GWAS Summary Statistics

Pantelis Bagos

2024

Introduction

- Genome-wide association studies (GWAS) enable the simultaneously testing of thousands of genetic variants, usually SNPs, across the genome in order to find variants associated with a trait or a disease
- In the first years of the development of the field, efforts were oriented towards the statistical aspects of the analysis
- However, it was soon clear that most variants discovered via GWAS have small overall effects on disease susceptibility. Thus, it became evident that integrating data from multiple sources and developing reliable bioinformatics tools was a necessary step in order to address the complexity of the underlying genetic basis of common human diseases

Historical overview

- Soon after the publication of the first GWAS it also became evident that, at least theoretically, **individuals could be identified** in such cohorts even if only the summary statistics are available. This led to imposing strict control access for sharing individual patients' data (IPD) from GWAS

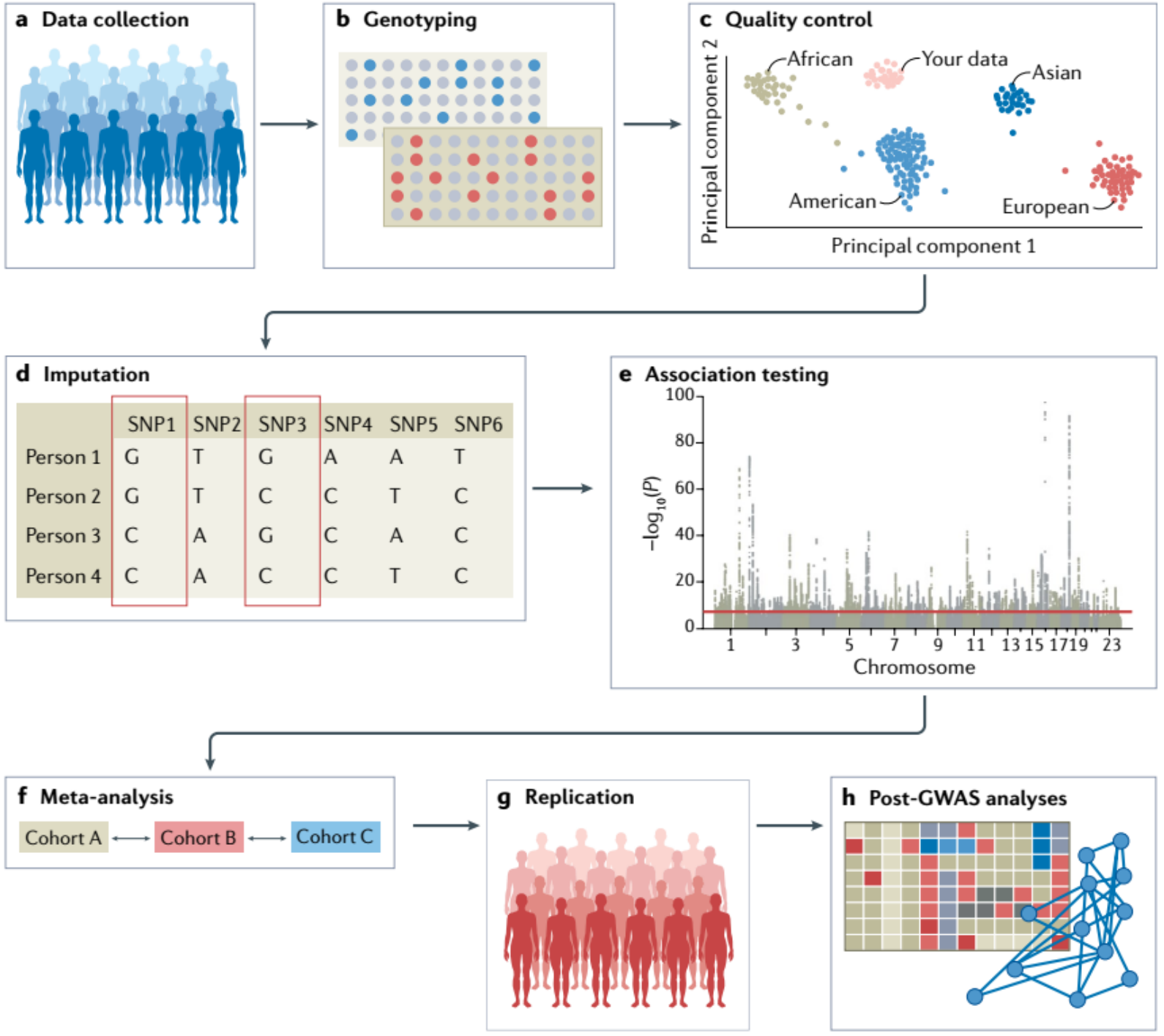
  Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. **Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays**. PLoS Genet. 2008;4(8): e1000167.

- Subsequent works found that privacy attacks are possible in theory but **unsuccessful and unconvincing in real practice**.

  Cai R, Hao Z, Winslett M, Xiao X, Yang Y, Zhang Z, et al. **Deterministic identification of specific individuals from GWAS results**. Bioinformatics. 2015;31(11):1701–7.

- In practice, however, **not all studies share their data**. It has been estimated that the proportion is only 13%, which increased from 3% in 2010 to 23% in 2017.

  Thelwall M, Munafo M, Mas-Bleda A, Stuart E, Makita M, Weigert V, et al. **Is useful research data usually shared? An investigation of genome-wide association study summary statistics**. PLoS ONE. 2020;15(2): e0229578.


a Data collection
b Genotyping
c Quality control
African
Your data
Asian
American
European
Principal component 2
Principal component 1
d Imputation
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
e Association testing
100
80
60
40
20
0
−log10(P)
1 3 5 7 9 11 13 15 17 19 23
Chromosome
f Meta-analysis
Cohort A
Cohort B
Cohort C
g Replication
h Post-GWAS analyses

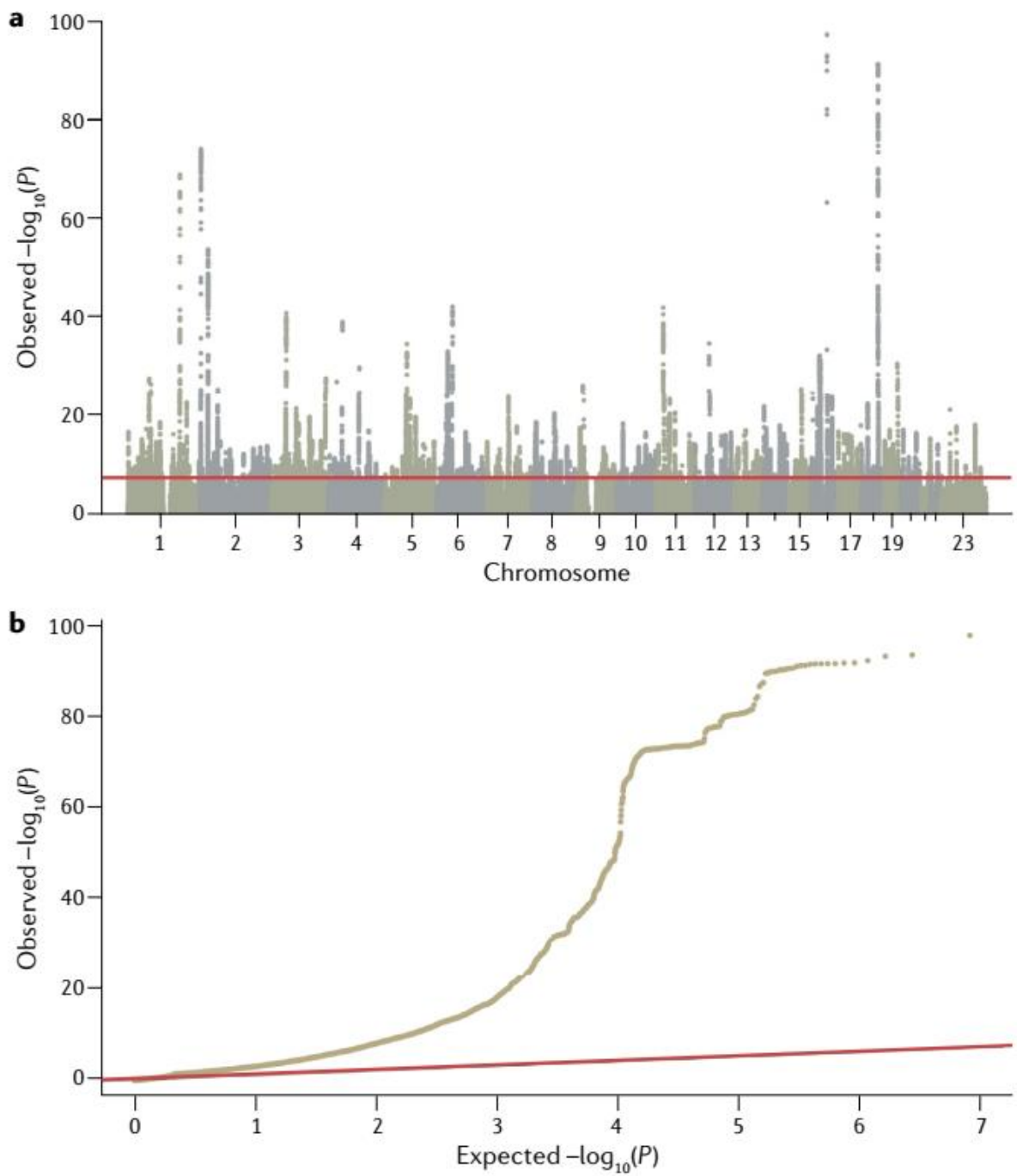
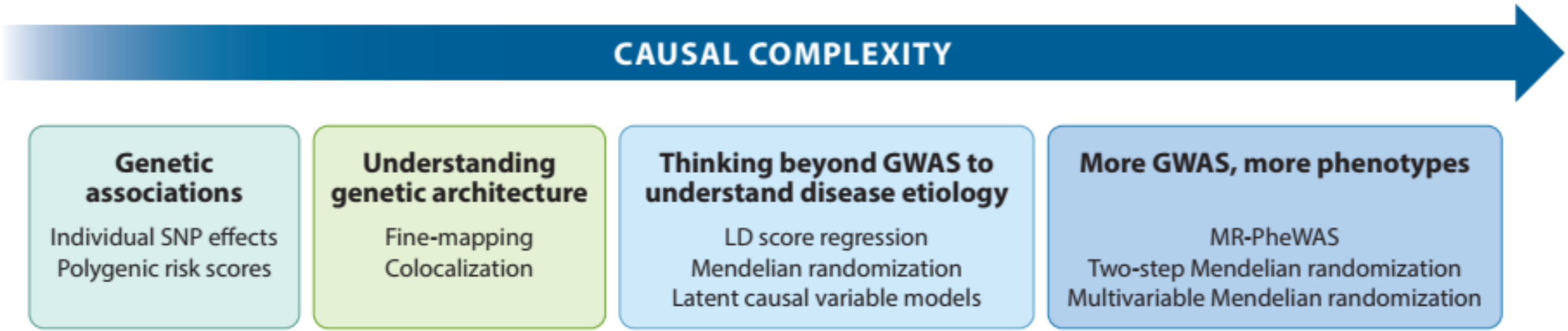


Fig. 2 | **Manhattan plot and quantile–quantile plot to visualize GWAS results.**


CAUSAL COMPLEXITY
Genetic associations
Individual SNP effects
Polygenic risk scores
Understanding genetic architecture
Fine-mapping
Colocalization
Thinking beyond GWAS to understand disease etiology
LD score regression
Mendelian randomization
Latent causal variable models
More GWAS, more phenotypes
MR-PheWAS
Two-step Mendelian randomization
Multivariable Mendelian randomization

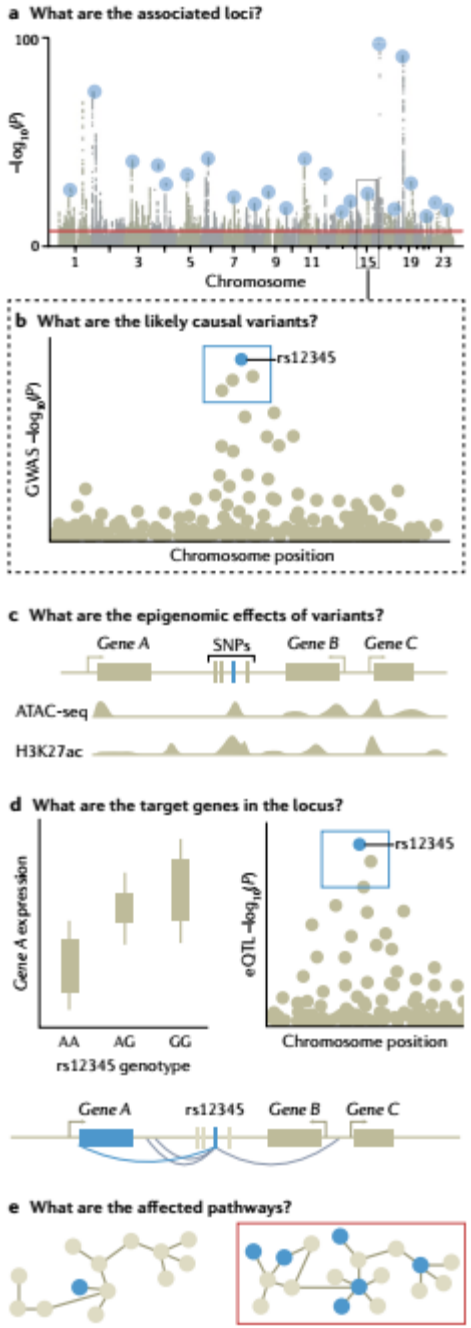


Fig. 3 | **Illustration of functional follow-up of GWAS.** **a** | Genome-wide association studies (GWAS) are conducted to identify associated variants, often visualized as a Manhattan plot to show their genomic positions and strength of association. **b** | To prioritize likely causal variants, statistical fine-mapping is applied to identify a set of variants that are likely to include the causal variant (blue box) as well as the most likely causal variant (rs12345; blue dot). Massively parallel reporter assays can be used to measure whether alleles differ in their ability to drive gene expression or other molecular activity for each variant (not shown). **c** | Functional annotations of the genome can be integrated with GWAS data to identify epigenetic mechanisms that may be perturbed by the causal variant, including enhancers, promoters or other functional elements. Additional approaches include mapping molecular quantitative trait loci (molQTL) or in vitro assays (not shown). **d** | Target gene for a GWAS locus can be prioritized by mapping expression quantitative trait loci (eQTLs) (left) and their co-localization (right) to identify loci where the causal variant from GWAS is also a causal variant affecting gene expression. For GWAS variants in enhancers, high-throughput chromosome conformation capture (Hi-C) data and maps of enhancer target genes can be used together with simple prioritization by distance to identify genes affected by the causal variant (below). **e** | To identify pathways whose perturbation may mediate the trait in question (red box), one can analyse the enrichment of multiple GWAS-implicated genes in predefined pathways. Additional approaches include *trans*-eQTL mapping and CRISPR perturbation of GWAS loci/genes followed by cellular phenotyping (not shown). For these analyses, the context of a relevant tissue, cell type and cell state needs to be carefully considered and analysed. ATAC-seq, assay for transposase-accessible chromatin using sequencing; H3K27Ac, histone H3 acetylated at K27; SNP, single-nucleotide polymorphism.

Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59.

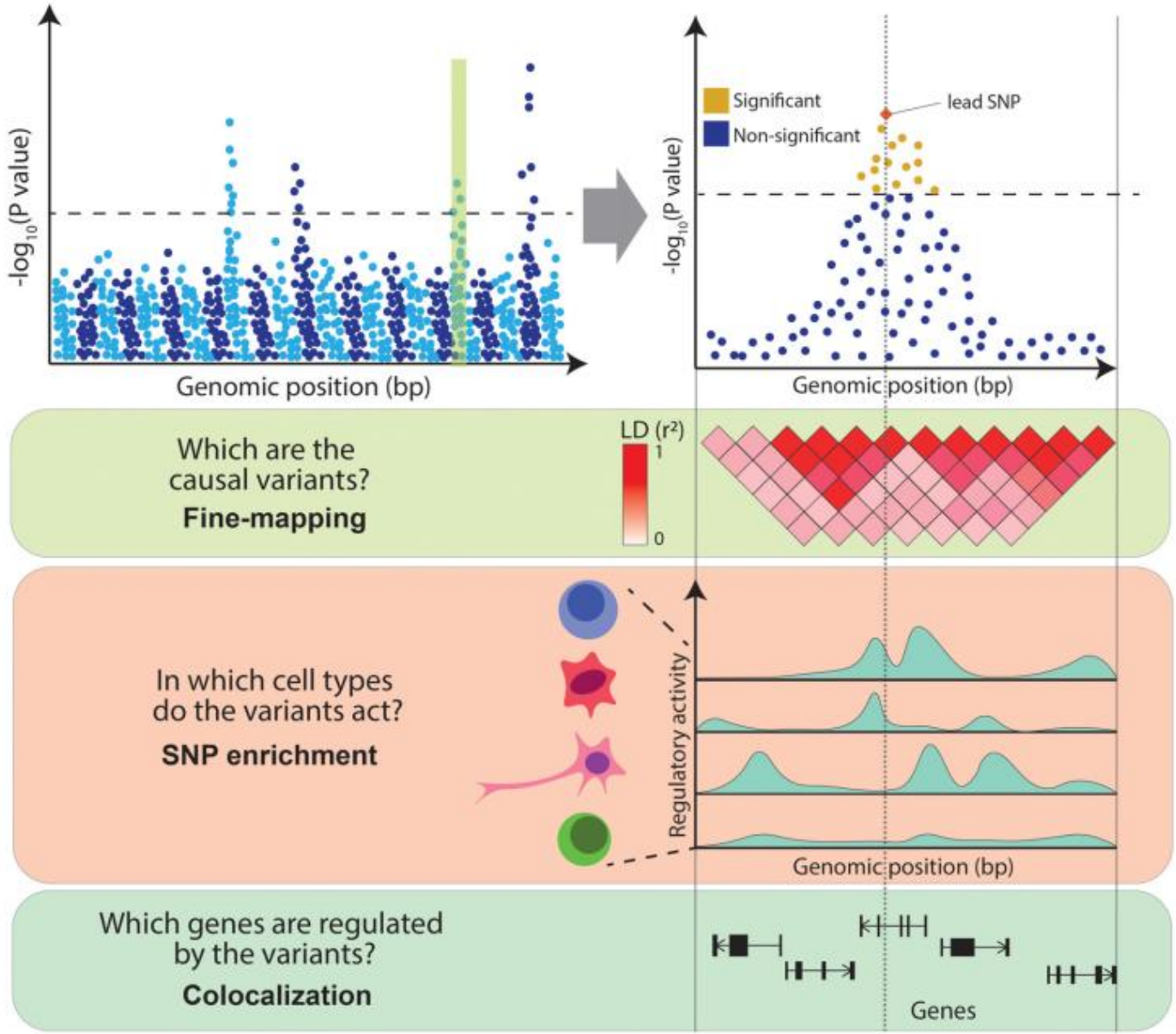
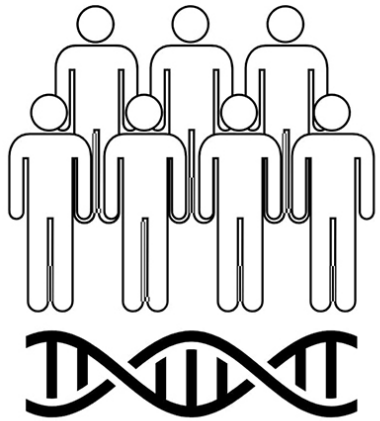
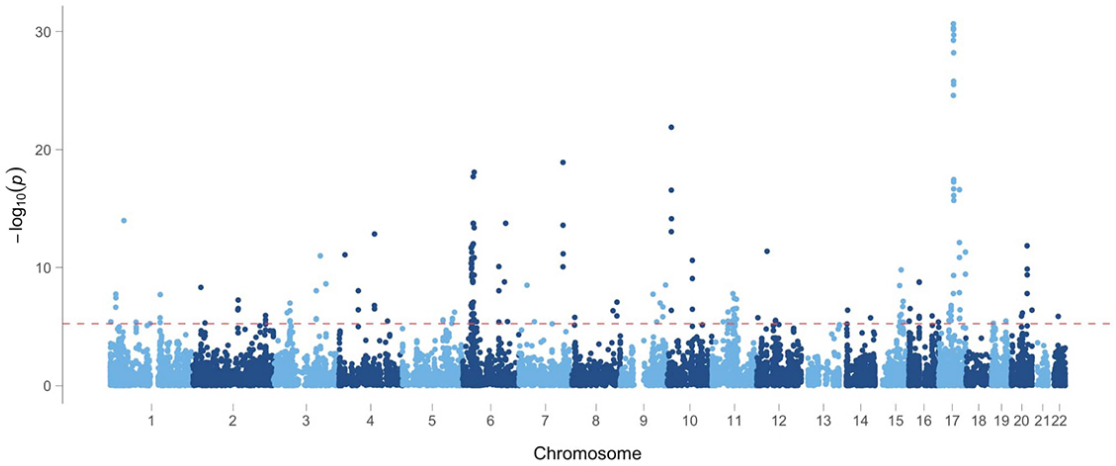


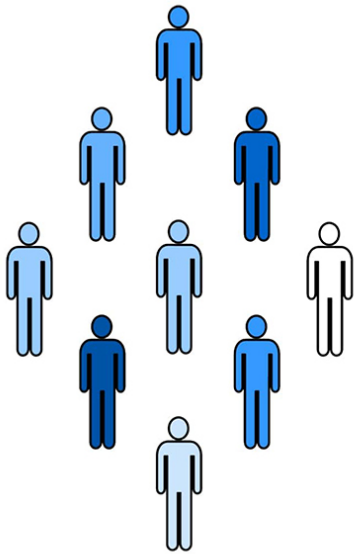
FIGURE 1 | Challenges in interpreting GWAS associations. From the top: Manhattan plot illustrates the association between genetic variants and a trait (e.g., a disease) at a genome-wide level (left panel) and within an example locus (right panel). Variants above the dotted line represent genome-wide significant associations. The panels below illustrate the main challenges in interpreting GWAS associations: high LD between variants (encoded in shades of red), variable levels of regulatory activity of the genomic regions across cell types (peaks of different heights represent different levels of activity of chromatin marks) and multiple genes within the associated locus.

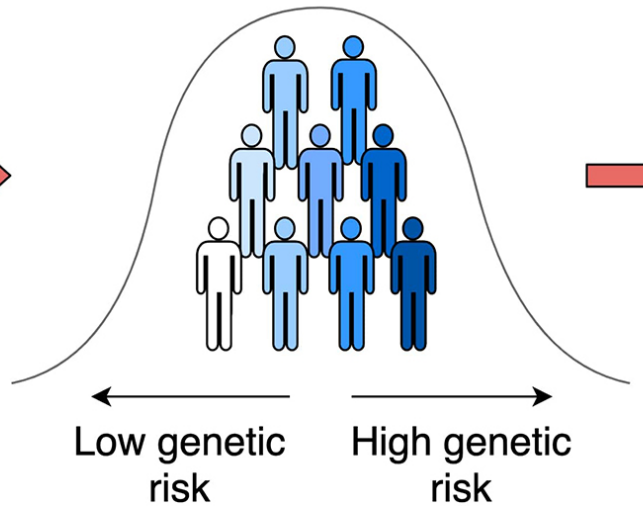
Step 1: Genome-wide association studies in adult populations from the UK Biobank

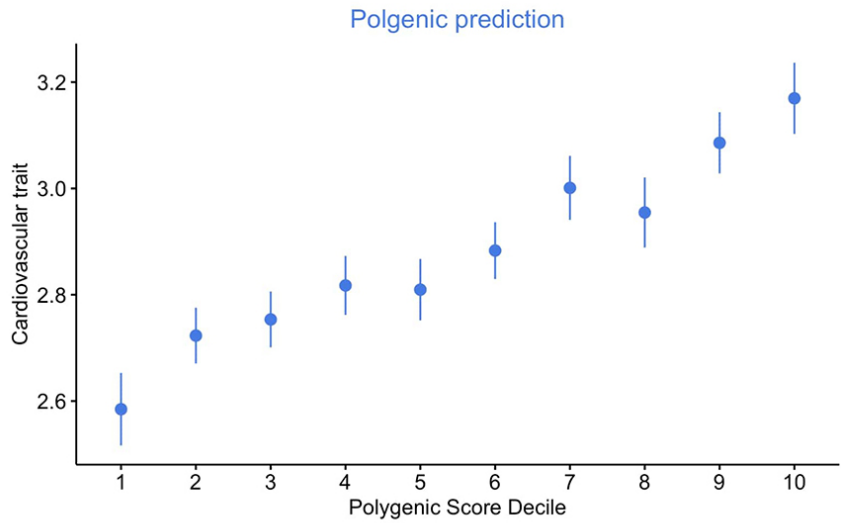


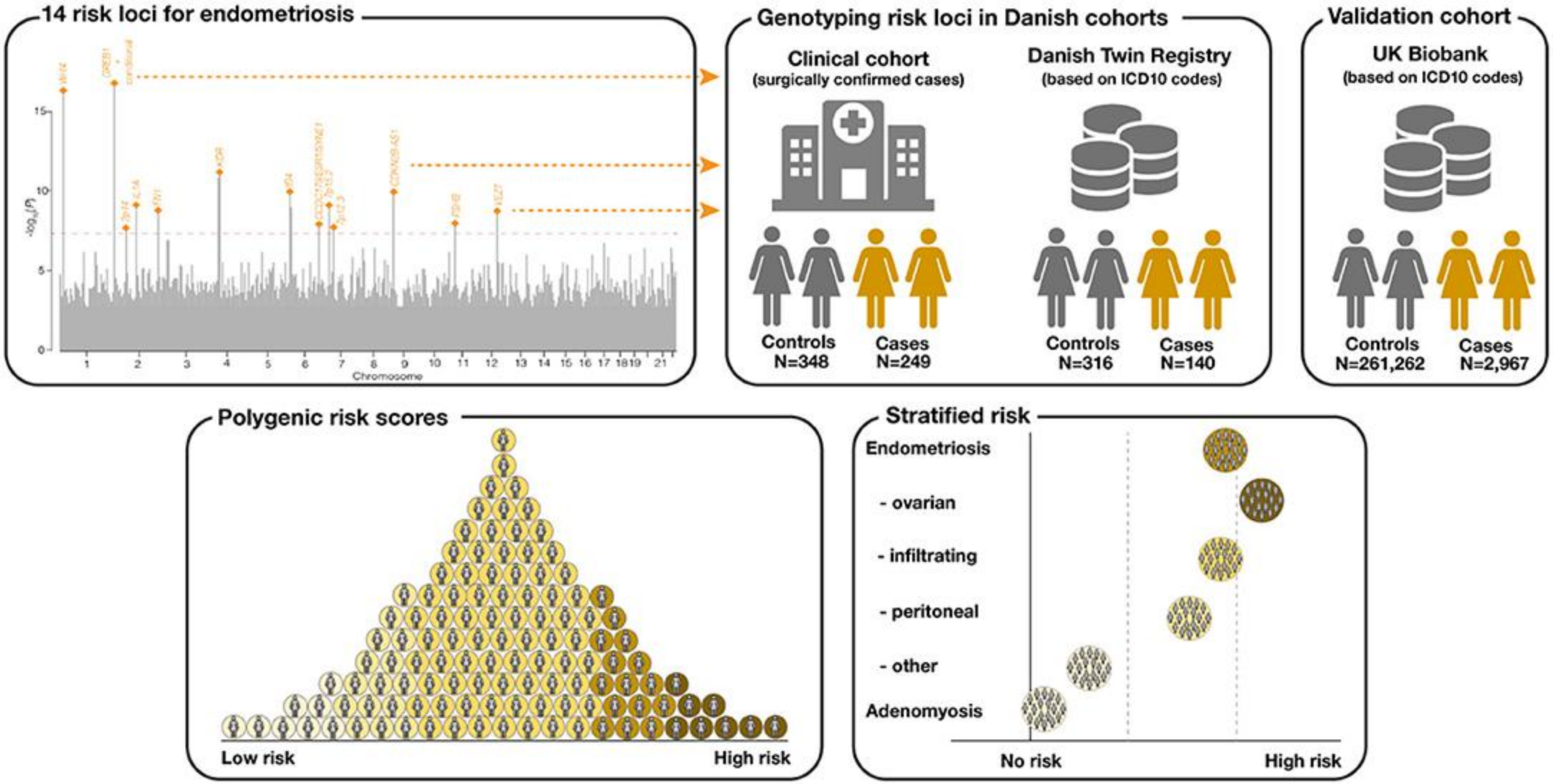


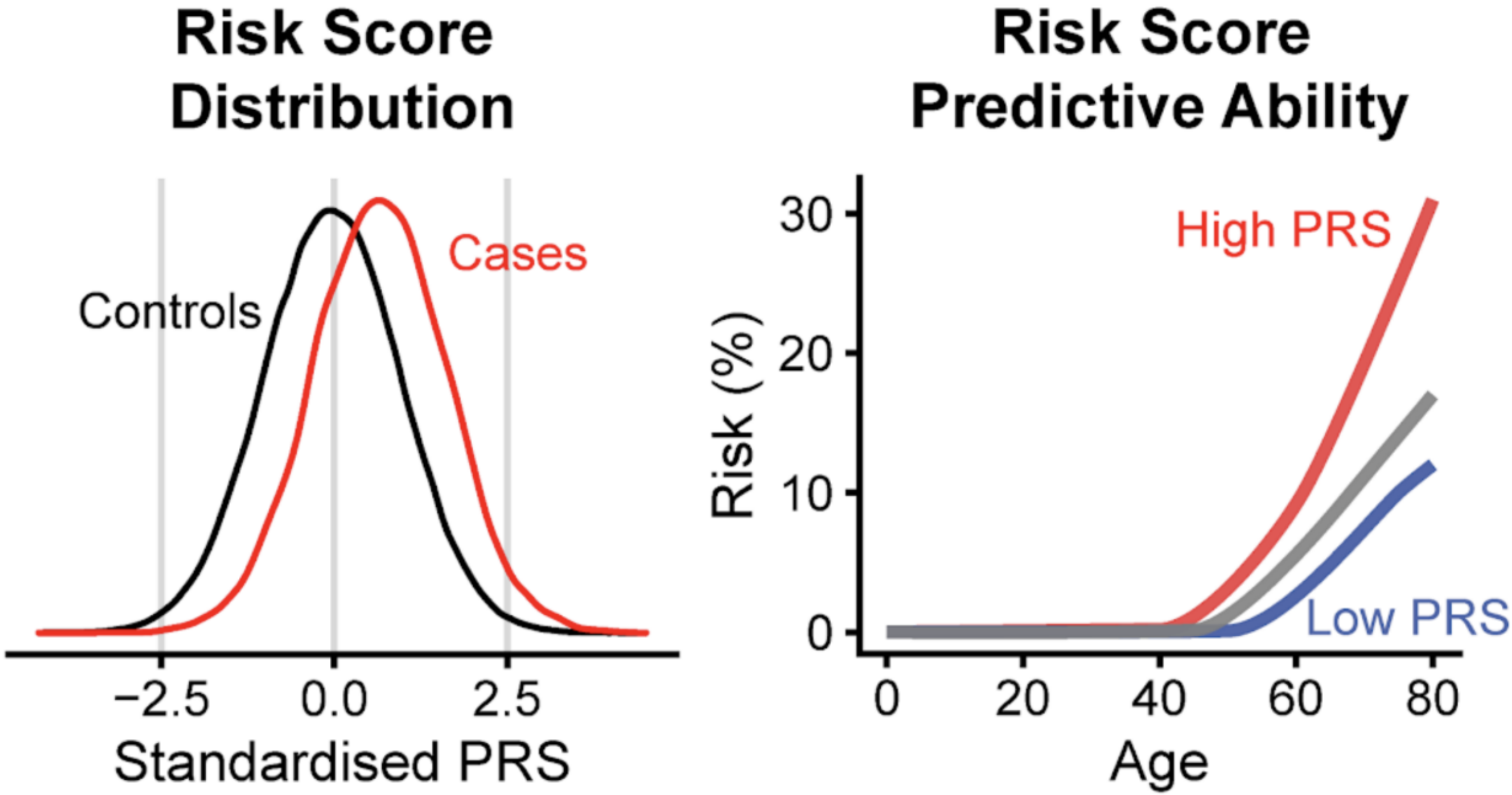
Step 2: Whole genome polygenic risk scores derived in children from the ALSPAC cohort





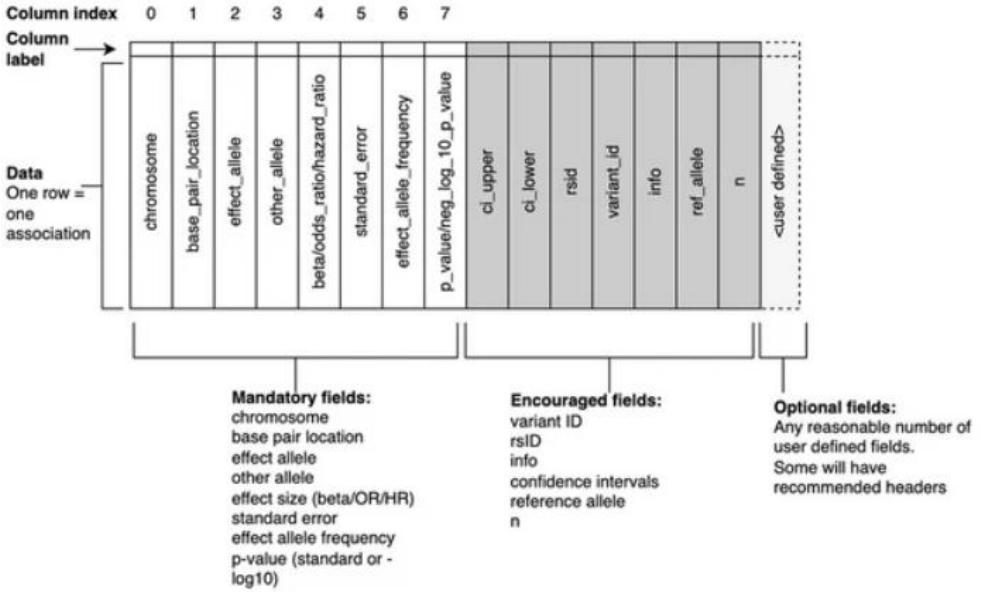



14 risk loci for endometriosis
Chromosome
Genotyping risk loci in Danish cohorts
Clinical cohort
(surgically confirmed cases)
Controls
N=348
Cases
N=249
Danish Twin Registry
(based on ICD10 codes)
Controls
N=316
Cases
N=140
Validation cohort
UK Biobank
(based on ICD10 codes)
Controls
N=261,262
Cases
N=2,967
Polygenic risk scores
Low risk
High risk
Stratified risk
Endometriosis
- ovarian
- infiltrating
- peritoneal
- other
Adenomyosis
No risk
High risk


Risk Score
Distribution
Controls
Cases
−2.5
0.0
2.5
Standardised PRS
Risk Score
Predictive Ability
Risk (%)
30
20
10
0
High PRS
Low PRS
0
20
40
60
80
Age

Summary statistics

- The **per-allele SNP effect sizes** (log odds ratios or betas)
  - along with their **standard errors**,
  - or equivalently the z-scores (per-allele effect sizes divided by their standard errors).
- Additional information required for mapping reasons, such as
  - the chromosome position,
  - the rsid,
  - the main and the alternative alleles and so on
- Not all studies share their data, and even when they do, **they do not share the same type of data**

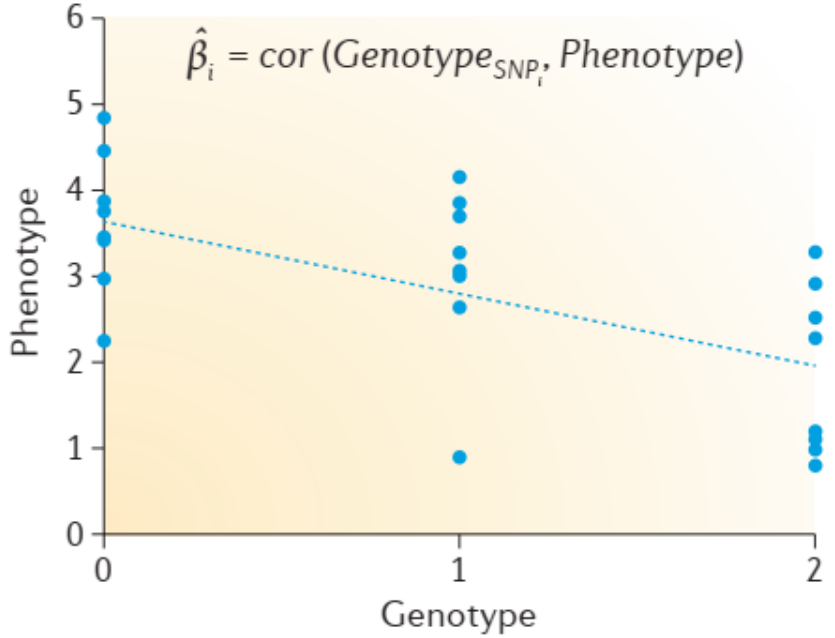


chromosome	base_pair_location	effect_allele	other_allele	beta	standard_error	effect_allele_frequency	p_value	variant_id	rsid	ref_allele
1	869388	A	G	-0.016619	0.00806496	0.997221	0.1	1_869388_A_G	#NA	EA
1	205813916	G	C	-0.0089589	0.00331941	0.983589	9.7E-03	1_205813916_G_C	rs74143855	EA
2	70478797	T	TG	0.0187528	0.00167685	0.934121	3.5E-30	2_70478797_T_TG	rs142640435	EA
7	8458030	TC	T	-0.0184003	0.00101051	0.78451	5.7E-76	7_8458030_TC_T	rs774624811	EA
23	24173186	A	C	0.00387762	0.08757958	0.627178	2.3E-08	23_24173186_C_A	rs5949233	OA

In this example, the summary statistics data file (TSV) has been pretty-printed to display the columns more clearly. The first line contains the column labels and every line thereafter are for variant-trait association data. Column labels and column order are in adherence to the definitions in Table 1. *variant_id* and *rsid* are optional (encouraged) they are simply placed anywhere after the mandatory 8 columns. Here the effect statistic is beta, so the column label of the effect size column is *beta*. The first data row represents a variant-trait association for a single-nucleotide polymorphism where the effect allele is an 'A' at the genomic location (genome assembly is given in the accompanying metadata file, see **??**). No rsID was provided for this first variant, so *#NA* was given as the value in the *rsid* column because there must be a value in all columns. The second data row shows an example where the p-value is given in scientific notation and rsID is provided. The third and fourth data rows are examples of deletions and insertions, respectively. The fifth example shows a variant located on the X-chromosome, which is mapped to *23* (Table 1).

https://github.com/EBISPOT/gwas-summary-statistics-standard

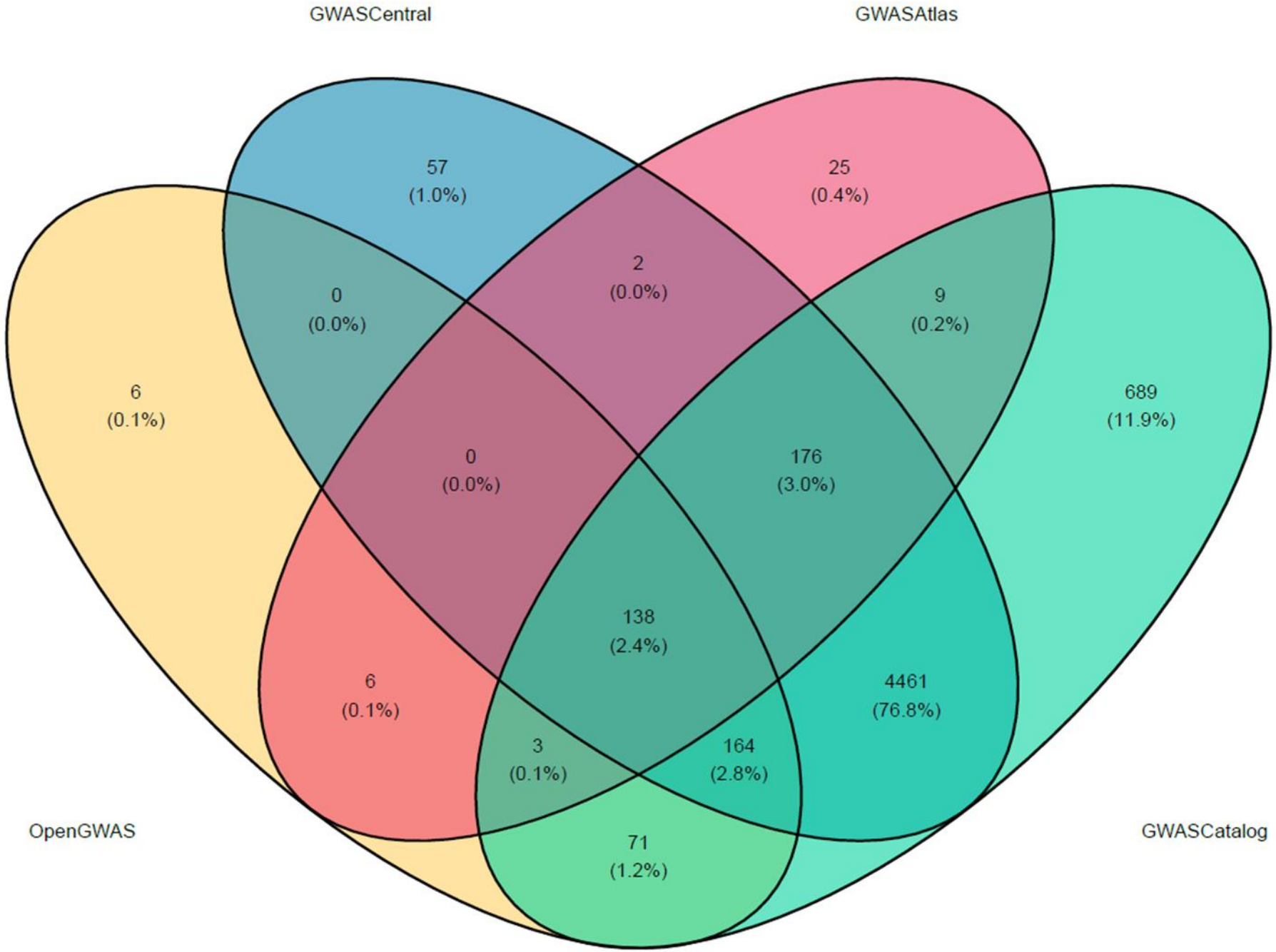
Effect size at SNP_i



z-scores

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)} \quad , \quad \cdots \quad , \quad \frac{\hat{\beta}_i}{s.e(\hat{\beta}_i)} \quad , \quad \cdots \quad , \quad \frac{\hat{\beta}_M}{s.e(\hat{\beta}_M)} \quad \sim MVN\,(0, V)$$

SNP LD (*V*)


GWASCentral
GWASAtlas
57
(1.0%)
25
(0.4%)
2
(0.0%)
0
(0.0%)
9
(0.2%)
6
(0.1%)
689
(11.9%)
0
(0.0%)
176
(3.0%)
138
(2.4%)
6
(0.1%)
4461
(76.8%)
3
(0.1%)
164
(2.8%)
71
(1.2%)
OpenGWAS
GWASCatalog

Summary statistics

- Summary statistics do not only offer the additional **protection of privacy**, but also offer significant advantages in **computational cost** when using the data in downstream analyses, which does not scale with the number of participants in the study
- Thus, it is of no surprise that during the last years a large variety of methods have been developed to perform a **so-called post-GWAS analysis using the summary results** of a single study, or of several studies, and in most cases integrating data from other sources

Post-GWAS analysis

- These methods seek to go a step further from the simple analysis, or re-analysis of a study, and aim to improve our understanding about the functional role of the identified variants
- Three are the key factors in these methods:
  - **linkage disequilibrium (LD)** information from a population reference panel such as HapMap or 1000 Genomes Project,
  - the gene expression variation in the form of **eQTL**, and
  - the integration of functional information **on biological pathways**

Kontou PI, Bagos PG. **The goldmine of GWAS summary statistics: a systematic review of methods and tools.** BioData Min. 2024 Sep 5;17(1):31

Types of tools

- Data
- Single trait analysis
- Multiple trait analysis

Data

- Database
- Quality Control
- Genotype Reconstruction
- Imputation

Single trait analysis

- Meta-analysis
- Heritability analysis
- Gene-based tests
- GSA
- Fine-mapping

Multiple trait analysis

- Genetic Correlation (GC)
- Pleiotropy analysis
- Mendelian Randomization (MR)
- Colocalization
- Transcriptome-wide association studies (TWAS)

Post-GWAS analysis

- These methods seek to go a step further from the simple analysis, or re-analysis of a study, and aim to improve our understanding about the functional role of the identified variants
- Three are the key factors in these methods:
  - **linkage disequilibrium (LD)** information from a population reference panel such as HapMap or 1000 Genomes Project,
  - the gene expression variation in the form of **eQTL**, and
  - the integration of functional information **on biological pathways**

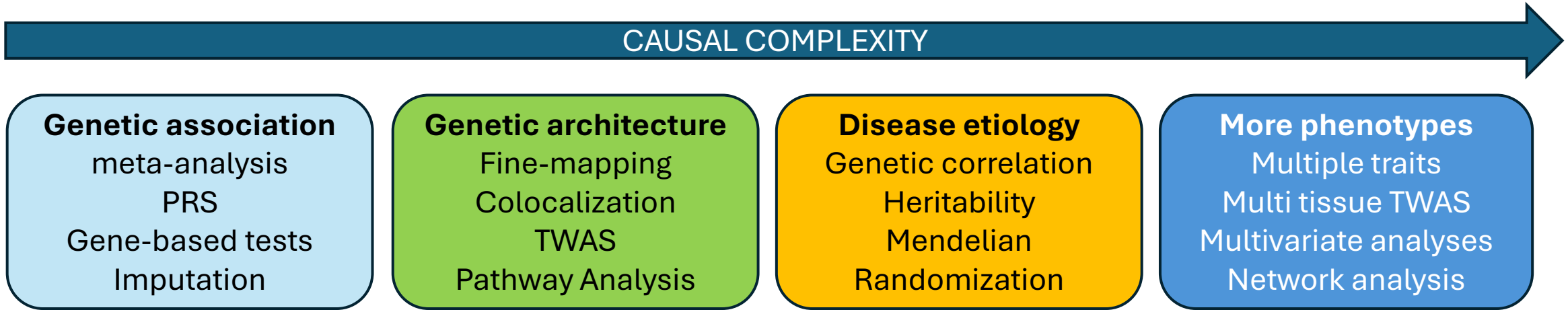
Kontou PI, Bagos PG. **The goldmine of GWAS summary statistics: a systematic review of methods and tools.** BioData Min. 2024 Sep 5;17(1):31

Single trait analysis

- Meta-analysis
- Heritability analysis
- Gene-based tests
- Fine-mapping
- Pathway Analysis

Multiple trait analysis

- Genetic Correlation (GC)
- Pleiotropy analysis
- Mendelian Randomization (MR)
- Colocalization
- Transcriptome-wide association studies (TWAS)

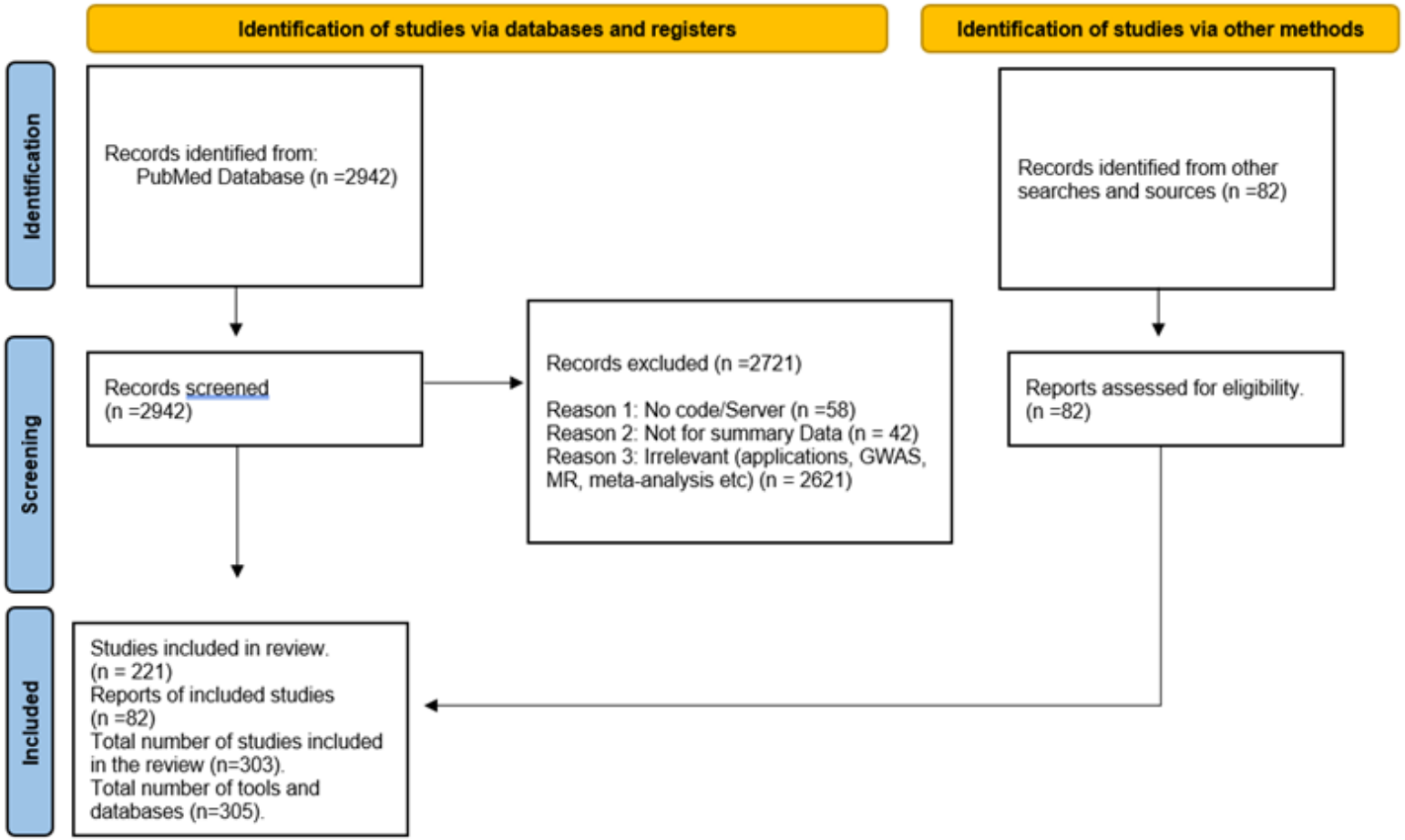


Kontou PI, Bagos PG. **The goldmine of GWAS summary statistics: a systematic review of methods and tools.** BioData Min. 2024 Sep 5;17(1):31

Types of methods

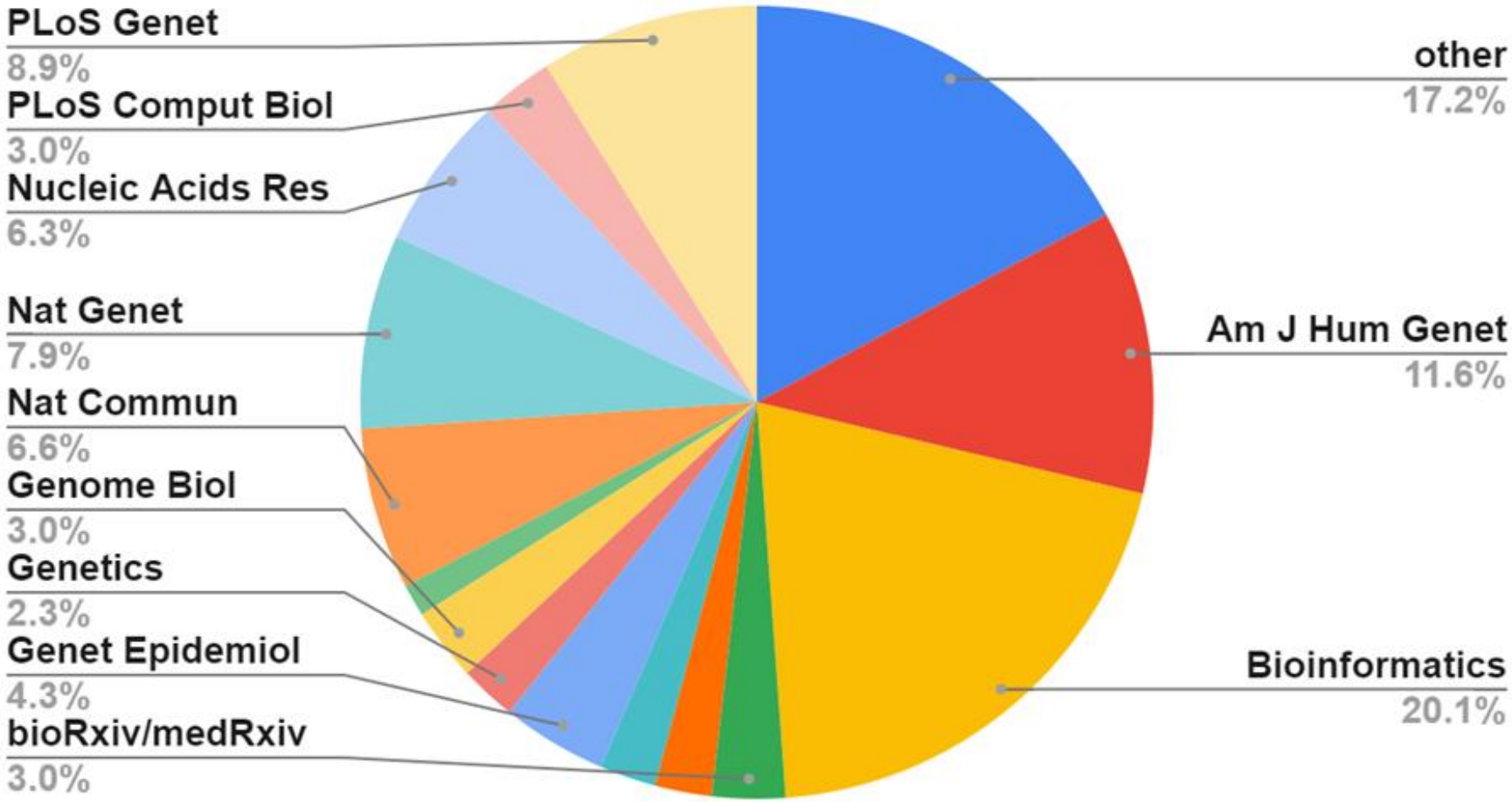
GWAS + GWAS	→	Replication, Meta-analysis (PLINK)
GWAS + LD	→	Imputation, heritability (LDSC), gene-based tests (GCTA), fine-mapping (CAVIAR)
GWAS + Pathways	→	Pathway analysis (g:Profiler)
GWAS + GWAS	→	Mendelian Randomization (MRBASE), Pleiotropy (ACA), Genetic Correlation (LDSC)
GWAS + eQTL	→	TWAS (FUSION), Colocalization (COLOC)
GWAS + xQTL+LD	→	Multi-tissue TWAS, Multi-omics analysis

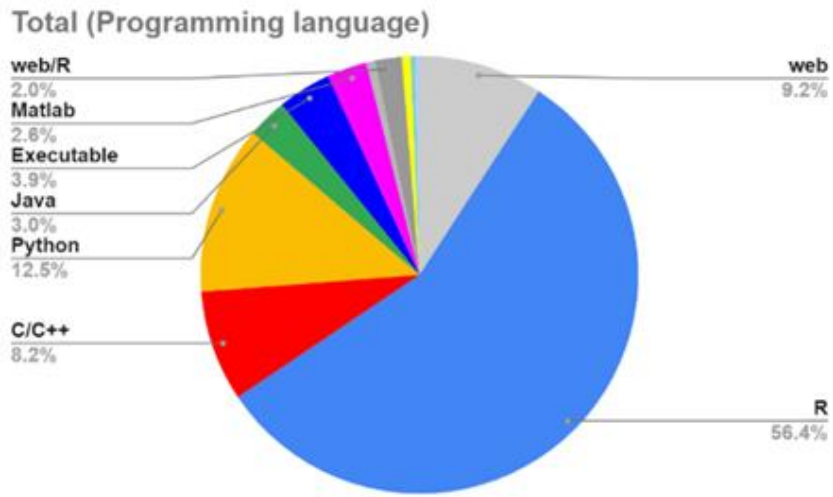
PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources.

