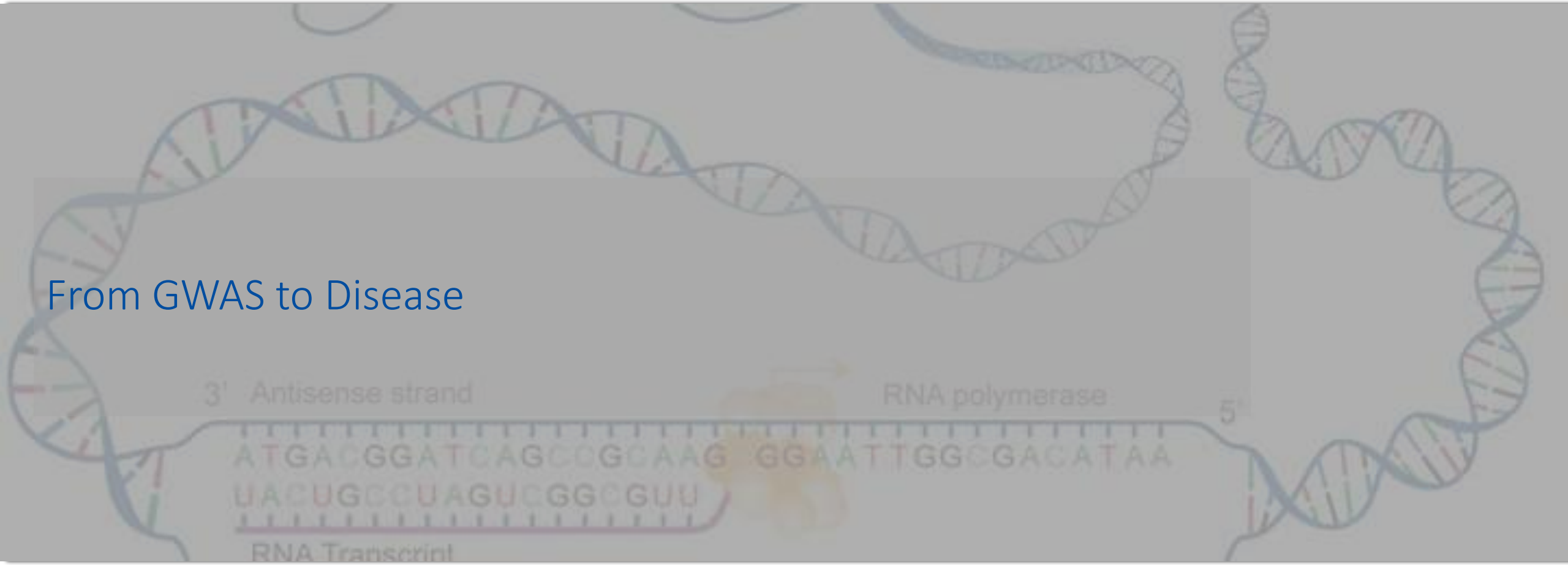
From GWAS to Disease



Christine Hsiung 熊嘉妮

Bioinfo@GGA.ASIA

Study design

- Define the phenotype
  - Phenotype is an observable trait
  - Phenotype must be measurable
    - Categorical
      - Hair color, Hypertension, Diabetes
    - Continuous
      - Height, Blood Pressure, Total Cholesterol
- Choosing a study design
  - Case-control study
  - Cohort study

Covariate adjustment

- Statistical tests should be adjusted for factors that are known to influence the trait, such as sex, age, and known clinical covariates.

- Covariate adjustment reduces spurious associations due to sampling artifacts or biases in study design.

Quality control

- Makers
  - Batch effect
  - Genotyping call rates
  - Minor allele frequencies (MAF)
  - Hardy-Weinberg Equilibrium (HWE)
  - Missing genotyping rate difference between case/control
- Samples
  - Filter samples by genotyping call rates
  - Remove individuals with extreme heterozygosity rate
  - Filter samples by relatedness

