上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

生命科学技术学院
School of Life Sciences and Biotechnology

GWAS and post-GWAS analysis

Sun Jianle

sjl-2017@sjtu.edu.cn

Genetic basis of human diseases

- Single major genes influencing rare Mendelian disorders
- Multiple genes (polygenetic) influencing common complex traits
  - Omnigenic model: a genetic architecture of regulatory networks composed of a small number of core genes that directly affect a trait but with a large number of genes outside the core that indirectly affect the trait.

Linkage disequilibrium (LD)

Definition

If we consider 2 genetically linked variants, each having 2 alleles, we expect to observe 4 pairwise combinations of these alleles, or haplotypes. If the alleles were independent or at linkage equilibrium, their frequencies should be the product of each allele frequency in the population. Likewise, for *n* biallelic variants, we expect to see 2^n haplotypes in the population. In fact, this random association of alleles is rarely observed, especially if the variants are physically close. We rather observe preferential allelic associations (ie, haplotype frequencies deviate from those expected at linkage equilibrium). The difference *D* between the observed and expected frequencies of a haplotype defines the LD of the 2 alleles on this haplotype. If 2 alleles are found together more often than expected, they are positively associated and their LD is positive. Conversely, the alternative combinations of alleles are less frequent than expected with the same absolute LD value but negative.

There are different measures of LD. The most intuitive one for biallelic variants is their correlation estimated by their coefficient of determination r^2 ranging between 0 and 1. The LD is "complete" when at least 1 haplotype is missing and the most extreme case of LD said "perfect" is when only 2 of the 4 haplotypes are observed, with an r^2 of 1. In the latter case, both variants are redundant such that the genotype at the second variant can be imputed from the genotype at the first variant for any individual in the population. This has major consequences for genetic association studies (see main text).

Forces shaping the LD are intimately connected to the history of the population: LD depends on when the variants appeared in the population and how they evolved across the subsequent generations (see below):

1. *Mutations/variations*

Initially, when mutation creates a new allele at a locus near an established variant, there are only 3 haplotypes and therefore the LD is complete.

2. *Meiotic recombination*

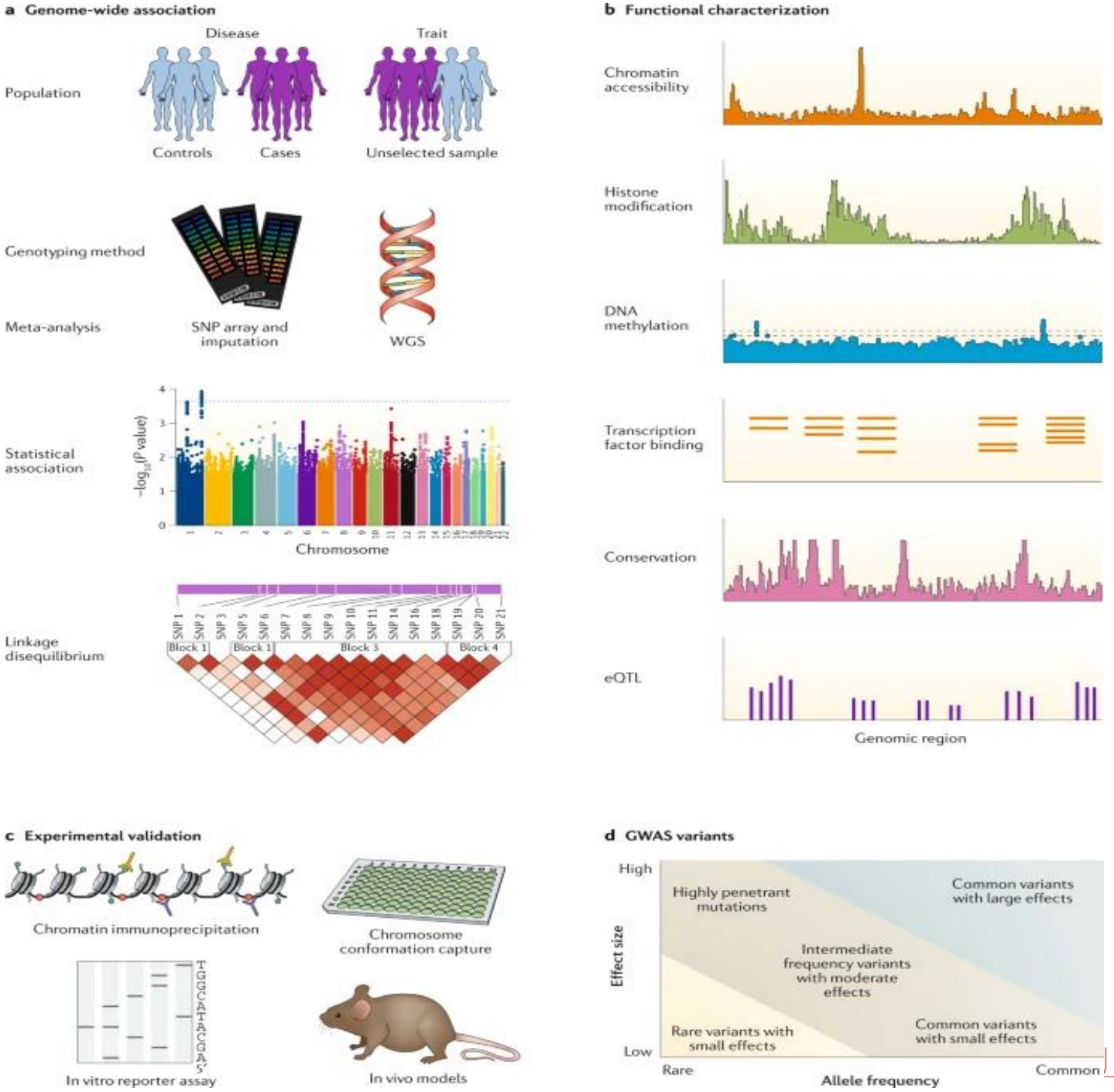
Following mutation, the major force shaping the LD is meiotic recombination that creates the fourth haplotypes and therefore dissipates the LD. The higher the number of generations since the mutation has appeared and the higher the recombination rate between 2 loci, the lower the LD between their alleles. Thus distant loci are generally in lower LD. However, the recombination activity is not homogeneous along the genome: at a small scale, hot-spots of recombination shape islands of LD or haplotype blocks where a smaller number of haplotypes than expected are observed in the population. Their average size is ~20 kb and varies between populations. However, it is important to remind that 2 variants in the same haplotype block are not always in LD and that, conversely, 2 variants may be in LD without being in the same haplotype block.

3. *Demographic and evolutionary forces*

By influencing variant allele frequencies, these forces also strongly impact the LD. For example, in a growing population, the LD decreases by increasing the number of recombinations. On the contrary, the genetic drift can fix some alleles, and thus the corresponding haplotypes, leading to new LD patterns. Migrations and population admixture also modify allele frequencies and LD. In population genetics, the time of a migration can be deduced from the LD decay. Recent positive selection can be detected with a higher LD around the selected variants. Epistatic interactions may also be selected and keep alleles in LD even if they are distant. This explains why LD depends on the considered haplotypes with different patterns for multi-allelic variants. For example, in the highly polymorphic human Major Histocompatibility Complex (MHC), LD may extend over long distances for some specific ancestral haplotypes. Finally, as allele frequencies vary substantially worldwide, both the extent and strength of LD also differ among populations.

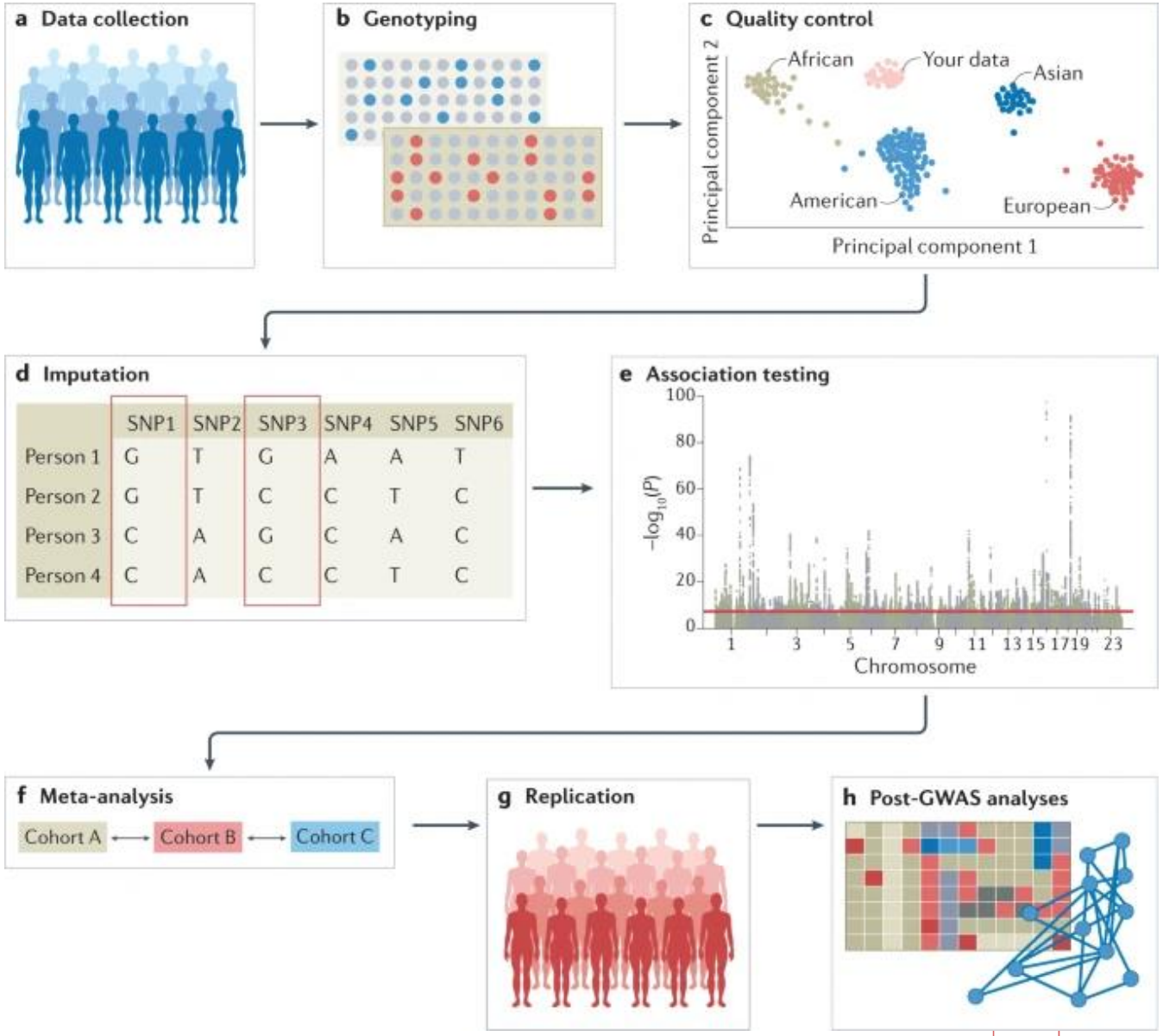
The genotypes of genetic variants are physically close together are not independent as they tend to be in linkage disequilibrium.

Study design



Overview of GWAS

- Genome-wide association studies
- Post-GWAS analysis will be discussed detailly.



Data collection

- Population-based GWAS
  - Cohort study
  - Case-control study: bias in control group
- Family-based GWAS
- Isolated population
- Biobank

- Population stratification & confounding

Table 2 | **Biobanks and large population-based studies with genetic and phenotype data available for research**

Data set	Ancestry
UK Biobank[76]	Predominantly white British
BioBank Japan[267]	Japanese
China Kadoorie Biobank[268]	Chinese
Genes & Health[269]	British South Asian
H3Africa[270]	Various African ancestries
BioMe[105]	Multiple ancestries (based in New York)
TOPMed[22]	Multiple ancestries (USA)
Million Veteran Programme[271]	Multiple ancestries (USA)
'All of Us' initiative[272]	Multiple ancestries (USA)
23andMe	Multiple ancestries (USA)

Genotyping

- Genotyping
  - Microarrays: includes common variants (tag SNPs)
  - Next-generation sequencing: also includes rare variants
    - Whole-genome sequencing (WGS)
    - Whole-exome sequencing (WES)

- Common and rare variants
  - Common Disease, Common Variant (CDCV): Cumulative effect of many common, low penetrance variants
  - Common Disease, Rare Variant (CDRV): Different single, rare, high penetrance variants

Genotyping

- GWAS using SNP arrays versus whole-genome sequencing (WGS)

Factor	SNP arrays	WGS
Cost	Relatively inexpensive (~US$40 per sample)	Expensive (>US$1,000 per sample)
Reliability	Reliable, highly accurate technology	Less mature and less accurate technology
Genomic coverage	• Mainly restricted to common and low-frequency variants, although imputation of rare variants is increasingly accurate (ultra-rare variants, however, can never be identified) • Biased towards variants discovered in well-studied or sequenced populations	From low-frequency, common variants to nearly all genetic variation in the genome, depending on the depth of sequencing
GWAS analysis	Well-established analytical pipeline and tools for data analysis	• Higher computational costs and greater analytical complexity • Eventually, larger multiple testing burden when conducting single-variant tests
Other considerations	Custom genotyping arrays can be extremely cost-effective	• As all variation is ascertained, fine-mapping is easier • Greater costs to store, process, analyse and interpret the resulting data
Suitable research objectives	• Analysing known or candidate associations in large cohorts • Detecting low-frequency, common variant associations in extremely large sample sizes	• Detecting and fine-mapping rare variants • Detecting ultra-rare risk variants when it becomes economically viable to perform WGS at a very large scale

Quality control

- Filtering of bad SNPs
  - Hardy-Weinberg equilibrium
  - Genotype call rate
  - Minor allele frequency: removing monomorphic variants
- Filtering of bad individuals
  - Sex check: ensure that phenotypes are well matched with genetic data (comparing self-reported sex versus sex based on X and Y chromosomes)
  - Genotype call rate
  - Sample call rate
  - Heterozygosity and relatedness checks

Imputation

- Using individual-level data: leverage LD information from a population reference panel.
  - Statistically **phase** individual genotypes (estimating whether genotyped alleles derive from the maternal or paternal allele)
  - Decide whether to use hard calls or weight for uncertainty
  - Select an appropriate **reference population panel**
  - Convert reference panel and target population into the same genomic build
  - Check strand issues, resolve issues between different platforms, possibly remove ambiguous SNPs
  - Check for unusual minor allele frequencies and patterns of linkage disequilibrium between reference panel and target data

Reference population panel

- Commonly used population reference panels

Table 1 | **Commonly used population reference panels**

Reference panel	Number of reference samples	Ancestry of reference samples	Number of variant sites	Indels available	Refs
Icelandic reference panel	15,220	European (Icelandic)	31.1 million	Yes	345
HapMap Project phase 3	1,011	Multi-ethnic	1.4 million	No	346
1000G phase 1	1,092	Multi-ethnic	28.9 million	Yes	347
1000G phase 3	2,504	Multi-ethnic	81.7 million	Yes	81
UK10K Project	3,781	European	42.0 million	Yes	348
HRC	32,470	Predominantly European (includes the 1000G reference panel samples)	40.4 million	No	71
TOPMed[a]	62,784	Multi-ethnic	463.0 million	Yes	85

1000G, 1000 Genomes; HRC, Haplotype Reference Consortium; indels, insertions or deletions; TOPMed, Trans-Omics for Precision Medicine. [a]Figures are based on the latest status of the reference panel.