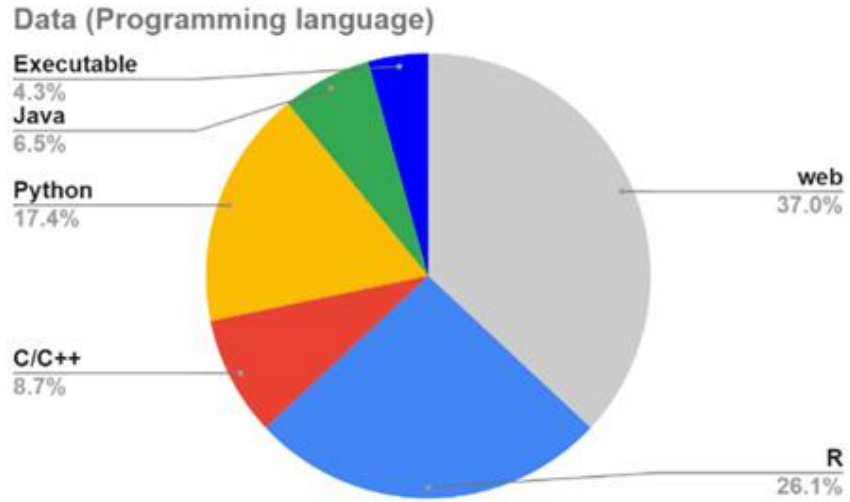


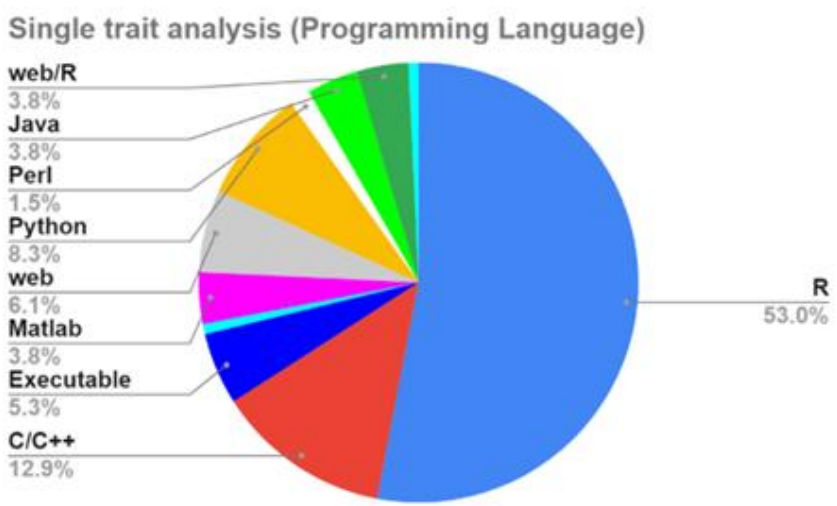
From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. For more information, visit: http://www.prisma-statement.org/

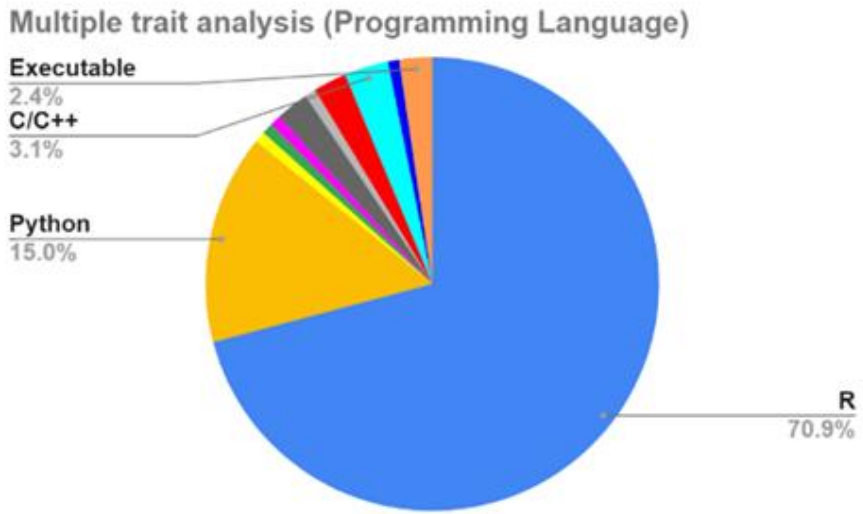
Journals

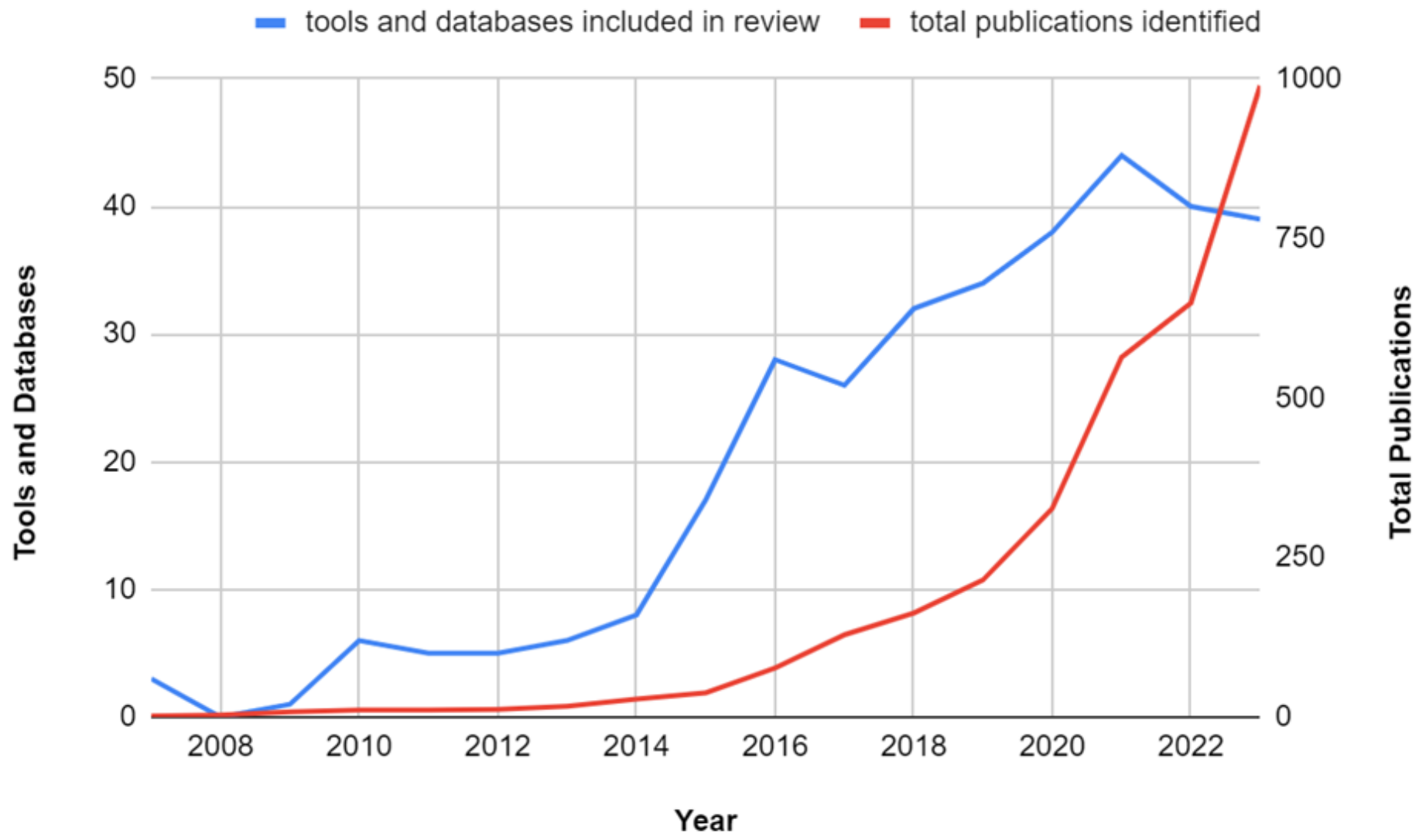

PLoS Genet
8.9%
PLoS Comput Biol
3.0%
Nucleic Acids Res
6.3%
Nat Genet
7.9%
Nat Commun
6.6%
Genome Biol
3.0%
Genetics
2.3%
Genet Epidemiol
4.3%
bioRxiv/medRxiv
3.0%
other
17.2%
Am J Hum Genet
11.6%
Bioinformatics
20.1%


Total (Programming language)
web/R
2.0%
Matlab
2.6%
Executable
3.9%
Java
3.0%
Python
12.5%
C/C++
8.2%
web
9.2%
R
56.4%


Data (Programming language)
Executable
4.3%
Java
6.5%
Python
17.4%
C/C++
8.7%
web
37.0%
R
26.1%


Single trait analysis (Programming Language)
web/R
3.8%
Java
3.8%
Perl
1.5%
Python
8.3%
web
6.1%
Matlab
3.8%
Executable
5.3%
C/C++
12.9%
R
53.0%


Multiple trait analysis (Programming Language)
Executable
2.4%
C/C++
3.1%
Python
15.0%
R
70.9%


tools and databases included in review
total publications identified
50
40
30
20
10
0
1000
750
500
250
0
2008
2010
2012
2014
2016
2018
2020
2022
Tools and Databases
Total Publications
Year

Standards and Quality Control

- Generally, it is acceptable that the minimum information contained in GWAS summary statistics should include these: the variant ID or chromosome plus base pair location, the p-value of the association, the risk allele and the other allele, the risk allele frequency, and an estimate of the effect size (odds ratio or beta) along with its standard error
- **GWAS-SSF** consists of a tab-separated data file with well-defined fields and an accompanying metadata file. Most repositories and programs use some variant of the GWAS-SSF
- This way, the VCF was adapted to include GWAS-specific metadata utilizing the sample column to store variant-trait association data. The **GWAS-VCF** is the standard used by the MRC-IEU OpenGWAS database and it comes with appropriate tools to map GWAS summary statistics to VCF with on-the-fly harmonization

Standards and Quality Control (2)

- Tools belonging to the former class were developed early and were focused mainly on harmonizing data in preparation of a meta-analysis. These include **QCGWAS**, **GWAtoolbox** and **EasyQC**. **GEAR** is very interesting in that it incorporates ideas from population genetics which allow verification of the genetic origin and geographic location of each cohort and identifying significant sample overlap. More recent tools like **MungeSumstats** and **GWASlab** perform standardization and quality control handling the most common formats, **SumStatsRehab** can be used for data validation, restoration of missing data, correction of errors or formatting, and **GWASinspector** provides extensive QC reports and perform harmonization being compatible with recent reference panels and by handling insertion/deletion and multi-allelic variants
- The latter class of methods, additionally, leverages information from the LD among SNPs. One such tool is **GQS** which Identifies suspicious regions and prevents erroneous interpretations by comparing the significance of the association for each SNP to its LD value for the reported index SNP. Similar functionalities are offered by **DENTIST** which uses LD to detect and eliminate errors and disagreements between GWAS data and the LD reference panel. **EXTminus23andMe** evaluates the quality of summary statistics after data removal and the suitability of the downsampled summary statistics for typical follow-up genetic analyses

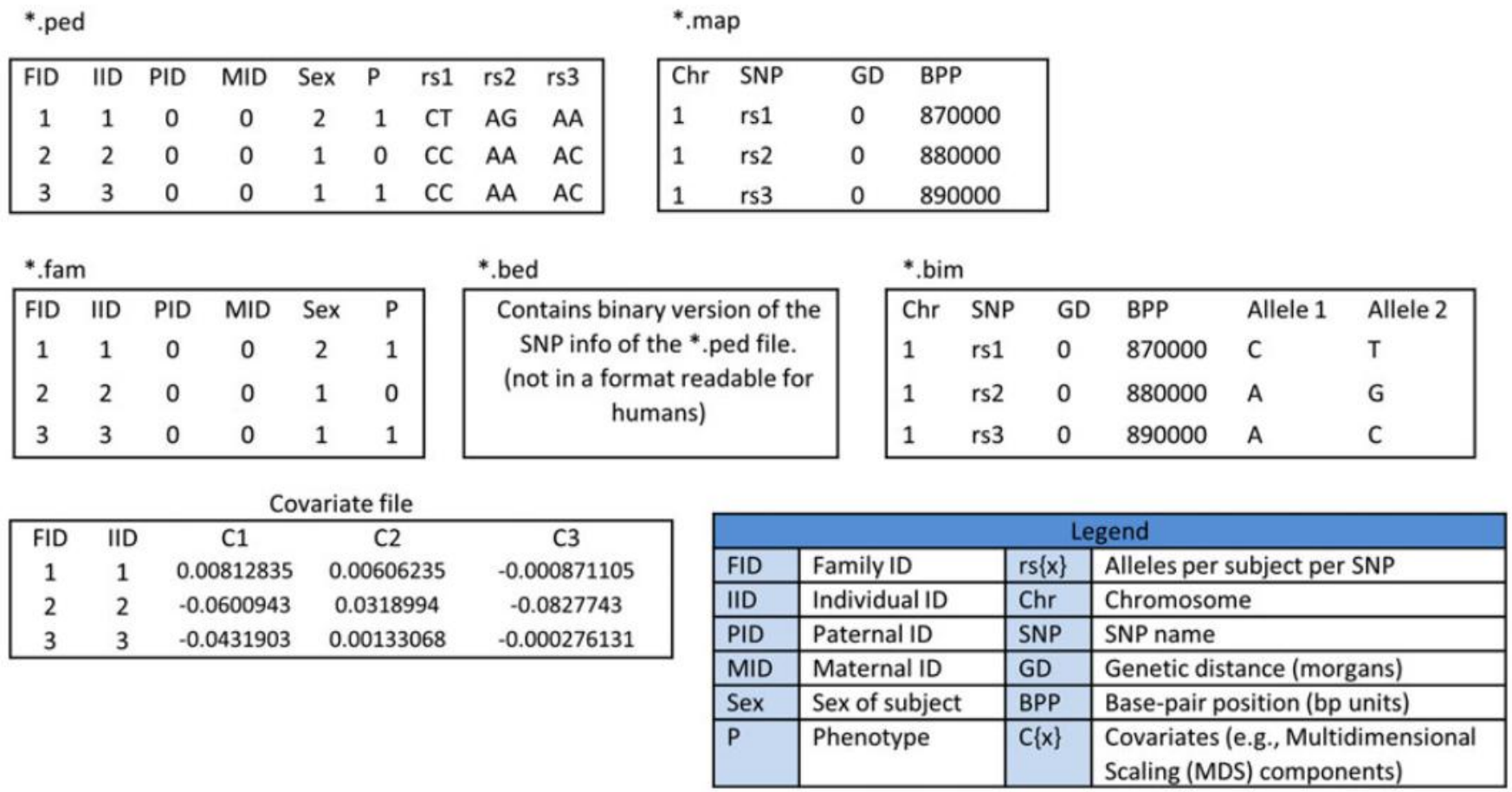
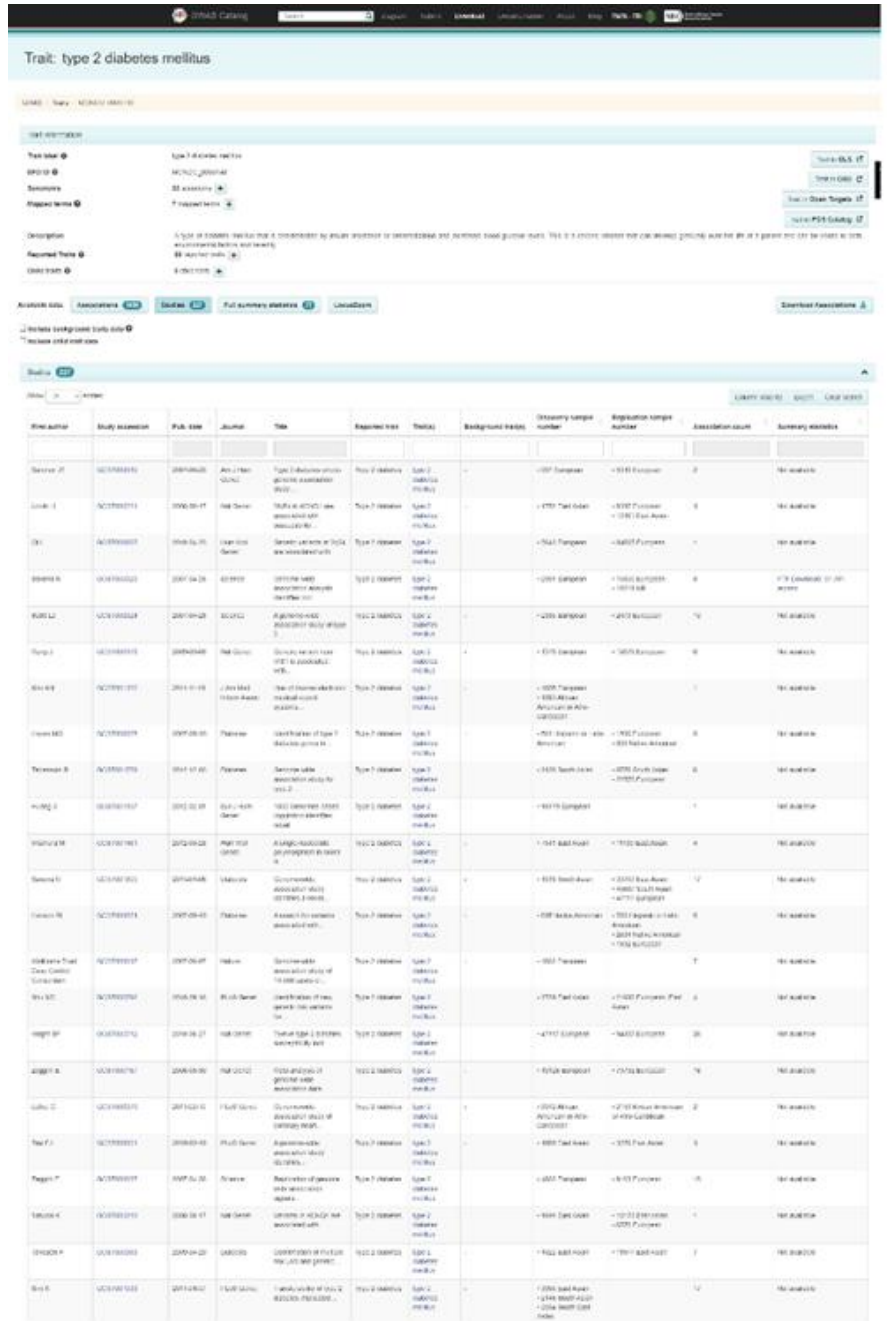
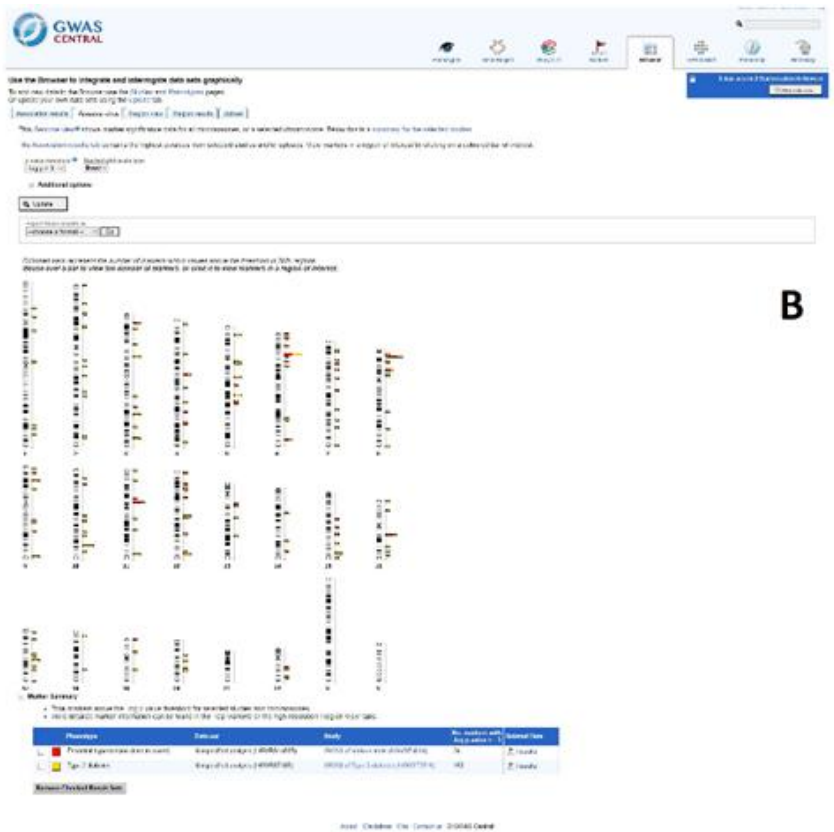


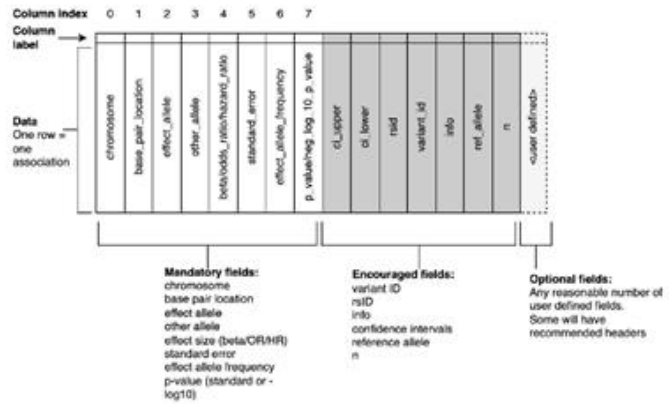
FIGURE 1 Overview of various commonly used PLINK files. SNP = single nucleotide polymorphism


Trait: type 2 diabetes mellitus

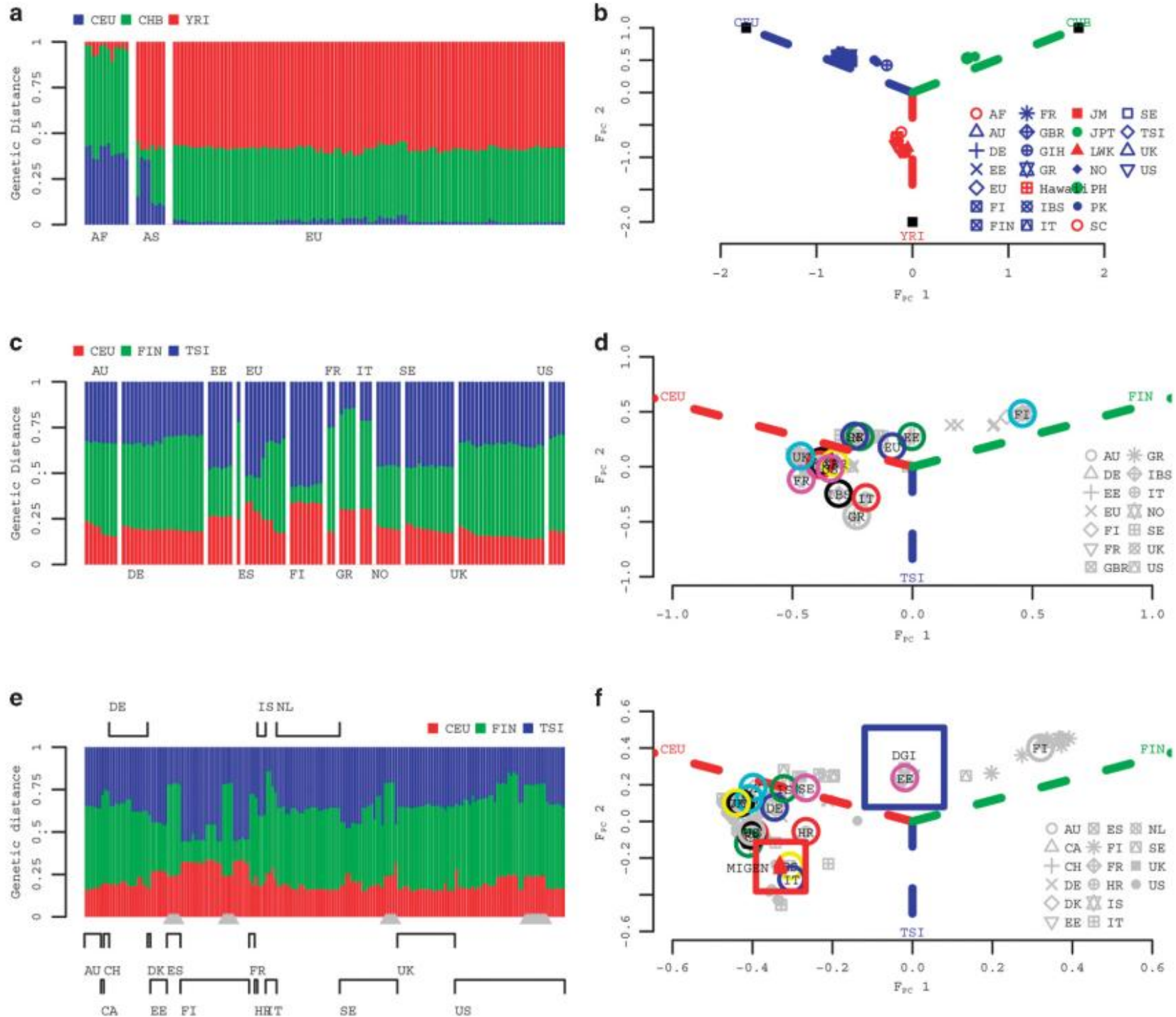
A


GWAS CENTRAL

B


Column index 0 1 2 3 4 5 6 7
Column label
Data
One row = one association
chromosome
base_pair_location
effect_allele
other_allele
beta/odds_ratio/hazard_ratio
standard_error
effect_allele_frequency
p_value/neg_log_10_p_value
ci_upper
ci_lower
rsid
variant_id
info
ref_allele
n
<user defined>
Mandatory fields:
chromosome
base pair location
effect allele
other allele
effect size (beta/OR/HR)
standard error
effect allele frequency
p-value (standard or -log10)
Encouraged fields:
variant ID
rsID
info
confidence intervals
reference allele
n
Optional fields:
Any reasonable number of user defined fields.
Some will have recommended headers

C


a
CEU CHB YRI
Genetic Distance
AF AS EU
b
CEU
CHB
YRI
AF AU DE EE EU FI FIN
FR GBR GIH GR Hawaii IBS IT
JM JPT LWK NO PH PK SC
SE TSI UK US
F_{PC} 1
F_{PC} 2
c
CEU FIN TSI
AU EE EU FR IT SE US
DE ES FI GR NO UK
d
CEU
FIN
TSI
AU DE EE EU FI FR GBR
GR IBS IT NO SE UK US
e
DE IS NL
CEU FIN TSI
AU CH DK ES FR UK
CA EE FI HR IT SE US
f
CEU
FIN
TSI
DGI
MIGEN
AU CA CH DE DK EE ES FI FR HR IS IT NL SE UK US

Databases (primary)

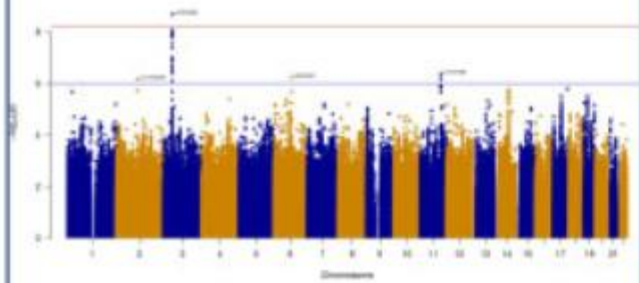
- NCBI's **dbGAP** was developed to contain the results of studies investigating the interaction of genotype and phenotype, which include GWAS. Summary statistics are generally available to the public, whereas access to IPD requires varying levels of authorization.
- The **NHGRI-EBI GWAS Catalog**, which was established in 2008 is considered for years the central repository of GWAS summary statistics. It is a high-quality curated collection of all published GWAS
- **GWAScentral** previously known as the Human Genome Variation (HGV) database of Genotype-to-Phenotype information is a database that contains over 72.5 million P-values for over 5,000 studies, with over 7.4 million unique genetic markers involved in more than 1,700 unique phenotypes. The database contains data from several sources
- The **IEU MRC OpenGWAS** is a new addition and contains 346 million genetic associations from 50,037 GWAS summary datasets. It contains complete data from various consortia and the UK Biobank and comes with a lot of tools for harmonizing the data and storing them in the GWAS-VCF format.
- **GeneATLAS and GBE** contain associations from the UK Biobank cohort. GBE contains summary statistics from over 750,000 individuals combining data from the UK Biobank, the Million Veterans Program and the Biobank Japan.
- **GTEx** and **QTLbase** are the primary resources for xQTL data

Databases (secondary)

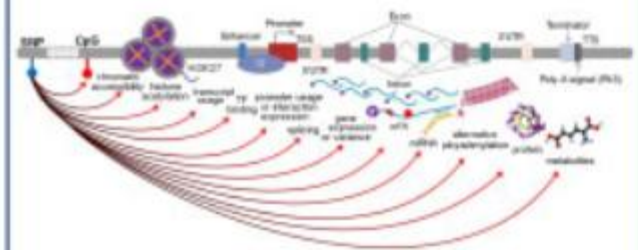
- **TSEA-DB** and **PCGA** use information from gene-expression in various tissues to perform tissue or cell-type enrichment analysis of the GWAS association statistics.
- **webTWAS** and **COLOCdb** also use information on eQTL but in different fashion. webTWAS contains data for GWAS for which it calculates the causal genes using single tissue expression imputation (using MetaXcan and FUSION), or cross-tissue expression imputation (using UTMOST). **COLOCdb** on the other hand is the most comprehensive colocalization analysis by integrating publicly available GWASs with different types of xQTL and different algorithms (COLOC, SMR).
- **GWAS ATLAS** contains results of GWAS accompanied by useful information obtained from downstream analysis. Each study is accompanied by MAGMA results, SNP heritability estimation and genetic correlations with other traits in the database.
- **GWASROCS** contains a large and comprehensive set of SNP-derived AUROCs and heritabilities.
- Phenome-wide association studies (PheWAS) invert the idea of a GWAS by searching for phenotypes associated with specific variants across the range of thousands of human phenotypes, or the "phenome. Thus, it is expected that a PheWAS will need large databases of GWAS results. **PhenoScanner** is the most complete such database with publicly available results. Similar functionalities are offered also by **OpenGWAS**, **GWAS ATLAS** and **PheWAS Catalog**.
- Lastly, we need to mention **LD Hub**, a centralized database of publicly available GWAS results for 173 diseases/traits which offers a web interface that automates the LD score regression (LDSC) analysis pipeline

Data

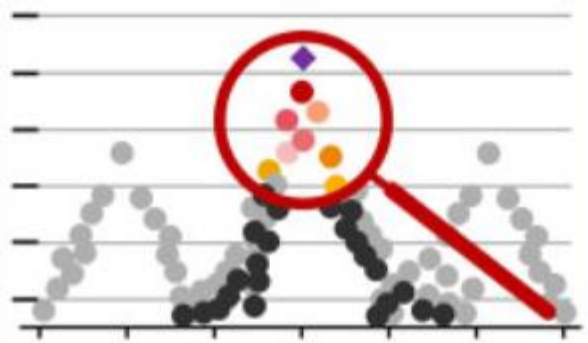
> 3000 GWASs



13 types of xQTL
188 xQTL datasets



Colocalization



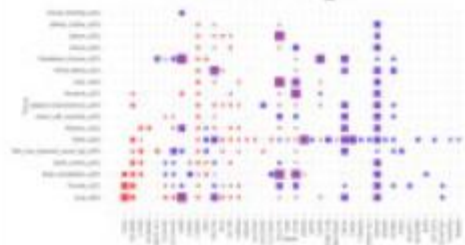
3 modules: GWAS-GWAS

GWAS-xQTL

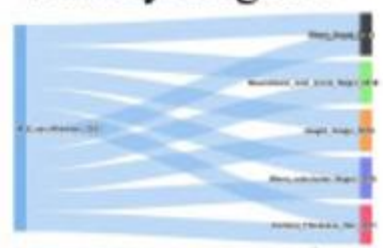
xQTL-xQTL

Visualization

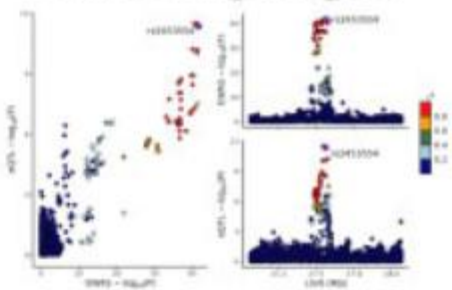
Heat map



Sankey diagram



Locus compare plot



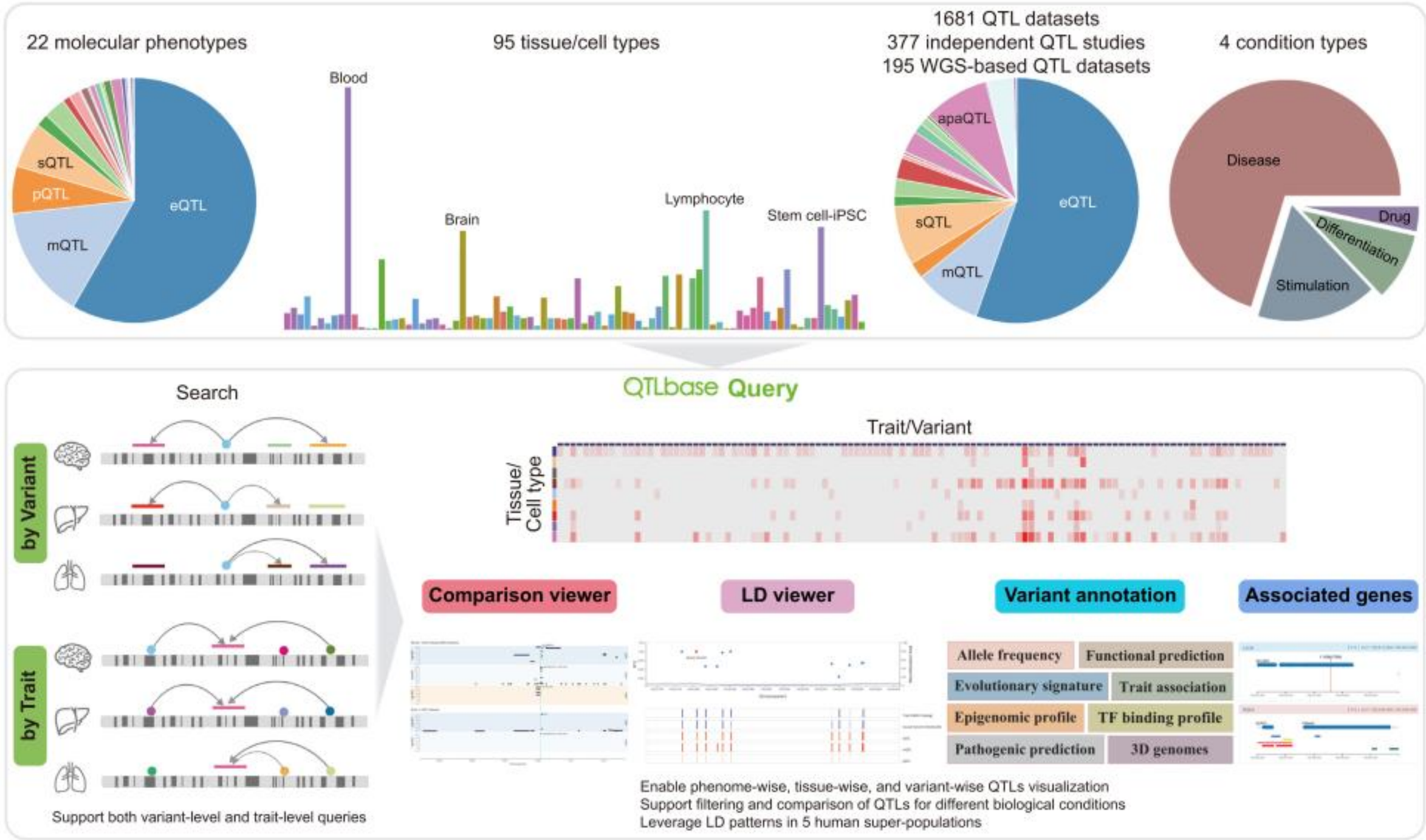
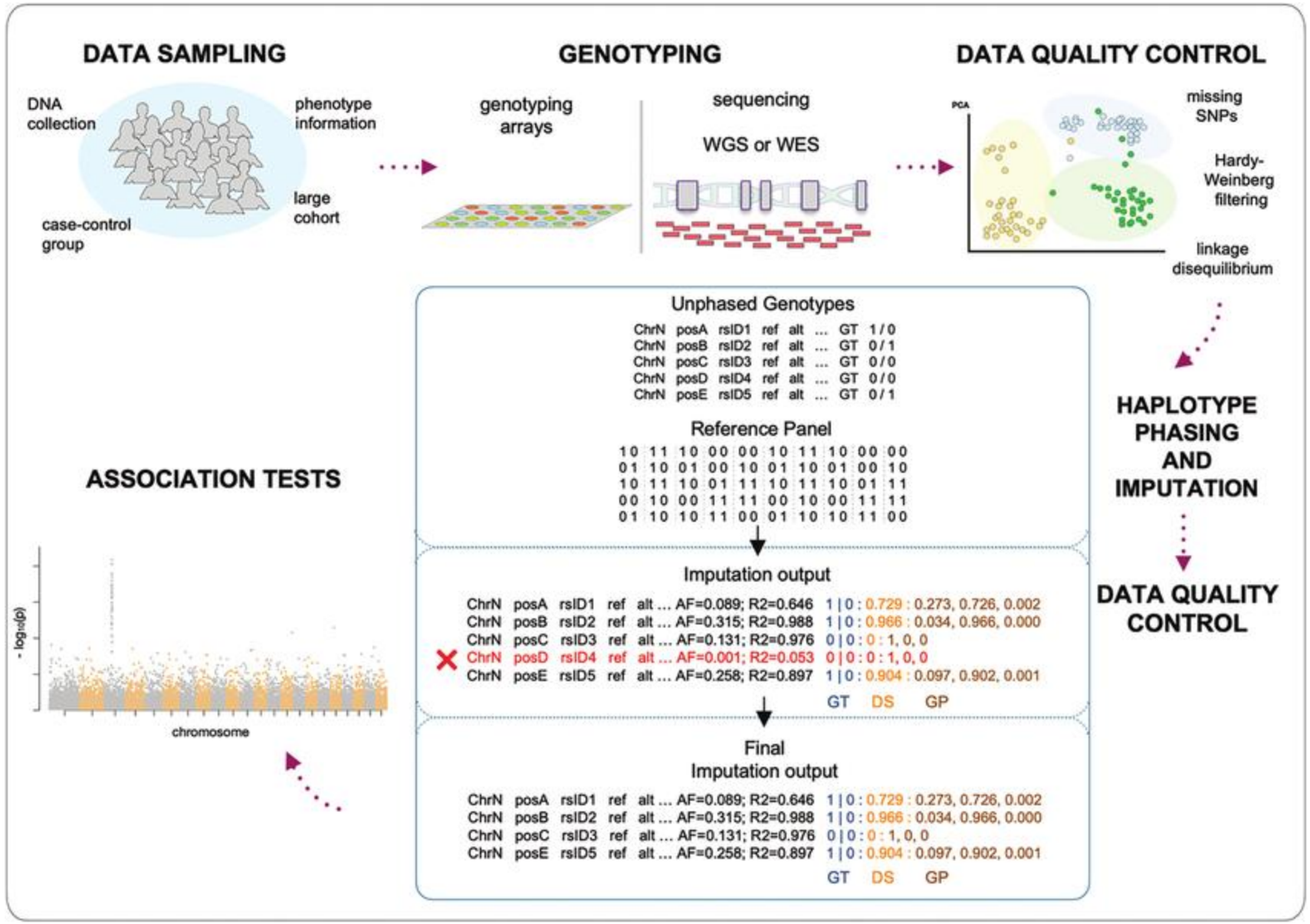


Figure 1. The database structure and newly added functions in QTLbase2.

Tool	URL	Reference	Use	Platform	Description
NHGRI-EBI GWAS Catalog	https://www.ebi.ac.uk/gwas/	https://pubmed.ncbi.nlm.nih.gov/30445434	Database	web	A high-quality curated collection of published GWAS. Currently contains 6,652 publications, 561,096 top associations and 66,522 full summary statistics
dbGAP	https://www.ncbi.nlm.nih.gov/gap/	https://pubmed.ncbi.nlm.nih.gov/17898773/	Database	web	Contains both IPD and summary data from GWAS. The latter are generally available to the public, while access to IPD require varying levels of authorization
GWAScentral	https://www.gwascentral.org	https://pubmed.ncbi.nlm.nih.gov/36350644/	Database	web	Previously known as the Human Genome Variation database of Genotype-to-Phenotype information. Currently contains over 72.5 million P-values for over 5000 studies testing over 7.4 million unique genetic markers investigating over 1700 unique phenotypes.
OpenGWAS	https://gwas.mrcieu.ac.uk	https://www.biorxiv.org/content/10.1101/2020.08.10.244293v1	Database	web	A database of 346,288,362,703 genetic associations from 50,037 GWAS summary datasets
GWAS ATLAS	http://atlas.ctglab.nl	https://pubmed.ncbi.nlm.nih.gov/31427789	Database	web	Currently contains 4,756 GWAS from 473 unique studies across 3,302 unique traits. Each GWAS is accompanied by the results of MAGMA (i.e. gene-based) results, SNP heritability and genetic correlations with other GWAS in the database.
GeneATLAS	http://geneatlas.roslin.ed.ac.uk	https://pubmed.ncbi.nlm.nih.gov/30349118	Database	web	A database of associations using the UK Biobank cohort. Currently contains data for 452,264 Individuals, 778 traits and 30 Million Variants
GWASROCS	https://gwasrocs.ca	https://pubmed.ncbi.nlm.nih.gov/31805043	Database	web	Database containing the largest and most comprehensive set of SNP-derived AUROCs. The database currently houses 579 simulated populations (corresponding to 219 different conditions) and SNP data (odds ratio, risk allele frequency, and p-values) for 2886 unique SNPs
GBE	biobankengine.stanford.edu	https://pubmed.ncbi.nlm.nih.gov/30520965	Database	web	Contains summary statistics from over 750,000 individuals across three population cohorts: UK Biobank, Million Veterans Program and Biobank Japan
GTEx	https://www.gtexportal.org/home/	https://pubmed.ncbi.nlm.nih.gov/32913098/	Database	web	The Genotype-Tissue Expression project contains data of gene expression and splicing in 838 individuals over 49 tissues (see the Perspective by Wilson).
QTLbase2	http://mulinlab.org/qtlbase	https://pubmed.ncbi.nlm.nih.gov/36330927	Database	web	Compiles genome-wide QTL summary statistics for many human molecular traits across over 95 tissue/cell types and multiple biological conditions. Contains tens of millions significant genotype-molecular trait associations under different conditions
COLOCdb	https://ngdc.cncb.ac.cn/colocdb	https://pubmed.ncbi.nlm.nih.gov/37941154	Database	web	The most comprehensive colocalization analysis by integrating publicly available GWASs, different types of xQTL and three different colocalization algorithms. Allows for GWAS-GWAS, GWAS-xQTL, and xQTL-xQTL comparisons
webTWAS	http://www.webtwas.net	https://pubmed.ncbi.nlm.nih.gov/34669946	Database	web	Currently contains data for over 1,389 full GWAS. It calculates the causal genes using single tissue expression imputation (MetaXcan and FUSION) or cross-tissue expression imputation (UTMOST). The users can also upload their own GWAS data
TSEA-DB	https://bioinfo.uth.edu/TSEADB/	https://pubmed.ncbi.nlm.nih.gov/33211888	Database	web	A database for trait-associated tissues. Uses TSEA to infer tissues in which trait-associated genes are enriched.Contains information of 5,019 GWAS summary statistics data sets for human complex traits and diseases (non-UKBB and UKBB)
PCGA	https://pmglab.top/pcga	https://pubmed.ncbi.nlm.nih.gov/35639771	Database	web	A web server to simultaneously estimate associated tissues/cell types and genes of complex diseases and traits. Includes data for 54 human tissues, 2,214 human single cell types and 4,384 mouse single cell types
LD Hub	http://ldsc.broadinstitute.org/	https://pubmed.ncbi.nlm.nih.gov/27663502	Database	web	A centralized database of GWAS results for 173 diseases/traits from publicly available resources and a web interface that automates the LD score regression analysis pipeline.
PheWAS Catalog	https://phewascatalog.org	https://pubmed.ncbi.nlm.nih.gov/24270849/	Database	web	The PheWAS catalog contains the PheWAS results for 3,144 single-nucleotide polymorphisms (SNPs) present in the NHGRI GWAS Catalog
Phenoscanner	http://www.phenoscanner.medschl.cam.ac.uk/	https://pubmed.ncbi.nlm.nih.gov/31233103/	Database	web	A curated database with publicly available results from GWAS used for facilitating “phenome scans”. Currently contains over 65 billion associations and over 150 million unique genetic variants. Comes with a Python command-line tool.

Imputation

- Although some of the methods for quality control mentioned previously can correct errors and alter the data, the methods used for imputation go one step further. In general, however, IPD methods are time-consuming since they process individuals one at a time, and thus methods that impute directly the summary statistics were developed
- These methods utilize only the information provided in the sample regarding the studied population (p-value, z-score or odds-ratio/beta) and require additional information regarding the LD structure. Nearly all methods perform a kind of multiple regression assuming the multivariate normal distribution for the test statistics and utilizing the theoretical result pointing that the correlation of such test statistics equals the correlation of the corresponding variables, that is the genotype correlation, available through the reference panel


DATA SAMPLING
DNA collection
phenotype information
case-control group
large cohort
GENOTYPING
genotyping arrays
sequencing
WGS or WES
DATA QUALITY CONTROL
PCA
missing SNPs
Hardy-Weinberg filtering
linkage disequilibrium
HAPLOTYPE PHASING AND IMPUTATION
DATA QUALITY CONTROL
ASSOCIATION TESTS
- log10(p)
chromosome
Unphased Genotypes
ChrN posA rsID1 ref alt ... GT 1 / 0
ChrN posB rsID2 ref alt ... GT 0 / 1
ChrN posC rsID3 ref alt ... GT 0 / 0
ChrN posD rsID4 ref alt ... GT 0 / 0
ChrN posE rsID5 ref alt ... GT 0 / 1
Reference Panel
10 11 10 00 00 10 11 10 00 00
01 10 01 00 10 01 10 01 00 10
10 11 10 01 11 10 11 10 01 11
00 10 00 11 11 00 10 00 11 11
01 10 10 11 00 01 10 10 11 00
Imputation output
ChrN posA rsID1 ref alt ... AF=0.089; R2=0.646 1 | 0 : 0.729 : 0.273, 0.726, 0.002
ChrN posB rsID2 ref alt ... AF=0.315; R2=0.988 1 | 0 : 0.966 : 0.034, 0.966, 0.000
ChrN posC rsID3 ref alt ... AF=0.131; R2=0.976 0 | 0 : 0 : 1, 0, 0
ChrN posD rsID4 ref alt ... AF=0.001; R2=0.053 0 | 0 : 0 : 1, 0, 0
ChrN posE rsID5 ref alt ... AF=0.258; R2=0.897 1 | 0 : 0.904 : 0.097, 0.902, 0.001
GT DS GP
Final Imputation output
ChrN posA rsID1 ref alt ... AF=0.089; R2=0.646 1 | 0 : 0.729 : 0.273, 0.726, 0.002
ChrN posB rsID2 ref alt ... AF=0.315; R2=0.988 1 | 0 : 0.966 : 0.034, 0.966, 0.000
ChrN posC rsID3 ref alt ... AF=0.131; R2=0.976 0 | 0 : 0 : 1, 0, 0
ChrN posE rsID5 ref alt ... AF=0.258; R2=0.897 1 | 0 : 0.904 : 0.097, 0.902, 0.001
GT DS GP

Typical imputation scenario

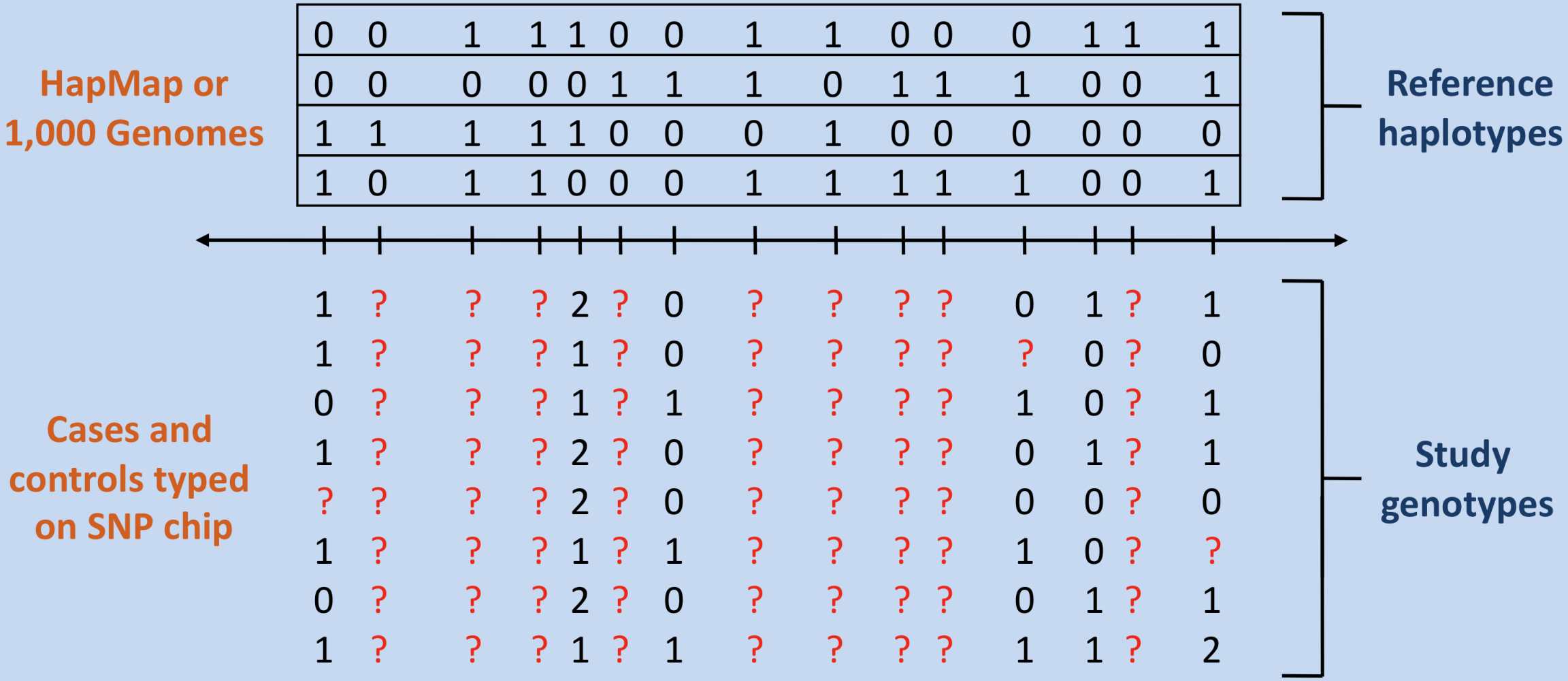


TABLE 1: COMMONLY USED IMPUTATION SOFTWARE PACKAGES

SOFTWARE NAME	INSTITUTION	URL
MACH	University of Michigan[1,2]	http://www.sph.umich.edu/csg/abecasis/MaCH/tour/imputation.html
BEAGLE	University of Auckland[3]	http://www.stat.auckland.ac.nz/~bbrowning/beagle/beagle.html
IMPUTE	Oxford University[4,5]	http://mathgen.stats.ox.ac.uk/impute/impute.html
PLINK	Massachusetts General Hospital / Broad Institute[6]	http://pngu.mgh.harvard.edu/~purcell/plink/

Imputation

- **IPD methods are time-consuming since they process individuals one at a time**, and thus methods that impute directly the summary statistics were developed
- These methods utilize only the information provided in the sample regarding the studied population (p-value, z-score or odds-ratio/beta) and **require additional information regarding the LD structure**.
- Nearly all methods **perform a kind of multiple regression** assuming the multivariate normal distribution for the test statistics and utilizing the theoretical result pointing that the correlation of such test statistics equals the correlation of the corresponding variables, that is the genotype correlation, available through the reference panel

Example with summary data

rsID	Beta	SE
rs001	0.3	0.02
rs002	-0.1	0.2
rs003	0.02	0.05
rs004	-0.01	0. 5
rs005	-	-
rs006	0.15	0.06
rs007	0.11	0. 5
rs008	0.32	0.07
rs009	-0.2	0.03

rsID	Beta	SE
rs001	0.3	0.02
rs002	-0.1	0.2
rs003	0.02	0.05
rs004	-0.01	0. 5
rs005	0.2	0.01
rs006	0.15	0.06
rs007	0.11	0. 5
rs008	0.32	0.07
rs009	-0.2	0.03

Summary statistic imputation using LD reference data

Let

$$Z = \frac{\hat{\beta}_M}{s.e.(\hat{\beta}_M)} = \frac{X^T Y}{\sqrt{(N)}}$$

be a vector of z-scores (estimated effect sizes divided by their standard errors) obtained by marginally testing each SNP one at a time. Under the null hypothesis of no association, $Z \sim N(0, V)$. Let Z_t and Z_i partition the vector Z into T typed SNPs and $M - T$ untyped SNPs, and let $V_{t,t}$ (covariances among typed SNPs), $V_{i,i}$ (covariances among untyped SNPs), and $V_{t,i}$ (covariances among typed and untyped SNPs) partition the matrix accordingly. It follows that $Z_i|Z_t \sim N(V_{i,t} V_{t,t}^{-1} Z_t, V_{i,i} - V_{i,t} V_{t,t}^{-1} V_{i,t}^T)$. The mean and variance of the conditional distribution can be used to impute summary association statistics at untyped SNPs.

$$\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2 \sim MVN\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{y}_2 - \boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

Summary statistic imputation using LD reference data

Let

$$Z = \frac{\hat{\beta}_M}{s.e.(\hat{\beta}_M)} = \frac{X^T Y}{\sqrt{(N)}}$$

be a vector of z-scores (estimated effect sizes divided by their standard errors) obtained by marginally testing each SNP one at a time. Under the null hypothesis of no association, $Z \sim N(0, V)$. Let Z_t and Z_i partition the vector Z into T typed SNPs and $M - T$ untyped SNPs, and let $V_{t,t}$ (covariances among typed SNPs), $V_{i,i}$ (covariances among untyped SNPs), and $V_{t,i}$ (covariances among typed and untyped SNPs) partition the matrix accordingly. It follows that $Z_i | Z_t \sim N(V_{i,t} V_{t,t}^{-1} Z_t, V_{i,i} - V_{i,t} V_{t,t}^{-1} V_{i,t}^T)$. The mean and variance of the conditional distribution can be used to impute summary association statistics at untyped SNPs.