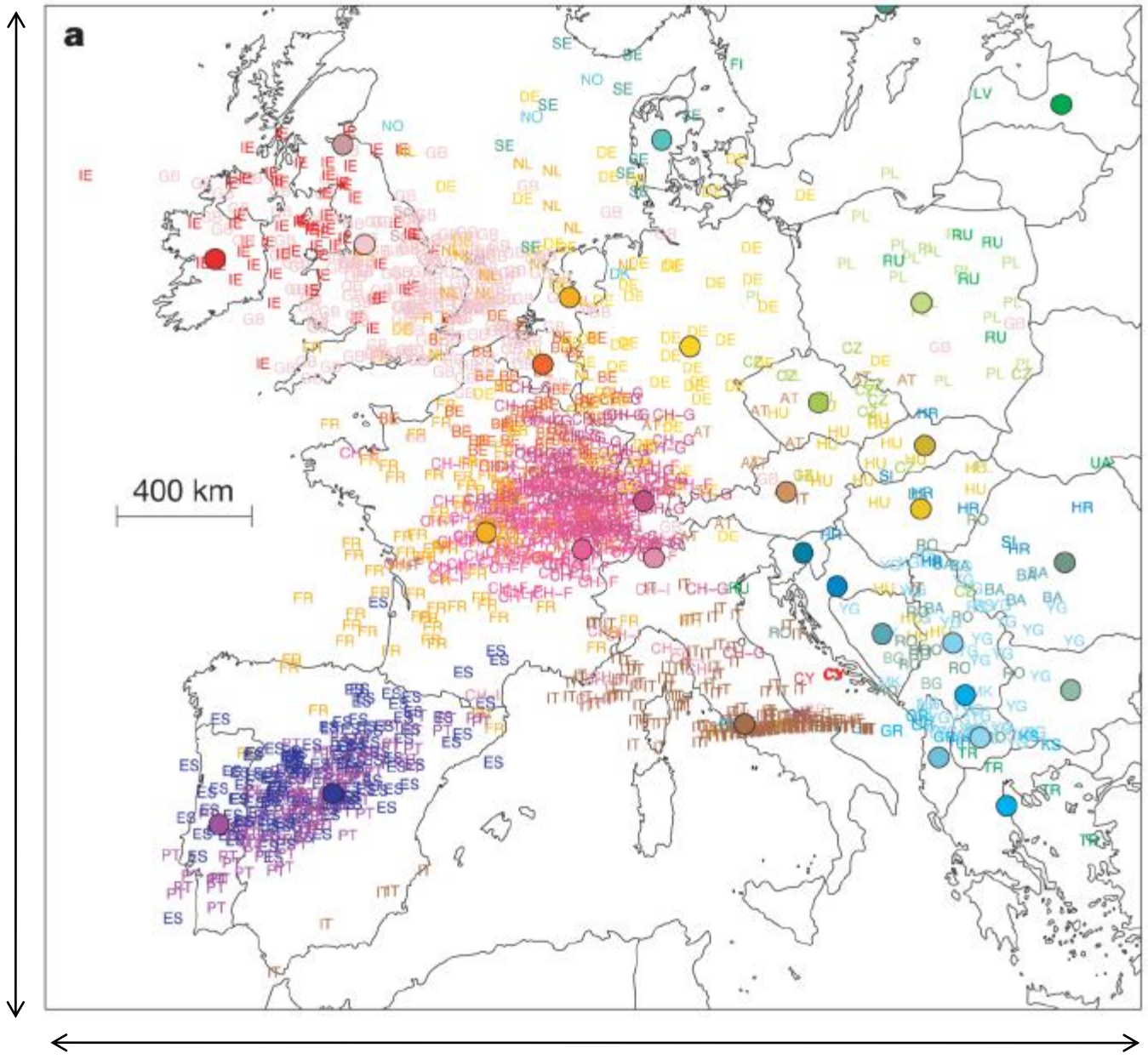
Population Structure

- One of the more important covariates to consider in GWAS is population structure.
- There are differences in phenotype prevalence due to ethnicity, and allele frequencies are variable across human subpopulations.
- In a population with multiple ethnicities, ethnic-specific SNPs will likely be associated to the trait due to population stratification.

Principal Components Analysis

- Principal Components Analysis is applied to genotype data to infer continuous axes of genetic variation
- Each axis explains as much of the genetic variance in the data as possible with the constraint that each component is orthogonal to the preceding components
- The top principal Components tend to describe population ancestry
- Include principal components in regression analysis => correct for the effects of stratification
- EIGENSTRAT, SHELLFISH,PLINK

Novembre et al, Nature (2008)