Association testing (Case-control)

- Allele counting to test for association
  - Fisher's exact test
  - Pearson's χ^2 test
  - Fisher's Exact Test
  - Continuous response: Armitage's trend test, ANOVA, t-test, etc.
- Models
  - Allele: G vs T
  - Dominance: GG+GT vs TT
  - Recessive: GG vs GT+TT
  - Additive or co-dominant model: GG vs GT vs TT

Odds ratio

- Genotype frequencies

	GG	GT	TT	Total
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

- Allele frequencies

Observed allele counts

	G	T	Total
Cases	$2r_0+r_1$	r_1+2r_2	$2R$
Controls	$2s_0+s_1$	s_1+2s_2	$2S$
Total	$2n_0+n_1$	n_1+2n_2	$2N$

Expected allele counts

G	T
$2R(2n_0+n_1)/(2N)$	$2R(n_1+2n_2)/(2N)$
$2S(2n_0+n_1)/(2N)$	$2S(n_1+2n_2)/(2N)$

Association testing

- Regression models
  - Linear or logistic regression models depends on the phenotype (continuous or binary)
  - **Covariates** are included to account for stratification and avoid confounding effects
    - Control of population stratification: Genomic principal components (PCs) as covariates
  - Including an additional **random effect term** (individual specific) to account for **genetic relatedness** among individuals

generalized linear model (GLM)

$$g(\mu) = \sum_j \beta_j X_j + uG + \sum_k \gamma_k PC_k$$

linear mixed model (LMM)

$$Y = \sum_j jX_j + uG + \sum_k v_k Z_k + \varepsilon,$$

$$\boldsymbol{Y} \sim \boldsymbol{W\alpha} + \boldsymbol{X}_s\boldsymbol{\beta}_s + g + e$$

$$g \sim N(0, \sigma^2_{\mathrm{A}}\boldsymbol{\psi})$$

$$e \sim N(0, \sigma^2_e\boldsymbol{I})$$

Let Y_i be the phenotype for individual i
$Y_i = 0$ for controls
$Y_i = 1$ for cases

Let X_i be the genotype of individual i at a particular SNP
TT $X_i = 0$
GT $X_i = 1$
GG $X_i = 2$

Association testing

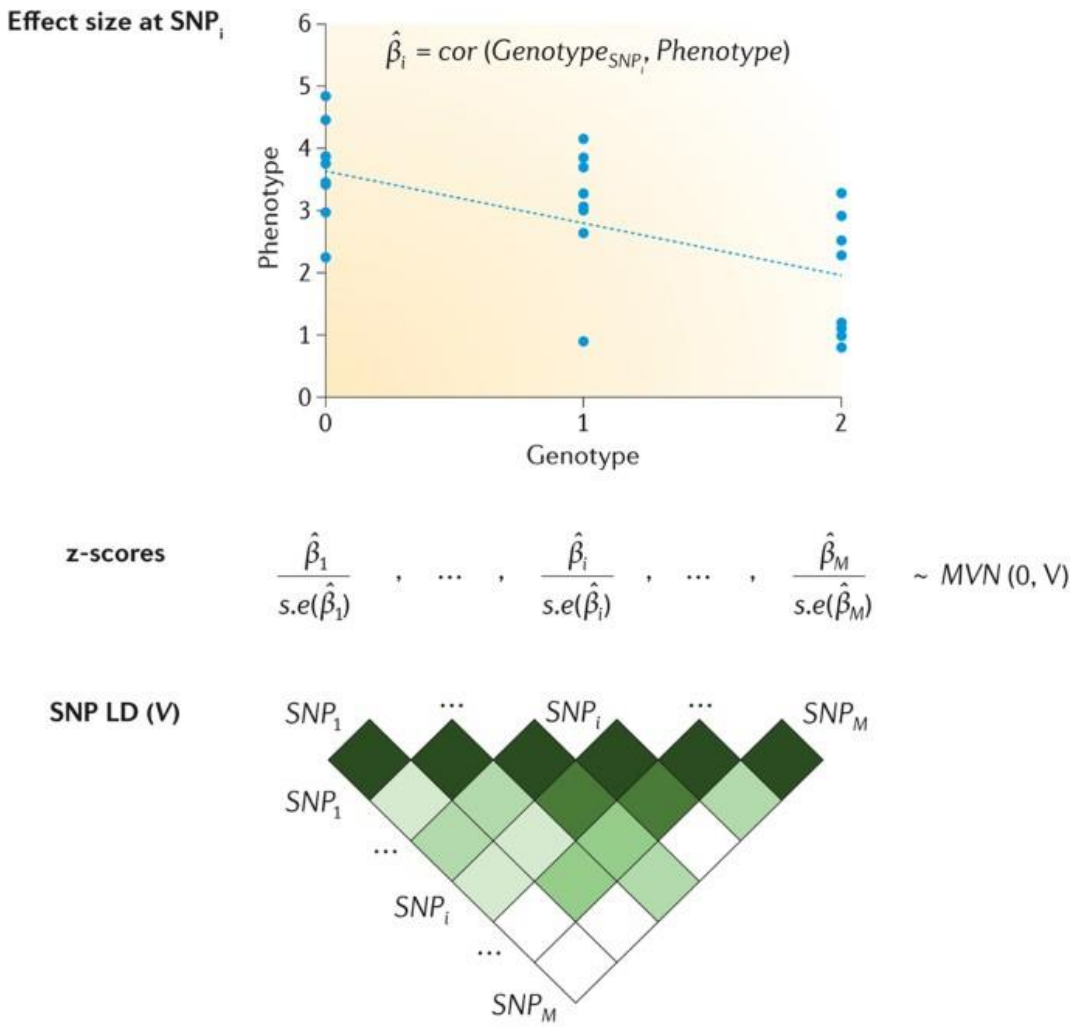
- Other methods:
  - Gene-based association analysis, especially for rare variants
    - regressing the disease status or trait value on the principal components of the SNP genotypes
    - Kernel-based methods such as SKAT
    - use summary statistics from individual SNPs (marginal *p*-values without individual genotype data) to derive gene-based tests: VEGAS and GATES
  - Pathway-based (gene set) analysis
  - Topology-based analysis: MAGMA
  - Incorporating single-nucleotide polymorphism annotations
  - Joint GWAS analysis from multiple traits
  - SNP-SNP and SNP-environment interactions

Association testing: accounting for false discovery

- A stringent multiple-testing threshold
  - There are millions of associations to be tested
  - Bonferroni testing threshold of $p < 5 \times 10^{-8}$
  - Depends on population size and minor allele frequency
- Winner's curse
  - The effect sizes of newly discovered alleles tend to be overestimated
- Comparing effect sizes between discovery and independent replication cohorts.

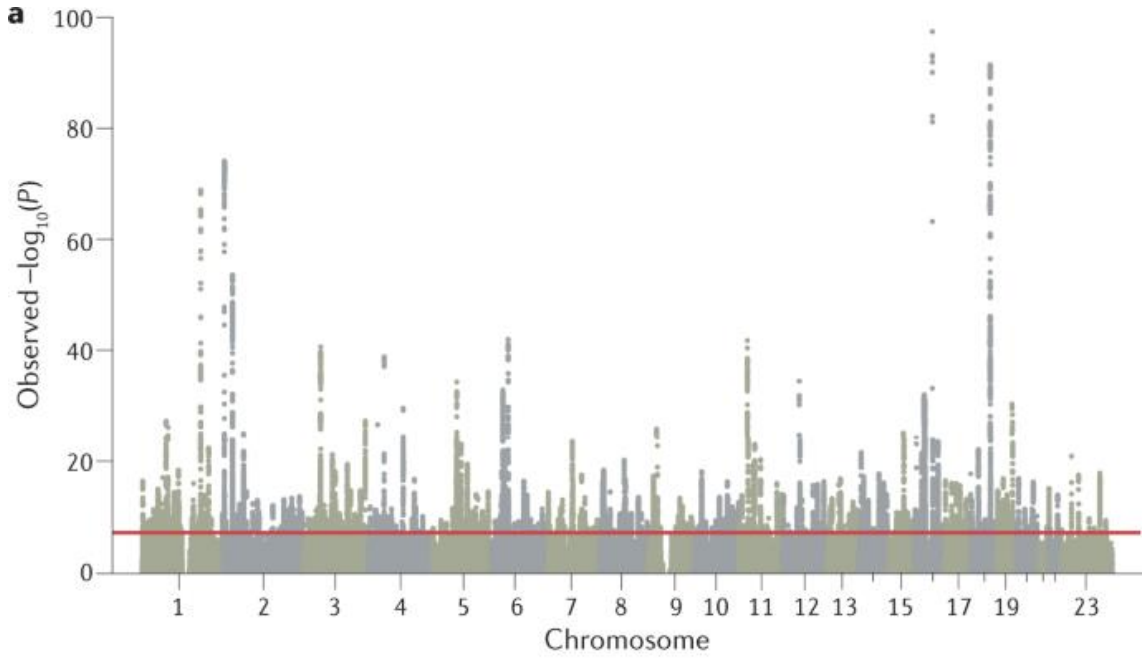
Association testing: GWAS summary statistics

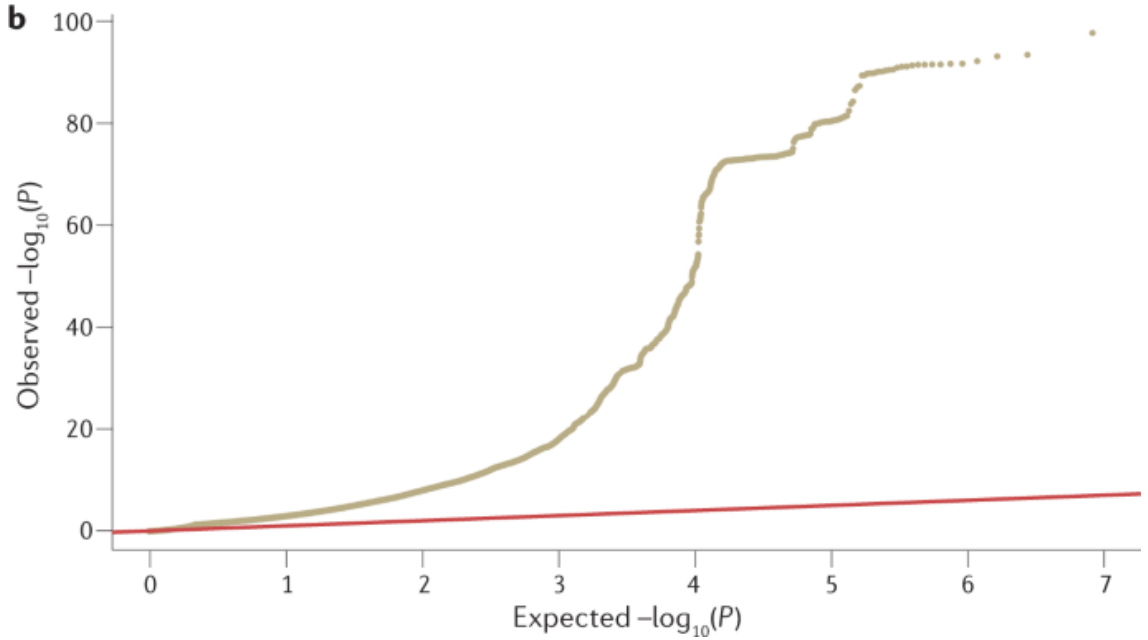
- The results of association testing: GWAS summary statistics
  - Effect sizes
  - Standard errors
  - Linage disequilibrium (LD) matrix



Association testing: visualizing

- Manhattan plots
- Quantile-quantile plots





Imputation of summary statistics

- Imputation using summary statistics and LD information from a population reference panel

Box 1 | **Conditional association and imputation from summary statistics**

Let X be an $N \times M$ matrix of genotypes, standardized to mean 0 and unit variance, and Y be an $N \times 1$ vector of standardized trait values, where M is the number of single nucleotide polymorphisms at the locus and N is the number of samples. Under a standard linear model, $Y = X\beta + \epsilon$. Let V be an $M \times M$ linkage disequilibrium (LD) matrix of pairwise LD; V is equal to X^TX if individual-level data are available but can otherwise be estimated from a population reference sample (with or without regularization).

Conditional association using LD reference data

We estimate the joint effects of all SNPs using least-squares as $\hat{\beta} = V^{-1}X^TY$ with $var(\hat{\beta}) = \sigma_J^2 V^{-1}$, where σ_J^2 is the residual variance in the joint analysis. However, in a standard genome-wide association study, each SNP is marginally tested one at a time, which can be expressed in matrix form as $\hat{\beta}_M = D^{-1}X^TY$ with $var(\hat{\beta}_M) = \sigma_M^2 D^{-1}$, where D is the (nearly constant) diagonal matrix of V and σ_M^2 is the residual variance in the marginal analysis. It follows that

$$\hat{\beta} = V^{-1} D\hat{\beta}_M$$

$$var(\hat{\beta}) = \sigma_J^2 V^{-1}$$

Summary statistic imputation using LD reference data

Let

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

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.

Resources of GWAS summary statistics

- Databases

Table 3 | **Databases of GWAS summary statistics**

Database	Content
GWAS Catalog[110]	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas[8]	UK Biobank GWAS summary statistics
Pan UKBB	UK Biobank GWAS summary statistics
GWAS Atlas[273]	Collection of publicly available GWAS summary statistics with follow-up in silico analysis
FinnGen results	GWAS summary statistics released from FinnGen, a project that collected biological samples from many sources in Finland
dbGAP	Public depository of National Institutes of Health-funded genomics data including GWAS summary statistics
OpenGWAS database	GWAS summary data sets
Pheweb.jp	GWAS summary statistics of Biobank Japan and cross-population meta-analyses

For a comprehensive list of genetic data resources, see REF.[13]. GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.