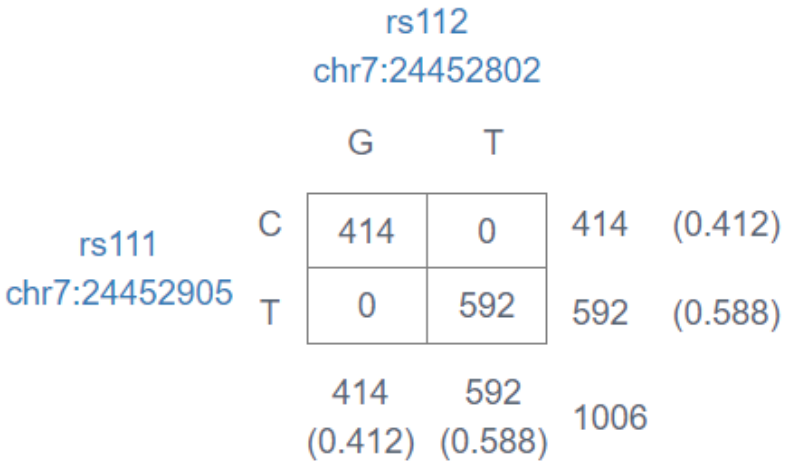
$$\mathbf{Y}_1 \mid \mathbf{Y}_2 = \mathbf{y}_2 \sim MVN\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{y}_2 - \boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

Imputation from summary data

- **FAPI**, **ImpG** , **RAISS**, **DIST** and **SSimp** are similar with most of the differences lying in the choice of the reference panel and the exact details of the mathematical methods used to handle matrix inversions in the multivariate normal.
- **DISSCO** allows for covariates.
- Extensions such as **DISTMIX** and **ARDISS** were developed to handle mixed ethnicity cohorts, improving the imputation performance.
- **Adapt-Mix** estimates the correlation structure in both admixed and non-admixed individuals using simulated and real data and allows the use of this matrix with other imputation methods.

Proposed method

- PRED-LD
  - ✓LD statistics (TOP-LD, HapMap, Pheno Scanner)
  - ✓TOP-MED Whole Genome Sequencing data (greater coverage)
  - ✓LD info for MAF < 1%
  - ✓**In contrary to other approaches, it uses a single point method**



Haplotypes		Statistics	
T_T:	592 (0.588)	D':	1.0
C_G:	414 (0.412)	R^2:	1.0
T_G:	0 (0.0)	Chi-sq:	1006.0
C_T:	0 (0.0)	p-value:	<0.0001

- D : is the linkage coefficient between the typed and the unmeasured SNP
- p_m:the allele frequency of the missing SNP
- p_t the allele frequency of the known SNP
- OR_t is the odds ratio of the typed SNP

Details

$$OR_u = 1 + \frac{D(OR_t - 1)}{p_u[(1 - p_u) + (p_t(1 - p_u) - D)(OR_t - 1)]}$$

$$\beta_u = \log\left(1 + \frac{D\left(e^{\beta_t} - 1\right)}{p_u[(1 - p_u) + (p_t(1 - p_u) - D)(e^{\beta_t} - 1)]}\right)$$

$$\widehat{\mathrm{var}}(f(\hat{\beta}_t)) \approx \left[f'(\hat{\beta}_t)\right]^2 \widehat{\mathrm{var}}(\hat{\beta}_t)$$

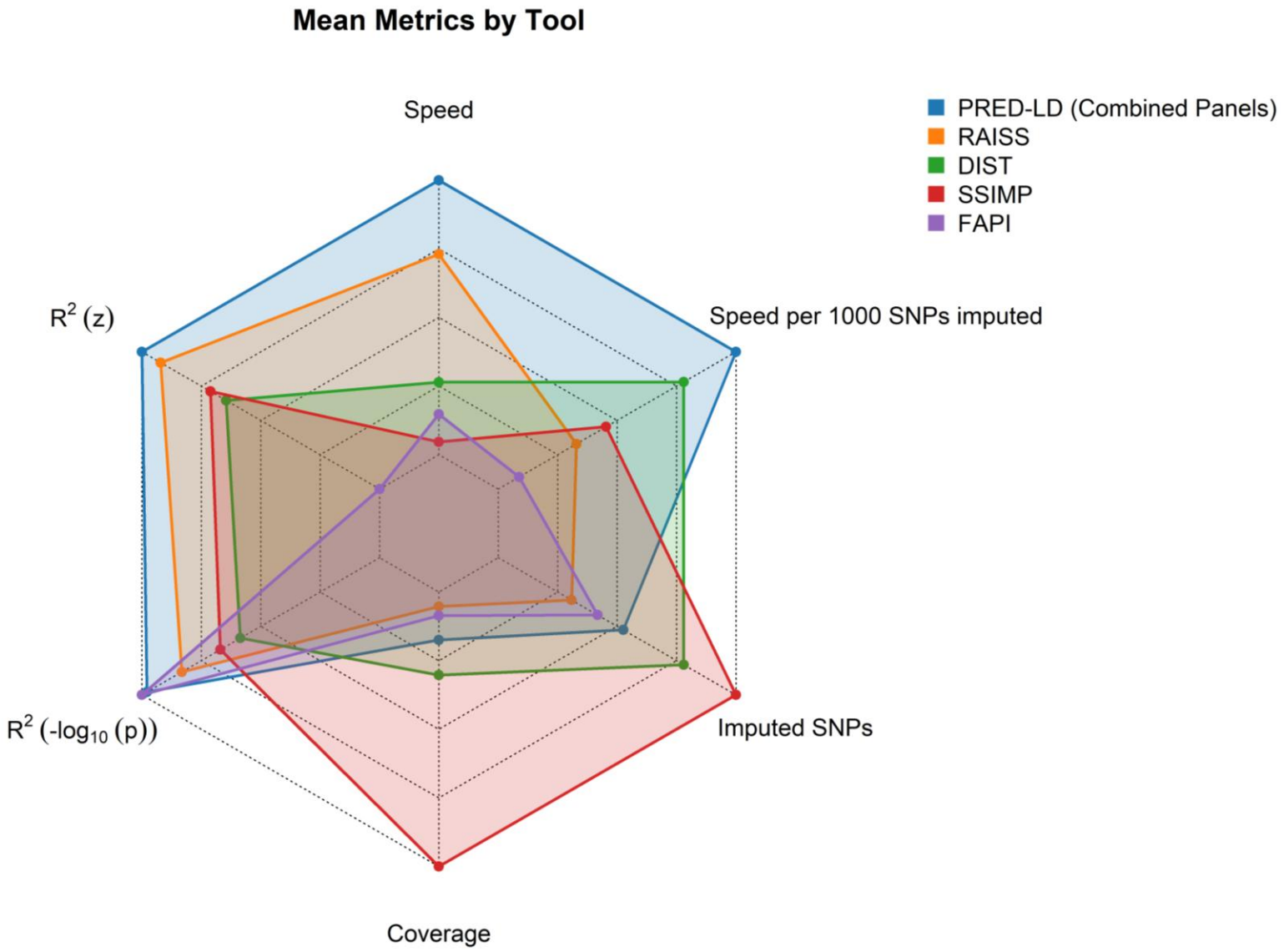
Zondervan KT, Cardon LR. **The complex interplay among factors that influence allelic association**. Nat Rev Genet. 2004;5:89–100.

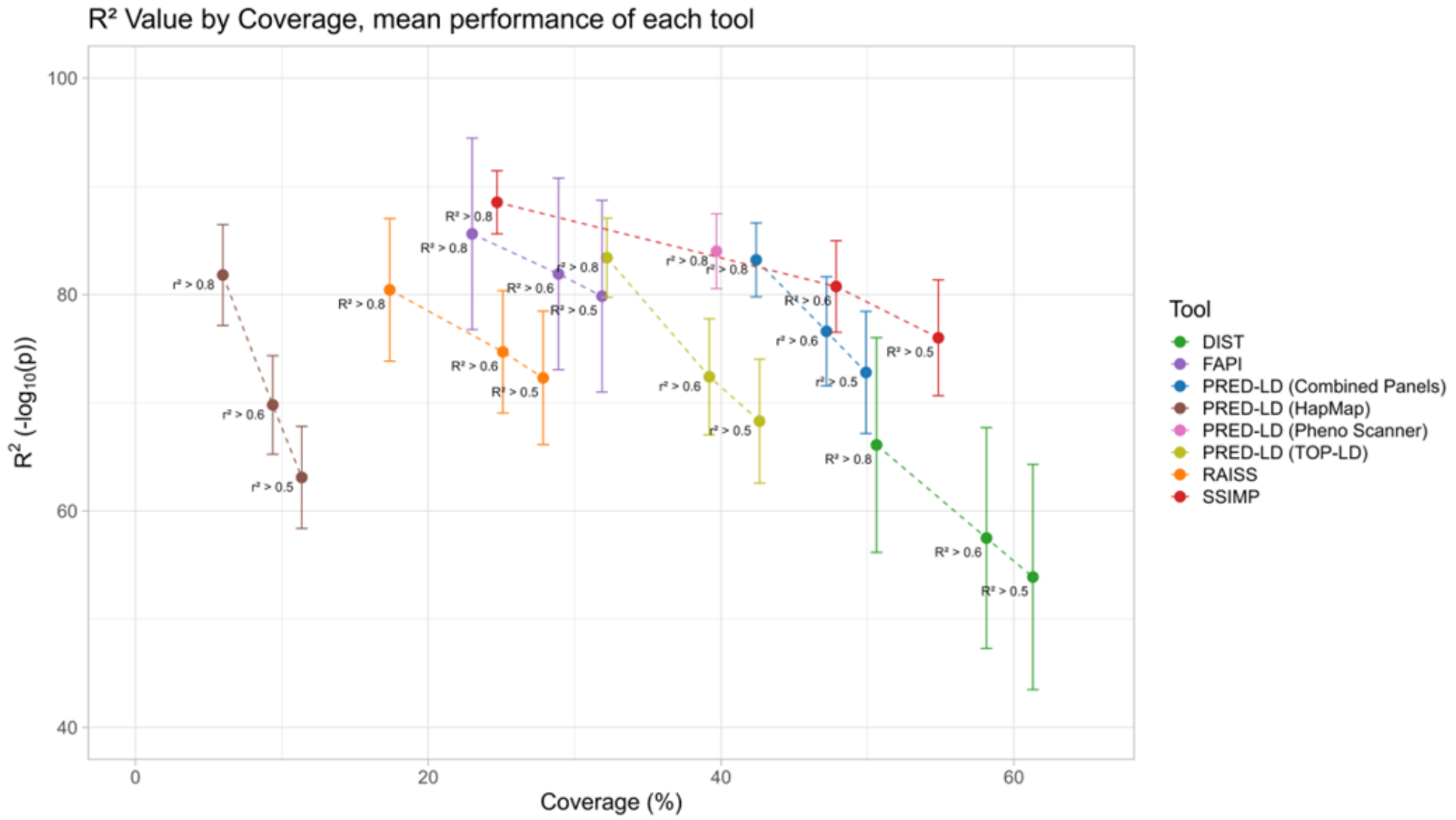
It is not only good in theory, but it actually works

	Tool	Imputed SNPs	R^2(z)	$R^2(-log_{10}(p))$	Imputation percentage in masked SNPs	Time	Time per 1000 SNPs imputed
Mean Performance	PRED-LD	338,803	0.817	0.728	49.90%	20m 12s	8.2s
	RAISS	130,089	0.752	0.619	35.86%	27m 53s	25.0s
	DIST	634,241	0.527	0.436	72.72%	76m 15s	10.5s
	SSIMP	1,944,061	0.581	0.498	93.25%	428m 9s	18.1s
	FAPI	168,365	-	0.744	38.86%	136m 21s	93.5s

Georgios A. Manios, Aikaterini Michailidi, Panagiota I. Kontou and Pantelis G. Bagos**. PRED-LD: Efficient imputation of GWAS summary statistics** . BMC Bioinformatics. 2025

Comparison of methods





Georgios A. Manios, Aikaterini Michailidi, Panagiota I. Kontou and Pantelis G. Bagos**. PRED-LD: Efficient imputation of GWAS summary statistics** . BMC Bioinformatics. 2025

Reviewer #2 intervened

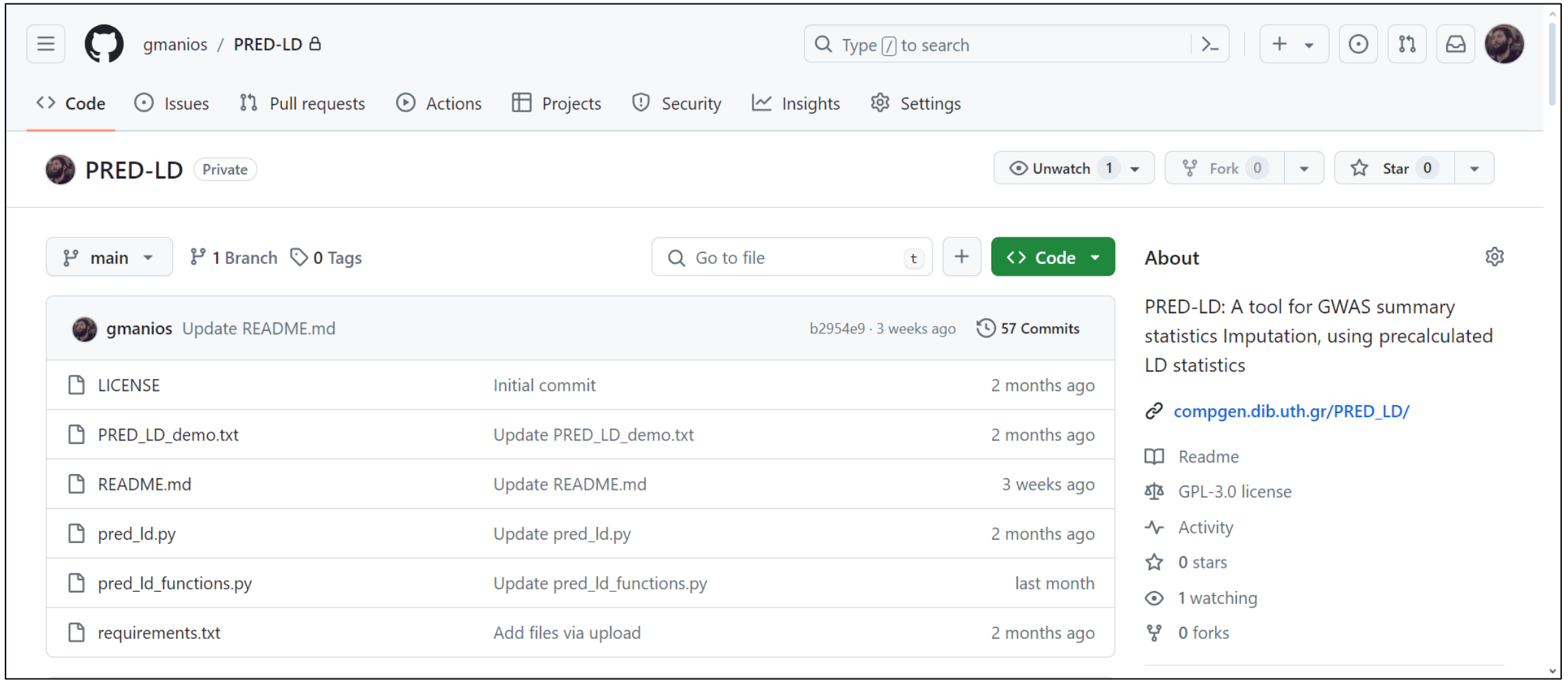
Table 5. Comparison of the mean performance of PRED-LD, DIST and SSIMP when the results of the latter two are filtered using (i) the reported coefficient of determination and keeping only the imputed SNPs with reported R^2>0.5, (ii) retaining the same number of imputed SNPs with PRED-LD (ranked in descending order of R^2 for DIST and SSIMP) and (iii) considering only the common imputed SNPs of PRED-LD across all intersection combinations. The three methods show comparable performance across all metrics except for speed. Execution time in other methods cannot be reduced since the R^2 can only be calculated after the imputation is performed.

Filter	Tool	Imputed SNPs	R^2(z)	$R^2(-log_{10}(p))$	Imputation percentage in masked SNPs	Time
$R^2 > 0.5$ (DIST, SSIMP) and $r^2 > 0.5$ (PRED-LD)	DIST	314,074	0.661	0.539	61.30%	76m 15s
	SSIMP	310,107	0.858	0.760	54.84%	428m 9s
	PRED-LD	338,803	0.817	0.728	49.90%	20m 12s
Same number of imputed SNPs of PRED-LD	DIST	338,803	0.765	0.628	52.69%	76m 15s
	SSIMP		0.879	0.808	53.86%	428m 9s
	PRED-LD		0.817	0.728	49.90%	20m 12s
Common SNPs (DIST-SSIMP-PRED-LD)	DIST	158,650	0.785	0.666	46.13%	76m 15s
	SSIMP		0.837	0.737		428m 9s
	PRED-LD		0.816	0.725		20m 12s
Common SNPs (DIST-PRED-LD)	DIST	162,354	0.787	0.666	46.47%	76m 15s
	PRED-LD		0.816	0.725		20m 12s
Common SNPs (SSIMP-PRED-LD)	SSIMP	219,732	0.837	0.739	49.39%	428m 9s
	PRED-LD		0.817	0.727		20m 12s

Advantages

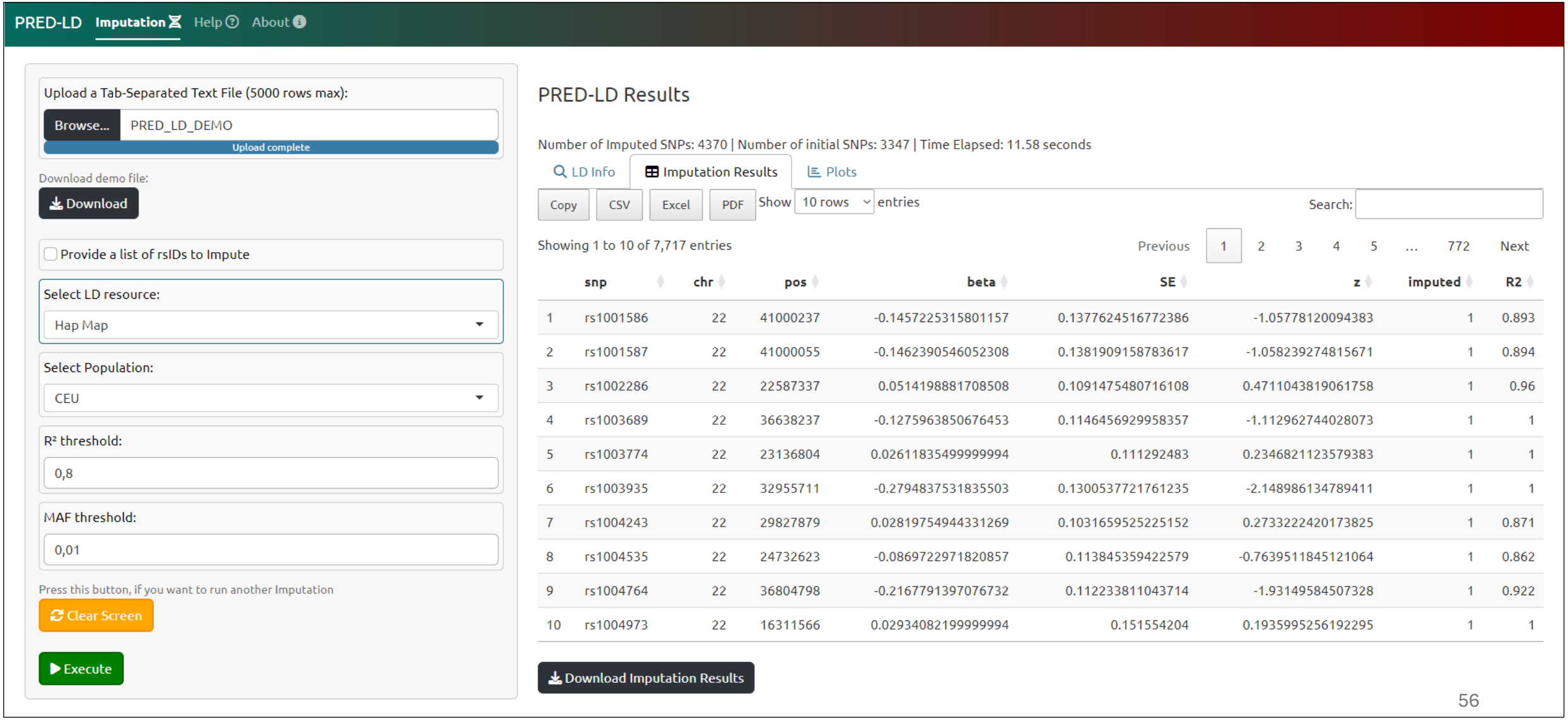
- This approach allows both for the definition of a strict r2 LD threshold or a lower one, resulting **either in a more accurate imputation or a broader coverage**, respectively
- The method **can use and combine different reference panels** with ease in a wide range of populations, since for each imputation only information from one SNP is used
- PRED-LD imputes beta and SE instead of Z's or P's
- The methods using the multivariate normal distribution need additional computations in order to regularize the variance-covariance matrix, or to avoid multicollinearity. Thus, it seems that **the single marker approach with the direct imputation is preferable**, especially when the SNPs in the GWAS and the panel are dense

GitHub Repository



Georgios A. Manios, Aikaterini Michailidi, Panagiota I. Kontou and Pantelis G. Bagos**. PRED-LD: Efficient imputation of GWAS summary statistics** . BMC Bioinformatics. 2025

PRED-LD (Web Interface)



	snp	chr	pos	beta	SE	z	imputed	R2
1	rs1001586	22	41000237	-0.1457225315801157	0.1377624516772386	-1.05778120094383	1	0.893
2	rs1001587	22	41000055	-0.1462390546052308	0.1381909158783617	-1.058239274815671	1	0.894
3	rs1002286	22	22587337	0.0514198881708508	0.1091475480716108	0.4711043819061758	1	0.96
4	rs1003689	22	36638237	-0.1275963850676453	0.1146456929958357	-1.112962744028073	1	1
5	rs1003774	22	23136804	0.02611835499999994	0.111292483	0.2346821123579383	1	1
6	rs1003935	22	32955711	-0.2794837531835503	0.1300537721761235	-2.148986134789411	1	1
7	rs1004243	22	29827879	0.02819754944331269	0.1031659525225152	0.2733222420173825	1	0.871
8	rs1004535	22	24732623	-0.0869722971820857	0.113845359422579	-0.7639511845121064	1	0.862
9	rs1004764	22	36804798	-0.2167791397076732	0.112233811043714	-1.93149584507328	1	0.922
10	rs1004973	22	16311566	0.02934082199999994	0.151554204	0.1935995256192295	1	1

Imputation

- **FAPI**, **ImpG** , **RAISS**, **DIST** and **SSimp** are similar with most of the differences lying in the choice of the reference panel and the exact details of the mathematical methods used to handle matrix inversions in the multivariate normal.
- **DISSCO** uses a similar framework but allows for covariates.
- Such methods may perform poorly in cases where the sample has a different LD structure compared to the reference panel. Thus, extensions such as **DISTMIX** and **ARDISS** were developed to handle mixed ethnicity cohorts, improving the imputation performance.
- **Adapt-Mix** estimates the correlation structure in both admixed and non-admixed individuals using simulated and real data and allows the use of this matrix with other imputation methods.
- Other methods such **LS-meta** and **LSimputing** offer additional advantages; LS-meta imputes both genetic and environmental components using information from additional omics-trait association summary data, whereas LSimputing implements a non-parametric method that allows for nonlinear SNP-trait associations and predictions in case a sample of IPD is available.
- Using the same principles, **simGWAS** allows simulation of whole GWAS summary data, without generating individual data as an intermediate step.

Genotype reconstruction

- Genotype reconstruction methods take a different approach. Given the summary statistics for a SNP (either directly measured or imputed), one can reconstruct the genotype counts that produced it. This will offer many advantages, since with the reconstructed genotypes the researchers could perform additional analyses using other statistical methods suitable for grouped data and test different hypotheses
- The details and the success of the reconstruction depend heavily on available summary statistics. As one can easily understand, p-values and z-scores cannot be used, and one must rely on available effect sizes such as the odds ratio (OR). When the OR, the standard error and the sample size is given, methods are available in epidemiology that allow the reconstruction of the allelic 2X2 table
- **React** uses an equivalent method relying on solving a system of nonlinear equations. If the allele frequency in one group (usually the controls) is also known, the allelic counts may easily be obtained with a simple calculation. In all cases the accuracy of the reconstruction may depend on the precision of the available summary statistics. After the allelic 2X2 table is reconstructed, it is straightforward to obtain the genotype counts, assuming HWE (which as one might expect adds another source of potential bias).
- **MetaSustract** is a tool that recreates analytically the results of the validation cohort from meta-analysis summary statistics, allowing the researchers to compute meta-analysis summary statistics that are independent of the validation cohort, without requiring access to the IPD.
- **Spkmt** works in similar fashion but in families; it can be used to derive the summary statistics of one parent from the data of the offspring and the other parent. Finally, we need to mention two tools that work in somewhat different modes.
- **OATH** is used to reproduce reported results from a GWAS and recover underreported results from other alternative models with a different combination of nuisance parameters, whereas **LMOR** performs transformations from the genetic effects estimated under the Linear Mixed Model to the Odds Ratio that only rely on summary statistics

QCGWAS	https://cran.r-project.org/web/packages/QCGWAS/index.html	https://pubmed.ncbi.nlm.nih.gov/24395754/	Quality Control	R	Automates the quality control of GWAS result files. Its main purpose is to facilitate the quality control of a large number of such files before meta-analysis.
DENTIST	https://zenodo.org/records/5516202	https://pubmed.ncbi.nlm.nih.gov/34880243	Quality Control	C/C++	Leverages LD among SNPs to detect and eliminate errors in GWAS or LD reference and heterogeneity between the two
GWAS-SSF	https://github.com/EBISPOT/gwas-summary-statistics-standard	https://www.biorxiv.org/content/10.1101/2022.07.15.500230v2.full	Quality Control	Python	Specifications for the first version of the GWAS-SSF format, which was developed to meet the requirements discussed with the community. GWAS-SSF consists of a tab-separated data file with well-defined fields and an accompanying metadata file
MungeSumstats	https://neurogenomics.github.io/MungeSumstats	https://pubmed.ncbi.nlm.nih.gov/34601555	Quality Control	R	A tool for the standardization and quality control of GWAS summary statistics. It can handle the most common summary statistic formats
GWASinspector	http://gwasinspector.com	https://pubmed.ncbi.nlm.nih.gov/33416854/	Quality Control	R	Developed to facilitate and streamline this process and provide the user with a comprehensive report. It will also generate cleaned, harmonized GWAS files ready for meta-analysis
VCF	https://github.com/MRCIEU/gwas-vcf-specification/releases/tag/1.0.0, https://github.com/mrcieu/gwas2vcf	https://pubmed.ncbi.nlm.nih.gov/33441155	Quality Control	Python	The variant call format is used to store GWAS summary statistics along with open-source tools to be uses in downstream analyses.
SumStatsRehab	https://github.com/Kukuster/SumStatsRehab	https://pubmed.ncbi.nlm.nih.gov/36284273	Quality Control	Python	A tool for data validation, restoration of missing data, correction and formatting
GQS	https://github.com/Xswapnil/GQS/	https://pubmed.ncbi.nlm.nih.gov/36651666	Quality Control	Python	Identifies suspicious regions and prevents erroneous interpretations. Aassesses all measured SNPs that are in LD and compares the significance of trait association of each SNP to its LD value for the reported index SNP
GWAtoolbox	https://github.com/cran/GWAtoolbox	https://pubmed.ncbi.nlm.nih.gov/22155946/	Quality Control	R	Contains three particular data quality aspects: data formatting, quality of the GWAS results and data consistency across studies
EasyQC	https://www.uni-regensburg.de/medizin/epidemiologie-praeventivmedizin/genetische-epidemiologie/software/index.html	https://pubmed.ncbi.nlm.nih.gov/24762786/	Quality Control	R	A general protocol for conducting meta-analysis and carrying out QC to minimize errors and to guarantee maximum use of the data
GEAR	https://github.com/syntheke/GEAR	https://pubmed.ncbi.nlm.nih.gov/27552965	Quality Control	Java	A tool that contains functions to identify significant sample overlap or heterogeneity between pairs of cohorts
EXTminus23andMe	https://github.com/Camzcamz/EXTminus23andMe	https://pubmed.ncbi.nlm.nih.gov/37713023	Quality Control	R	A tool to evaluate the quality of summary statistics after data removal and the suitability of these downsampled summary statistics for typical follow-up genetic analyses
GWASlab	https://github.com/Cloufield/gwaslab	https://jxiv.jst.go.jp/index.php/jxiv/preprint/view/370	Quality Control	Python	A toolkit for handling GWAS summary statistics. Offers functionalities for converting most formats, standardization, normalization, harmonization, filtering and visualization.

OATH	https://github.com/gc5k/GEAR	https://pubmed.ncbi.nlm.nih.gov/28122950	Reconstruction	Java	Reproduces reported results from a GWAS and recovers underreported results from other alternative models with a different combination of nuisance parameters
Metasubtract	https://cran.r-project.org/web/packages/MetaSubtract	https://pubmed.ncbi.nlm.nih.gov/32696040	Reconstruction	R	Subtracts the results of the validation cohort from meta-GWAS summary statistics analytically
LMOR	https://github.com/lukelloydjones/ORShiny	https://pubmed.ncbi.nlm.nih.gov/29429966	Reconstruction	R	Performs transformations from the genetic effects estimated under the Linear Mixed Model to the Odds Ratio that only rely on summary statistics
ReACt	https://github.com/Paschou-Lab/ReACt	https://pubmed.ncbi.nlm.nih.gov/35581276	Reconstruction	C/C++	Performs genotype reconstruction for case-control GWAS summary statistics. It includes three modules: Meta-analysis, group PRS and case-case GWAS.
simGWAS	http://github.com/chr1swallace/simGWAS	https://pubmed.ncbi.nlm.nih.gov/30371734	Reconstruction	R	Simulates GWAS summary data without individual data as an intermediate step
spkmt	https://osf.io/spkmt/	https://pubmed.ncbi.nlm.nih.gov/37518004	Reconstruction	R	Method to derive GWAS summary statistics for one parent when observations have only been made on the offspring and another parent
FAPI	https://pmglab.top/fapi/	https://pubmed.ncbi.nlm.nih.gov/26306642	Imputation	Executable	Fast and accurate P-value imputation method that utilizes summary statistics of common variants. Its computational cost is linear with the number of untyped variants
impG	https://bogdan.dgsom.ucla.edu/pages/impg/	https://pubmed.ncbi.nlm.nih.gov/24990607	Imputation	C/C++	Uses the multivariate normal distribution and LD from external source
SSimp	https://github.com/zkutalik/ssimp_software	https://pubmed.ncbi.nlm.nih.gov/29782485/	Imputation	C/C++	Uses the multivariate normal distribution and LD from external source
RAISS	https://gitlab.pasteur.fr/statistical-genetics/raiss	https://pubmed.ncbi.nlm.nih.gov/31173064	Imputation	Python	Uses LD and the multivariate normal distriburion along with several optimizations
LS-META	https://github.com/ren328/LS-Meta	https://pubmed.ncbi.nlm.nih.gov/37369060	Imputation	R	Imputes both genetic and environmental components of a trait using both SNP-trait and omics-trait association summary data
DIST/DISTMIX	https://dleelab.github.io/distmix/	https://pubmed.ncbi.nlm.nih.gov/26059716, https://pubmed.ncbi.nlm.nih.gov/23990413	Imputation	Executable	Uses the multivariate normal distribution and the correlation structure from a relevant reference population. DISTMIX is the extension to mixed ethnicity cohorts
ARDISS	https://github.com/BorgwardtLab/ARDISS	https://pubmed.ncbi.nlm.nih.gov/30423082	Imputation	Python	Imputes missing summary statistics in mixed-ethnicity cohorts through Gaussian Process Regression and automatic relevance determination
LSimputing	https://github.com/ren328/LSimputing	https://pubmed.ncbi.nlm.nih.gov/37181332	Imputation	R	A nonparametric method for large-scale imputation of the genotype effects. If a sample of IPD is available the method allows for nonlinear SNP-trait associations and predictions
DISSCO	https://yunliweb.its.unc.edu/DISSCO/	https://pubmed.ncbi.nlm.nih.gov/25810429	Imputation	Java	Uses the multivariate normal distribution and LD, and allows for covariates
Adapt-Mix	https://github.com/dpark27/adapt_mix	https://pubmed.ncbi.nlm.nih.gov/26072481	Imputation	Python	Combines information across all available reference panels to produce estimates of local genetic correlation structure for summary statistics-based methods in arbitrary populations

Single trait analysis

- Meta-analysis
- Heritability analysis
- Gene-based tests
- GSA
- Fine-mapping

Meta-analysis

- meta-analysis can be performed with various methods using IPD or summary data; the former offers many advantages, but the latter is far more easy to be performed taking into account the various restrictions imposed on sharing GWAS IPD and the difficulties in the logistics of such a project
- A compromise between these two extremes arises when a research group has access to individual-level genotype data of a limited sample size and wants to integrate these with existing summary data available in the databases. Such methods are in use in epidemiology for years and several tools have been developed especially for handling GWAS data, for instance **IGESS**, **metaGIM** and **LEP**. **PolyGIM** can be applied with or without IPD and uses polytomous logistic regression to investigate disease subtype heterogeneity in situations when only summary data is available

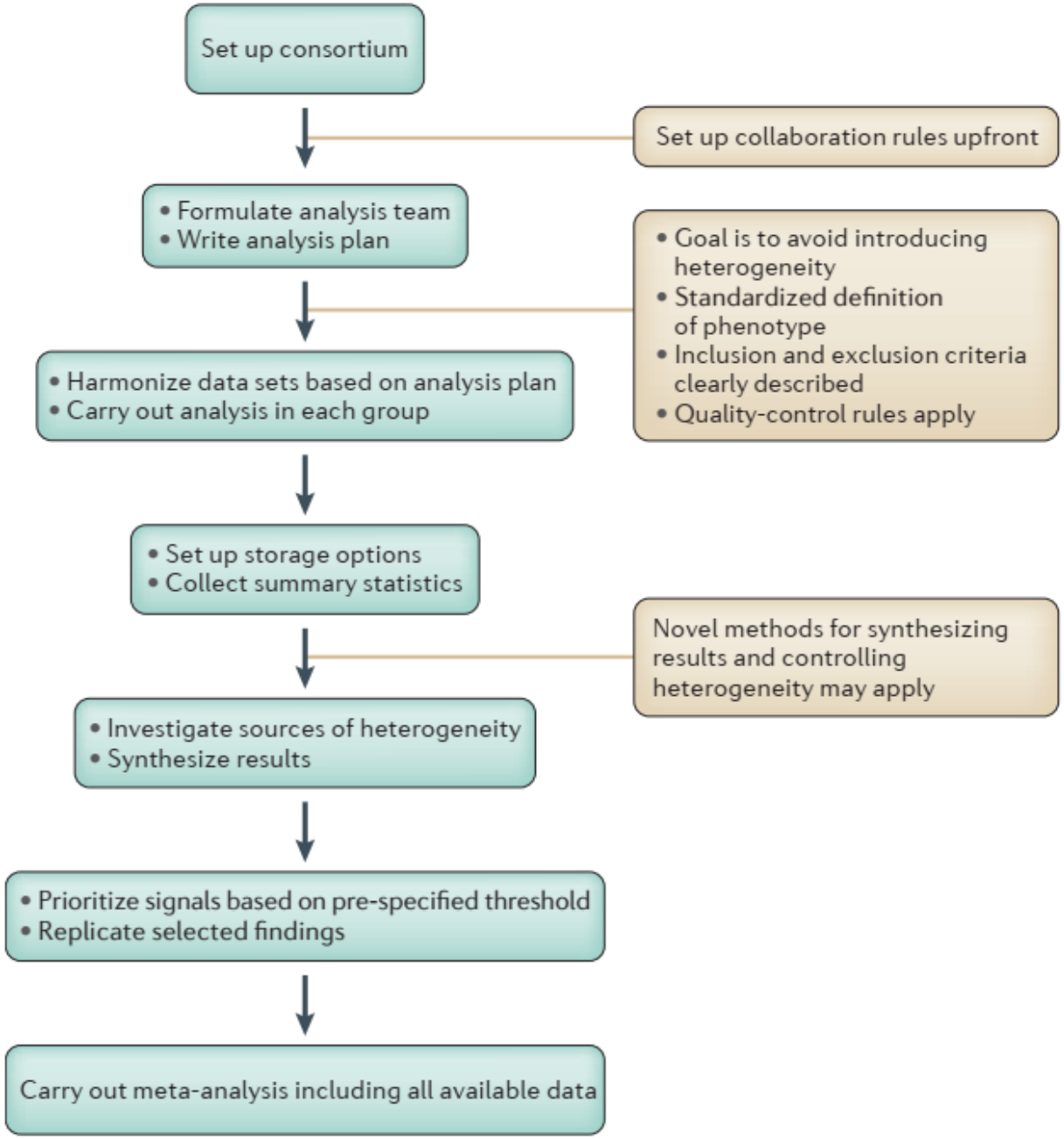


Figure 1 | **Stages in a meta-analysis.** A typical plan for a meta-analysis of genome-wide and next-generation sequence data.

Table 3 | **Summary of methods for meta-analysis of genome-wide data**

Method	Description	Advantages	Disadvantages	Main software used
P value meta-analysis	Simplest meta-analytical approach	Allows meta-analysis when effects are not available	Direction of effect is not always available; inability to provide effect sizes; difficulties in interpretation	METAL, GWAMA, R packages
Fixed effects	Synthesis of effect sizes. Between-study variance is assumed to be zero	Effects readily available through specialized software	Results may be biased if a large amount of heterogeneity exists	METAL, GWAMA, R packages
Random effects	Synthesis of effect sizes. Assumes that the individual studies estimate different effects	Generalizability of results	Power deserts in discovery efforts; may yield spuriously large summary effect estimates when there are selection biases	GWAMA, R packages
Bayesian approach	Incorporates prior assessment of the genetic effects	Most direct method for interpretation of results as posterior probabilities given the observed data	Methodologically challenging; GWAS-tailored routine software not available; subjective prior information used	R packages
Multivariate approaches	Incorporates the possible correlation between outcomes or genetic variants	Increased power can identify variants that conventional meta-analysis do not reveal using the same data sets	Computationally intensive; software not available for all analyses; some may require individual-level data	GCTA for multi-locus approaches
Other extensions	A set of different approaches that allows for the identification of multiple variants across different diseases	Summary results of previous meta-analyses can be used	May need additional exploratory analyses for the identification of variants; prone to systematic biases	Software developed by the authors of the proposed methodologies

GCTA, genome-wide complex trait analysis; GWAS, genome-wide association study.

Box 3 | **Statistical properties of common GWAS meta-analysis approaches**

The simplest genome-wide association study (GWAS) meta-analysis approach is to combine *P* values using Fisher's method. The formula for the statistic is

$$X^2 = -2\sum_{i=1}^{k}\log(P_i)$$

where P_i is the *P* value for the i^{th} study, and *k* is the number of studies in the meta-analysis. Under the null hypothesis, X^2 follows a χ^2 distribution with $2k$ degrees of freedom. The *Z* scores meta-analysis can be implemented using the equation

$$Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$$

where w_i is the square root of sample size of the i^{th} study and

$$Z_i = \Phi^{-1}\left(1-\frac{P_i}{2}\right)^{*}(\text{effect direction for study } i)$$

where Φ is the standard normal cumulative distribution function. For fixed effects models, inverse variance weighting is widely used. The weighted average of the effect sizes can be calculated as

$$\hat{\theta}_F = \frac{\sum_i w_i \hat{\theta}_i}{\sum_i w_i}$$

and the variance is

$$\mathrm{var}(\hat{\theta}_F) = \frac{1}{\sum_i w_i}$$

where $\hat{\theta}_i$ is the i^{th} study normalized effect (for example, logarithm of odds ratio or β-coefficient for a logistic regression for a binary phenotype or mean difference or standardized mean difference for a continuous phenotype), and w_i is the reciprocal of the estimated variance of the effect study. The random effects model incorporates the between-study variance of heterogeneity, and therefore the weight for the random effects model is calculated as

$$w_i^R = \frac{1}{\left(\frac{1}{w_i} + \hat{\tau}^2\right)}$$

where

$$\tau^2 = \frac{(Q-(k-1))}{\left(\sum_i w_i - \left(\frac{\sum_i w_i^2}{\sum_i w_i}\right)\right)}$$

and *Q* is Cochran's *Q* statistic, which is given by

$$Q = \sum_i w_i(\hat{\theta}_i - \hat{\theta}_F).$$

Another popular heterogeneity metric, I^2, is given by

$$I^2 = \frac{100^{*}(Q-(k-1))}{Q}.$$

The multivariate meta-analysis approaches are based on the calculation of a variance–covariance matrix for the correlated phenotypes or the single-nucleotide polymorphisms in linkage disequilibrium that will allow the calculation of the marginal effects. In cross-phenotype meta-analysis, the developed statistic measures the likelihood of the null hypothesis, given the data. The test is asymptotically distributed as

$$\chi^2_{df=1}.$$

Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379-89.

Meta-analysis

- Combining the estimates of several studies
- The methodology dates back to Fisher
- The term appeared for the first time in Psychology (Glass, 1976)
- In its simpler form, it is a weighted average of the estimates
- Improves the statistical power to detect weak effects
- Tools like PLINK, GWAMA facilitate the analysis of GWAS

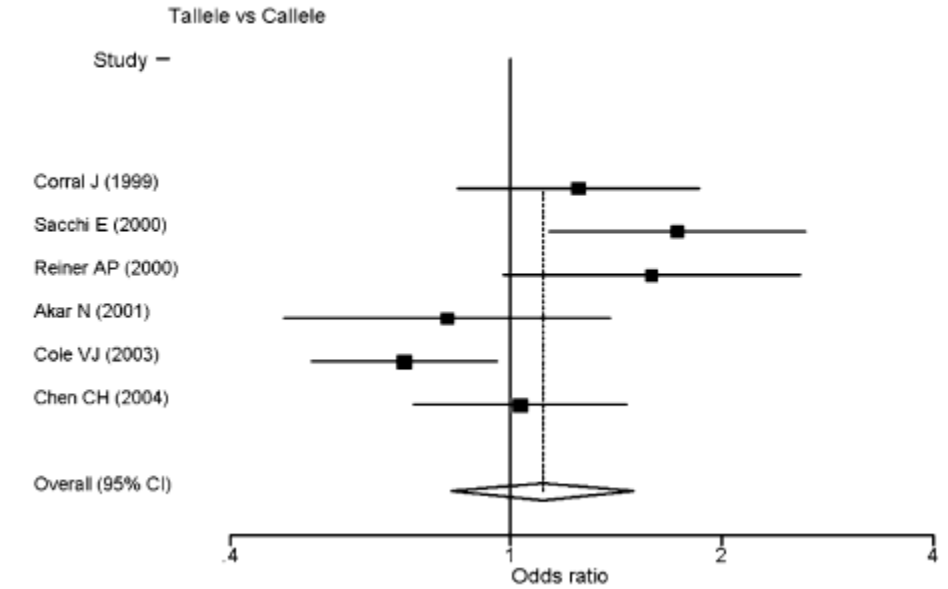
$$Y_i = \log OR_i = \log\left(\frac{\alpha\delta}{\beta\gamma}\right), \quad s_i = \sqrt{\frac{1}{\alpha}+\frac{1}{\beta}+\frac{1}{\gamma}+\frac{1}{\delta}}$$

$$\hat{\theta}_{\text{MLE}} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \quad \text{with } W_i = \frac{1}{s_i^2}$$

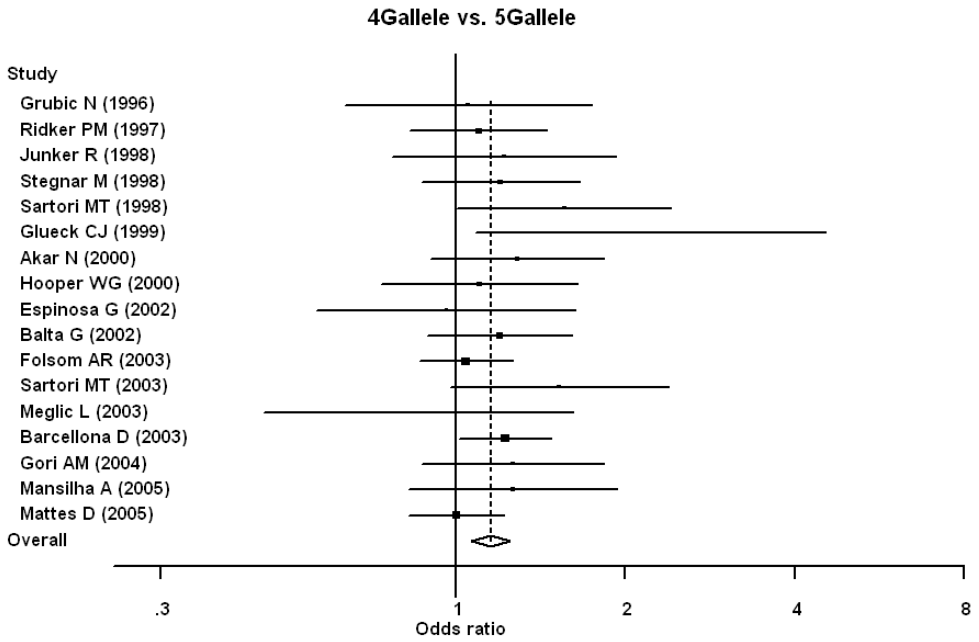
	Allele B	Allele A
Cases	α	β
Controls	γ	δ

Glass GV. **Primary, secondary, and meta-analysis of research**. Educational Researcher, 1976; 5: 3-8

Nikolopoulos GK, Tsantes AE, Bagos PG, Travlou A, Vaiopoulos G**. Integrin, alpha 2 gene C807T Polymorphism and Risk of Ischemic Stroke: a Meta-Analysis.** 2007, *Thrombosis Research*; 119 **(4)**: 501-510



Tsantes AE, Nikolopoulos GK, Bagos PG, Rapti E, Mantzios G, Kapsimali V, Travlou A**. Association between the Plasminogen Activator Inhibitor-1 4G/5G Polymorphism and Venous Thrombosis: a Meta-Analysis.** 2007, *Thrombosis and Haemostasis*



PLOS COMPUTATIONAL BIOLOGY

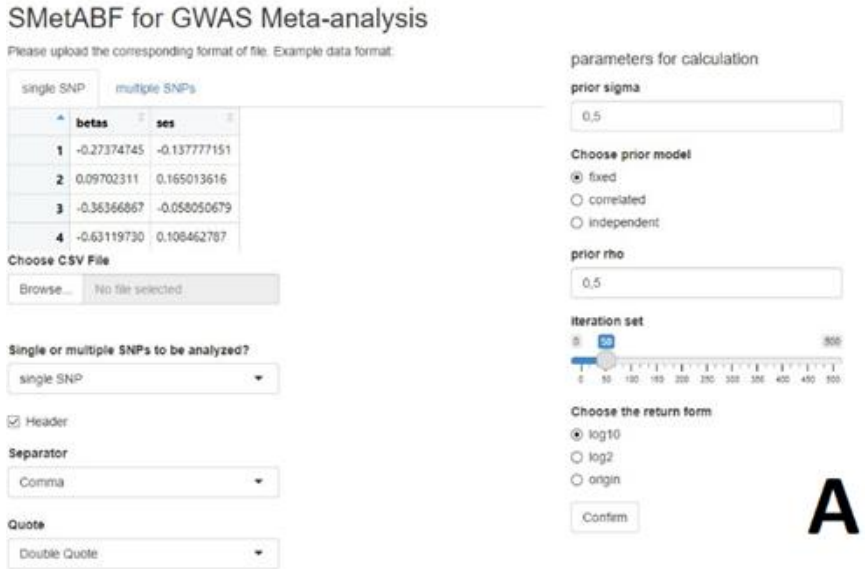
EDITORIAL

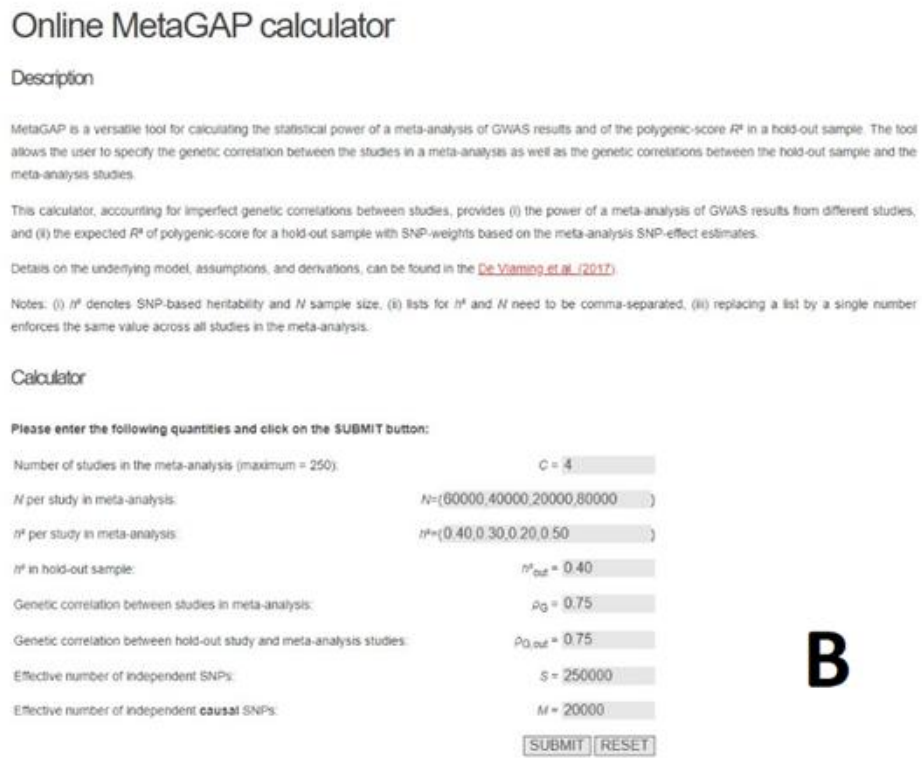
Ten simple rules for carrying out and writing meta-analyses

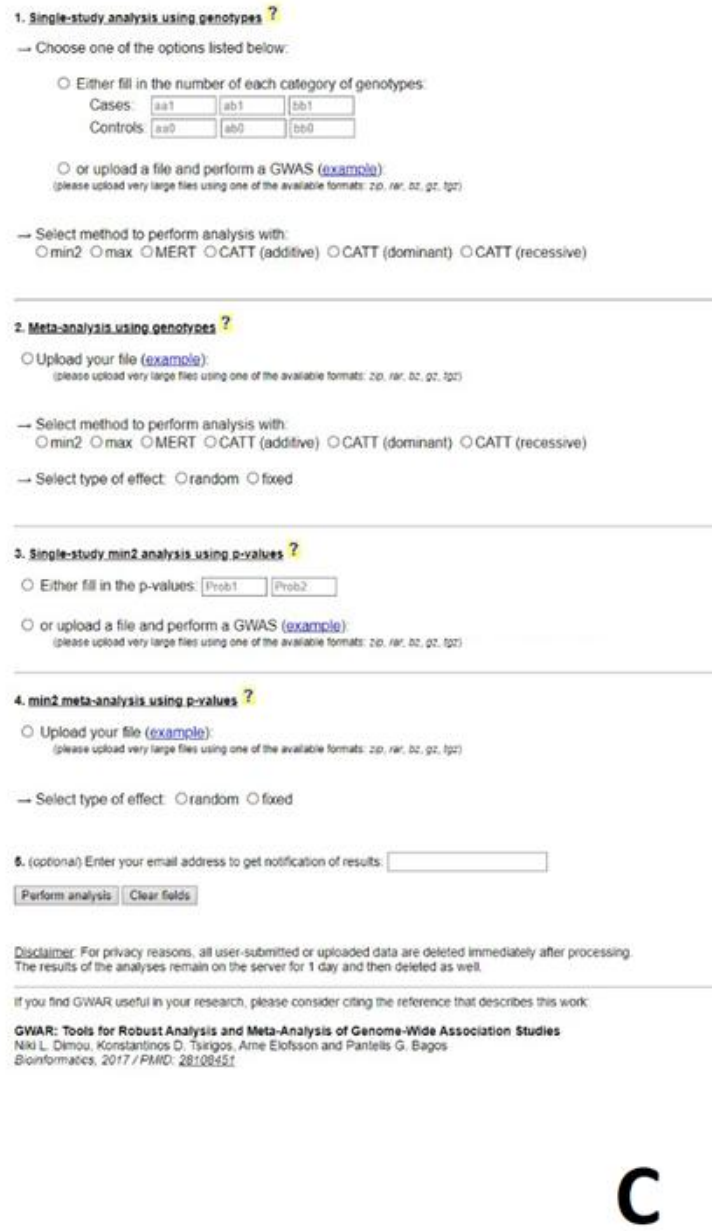
Diego A. Forero[1,2]*, **Sandra Lopez-Leon**[3], **Yeimy González-Giraldo**[4], **Pantelis G. Bagos**[5]

1 Laboratory of NeuroPsychiatric Genetics, Biomedical Sciences Research Group, School of Medicine, Universidad Antonio Nariño, Bogotá, Colombia, **2** PhD Program in Health Sciences, School of Medicine, Universidad Antonio Nariño, Bogotá, Colombia, **3** Novartis Pharmaceuticals Corporation, East Hanover, New Jersey, United States of America, **4** Departamento de Nutrición y Bioquímica, Facultad de Ciencias, Pontificia Universidad Javeriana, Bogotá., Colombia, **5** Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece

- Regarding summary-data meta-analysis of GWAS, the most commonly used methods includes standard methods, such as combining p-values, z-statistics or effects sizes like Odds Ratio (for binary traits) or mean differences (for continuous traits) using fixed or random effects models. These statistical methods are straightforward to implement, and are available in general purpose statistical packages such as STATA and R.
- However, there are several specialized tools that facilitate the process and provide integration with useful bioinformatics or visualization functions. Such widely used tools include **METAL**, **GWAMA** and **PLINK**.
- **YAMAS** performs meta-analysis including missing SNPs identified with LD without performing imputation and **rareMETALS** uses a partial correlation based score to perform meta-analysis in the presence of large amounts of missing values.
- There is also a class of tools which focus on the replication of GWAS and the combined analysis of data from primary and replication studies. Such tools include **rfdr** and **Jlfdr** which control for False Discovery Rate (FDR), **Rrate**, which determines the sample size of the replication study and checks the consistency between the primary and the replication study, and **MAJAR** which jointly test prognostic and predictive effects in meta-analysis without the need of using an independent cohort.
- **metaGAP** is an online tool for calculating the statistical power of a meta-analysis
- **METACARPA** works with overlapping or related samples, even when details of the overlap or relatedness are unknown,
- **MAGENTA** performs meta-analysis with gene set enrichment analysis (GSEA),
- **GWASmeta** and **MetABF** work in a bayesian framework calculating the Approximate Bayes Factor (ABF).
- Other tools offer more advanced options such as meta-analysis with multiple traits, like **nGWAMA**, **metaCCA**, **CPASSOC**, **metaUSAT** and **CPBayes** (and its extension **GCPBayes**), and others are designed for meta-analysis under different genetic models, like **GWAR** which uses robust methods in order to handle the uncertainty in the underlying genetic model, Finally, we need to mention **sPLINK** which performs privacy-aware GWAS on distributed datasets, and **XPEB** which is an empirical Bayes approach designed to improve the power GWAS in minority populations by exploiting information from GWASs performed in populations of different origin.


