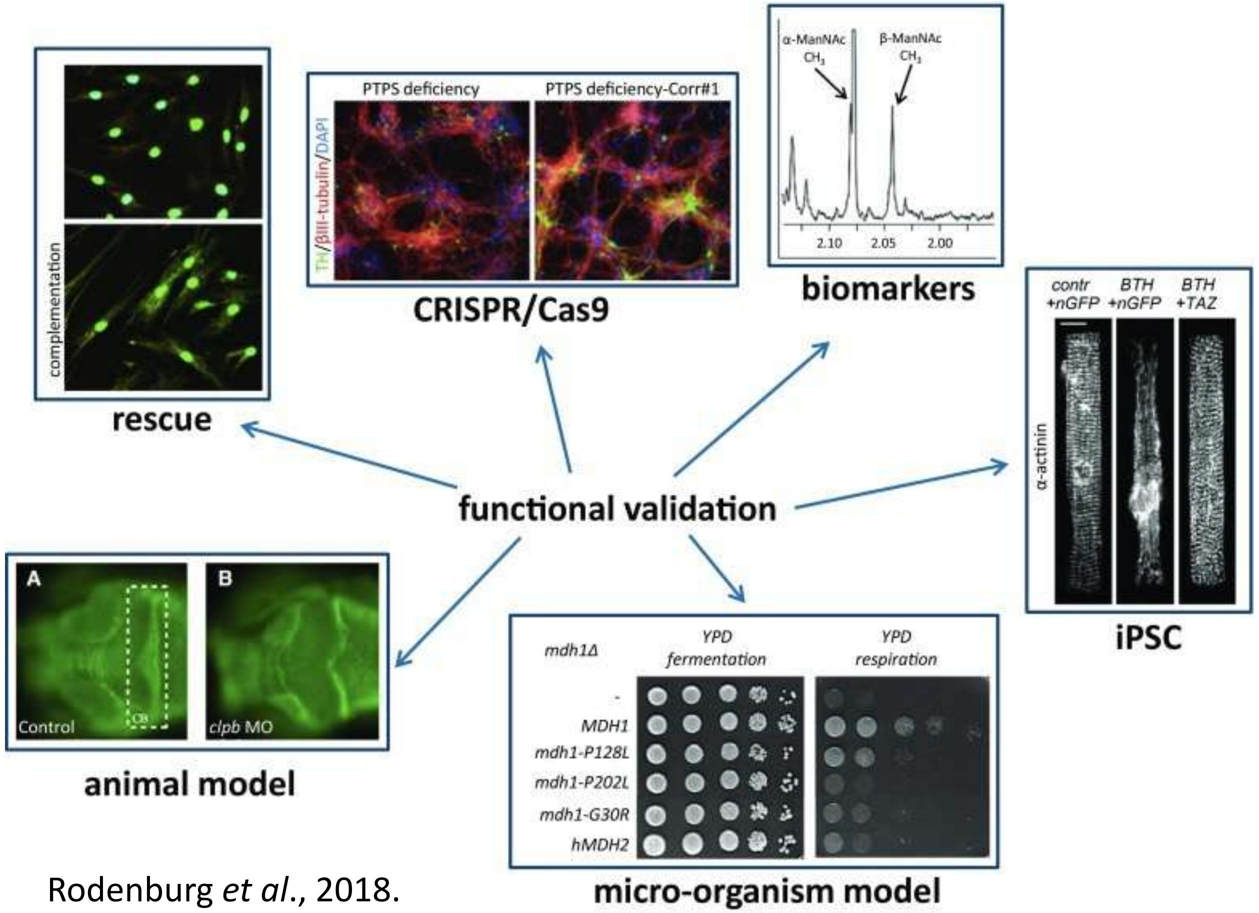
kidney function



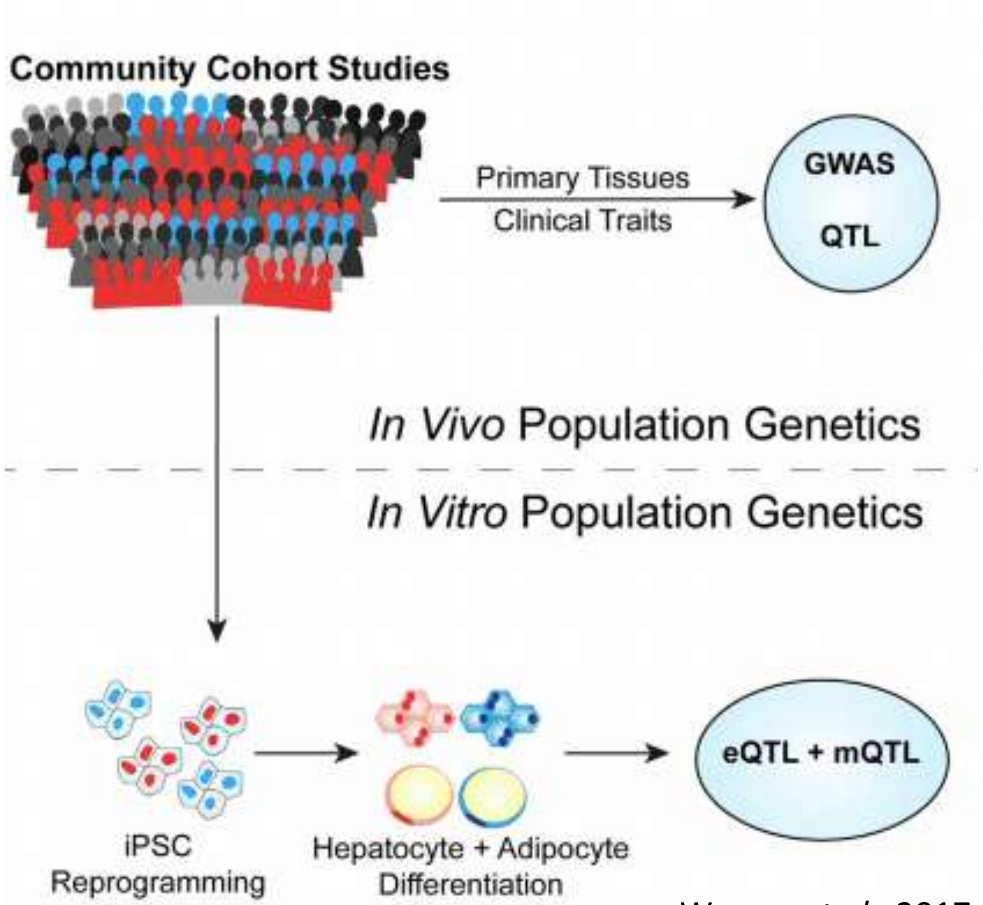
NOSTRIN
(Nitric Oxide Synthase Trafficking)

- neurotransmission
- inflammatory response
- vascular homeostasis

In vivo and in vitro follow-up of GWAS results



Rodenburg *et al*., 2018.


Community Cohort Studies
Primary Tissues
Clinical Traits
GWAS
QTL
In Vivo Population Genetics
In Vitro Population Genetics
iPSC Reprogramming
Hepatocyte + Adipocyte Differentiation
eQTL + mQTL

Warren *et al*., 2017.

Take-home message

- Genome-Wide Association Study: whole-genome SNP genotyping data analyzed without prior hypothesis
- GWAS follow-up analysis: do these SNPs have any functional consequence, causality, etc.?
- Computational analysis vs *in vitro* and *in vivo* studies

Thank you for your attention!

Contact

E-mail: dorakoller@ub.edu

@DoraKoller

Personal website