Resources of GWAS summary statistics

- GWAS summary statistics for various traits

Table 1 | **Publicly available summary association statistics***

Trait	N	URL	Ref.
Age at menarche	127,884	http://www.reprogen.org/	119
Alzheimer disease	54,162	http://www.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php	120
Bone mineral density	53,236	http://www.gefos.org/?q=content/data-release-2015	121
Body mass index	122,033	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	122
Body mass index‡	322,154	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	9
Coronary artery disease	77,210	http://www.cardiogramplusc4d.org/	123
Crohn's disease	20,883	http://www.ibdgenetics.org/downloads.html	124
Crohn's disease‡	51,874	http://www.ibdgenetics.org/downloads.html	125
Depressive symptoms	161,460	http://www.thessgac.org/data	126
Ever smoked	74,035	http://www.med.unc.edu/pgc/downloads/	127
Fasting glucose	58,074	http://www.magicinvestigators.org/downloads/	128
HbA_{1c} (glycated haemoglobin)	46,368	http://www.magicinvestigators.org/downloads/	129
High-density lipoprotein	97,749	http://www.broadinstitute.org/mpg/pubs/lipids2010/	130
High-density lipoprotein‡	188,577	http://csg.sph.umich.edu//abecasis/public/lipids2013/	131
Height	131,547	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	132
Height‡	253,288	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	8
Hip circumference	213,038	http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	133
Irritable bowel syndrome (Crohn's disease or ulcerative colitis)	34,652	http://www.ibdgenetics.org/downloads.html	124

Resources of GWAS summary statistics

- GWAS summary statistics for various traits

Irritable bowel syndrome (Crohn's disease or ulcerative colitis)‡	65,643	http://www.ibdgenetics.org/downloads.html	125
Low-density lipoprotein	93,354	http://www.broadinstitute.org/mpg/pubs/lipids2010/	130
Low-density lipoprotein‡	188,577	http://csg.sph.umich.edu//abecasis/public/lipids2013/	131
Neuroticism	170,911	http://www.thessgac.org/data	126
Rheumatoid arthritis (Europeans)	38,242	http://plaza.umin.ac.jp/~yokada/datasource/software.htm	134
Rheumatoid arthritis (Europeans) ‡	58,284	http://plaza.umin.ac.jp/~yokada/datasource/software.htm	134
Rheumatoid arthritis (East Asians)	22,515	http://plaza.umin.ac.jp/~yokada/datasource/software.htm	134
Schizophrenia	70,100	http://www.med.unc.edu/pgc/downloads/	135
Subjective well-being	298,420	http://www.thessgac.org/data	126
Triglycerides	94,461	http://www.broadinstitute.org/mpg/pubs/lipids2010/	130
Triglycerides‡	188,577	http://csg.sph.umich.edu//abecasis/public/lipids2013/	131
Type 2 diabetes	60,786	http://diagram-consortium.org/	136
Ulcerative colitis	27,432	http://www.ibdgenetics.org/downloads.html	124
Ulcerative colitis‡	47,746	http://www.ibdgenetics.org/downloads.html	125
Waist circumference	232,101	http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	133
Waist/hip ratio	212,248	http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files	133
Years of education	328,917	http://www.thessgac.org/data	126

*We provide a selected list of publicly available summary statistics from genome-wide association studies with sample sizes of at least 20,000. A more complete list can be found in REF. 137. ‡Includes specialty genotyping array data; not suitable for analysis using linkage disequilibrium score regression and its extensions.

Meta-analysis & Mega-analysis

- Combining data from different studies
  - Summary association statistics: meta-analysis
  - Individual-level data: mega-analysis
- Fixed effects meta-analysis
  - Assuming that true effect size are the same across studies
- Random effects meta-analysis
  - Assuming that true effect size may differ across studies
- Subset-based meta-analysis
  - Evaluating all possible combinations of non-null models for association, selecting the strongest association and adjusting for the multiple comparisons.

Post-GWAS analysis

- Fine-mapping
- Functional inference
  - Determining the affected gene
  - Determining regulatory pathways and cellular effects
- Polygenicity analysis of complex traits
  - Polygenic risk prediction
  - Understanding trait genetic architecture
- Cross-trait analysis

Fine-mapping

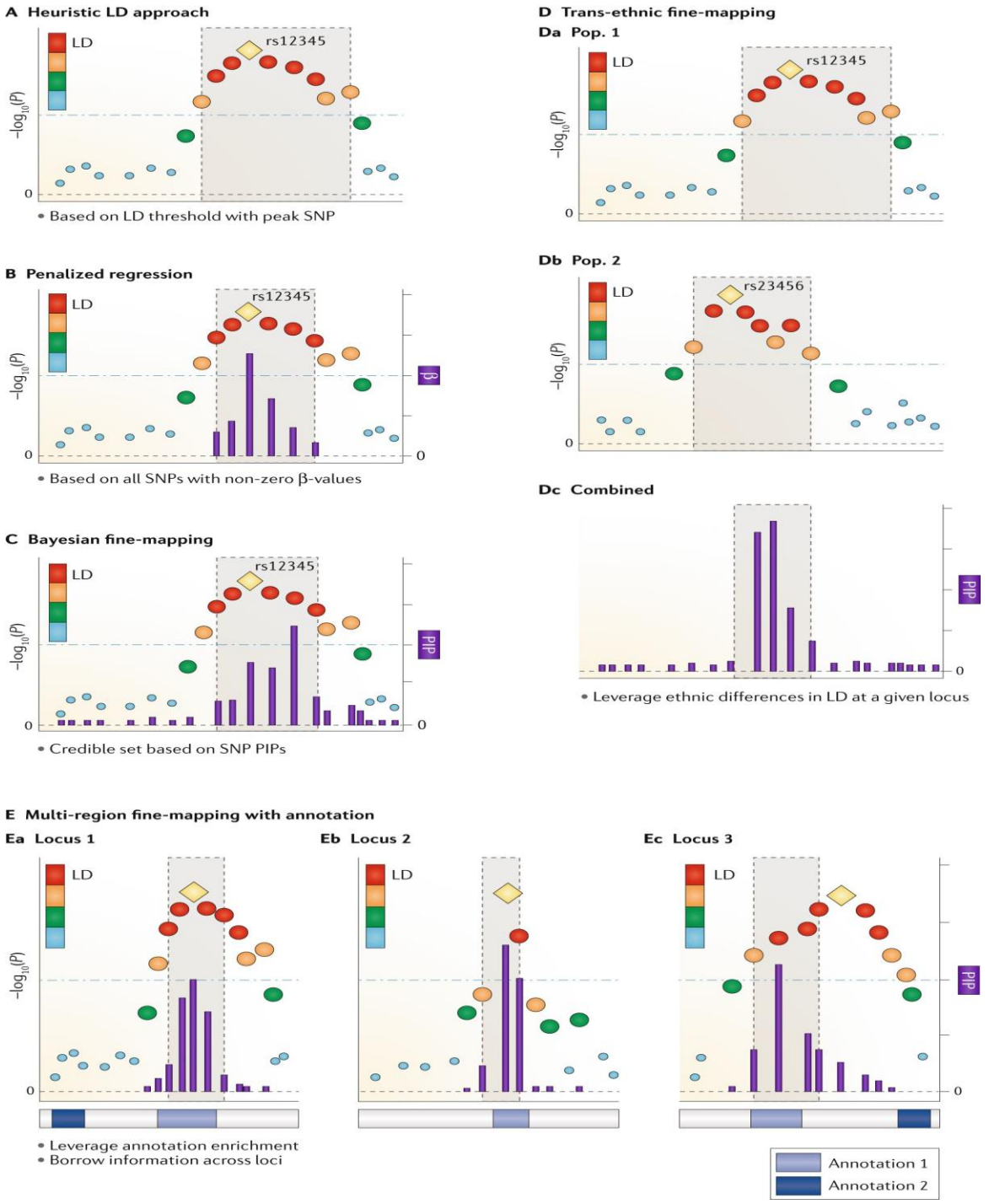
- To identify the causal variant(s) that is driving a GWAS association signal
  - Many non-causal variants are significantly associated with a trait of interest owing to linkage disequilibrium.
  - The most significant association may be non-causal.
- SNP to gene mapping
  - Find credible variants that modulate the expression patterns and functions of causal genes.
- SNP to biology mapping
  - Find credible variants that contribute to the development of the target phenotype.

Fine-mapping

- Using posterior probabilities
  - Prioritize the variants based on the strength of marginal association statistics
  - Conditional association analysis
- Conditional association analysis
  - The association between a SNP and a trait is evaluated after conditioning on the top SNP at a locus (including the lead variant as a covariate in genotype-phenotype regression model)
  - Stepwise conditional analysis: forward stepwise selection
  - Using individual-level data or only using summary statistics with LD information from population reference panel

Fine-mapping

- Fine-mapping strategies
  - Heuristic fine-mapping approaches
    - Filter SNPs according to pairwise correlation (r^2) with the lead SNP
    - Pairwise LD among SNPs within haplotypes
  - Penalized regression models
  - Bayesian methods
    - Posterior inclusion probability (PIP)
      - The sum of the posteriors over all models that include SNP j as causal
    - Credible sets



Bayesian fine-mapping

- The effect size of the causal SNP on a trait (multiple regression R^2)
- The sample size (N)
- Assume one causal SNP and m non causal SNPs
- All SNPs are equally correlated with correlation ρ
- The posterior probability for a causal SNP can be expressed as

$$post_c = \frac{pr_c}{pr_c + \sum_{i=1 i \neq c}^{m} pr_i \cdot \exp\{-(1-\rho)NR^2/(1-R^2)\}}$$

Bayesian fine-mapping procedure

Table 1 | Commonly used Bayesian fine-mapping software

Software	Trait type[a]	Input covariates[b]	Uses summary statistics?	Maximum number of causal variants[c]	Input annotation?	Causal search	Main output	Refs
BIMBAM v1.0	qt and binary	No	No	Fixed	No	Exhaustive	Bayes factor	113,114
mvBIMBAM v1.0.0	mqt	No	Yes	1	No	Exhaustive	Bayes factor	115,116
SNPTEST v2.5.4-beta3	qt, binary, mqt and multinomial	No	No	1	No	Exhaustive	Bayes factor	117
piMASS v0.9	qt and binary	No	No	Computed	No	MCMC	Bayes factor and PIP	65
BVS v4.12.1	Binary	Yes	No	Computed	Yes	MCMC	Bayes factor and PIP	63,118,119
FM-QTL	qt	No	No	Computed	Yes	MCMC	Bayes factor and PIP	96
DAP v1.0.0	qt	Yes	Yes	1, fixed and computed	Yes	Exhaustive	Bayes factor and PIP	52
Fine-mapping	Multinomial	Yes	No	Computed	No	Greedy	PIP	98
Trinculo	Multinomial	Yes	No	Computed	No	Greedy	Bayes factor and PIP	98,120
BayesFM	Binary	Yes	No	20	No	MCMC	PIP	98
ABF	qt and binary[d]	Yes	Yes	1	No	Exhaustive	Bayes factor	121
fgwas v0.3.6	qt and binary[d]	No	Yes	1	Yes	Exhaustive	Bayes factor and PIP	95
CAVIAR/eCAVIAR	qt and binary[d]	No	Yes	Fixed	No	Exhaustive	ρ probability confidence set and PIP	46,101
PAINTOR v3.0	qt, binary[d] and mqt	No	Yes	Fixed and computed	Yes	Exhaustive and MCMC	Bayes factor and PIP	54,57,71
CAVIARBF v0.2.1	qt and binary[d]	No	Yes	Fixed	Yes	Exhaustive	Bayes factor and PIP	47,94
FINEMAP v1.1	qt and binary[d]	No	Yes	Fixed	No	Shotgun stochastic search	Bayes factor and PIP	53
JAM in R2BGLIMS v0.1	qt and binary[d]	No	Yes	Fixed and computed	No	Exhaustive and MCMC	Bayes factor and PIP	55

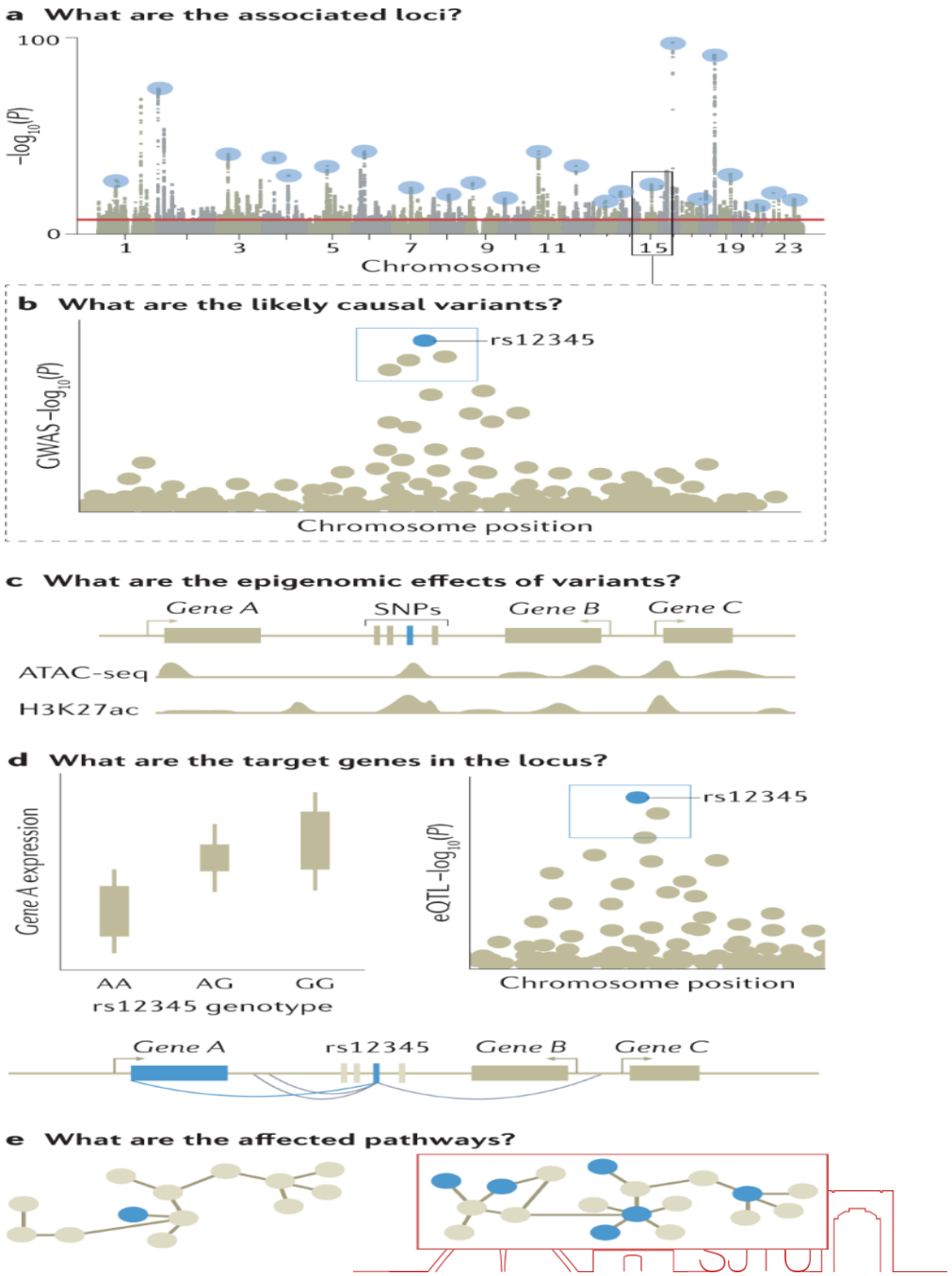
MCMC, Markov chain Monte Carlo; PIP, posterior inclusion probability. [a]Trait types are binary, single binary trait; mqt, multiple quantitative traits; multinomial, trait with more than two categories; and qt, single quantitative trait. [b]For software that does not allow covariates to be input, the traits can be adjusted for covariates by first regressing out the covariates (that is, subtracting trait predicted by covariates from trait values). [c]A fixed number is specified by the user to reduce computational cost. It is usually small (for example, three) when the number of candidate variants is large. When computed, the number of causal variants is determined by the software. As indicated, some software allow different options for whether the maximum number of causal variants is fixed by the user or computed by the software. [d]Application to binary traits is based on linear regression, an approximation assuming small effect sizes and large sample sizes.