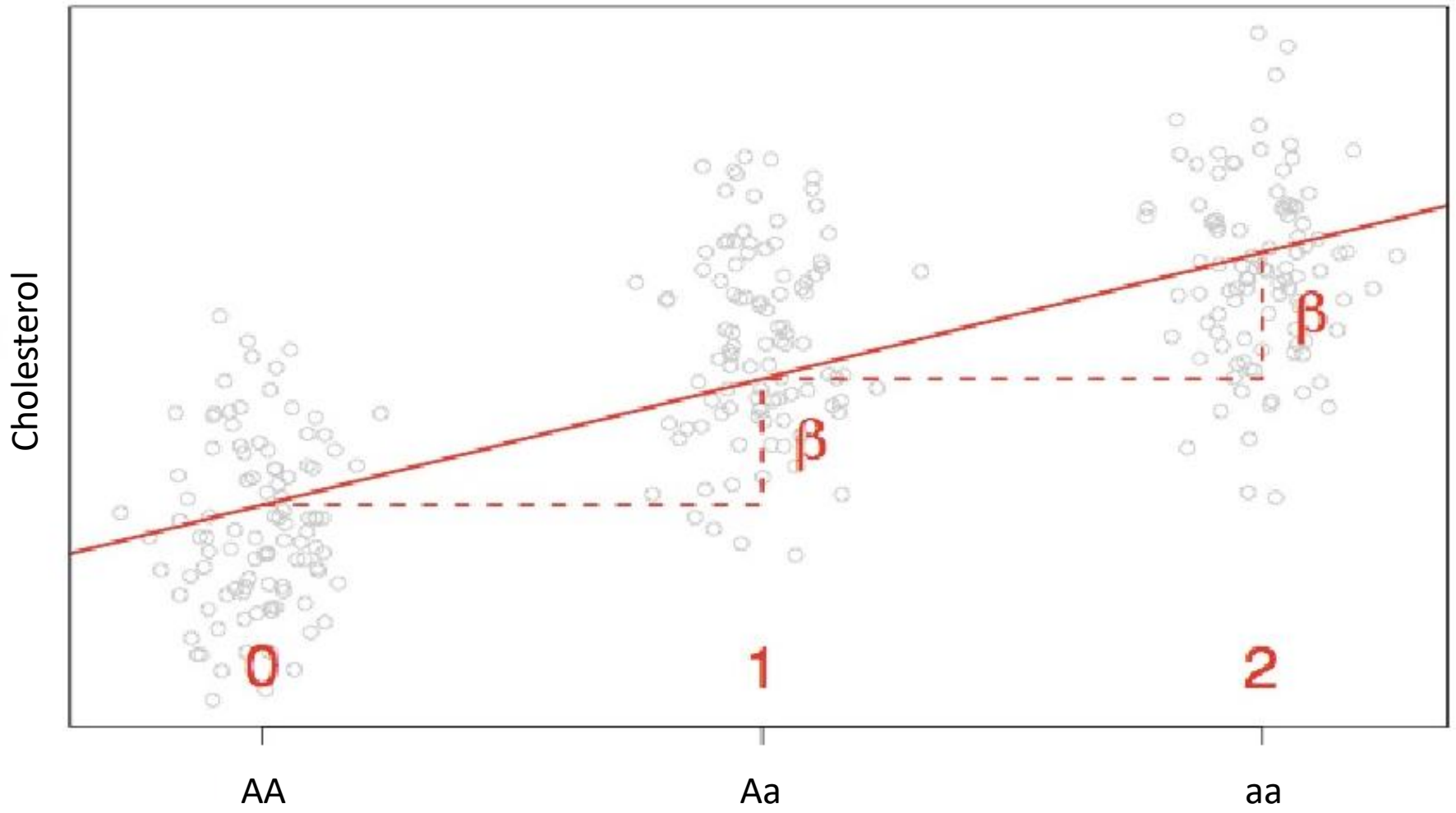


Statistical test

- Quantitative traits
  - Linear regression
  - H_0 : there is no difference between the trait means of genotype groups.

- Dichotomous case/control traits
  - Logistic regression
  - H_0 : there is no association between the phenotype and genotype groups.

Additive model



Multiple testing

- In GWAS, hundreds of thousands to millions of tests are conducted.
  - The likelihood of finding false positive over the entire GWAS is therefore much higher.

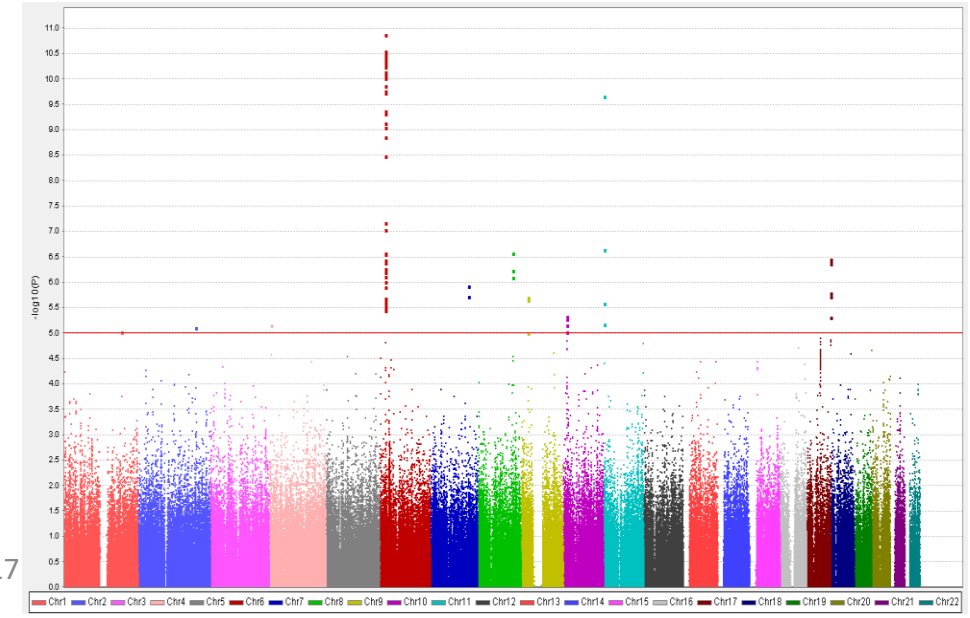
- Correction for multiple testing
  - Bonferroni correction