SMetABF for GWAS Meta-analysis
Please upload the corresponding format of file. Example data format:
single SNP
multiple SNPs
betas ses
1 -0.27374745 -0.137777151
2 0.09702311 0.165013616
3 -0.36366867 -0.058050679
4 -0.63119730 0.108462787
Choose CSV File
Browse...
No file selected
Single or multiple SNPs to be analyzed?
single SNP
Header
Separator
Comma
Quote
Double Quote
parameters for calculation
prior sigma
0.5
Choose prior model
fixed
correlated
independent
prior rho
0.5
iteration set
Choose the return form
log10
log2
origin
Confirm
A


Online MetaGAP calculator
Description
MetaGAP is a versatile tool for calculating the statistical power of a meta-analysis of GWAS results and of the polygenic-score R^2 in a hold-out sample. The tool allows the user to specify the genetic correlation between the studies in a meta-analysis as well as the genetic correlations between the hold-out sample and the meta-analysis studies.
This calculator, accounting for imperfect genetic correlations between studies, provides (i) the power of a meta-analysis of GWAS results from different studies, and (ii) the expected R^2 of polygenic-score for a hold-out sample with SNP-weights based on the meta-analysis SNP-effect estimates.
Details on the underlying model, assumptions, and derivations, can be found in the De Vlaming et al. (2017).
Notes: (i) h^2 denotes SNP-based heritability and N sample size, (ii) lists for h^2 and N need to be comma-separated, (iii) replacing a list by a single number enforces the same value across all studies in the meta-analysis.
Calculator
Please enter the following quantities and click on the SUBMIT button:
Number of studies in the meta-analysis (maximum = 250): C = 4
N per study in meta-analysis: N=(60000,40000,20000,80000)
h² per study in meta-analysis: h²=(0.40,0.30,0.20,0.50)
h² in hold-out sample: h^2_{out} = 0.40
Genetic correlation between studies in meta-analysis: ρ_G = 0.75
Genetic correlation between hold-out study and meta-analysis studies: $\rho_{G,out}$ = 0.75
Effective number of independent SNPs: S = 250000
Effective number of independent causal SNPs: M = 20000
SUBMIT
RESET
B


1. Single-study analysis using genotypes
→ Choose one of the options listed below:
Either fill in the number of each category of genotypes:
Cases: aa1 ab1 bb1
Controls: aa0 ab0 bb0
or upload a file and perform a GWAS (example)
(please upload very large files using one of the available formats: zip, rar, bz, gz, tgz)
→ Select method to perform analysis with:
min2 max MERT CATT (additive) CATT (dominant) CATT (recessive)
2. Meta-analysis using genotypes
Upload your file (example):
(please upload very large files using one of the available formats: zip, rar, bz, gz, tgz)
→ Select method to perform analysis with:
min2 max MERT CATT (additive) CATT (dominant) CATT (recessive)
→ Select type of effect: random fixed
3. Single-study min2 analysis using p-values
Either fill in the p-values: Prob1 Prob2
or upload a file and perform a GWAS (example)
(please upload very large files using one of the available formats: zip, rar, bz, gz, tgz)
4. min2 meta-analysis using p-values
Upload your file (example):
(please upload very large files using one of the available formats: zip, rar, bz, gz, tgz)
→ Select type of effect: random fixed
5. (optional) Enter your email address to get notification of results:
Perform analysis
Clear fields
Disclaimer: For privacy reasons, all user-submitted or uploaded data are deleted immediately after processing.
The results of the analyses remain on the server for 1 day and then deleted as well.
If you find GWAR useful in your research, please consider citing the reference that describes this work:
GWAR: Tools for Robust Analysis and Meta-Analysis of Genome-Wide Association Studies
Niki L. Dimou, Konstantinos D. Tsirigos, Arne Elofsson and Pantelis G. Bagos
Bioinformatics, 2017 / PMID: 28108451
C

IGESS	https://github.com/daviddaigithub/IGESS	https://pubmed.ncbi.nlm.nih.gov/28498950	IPD+SD	R	Uses variational inference to increase statistical power and improve accuracy of risk prediction by integrating individual level genotype data and summary statistics
MetaGIM	https://github.com/fushengstat/MetaGIM	https://pubmed.ncbi.nlm.nih.gov/36964712	IPD+SD	R	A divide and conquer method to increase inference efficiency by incorporating aggregated summary information from other sources to an IPD analysis
LEP	https://github.com/daviddaigithub/LEP	https://pubmed.ncbi.nlm.nih.gov/30307540	IPD+SD	R	Integrates IPD and SD by Leveraging Pleiotropy to increase the statistical power of risk variants identification and the accuracy of risk prediction
PolyGIM	https://github.com/fushengstat/PolyGIM	https://pubmed.ncbi.nlm.nih.gov/37437002	IPD+SD	R	Uses polytomous logistic regression to investigate disease subtype heterogeneity in situations when only summary data is available
GWASmeta	https://github.com/sjl-sjtu/GWAS_meta	https://pubmed.ncbi.nlm.nih.gov/35286307	Meta-analysis	R	A method for the optimal ABF in the GWAS meta-analysis. Uses shotgun stochastic search to improve the Bayesian GWAS meta-analysis framework
nGWAMA	https://github.com/baselmans/multivariate_GWAMA	https://pubmed.ncbi.nlm.nih.gov/30643256/	Meta-analysis	R	Performs multivariate meta-analysis correcting for sample overlap
METAL	http://csg.sph.umich.edu/abecasis/Metal/	https://pubmed.ncbi.nlm.nih.gov/20616382/	Meta-analysis	C/C++	A versatile and efficient tool for meta-analysis of GWAS. It can combine test statistics and standard errors, or p-values across studies
PLINK	https://www.cog-genomics.org/plink/	https://pubmed.ncbi.nlm.nih.gov/17701901/	Meta-analysis	C/C++	A versatile program which supports data management, quality control, and common statistical computations including meta-analysis
GWAMA	https://genomics.ut.ee/en/tools	https://pubmed.ncbi.nlm.nih.gov/20509871/	Meta-analysis	Executable	A flexible, open-source tool for meta-analysis of GWAS. It incorporates a variety of error trapping facilities, and provides a range of meta-analysis summary statistics
YAMAS	https://github.com/cmeesters/yamas	https://pubmed.ncbi.nlm.nih.gov/22971100/	Meta-analysis	C/C++	Meta-analysis including missing SNPs identified with LD (proxy SNPs)
GWAR	http://www.compgen.org/tools/GWAR	https://pubmed.ncbi.nlm.nih.gov/28108451/	Meta-analysis	Stata/web	Analysis and meta-analysis of GWAS using standard as well as robust methods (MAX, MIN2, MERT)
MAGENTA	https://software.broadinstitute.org/mpg/magenta/	https://pubmed.ncbi.nlm.nih.gov/20714348/	Meta-analysis	Matlab	Meta-analysis with gene set enrichment analysis (GSEA)
CPBayes	https://cran.r-project.org/web/packages/CPBayes/index.html	https://pubmed.ncbi.nlm.nih.gov/29432419	Meta-analysis	R	Bayesian method for studying cross-phenotype genetic associations.
metaSKAT	https://cran.r-project.org/web/packages/MetaSKAT/index.html	https://pubmed.ncbi.nlm.nih.gov/23768515	Meta-analysis	R	Extensions of the Burden Test, SKAT and Optimal SKAT (SKAT-O) for multiple studies.
METACARPA	https://github.com/hmgu-itg/metacarpa	https://pubmed.ncbi.nlm.nih.gov/23424128	Meta-analysis	C/C++	Meta-analysis of GWAS with overlapping or related samples, when details of the overlap or relatedness are unknown
metaCCA	https://www.bioconductor.org/packages/release/bioc/html/metaCCA.html	https://pubmed.ncbi.nlm.nih.gov/27153689/	Meta-analysis	R	Multivariate analysis and meta-anaysis of GWAS. It uses canonical correlation analysis and employs a covariance shrinkage algorithm to achieve robustness
metaUSAT	https://github.com/RayDebashree/metaUSAT	https://pubmed.ncbi.nlm.nih.gov/29226385	Meta-analysis	R	A method for multiple traits. It is robust to the association structure of correlated traits. It can also be used to analyze a single trait over multiple studies, accounting for overlapping samples
CPASSOC	http://hal.case.edu/~xxz10/zhu-web/	https://pubmed.ncbi.nlm.nih.gov/25500260/	Meta-analysis	R	Method applicable to a multivariate phenotype containing any type of components including continuous, categorical and survival phenotypes, as well as to samples consisting of families or unrelated samples
GCPBayes	https://github.com/tbaghfalaki/GCPBayes	https://pubmed.ncbi.nlm.nih.gov/33368447	Meta-analysis	R	Bayesian meta-analysis methods for pleiotropy that extend CPBayes to the gene or pathway level
MetABF	https://github.com/trochet/metabf	https://pubmed.ncbi.nlm.nih.gov/30920090	Meta-analysis	R	A simple Bayesian framework for performing integrative meta-analysis across multiple GWAS
rareMETALS	https://genome.sph.umich.edu/wiki/Rare_Variant_Analysis_and_Meta-Analysis	https://pubmed.ncbi.nlm.nih.gov/30016313	Meta-analysis	R	Works even when the data contain large amounts of missing values. Uses a score statistic called PCBS (partial correlation based score statistic) for conditional analysis of single-variant and gene-level associations
meta-simulation	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863759/	https://pubmed.ncbi.nlm.nih.gov/29577025	Meta-analysis	R	A tool to implement an alternate strategy for the additive model based on simulating data for the individual studies
metaGAP	https://www.devlaming.eu/metagap.html	https://pubmed.ncbi.nlm.nih.gov/28095416	Meta-analysis	web	A versatile tool for calculating the statistical power of a meta-analysis of GWAS results and of the polygenic-score R^2 in a hold-out sample
rfdr	http://bioinformatics.ust.hk/RFdr.html	https://pubmed.ncbi.nlm.nih.gov/28067114	Meta-analysis	R	A method for replication of GWAS. Provides the most powerful significance levels when controlling the FDR in the two-stage study
jlfdr	https://bioinformatics.hkust.edu.hk/Jlfdr.html	https://pubmed.ncbi.nlm.nih.gov/28011772	Meta-analysis	R	A joint analysis method based on controlling the joint local false discovery rate
RRate	http://bioinformatics.ust.hk/RRate.html	https://pubmed.ncbi.nlm.nih.gov/27687799	Meta-analysis	R	A Bayesian probabilistic measure of the Replication Rate with which we can determine the sample size of the replication study and to check the consistency between the primary and the replication study.
MAJAR	https://github.com/JustinaXie/MAJAR	https://pubmed.ncbi.nlm.nih.gov/37519295	Meta-analysis	R	Method to jointly test prognostic and predictive effects in meta-analysis without the need of using an independent cohort for replication of the detected biomarkers
sPLINK	https://exbio.wzw.tum.de/splink/	https://pubmed.ncbi.nlm.nih.gov/35073941	Meta-analysis	Python	Performs privacy-aware GWAS on distributed datasets while preserving accuracy
XPEB	https://med.stanford.edu/tanglab/software/XPEB.html	https://pubmed.ncbi.nlm.nih.gov/25892113	Meta-analysis	R	An empirical Bayes approach to improve the power of GWAS in a minority population by exploiting information from another ethnic population

Heritability analysis

- Heritability is generally defined as the fraction of phenotypic variation explained by genetic variation. Heritability is a dimensionless parameter of the population, and it was introduced by Sewall Wright and Ronald Fisher in the previous century. Traditionally, heritability is estimated using family-based designs such as twin studies. However, there are controversies regarding the various methodologies for estimation and interpretation of the results
- heritability can also be estimated via the results obtained in a traditional GWAS using unrelated individuals. The gap between these estimates and those obtained from classical heritability estimation methods has been termed the "missing heritability problem" and it is an important open question in current research

Box 3 | **Historical background**

It has become standard to use the symbol h^2 for heritability because Sewall Wright[87] used h (for heredity) to denote the correlation between genotype and phenotype in his path coefficient model[87]. The square of that correlation (that is, h^2) is, per definition, the proportion of variation in the phenotype that is attributable to the path from genotype to phenotype. Ronald Fisher, in his classical 1918 paper, parameterized the resemblance between relatives in terms of correlation and regression coefficients, but also gives an example of the percentage of the total variance in stature in humans that can be ascribed to genotypes and to 'essential genotypes'[88]. These percentages correspond to what we now call broad-sense and narrow-sense heritability (BOX 1). It is thought that J. L. Lush was the first to formally use the term 'heritability' to describe the proportion of variation that is due to hereditary factors[89].

Box 1 | **Heritability and the partitioning of total variance**

Population parameters

Observed phenotypes (P) of a trait of interest can be partitioned, according to biologically plausible nature–nurture models, into a statistical model representing the contribution of the unobserved genotype (G) and unobserved environmental factors (E):

$$\text{Phenotype (P)} = \text{Genotype (G)} + \text{Environment (E)} \quad (1)$$

The variance of the observable phenotypes (σ^2_P) can be expressed as a sum of unobserved underlying variances (σ^2_G and σ^2_E):

$$\sigma^2_P = \sigma^2_G + \sigma^2_E \quad (2)$$

Heritability is defined as a ratio of variances, by expressing the proportion of the phenotypic variance that can be attributed to variance of genotypic values:

$$\text{Heritability (broad sense)} = H^2 = \frac{\sigma^2_G}{\sigma^2_P}$$

The genetic variance can be partitioned into the variance of additive genetic effects (breeding values; σ^2_A), of dominance (interactions between alleles at the same locus) genetic effects (σ^2_D), and of epistatic (interactions between alleles at different loci) genetic effects (σ^2_I):

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_I$$

$$\text{and heritability (narrow or strict sense)} = h^2 = \frac{\sigma^2_A}{\sigma^2_P}$$

In general, σ^2_E can be broken down into any number of identifiable, but random, contributing factors that can be specific to the phenotype. Examples include the environmental variance that is common to specified groups, for example, siblings and litters (σ^2_{CE}), and the non-genetic variance that is common to repeated measures of individuals (σ^2_{PE}).

We define the remainder of the environmental variance, which cannot be attributed to other factors, as the environmental residual variance, which includes individual stochastic error variance and measurement error (σ^2_{RE}):

$$\sigma^2_E = \sigma^2_{CE} + \sigma^2_{PE} + \sigma^2_{RE}$$

In the simplest partitioning, no specific factors that contribute to σ^2_E are identified and $\sigma^2_{RE} = \sigma^2_E$. Both the genetic and environmental variances can be partitioned further for a trait such as birth weight of the offspring to include genetic and environmental maternal effects that are attributable to the mother[75].

The partitioning of the phenotypic variance (equation 2) assumes the absence of genotype by environment covariance ($\sigma_{G,E}$). Examples leading to a positive covariance are parents with a high intelligence quotient (IQ) providing an IQ-stimulating environment for their children, and dairy cattle being fed according to production. A further term that is ignored in equations 1 and 2 is the interaction between genotype and environment (G*E), when the effect of the genotype depends on the environment. The most studied, yet still controversial, example of G*E in humans is the interaction between stressful life events (the environment) and the length polymorphism of the serotonin transporter gene (the genotype) and their effects on major depression (the phenotype)[76]. If G*E exists, P = G + E + G*E, so a more complete partitioning of phenotypic variance is:

$$\sigma^2_P = \sigma^2_G + \sigma^2_E + 2\sigma_{G,E} + \sigma^2_{G*E}$$

Both G and E covariation and G*E interaction are often ignored, usually because they cannot be estimated. If either is present, ignoring the former will inflate estimates of σ^2_G and ignoring the latter will inflate estimates of σ^2_E (REF. 3).

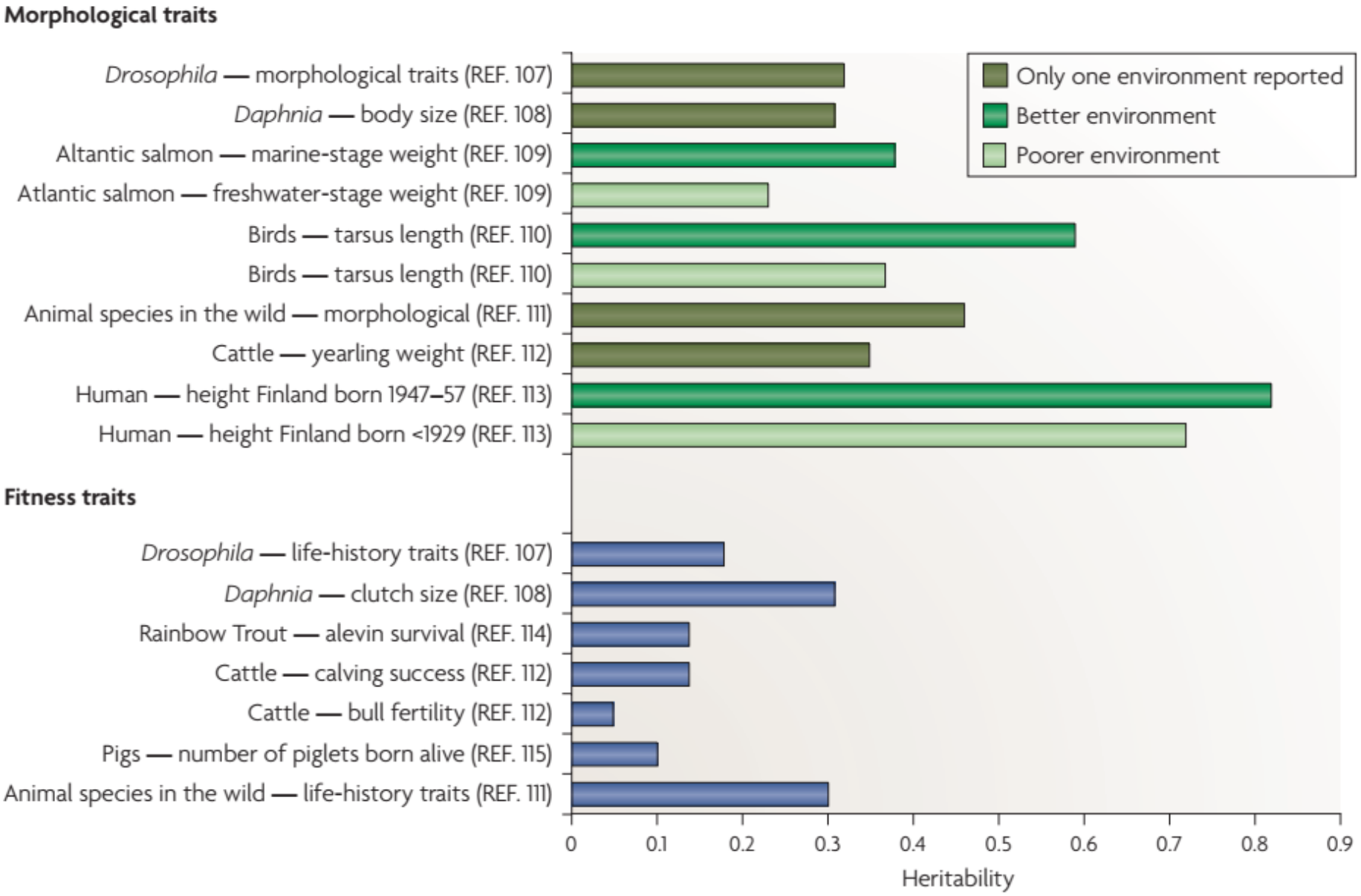


Figure 1 | **Examples of estimates of heritabilities of morphological and fitness traits.** Where possible, the estimates of heritability were taken from Reviews, and are the mean across a number of studies. The examples show that, on average, heritability estimates are larger for morphological traits than for fitness-related traits, and that heritability tends to be larger in better environments when compared with poorer environments.

Box 5 | **Heritability of categorical traits**

The heritability of all-or-none (0/1) traits, such as disease status, twinning rate or survival, can be defined in the usual way, that is, by the proportion of variation on an observed scale, for example, 0 and 1, that is due to additive genetic factors, and can be estimated as for continuous traits by, for example, parent–offspring regression or sibling correlation. However, variances and heritabilities calculated on this observed scale (h^2_O) are a function of the incidence of the trait in the population[2,3]. For example, the phenotypic variance on the observed scale for a 0/1 trait with an incidence of K is K(1–K), with a maximum at K = 0.5. This relationship between mean and variance obscures the comparison of the importance of genetic factors in different environments or in different populations that differ in incidence.

Because most quantitative traits follow a normal bell-shaped distribution, it is reasonable to assume that all-or-none traits can be represented by an underlying normally distributed liability trait, which, as for other traits, is the sum of independent normally distributed genetic and environmental components[16,19]. This assumption implies that liability to disease is multifactorial and that contributions from individual genetic or environmental risk factors are small. If the score on the liability scale exceeds a threshold then the individual has a phenotypic value of 1, otherwise it is 0, with the proportion of the normal distribution that exceeds the threshold being equal to the trait incidence. The relationship between h^2_O and the narrow-sense heritability on the underlying continuous liability scale (h^2) is:

$$h^2_O = h^2 z^2/[K(1–K)]$$

where z is the height of the standard normal curve at the threshold that truncates the proportion K (REF. 99). Heritability on the observed scale is always smaller than that on the liability scale because information is lost by the grouping into two categories, and the maximum value for h^2_O is 0.64 when K = 0.5 and $h^2 = 1$. For categorical traits with more than two classes, heritability can be estimated by assuming that the categories relate to multiple thresholds across an underlying liability scale[2,3]. Estimation of heritability for susceptibility to disease in human populations is often based on the threshold liability model.

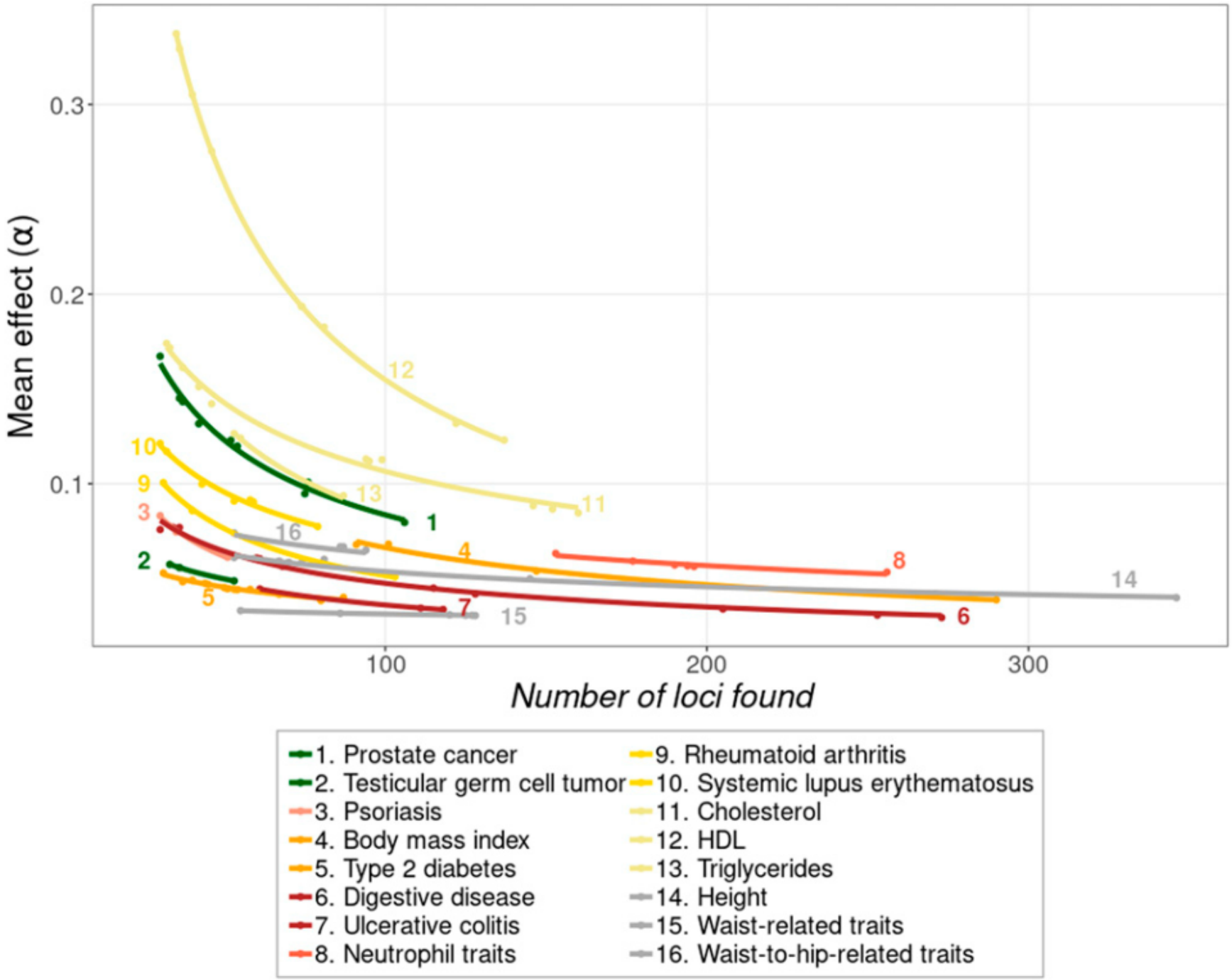


Figure 1 Decline of the average locus effect (α) with the number of loci found. The points represent the cumulated results of successive GWAS with increasing larger sample sizes. The first point at the left of the series is the mean effect of loci found in the GWAS with the lowest sample size (conditional on finding at least 30 loci), and the following points give the mean effect of loci as additional ones are found by studies with larger sample sizes (usually, but not always, by more recent studies). The lines are the fit of the observations to an exponential model (average R^2 = 0.96). Traits are colored depending on the functional domain they belong: cancer (green), dermatological (pink), endocrine (orange), gastrointestinal (brown), hematological (red), immunological (yellow), metabolic (beige), skeletal (gray). The final set of data corresponding to the last (right-hand side) points for each line are given in Table S3.

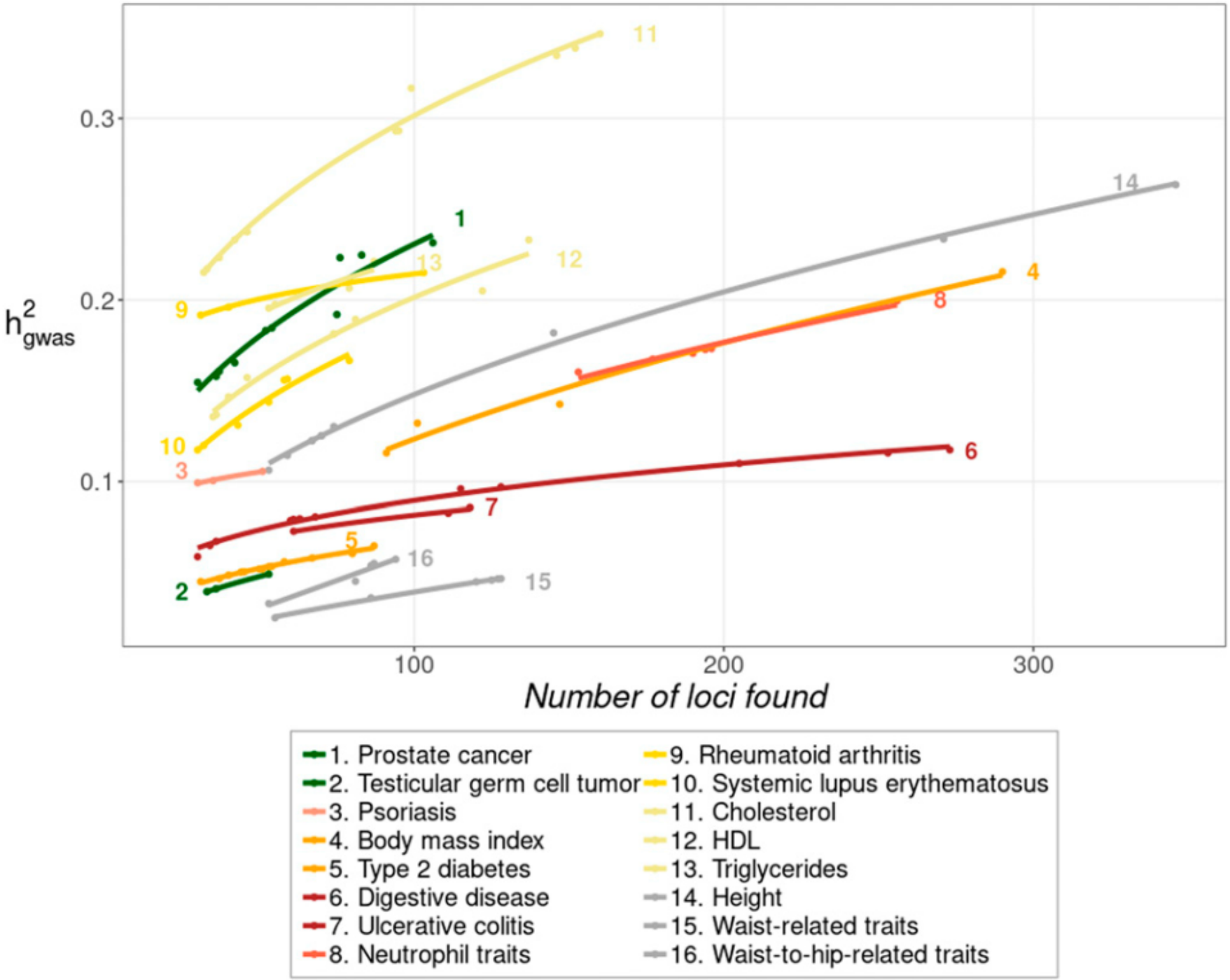


Figure 2 Increase of heritability explained by loci found (h^2_{gwas}) as the number of these increases. The points represent the observed values, while the lines are the fit to an exponential model (average R^2 = 0.97). Traits are colored depending on the functional domain they belong: cancer (green), dermatological (pink), endocrine (orange), gastrointestinal (brown), hematological (red), immunological (yellow), metabolic (beige), skeletal (gray). The final set of data corresponding to the last (right-hand side) points for each line are given in Table S3.

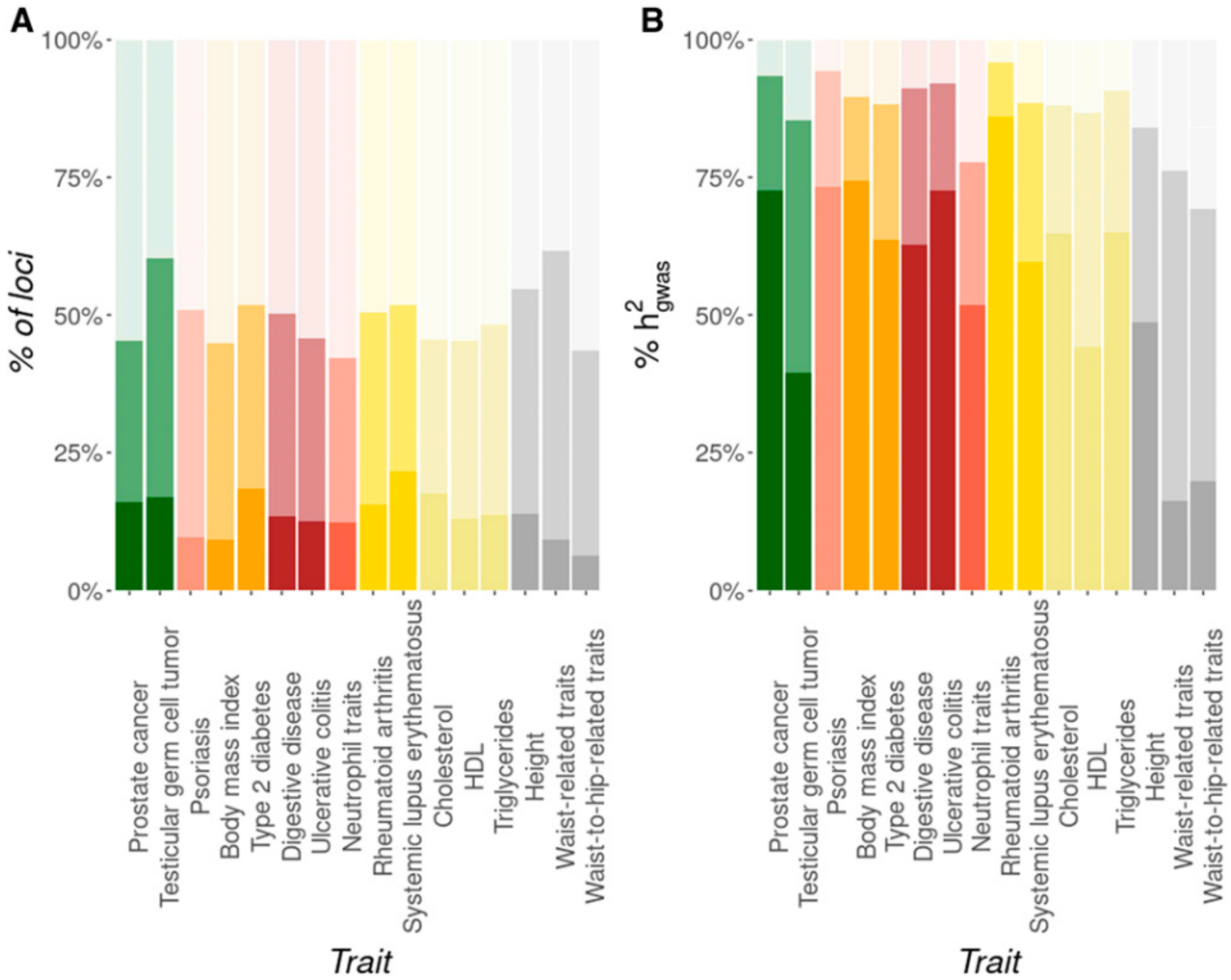


Figure 3 Percentage of loci for different classes of effect sizes and their contributions to heritability (in %). (A) Three arbitrary classes of locus effect sizes (high, medium, and low effects) are assumed such that ~50% of loci are within the low-effect class (high transparency), ~36% within the medium-effect class (low transparency), and ~14% within the large-effect class (solid colors). (B) Contribution (in percentage) of the three classes to heritability. Traits are ordered and colored by functional domain: Cancer (green), dermatological (pink), endocrine (orange), gastrointestinal (brown), hematological (red), immunological (yellow), metabolic (beige), skeletal (gray).

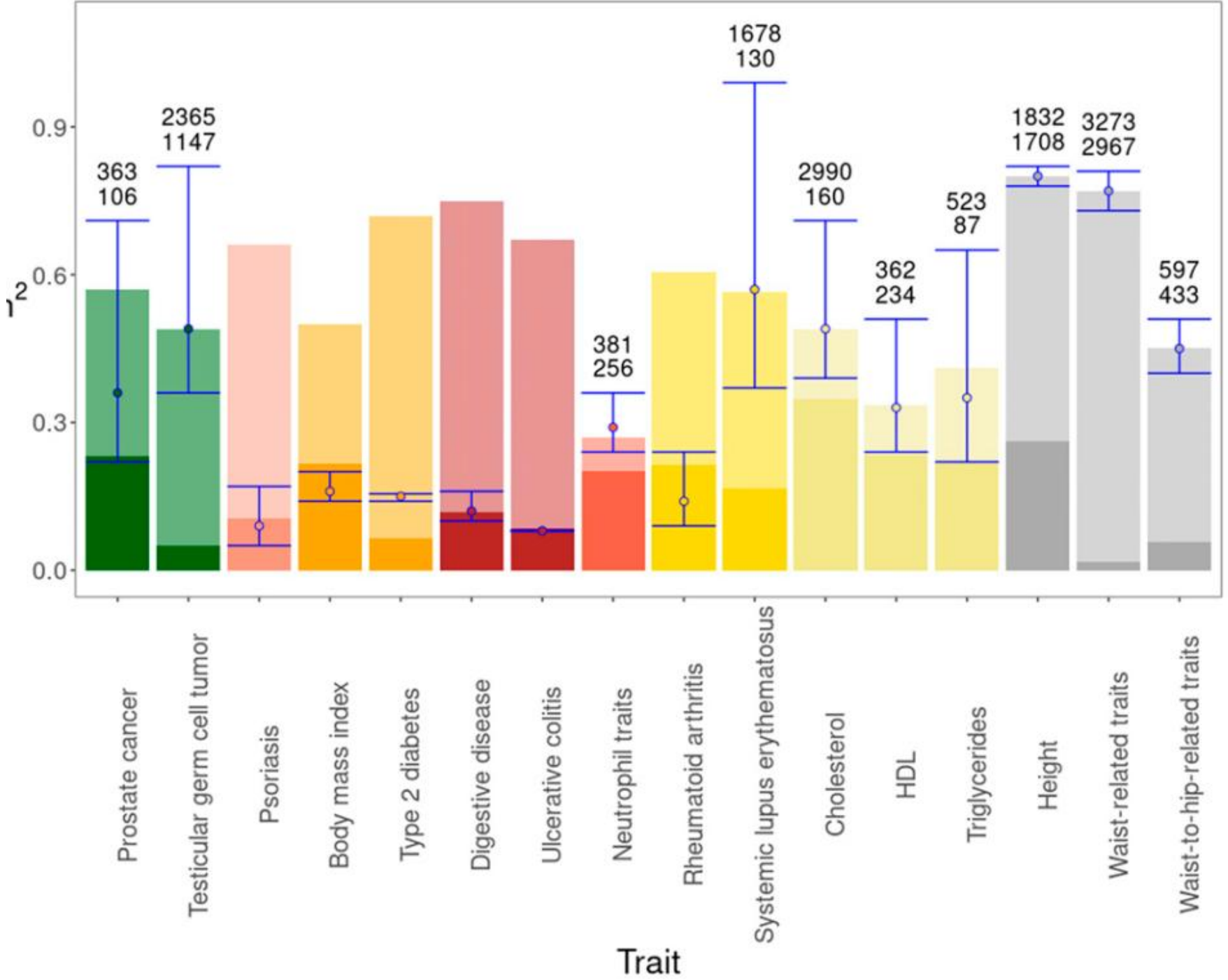


Figure 4 Observed and expected values of heritability. The full length of bars indicate the mean familial heritability (h^2_{fam}) for the studied traits (average values are shown when there is a range of estimates from the Literature, Table S2). In solid color it is shown the heritability explained by the loci already found and available from the Catalog (h^2_{gwas}). The blue error bar gives the inferred value of heritability (the dot corresponds to the median value) that approaches most to the familial heritability with a 95% confidence interval, using data from the expected distribution of locus effects. The expected number of loci for each trait required to explain the familial heritabilities within the error bars assuming an additive contribution of single loci are given over the bars. Traits are colored depending on the functional domain they belong: Cancer (green), dermatological (pink), endocrine (orange), gastrointestinal (brown), hematological (red), immunological (yellow), metabolic (beige), skeletal (gray).

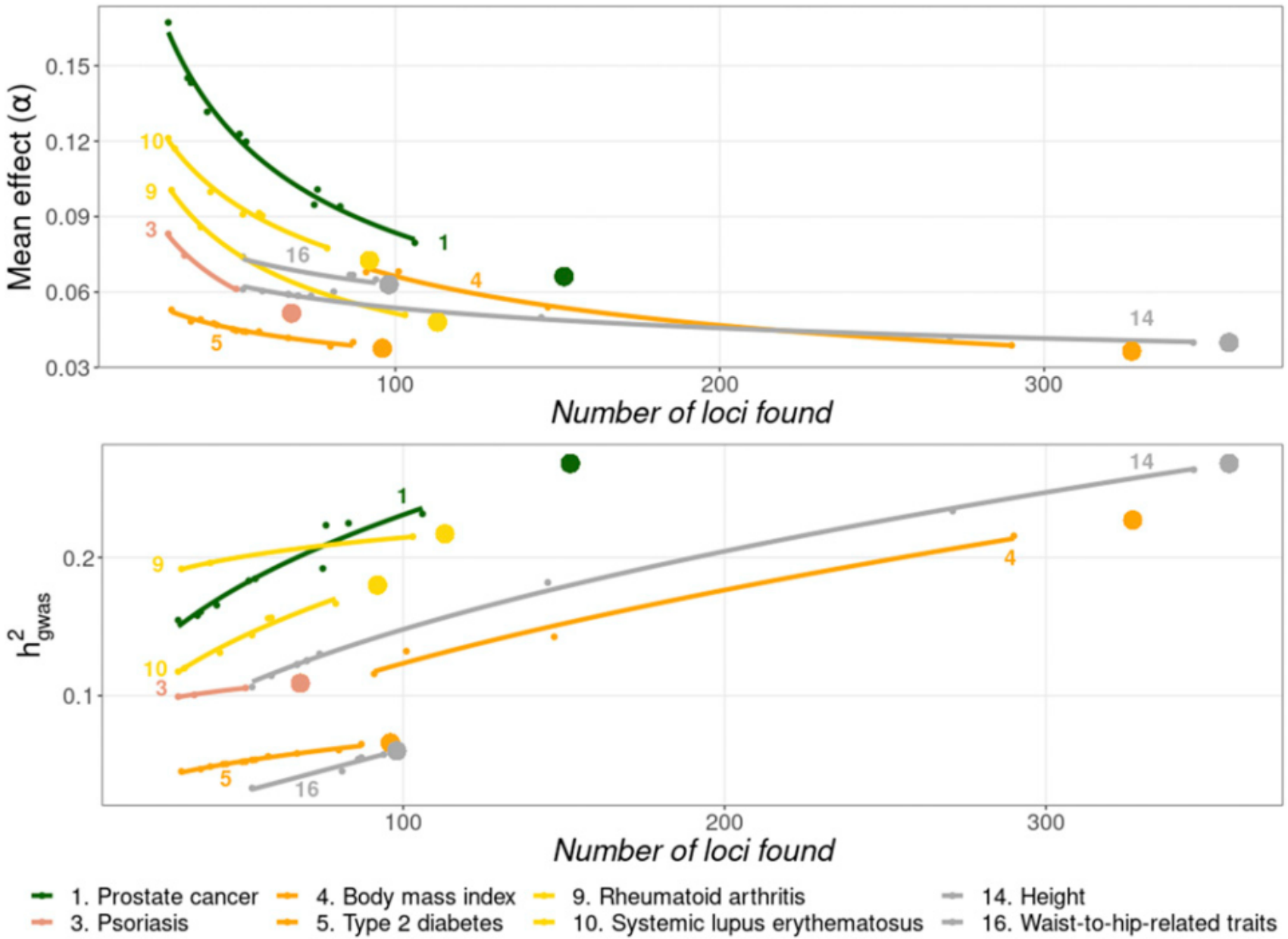


Figure 5 Decline of the average locus effect (upper graph), and increase of the heritability explained (h^2_{gwas}) (lower graph) as the number of loci found is increasing. The small points represent the observed values of the previous analyses (Figure 1 and Figure 2) and the large points those of a more recently collected set. Lines are the fit to an exponential model. Traits are colored depending on the functional domain they belong: Cancer (green), dermatological (pink), endocrine (orange), immunological (yellow), skeletal (gray).

- One of the first and simplest methods to calculate heritability from allele frequency, odds ratio and prevalence of the disease was implemented in the **SumVg** package. This method, however, utilizes only the significant SNPs. The same authors extended the method later in order to allow calculation using the z-statistics from the whole GWAS sample. A disadvantage of this method is that LD is not taken care of, and highly correlated SNPs need to be filtered manually.
- **AVENGEME** is a tool that treats causal effect sizes as fixed effects and models the genotypes as random correlated variables.
- **HESS** which was presented later built upon the same ideas and can be viewed as a weighted sum of the squares of the projection of effect sizes onto the eigenvectors of the LD matrix at the particular locus, with weights inversely proportional to the corresponding eigenvalues.
- LD Score Regression (**LDSC**) has been frequently applied to summary statistics from GWAS and one of its functionalities is to estimate the SNP heritability of a trait. **LDER** extends LDSC making full use of the information from the LD matrix providing more accurate estimates, whereas **s-LDSC** is an extension suitable for partitioning heritability.
- **SumHer** presented later and offers the same functionalities, with the main difference being that it allows for different so called “heritability models”. According to these, a SNP with high MAF is expected to contribute more to the total heritability compared to one with low MAF, whereas on the other hand, a SNP in a region of low LD is expected to contribute more compared to one in a region of high LD. On the contrary, LDSC estimates are obtained by assuming that all SNPs contribute equally.
- **HEELS** is a new tool using REML to produce accurate and precise local heritability estimates and **RSS**, is a multiple regression-based fine-mapping tool can also calculate SNP heritability from the regression model.
- **VarExp** and **GxESum** are methods for estimating the phenotypic variance explained by genome-wide gene-environment (GxE) interactions.

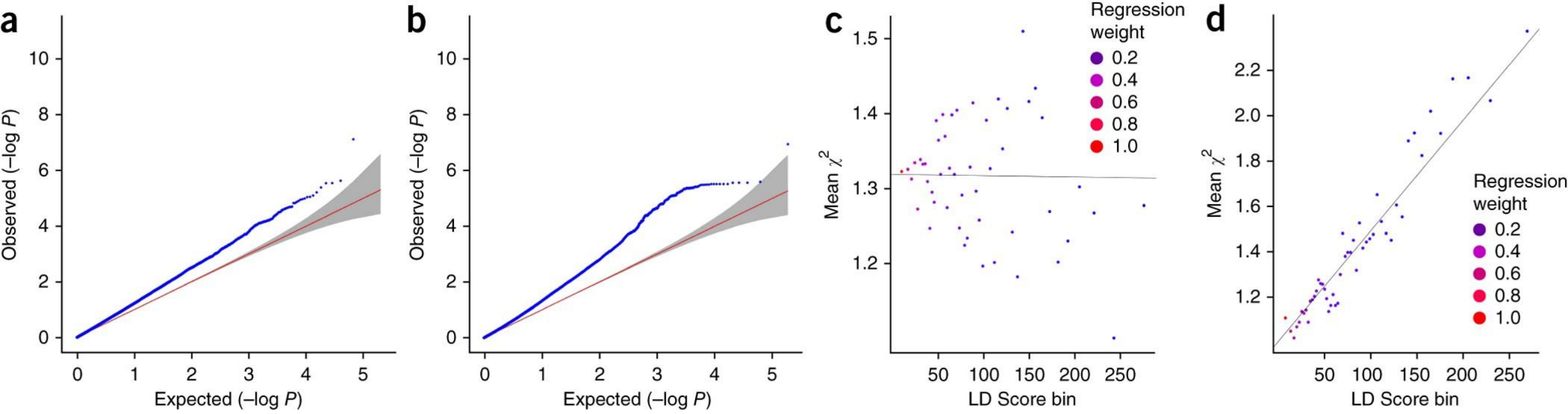
- There are also tools like **GWIZ** and **SummaryAUC** that calculate the Receivers Operator's Characteristic (ROC) curve and the associated Area Under the Curve (AUC). GWIZ generates ROC curves and the AUC using simulations and then estimates heritability using the square of the Somers' rank correlation D. SummaryAUC on the other hand approximates the AUC of a PRS and its variance.
- **HAMSTA** is a tool that, among others, estimates heritability explained by local ancestry using data from admixture mapping studies.
- Estimating the Effect size distribution is also a related important concept. **GENESIS** uses LD and a Likelihood-based approach to estimate effect-size distributions. It also allows predictions regarding yield of future GWAS with larger sample sizes. **GWEHS** calculates the distribution of effect sizes of SNPs, as well as their contribution to trait heritability. Furthermore, it performs predictions for the change in the effect size as well as in the heritability when new variants are identified. **FMR** is a method-of-moments for calculating the effect-size distribution and **GWAS-Causal-Effects-Model** is a random effects model for estimating the causal variants and their effect size distribution.
- Finally, there are tools to implicate gene-expression in heritability analysis: **MESC** which estimates the proportion of heritability mediated by gene expression levels using linkage disequilibrium (LD) scores and eQTL, and **GCSC** which uses results from a TWAS (see "TWAS and Colocalization") in the so-called gene co-regulation score regression, to identify gene sets enriched for disease heritability

Under a polygenic model, such that effect sizes are drawn independently from distributions with variance proportional to $p(1\text{-}p)^{-1/2}$ where p is minor allele frequency (MAF), then the expected χ^2-statistic of variant j is

$$E[\chi^2|\ell_j]=Nh^2\ell_j/M+Na+1, \quad (1)$$

where N is sample size; M is the number of SNPs, such that h^2/M is the average heritability explained per SNP; a measures the contribution of confounding biases, such as cryptic relatedness and population stratification; and $\ell_j := \Sigma_k r_{jk}^2$ is the *LD Score* of variant j, which measures the amount of genetic variation tagged by j and (a full derivation of this equation is provided in the Supplementary Note). This relationship holds for meta-analyses, and also for ascertained studies of binary phenotypes, in which case h^2 is on the observed scale. Consequently, if we regress χ^2statistics from GWAS against LD Score (LD Score regression), the intercept minus one is an estimator of the mean contribution of confounding bias to the inflation in the test statistics.

Results from selected simulations. (a) QQ plot with population stratification (λGC = 1.32, LD Score regression intercept = 1.30). (b) QQ plot with polygenic genetic architecture with 0.1% of SNPs causal (λGC = 1.32, LD Score regression intercept = 1.006) (c) LD Score plot with population stratification. Each point represents an LD Score quantile, where the x-coordinate of the point is the mean LD Score of variants in that quantile and the y-coordinate is the mean χ2 of variants in that quantile. Colors correspond to regression weights, with red indicating large weight. The black line is the LD Score regression line. (d) As in panel c but LD Score plot with polygenic genetic architecture.