Fine-mapping

- Integrating functional annotation data
  - Jointly estimate functional enrichment and update posterior probabilities of causality using functional annotations.
  - Help to understand polygenic architectures by identifying tissue-specific functional annotations.
  - Protein-coding & non-protein-coding annotations
  - Gene expressions
- Trans-ethic fine-mapping
  - Meta-analysis

Functional inference

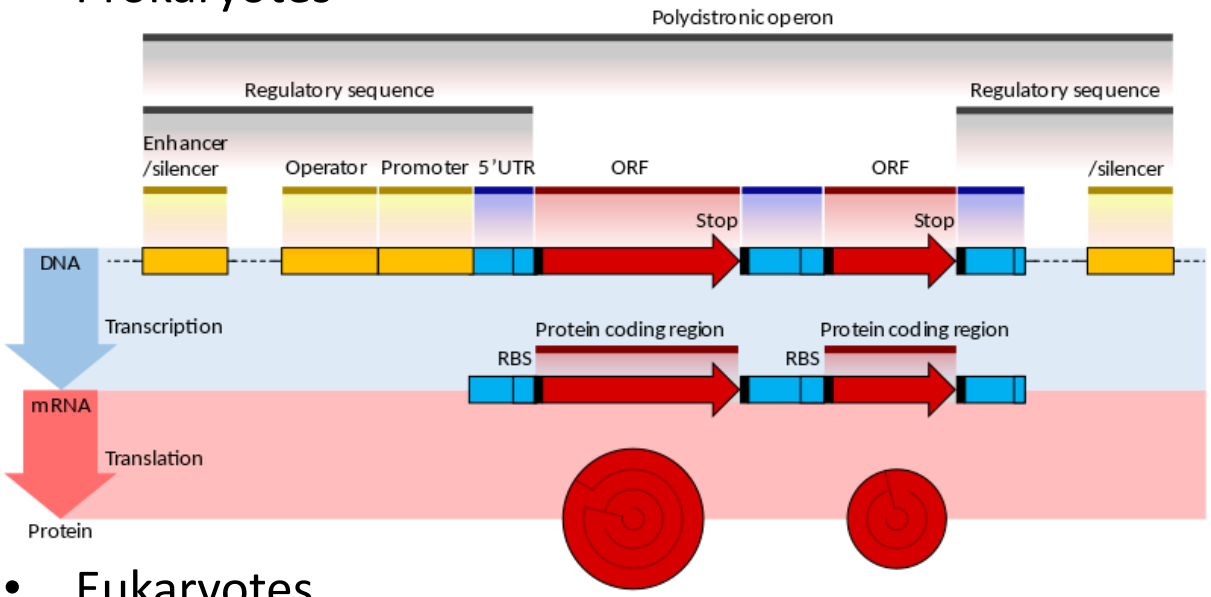
- To identify
  - The immediate effects of causal variants (on protein or enhancer function)
  - The affected gene or genes in the locus that mediate the disease association
  - The downstream network or pathway effects that lead to changes in cellular and physiological function
  - The relevant tissue, cell type and cell state for all these effects



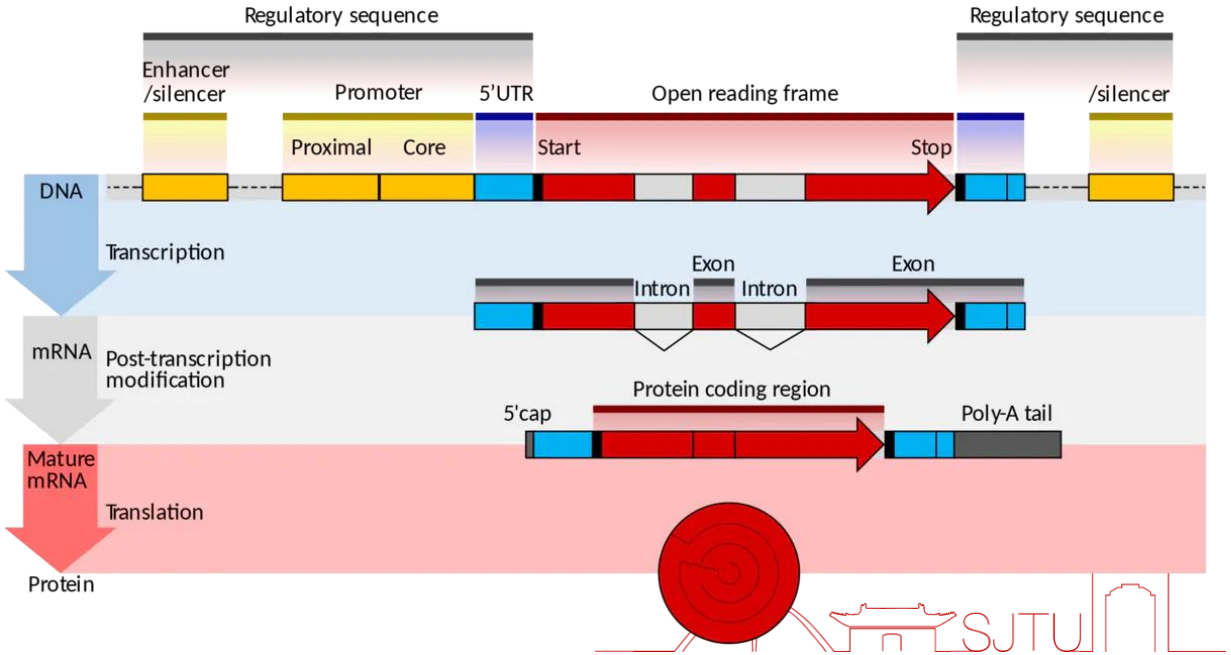
Functional analysis

- Determine the affected gene
  - expression quantitative trait loci (eQTLs)
  - molecular quantitative trait loci (molQTLs) analysis

- Prokaryotes

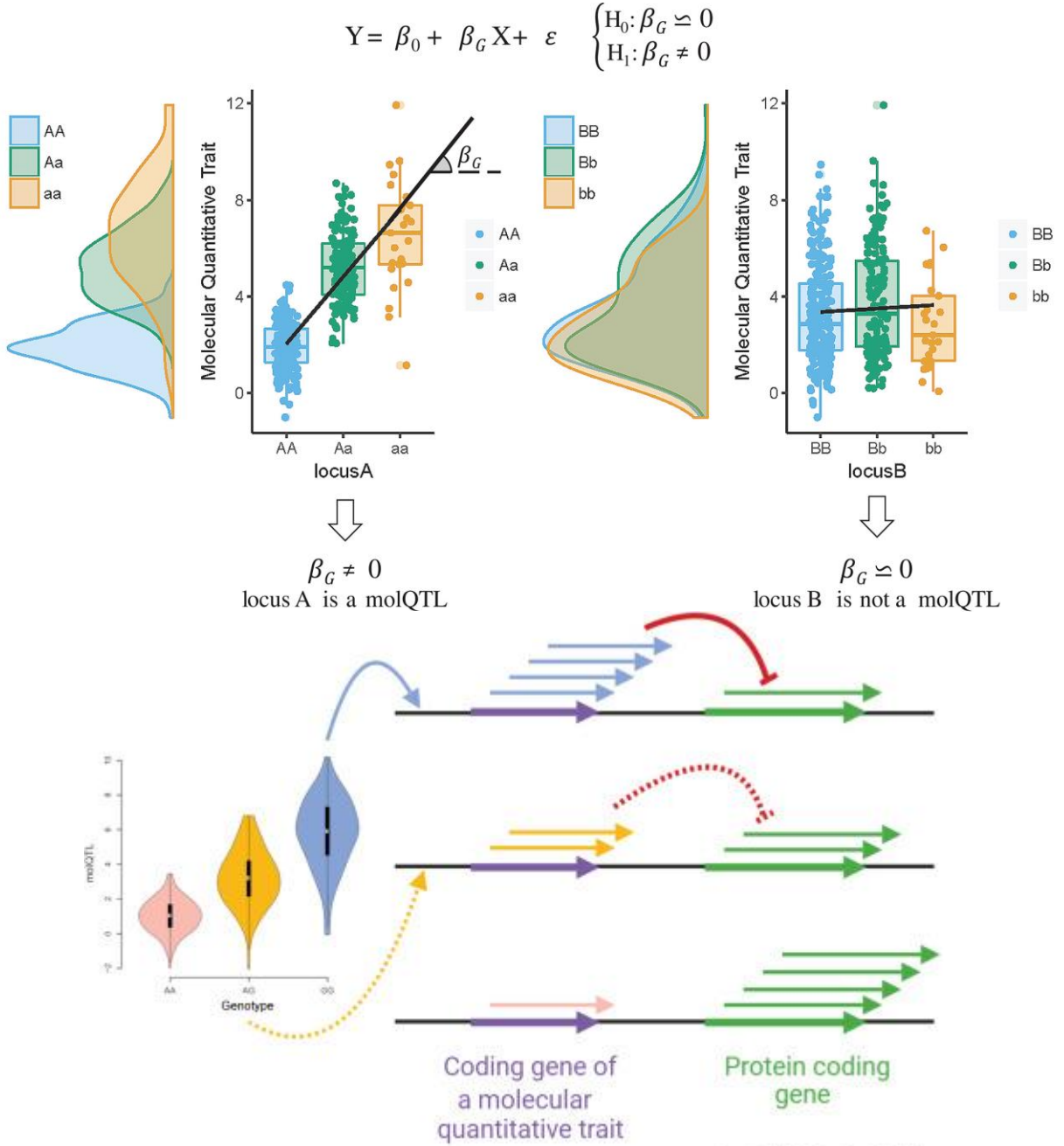


- Eukaryotes



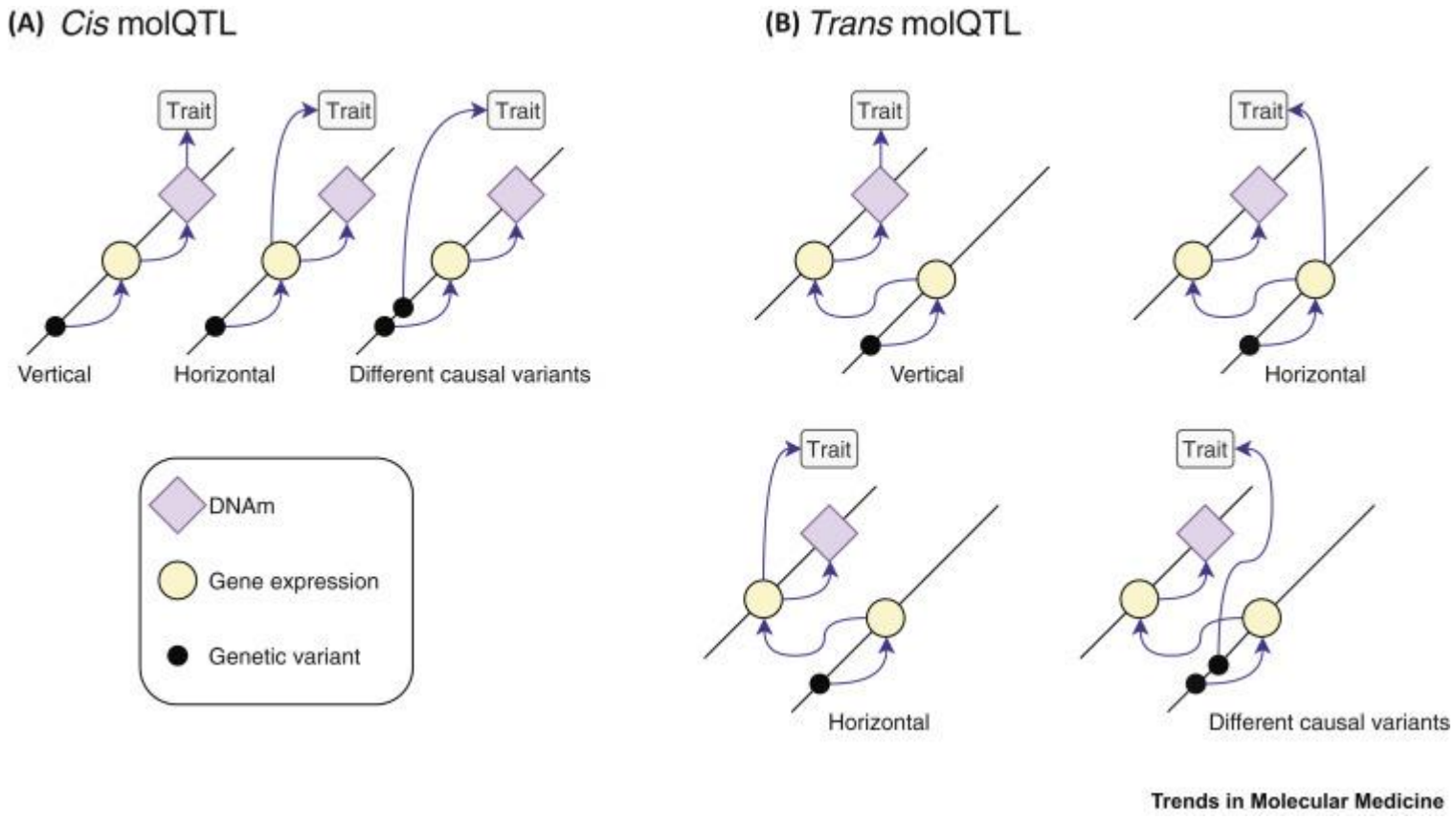
molQTL

- The idea of expression phenotypes (eQTL) can be extended to non-coding genes, to post-transcriptional RNA modifications and to the post-translational level, introducing the general concept of molecular QTLs