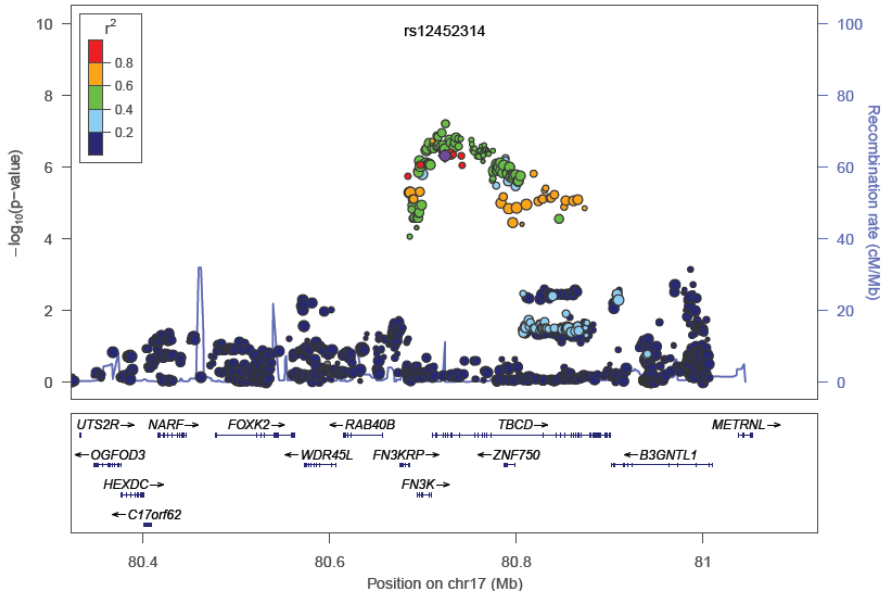
$$\alpha = \frac{0.05}{the\ number\ of\ tests\ conducted}$$

$\alpha = 0.05 / 10^7 = 5e^{-8}$ (genome-wide significance)

Visualization of result

- Manhattan plot
  - Genome-wide p-values
- Regional plot
  - Gene-level p-values