MESC	https://github.com/douglasyao/mesc	https://pubmed.ncbi.nlm.nih.gov/32424349	Heritability	Python	Estimates the proportion of heritability mediated by assayed gene expression levels using linkage disequilibrium (LD) scores and eQTL
GENESIS	https://github.com/yandorazhang/GENESIS	https://pubmed.ncbi.nlm.nih.gov/30104760	Heritability	R	Uses LD information and a Likelihood-based approach to estimate variants effect-size distributions. It also allows users to make predictions regarding yield of future GWAS.
GWEHS	https://gitlab.com/elcortegano/GWEHS	https://pubmed.ncbi.nlm.nih.gov/31344961	Heritability	R	Calculates the distribution of effect sizes of SNPs affecting traits, as well as their contribution to heritability. It also allows for predictions as new loci are found
GWIZ	https://github.com/jonaspatronjp/GWIZ-Rscript/ https://gwasrocs.ca	https://pubmed.ncbi.nlm.nih.gov/31805043	Heritability	R	A method to generate ROC curves and calculate the AUROC
SummaryAUC	https://github.com/lsncibb/SummaryAUC	https://pubmed.ncbi.nlm.nih.gov/30911754	Heritability	R	A method for approximating the AUC and its variance of a PRS when only the summary level data of the validation dataset are available.
GxEsum	https://github.com/honglee0707/GxEsum	https://pubmed.ncbi.nlm.nih.gov/34154633	Heritability	R	A method for estimating the phenotypic variance explained by genome-wide GxE
VarExp	https://gitlab.pasteur.fr/statistical-genetics/VarExp	https://pubmed.ncbi.nlm.nih.gov/29726908	Heritability	R	A method that allows for the estimation of the proportion of phenotypic variance explained. It allows for a range of models to be evaluated, including marginal genetic effects, GxE interaction effects and both effects jointly
SumVG	https://github.com/lab-hcso/SumVg	https://pubmed.ncbi.nlm.nih.gov/21618601/	Heritability	R	Provides estimates of the sum of heritability explained by all true susceptibility variants in GWAS. It also estimates the standard error based on re-sampling approaches
SumHer	http://dougspeed.com/sumher/	https://pubmed.ncbi.nlm.nih.gov/30510236/	Heritability	Executable	Estimates the SNP Heritability of a trait, Heritability Enrichments and Genetic Correlations between traits
AVENGEME	https://sites.google.com/site/fdudbridge/software/	https://pubmed.ncbi.nlm.nih.gov/26189816	Heritability	R	A method to estimate the variance in disease liability explained by large sets of genetic markers. Uses polygenic scores, based on the formula for the non-centrality parameter of the association test of the score.
HESS	https://github.com/huwenboshi/hess	https://pubmed.ncbi.nlm.nih.gov/26189816	Heritability	Python	Provides utilities for estimating and analyzing local SNP-heritability and genetic covariance
HAMSTA	https://github.com/tszfungc/HAMSTA	https://pubmed.ncbi.nlm.nih.gov/37875120/	Heritability	Python	Estimates the heritability explained by local ancestry in admixture mapping studies. It also quantifies inflation in test statistics that is not contributed by local ancestry effects, and determines significance threshold for admixture mapping
HEELS	https://github.com/huilisabrina/HEELS	https://pubmed.ncbi.nlm.nih.gov/38040712/	Heritability	Python	Uses REML to produce accurate and precise local heritability estimates
LDER	https://github.com/shuangsong0110/LDER	https://pubmed.ncbi.nlm.nih.gov/35421325	Heritability	R	Extends the LDSC method making full use of the information from the LD matrix and provides more accurate estimates of heritability and confounding inflation
s-LDSC	https://github.com/bulik/ldsc	https://pubmed.ncbi.nlm.nih.gov/26414678	Heritability	R	Extension of LDSC for partitioning heritability
FMR	https://github.com/lukejoconnor/FMR	https://pubmed.ncbi.nlm.nih.gov/34326547	Heritability	Matlab	A method-of-moments estimator of the effect-size distribution. The coefficients quantify the heritability explained by components of a mixture model for the effect-size distribution
GWAS-Causal-Effects-Model	https://github.com/dominicholland/GWAS-Causal-Effects-Model	https://pubmed.ncbi.nlm.nih.gov/32427991	Heritability	Matlab	Random effects model for estimating the causal variants and their effect size distribution from a dense panel
GCSC	https://github.com/ksiewert/GCSC	https://pubmed.ncbi.nlm.nih.gov/35108496	Heritability	Python	Uses TWAS results in a gene co-regulation score regression, to identify gene sets that are enriched for disease heritability explained by predicted expression

Gene-based tests

- Gene-based tests aggregate individual variant associations within a gene, providing a more comprehensive assessment of the gene's overall contribution to a trait or disease. This approach helps prioritize genes with multiple associated variants, enhancing the biological relevance of findings, and it has proven to be useful particularly in case of low frequency variants
- There are plenty of different methods for combining the association statistics or p-values within a gene, ranging from simple Fisher's method or the minimum p-value approach, to more advanced methods like the Burden Test (BT) or quadratic tests like SKAT with variations in power. Nevertheless, there is a consensus regarding the importance of incorporating LD information of the nearby variants into the methods for controlling the type I error rate at the desired level

Box 2 | **Rare variant association tests using summary association statistics**

Let X be an $N \times M$ matrix of genotypes, standardized to mean 0 and variance 1, and Y be an $N \times 1$ matrix of standardized trait values, where M is the number of rare variants (for example, in a given gene being tested for association) and N is the number of samples. An $M \times 1$ vector of z-scores (estimated effect sizes divided by their standard errors) can be computed as

$$Z = \frac{X^T Y}{\sqrt{N}}$$

with multivariate normal null distribution $Z \sim N(0, V)$, where V is an in-sample linkage disequilibrium matrix.

Burden tests

Burden tests assume that all rare variants in a candidate gene have the same direction of effect. Burden tests may either assume that standardized effect sizes are the same for each rare variant[112] (that is, per-allele effect sizes are proportional to

$$\frac{1}{\sqrt{p_i(1-p_i)}}$$

where p_i is the allele frequency), or apply weights or thresholds based on allele frequency or functional information[113,114]. If w is an $M \times 1$ vector of weights for each rare variant (including zero weights for rare variants excluded by a threshold), the test statistic for a weighted burden test is $T_{burden} = w^T Z$ with null distribution $T_{burden} \sim N(0, w^T V w)$. This test statistic can naturally be extended to a meta-analysis of burden tests from multiple cohorts (via inverse-variance weighting), and can be extended to variable threshold tests and binary traits[40–42].

Overdispersion tests

Overdispersion tests assume that rare variants in a candidate gene can affect a complex trait in either direction, and can be computed as weighted sums of squared single-variant test statistics[115,116]. If $W = \mathrm{diag}(w_1, \ldots, w_M)$ is an $M \times M$ diagonal matrix of weights for each rare variant, the test statistic for a weighted overdispersion test is $T_{overdispersion} = Z^T W Z$ with null distribution $T_{burden} \sim \Sigma_i \mu_i \chi_i^2$, where weights μ_i for each χ^2 (1 d.f.) distribution χ_i^2 are given by eigenvalues of the matrix $V^{1/2} W V^{1/2}$. This test statistic can be extended to a meta-analysis of overdispersion tests from multiple cohorts (via inverse-variance weighting), and can be extended to binary traits[40–42].

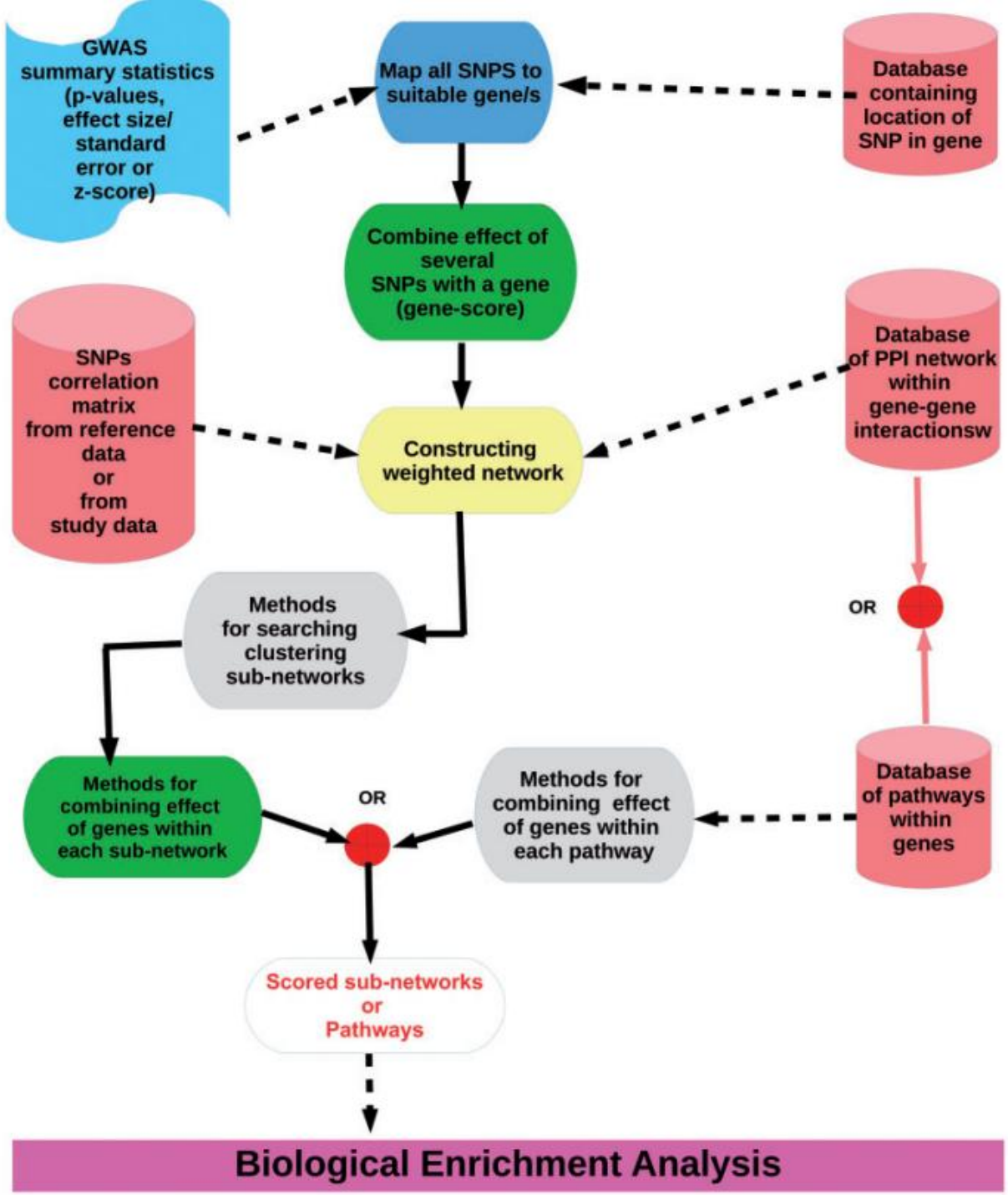
- **SKAT**, **VEGAS**, **GATES**, **fastBAT** and **GCTA** are among the oldest tools, which remain efficient and widely used.
  - **SKAT** (Sequence Kernel Association Test) is a regression method for testing association between variants and traits adjusting for covariates. As a score-based variance-component test, SKAT calculates p-values analytically by fitting the null model containing only the covariates.
  - **GCTA** and **VEGAS** also use the multivariate normal framework adjusting the estimates for LD using a reference panel. Of note, **GCTA** also offers methods for conditional analysis (see "Fine mapping"), and same also holds for **KGG**, whereas **VEGAS's** new version allows for mixed ethnicity populations.
  - GATES, on the other hand, uses an extended Simes procedure that integrates functional information and association evidence to combine p-values, whereas **fastBAT** offers fast analytical p-value computations.
- The gene analysis in **MAGMA** (Multi-marker Analysis of GenoMic Annotation) is based on a multiple linear principal components' regression model to account for LD and uses an F-test to compute the overall gene p-value.
  - Its extension, **nMAGMA**, extends the lists of genes that can be annotated by integrating local signals, long-range regulation signals, and tissue-specific gene networks. It also provides tissue-specific risk signals, which are useful for understanding disorders with multi-tissue origins.
  - **H-MAGMA** and **eMAGMA** are two other extensions. The former integrates 3D chromatin configuration, whereas the latter leverages significant tissue-specific cis-eQTL information to assign SNPs to putative genes.
- **EPIC** and **GAMBIT** also utilize functional data for gene-based analysis; the former using cell-type-specific gene expression data obtained from single-cell RNA sequencing and the latter using coding and tissue-specific regulatory annotations. Such methods share several features in common with TWAS methods (see respective section).

- **AgglomerativLD** also captures LD between SNPs of nearby genes, which induces correlation of the gene-based test statistics.
- **DOT** is one of the few methods that applies a decorrelation-based approach before combining SNP-level statistics or p-values.
- Tools like **GPA**, **oTFisher**, **TS** and **aSPU** implement some type of so-called adaptive tests (AT), that is, they account for possibly varying association patterns across SNPs,
- some modern tools like **MKATR**, **COMBAT**, **MCA**, **OWC**, **FST**, **ACAT**, **HYST**, **GBJ** and **sumFREGAT** perform analysis with multiple statistical methods and test and combine the results.
- Notably, tools like **aSPU**, **snpGeneSets**, **Pascal/PascalX**, **MAGMA**, **chromMAGMA** and **FUMA**, also offer the option of performing gene-set analysis after performing the gene-based analysis (see next section), whereas **HSVS-M** tests the association of a gene with multiple correlated traits.

JEPEG/JEPEGMIX	http://dleelab.github.io/jepegmix/ , http://dleelab.github.io/jepeg	https://pubmed.ncbi.nlm.nih.gov/26428293 , https://pubmed.ncbi.nlm.nih.gov/25505091	Gene-based tests	Executable	A gene-based method for testing the joint effects on trait for SNPs functionally associated with a gene (eQTLs)
VEGAS	https://cran.r-project.org/web/packages/snpsettest/index.html	https://pubmed.ncbi.nlm.nih.gov/20598278/	Gene-based tests	R	One of the first multivariate methods. Takes account of LD between markers in a gene by using simulation based on the LD of a reference panel
AgglomerativLD	https://github.com/ryurko/Agglomerative-LD-loci-testing	https://pubmed.ncbi.nlm.nih.gov/34459489	Gene-based tests	R	Captures LD of SNPs falling in nearby genes, which induces correlation of gene-based test statistics
sumFREGAT	https://cran.r-project.org/web/packages/sumFREGAT	https://pubmed.ncbi.nlm.nih.gov/35653402	Gene-based tests	R	Offers a wide range of gene-based methods to combine. It allows the user to arbitrarily define a set of these methods, weighting functions and probabilities of genetic variants being causal.
LDAK-GBAT	https://dougspeed.com/ldak-gbat/	https://pubmed.ncbi.nlm.nih.gov/36480927	Gene-based tests	Executable	A computationally efficient method for gene-based association testing. Produces well-calibrated p values and is significantly more powerful than existing tools
nMAGMA	https://github.com/sldrcyang/nMAGMA	https://pubmed.ncbi.nlm.nih.gov/33230537	Gene-based tests	R	An extension of MAGMA which extends the lists of genes that can be annotated to SNPs by integrating local signals, long-range regulation signals, and tissue-specific gene networks. It also provides tissue-specific risk signals, which are useful for understanding disorders with multitissue origins
MKATR	http://www.github.com/baolinwu/mkatr.	https://pubmed.ncbi.nlm.nih.gov/29558699	Gene-based tests	R	The method calculates the correlation of the the test Z-statistics across variants using LD from a population reference panel. Incorporates various tests (sum test, SKAT, adaptive test)
GPA	https://github.com/Biocomputing-Research-Group/GPA	https://pubmed.ncbi.nlm.nih.gov/31392781	Gene-based tests	C/C++	A general gene-based p-value adaptive combination approach (GPA) which can integrate association evidence of multiple SNPs. It is applicable to both continuous and binary traits and also to multiple studies
OWC	https://github.com/Xuexia-Wang/OWC-R-package	https://pubmed.ncbi.nlm.nih.gov/36597047	Gene-based tests	R	A gene-based test that incorporates different weighting schemes and includes several popular methods as its special cases (burden test, weighted sum of squared score test [SSU], weighted sum statistic [WSS], SNP-set Kernel Association Test [SKAT], and score test)
MCA	https://github.com/biostatpzeng/MCA	https://pubmed.ncbi.nlm.nih.gov/36042399	Gene-based tests	R	Implements 22 different gene-based methods, including linear regression, higher criticism tests, Berk-Jones tests, burden test; SKAT and SKAT-O, Simes and GATES, aggregated Cauchy association test and more
SKAT	https://cran.r-project.org/web/packages/SKAT	https://pubmed.ncbi.nlm.nih.gov/21737059/	Gene-based tests	R	A supervised and computationally efficient regression method to test for association between genetic variants and a trait adjusting for covariates. As a score-based variance-component test, itcan quickly calculate p values analytically
COMBAT	https://cran.r-project.org/web/packages/COMBAT/	https://pubmed.ncbi.nlm.nih.gov/28878002	Gene-based tests	R	A combined association test for genes, which incorporates strengths from existing gene-based tests and shows higher overall performance than individual tests
oTFisher	https://cran.r-project.org/web/packages/TFisher/index.html	https://pubmed.ncbi.nlm.nih.gov/36468009	Gene-based tests	R	The omnibus thresholding Fisher's method for performing SNP-set and gene-based tests
H-MAGMA	https://github.com/thewonlab/H-MAGMA	https://pubmed.ncbi.nlm.nih.gov/36289406	Gene-based tests	R	Extends MAGMA by incorporating 3D chromatin configuration in assigning variants to their putative target genes
eMAGMA	https://github.com/eskederks/eMAGMA-tutorial	https://pubmed.ncbi.nlm.nih.gov/33624746	Gene-based tests	C/C++	Gene-based approach with a modification of MAGMA, leverages significant tissue-specific cis-eQTL information to assign SNPs to putative genes
EPIC	https://github.com/rujinwang/EPIC	https://pubmed.ncbi.nlm.nih.gov/35709291	Gene-based tests	R	Method that relates large-scale GWAS summary statistics to cell-type-specific gene expression measurements from single-cell RNA sequencing
GAMBIT	https://github.com/corbinq/GAMBIT	https://pubmed.ncbi.nlm.nih.gov/33320851	Gene-based tests	C/C++	Integrates heterogeneous annotations with GWAS summary statistics for gene-based analysis, using various coding and tissue-specific regulatory annotations. Allows various tests like SKAT, minP, ACAT, HMP etc
MARS	https://github.com/junghyunJJ/marsR	https://pubmed.ncbi.nlm.nih.gov/33931127	Gene-based tests	R	Finds associations between variants in risk loci and a phenotype, considering the causal status of variants
GBJ	https://cran.r-project.org/web/packages/GBJ/index.html	https://pubmed.ncbi.nlm.nih.gov/33041403	Gene-based tests	R	Generalized Berk-Jones test for the association between a SNP-set and outcome by accounting for LD. Includes also tests for Berk-Jones (BJ), Higher Criticism (HC), Generalized Higher Criticism (GHC), Minimum p-value (minP), and an an omnibus test (OMNI) which integrates information from each of the tests.
GENE-E	https://github.com/ramachandran-lab/genee	https://pubmed.ncbi.nlm.nih.gov/32542026	Gene-based tests	R	A gene-based test using an empirical Bayesian approach and a mixture of normal distributions over the (regularized) effect size estimates
PEGASUS	https://github.com/ramachandran-lab/PEGASUS	https://pubmed.ncbi.nlm.nih.gov/27489002	Gene-based tests	Perl	Gene-based method that uses an analytical approach to compute gene-level P-values of observed gene scores according to a null distribution modeling LD
FST	https://cran.r-project.org/web/packages/FSTpackage/	https://pubmed.ncbi.nlm.nih.gov/28844485	Gene-based tests	R	Combining dispersion and burden tests and an efficient perturbation method for individual gene/large gene-set/genome wide analysis
ACAT	https://github.com/yaowuliu/ACAT	https://pubmed.ncbi.nlm.nih.gov/30849328	Gene-based tests	R	A gene-based method using the Cauchy Combination Test. Includes also an omnibus procefure combining SKAT, BT and ACAT
DOT	https://github.com/dmitri-zaykin/Total_Decor	https://pubmed.ncbi.nlm.nih.gov/32287273	Gene-based tests	R	Decorrelation-based approach (DOT) for combining SNP-level summary statistics (or, equivalently, P-values)
fastBAT	https://yanglab.westlake.edu.cn/software/gcta	https://pubmed.ncbi.nlm.nih.gov/27604177	Gene-based tests	C/C++	Performs a fast set-based association analysis for human complex traits using summary-level data from GWAS and LD
KGG	http://pmglab.top/kggsee/#/	https://pubmed.ncbi.nlm.nih.gov/30101339	Gene-based tests	Java	Conditional test that uses a sequential analysis with a linear combination of chi-square statistics
TS	https://github.com/Jianjun-CN/c-code-for-TS	https://pubmed.ncbi.nlm.nih.gov/32366212	Gene-based tests	Executable	Uses a truncated method to find the genes that have a true contribution to the genetic association
HSVS-M	https://github.com/yiyangphd/HSVSM	https://pubmed.ncbi.nlm.nih.gov/34787916	Gene-based tests	R	Multivariate hierarchically structured variable selection model, a flexible Bayesian model that tests the association of a gene with multiple correlated traits
GATES	http://pmglab.top/kggsee/#/	https://pubmed.ncbi.nlm.nih.gov/21397060/	Gene-based tests	Java	An extended Simes test that integrates functional information and association evidence to combine the p-values of the SNP within a gene to obtain an overall p-value
HYST	http://pmglab.top/kgg/	https://pubmed.ncbi.nlm.nih.gov/22958900/	Gene-based tests	Java	A set-based statistical mwthod combining the extended Sime's test and the scaled chi-square test to examine the overall association significance in a set of SNPs
GCTA	https://yanglab.westlake.edu.cn/software/gcta/	https://pubmed.ncbi.nlm.nih.gov/22426310	Gene-based tests	C/C++	An approximate conditional and joint association analysis that uses LD from a reference sample

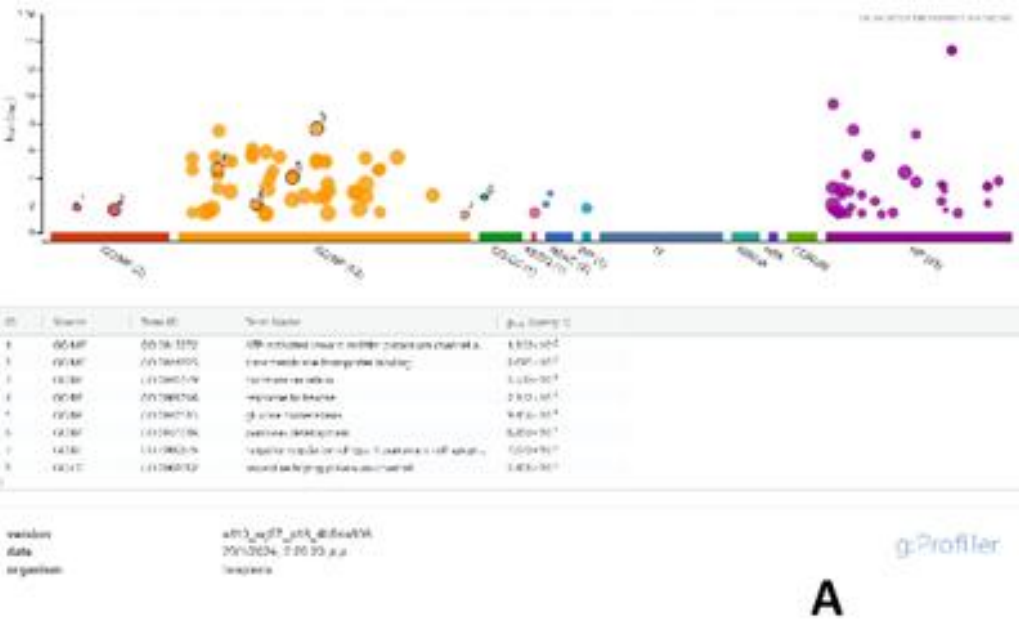
Gene Set Analysis

- Gene set analysis (GSA), or Pathway Analysis, extends the concept of gene-based methods by jointly analyzing groups of functionally related genes and identifying biological pathways enriched with trait-associated genes. By considering the collective impact of multiple genes within a pathway, researchers can obtain a clearer picture of the underlying biological mechanisms influencing the phenotype under investigation.
- The first applications of such methods borrowed ideas from the microarray data analysis literature, and since then they became widespread in analysis of GWAS.
- Any GSA method needs to address some issues.
  - Firstly, how to handle SNPs of the same gene; secondly, how to define the appropriate gene-set or pathway, and finally how to combine the effects from multiple SNPs/genes within the same set/pathway. Thus, the choices made by different methods can be very diverse leading to a wide variety of different approaches. For instance, some methods operate with SNP-level statistics (effect sizes, z, or p-values) assigning the SNP to the closest gene (usually within a range of ±20K bases), whereas others take as input a gene-level statistic or simply a gene list obtained by a gene-based method (of course, several tools allow for both a gene-based and a GSA approach).
  - Regarding the choice of set there is a plethora of databases containing biological pathways (KEGG, PANTHER etc), or other types of gene-set representation like PPI interactions, ontologies and so on.
  - Finally, regarding the statistical method used to aggregate evidence there is also a wide range of different methods that handle with different approaches the gene set size and gene length, the LD patterns and the presence of overlapping genes within pathways, or apply different statistical approaches such as those using the so-called competitive null hypothesis, or those using the self-containing one.


GWAS summary statistics (p-values, effect size/ standard error or z-score)
Map all SNPS to suitable gene/s
Database containing location of SNP in gene
Combine effect of several SNPs with a gene (gene-score)
SNPs correlation matrix from reference data or from study data
Constructing weighted network
Database of PPI network within gene-gene interactionsw
OR
Methods for searching clustering sub-networks
Methods for combining effect of genes within each sub-network
OR
Methods for combining effect of genes within each pathway
Database of pathways within genes
Scored sub-networks or Pathways
Biological Enrichment Analysis

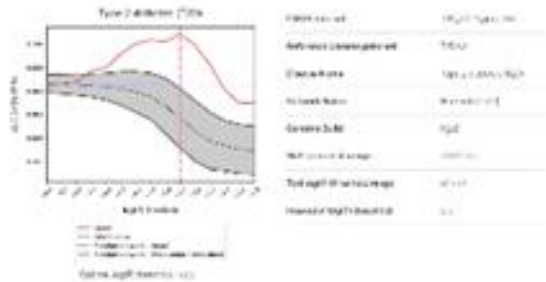
- Among the most easily used and frequently cited are the tools that utilize a webserver. **FUMA** and **iGSE4GWAS** are tools specialized in GWAS and use SNP-level statistics as inputs, differing in the subsequent analyses: FUMA uses MAGMA for gene-based testing and allows for ORA and Kologorov-Smirnov test (GSEA), whereas iGSE4GWAS maps the most significant SNP to a gene and then performs an improved GSEA with label permutation to obtain accurate p-values.
- Tools like **Enrichr**, **g:Profiler**, **DAVID**, **WebGestalt** and **PANTHER** are general purpose enrichment tools that provide functionalities for different types of omics data. They accept gene or SNP-list as input and provide Application Programming Interface (API) ensuring interoperability, whereas for the statistical analysis they all use some version of ORA and/or GSEA (WebGestalt also uses Network Topology-based Analysis). A major feature of these tools is that they incorporate a large number of biological and pathway databases, with g:Profiler and Enrichr offering the most complete collection.
- **GSA-SNP2** is one of the first methods to be developed for GWAS and has seen several improvements regarding the calculation of the combined gene score and the execution time, being among the fastest methods.
- **aSPUpath2** and **GIGSEA** are two methods that integrate expression data (eQTL) in the pathway analysis. The former uses an adaptive test that extends the aSPU methodology based on chi-square, whereas the latter uses a regression-based approach coupled with permutations to calculate accurate p-values. In a similar fashion, **deTS** and **PGCA** perform tissue-specific enrichment analysis (TSEA) for detecting tissue-specific genes and for enrichment test of different forms of query data

- Other methods use different definitions of the gene-sets, in some cases utilizing additional information. For instance, **dmGWAS** integrates PPI networks and uses a search method to identify subnetworks. Compared with standard pathway methods it offers to the users the flexibility in the definition of a gene set and can utilize local PPI information.
- **GEMB** defines the gene-sets using gene weights from model predictions and gene ranks from GWAS, and **GENOMICper** uses permutations of the identified SNPs by rotation with respect to the genomic locations.
- **GWAB** uses network connections to reprioritize candidate genes by integrating the GWAS and network data, whereas **GenToS** searches for trait-associated variants in existing human GWAS.
- We also need to mention **PAPA** which is a flexible tool for pleiotropic pathway analysis.
- As we already mentioned, **aSPU**, **snpGeneSets**, **PascalX/PASCAL** and **MAGMA/chromMAGMA** are gene-based methods that also perform GSA, whereas **MAGENTA** is a tool that performs meta-analysis and subsequently GSA (see “meta-analysis”).
- Lastly, we need to mention **Inferno** and **Mergeomics** which are webservers offering a variety of options, extending typical GSA applications. Inferno integrates a variety of functional genomics sources to identify causal noncoding variants using COLOC, WebGestalt, LDSC and MetaXcan. Mergeomics uses summary statistics of multi-omics association studies (GWAS, EWAS, TWAS, PWAS, etc.) and performs correction for LD, GSEA, meta-analysis and identification of regulators of disease-associated pathways and networks.

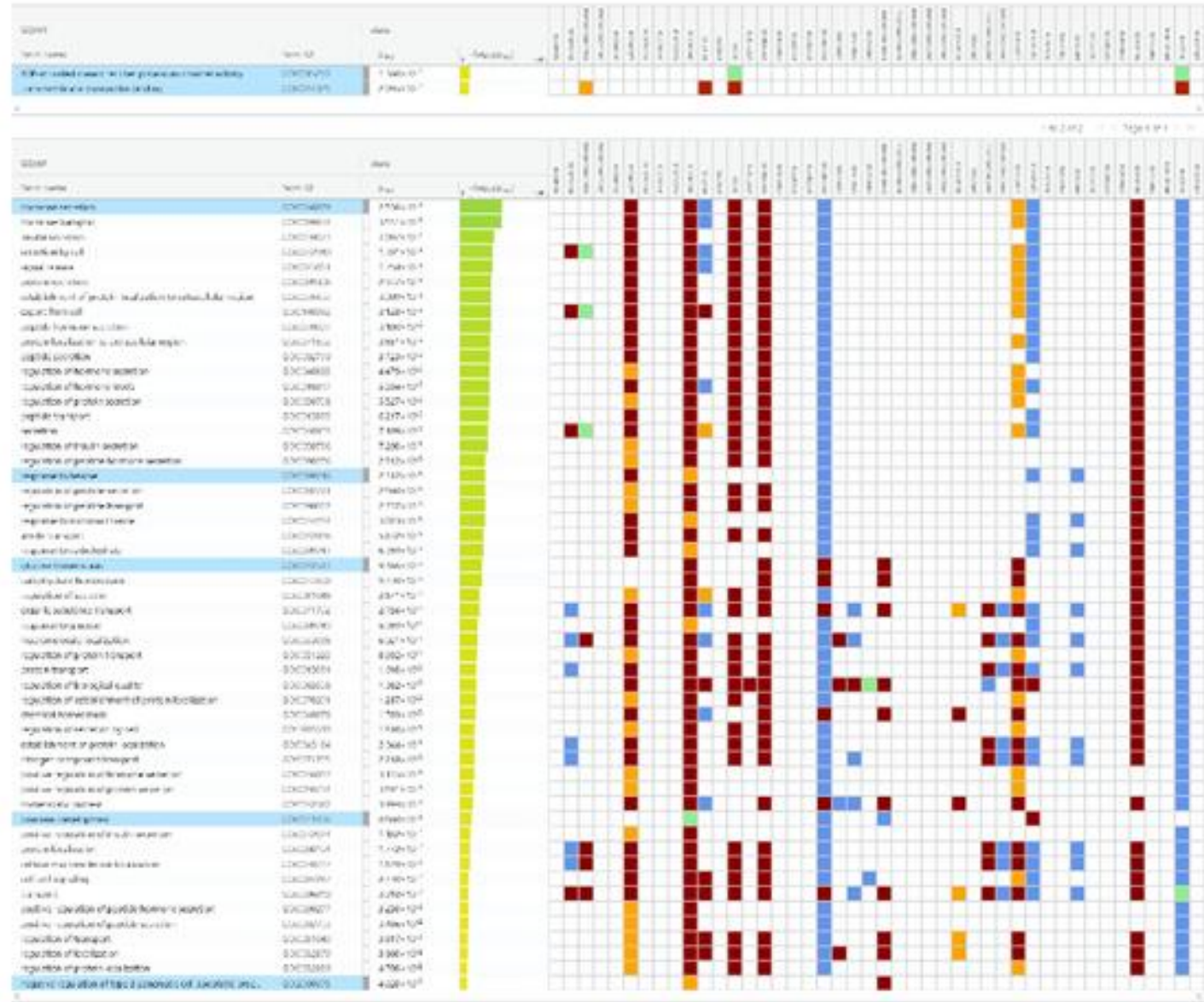


A

B



C



D

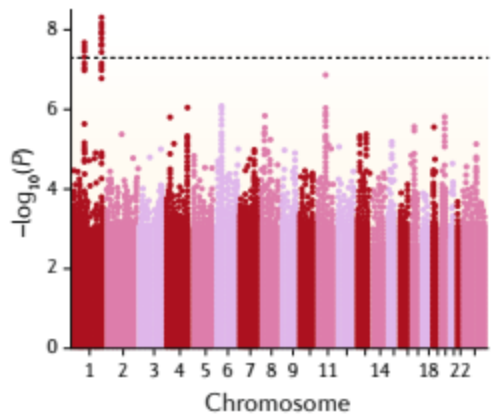
aSPUpath2	https://github.com/ChongWu-Biostat/aSPUpath2	https://pubmed.ncbi.nlm.nih.gov/29411426/	GSA	R	Integrates gene expression reference weights, GWAS summary data, LD information, and candidate pathways to identify pathways whose expression is associated with complex traits
GIGSEA	www.github.com/zhushijia/GIGSEA	https://pubmed.ncbi.nlm.nih.gov/30010968	GSA	R	Uses GWAS and eQTL to infer differential gene expression and interrogate gene set enrichment for the trait-associated SNPs. By incorporating expression data it naturally accounts for factors such as gene size, gene boundary, SNP distal regulation and multiple-marker regulation
GAUSS	https://github.com/diptavo/GAUSS	https://pubmed.ncbi.nlm.nih.gov/33730541	GSA	R	Tests for any self-contained association between a phenotype and a gene-set and produces a p-value for the association.
GSA-SNP2	https://sourceforge.net/projects/gsasnp2	https://pubmed.ncbi.nlm.nih.gov/29562348	GSA	C/C++	A method for pathway enrichment analysis of GWAS P-value data. It accepts also gene-wise p-values (obtained from other methods) and outputs pathway gene sets 'enriched' with genes associated with the given phenotype
VSEAMS	https://github.com/ollyburren/vseams	https://pubmed.ncbi.nlm.nih.gov/25170024	GSA	R	A non-parametric SNP set enrichment method to test for enrichment of GWAS signals in functionally defined loci using P-values
dmGWAS	https://bioinfo.uth.edu/dmGWAS/	https://www.ncbi.nlm.nih.gov/pubmed/21045073/	GSA	Executable	A dense module searching method to identify candidate subnetworks or genes for complex diseases by integrating PPI network. Extensively searches for subnetworks enriched with low P-value genes.
iGSE4GWAS	http://gsea4gwas.psych.ac.cn	https://pubmed.ncbi.nlm.nih.gov/20435672/	GSA	web	Detects pathways associated with traits by applying an improved gene set enrichment analysis. Implements also a follow-up functional analysis for SNPs in trait-associated pathways identified. Uses LD and putative functional annotation from Ensembl,ENCODE and eQTLs
Enrichr	https://maayanlab.cloud/Enrichr	https://pubmed.ncbi.nlm.nih.gov/27141961/	GSA	web	A gene set search engine that enables the querying of hundreds of thousands of annotated gene sets. Enrichr uniquely integrates knowledge from many high-profile projects to provide synthesized information about mammalian genes and gene sets.
SNPratio test	https://sourceforge.net/projects/snpratiotest/	https://pubmed.ncbi.nlm.nih.gov/19620097/	GSA	Perl	Compares the proportion of significant to all SNPs within genes that are part of a pathway and computes an empirical P-value based on comparisons to ratios in datasets where the assignment of case/control status has been randomized.
g:Profiler	https://biit.cs.ut.ee/gprofiler/gost	https://pubmed.ncbi.nlm.nih.gov/37144459/	GSA	web/R	Iintegrates many databases, including Gene Ontology, KEGG and TRANSFAC, to provide a comprehensive and in-depth analysis of gene lists. It also provides interactive and intuitive user interfaces and supports ordered queries and custom statistical backgrounds, among other settings.
DAVID	https://david.ncifcrf.gov/	https://pubmed.ncbi.nlm.nih.gov/35325185/	GSA	web/R	An enrichment tool with functionalities for different types of omics data including GWAS. It accepts gene or SNP-list as input and provide API ensuring interoperability. For analysis it uses ORA and GSEA
WebGestalt	http://www.webgestalt.org/	https://pubmed.ncbi.nlm.nih.gov/31114916/	GSA	web/R	An enrichment tool with functionalities for different types of omics data including GWAS. It accepts gene or SNP-list as input and provide API ensuring interoperability. For analysis it uses ORA, GSEA and Network Topology-based Analysis
PANTHER	http://www.pantherdb.org/	https://pubmed.ncbi.nlm.nih.gov/33290554/	GSA	web	An enrichment tool with functionalities for different types of omics data including GWAS. It accepts gene or SNP-list as input and provide API ensuring interoperability. For analysis it uses ORA and GSEA
deTS	https://cran.r-project.org/web/packages/deTS/index.html	https://pubmed.ncbi.nlm.nih.gov/30824912/	GSA	web/R	Performs tissue-specific enrichment analysis (TSEA) for detecting tissue-specific genes and for enrichment test of different forms of query data.
DESE	https://pmglab.top/pcga	https://pubmed.ncbi.nlm.nih.gov/31694669/	GSA	web	Detects the causal tissues of complex traits according to selective expression of disease-associated genes
PAPA	https://sourceforge.net/projects/papav1/files/	https://pubmed.ncbi.nlm.nih.gov/26568630	GSA	C/C++	A flexible tool for pleiotropic pathway analysis utilizing GWAS summary results
GEMB	https://github.com/cochran4/GEMB	https://pubmed.ncbi.nlm.nih.gov/33034635	GSA	Matlab	A method that combines gene weights from model predictions and gene ranks from genome-wide association studies into a weighted gene-set test
GENOMICper	https://cran.r-project.org/web/packages/genomicper/index.html	https://pubmed.ncbi.nlm.nih.gov/22973544/	GSA	R	Uses SNP association p-values and permutes them by rotation with respect to the genomic locations. The joint gene p-values are calculated using Fisher's combination test and pathways' association tested using the hypergeometric test
GWAB	https://www.inetbio.org/gwab/	https://pubmed.ncbi.nlm.nih.gov/28449091	GSA	web	Trait-associated genes with sub-threshold significance score can be rescued by network connections to other significant candidates
Inferno	http://inferno.lisanwanglab.org/	https://pubmed.ncbi.nlm.nih.gov/30113658	GSA	web/Python	Method which integrates diverse functional genomics data sources to identify causal noncoding variants. Characterizes the relevant tissue contexts, target genes, and downstream biological processes affected by functional variants.Uses COLOC, WebGestalt, LDSC and MetaXcan.
Mergeomics	http://mergeomics.research.idre.ucla.edu	https://pubmed.ncbi.nlm.nih.gov/34048577	GSA	web/R	Web server which uses summary statistics of multi-omics association studies (GWAS, EWAS, TWAS, PWAS, etc) and performs correction for LD, GSEA, meta-analysis and identification of essential regulators of disease-associated pathways and networks
GenToS	https://github.com/genepi-freiburg/gentos	https://pubmed.ncbi.nlm.nih.gov/27612175	GSA	Java	Calculates an appropriate statistical significance threshold and then searches for trait-associated variants in summary statistics from human GWAS
aSPU	https://cran.r-project.org/web/packages/aSPU/	https://pubmed.ncbi.nlm.nih.gov/27592708	GSA/Gene-Based	R	Performs adaptive gene-based test and pathway-based test for association analysis of multiple traits. The tests are adaptive at both the SNP- and trait-levels, thus maintaining high power across a wide range of situations. The methods can be applied to mixed types of traits, and to Z-statistics or P-values
snpGeneSets	https://www.umc.edu/SoPH/Departments-and-Faculty/Data-Science/Research/Services/Software.html/	https://www.ncbi.nlm.nih.gov/pubmed/27807048/	GSA/Gene-Based	R	Integrates local genomic annotation databases and provides genome-wide annotation for SNP, Gene and gene sets. It aims to support interpretation of GWAS results and performing post-analysis
PascalX	https://github.com/BergmannLab/PascalX	https://pubmed.ncbi.nlm.nih.gov/37137228	GSA/Gene-Based	Python	Provides fast and accurate mapping of SNP-wise GWAS data. It allows for scoring genes and annotated gene sets for enrichment signals based on data from, both, single GWAS and pairs of GWAS
PASCAL	https://www2.unil.ch/cbg/index.php?title=Pascal	https://pubmed.ncbi.nlm.nih.gov/26808494	GSA/Gene-Based	Java	Computes gene and pathway scores from SNP-phenotype associations. For gene score computation, implements analytic and efficient numerical solutions to calculate test statistics. For pathway scoring, it uses a modified Fisher method
MAGMA	https://ctg.cncr.nl/software/magma	https://pubmed.ncbi.nlm.nih.gov/25885710/	GSA/Gene-Based	C/C++	Uses p-values and performs gene-based and gene-set analysis as well as meta-analysis
FUMA	https://fuma.ctglab.nl	https://pubmed.ncbi.nlm.nih.gov/29184056/	GSA/Gene-Based	web	An integrative web-based platform using information from multiple biological resources to facilitate functional annotation of GWAS results, gene prioritization and interactive visualization. It accommodates positional, expression quantitative trait loci (eQTL) and chromatin interaction mappings, and provides gene-based, pathway and tissue enrichment results.

Fine-mapping

- Fine mapping, working in the opposite direction of that of the gene-based approaches, is a process aimed at narrowing down and identifying causal variants, that is the specific genetic variants responsible for the observed associations between genomic regions and traits of interest.
- The plethora of statistical methods and study designs makes it dificult to choose an optimal approach. The different approaches that have been proposed to perform fine-mapping can be divided in three broad categories:
  - heuristic methods that select SNPs based on LD patterns,
  - conditional or penalized regression models that perform variable selection, and
  - Bayesian methods that calculate posterior probabilities or Bayes Factors.
- Based on theoretical and empirical evidence it seems that Bayesian methods have superior performance. Several factors may influence the performance of fine-mapping approaches, including the true number of causal SNPs in a region and their effect sizes, the local LD structure, the sample size, and the SNP density.
- Functional annotations are also of great importance leading to the so-called functionally informed fine-mapping (FIFM) methods. The hypothesis of a single causal variant is also very restrictive, and several methods have been developed to allow multiple causal variants in a region as well as to incorporate additional layers of functional annotations, like eQTL. Moreover, methods for fine-mapping of multiple datasets have been proposed, either exploiting different LD patterns across ethnic groups or borrowing information between different traits.

GWAS
- Trait and genotypes (imputed)

Manhattan plot



List of associated SNPs

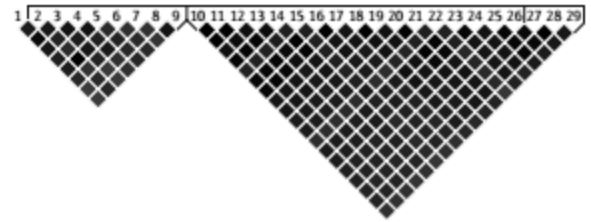
Explore each region

Partition to independent regions (forward regression)

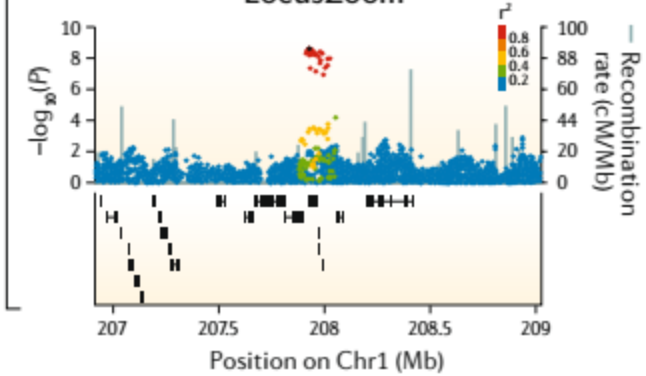
Fine-map each region
- Single study
- Multiple studies (meta-analyses)

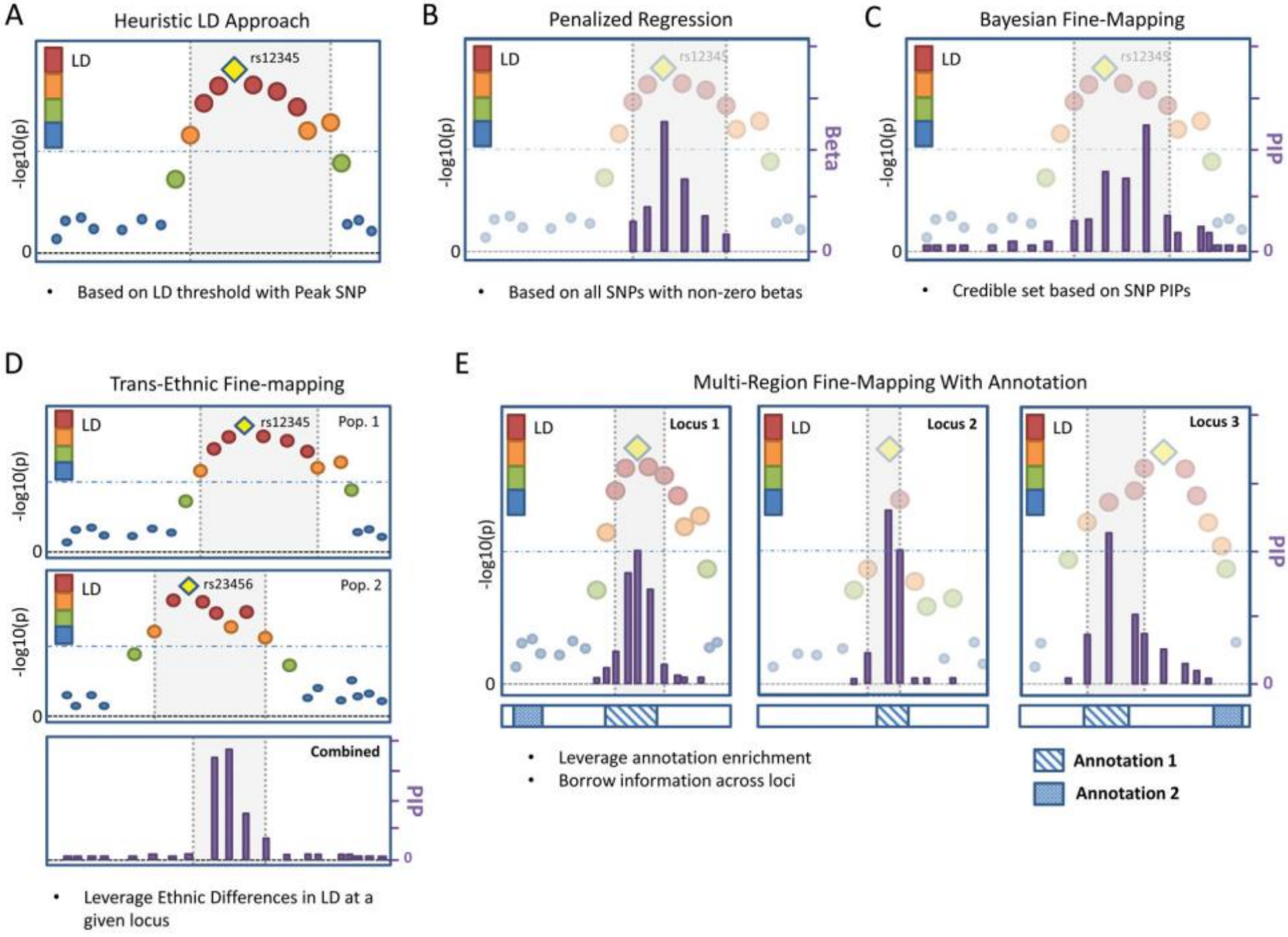
Consider likely function of selected SNPs
- Annotation from databases
- Integrate with gene expression (eQTL)
- Use annotation either after selection or as prior information for Bayesian methods

LD Haploview



LocusZoom




A
Heuristic LD Approach
LD
rs12345
-log10(p)
0
• Based on LD threshold with Peak SNP
B
Penalized Regression
LD
rs12345
-log10(p)
Beta
0
0
• Based on all SNPs with non-zero betas
C
Bayesian Fine-Mapping
LD
rs12345
-log10(p)
PIP
0
0
• Credible set based on SNP PIPs
D
Trans-Ethnic Fine-mapping
LD
rs12345
Pop. 1
-log10(p)
0
LD
rs23456
Pop. 2
-log10(p)
0
Combined
PIP
0
• Leverage Ethnic Differences in LD at a given locus
E
Multi-Region Fine-Mapping With Annotation
LD
Locus 1
LD
Locus 2
LD
Locus 3
-log10(p)
0
PIP
0
• Leverage annotation enrichment
• Borrow information across loci
Annotation 1
Annotation 2

- Bayesian methods seem to have superior performance and thus it is of no surprise that most of the currently available methods operate in a Bayesian framework calculating Posterior Inclusion Probabilities (PIP) and/or Bayes Factors (BFs) in various settings: **PAINTOR**, **DAP**, **fgwas**, **FINEMAP**, **flashfm**, **FINMOM**, **CARMA** and **CAVIAR/CAVIARBF**. **MsCAVIAR** is an extension of the latter method leveraging information from multiple studies, useful in trans-ethnic fine mapping.
- Similarly, **XMAP** performs cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias. **BEATRICE** is a unique method that combines a hierarchical Bayesian model with a deep learning-based inference procedure, whereas **RIVIERA-beta** performs Bayesian fine-mapping using Epigenomic Reference Annotation.
- On a different level, **PolyFun/PolyLoc** do not perform fine-mapping per se but are used for estimating the prior causal probabilities of SNPs, which can then be used by other Bayesian fine-mapping methods.
- **SusieR**, **BVS-PICA** and **JAM**, operate also in a Bayesian regression framework performing variable selection and penalized regression.
- Other regression-based methods, like **SOJO** and **ANNORE** work in a frequentist framework and perform lasso-type and differential shrinkage via random effects, respectively, whereas **GSR** utilizes a gene score regression approach and **RSS** performs multiple regression utilizing the so-called summary statistics likelihood.
- **AHIUT** performs an intersection-union test based on a joint/conditional regression model with all the SNPs in a region. Lastly, we need to mention **PICS2**, which performs probabilistic identification of causal SNPs and is the only of the methods that is available as a web-server, and **echocolatoR** which requires minimal input from users and integrates a suite of fine-mapping tools to identify consensus variants, test enrichment and visualize the results.

SOJO	https://github.com/zhenin/sojo	https://pubmed.ncbi.nlm.nih.gov/29198721/	Fine-mapping	R	Penalized selection operator for jointly analyzing multiple variants (SOJO) within each mapped locus on the basis of LASSO regression derived from summary association statistics
PolyFun/PolyLoc	https://github.com/omerwe/polyfun	https://pubmed.ncbi.nlm.nih.gov/33199916	Fine-mapping	Python	Estimates prior causal probabilities for SNPs, which can then be used by fine-mapping methods like SuSiE or FINEMAP. It can aggregate polygenic data from across the entire genome and hundreds of functional annotations.
SusieR	https://github.com/stephenslab/susieR	https://pubmed.ncbi.nlm.nih.gov/35853082	Fine-mapping	R	Performs variable selection in multiple regression using a Bayesian version of stepwise selection approach, and is particularly well-suited to settings where some of the variables are highly correlated
FINEMAP	http://www.christianbenner.com	https://pubmed.ncbi.nlm.nih.gov/26773131	Fine-mapping	Executable	Applies a shotgun stochastic search algorithm and can identify causal SNPs, estimate their effect sizes and the heritability contribution of causal SNPs in genomic regions associated with complex traits
flashfm	https://jennasimit.github.io/flashfm/	https://pubmed.ncbi.nlm.nih.gov/34686674	Fine-mapping	R	Uses summary statistics to jointly fine-map genetic associations for several related quantitative traits in a Bayesian framework that leverages information between the traits
MsCAVIAR	https://github.com/nlapier2/MsCAVIAR	https://pubmed.ncbi.nlm.nih.gov/34543273	Fine-mapping	C/C++	A method for fine-mapping by leveraging information from multiple studies. One important application area is trans-ethnic fine mapping.
CAVIARBF	https://bitbucket.org/Wenan/caviarbf/src/master/	https://pubmed.ncbi.nlm.nih.gov/25948564	Fine-mapping	C/C++	A fine-mapping method that combines CAVIAR with Bayesian inference using marginal test statistics
PICS2	https://pics2.ucsf.edu	https://pubmed.ncbi.nlm.nih.gov/33624747	Fine-mapping	web	Probabilistic Identification of Causal SNPs is a fine-mapping tool for determining the likelihood that each SNP in LD with a reported index SNP is a true causal polymorphism
ANNORE	https://github.com/vafisher/AnnoRE	https://pubmed.ncbi.nlm.nih.gov/34302344	Fine-mapping	R	Uses local LD structure and functional annotation, accross many categories, to prioritize the most plausible causal SNPs. It is based on multiple regression with differential shrinkage via random effects
JOINTSUM	https://github.com/yangq001/conditional	https://pubmed.ncbi.nlm.nih.gov/32275709	Fine-mapping	R	A simple and general approach based on conditional analysis of a locus on multiple traits, overcoming the shortcomings of other methods.
HAPRAP	http://apps.biocompute.org.uk/haprap/	https://pubmed.ncbi.nlm.nih.gov/27591082	Fine-mapping	Python	An empirical iterative method, that enables fine mapping using haplotype information from an individual-level reference panel.
GSR	https://github.com/li-lab-mcgill/GSR	https://pubmed.ncbi.nlm.nih.gov/32817676	Fine-mapping	Python	Detects causal gene sets for complex traits using gene score regression while accounting for gene-to-gene correlations. It can operate either on GWAS summary statistics or gene expression
RSS	https://github.com/stephenslab/rss	https://pubmed.ncbi.nlm.nih.gov/29399241	Fine-mapping	Matlab	It is a generally-applicable framework for multiple-SNP analysis. Uses a “Regression with Summary Statistics” (RSS) likelihood, which relates the multiple regression coefficients to univariate regression results
JAM	https://github.com/pjnewcombe/R2BGLiMS	https://pubmed.ncbi.nlm.nih.gov/27027514	Fine-mapping	R	Bayesian penalized regression that accounts for SNP correlation and finds SNPs that best explain the complete joint pattern of marginal effects
PAINTOR	https://github.com/gkichaev/PAINTOR_V3.0	https://pubmed.ncbi.nlm.nih.gov/27663501/	Fine-mapping	C/C++	Integrates functional genomic data with association strength from potentially multiple populations (or traits) to prioritize variants for follow-up analysis
DAP	https://github.com/xqwen/dap	https://pubmed.ncbi.nlm.nih.gov/27236919/	Fine-mapping	C/C++	Deterministic approximation of posteriors enables highly efficient and accurate joint enrichment analysis and identification of multiple causal variants
fgwas	https://github.com/joepickrell/fgwas	https://pubmed.ncbi.nlm.nih.gov/24702953/	Fine-mapping	C/C++	Integrates functional genomic information into a GWAS
echocolatoR	https://github.com/RajLabMSSM/echolocatoR	https://pubmed.ncbi.nlm.nih.gov/34529038/	Fine-mapping	R	Integrates a diverse suite of statistical and functional fine-mapping tools to identify, test enrichment in, and visualize high-confidence causal consensus variants in any phenotype.
RiVIERA-beta	https://github.com/yueli-compbio/RiVIERA-beta	https://pubmed.ncbi.nlm.nih.gov/27407109/	Fine-mapping	R	Bayesian fine-mapping using Epigenomic Reference Annotation
BEATRICE	https://github.com/sayangsep/Beatrice-Finemapping	https://pubmed.ncbi.nlm.nih.gov/36993396/	Fine-mapping	Python	Combines a hierarchical Bayesian model with a deep learning-based inference procedure
XMAP	https://github.com/YangLabHKUST/XMAP	https://pubmed.ncbi.nlm.nih.gov/37898663	Fine-mapping	R	A variational EM method for cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias
FINMOM	https://vkarhune.github.io/finimom/	https://pubmed.ncbi.nlm.nih.gov/37348543	Fine-mapping	R	A Bayesian method that allows for multiple causal variants using product inverse-moment prior which is a natural prior distribution to model non-null effects in finite samples
CARMA	https://github.com/Iuliana-Ionita-Laza/CARMA	https://pubmed.ncbi.nlm.nih.gov/37169873	Fine-mapping	R	Bayesian model that allows flexible specification of the prior distribution, joint modeling of summary statistics and functional annotations, and accounting for discrepancies between summary statistics and external LD
AHIUT	https://figshare.com/articles/dataset/AHIUT/6615470	https://pubmed.ncbi.nlm.nih.gov/29959179	Fine-mapping	R	An intersection-union test based on a joint/conditional regression model with all the SNPs in a locus to infer AH
BVS-PICA	https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fbiom.12620&file=biom12620-sup-0001-SuppData.zip	https://pubmed.ncbi.nlm.nih.gov/27858978	Fine-mapping	R	Bayesian variable selection for classifying genomic class level associations

Multiple trait analysis

- Genetic Correlation (GC)
- Pleiotropy analysis
- Mendelian Randomization (MR)
- Colocalization
- Transcriptome-wide association studies (TWAS)

Genetic Correlation (GC)

- Genetic correlation is related to pleiotropy and describes the relationship between two traits, that is, the extent to which the genetic variants influencing one trait overlap with the genetic variants associated with the other. It thus can quantify the overall genetic similarity and provide insights into the polygenic genetic architecture of complex traits.
- As we already saw, analyzing simultaneously multiple traits may increase power in case of horizontal pleiotropy; an additional potential application is to use the estimated correlation in order to establish causality between traits in case of vertical pleiotropy (see also next sections).
- Since heritability is the proportion of the phenotypic variance explained by genotypic variation it is of no surprise that genetic correlation (or, the genetic covariance) is related to the traits' heritabilities. Thus, several of the methods for estimating heritability discussed earlier, like HESS and SumHer can also calculate the correlation between traits.