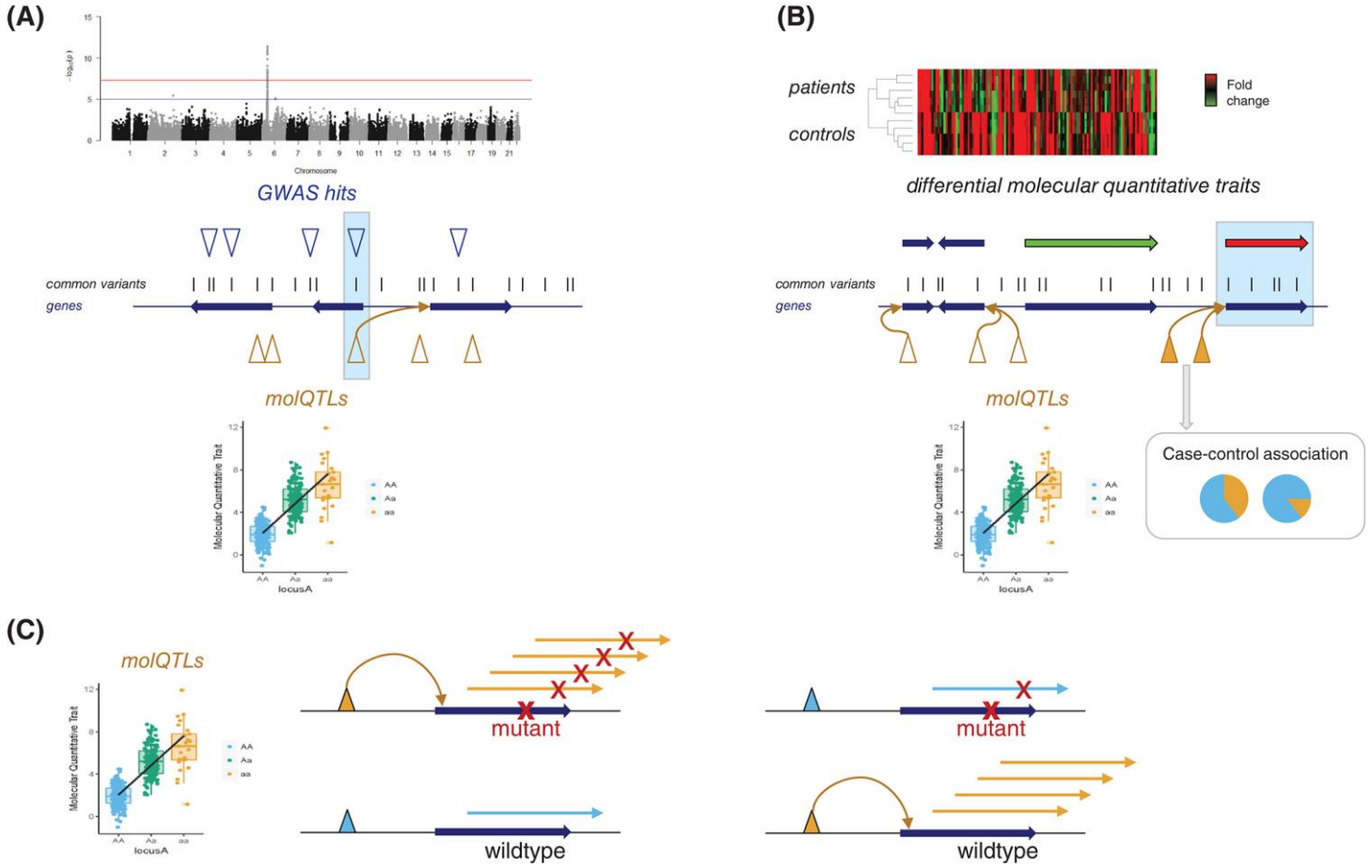
molQTL

- *cis*-acting
  - the regulation of genes within 1 Mb
- *trans*-acting
  - molQTLs affecting genes further away or on different chromosomes



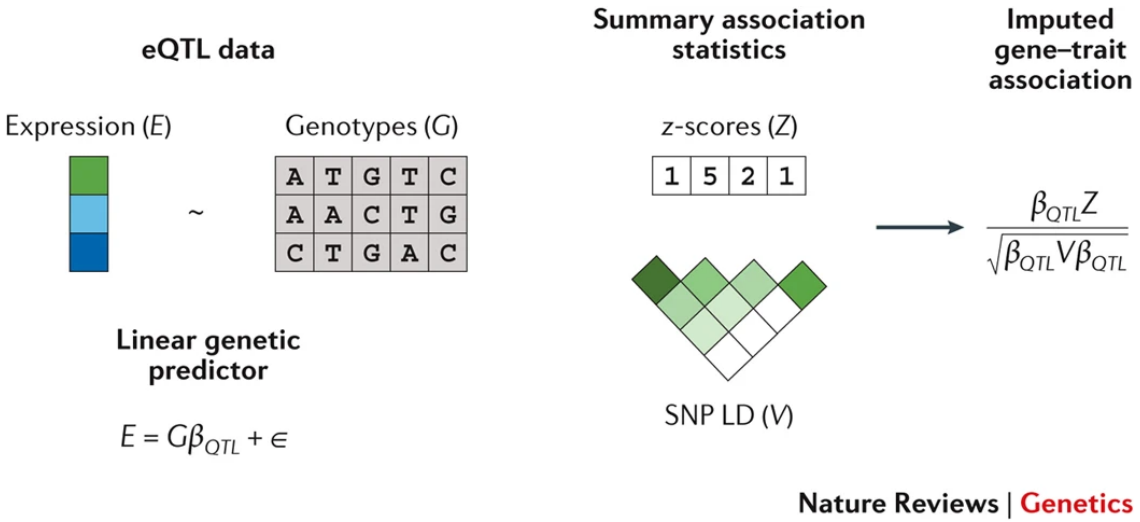
molQTL

- The most credible GWAS variants are prioritized by statistical methods of colocalization with molQTLs.
- Differential molecular traits are filtered by the presence of molQTLs regulating them.
- Modulation of a Mendelian trait by a molQTL.



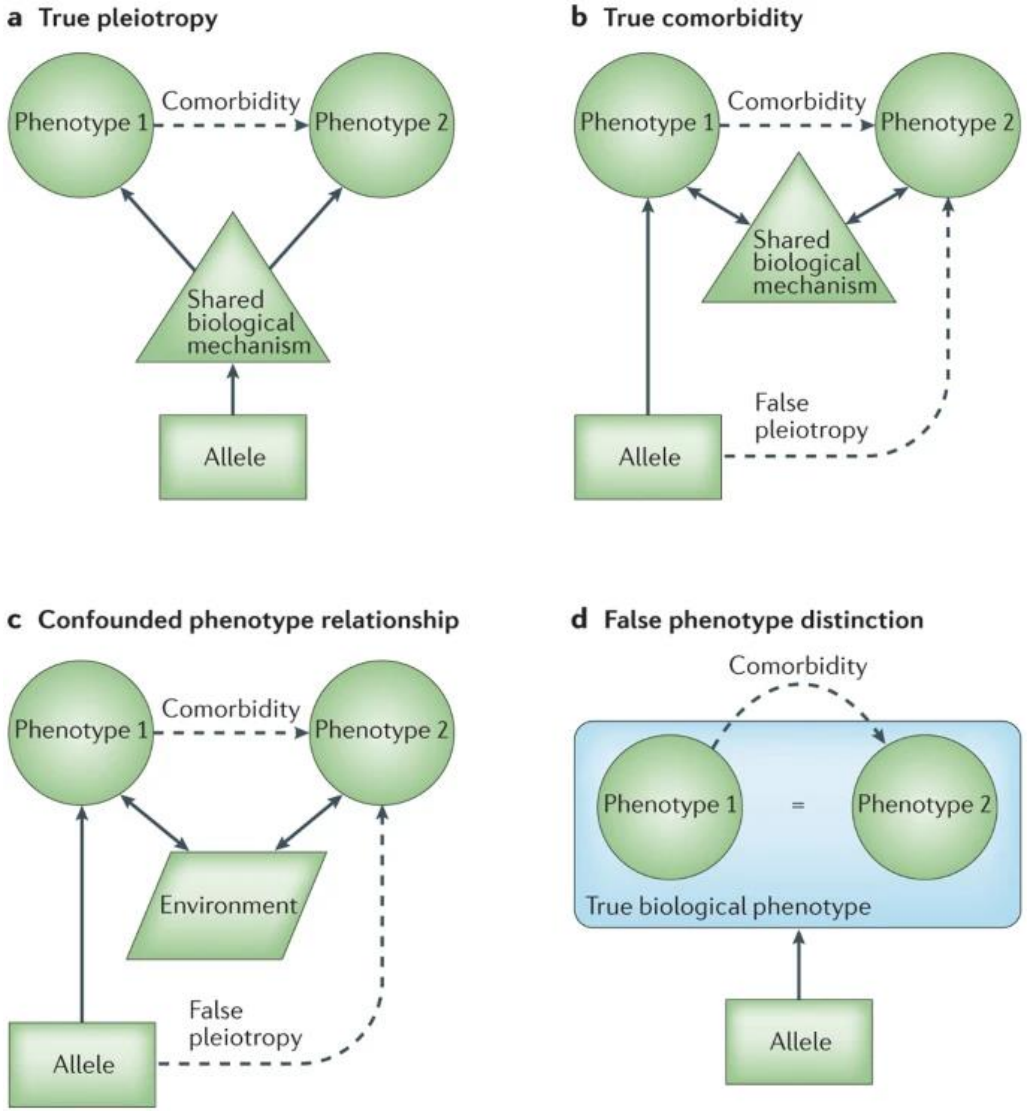
TWAS

- Transcriptome-wide association studies
  - Transcriptome reference data are used to build a linear predictor for gene expression, typically using SNPs from 1 Mb local region around the gene with regularized effect sizes.
  - The 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.



PheWAS

- phenome-wide association studies



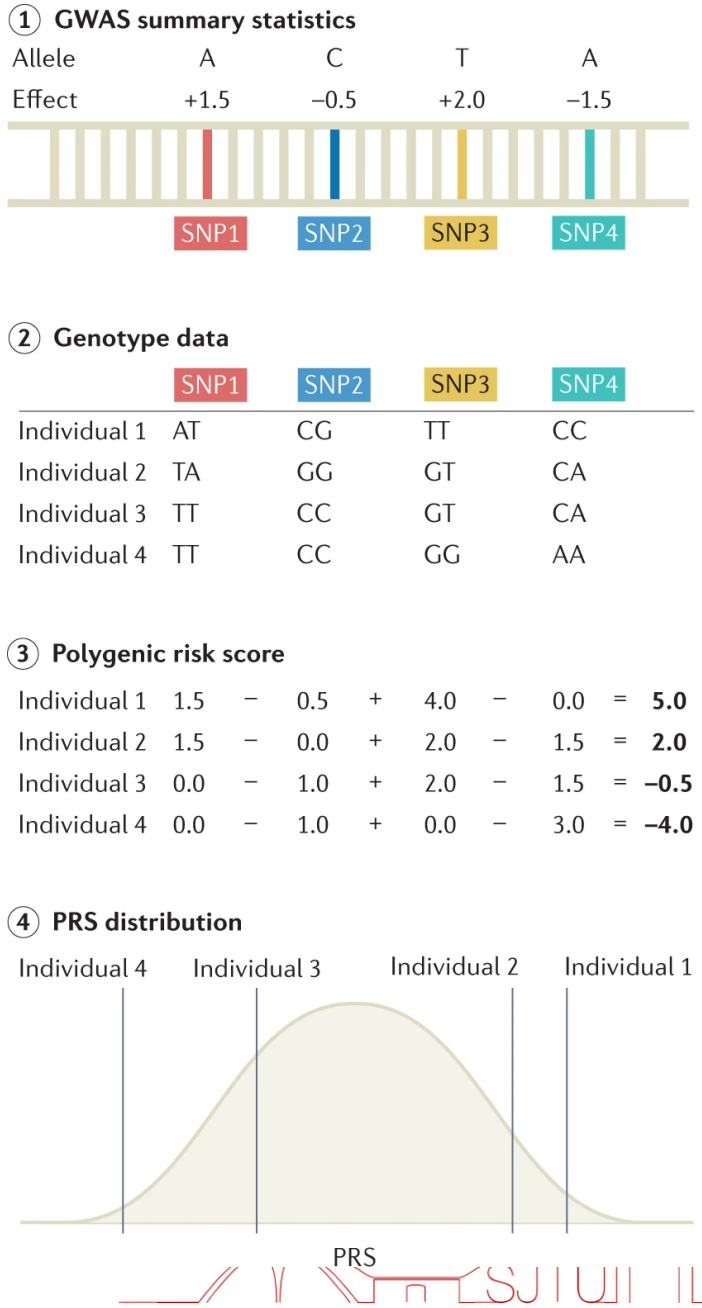
Nature Reviews | Genetics

Polygenic risk prediction

- Polygenic risk scores (PRS): weighted sum scores of risk alleles
  a) Obtain GWAS summary statistics of each SNP: pruning & thresholding
  b) Individuals' genotype data are referenced against GWAS summary statistics
  c) Sum up the effect sizes of all alleles for each individual
  d) Linear regression on PRS to measure the effect of PRS on the outcome

$$H_0 : \text{Phenotype} \sim \text{covariates} + e$$

$$H_1 : \text{Phenotype} \sim \text{PRS} + \text{covariates} + e$$



Polygenic risk prediction

- Fit effect sizes of all markers simultaneously using best linear unbiased prediction (BLUP) methods
- Assume infinitesimal (Gaussian) architectures in which all markers are causal
- Require individual-level training data
- Restrict markers to those below a *p*-value threshold or estimate posterior mean causal effect sizes under a point-normal prior