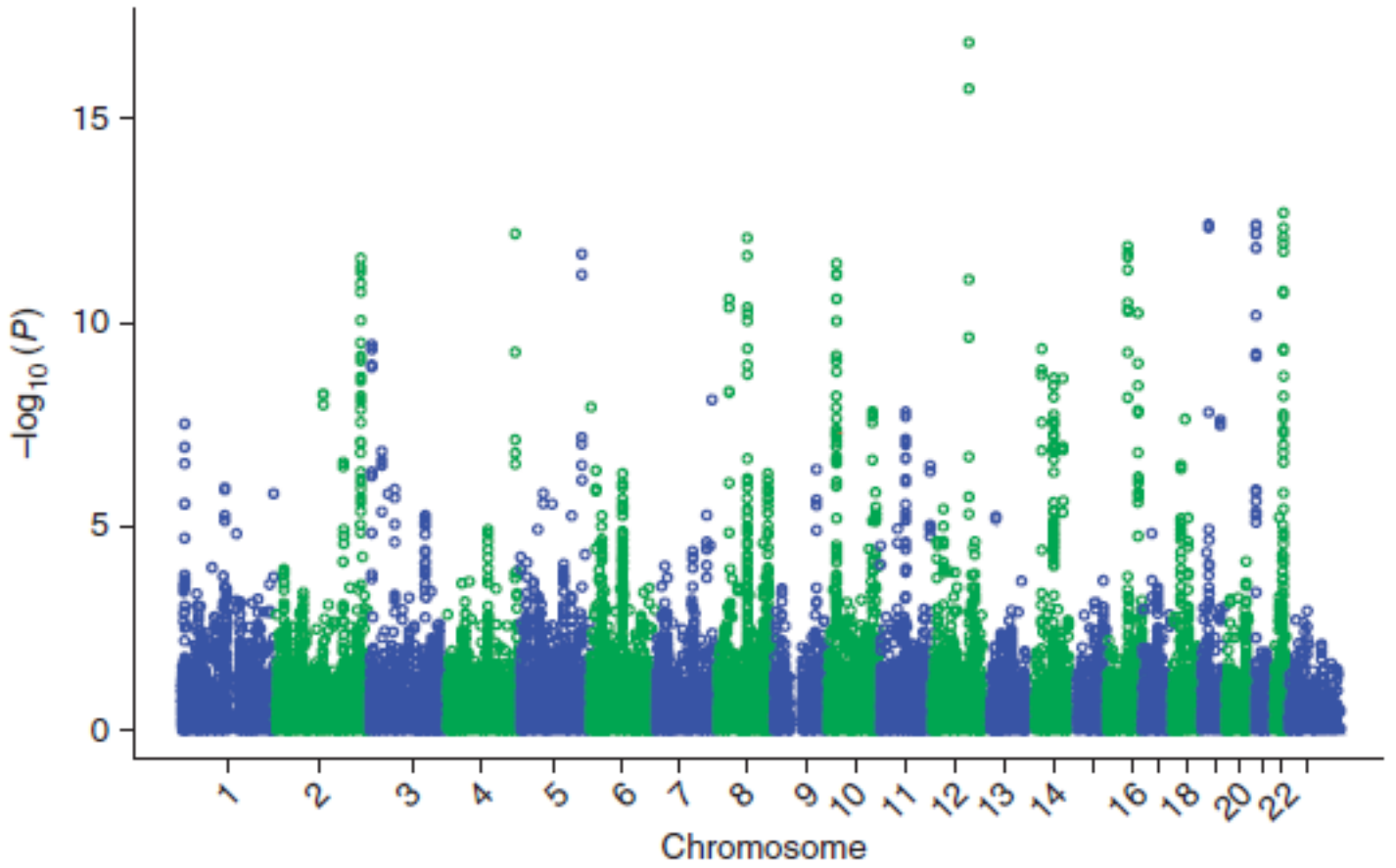


Figure 1 One-degree-of-freedom trend-test statistics for 29,807 iCOGS SNPs selected from the combined GWAS, excluding those occurring in known susceptibility regions. The red horizontal line represents $P = 5 \times 10^{-8}$. The blue horizontal line represents $P = 1 \times 10^{-5}$.