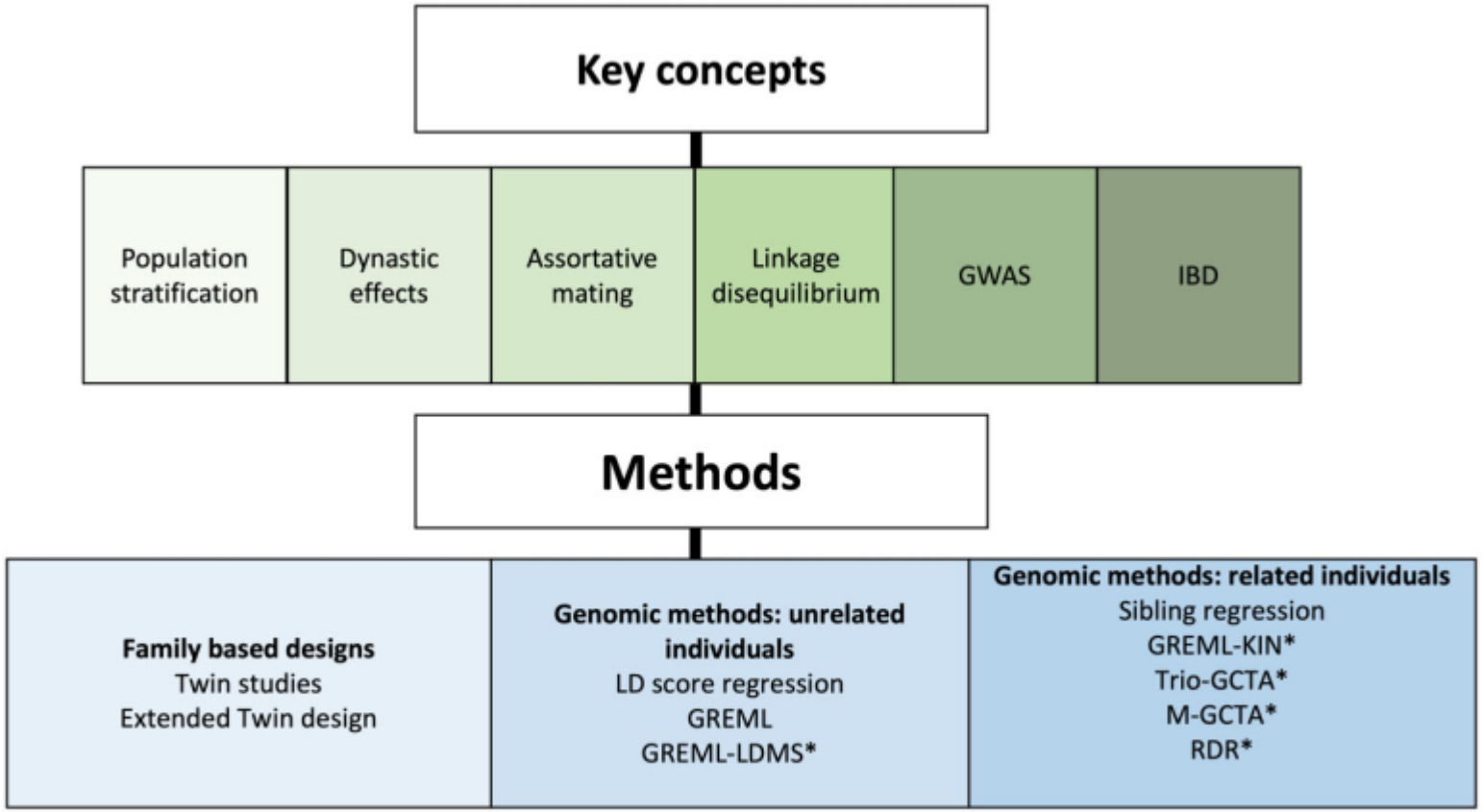


Figure 1 A flowchart detailing the outline of the review. * indicates methods that can be found in the Supplementary Notes (available as Supplementary data at *IJE* online)

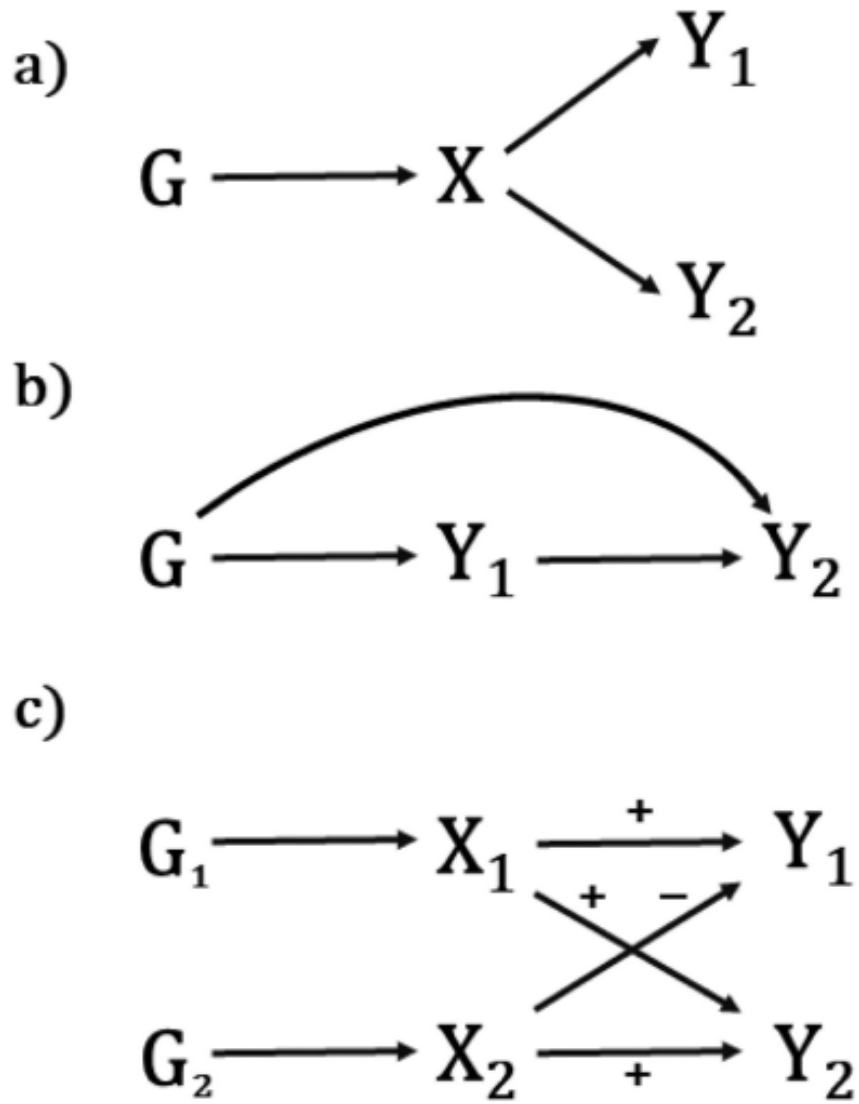


Figure 1:
Potential sources of genetic correlation. **(a)** Traits Y_1 and Y_2 share a common cause X under genetic control. **(b)** Trait Y_1 causes trait Y_2. **(c)** Traits Y_1 and Y_2 share two common causes, X_1 and X_2, one of which has the same directions of effect on both traits, the other of which has opposite directions. In this case, the genome-wide genetic correlation may be close to 0, although when restricted to the loci in G_1 or G_2 the magnitude could be away from 0.

- The most commonly used method for calculating genetic correlation is **LDSC** (LD Score Regression). The method originally developed for distinguishing polygenicity from bias by examining the relationship between test statistics and LD score, but it is also used for estimating heritability and genetic correlation. LDSC is also available through the LD Hub server.
- **PCGC-s** is an adaptation of stratified LDSC for case-control studies and can also estimate genetic heritability, genetic correlation, and functional enrichment.
- Another popular tool is **GNOVA** which calculates annotation-stratified covariance using the method of moments and allows for sample overlap. Its extension, **SUPERGNOVA** identifies global and local genetic correlations that could provide new insights into the shared genetic basis of many phenotypes. Local correlations, among others, can be also computed using **LAVA**.
- **HDL** is a likelihood-based method which produces more precise estimates. A recent comparison found that LDSC and GNOVA are more similar and robust to LD and sample overlap compared to HDL. HDL provides biased estimates of the genetic covariance in most cases and could not distinguish genetic from non-genetic correlation. Moreover, HDL restricts the users to using the built-in reference panel, and its performs poorly when the number of shared SNPs between reference panel and GWAS is small.
- Other tools provide somewhat different types of analyses. For instance **Popcorn** estimates transethnic genetic correlation, **GECKO** estimates both genetic and environmental covariances, **PhenoSD** uses LDSC for estimating phenotypic correlations and then performs correction for multiple testing using the spectral decomposition of matrices, whereas **LPM** is a latent probit model scalable to hundreds of annotations and phenotypes that integrates functional annotations.
- **ccGWAS** is a tool for comparing two different disorders with small genetic correlation providing a case-case association test, and **RHOGE** estimates the genetic correlation between two traits as a function of predicted gene expression effect. **LOGOdetect** uses scan statistics with an LD score-weighted inner product of local z-scores to identify small segments that harbor local genetic correlation between two traits. **DONUTS** is a unique method since it operates on summary statistics from families.

RHOGE	https://github.com/bogdanlab/RHOGE	https://pubmed.ncbi.nlm.nih.gov/28238358/	Genetic correlation	R	Estimates the genetic correlation between two complex traits as a function of predicted gene expression effect
LDSC	https://github.com/bulik/ldsc/	https://pubmed.ncbi.nlm.nih.gov/25642630	Genetic correlation	Python	Distinguishes polygenicity from bias by examining the relationship between test statistics and LD score. Used allso for estimating heritability and genetic correlation
HDL	https://github.com/zhenin/HDL	https://pubmed.ncbi.nlm.nih.gov/32601477/	Genetic correlation	R	A likelihood-based method for estimating genetic correlation. Compared to LD Score regression (LDSC), It reduces the variance of a genetic correlation estimate by about 60%
PCGC-s	https://github.com/omerwe/PCGCs	https://pubmed.ncbi.nlm.nih.gov/29979983	Genetic correlation	Python	It is an adaptation of stratified LD score regression (S-LDSC) for case-control studies. It can estimate genetic heritability, genetic correlation and functional enrichment.
PhenoSpD	https://github.com/MRCIEU/PhenoSpD	https://pubmed.ncbi.nlm.nih.gov/30165448	Genetic correlation	R	Uses LDSC to estimate phenotypic correlations and then performs correction of multiple testing using the spectral decomposition of matrices
GNOVA	https://github.com/xtonyjiang/GNOVA	https://pubmed.ncbi.nlm.nih.gov/29220677	Genetic correlation	Python	A method that calculates annotation-stratified covariance between arbitrary number of traits and enables researchers to dissect both the shared and distinct genetic architecture across traits
LPM	https://github.com/mingjingsi/LPM	https://pubmed.ncbi.nlm.nih.gov/31860024	Genetic correlation	R	A latent probit model that can integrate functional annotations. It is scalable to hundreds of annotations and phenotypes
Popcorn	https://github.com/brielin/Popcorn	https://pubmed.ncbi.nlm.nih.gov/27321947/	Genetic correlation	Python	Method for estimating the transethnic genetic correlation: the correlation of causal-variant effect sizes at SNPs common in populations
LOGOdetect	https://github.com/ghm17/LOGODetect	https://pubmed.ncbi.nlm.nih.gov/33795679/	Genetic correlation	R	A tool to identify small segments that harbor local genetic correlation between two traits
cc-GWAS	https://github.com/wouterpeyrot/CCGWAS	https://pubmed.ncbi.nlm.nih.gov/33686288	Genetic correlation	R	Tool for case-case association testing of two different disorders
DONUTS	https://github.com/qlu-lab/DONUTS	https://pubmed.ncbi.nlm.nih.gov/34131076	Genetic correlation	R	A statistical framework that can estimate direct and indirect genetic effects at the SNP level and calculate genetic correlation between traits
SUPERGNOVA	https://github.com/qlu-lab/SUPERGNOVA	https://pubmed.ncbi.nlm.nih.gov/34493297	Genetic correlation	Python	Extension of GNOVA to identify global and local genetic correlations that could provide new insights into the shared genetic basis of many phenotypes
GECKO	https://github.com/borangao/GECKO	https://pubmed.ncbi.nlm.nih.gov/33395406	Genetic correlation	R	Method based on composite likelihood for estimating genetic and environmental covariances
LAVA	https://ctg.cncr.nl/software/lava	https://pubmed.ncbi.nlm.nih.gov/35288712/	Genetic correlation	R	An integrated framework for local genetic correlation analysis that can also evaluate local heritabilities and analyze conditional genetic relations between several phenotypes using partial correlation and multiple regression

Pleiotropy analysis

- Pleiotropy is the phenomenon in which a single variant influences several traits. Such methods are of great importance in genetic research and several methods have been developed during the last years.
- A major goal of such methods is to increase the statistical power over single trait methods. Imagine for instance a variant that produces a near-significant effect when analyzed separately for two or three traits. A method that can combine these estimates may produce significant results.
- Another application of a joint analysis would be to identify variants that influence both traits, or variants that influence only one of them. When all the relevant variants are considered, one can also estimate the kind of relationship between the traits (see “genetic correlation”).
- Usually, the methods that allow for multiple trait analysis are oriented toward quantitative traits like BMI, SBP, DBP and so on, that traditionally are measured on a single cohort, resulting in the existence of cross-trait correlation that needs to be taken into account in the analysis.
- However, there are also methods for performing the same analysis with summary estimates derived from different cohorts, as well as methods that allow for binary traits with the case-control design, using overlapped or non-overlapped controls.

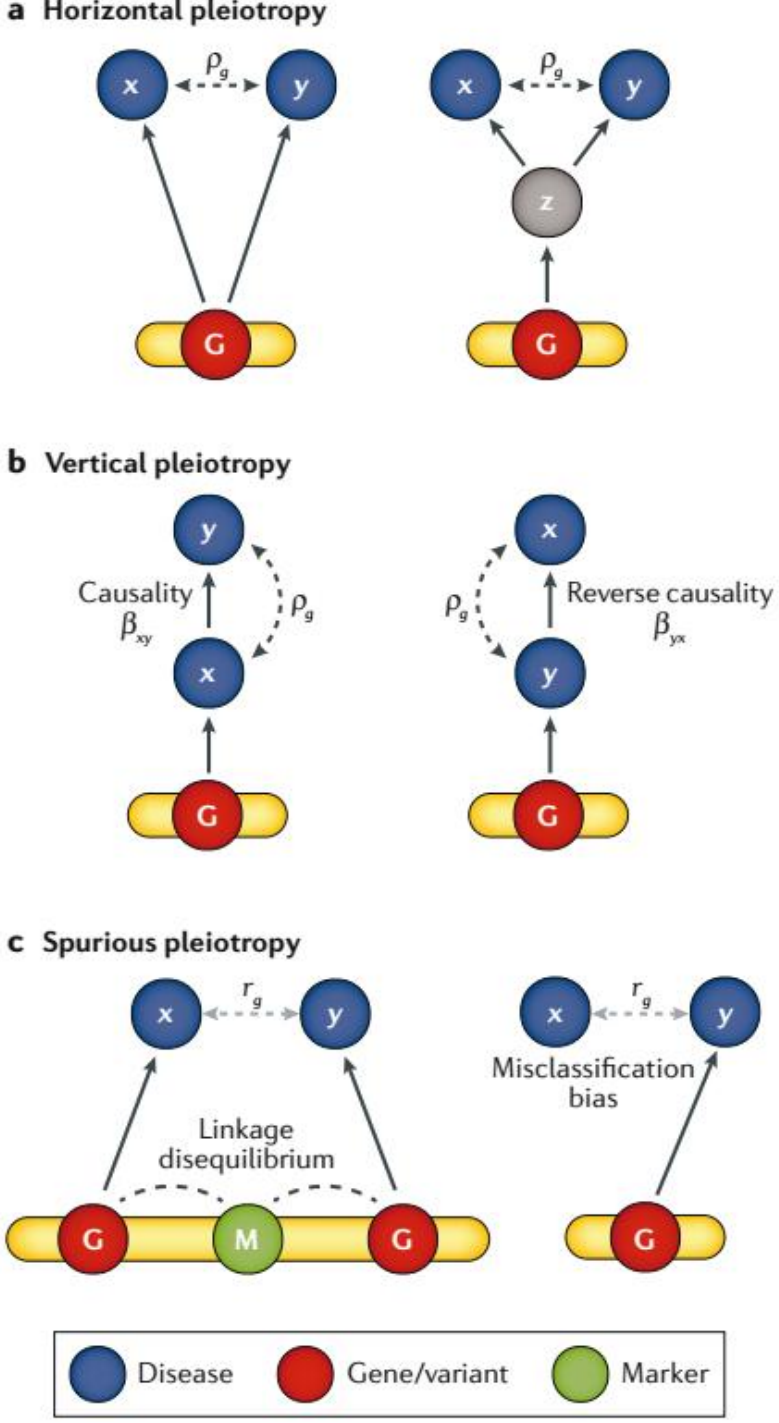


Fig. 1 | **Different mechanisms of pleiotropy between two diseases. a** | In horizontal pleiotropy, the genetic variant (G) contributes directly to risk of both diseases (*x* and *y*) or indirectly through an intermediate (endo)phenotype (*z*). **b** | In vertical pleiotropy, there is a causal relationship between disease *x* and *y* where disease *x* itself leads to an increased risk of disease *y* (left); in some circumstances, reverse causation may be observed (here we assume that there is a natural order in the traits to expect causation from trait *x* to trait *y*, with the reverse causation from *y* to *x* less expected) (right). In some circumstances, there may be causality, reverse causality and genetic correlation. **c** | Spurious pleiotropy at a locus can occur when a measured genetic marker (M) 'tags', via linkage disequilibrium, two distinct causal variants (left); however, this sort of spurious pleiotropy has to be consistent across loci to result in a non-zero estimate of genome-wide genetic correlation. Different sources of bias (such as disease misclassification) may also lead to the incorrect assumption of pleiotropy between disease *x* and *y* (right). In early work on pleiotropy, the term 'spurious pleiotropy' was used for what we term 'vertical pleiotropy'. β_{xy} is the expected change in trait y caused by each unit increase in trait x; β_{yx} is the expected change in trait x caused by each unit increase in trait y. ρ_g, genetic correlation; r_g, estimated genetic correlation.

van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet. 2019;20(10):567-81.

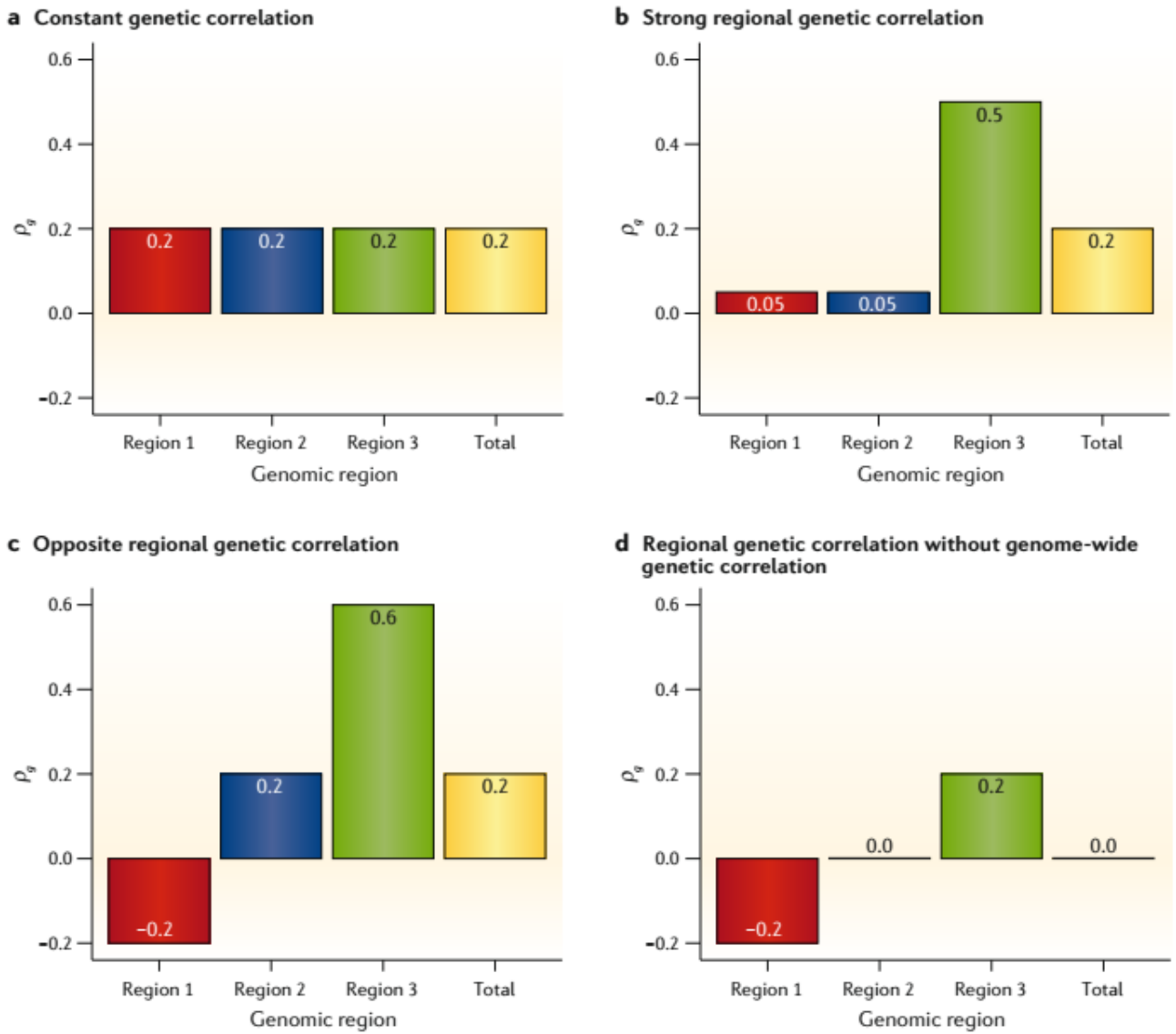


Fig. 2 | **Genome-wide genetic correlation versus regional genetic correlation.** The overall genetic correlation (as estimated from pedigree or genome-wide association study data) between two traits has been fixed at 0.2, but the underlying regional architecture of the genetic correlation can vary widely[11]. In part **a**, the genetic correlation is constant across the genome. Alternative scenarios include strong regional genetic correlation (part **b**) and a combination of both positive and negative regional correlations (part **c**); in both of these cases, the regional genetic correlation can far exceed the overall genome-wide genetic correlation. In part **d**, both positive and negative regional correlations occur in the absence of an overall genetic correlation. Regions can be interpreted as physical genomic loci, as allele-frequency bins or as functionally annotated categories (such as coding versus non-coding, biological pathways or tissue-specific expression). ρ_g, genetic correlation.

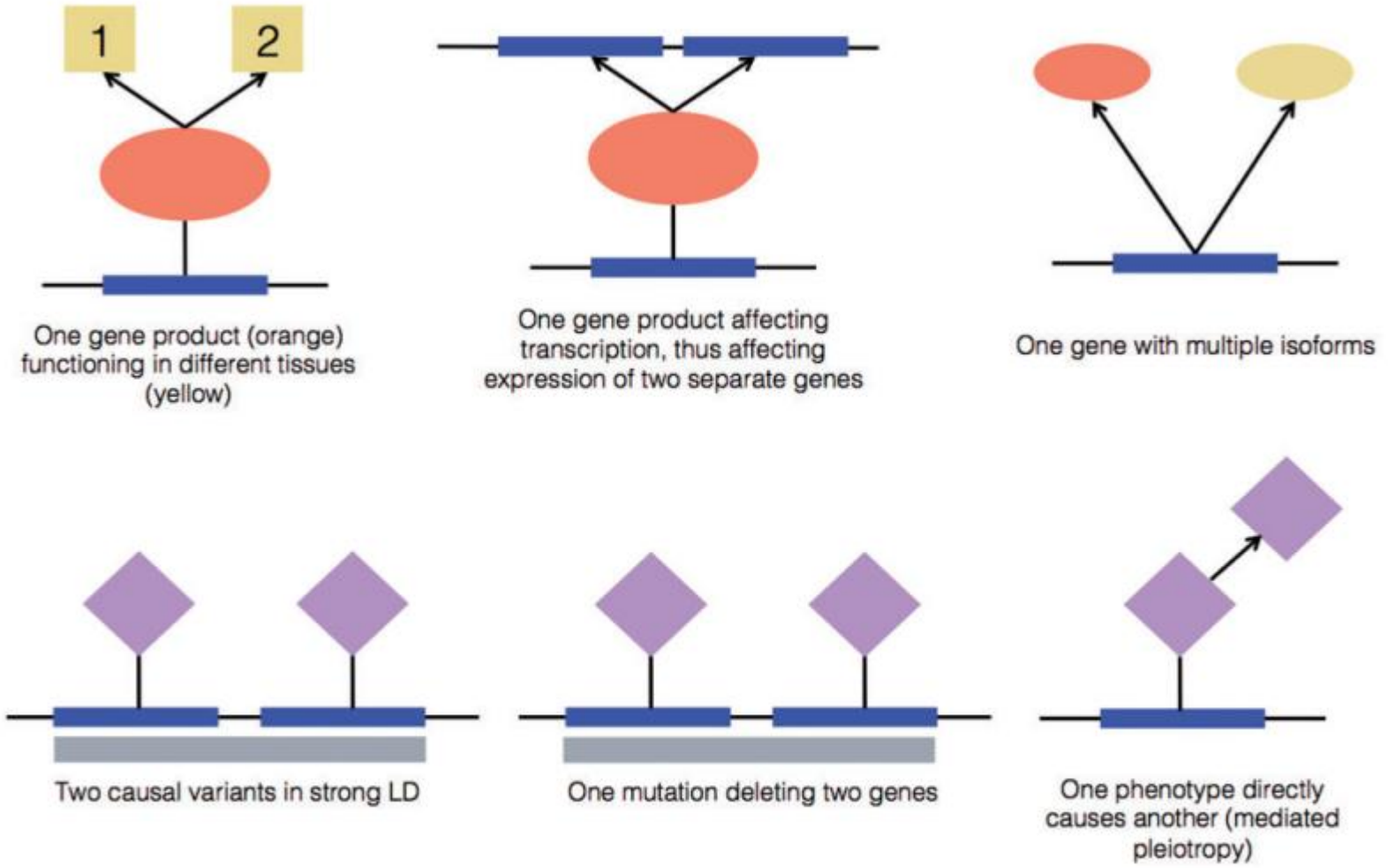


Figure 1. Example mechanisms of cross-phenotype associations and pleiotropy. Both top and bottom panels illustrate examples of cross-phenotype associations in which a region of the genome is associated with multiple traits regardless of the underlying mechanism. Top: This figure illustrates a subset of the many molecular mechanisms by which a single gene can influence multiple traits. These mechanisms are often considered 'true' pleiotropy. Bottom: Pleiotropy is often considered 'spurious' when the multiple affected traits cannot be connected directly to a single gene. For example, two genes may be strongly linked, resulting in the observation of a single genetic locus influencing multiple phenotypes. Additionally, a single mutation may affect more than one gene. Not shown are non-genic loci that may influence multiple phenotypes through regulation of multiple genes.

Tyler AL, Crawford DC, Pendergrass SA. The detection and characterization of pleiotropy: discovery, progress, and promise. Brief Bioinform. 2016;17(1):13-22.

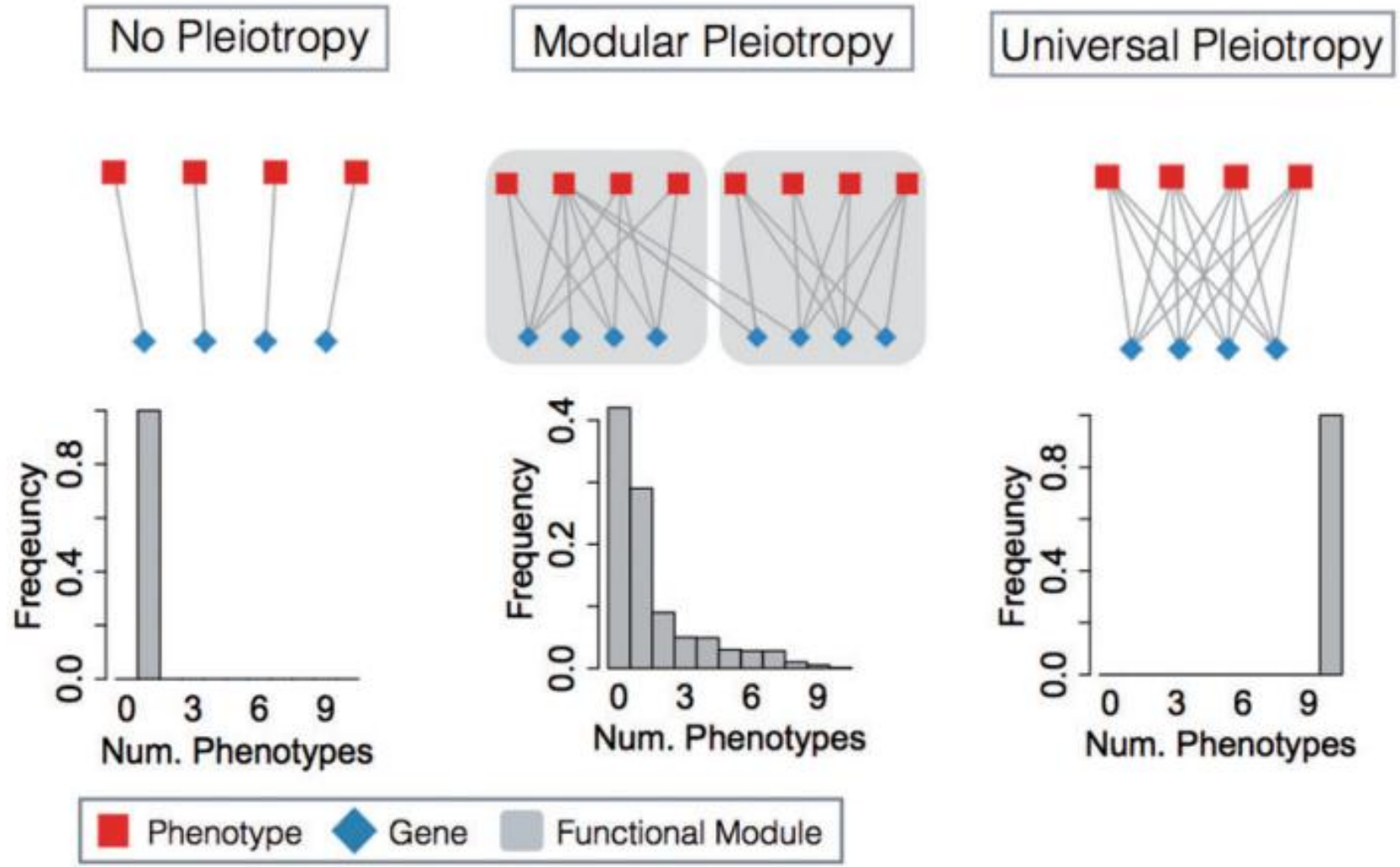


Figure 2. The genetic architecture of pleiotropy. Top: depiction of gene (diamond) to phenotype (square) effects if genes affect only one, all or a subset of phenotypes. In the case of modular pleiotropy, genes influence a subset of traits and the genes and traits cluster into functional modules (rounded corner boxes). Bottom: Distribution of counts of phenotypic effects. If no pleiotropy exists (left), all genes affect one phenotype. If pleiotropy is universal (right), all genes affect all phenotypes. If pleiotropy is modular (middle), we expect to see a continuous distribution of the number of genetic effects. Multiple experiments have demonstrated that most genes have zero or one phenotypic effect, while few have many phenotypic effects. This is consistent with a modular view of pleiotropy.

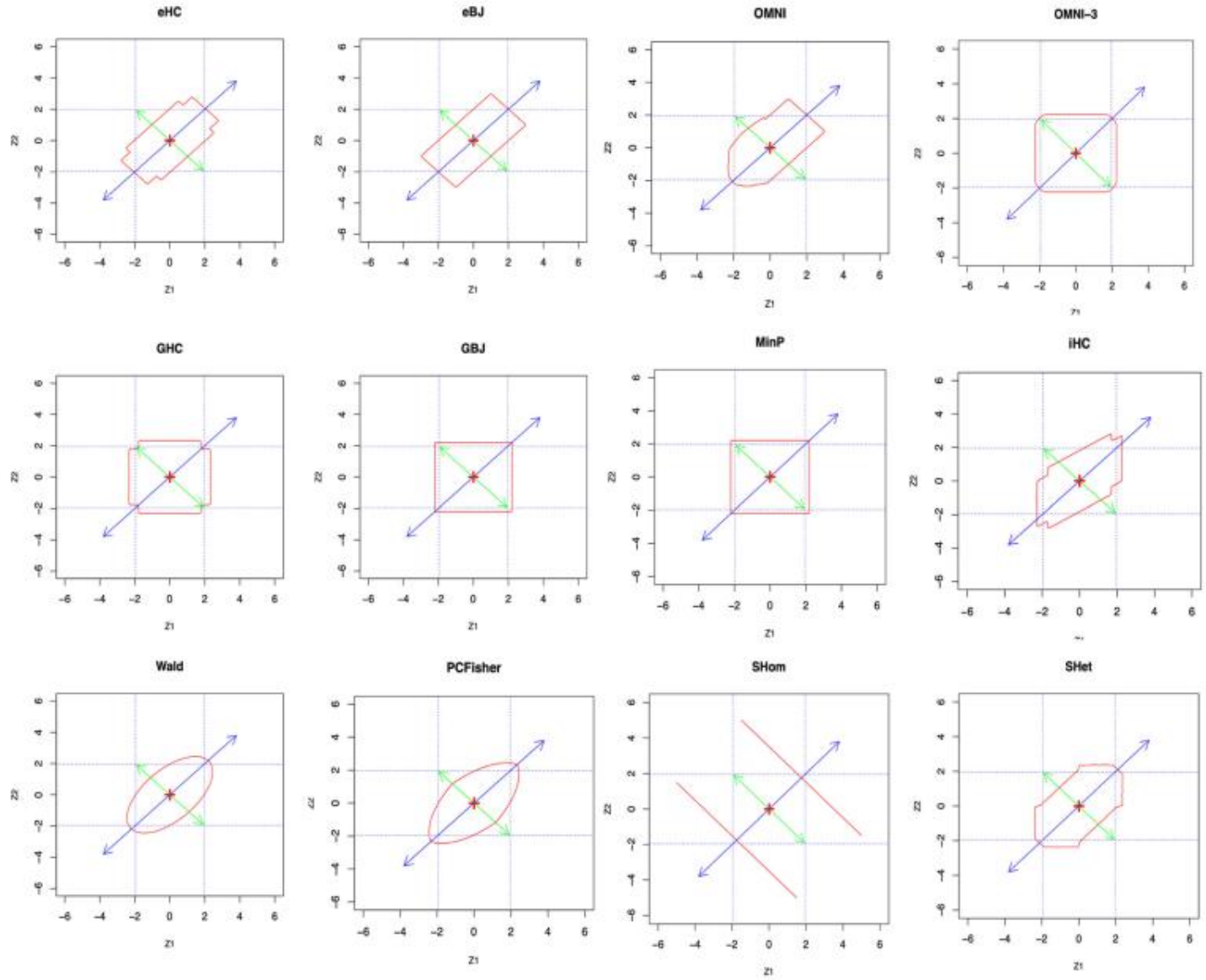


FIGURE 1 The rejection boundaries (solid lines without arrows or curves) of the OMNI, eHC, eBJ, iHC, GHC, GBJ, MinP, Wald, PCFisher, OMNI-3, SHom, and SHet tests at the significance level $\alpha = 0.05$ for a bivariate normal $\boldsymbol{Z} = (Z_1, Z_2)^T$ test statistics with correlation coefficient $\rho = 0.6$ under H_0. The longer solid lines with arrows represent the direction where $\boldsymbol{Z}$ has the largest variation and the shorter solid lines with arrows represent the direction where $\boldsymbol{Z}$ has the secondlargest variation under the null. The dotted lines mark the univariate critical values at ± 1.96 for Z_1 and Z_2, respectively. eBJ, eigen Berk–Jones; eHC, eigen higher criticism; GBJ, generalized Berk–Jones; GHC, generalized higher criticism; iHC, innovated higher criticism; MinP, minimum of the p values; OMNI, omnibus test

- All methods base their inference on the assumption that the z-statistics follow a multivariate normal distribution (MVN) and perform different types of tests and/or different procedures to estimate or approximate the correlation structure.
- **ACA** one of the first methods, estimates the traits covariance from a subset of the phenotypic data or from published studies, **p_ACT** integrates the MVN using the trait correlation, **PAT** uses a likelihood-ratio test, and **PLEI** uses the union-intersection testing method, but in addition to the likelihood ratio test, it also applies generalized estimating equations under the working independence model; it can be applied for both marginal analysis and conditional analysis. **USAT** uses a score-based test, **JaSPU** uses an adaptive test which is robust to violations of the MVN assumptions and **MTAR** uses a Principal Components (PC)-based test.
- **BMASS** on the other hand is a Bayesian multivariate method, whereas **TWT**, **MTAFS** and **EBMMT**, which are among the newer tools, perform a Cauchy Combined Test (CCT) to handle the correlation structure and obtain accurate p-values. **SHAHER** uses a linear combination of traits by maximizing the proportion of its genetic variance explained by the shared variants and allows both shared and unshared variants to be effectively analyzed and **HIPO** performs heritability-informed power optimization for conducting multi-trait association analysis. **HOPS** computes a horizontal pleiotropy score by removing correlations between traits caused by vertical pleiotropy and normalizing effect sizes across all traits and **PDR** performs a pleiotropic decomposition regression to identify shared components and their underlying genetic variants.
- We also need to mention methods like **MTAG** and **PLEIO** which use LDSC and apart from sample overlap also allow data from multiple studies, something that can be considered meta-analysis and methods like **MSKAT**, **multiSKAT**, **MGAS**, **MAIUP** and **MTAR** (multi-trait analysis of rare variants) which are gene-based methods specialized for multiple traits. Finally, methods like **iMAP** and **graphGPA2** use graphical models and are capable of performing analysis of large number of traits.

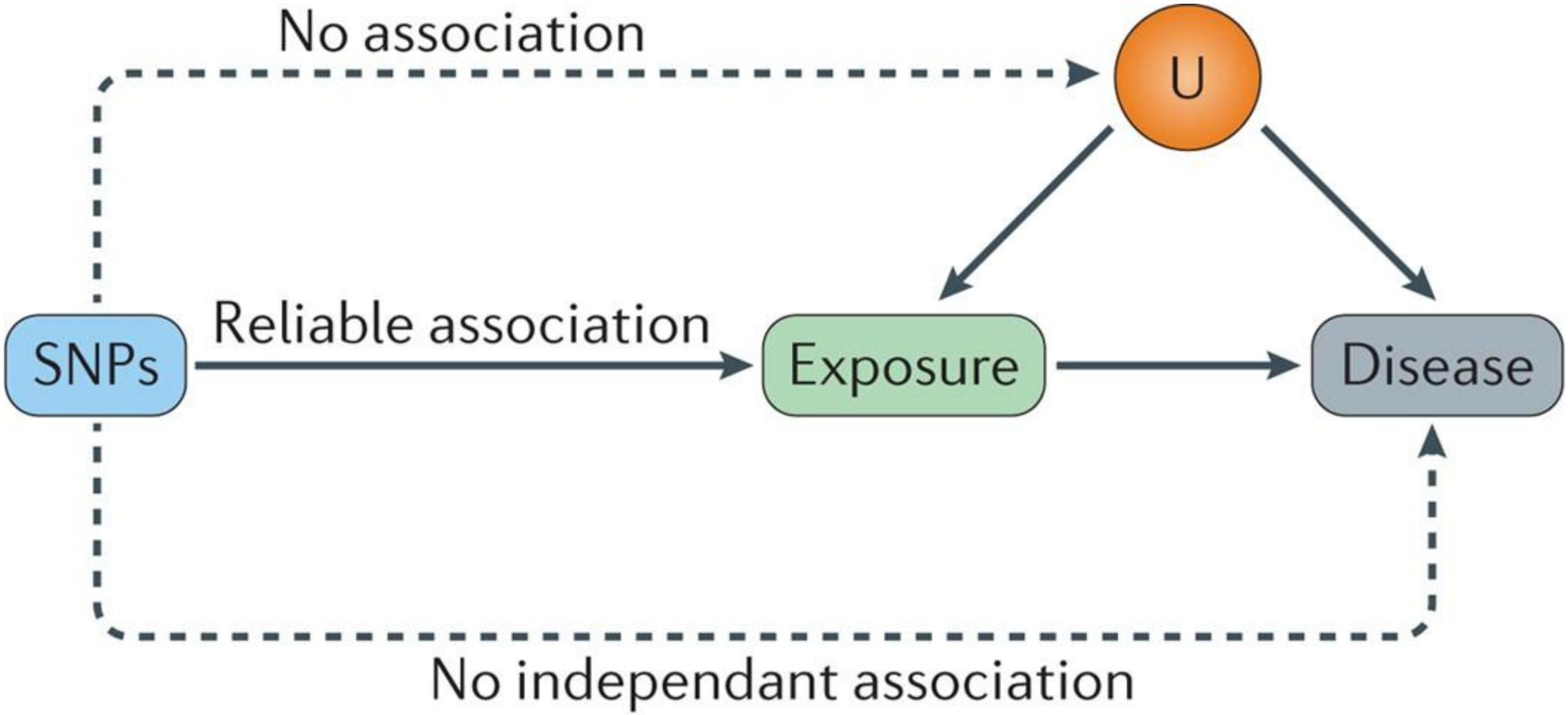
JaSPU	https://github.com/kaskarn/JaSPU	https://pubmed.ncbi.nlm.nih.gov/33949650	Pleiotropy (overlapped)	Julia	Evaluates the effect of SNPs across k traits using z-scores from previous regression analyses. It performs simulations to produce p-values, using the empirical multivariate-normal distribution of null z-scores.
HIPO	https://github.com/gqi/hipo	https://pubmed.ncbi.nlm.nih.gov/30289880/	Pleiotropy (overlapped)	R	Performs heritability informed power optimization for conducting multi-trait association analysis
MTAR	https://github.com/baolinwu/MTAR	https://pubmed.ncbi.nlm.nih.gov/30476000	Pleiotropy (overlapped)	R	Uses principal component (PC)-based association test which has optimal power when the underlying multi-trait signal can be captured by the first PC. Performs an adaptive test by optimally weighting the PC-based test and the omnibus chi-square test to achieve robust performance
PAT	https://github.com/koditaraszka/pat	https://pubmed.ncbi.nlm.nih.gov/36342933	Pleiotropy (overlapped)	Python	The pleiotropic association test (PAT) is used for joint analysis of multiple traits. Uses the decomposition of phenotypic covariation into genetic and environmental components to create a likelihood ratio test statistic for each genetic variant.
TATES	https://ctg.cncr.nl/software/	https://pubmed.ncbi.nlm.nih.gov/23359524	Pleiotropy (overlapped)	FORTRAN	Trait-based Association Test that uses Extended Simes procedure combines the p-values obtained in standard univariate GWAS to acquire one trait-based p-value, while correcting for correlations between components
CONFIT	https://github.com/lgai/CONFIT	https://pubmed.ncbi.nlm.nih.gov/29949991	Pleiotropy (overlapped)	Python	The method estimates the degree of shared effects between traits from the data. The test statistic is a sum of the relative likelihoods for each alternate configuration.
cFDR	https://github.com/jamesliley/cFDR-common-controls	https://pubmed.ncbi.nlm.nih.gov/25658688	Pleiotropy (overlapped)	R	Calculates an upper bound on the expected false discovery rate (FDR) across a set of SNPs whose p-values for two diseases are both less than two disease-specific threshold
USAT	https://github.com/RayDebashree/USAT	https://pubmed.ncbi.nlm.nih.gov/26638693	Pleiotropy (overlapped)	R	Uses a data-adaptive weighted score-based test statistic for testing association of multiple continuous phenotypes with a single SNP
bmass	https://github.com/mturchin20/bmass	https://pubmed.ncbi.nlm.nih.gov/31596850	Pleiotropy (overlapped)	R	Bayesian multivariate analysis of GWAS data using univariate association summary statistics
ACA	https://sites.bu.edu/fhspl/publications/approximate-conditional-analysis/	https://pubmed.ncbi.nlm.nih.gov/33510268	Pleiotropy (overlapped)	R	Relies on an approximate conditional phenotype analysis. The traits covariance may be estimated either from a subset of the phenotypic data; or from published studies.
TWT	https://github.com/bschilder/ThreeWayTest	https://pubmed.ncbi.nlm.nih.gov/37027223	Pleiotropy (overlapped)	R	Uses the correlation coefficients between Wald statistics obtained from linear regression with covariates. Then, a test is applied by integrating three-level information including the intrinsic genetic structure, pleiotropy, and the potential information combinations
EBMMT	https://github.com/Vivian-Liu-Wei64/EBMMT	https://pubmed.ncbi.nlm.nih.gov/35192735	Pleiotropy (overlapped)	R	Uses the eigen higher criticism and the eigen Berk-Jones testing procedures to test the association between SNPs and multiple correlated traits. Then uses the aggregated Cauchy association test .
Plei	https://github.com/yangq001/Plei	https://pubmed.ncbi.nlm.nih.gov/28971959	Pleiotropy (overlapped)	R	A procedure that can be applied for both marginal analysis and conditional analysis. Uses the union-intersection testing methods, but in addition to the likelihood ratio test, it also applies generalized estimating equations under the working independence model
p_ACT	http://csg.sph.umich.edu/boehnke/p_act.php	https://pubmed.ncbi.nlm.nih.gov/17966093/	Pleiotropy (overlapped)	R	A method of computing P values adjusted for correlated tests that attains the accuracy of permutation or simulation-based tests in much less computation time
SHAHER	https://github.com/Sodbo/shared_heredity	https://pubmed.ncbi.nlm.nih.gov/36292579	Pleiotropy (overlapped)	R	It is based on the construction of a linear combination of traits by maximizing the proportion of its genetic variance explained by the shared genetic factors.
MTAG	https://github.com/omeed-maghzian/mtag	https://pubmed.ncbi.nlm.nih.gov/29292387	Pleiotropy (overlapped)	Python	A method for joint analysis of GWAS of different traits, using a weigthed sum and LDSC
PLEIO	https://github.com/cuelee/pleio	https://pubmed.ncbi.nlm.nih.gov/33352115	Pleiotropy (overlapped)	Python	A framework to map and interpret pleiotropic loci in a joint analysis of multiple diseases and complex traits. It maximizes power by systematically accounting for genetic correlations and heritabilities of the traits using LDSC.
MSKAT	https://github.com/baolinwu/MSKAT	https://pubmed.ncbi.nlm.nih.gov/30239606	Pleiotropy (overlapped)	R	Various types of multi-trait SNP-set association tests (variance component test, burden test and adaptive test), and efficient numerical calculation of P-values
multiSKAT	https://github.com/diptavo/MultiSKAT	https://pubmed.ncbi.nlm.nih.gov/30298564/	Pleiotropy (overlapped)	R	A general framework for testing pleiotropic effects of rare variants on multiple continuous phenotypes using multivariate kernel regression. Many existing tests are equivalent to specific choices of parameters within this framework
MTAR	https://cran.r-project.org/web/packages/MTAR	https://pubmed.ncbi.nlm.nih.gov/32503972	Pleiotropy (overlapped)	R	Joint analysis of association summary statistics between multiple rare variants and different traits. Leverages the genome-wide genetic correlation to inform the degree of gene-level effect heterogeneity across traits.
MGAS	http://pmglab.top/kgg/	https://pubmed.ncbi.nlm.nih.gov/25431328/	Pleiotropy (overlapped)	Java	A multivariate gene-based association test by extended Simes procedure (MGAS), that allows gene-based testing of multivariate phenotypes
iMAP	https://github.com/biostatpzeng/iMAP	https://pubmed.ncbi.nlm.nih.gov/29635306	Pleiotropy (overlapped)	R	Performs integrative mapping of pleiotropic association and functional annotations using penalized Gaussian mixture models. Uses a multinomial logistic regression model
graphGPA2	https://dongjunchung.github.io/GGPA2/	https://pubmed.ncbi.nlm.nih.gov/37501720	Pleiotropy (overlapped)	R	Bayesian graphical model which allows to integrate functional annotations with GWAS datasets for multiple phenotypes within a unified framework
MTAFS	https://github.com/Qiaolan/MTAFS	https://pubmed.ncbi.nlm.nih.gov/37237036	Pleiotropy (overlapped)	R	An efficient and robust adaptive method for multi-trait analysis of GWAS.
PDR	https://github.com/jballard28/PDR	https://pubmed.ncbi.nlm.nih.gov/35477001	Pleiotropy (overlapped)	Matlab	Pleiotropic decomposition regression using method of moments to identify shared components and their underlying genetic variants
PLACO	https://github.com/RayDebashree/PLACO	https://pubmed.ncbi.nlm.nih.gov/33290408	Pleiotropy (overlapped)	R	Implements a variant-level formal statistical test of pleiotropy of two traits inspired from mediation analysis.
HOPS	https://github.com/rondolab/HOPS	https://pubmed.ncbi.nlm.nih.gov/31653226	Pleiotropy (overlapped)	R	Allows to compute the horizontal pleiotropy score by removing correlations between traits caused by vertical pleiotropy and normalizing effect sizes across all traits
					Test constructed based on the traditional intersection–union test with two sets of independent P-values as input and follows

- On the other hand, there are several methods that assume independence of the studied samples.
- Most of them are designed for larger analyses of many traits from multiple studies, for instance **PolarMorphism**, **JASS**, **gwas-pw** and **FactorGo**, **sumDAG**, **combGWAS** and **GCPBayes pipeline**.
  - GCPBayes_pipeline uses the functionality of GCPBayes to perform cross-phenotype gene-set analysis between two traits.
  - gwas-pw is used for the joint analysis of two GWAS in order to identify variants influencing both traits. PolarMorphism is based on a transform from Cartesian to polar coordinates and reports a per variant degree of 'sharedness' across traits, whereas
  - FactorGo provides scalable variational factor analysis model that is computationally efficient for large number of traits.
  - JASS provides interactive exploration and visualization of the results of comparison of many traits through a web interface,
  - sumDAG goes one step further and constructs phenotype networks by using a Gaussian linear model and a directed acyclic graph, and
  - combGWAS identifies susceptibility variants for comorbid disorders and calculate genetic correlations.
- **EPS** and **GPA** differ in integrating Pleiotropy and functional annotation from eQTL.