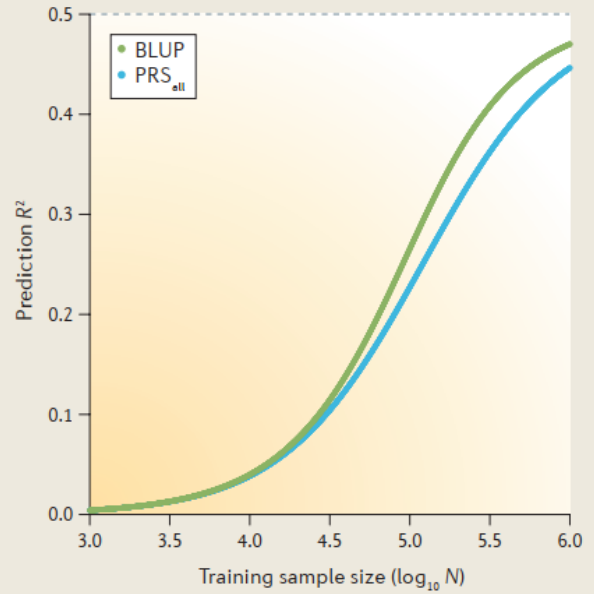
Box 3 | **Polygenic risk prediction using summary versus individual-level data**

Suppose that polygenic risk prediction for a quantitative trait is conducted using a training cohort with N unrelated samples, using M unlinked markers with single nucleotide polymorphism (SNP) heritability[7] equal to h_g^2. We initially consider two polygenic risk prediction methods that assume infinitesimal (Gaussian) architectures: polygenic risk scores computed using marginal effects at all markers with no P value thresholding (PRS_{all}), and fitting effect sizes of all markers simultaneously via best linear unbiased prediction (BLUP). We note that PRS_{all} requires only summary statistics from the training cohort, whereas BLUP requires individual-level data. Prediction accuracy (coefficient of determination; R^2) for each method is given by[83,117]

$$R^2_{PRS_{all}} = \frac{h_g^2}{1 + \frac{M}{Nh_g^2}}$$

$$R^2_{BLUP} = \frac{h_g^2}{1 + \frac{M}{Nh_g^2}(1 - R^2_{BLUP})}$$

These equations can naturally be extended to linked markers (using the effective number of unlinked markers[108]) and case–control traits (using observed-scale SNP heritability[118]). The relative advantage of BLUP over PRS_{all} is small when prediction R^2 is small in absolute terms, but grows larger when prediction R^2 is larger. This effect is illustrated in the figure, which shows prediction R^2 at various training sample sizes based on $M = 60{,}000$ unlinked markers and a SNP heritability of $h_g^2 = 0.5$. These results generalize to non-infinitesimal extensions of polygenic risk scores[75,77] and BLUP[81,82]; in the latter case, the noise reduction from fitting all markers simultaneously remains equal to $1 - R^2$, corresponding to an increase in training sample size of $1/(1 - R^2)$.



Inferring polygenic architectures

- Polygenic architectures of complex traits
  - A large number of causal variants with small effects
- Determining the genetic architecture of a trait involves
  - The number of causal variants
  - Their corresponding effect sizes
  - Allele frequencies
  - Heritability: the proportion of variation in the trait that can be explained by genetic variation in the population
    - Broad-sense heritability (H^2): the fraction of phenotypic variation explained by both additive and dominance effects
    - Narrow-sense heritability (h^2): the fraction of phenotypic variation explained by additive effects only

Inferring polygenic architectures: heritability

- missing heritability: the observation that association signals from early GWAS results often only explain a small proportion of overall heritability
- residual maximum likelihood (REML) estimator
- chip-based heritability: it does not account for variants that cannot be captured by the SNPs on the genotyping platform; it is likely mis-specified since it is unlikely that all the SNPs will contribute to the observed trait

$$Y = X\beta + \varepsilon,$$
$$\beta \sim N\left(0, \frac{h^2}{m} I\right),$$
$$\varepsilon \sim N(0, (1 - h^2)I),$$

Y are the standardized trait values, X is the standardized genotype matrix, β are the genetic effects, h^2 is the overall heritability, and m is the number of SNPs.

Inferring polygenic architectures: heritability

- LD score regression
  - Regressing χ^2 statistics ($\left(\frac{\beta}{SE}\right)^2$) against linkage disequilibrium (LD) scores for each SNP.
  - LD scores are computed as sums of squared correlation of each SNP with all SNPs including itself. (parameters: window size, r^2 cutoff, excluded singletons (MAF))
  - Can distinguish between polygenicity and confounding.
  - Extension:
    - Stratified LD score regression
    - cross-trait LD score regression

$$E[\chi^2 \mid \ell_j] = Nh^2\ell_j / M + Na + 1 \qquad l_j = \sum_{k=1}^{m} r_{jk}^2.$$

N: sample size

l_j : the LD Score of variant j

M: the number of SNPs

h^2/M: the average heritability explained per SNP

a : the contribution of confounding biases

Cross-trait analysis

- Correlation
  - Genetic correlation / cross-phenotype (CP) associations
  - The distinguish between a CP association and (biological) pleiotropy is important to define.
- Causality
  - Mendelian randomization

Genetic correlations between traits

- cross-trait LD score regression

$$Y_1 = X\beta + \varepsilon, \qquad \beta \sim N\left(0, \frac{h_1^2}{m} I\right), \qquad \varepsilon \sim N\left(0, \left(1 - h_1^2\right) I\right),$$

$$Y_2 = Z\gamma + \delta, \qquad \gamma \sim N\left(0, \frac{h_2^2}{m} I\right), \qquad \delta \sim N\left(0, \left(1 - h_2^2\right) I\right),$$

$$E\left(\beta\gamma^T\right) = \frac{\rho_g}{m} I,$$

ρ_g is the genetic covariance between traits Y_1 and Y_2

$$E\left((z_1)_j (z_2)_j\right) = \frac{\sqrt{n_1 n_2}\,\rho_g}{m} l_j,$$

no shared samples between the two GWAS

$$\text{corr} = \frac{\rho_g}{h_1 h_2}.$$

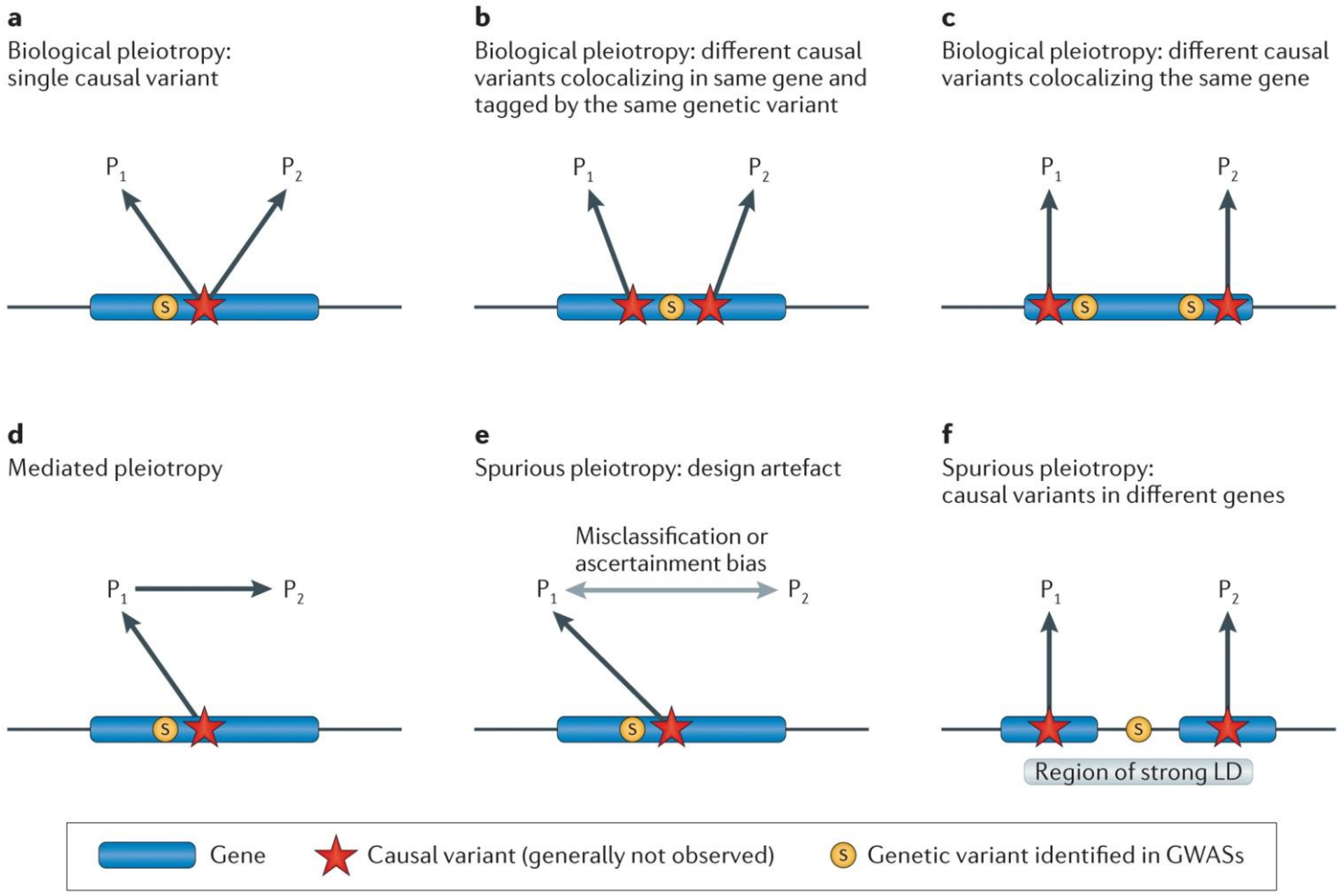
genetic correlation

Pleiotropy

- Horizontal pleiotropy / Biological pleiotropy
  - A genetic variant or gene that has a direct biological influence on more than one phenotypic trait.
- Vertical pleiotropy / Mediated pleiotropy
  - One phenotype is itself causally related to a second phenotype so that a variant associated with the first phenotype is indirectly associated with the second.
- Spurious pleiotropy / LD-induced pleiotropy
  - Two different variants that are in linkage disequilibrium each influence one of two traits.

Pleiotropy

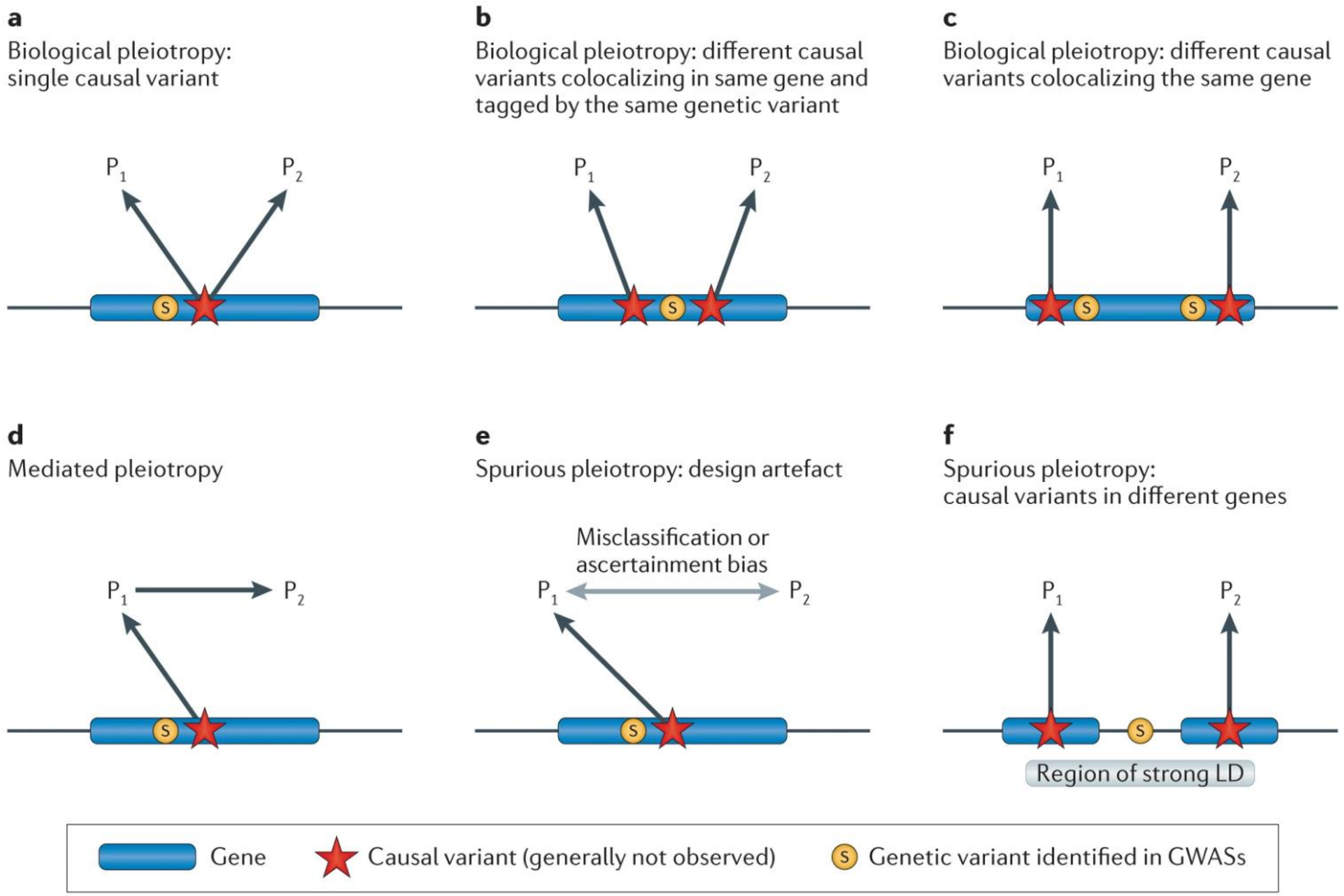
- A CP association can be observed at different levels
- Biological pleiotropy
  - Allelic level: a single causal variant is related to multiple phenotypes
  - Gene or region level: multiple variants in the same gene or region are associated with different phenotypes



Nature Reviews | Genetics

Pleiotropy

- Spurious pleiotropy
  - Defects in studies
    - Ascertainment bias
    - Phenotypic misclassification
    - Shared controls
    - Population stratification
    - Batch effects
  - Linkage disequilibrium