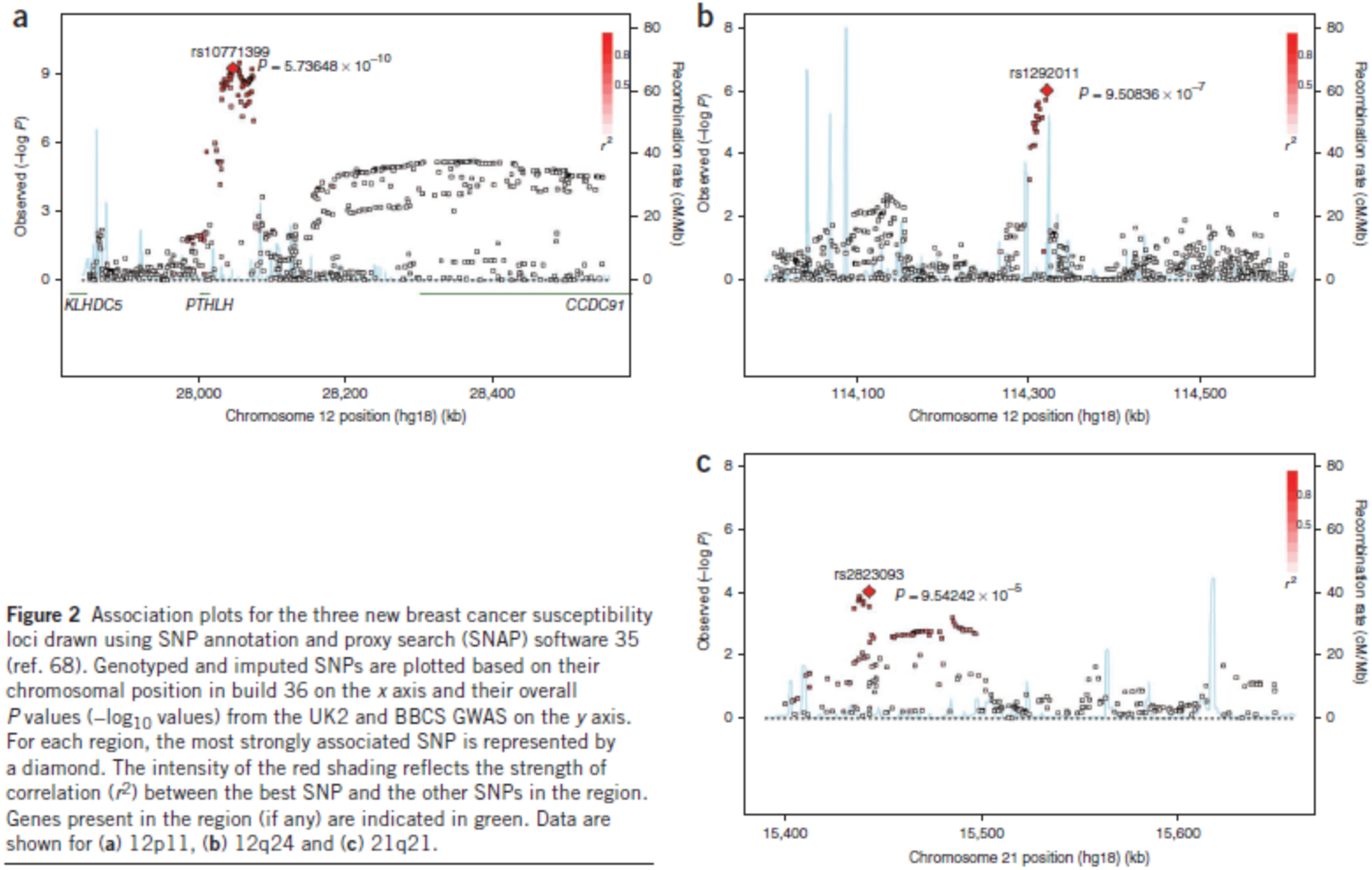
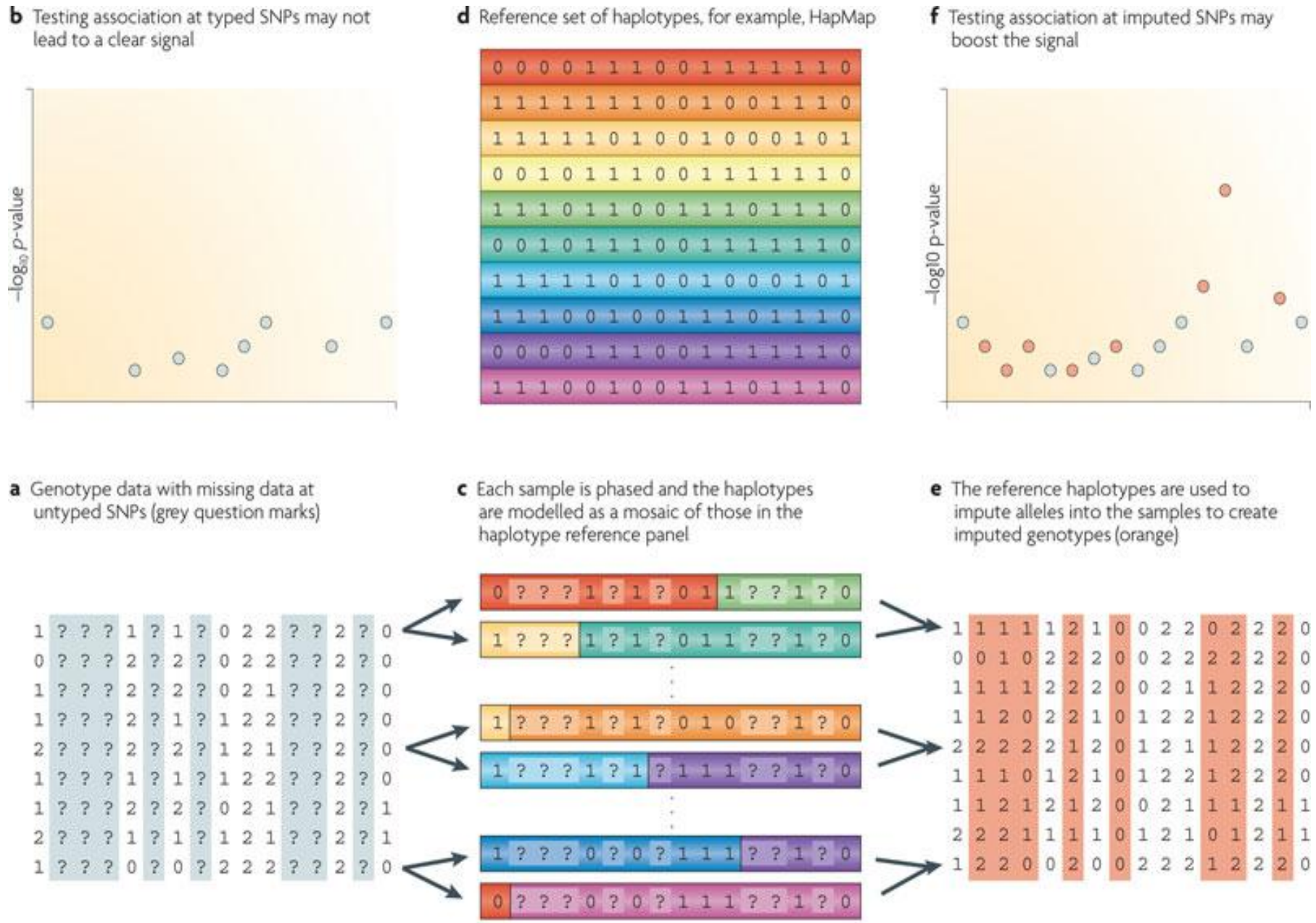


Figure 2 Association plots for the three new breast cancer susceptibility loci drawn using SNP annotation and proxy search (SNAP) software 35 (ref. 68). Genotyped and imputed SNPs are plotted based on their chromosomal position in build 36 on the *x* axis and their overall *P* values ($-\log_{10}$ values) from the UK2 and BBCS GWAS on the *y* axis. For each region, the most strongly associated SNP is represented by a diamond. The intensity of the red shading reflects the strength of correlation (r^2) between the best SNP and the other SNPs in the region. Genes present in the region (if any) are indicated in green. Data are shown for (**a**) 12p11, (**b**) 12q24 and (**c**) 21q21.