GPA	https://github.com/dongjunchung/GPA	https://pubmed.ncbi.nlm.nih.gov/2539367	Pleiotropy (independent)	R	Uses the EM algorithm to integrate pleiotropy and functional annotation (eQTL etc)
EPS	https://github.com/gordonliu810822/EPS	https://pubmed.ncbi.nlm.nih.gov/27153687/	Pleiotropy (independent)	Matlab	An Empirical Bayes approach to integrating Pleiotropy and Tissue-Specific information (EPS) for prioritizing risk genes
PolarMorphism	https://github.com/UMCUGenetics/PolarMorphism	https://pubmed.ncbi.nlm.nih.gov/35758773	Pleiotropy (independent)	R	The method is based on a transform from Cartesian to polar coordinates. Analyzes multiple related phenotypes and reports (per SNP) the degree of 'sharedness' across them, its overall effect size, as well as p-values
FactorGo	https://github.com/mancusolab/FactorGo	https://pubmed.ncbi.nlm.nih.gov/37879338	Pleiotropy (independent)	Python	A scalable variational factor analysis model used to identify and characterize pleiotropic components. Works well in capturing latent pleiotropic factors across phenotypes while at the same time being computationally efficient
gwas-pw	https://github.com/joepickrell/gwas-pw	https://pubmed.ncbi.nlm.nih.gov/27182965/	Pleiotropy (independent)	R	A tool for jointly analysing two GWAS to identify loci that influence both traits. Instead of using two P-value thresholds to identify variants that influence both traits, the algorithm learns reasonable thresholds from the data
UNITY	https://github.com/bogdanlab/UNITY	https://pubmed.ncbi.nlm.nih.gov/29949958	Pleiotropy (independent)	Python	A Bayesian framework for estimating the proportion of causal variants shared between a pair of complex traits
sumDAG	https://github.com/chunlinli/sumdag	https://pubmed.ncbi.nlm.nih.gov/38045347	Pleiotropy (independent)	R	Constructs a phenotype network by assuming a Gaussian linear structure model embedding a directed acyclic graph
GCPBayes	https://github.com/CESP-ExpHer/GCPBayes-Pipeline	https://pubmed.ncbi.nlm.nih.gov/37416786	Pleiotropy (independent)	R	Pipeline to perform cross-phenotype gene-set analysis between two traits using GCPBayes
combGWAS	https://github.com/LiangyingYin/CombGWAS	https://pubmed.ncbi.nlm.nih.gov/34050728	Pleiotropy (independent)	R	A statistical framework to uncover susceptibility variants for comorbid disorders and calculate genetic correlations
JASS	https://gitlab.pasteur.fr/statistical-genetics/jass	https://pubmed.ncbi.nlm.nih.gov/32002517	Pleiotropy (independent)	web/Python	Incorporates various joint tests such as the omnibus approach and weighted sum of Z-score tests while offering data cleaning and harmonization, fast derivation of joint statistics, and optimized data management process

Mendelian Randomization

- Mendelian Randomization (MR) is a method suggested in the pre-GWAS era to investigate causal relationships between two traits, usually a phenotype and a disease using genotype–trait associations to make inferences about environmentally modifiable causes of the traits. In technical terms, MR uses genetic variants as instrumental variables to mimic the random assignment of exposures in a randomized controlled trial, similar to the way Mendel's laws of inheritance dictate the random assortment of alleles during gamete formation. By utilizing the natural randomization of genetic inheritance, MR aims to minimize biases introduced by confounding factors that usually affect observational studies when investigating the association of two traits.
- Usually, we are interested in a disease and some other intermediate phenotype, or another disease. For instance, the MR approach may involve the relationship between hypertension and BMI, or between hypertension and diabetes.
- Traditionally MR was performed with one sample (1SMR) using a single variant (usually referred to IPD methods), and subsequently multivariate methods for MR meta-analysis were developed. With the emergence of GWAS these methods evolved to the most commonly used two-sample MR (2SMR) methods that utilize summary data estimates from several variants regarding the genotype-phenotype and genotype-disease association from different samples. To establish connection with the previous sections, MR seeks to analyze correlated traits and to provide evidence for causation, in other words to distinguish vertical from horizontal pleiotropy.



Nature Reviews | Cardiology

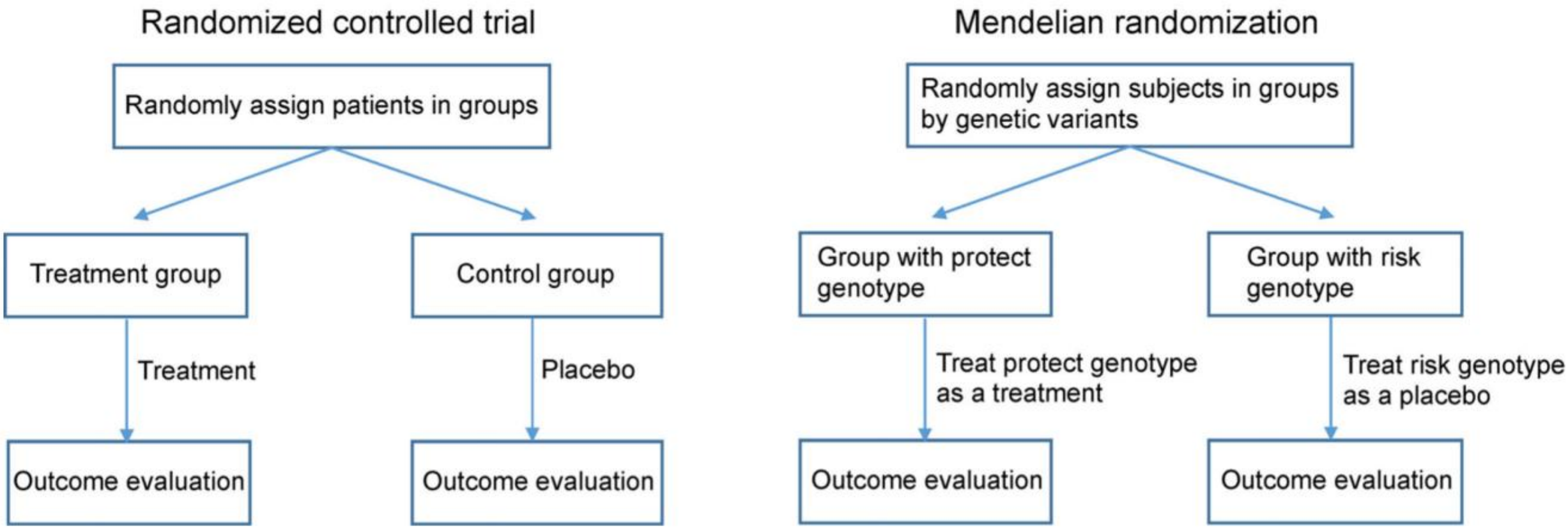
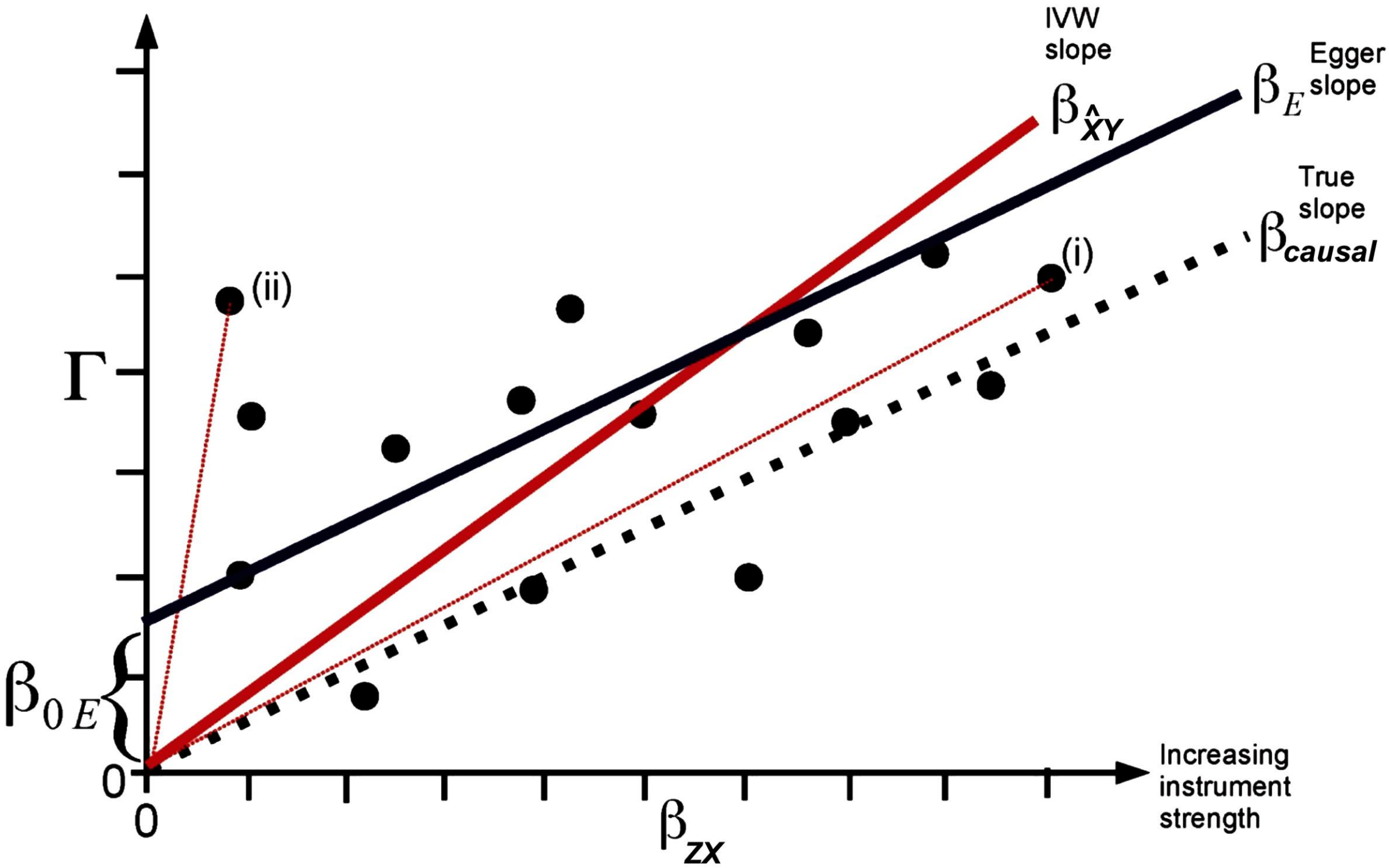


Figure 1. Comparison between RCT and MR designs.


IVW slope
$\beta_{\hat{X}Y}$
β_E Egger slope
True slope
β_{causal}
(i)
(ii)
Γ
β_{0E}
0
0
β_{ZX}
Increasing instrument strength

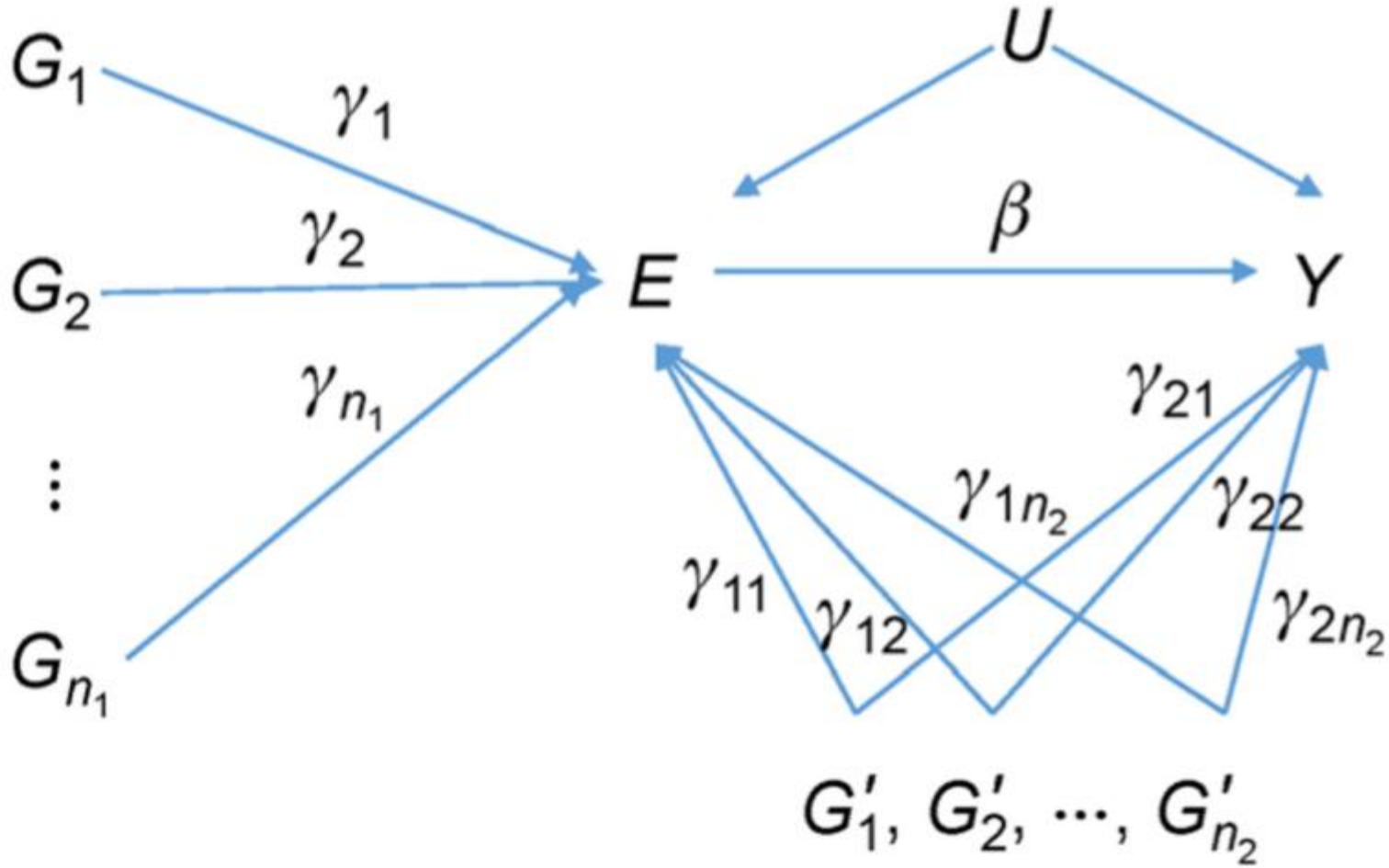


Figure 2. A causal path diagram for multiple instrumental variables.

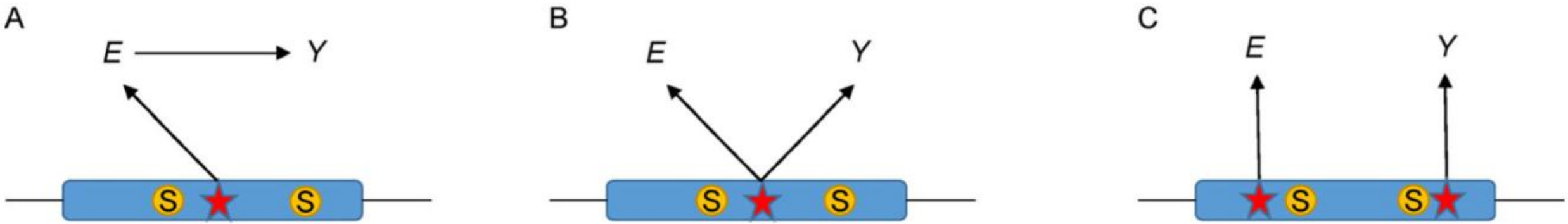


Figure 3. The relationships of genetic variants, exposure and outcome.
(A) Mediation: the causal variant lies on the causal path to *Y*. (B) Horizontal pleiotropy: the causal variant affects both *E* and *Y*. (C) Colocalization: two different causal variants at one locus affect *E* and *Y*. The red star represents the causal variant and S represents genetic markers.

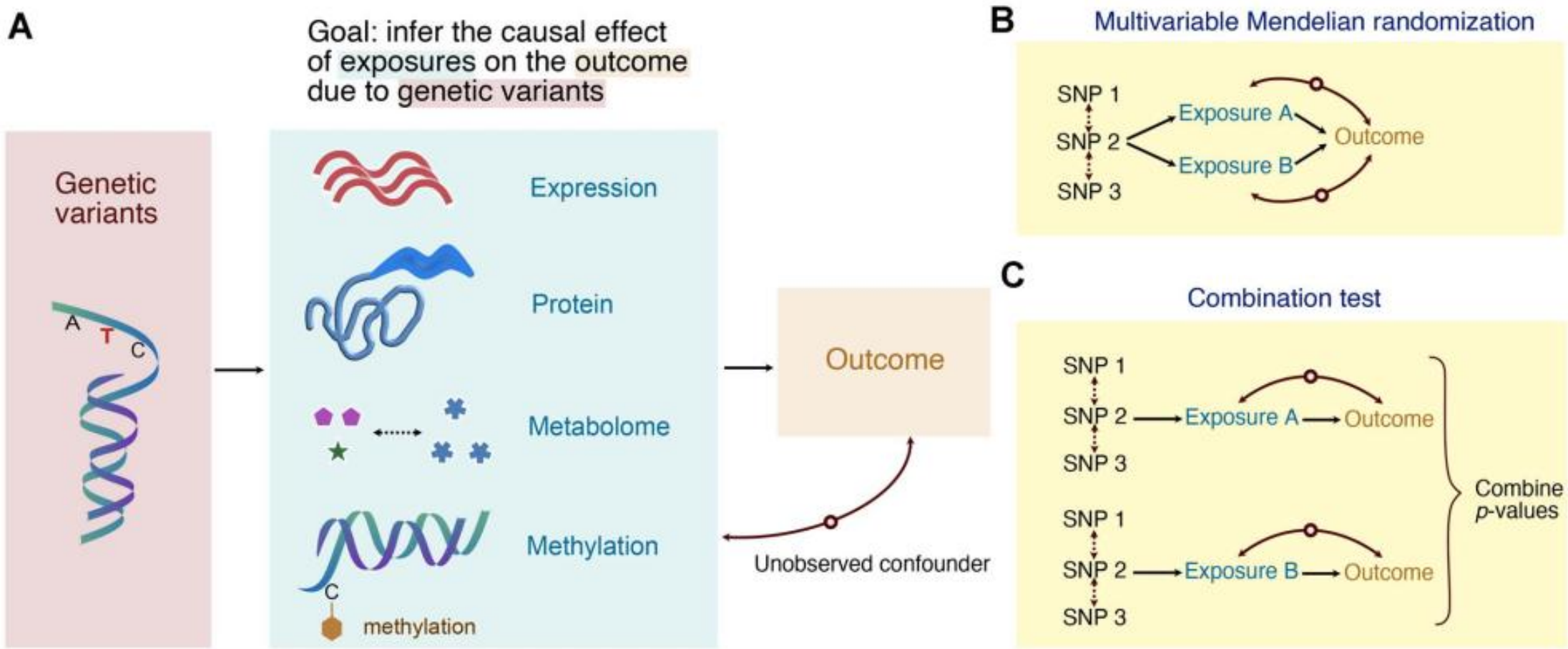


Figure 1. (**A**) A schematic diagram illustrating MR analyses adapted to a multi-omics setting through analyzing QTLs and GWAS summary data to infer the causal effect (or pleiotropic effect) between the exposures and the outcome. (**B**) Multivariable MR is a strategy to interrogate the effect of multiple exposures on the outcome. (**C**) We propose combination tests that combine P-values from MR experiments based on multiple exposures.

- Several standard methods for MR in GWAS with summary data have been made available during the last years: the inverse-variance weighted method (IVW), the various types of median estimators (simple or weighted) and the MR-Egger regression approach. IVW gives consistent estimates only if all the genetic variants in the analysis are valid instruments. The median estimator is consistent even when up to 50% of the information comes from invalid instrumental variables, whereas MR-Egger performs equally well but provides somewhat less precise estimates. These methods are readily available in standard packages like **TwoSampleMR** and **MR**. The functionalities of TwoSampleMR are also offered, at least partially, through the webserver of **MRBASE**, which is the only method available as such.
- **BWMR** is a tool that performs MR in a Bayesian framework. Besides the issue of weak instruments which is of importance, most modern methods also aim to perform the MR analysis accounting or correcting for horizontal pleiotropy. For instance, **pIVW** is an extension of the IVW that accounts simultaneously for weak instruments and balanced horizontal pleiotropy and **MRmix** uses a mixture approach allowing a fraction of the instruments to have pleiotropic effect on the outcome. Similarly, **MRcML**, **MR-LDP**, **MR-Corr2** and **MR-PRESSO** provide functionalities to account for horizontal pleiotropy, whereas **IMRP** takes a different approach and searches iteratively for horizontally pleiotropic variants and causal effects.
- **MR-APSS** differs in that it performs MR accounting for both pleiotropy and sample structure which seems to be another important confounder (and includes population stratification, cryptic relatedness, and sample overlap); **MRlap** considers both weak instrument bias and winner's curse, accounting for sample overlap. **MR.CUE** and **TS_LMM** offer additional functionality for handling variability of the estimates. **LCV** is a method that estimates causal associations between traits avoiding confounding by genetic correlation, whereas **OMR** uses information from all GWAS SNPs for causal inference and **JAM-MR** performs variable selection and causal effect estimation in MR.

- **CS**, **BiDirectCausal**, **MRCI** and **LHC-MR** constitute another important class of methods since they can identify bidirectional causal effects.
- Another important extension is offered by methods like **MR2**, **MV-MR**, **MRBEE**, **MVMR-cML** and **adOMICs** which extend the MR framework in the multivariate setting allowing more than one exposures or outcomes, as well as **MR-BMA** which go one step further performing multivariate MR in a Bayesian framework.
- Finally, other methods like **hJAM**, **MR.RAPS** and **MRPEA** offer more advanced options. hJAM unifies the framework of MR and TWAS and can be applied to correlated instruments and multiple intermediates, MR.RAPS uses a three-sample genome-wide design with many independent genetic instruments across the genome to handle many weak genetic instruments and pleiotropy, whereas MRPEA uses pathway association MR analysis approach using data of environmental exposures.

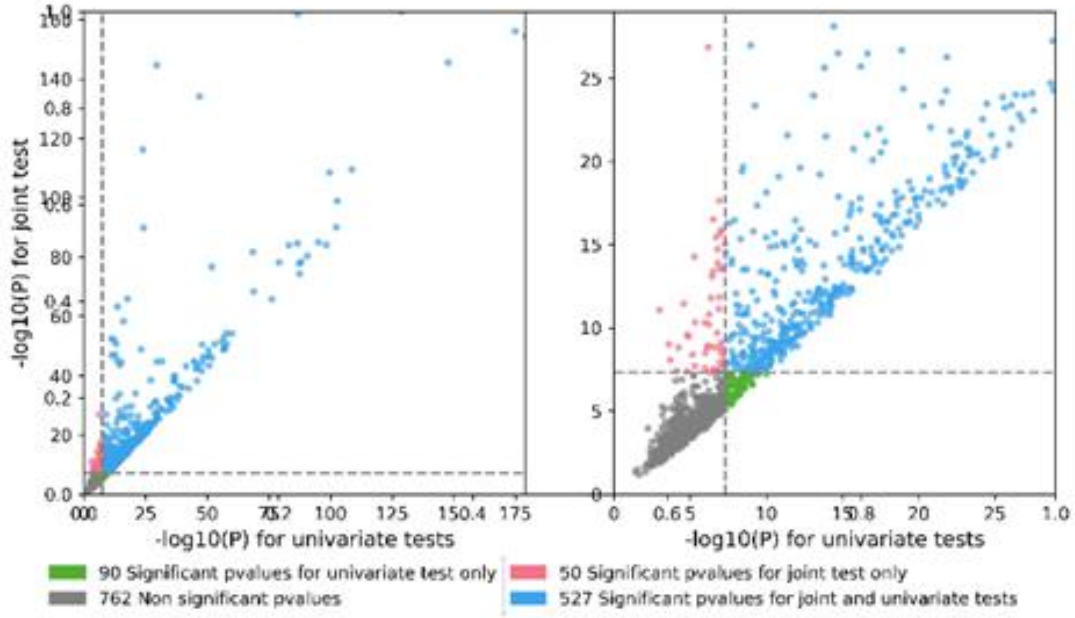
IMRP	https://github.com/XiaofengZhuCase/IMRP	https://pubmed.ncbi.nlm.nih.gov/33226062	MR	R	Performs Iterative MR and Pleiotropy analysis to simultaneously search for horizontal pleiotropic variants and estimate causal effect
CS	https://github.com/xue-hr/Causal_Direction	https://pubmed.ncbi.nlm.nih.gov/33137120	MR	R	Infers causal direction between two traits in the presence of horizontal pleiotropy
MR2	https://github.com/lb664/MR2/	https://pubmed.ncbi.nlm.nih.gov/37419091	MR	R	Performs MR for multiple outcomes to identify exposures that cause more than one outcome or, conversely, exposures that exert their effect on distinct responses
MRmix	https://github.com/gqi/MRMix	https://pubmed.ncbi.nlm.nih.gov/31028273/	MR	R	MR analysis using an underlying mixture model incorporating a fraction of the genetic instruments to have direct effect on the outcome (horizontal pleiotropy)
MR-PRESSO	https://github.com/rondolab/MR-PRESSO	https://pubmed.ncbi.nlm.nih.gov/29686387/	MR	R	Allows for the evaluation of horizontal pleiotropy in multi-instrument MR
MV-MR	https://mrcieu.github.io/software/mvmr-r-package/	https://pubmed.ncbi.nlm.nih.gov/30535378/	MR	R	Performs multivariable MR analyses, including heterogeneity statistics for assessing instrument strength and validity
MR-BMA	https://github.com/verena-zuber/demo_AMD	https://pubmed.ncbi.nlm.nih.gov/31911605/	MR	R	Bayesian algorithm to perform risk factor selection in multivariable MR
MR	https://cran.r-project.org/web/packages/MendelianRandomization/index.html	https://pubmed.ncbi.nlm.nih.gov/31953392/	MR	R	Several standard methods (simple and weighted median, IVW, and MR-Egger) for performing MR analyses with summary data
pIVW	https://cran.r-project.org/web/packages/mr.pivw/index.html	https://pubmed.ncbi.nlm.nih.gov/35942938	MR	R	An extension to IVW that accounts for weak instruments and balanced horizontal pleiotropy simultaneously
MR-APSS	https://github.com/YangLabHKUST/MR-APSS	https://pubmed.ncbi.nlm.nih.gov/35787050/	MR	R	Performs MR accounting for both pleiotropy and sample structure (which includes population stratification, cryptic relatedness, and sample overlap)
BWMR	https://github.com/jiazhao97/BWMR	https://pubmed.ncbi.nlm.nih.gov/31593215/	MR	R	Bayesian methods for MR
TwoSampleMR	https://mrcieu.github.io/TwoSampleMR/	https://pubmed.ncbi.nlm.nih.gov/29149188/	MR	R	Standard methods for performing MR. It uses the IEU GWAS database and to the MR-Base web app
MRbase	https://www.mrbase.org	https://pubmed.ncbi.nlm.nih.gov/29846171/	MR	web	A database and analytical platform for Mendelian randomization. It is coupled to TwoSampleMR and to MRC IEU OpenGWAS database.
LCV	https://github.com/lukejoconnor/LCV	https://pubmed.ncbi.nlm.nih.gov/30374074/	MR	R	Estimates causal associations between traits avoiding confounding by genetic correlation. Uses LDSC
MRcML	https://github.com/xue-hr/MRcML	https://pubmed.ncbi.nlm.nih.gov/34214446	MR	R	Uses constrained maximum likelihood and model averaging, that is robust to invalid IVs with uncorrelated or correlated pleiotropic effects
MR-LDP	https://github.com/QingCheng0218/MR.LDP	https://pubmed.ncbi.nlm.nih.gov/33575584	MR	R	Probabilistic model for MR in the presence of LD and to account for horizontal pleiotropy
MR-Corr2	https://github.com/QingCheng0218/MR.Corr2	https://pubmed.ncbi.nlm.nih.gov/34499127	MR	R	Bayesian approach that uses the orthogonal projection to reparameterize the bivariate normal distribution for effects of variants on exposure and horizontal pleiotropy
MR.CUE	https://github.com/QingCheng0218/MR.CUE	https://pubmed.ncbi.nlm.nih.gov/36310177	MR	R	Estimates causal effect while identifying IVs with correlated horizontal pleiotropy and accounting for estimation uncertainty
TS_LMM	https://github.com/mingding-hsph/TS_LMM	https://pubmed.ncbi.nlm.nih.gov/37162968	MR	R	Performs two-stage linear mixed model for MVMR that accounts for variance of summary statistics not only in outcome, but also in all of the risk factors
MRlap	https://github.com/n-mounier/MRlap/	https://pubmed.ncbi.nlm.nih.gov/37036286	MR	R	Simultaneously considers weak instrument bias and winner's curse while accounting for potential sample overlap and corrects IVW-MR
MRCI	https://github.com/zpliu/MRCI	https://pubmed.ncbi.nlm.nih.gov/36854672	MR	R	Estimates reciprocal causation between two phenotypes simultaneously using reference LD information
BiDirectCausal	https://github.com/xue-hr/BiDirectCausal	https://pubmed.ncbi.nlm.nih.gov/35576237	MR	R	Infers possibly bi-directional causal effects between two traits
LHC-MR	https://github.com/LizaDarrous/lhcMR	https://pubmed.ncbi.nlm.nih.gov/34907193	MR	R	Estimates bi-directional causal effects, direct heritabilities, and confounder effects while accounting for sample overlap
MRBEE	https://github.com/noahlorinczcomi/MRBEE	https://pubmed.ncbi.nlm.nih.gov/37066391	MR	R	Multivariable MR method capable of simultaneously removing measurement error bias and identifying horizontal pleiotropy
MVMR-cML	https://github.com/ZhaotongL/MVMR-cML	https://pubmed.ncbi.nlm.nih.gov/36948188	MR	R	An efficient and robust MVMR method based on constrained maximum likelihood (cML)
adOMICs	https://github.com/lshen/adomics	https://pubmed.ncbi.nlm.nih.gov/36094096	MR	Python	Used to investigate the causal effects of multiple omics biomarkers on an outcome. The method first tests the effect of each omics biomarker on the outcome separately using an MR method and then combines the p-values using various methods
OMR	https://github.com/wanglu205/OMR	https://pubmed.ncbi.nlm.nih.gov/34379090	MR	R	MR method that uses all GWAS SNPs for causal inference. The method accommodates the commonly encountered horizontal pleiotropy effects and relies on a composite likelihood framework for scalable computation
JAM-MR	https://github.com/pjnewcombe/R2BGLiMS	https://pubmed.ncbi.nlm.nih.gov/34155684	MR	R	Performs variable selection and causal effect estimation in MR as an extension of the JAM algorithm
hJAM	https://cran.r-project.org/web/packages/hJAM/	https://pubmed.ncbi.nlm.nih.gov/33404048	MR	R	A two-stage hierarchical model that unifies the framework of MR and TWAS and can be applied to correlated instruments and multiple intermediates
MR.RAPS	https://github.com/qingyuanzhao/mr.raps	https://pubmed.ncbi.nlm.nih.gov/31298269	MR	R	A three-sample genome-wide design with many independent genetic instruments across the whole genome. The method is efficient with many weak genetic instruments and robust to balanced and/or sparse pleiotropy
MRPEA	https://sourceforge.net/projects/mrpea/files/	https://pubmed.ncbi.nlm.nih.gov/28334273	MR	R	A pathway association MR analysis approach, which was capable of correcting the genetic confounding effects of environmental exposures, using data of environmental exposures

INSTITUT PASTEUR JASS

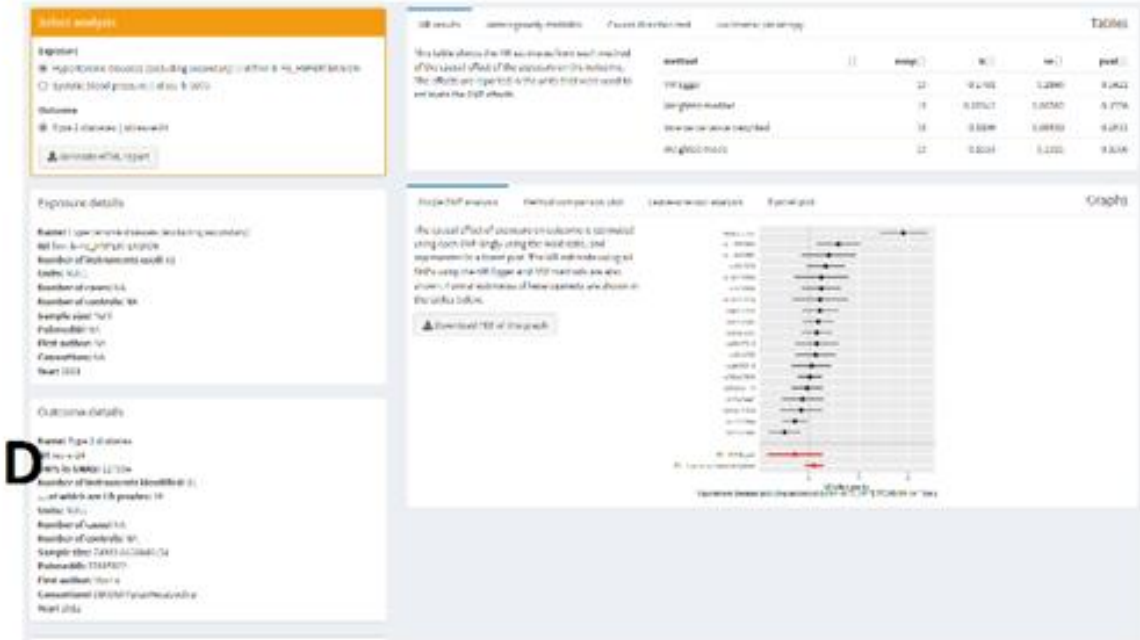
A

B

C

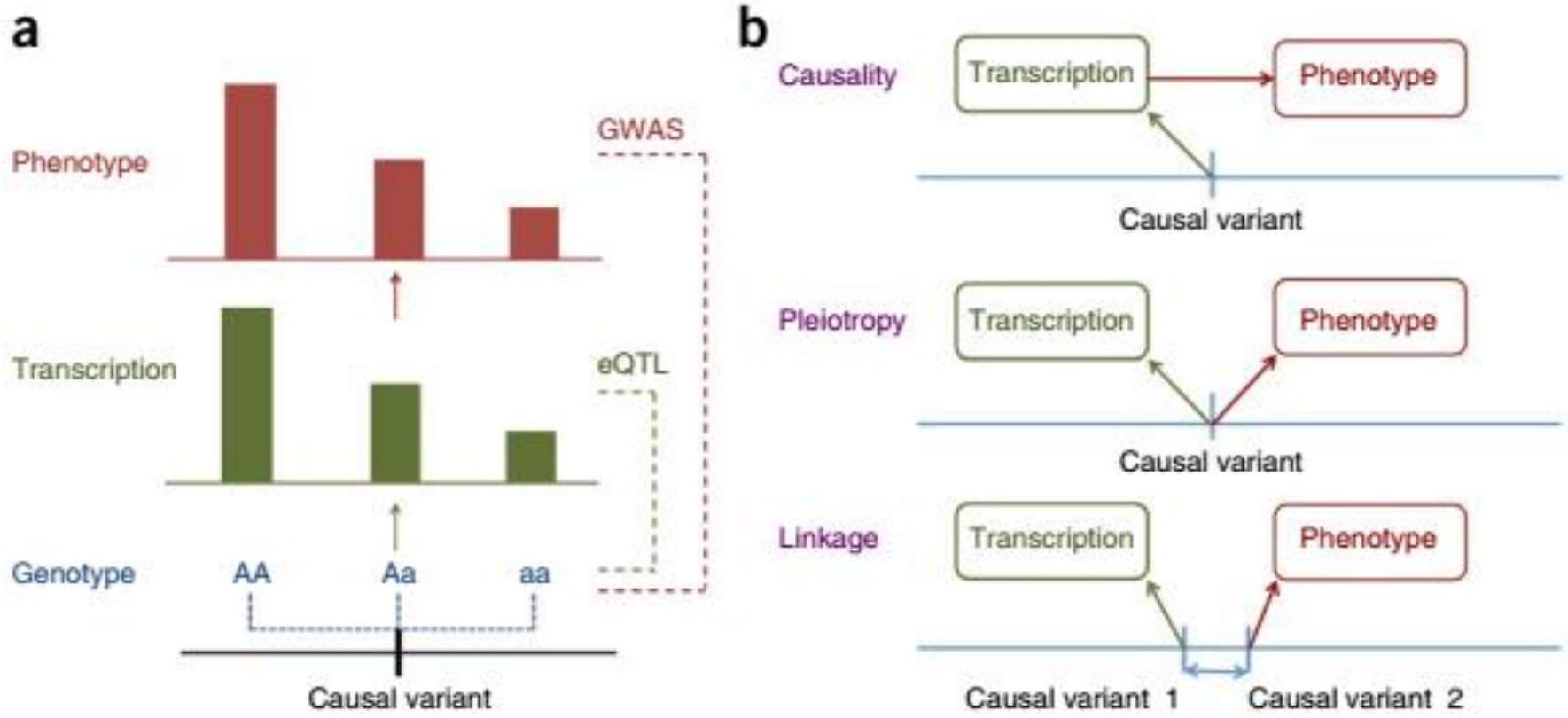


D

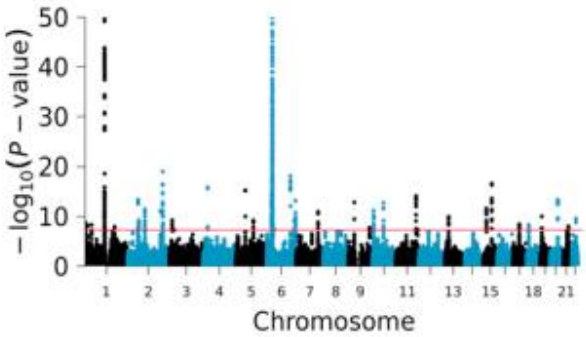


TWAS

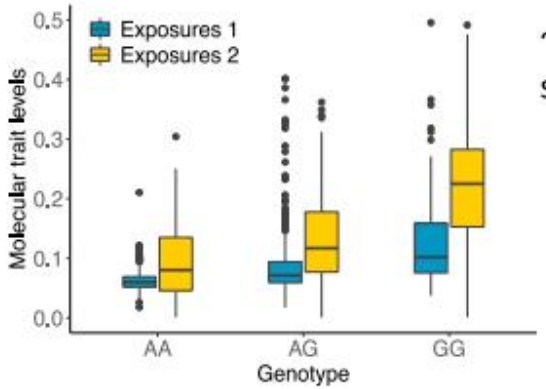
- As we already described, the MR approach involves the combination of two types of data, a genotype-disease association, and a genotype-phenotype association. If the phenotype involves gene-expression, that is the result of an eQTL study, then we have two distinct but fundamentally related methods, the Transcriptome-wide association study (TWAS) and the colocalization approach.
- TWAS is based on the idea that genetic variants can influence gene expression, which subsequently can affect complex traits or diseases. Thus, the approach uses information from eQTL to identify associations between predicted gene expression levels and complex traits/diseases.
- Even though there are several different methods, the resemblance to MR is obvious; in fact several methods like **SMR** that uses a single variant, **GSMR** that uses multiple variants, and **PMR** which can account for correlated instruments, horizontal pleiotropy, and can accommodate both single traits and multiple correlated outcomes, all use the term MR, whereas the authors of **TScML**, which uses two-stage constrained maximum likelihood, which is an extension of 2SLS, explicitly state that can be used for both MR and TWAS analyses.


a
Phenotype
GWAS
Transcription
eQTL
Genotype
AA
Aa
aa
Causal variant
b
Causality
Transcription
Phenotype
Causal variant
Pleiotropy
Transcription
Phenotype
Causal variant
Linkage
Transcription
Phenotype
Causal variant 1
Causal variant 2

Genome-wide association studies (GWAS)



Molecular quantitative trait loci (xQTL)



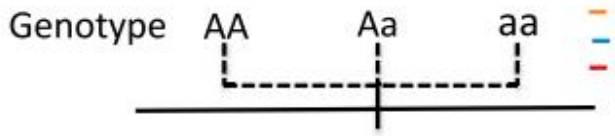
~50% GWAS signals shared with at least one molecular phenotype

LD

Complex trait — GWAS

Molecular phenotype 1 — $xQTL_1$

Molecular phenotype k — $xQTL_k$



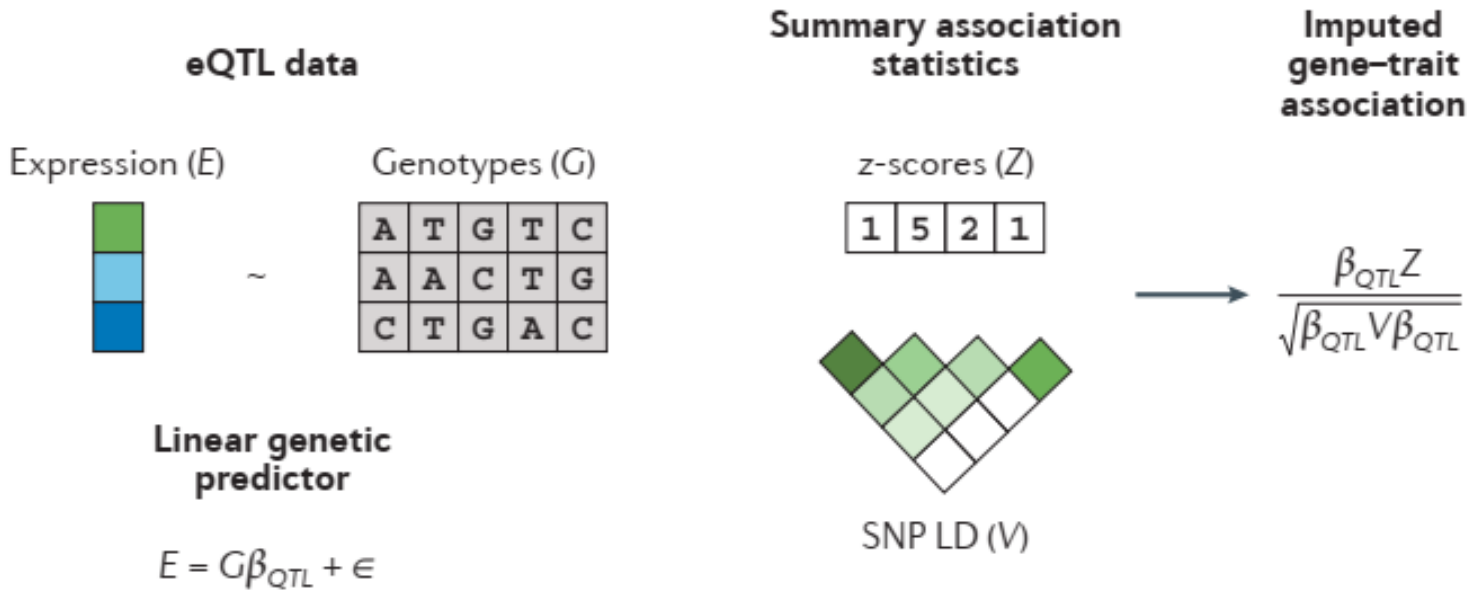
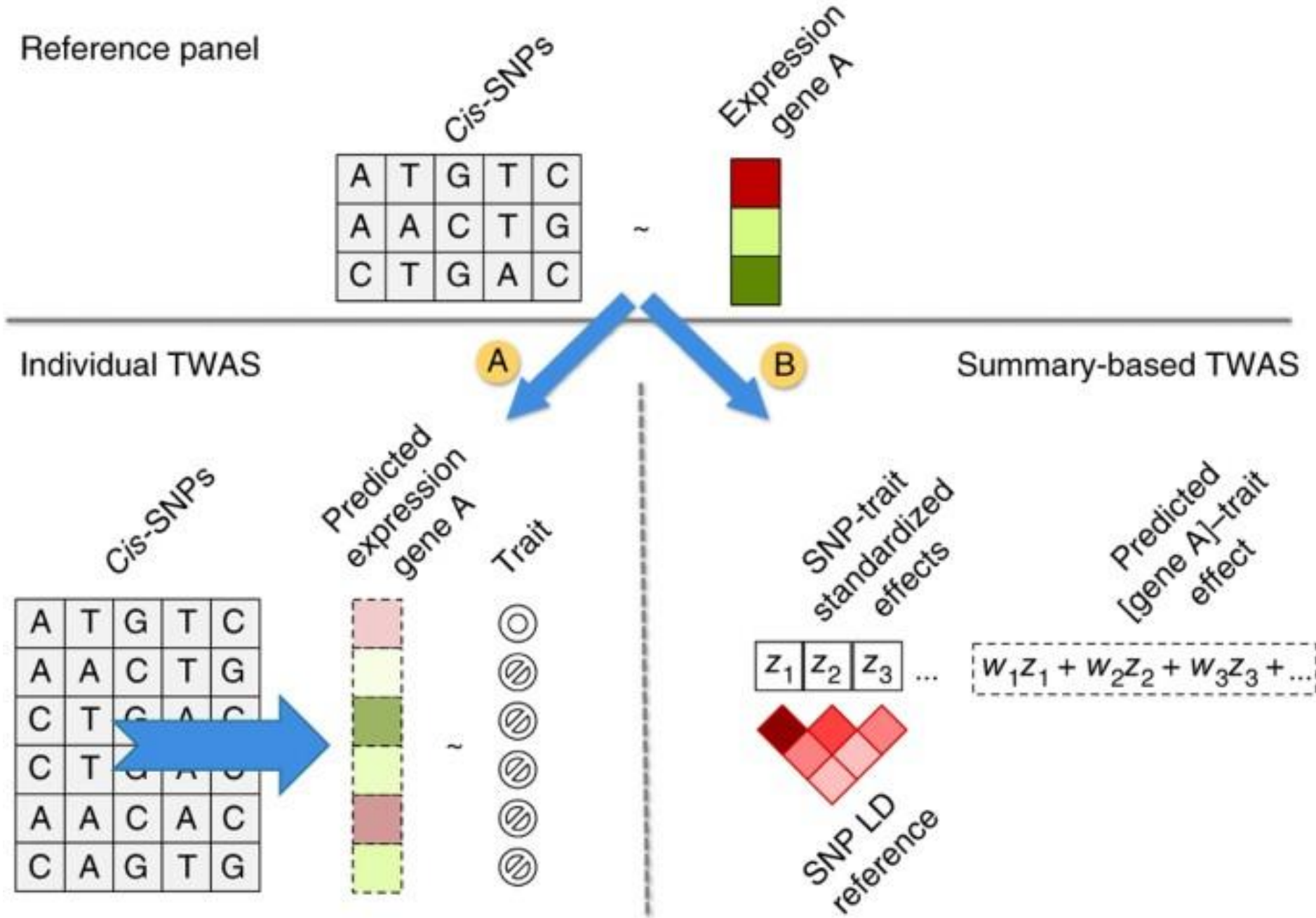


Figure 2 | **TWAS using predicted expression and summary data.** Transcriptome-wide association studies (TWAS) using predicted expression and summary data follow two steps. First, transcriptome reference data are used to build a linear predictor for gene expression, typically using single nucleotide polymorphisms (SNPs) from the 1 Mb local region around the gene with regularized effect sizes (for example, using a Bayesian sparse linear mixed model[101]). Second, this predictor is applied to summary genome-wide association *z*-scores, and gene–trait association *z*-scores are computed, testing the null model of no association between a gene and a trait. eQTL, expression quantitative trait loci; LD, linkage disequilibrium.


Reference panel
Cis-SNPs
Expression gene A
A T G T C
A A C T G
C T G A C
~
Individual TWAS
A
B
Summary-based TWAS
Cis-SNPs
Predicted expression gene A
Trait
SNP-trait standardized effects
Predicted [gene A]–trait effect
A T G T C
A A C T G
C T G A C
C T G A C
A A C A C
C A G T G
~
z_1 z_2 z_3 ...
$w_1z_1 + w_2z_2 + w_3z_3 + ...$
SNP LD reference

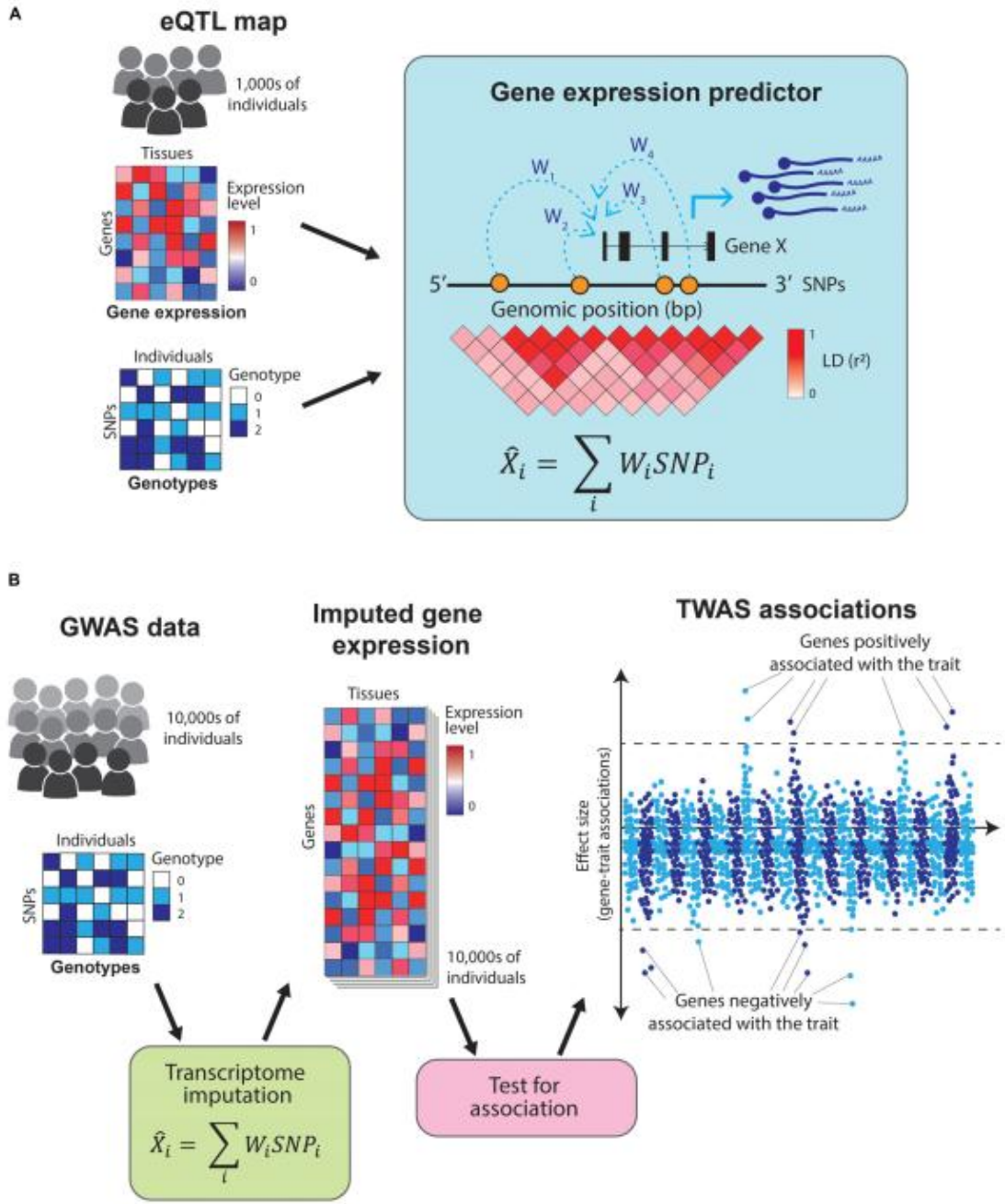


FIGURE 4 | Overview of transcriptome-wide association studies TWAS leverage information from eQTL catalogs and GWAS studies to directly associate traits to genes. **(A)** TWAS use eQTL maps (which contain tissue-specific gene expression and genotypes for thousands of individuals) as a training set to build gene expression predictors. These predictors take the SNPs *in cis* to a gene and estimate its expression levels. **(B)** The resulting predictors are used to impute gene expression values across the hundreds of thousands of individuals in a GWAS study (which contains genotypes but no gene expression data). Finally, the imputed gene expression values are directly tested for association with the GWAS trait, resulting in a set of genes which positively or negatively influence it.