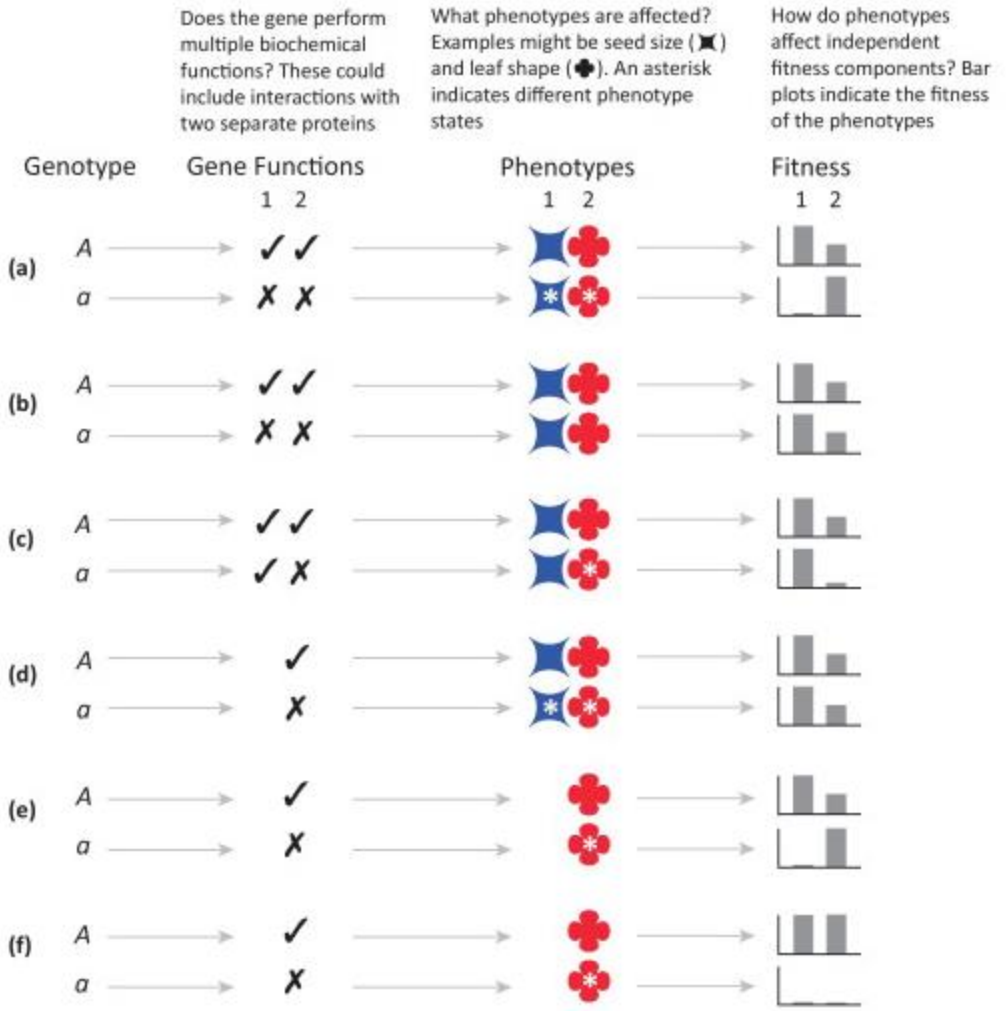


Biological meanings of pleiotropy

- At its essence, pleiotropy implies a mapping from one thing at the genetic level to multiple things at a phenotypic level.
- Molecular gene pleiotropy
  - The question is about the number of functions a **molecular gene** has.
  - These functions can be defined not only **genetically**, but also **biochemically**.
- Developmental pleiotropy
  - **Mutations** rather than molecular genes are the relevant units.
- Selectional pleiotropy
  - The question is about the number of separate components of fitness a mutation affects.
  - A key feature of selectional pleiotropy is that traits are defined by the action of selection and not by the intrinsic attributes of the organism.

Biological meanings of pleiotropy

- When considering the relevance of data to each of these classes of pleiotropy, four issues are critical.
  - Are we discussing the genotype-phenotype map or the genotype-fitness map?
  - Are we discussing a molecular gene or a mutation?
  - How are we enumerating traits?
  - What do we mean when we say that a gene or mutation “affects” multiple traits?



TRENDS in Genetics

Detecting CP associations

- Univariate approaches
  - Combine the associations across various phenotypes
- Multivariate approaches
  - Jointly analyze more than one phenotype in a unified framework and test for the association of multiple phenotypes with a genetic variant.
  - Multivariate regression framework: generalized estimating equations (GEE), log-linear model, ordinal regression
  - Bayesian framework
  - Dimension reduction: principal components analysis, canonical correlation analysis

Detecting CP associations

- Univariate approaches

Table 2 | **Univariate approaches for detecting CP associations**

	Input	Explicit test of CP association	Allows effect heterogeneity	Types of phenotype (such as continuous or categorical)	Accommodates overlapping subjects	Combine data across multiple studies	Identify subset of associated phenotypes	Genetic variant versus region	Refs
Fisher	*P* value	No	Yes	Any	No	Yes	No	Variant	56
CPMA	*P* value	Yes	Yes	Any	No	Yes	No	Variant	14
Fixed effects meta-analysis	Effect estimate	No	No	Same type; need to standardize continuous phenotypes	No	Yes	No	Variant	54,57, 58‖
Random effects meta-analysis	Effect estimate	No	Moderate level; not opposite effects	Same type; need to standardize continuous phenotypes	No	Yes	No	Variant	54,57, 58‖
Subset-based meta-analysis	Effect estimate	No	Yes	Same type; need to standardize continuous phenotypes	No; offer extension to account for some overlap	Yes	Yes	Variant	59
Extensions to O'Brien	Effect estimate	No	Yes	Any	Yes; all subjects overlap*	No§	No	Variant	61,62
TATES	*P* value	No	Yes	Any	Yes; all subjects overlap‡	No§	No	Variant	63
PRIMe	*P* value	No	Yes	Any	Yes	Yes	No	Region	64

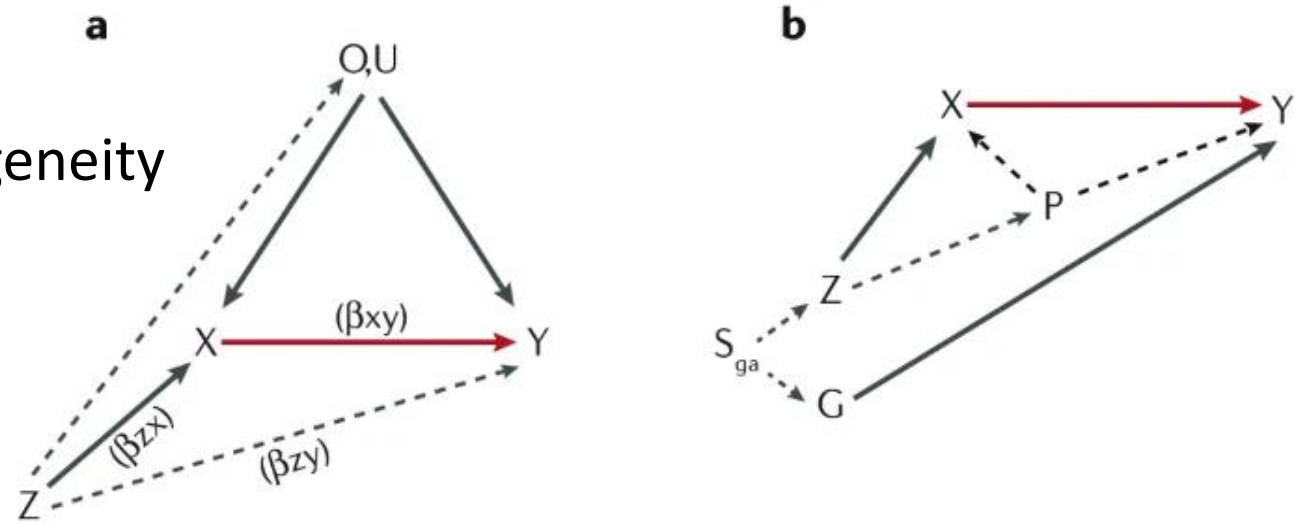
CP, cross-phenotype; CPMA, cross-phenotype meta-analysis; PRIMe, Pleiotropy Regional Identification Method; TATES, Trait-based Association Test that uses Extended Simes. *Can accommodate values missing completely at random. ‡Can accommodate values missing completely at random and blockwise missingness. §Can combine across multiple studies if all subjects have non-missing values for all phenotypes; TATES can accommodate situations in which a subset of studies have missing values for a subset of the phenotypes. ‖References are given for meta-analytical methods typically used in genome-wide association studies.

Distinguish CP effects

- Fine mapping
  - to distinguish biological and spurious pleiotropy
- Identifying mediated pleiotropy
  - The association between the variant and the first phenotype can be tested by adjusting or stratifying the first phenotype.
  - May be biased at the presence of confounding factors.
  - Mendelian randomization

Mendelian randomization

- The assumptions of MR
  - Relevance
  - Ignorability / Exchangability / Exogeneity
  - Exclusion restriction
- Challenges
  - Weak IV: polygenicity
  - Invalid IV: pleiotropy
    - Vertical pleiotropy: causality
    - Horizon pleiotropy: directional (unbalanced) or indirectional (balanced) – confounding
    - Spurious pleiotropy



Ignorability / Exchangability

- Ignorability means that the potential outcomes are independent with the treatment assignments (observed exposures).

$$Y(a) \perp A \text{ for all } a$$

- Exchangability means that the expected outcome in the non-treated group would have been the same as the outcome in the treated group if they had received the treatment.
- Conditional ignorability / exchangability:

$$Y(a) \perp A|X \text{ for all } a$$

- It is satisfied by randomization (in RCT), matching or exogeneity.

Mendelian randomization with summary statistics

- IVW-based methods
  - A weighted linear regression of **SNP effects on the outcome** on **SNP effects on the risk factor**
- MR Egger
  - An intercept term to deal with directional horizonal pleiotropy
  - **InSIDE assumption**: pleiotropy effects are independent of the effects on the exposure.

Table 1 | **Estimators for Mendelian randomization using summary statistics**

Method	Implementation	Limitations	Refs
IVW-based methods	The IVW method involves a weighted linear regression of SNP effects on the outcome on SNP effects on the risk factor, without an intercept term. The regression slope is equivalent to a weighted average of the ratio estimates (BOX 3) based on the precision of the causal estimate for each SNP used as an instrument[a,b]. IVW methods are more powerful than other methods (for example, MR-Egger)	Unlike the other methods described below, IVW cannot account for directional (unbalanced) pleiotropy[c]. Balanced pleiotropic effects[d] can be accounted for in random-effects IVW models (by allowing for heterogeneity) if the InSIDE assumption[e] holds true	129
Methods based on Egger regression	Linear regression with an intercept term using inverse variance weights[a,b]. MR-Egger regression provides consistent estimates even if all genetic instrumental variables are invalid under the InSIDE assumption[c]. This analysis is robust to directional (unbalanced) pleiotropy[c]. The intercept can be interpreted as the average pleiotropic effect across the genetic instrumental variables. Significance of the intercept term indicates the presence of unbalanced pleiotropy or violation of the InSIDE assumption[e]	Egger regression is less efficient and powerful than other methods because it allows for heterogeneity due to pleiotropy. It requires the InSIDE assumption[e]	60

Mendelian randomization with summary statistics

- Median-based methods
  - Calculate the ratio causal estimate for each instrument and then take the median
  - Assume that at least 50% IVs or IVs representing at least 50% weights are valid
- Mode-based methods
  - **ZEMPA assumption**: the largest subset of instruments with the same ratio estimate comprises the valid instruments

Median-based methods	Median-based methods allow some (but not all) instrumental variables to be invalid instruments. The median estimate is obtained by first calculating the ratio causal estimate for each instrumental variable and then taking their median. In the unweighted version, each genetic instrumental variable receives equal weight in the analysis. In the weighted version, the median is calculated using the inverse variance weights[b]. Median-based methods are more robust to directional pleiotropy than IVW and are more robust to individual genetic variants with outlying causal estimates than IVW and MR-Egger regression	These methods assume that at least 50% of the instrumental variables are valid instruments (unweighted median estimates) or that the instrumental variables that represent 50% of the weight in the analysis are valid instruments (weighted median estimates)	61
Mode-based methods	These methods allow the majority of the genetic instrumental variables to be invalid instruments under the ZEMPA assumption[f]. In the unweighted version of the mode estimate, each genetic instrumental variable receives equal weight in the analysis. In the weighted version, the mode is calculated using the inverse variance weights[b]. Mode-based methods are more robust to directional pleiotropy than IVW and more powerful than MR-Egger regression	The methods assume that the largest number of instrumental variable estimates comes from valid instruments (ZEMPA assumption[f]), that is, that the invalid instrumental variables have heterogeneous effect estimates. They have less power than IVW and median methods	130,131
Multiple methods	In practice, it is recommended to apply each of these methods to assess the robustness of the assumptions relevant for the different estimators, including the IVW estimator (all instruments are valid), the Egger estimator (all instruments may be invalid if the InSIDE assumption[e] is verified) and the median and modal estimators (a subset of genetic variants are valid instruments)		132

Summary

Process of GWAS

- Data collection
- Genotyping
- Quality control
- Imputation
- Association testing
- Visualization