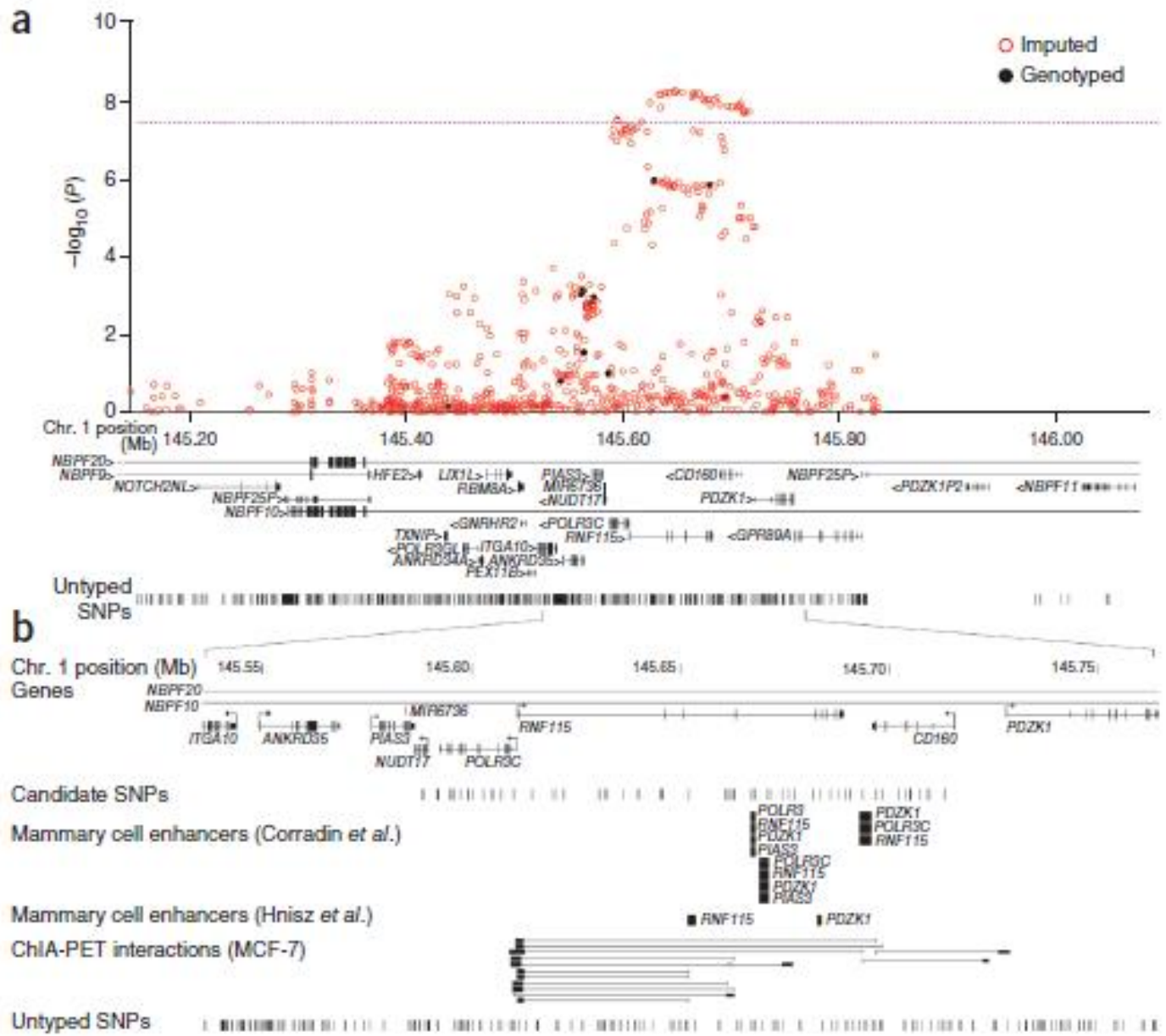
Imputation of SNP Genotypes

- Estimate unmeasured or missing genotypes.
- Based on measured SNPs and external info (e.g., haplotype structure of HapMap).
- Increase GWAS power.
- Allow for combining data across different platforms (e.g., Affy & Illumina) (for replication / meta-analysis).


b Testing association at typed SNPs may not lead to a clear signal
−log10 p-value
d Reference set of haplotypes, for example, HapMap
f Testing association at imputed SNPs may boost the signal
−log10 p-value
a Genotype data with missing data at untyped SNPs (grey question marks)
c Each sample is phased and the haplotypes are modelled as a mosaic of those in the haplotype reference panel
e The reference haplotypes are used to impute alleles into the samples to create imputed genotypes (orange)

Functional annotation

- Methods of discovery
  - Lab
    - Chromatin conformation capture (3C)
    - Luciferase assay
  - Computation
    - Coding / UTR/ Intron
    - Transcription factor binding
    - eQTL (GTEx protal)
    - Methylation/histone modification


a
Imputed
Genotyped
−log10 (P)
Chr. 1 position (Mb)
145.20
145.40
145.60
145.80
146.00
Untyped SNPs
b
Chr. 1 position (Mb)
Genes
Candidate SNPs
Mammary cell enhancers (Corradin et al.)
Mammary cell enhancers (Hnisz et al.)
ChIA-PET interactions (MCF-7)
Untyped SNPs

Prediction from GWAS

- The vision, the doctor will have a “desktop predictor”
  - Input: patient’s genome
  - Output: risk for one (or many) diseases