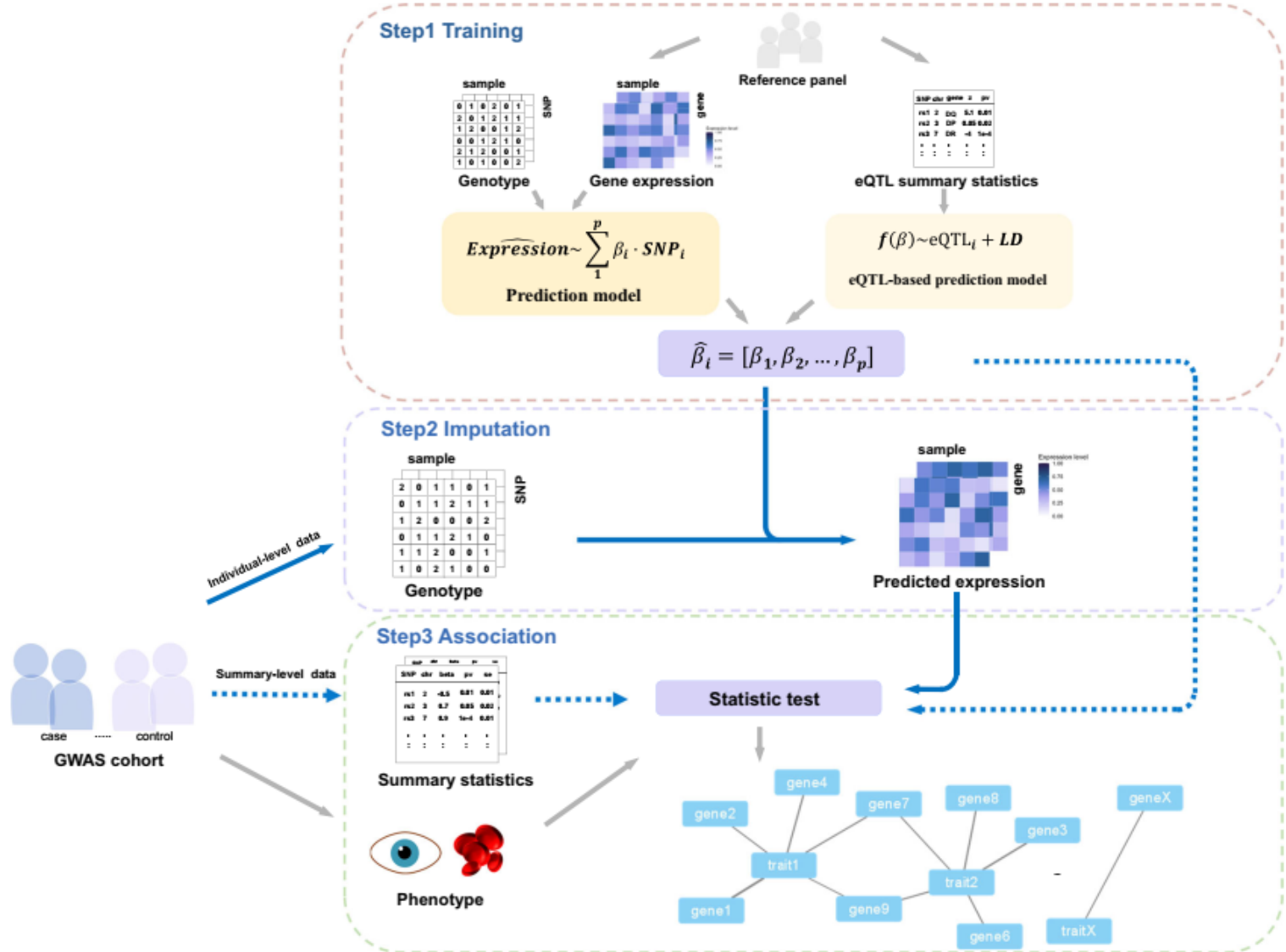
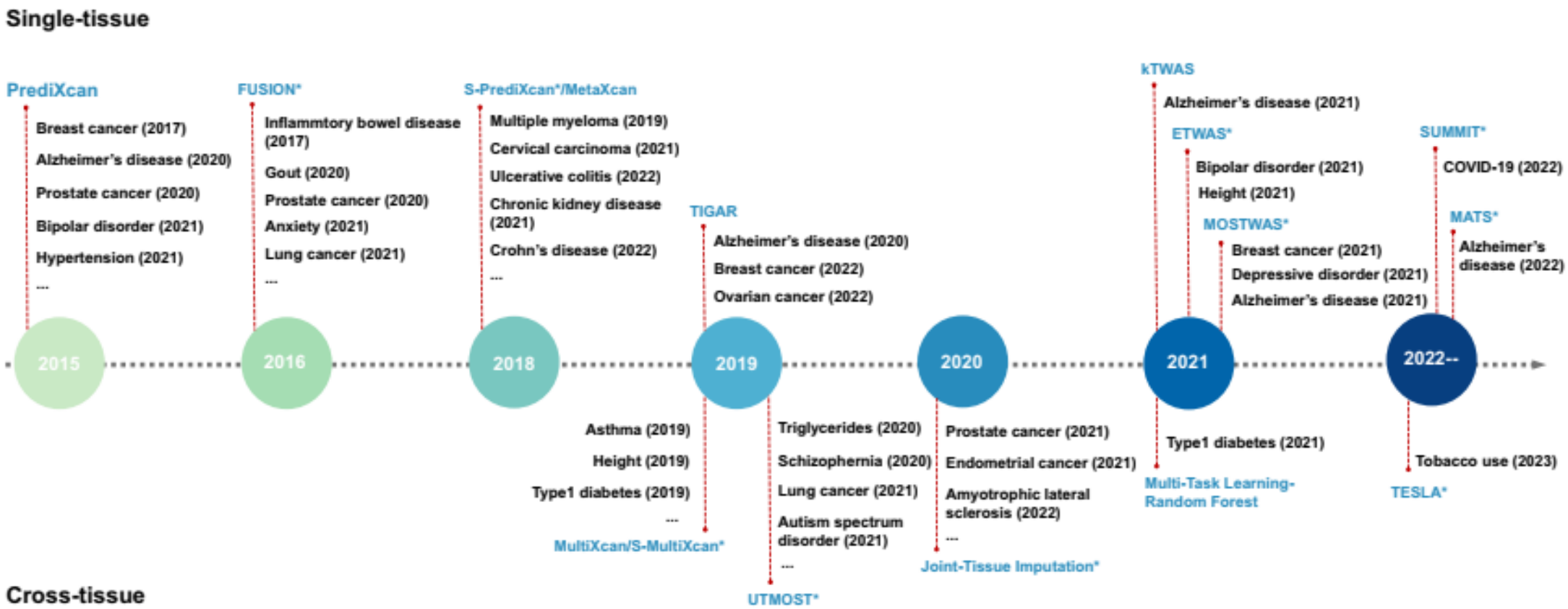
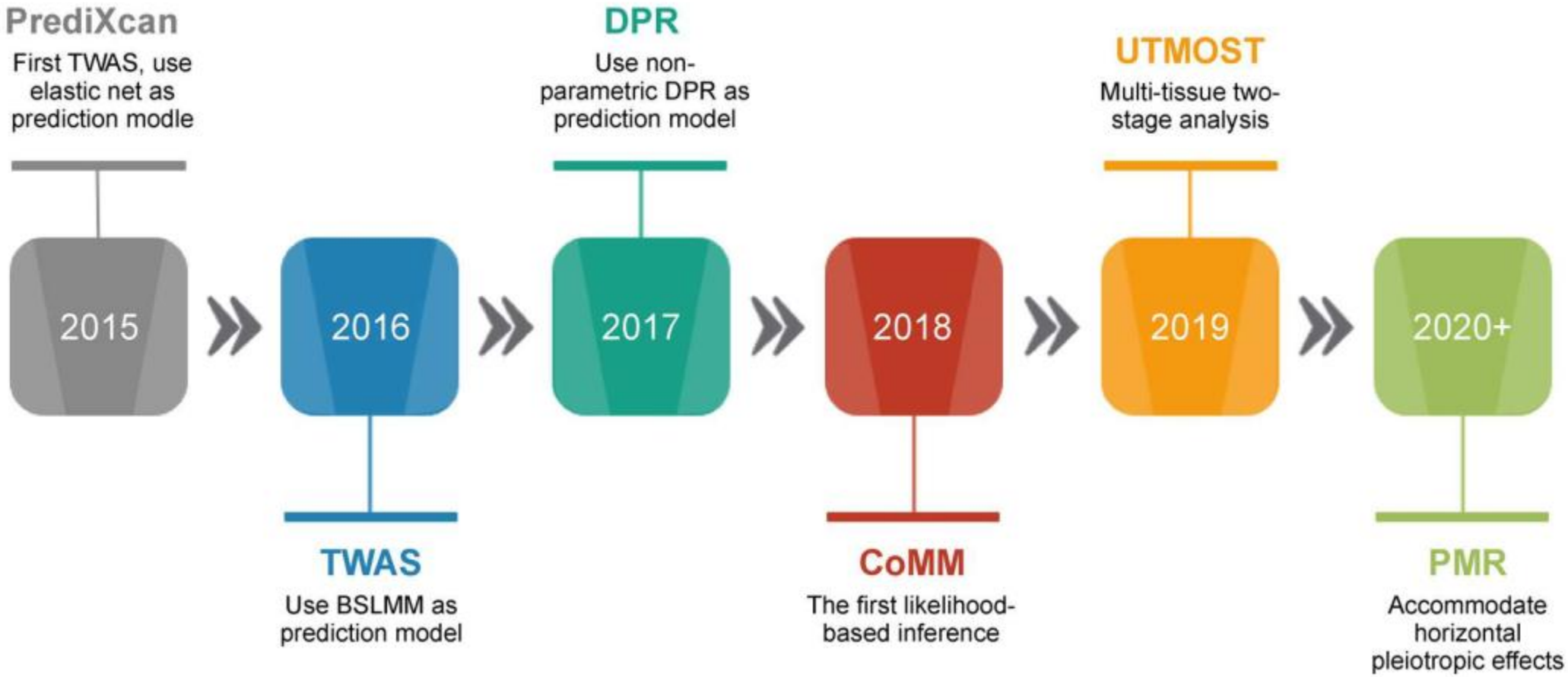


Fig. 1 Schematic workflow of TWAS analysis. The first step is to train expression predictive models by inputting either genotype data and corresponding expression data of reference panel, or eQTL summary statistics with specific TWAS methods. Next, for individual-level GWAS data, the second step imputes the expression data of GWAS individuals using the fitted predictive models. The third step analyzes the association between each phenotype-imputed gene expression pair (as the blue solid line shows). The predicted SNP weight vector is combined to calculate association statistics (as the blue dashed line shows) for GWAS summary statistics data.



Mai J, Lu M, Gao Q, Zeng J, Xiao J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol. 2023;6(1):899.



Zhu H, Zhou X. Transcriptome-wide association studies: a view from Mendelian randomization. Quant Biol. 2021;9(2):107-21.

- **FUSION** and **S-PrediXcan** are the oldest and most widely known methods. FUSION is the current implementation of the first TWAS method, whereas S-PrediXcan is the summary-data version of PrediXcan. Xu et al noted that PrediXcan and TWAS can be viewed as a special case of general association testing with multiple SNPs in a GLM and proposed the so-called sum of powered score (SPU) test implemented in **aSPU-TWAS**. A subsequent evaluation has shown that the original TWAS statistic is equivalent to an LD-aware version of standard MR.
- **iFunMed** and **sMIST** formulate the problem within the framework of mediator analysis, and similarly **PTWAS** applies principles from instrumental variables analysis. **Comm-S*** uses a variational Bayesian EM algorithm and a likelihood ratio test to assess expression-trait association. Its extension **Tiss-Comm** leverages the co-regulation of genetic variations across different tissues explicitly via a unified probabilistic model and also detects the tissue-specific role of candidate target genes in complex traits. Similar multi-tissue approaches are followed by **fQTL**, **sCCA** and **UTMOST**.
- **Primo** and **OPERA** extend further the integration by allowing different types of xQTL data (eQTL, pQTL, mQTL etc) to allow estimation under different conditions, whereas **SUMMIT** uses a large eQTL summary-level dataset, penalized regression and Cauchy Combination Test and **HMAT** aggregates TWAS association tests obtained across multiple gene expression prediction models using the harmonic mean P-value combination (HMP).
- **BGW** and **ARCHIE** are two methods that utilize trans-regulated eQTLs. Other tools use combination of methods, like **TIGAR** which combines DPR and PrediXcan, whereas others, like **JEPEGMIX2-P** or **FOCUS**, perform TWAS using pathway information, or use LD to perform fine-mapping over the gene–trait association signals obtained from TWAS, respectively.

- Even though the various methods discussed here have different modeling assumptions and many were initially developed to answer different biological questions, a recent technical review of the TWAS methods showed that all can be viewed as versions of the two-sample MR analysis.
- Indeed, several recent tools like **MRLocus**, **TWMR**, and **Mr.MtRobin** make explicit use of the MR methodology and jargon in order to perform a sophisticated TWAS.
- **MRLocus** performs first a colocalization step to each nearly-LD-independent eQTL, and then performs an MR analysis step across eQTLs.
- **TWMR** performs a multi-gene multi-instrument MR approach to identify genes whose expression influence the phenotype.
- Finally, **Mr.MtRobin** uses multi-tissue eQTL and a reverse regression random slope mixed model to infer whether a gene is associated with a complex trait.
- As we have already noticed, webTWAS, apart from the database, also offers a webserver for accessing S-PrediXcan, SMR and UTMOST with user supplied datasets.

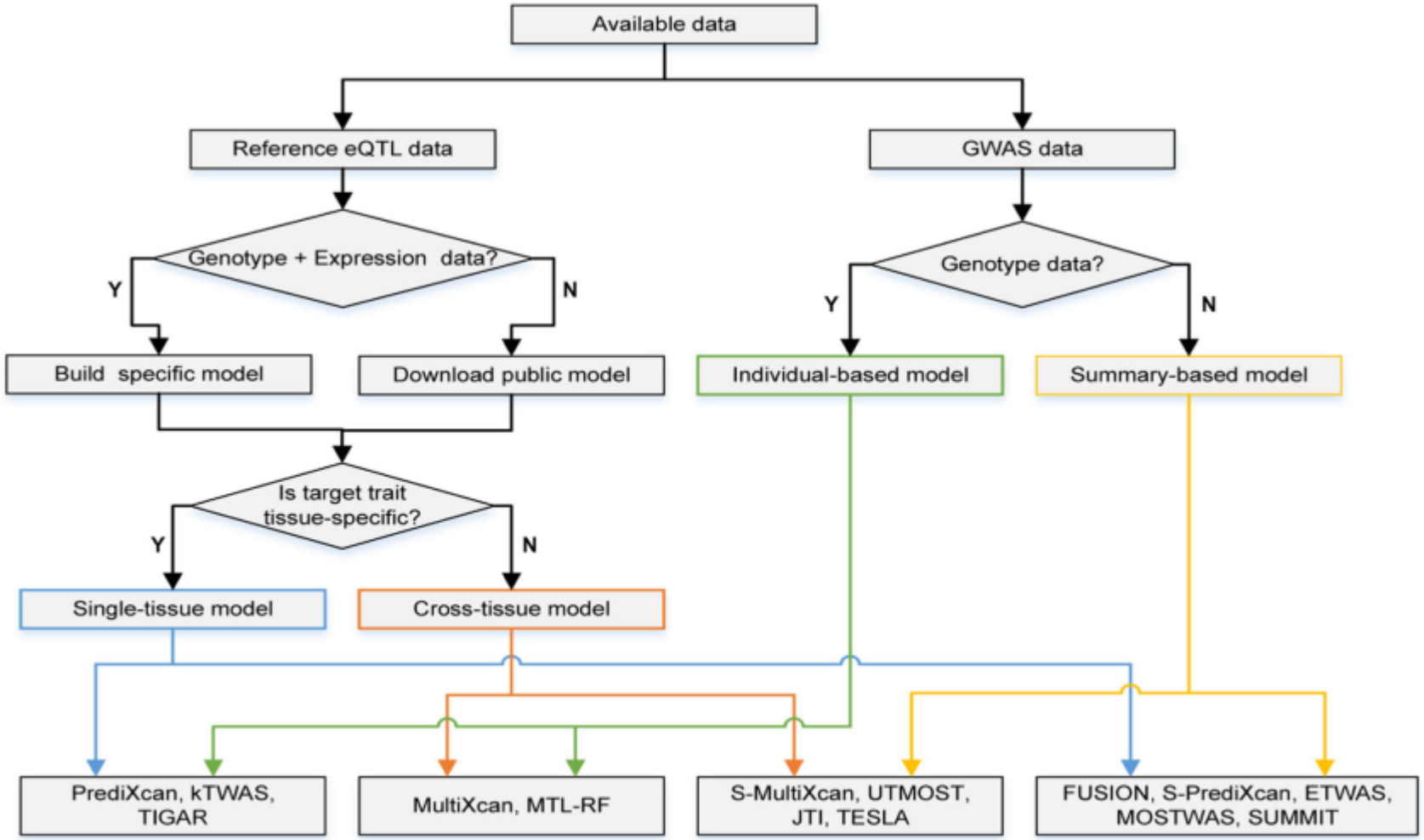
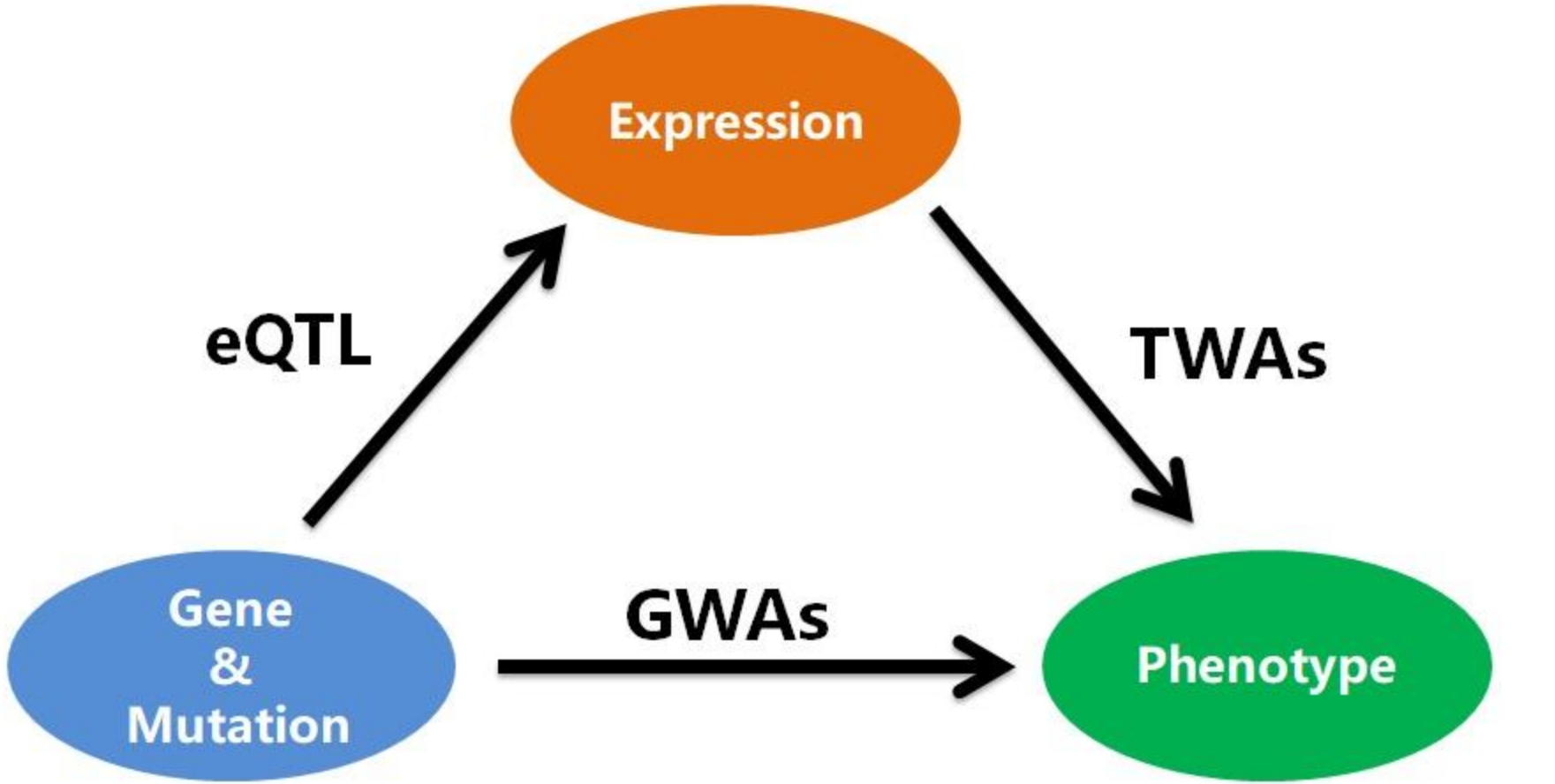


Fig. 3 Suggestions for TWAS methods selection. The tree shows the selection process of TWAS models decided by available data and trait characteristics. It involves three main aspects, including the availability of reference panel data, the category of GWAS data, and the preference of tissue type. The joint selection determines the suggested TWAS methods listed at the bottom of the tree.

Mai J, Lu M, Gao Q, Zeng J, Xiao J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol. 2023;6(1):899.

Colocalization

- Another method that also uses GWAS results along with eQTL data is colocalization. Colocalization approaches are used to assess whether two different traits or diseases share a common causal genetic variant or set of variants at a specific genomic locus.
- Colocalization analysis identifies genetic variants that show significant association in both GWAS and eQTL studies.
- However, unlike TWAS, it does not perform gene expression prediction and gene-trait association tests, but it focuses on the colocalized SNPs
- TWAS and colocalization are related approaches but not identical, since it has been shown that may give different results under different conditions (for instance in case of horizontal pleiotropy) and thus it has been suggested that they should be used complementary



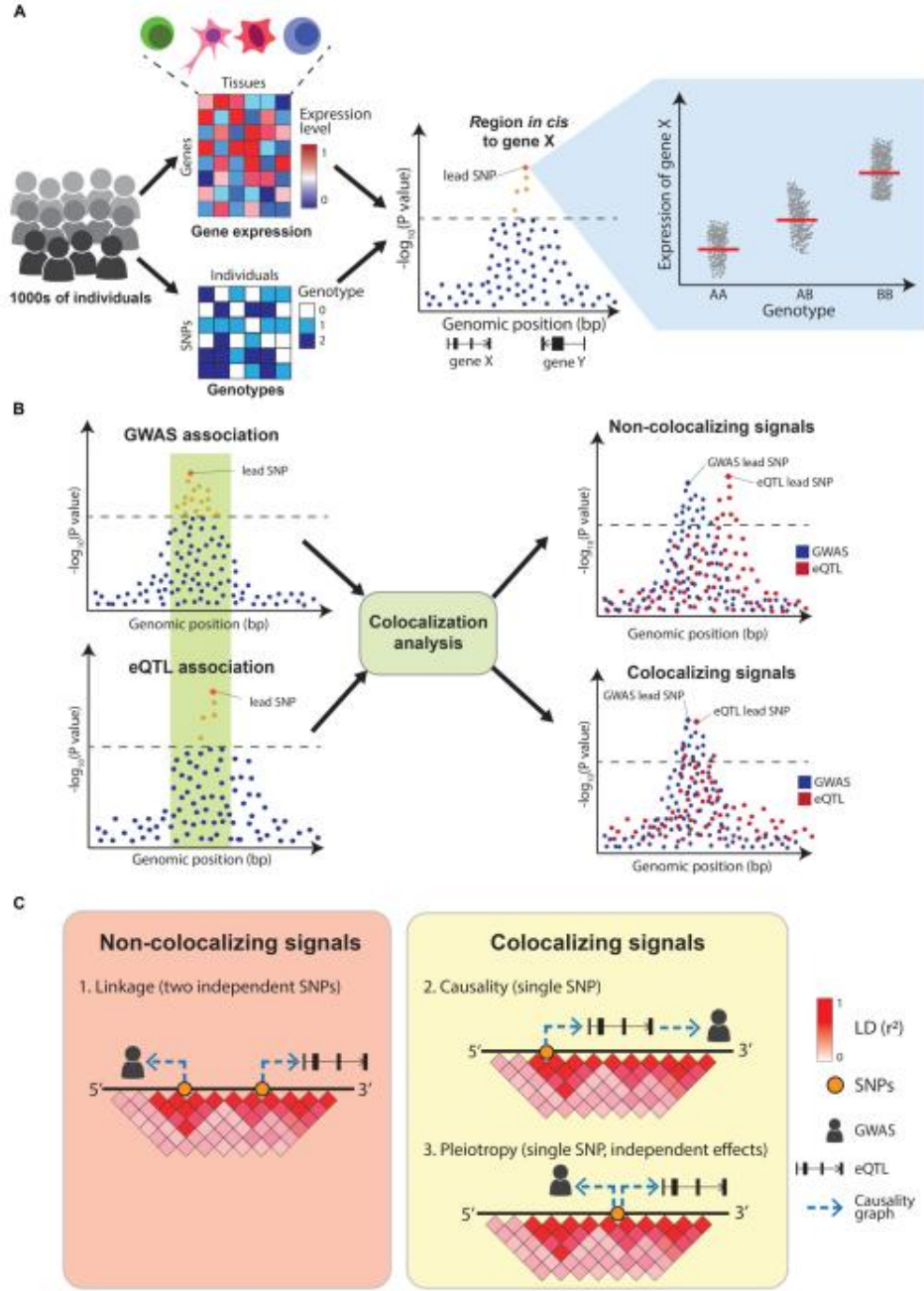


FIGURE 3 | Overview of eQTL-mapping and colocalization. **(A)** In eQTL-mapping gene expression is profiled in thousands of individuals and the expression level of each gene is tested for association with genotypes at nearby (*cis*) SNPs. **(B)** Colocalization compares the association patterns of GWAS and eQTLs at a locus to find if both signals are driven by the same causal variants. **(C)** GWAS and eQTL signals can overlap for three reasons: two independent causal variants in LD (linkage), a single causal variant affecting the GWAS trait via gene expression modulation (causality) or a single causal variant affecting both traits independently (pleiotropy). A positive colocalization supports causality or pleiotropy in favor of linkage.

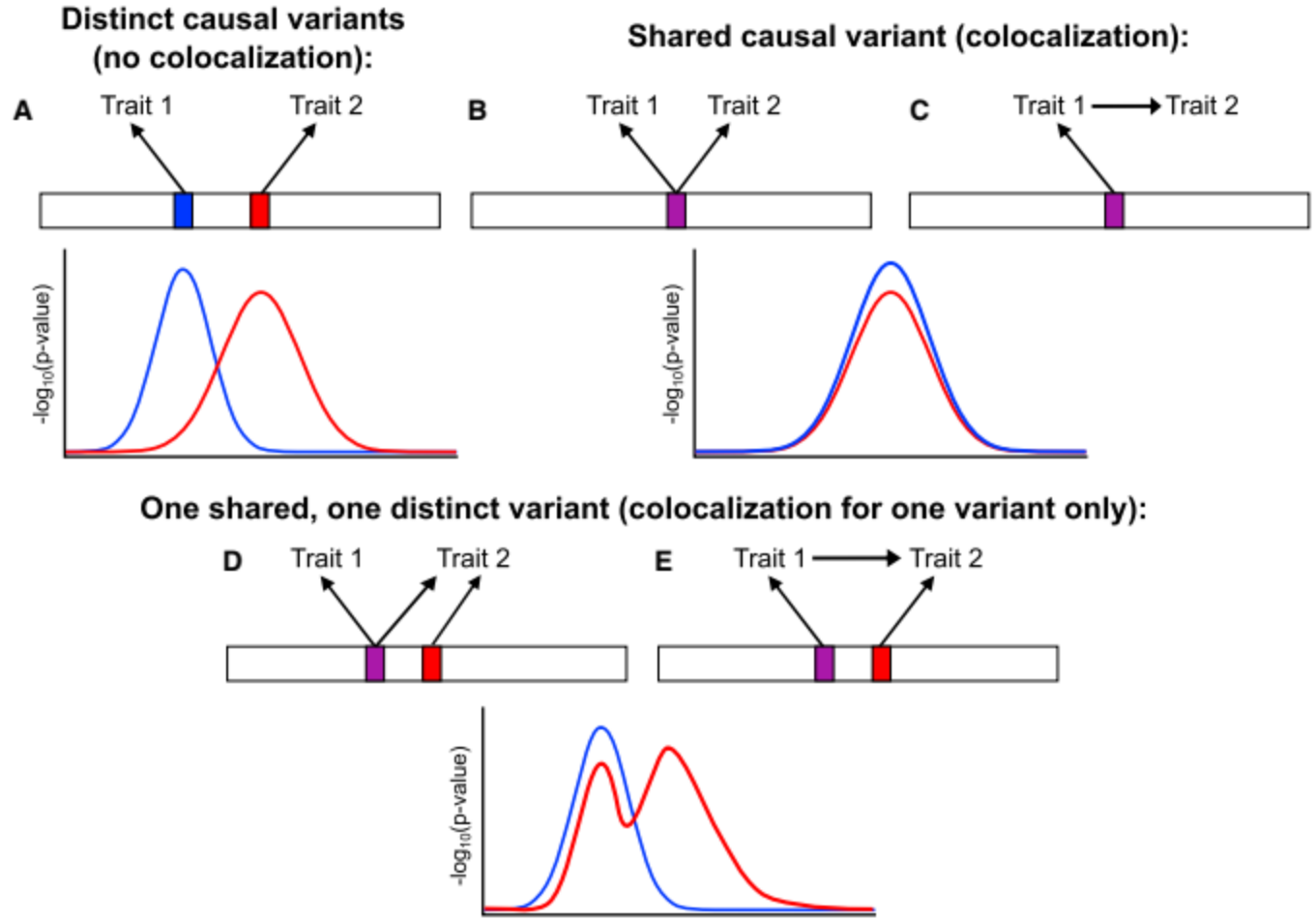


Figure 2. Schematic diagrams illustrating colocalization in five scenarios

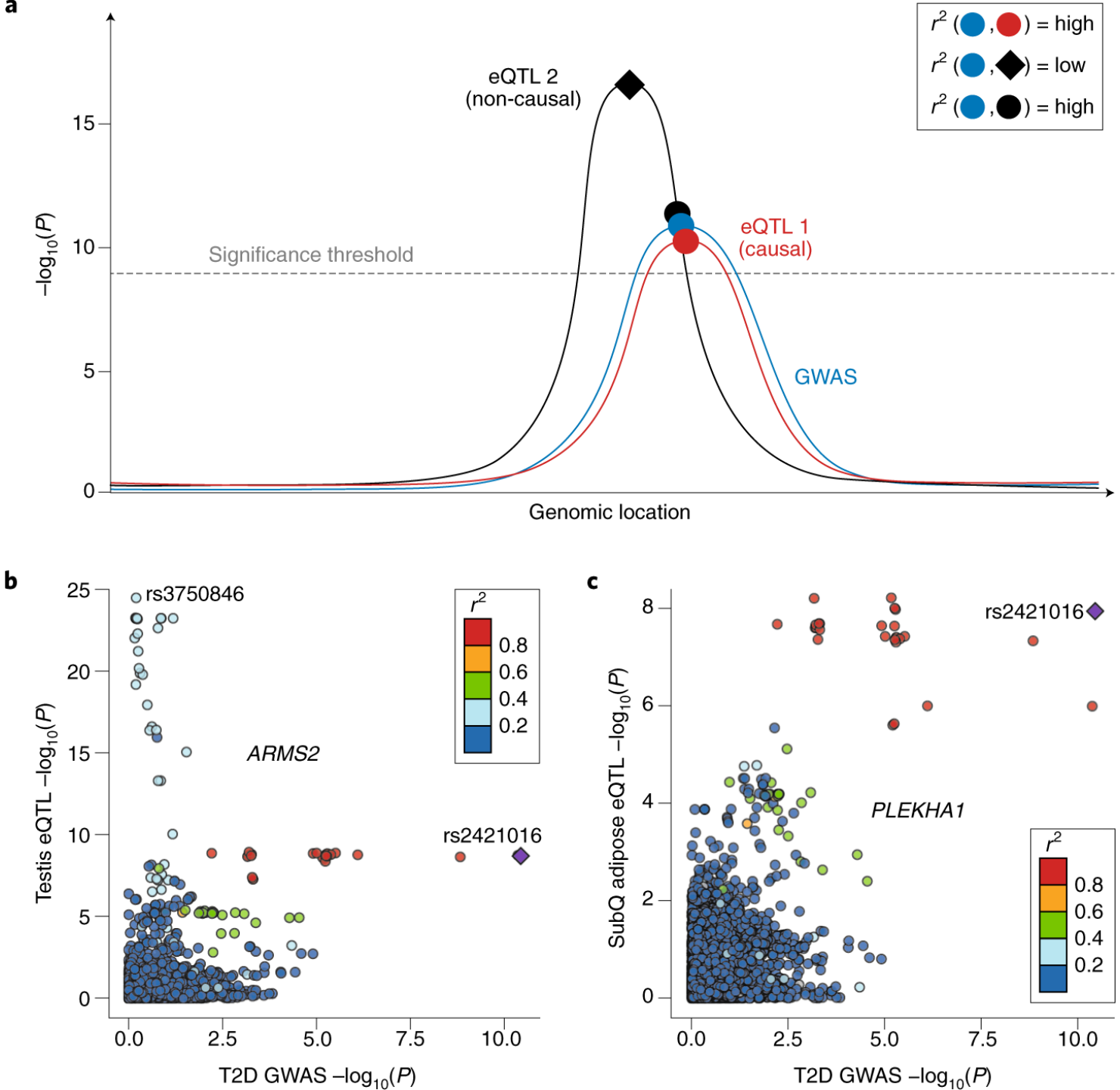
(A) Two traits with distinct causal variants in linkage disequilibrium.

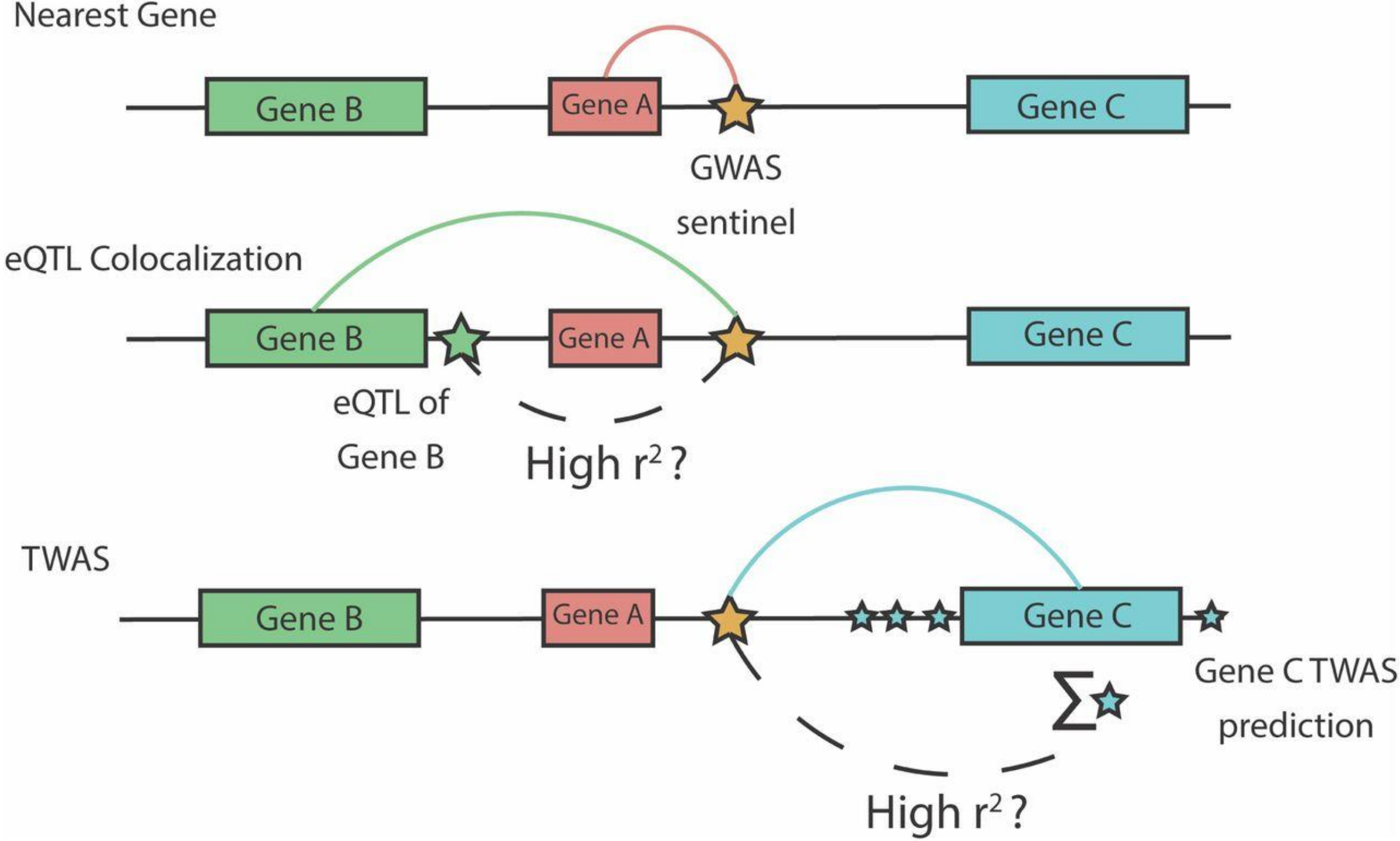
(B) Two unrelated traits with a shared causal variant.

(C) Two traits with a shared causal variant where the first trait influences the second trait.

(D and E) One shared causal variant and one distinct causal variant for trait 2.

Scenarios (B) and (C) are examples of colocalization. For scenarios (D) and (E), there is colocalization at the shared variant, but not at the distinct variant. Colocalization is unable to distinguish between the scenarios in which trait 1 and trait 2 are causally unrelated (scenarios B and D), and in which trait 1 has a causal effect on trait 2 (scenarios C and E). Illustrative regional association plots for each scenario represent the negative $\log_{10}$ p values for associations of variants with each trait (blue for trait 1, red for trait 2) plotted against chromosomal position.


a
r^2 (,) = high
r^2 (,) = low
r^2 (,) = high
eQTL 2
(non-causal)
eQTL 1
(causal)
GWAS
Significance threshold
$-\log_{10}(P)$
Genomic location
b
rs3750846
ARMS2
rs2421016
r^2
0.8
0.6
0.4
0.2
Testis eQTL $-\log_{10}(P)$
T2D GWAS $-\log_{10}(P)$
c
rs2421016
PLEKHA1
SubQ adipose eQTL $-\log_{10}(P)$
T2D GWAS $-\log_{10}(P)$


Nearest Gene
Gene B
Gene A
Gene C
GWAS
sentinel
eQTL Colocalization
Gene B
Gene A
Gene C
eQTL of
Gene B
High r² ?
TWAS
Gene B
Gene A
Gene C
Gene C TWAS
prediction
High r² ?

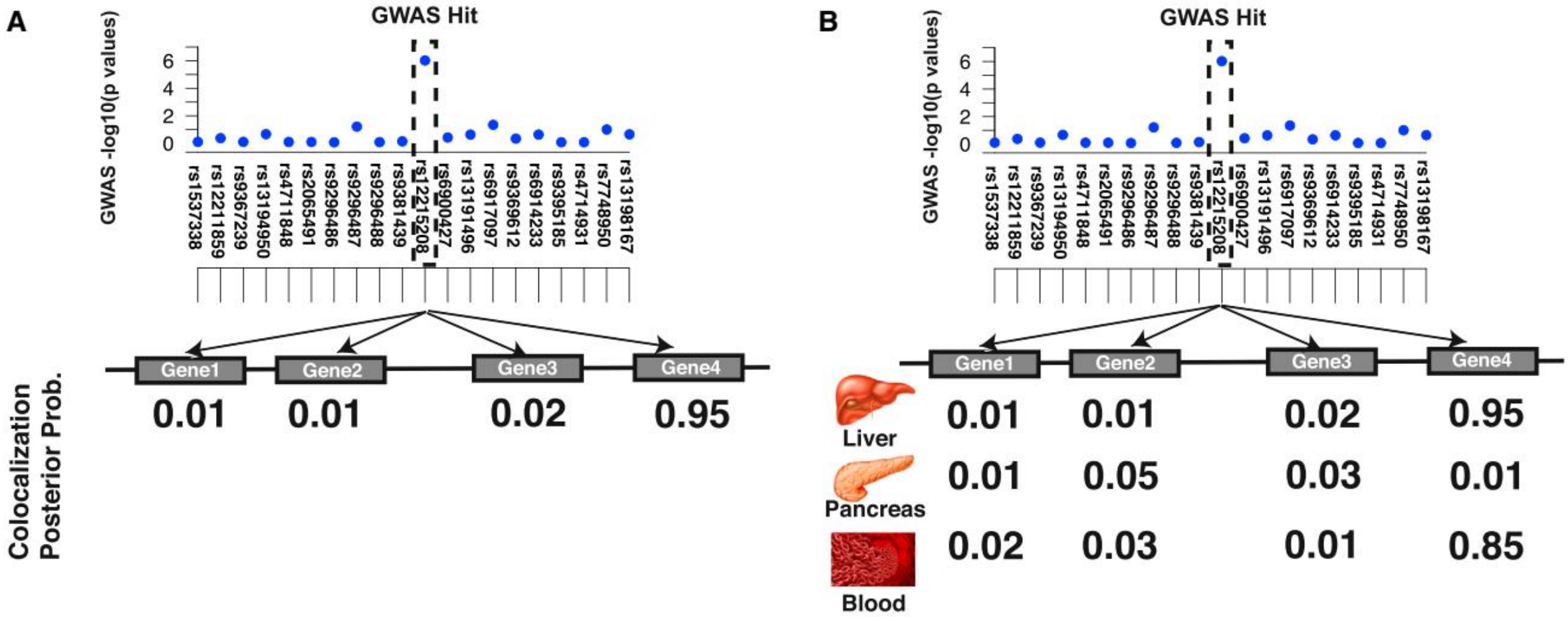


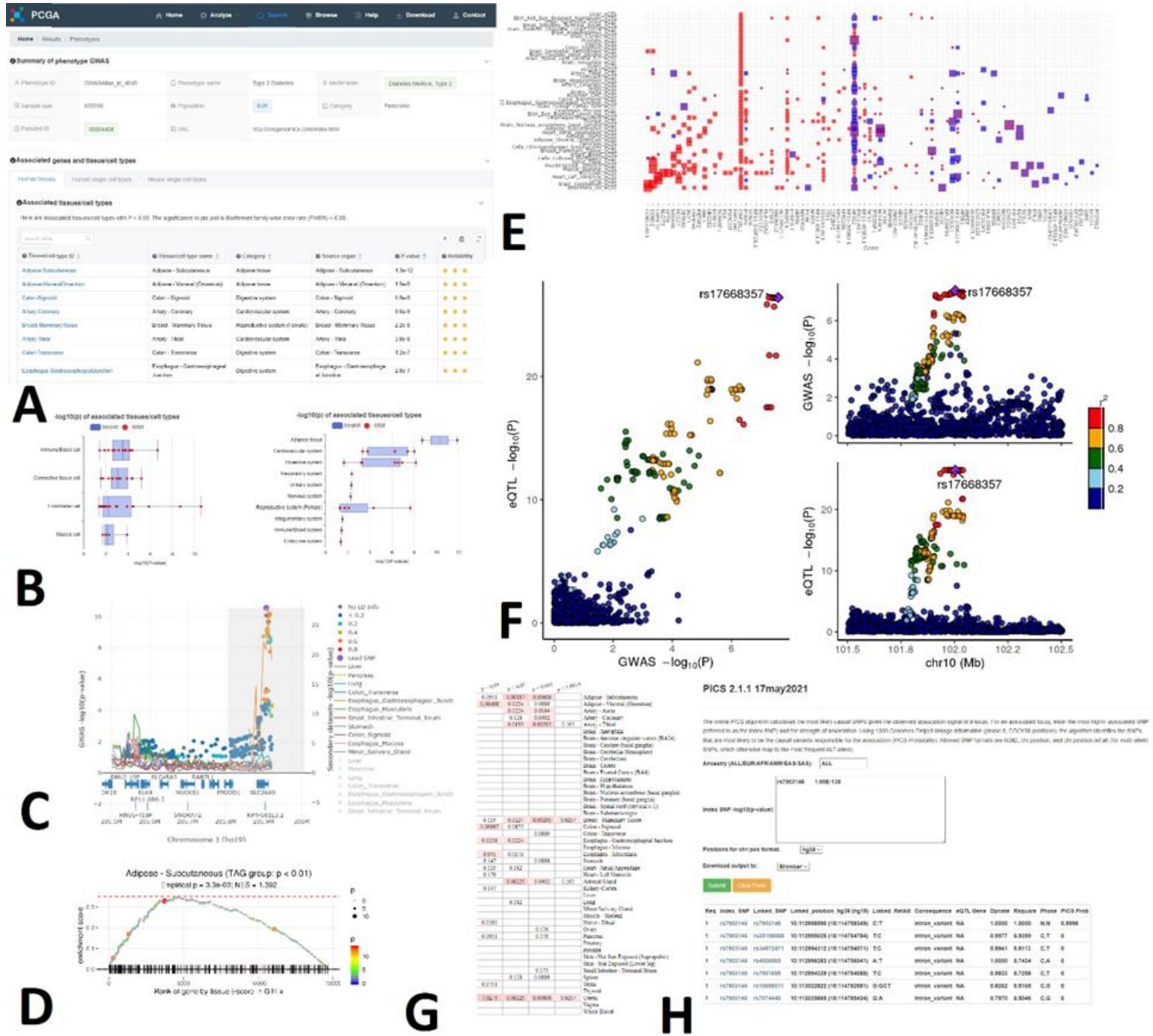
Figure 1. Overview of Our Method for Detecting the Target Gene and Most Relevant Tissue

We compute the CLPP for all genes and all tissues.

(A) A simple case where we have only one tissue and want to find the target gene. We consider all genes for this GWAS risk locus and observe that gene 4 has the highest CLPP. Thus, the target gene is gene 4.

(B) We have three tissues and utilize the quantity of CLPP. Thus, the target gene is gene 4 again. Moreover, in this example, liver and blood are considered the relevant tissues for this GWAS risk locus, whereas the pancreas is not relevant.

- **COLOC** was one of the first methods for colocalization and has seen several improvements. The latest version uses SuSiE and allows evidence for association at multiple causal variants to be evaluated simultaneously, while at the same time separating the statistical support for each variant conditional on the causal signal being considered. **MOLOC** is multiple-trait version of COLOC, operating in a Bayesian framework that integrates GWAS summary data with multiple xQTL data to identify regulatory effects, **HyPrColoc** is a deterministic Bayesian method that detects colocalization across large numbers of traits, and **SS2** operates across any number of gene-tissue pairs allowing for sample overlap.
- **LLR** works for colocalizing genetic risk variants in multiple GWAS and phenotypes, whereas **POEMcoloc** is an approximation to the COLOC method that can be applied when limited data are available. **SparkINFERNO**, **PwCoCo** and **ColocQuiaL** are pipelines offering additional functionalities, all using COLOC.
- **eCAVIAR** is another popular method that uses a probabilistic model that accounts for more than one causal variant at a given locus. **MSG** increases the power using a spliced gene approach and **SharePro** integrates LD modeling and colocalization assessment to account for multiple causal variants in colocalization analysis. **PESCA** uses estimates of LD that are ancestry-matched, in order to infer proportions of population-specific and shared causal variants in two populations. These estimates are then used as priors in an empirical Bayes framework for colocalization and test for enrichment of these causal variants in loci of interest. Lastly, we have to mention the methods that operate as webservers offering ease of use.
- **Sherlock** which is also one of the oldest methods, uses a database of eQTL associations from different tissues to identify genetic signatures that match those for specific genes. Unlike other methods it incorporates information from both cis- and trans- eQTL SNPs. **LocusFocus** is a web-based colocalization tool that tests colocalization using the Simple Sum method to identify relevant genes and tissues for a particular GWAS locus in the presence of high linkage disequilibrium and/or allelic heterogeneity.
- Regarding the analysis of eQTL data, **ezQTL** is a webserver performing various tasks like data quality control for variants matched between different datasets, LD visualization, and colocalization analysis using eCAVIAR and HyPrColoc, whereas **BAGEA** uses a variational Bayes framework to model cis-eQTLs using directed and undirected genomic annotations


PCGA
Home
Analyze
Search
Browse
Help
Download
Contact
Summary of phenotype GWAS
Associated genes and tissue/cell types
Associated tissues/cell types
A
-log10(p) of associated tissues/cell types
B
C
Chromosome 1 (hg19)
D
Adipose - Subcutaneous (TAG group: p < 0.01)
Rank of gene by tissue specificity (TSI)
E
F
rs17668357
GWAS −log10(P)
eQTL −log10(P)
chr10 (Mb)
G
H
PICS 2.1.1 17may2021

Sherlock	http://sherlock.ucsf.edu	https://pubmed.ncbi.nlm.nih.gov/23643380/	Colocalization	web	It uses a database of eQTL n different tissues to identify patterns in GWAS that match those for specific genes. information from both cis- and trans- eQTL SNPs
COLOC	https://github.com/chr1swallace/coloc	https://pubmed.ncbi.nlm.nih.gov/34587156	Colocalization	R	Allows evidence for association at multiple causal variants to be evaluated simultaneously, whilst separating the statistical support for each variant conditional on the causal signal being considered.
eCAVIAR	https://github.com/fhormoz/caviar	https://pubmed.ncbi.nlm.nih.gov/27866706	Colocalization	C/C++	Colocalization of GWAS and eQTL with a probabilistic method that accounts for more than one causal variant in any given locus
moloc	https://github.com/clagiamba/moloc	https://pubmed.ncbi.nlm.nih.gov/29579179/	Colocalization	R	Multiple-trait-coloc, a Bayesian statistical framework that integrates GWAS summary data with multiple molecular QTL data to identify regulatory effects at GWAS risk loci
POEMColoc	https://github.com/AbbVie-ComputationalGenomics/POEMColoc	https://pubmed.ncbi.nlm.nih.gov/34000989	Colocalization	R	An approximation to the coloc method that can be applied when limited summary statistics are available
HyPrColoc	https://github.com/cnfoley/hyprcoloc	https://pubmed.ncbi.nlm.nih.gov/33536417	Colocalization	R	An efficient deterministic Bayesian algorithm that can detect colocalization across vast numbers of traits simultaneously
SparkINFERNO	https://bitbucket.org/wanglab-upenn/SparkINFERNO	https://pubmed.ncbi.nlm.nih.gov/32330239	Colocalization	Python	A scalable bioinformatics pipeline characterizing non-coding GWAS. It prioritizes causal variants underlying association signals and reports relevant regulatory elements, tissue contexts and plausible target genes
ColocQuiaL	https://github.com/bvoightlab/ColocQuiaL	https://pubmed.ncbi.nlm.nih.gov/35894642/	Colocalization	R	Pipeline that provides a framework to perform eQTL and sQTL colocalization analyses at scale across the genome with COLOC
MSG	https://github.com/yingji15/MSG_public	https://pubmed.ncbi.nlm.nih.gov/35771864	Colocalization	R	A multidimensional splicing gene approach
LocusFocus	https://locusfocus.research.sickkids.ca/	https://pubmed.ncbi.nlm.nih.gov/33090994/	Colocalization	web	A web-based tool that tests colocalization using the Simple Sum method to identify the most relevant genes and tissues for a particular locus in the presence of high LD and/or allelic heterogeneity
SharePro	https://github.com/zhwm/SharePro_coloc	https://www.biorxiv.org/content/10.1101/2023.07.24.550431v1	Colocalization	Python	Takes an effect group-level approach to integrate LD modelling and colocalization assessment to account for multiple causal variants in colocalization analysis
pwCoCo	https://github.com/jwr-git/pwcoco	https://pubmed.ncbi.nlm.nih.gov/32895551/	Colocalization	Python	A fast tool that integrates methods from GCTA-COJO and the coloc R package
ezQTL	https://analysistools.cancer.gov/ezqtl/#/home	https://pubmed.ncbi.nlm.nih.gov/35643189/	Colocalization	web/R	Performs data quality control for variants matched between different datasets, LD visualization, and two-trait colocalization analyses using two state-of-the-art methodologies (eCAVIAR and HyPrColoc)
PESCA	https://github.com/huwenboshi/pesca	https://pubmed.ncbi.nlm.nih.gov/32442408	Colocalization	C/C++	Uses ancestry-matched estimates of LD to infer genome-wide proportions of population-specific and shared causal variants for a single trait in two populations. These estimates are then used as priors in an empirical Bayes framework to localize and test for enrichment of population-specific/shared causal variants in regions of interest
LLR	https://github.com/gordonliu810822/LLR	https://pubmed.ncbi.nlm.nih.gov/28961754	Colocalization	Matlab	A latent low-rank approach to colocalizing genetic risk variants in multiple GWAS and phenotypes
SS2	https://github.com/FanWang0216/SimpleSum2Colocalization	https://pubmed.ncbi.nlm.nih.gov/35065708	Colocalization	R	Integrates GWAS summary statistics with eQTL summary statistics across any number of gene-by-tissue pairs, is applicable when there are overlapping participants in the two studies

Conclusions

- Summary statistics offer protection of privacy over IPD, as well as significant advantages in computational cost, which does not scale with the number of individuals in the study.
- Naturally, in the post-GWAS era it is expected that a large number of methods would be developed to perform analysis using the summary results of GWAS. The particular methods, integrating data from multiple sources such as LD, gene expression and biological pathways, aim to provide biological insight and improve our understanding about the functional role of identified variants.

- We categorized the tools and databases by their functionality, in categories related to data, single-trait analysis, and multiple-trait analysis, along with their sub-categories which we analyzed and reviewed. We also compared the tools and databases based on their features, limitations, and user-friendliness. Our review identified a wide range of tools, each with unique strengths and limitations. We provided descriptions of the key features of each tool and database, including their input/output formats, data types, and computational requirements. We also discussed the overall usability and applicability of each tool for different research scenarios.

- among the tools we identified the majority are written in R and Python, but only a handful is available as a webserver: ten of the tools for GSA, three tools for colocalization, two tools for meta-analysis, and one for pleiotropy analysis, MR and fine-mapping.
- Of course, several of the secondary databases we identified also provide the functionality of performing the analysis using data provided by the user (webTWAS, TSEA-DB, PCGA), but even counting these the proportion of web-tools is rather low (<10%). Web servers and web services have become of high relevance to the field of bioinformatics during the last 20 years, so it is expected to have an increasing number of relevant webservers in the near future as relevant tools are available to facilitate the incorporation of existing applications.
- On the other hand, some tools may be too computationally demanding, so other solutions must be found. Container-based applications such as Docker can simplify maintenance procedures and add to the reproducibility of research. Community efforts such as udocker may promote usability of complex software tools by non-experts in multi-user environments

- As data accumulates it is unavoidable to head to analyses on an even larger scale. Traditionally the large-scale analysis of many gene-disease associations is modeled by the so-called diseasome using graph theoretic methods. The gene-disease network is composed of pairwise associations obtained from public databases and is a bipartite network consisting of two separate sets of nodes and the interactions between nodes belonging to the different sets. The projection to the one or the other of the sets may lead to the gene-gene or the disease-disease projected networks that inform us about the associations between members of the same set (for instance, two diseases are connected if they share common genes, and so on).
- Such methods are available for years, but they treat the associations as fixed inputs to the graph. As data accumulate and even more complex statistical methods are developed that allow cross-trait comparisons and combined analyses of multiple traits, along with the integration of different types of data such as xQTL, it is tempting to speculate that a fusion of these two traditions may come, in which the statistical formalism of the tools presented in this review will merge with the graph theoretic approaches developed in the systems biology literature. For instance, we may see network approaches leading to causal analyses (similar to MR) that consider simultaneously all the diseases and traits for which we have GWAS summary data, or similar approaches that integrate xQTL data of various types, different tissues and so on.

References

1. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1):59.
2. MacArthur JAL, Buniello A, Harris LW, Hayhurst J, McMahon A, Sollis E, et al. Workshop proceedings: GWAS summary statistics standards and sharing. Cell Genom. 2021;1(1).
3. Ziegler A, Konig IR, Thompson JR. Biostatistical aspects of genome-wide association studies. Biom J. 2008;50(1):8-28.
4. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev Genet. 2017;18(2):117-27
5. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genet. 2018;102(5):717-30.
6. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11(7):499-511.
7. Hassani-Pak K, Rawlings C. Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes. J Integr Bioinform. 2017;14(1).
8. Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res. 2012;40(9):3777-84.
9. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet. 2013;14(6):379-89.
10. Tang M, Wang T, Zhang X. A review of SNP heritability estimation methods. Brief Bioinform. 2022;23(3).
11. Cinar O, Viechtbauer W. A Comparison of Methods for Gene-Based Testing That Account for Linkage Disequilibrium. Front Genet. 2022;13:867724.
12. Mooney MA, Wilmot B. Gene set analysis: A step-by-step guide. Am J Med Genet B Neuropsychiatr Genet. 2015;168(7):517-27.
13. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat Rev Genet. 2018;19(8):491-504.
14. van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. Nat Rev Genet. 2019;20(10):567-81.
15. Zhang Y, Cheng Y, Jiang W, Ye Y, Lu Q, Zhao H. Comparison of methods for estimating genetic correlation between complex traits using GWAS summary statistics. Brief Bioinform. 2021;22(5).
16. Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. Open Biol. 2017;7(11).
17. Tyler AL, Crawford DC, Pendergrass SA. The detection and characterization of pleiotropy: discovery, progress, and promise. Brief Bioinform. 2016;17(1):13-22.
18. Boehm FJ, Zhou X. Statistical methods for Mendelian randomization in genome-wide association studies: A review. Comput Struct Biotechnol J. 2022;20:2338-51.
19. Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, et al. Opportunities and challenges for transcriptome-wide association studies. Nat Genet. 2019;51(4):592-9.
20. Mai J, Lu M, Gao Q, Zeng J, Xiao J. Transcriptome-wide association studies: recent advances in methods, applications and available databases. Commun Biol. 2023;6(1):899.
21. Zhu H, Zhou X. Transcriptome-wide association studies: a view from Mendelian randomization. Quant Biol. 2021;9(2):107-21.
22. Hukku A, Sampson MG, Luca F, Pique-Regi R, Wen X. Analyzing and reconciling colocalization and transcriptome-wide association studies from the perspective of inferential reproducibility. Am J Hum Genet. 2022;109(5):825-37.