Post-GWAS analysis

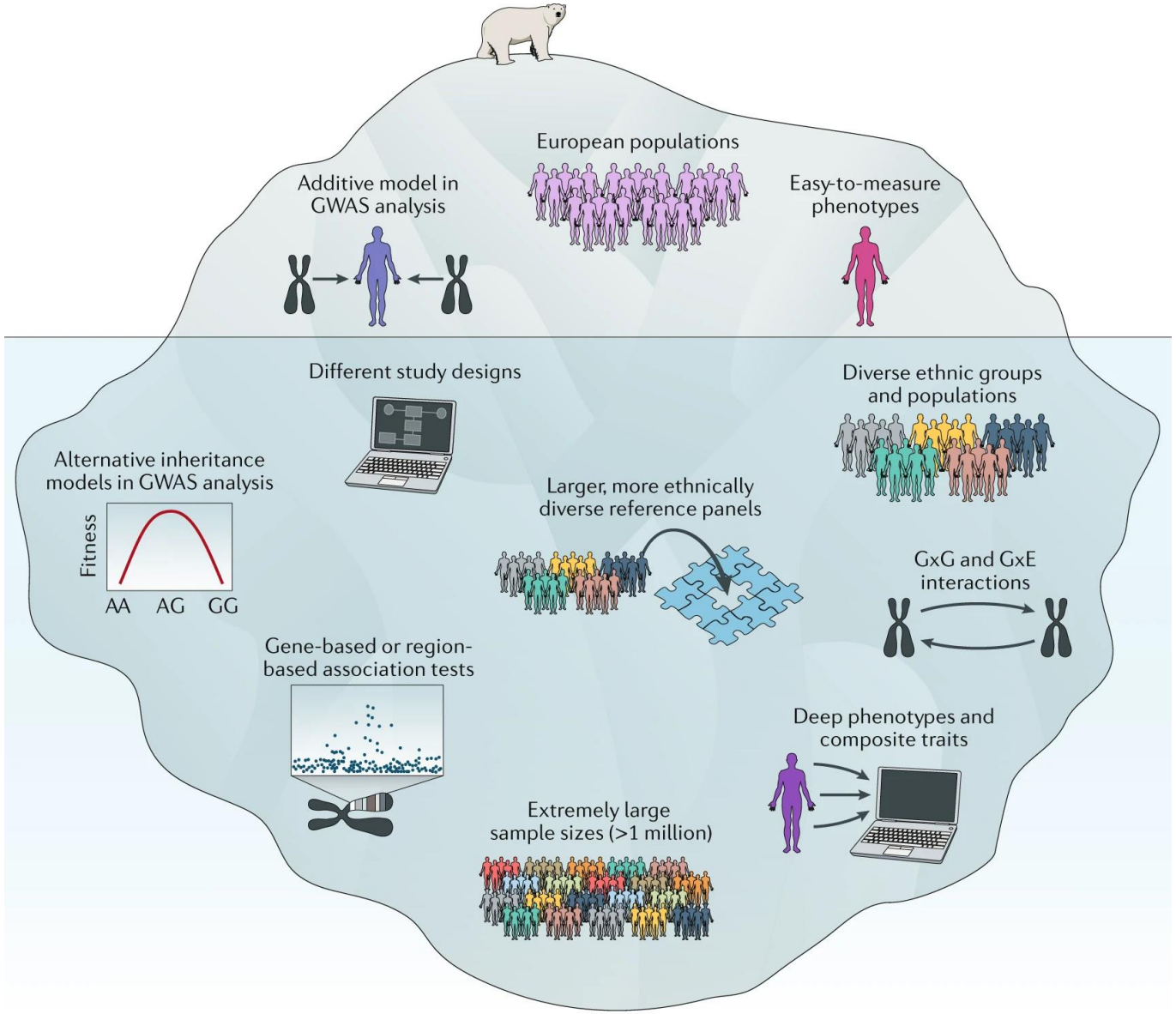
- Fine-mapping
- Functional analysis
- Risk prediction
- Determining polygenetic architecture
- Cross-trait analysis & causal inference

Prospect

- Extending the phenotypes studied in GWAS
  - Large prospective cohort studies with longitudinally measured clinical, demographic, lifestyle and environmental exposure data are needed
  - Electronic health data, behavioral health-tracking data, genetic data
- Expansion in scale at multiple levels
  - Sample size
  - Population studied: multi-ethic, admixed groups, isolated (founder and highly consanguineous populations)
  - Methods and study design used:
    - autosomal additive model → recessive, dominant, over-dominant, multiplicative, parent-of-origin-specific & X-linked inheritance models
    - Gene-gene & gene-environment interactions
    - Study designs: case-control, case-only, intervention & hypothesis-driven
    - Genomic-region based or gene-based association test
    - Bayesian analyses, machine learning, etc.
    - Heterogeneous traits

Prospect

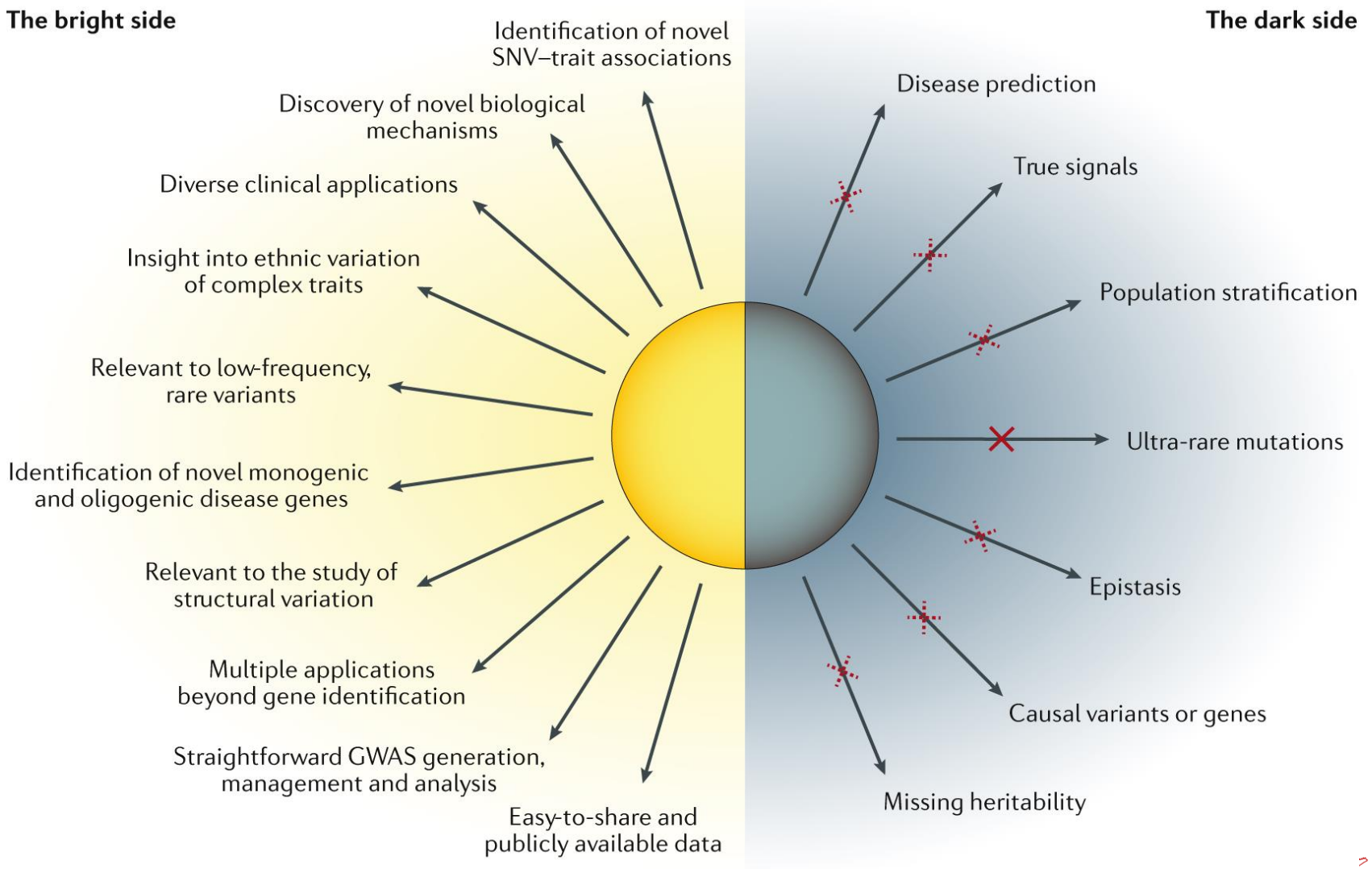
- GWAS performed to date represent **the tip of the iceberg**.



Challenges of GWAS

- Methodological challenges
  - Population stratification: spurious or biased associations
  - Fine-mapping: complex structure of genes
  - Polygenicity
  - Multiple testing burden
  - Causality
  - Unsuccessful in detecting epitasis
- Ethical challenges

Benefits & limitations



Software

- Open access tools for each stage of GWAS

Table 1 | **Open access tools that can be applied at each stage of GWAS**

Software	Use
Quality control	
PLINK/PLINK2 (REF.[20])	Can be used for many key steps in quality control, including filtering of bad SNPs (based on deviation from Hardy–Weinberg equilibrium, genotyping call rate and minor allele frequency) and bad individuals (based on sex check, genotyping call rate, sample call rate, heterozygosity and relatedness checks)
RICOPILI[23]	Quality control of raw genetic data and summary statistics used for input in meta-analyses
SMARTPCA	Principal component analysis of raw genotyping data; provides individual-level principal components that can be used to correct for population stratification
FlashPCA[255]	Similar to SMARTPCA; faster and more scalable with increasing sample sizes
Imputation	
IMPUTE2 (REFS[256,257])	Imputation of missing genotypes against an existing reference panel matched for ancestry; tends to use more memory than other imputation tools
BEAGLE[258]	Imputation of missing genotypes against an existing reference panel matched for ancestry
MACH/Minimac[259]	Imputation of missing genotypes against an existing reference panel matched for ancestry; Minimac includes pre-phasing, which speeds up imputation time
Association	
PLINK/PLINK2 (REF.[20])	Most widely known tool for conducting genetic associations
SNPTEST[260]	Genetic association testing; works well with IMPUTE2
GEMMA[55]	Genetic association testing based on linear mixed models
SAIGE[35]	Genetic association for binary phenotypes; analyses very large samples ($N > 100,000$)
BOLT-LMM[261]	Genetic association testing based on the BOLT-LMM algorithm for mixed model association testing and the BOLT-REML algorithm for variance components analysis (partitioning of SNP-based heritability and estimation of genetic correlations)
REGENIE[56]	Genetic association testing; analyses very large samples ($N > 100,000$); can assess multiple phenotypes at once; fast and memory efficient
BGENIE[76]	Genetic association for continuous phenotypes; analyses very large samples ($N > 100,000$); custom-made for the UK Biobank BGENv1.2 file format
fastGWA[37]	Mixed-model genetic association analysis

Software

- Open access tools for each stage of GWAS

Statistical fine-mapping	
CAVIAR[127]	Estimates the probability of each variant in a locus to be causal based on the observed pattern of *P* values and the level of linkage disequilibrium; allows for an arbitrary number of causal variants
PAINTOR[95]	Statistical fine-mapping using GWAS summary statistics and functional genomic data to prioritize likely causal variants
SuSIE[96]	Statistical fine-mapping using GWAS summary statistics and linkage disequilibrium information from a reference panel; based on a Bayesian modification of a forward selection model
FINEMAP[94]	Statistical fine-mapping using GWAS summary statistics as input; calculates effect sizes and heritability owing to likely causal SNPs
Meta-analysis	
GWAMA[262]	Fixed and random effects meta-analysis; allows the specification of different genetic models
METAL[39]	Weighted meta-analysis using GWAS summary statistics as input
Variant annotation	
VEP[115]	Functional annotation of genetic variants with their effect on genes, transcripts and protein sequence as well as regulatory regions
ANNOVAR[114]	Functional annotation of genetic variants with their effect on genes, transcripts and protein sequence as well as regulatory regions
FUMA[88]	Functional annotation of genetic variants with their effect on genes, transcripts and protein sequence as well as regulatory regions; includes chromatin interaction information and integrates and visualizes all output
Enrichment or gene-set analysis	
MAGMA[136]	Gene-based and gene-set analysis using competitive testing with a regression framework; allows testing of custom gene sets and includes options for conditional and interaction testing between gene sets
DEPICT[137]	Systematic prioritization of genes and assessment of enriched pathways using predicted gene functions
LDSC[174]	Partitioned SNP-based heritability analyses showing enrichment in sets of functionally related SNPs

Software

- Open access tools for each stage of GWAS

Table 1 (cont.) | **Open access tools that can be applied at each stage of GWAS**

Software	Use
QTL analysis	
QTLTools[263]	Molecular QTL discovery and analysis; uses raw genomic (sequence) data as input
Genetic correlations	
LDSC[174]	Assessment of genetic correlation between phenotypes using summary statistics as input; has various other functions, including partitioned SNP-based heritability and assessment of selection bias
GCTA[173]	Assessment of genetic correlation between phenotypes using raw genotypic data as input
SumHer[264]	Assessment of genetic correlation between phenotypes using summary statistics as input; has various other functions, including partitioned SNP-based heritability and assessment of selection bias
superGNOVA[183]	Assessment of local genetic correlations using GWAS summary statistics
ρ-HESS[184]	Assessment of local SNP-based heritability and genetic correlations using GWAS summary statistics
LAVA[185]	Assessment of local multivariate genetic correlations using GWAS summary statistics
GenomicSEM[265]	Assessment of multivariate genetic correlations based on GWAS summary statistics
Causality	
Mendelian randomization[266]	Assessment of causal relation between traits based on genetic overlap, using GWAS summary statistics as input.
PRS analysis	
PRScs[146]	Estimation of posterior effect sizes of SNPs using a Bayesian shrinkage approach
LDPred[151]/ LDPred-2 (REF.[150])	Estimation of posterior effect sizes of SNPs using a Bayesian shrinkage approach
SBayesR[147]	Estimation of posterior effect sizes of SNPs using a Bayesian shrinkage approach
PRSice[144]	PRS analysis using a *P* value thresholding and clumping approach
TWAS	
FUSION[125]	Performing TWAS by predicting functional/molecular phenotypes based on reference data; uses GWAS summary statistics as input
PrediXcan[126]	Prioritizing likely causal genes based on transcription data; uses GWAS summary statistics as input
SMR	Testing whether SNP-trait associations are mediated by gene expression levels using a Mendelian randomization approach

GWAMA, genome-wide association meta-analysis; GWAS, genome-wide association studies; PRS, polygenic risk score; QTL, quantitative trait locus; SNP, single-nucleotide polymorphism; TWAS, transcriptome-wide association studies.

Software

Term/approach	Objective	Selected resources and tools
Genome annotations	Prioritize genes, regions, and variants based on their potential functions	Protein-coding region: PolyPhen Sequence conservation: GERP, phyloP Supervised methods: CADD, GWAVA, DeltaSVM, DeepSea Unsupervised methods: ChromHM, Segway, GenoCanyon, Eigen Tissue-specific annotations: GenoSkyline, FUN-LDA Integration of disease information: Phevor, Phen-Gen, PINES Integrated resources: RegulomeDB, ANNOVAR
Meta-analysis	Integrate results from different studies	Stata, METAL
Pleiotropic analysis	Integrate information from genetically correlated traits	GPA, MTAG
Multilocus or gene-based analysis	Integrate information from different markers	PCA, MAGMA, SKAT, VEGAS, GATES
Pathway- and network-based analysis	Integrate information from multiple related genes	MAGMA, ToppGene, MRF-based analysis
Fine mapping	Localize functional genes and variants	PAINTOR, CAVIAR, FINEMAP
TWAS	Integrate information from eQTL studies	PrediXcan, FUSION, UTMOST
Heritability	Estimate the overall genetic contribution from a set of markers to a trait	GCTA, LDSC, GNOVA
Genetic correlation	Estimate the genetic correlations between traits for a set of markers	GCTA, LDSC, GNOVA
Polygenic risk score	Predict disease risk based on genetic information	P+T, LDPred, AnnoPred, PleioPred, PRS-CS
Mendelian randomization	Infer the causal relationship among traits	IVW, MR-Egger

Any Questions?