- Building prediction models is a very different use of GWAS information
  - Non-genetic risk factors that are correlated with the genome (like diet) are also legitimate for prediction
  - Construct the PRS (polygenetic risk score) and other risk factor model
- Estimate the prediction model accuracy
  - Receiver operating characteristic curve (ROC)/Area under curve (AUC)

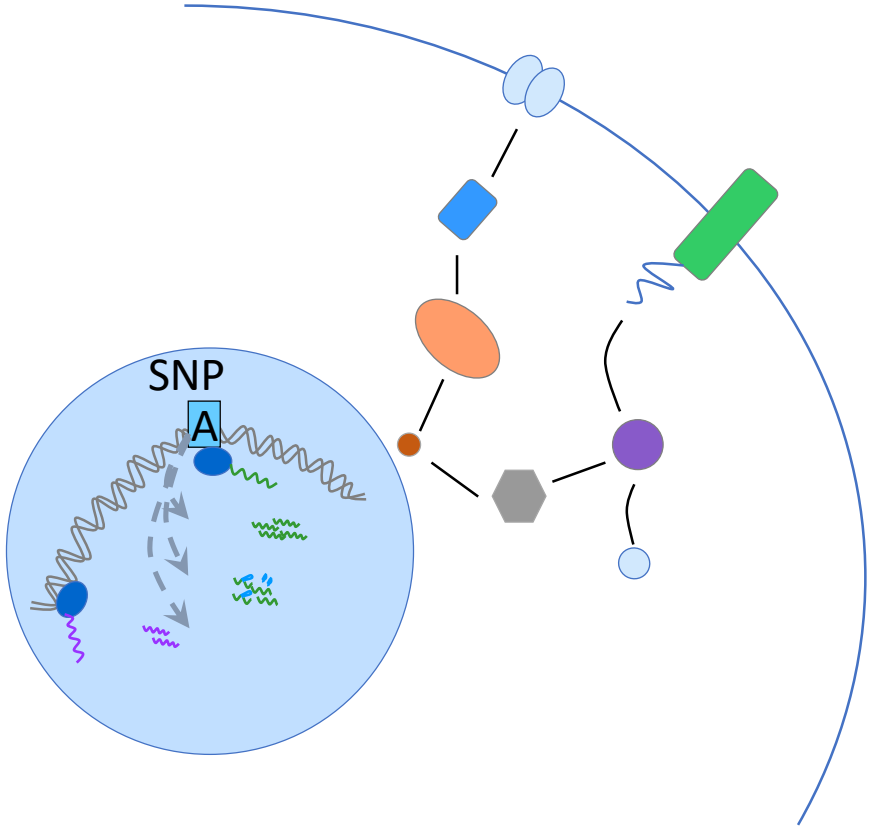
Construct PRS

- $PRS_j = \sum [\beta_{i,discovery} * SNP_{ij}]$
  - $\beta_{i,discovery}$ = effect size in discovery sample from OLS (continuous trait) or logistic reg (binary trait; $\beta = \log(OR)$)
  - SNP_{ij} = # alleles (0,1,2) for SNP *i* of person *j* in target sample
- In PLINK, --score/--score sum.

Motivation and introduction

How can we use gene expression and epigenetics to help us understand complex trait genetics?

- Majority of trait-associated variation is non-coding.
- Common hypothesis is that most of these function by altering gene expression.

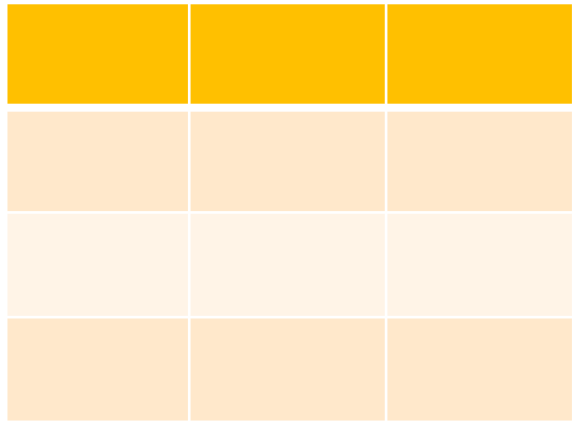


eQTL analysis Math

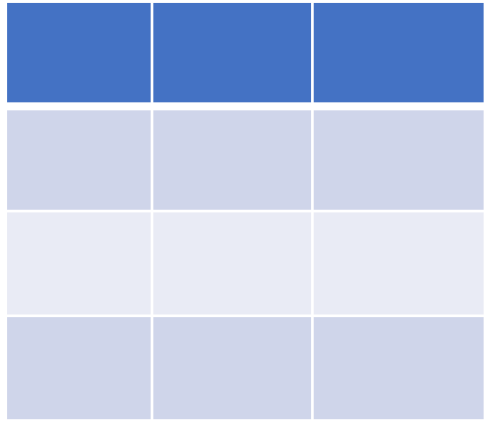
- Linear regression: find the coefficients for the effect of expression on genotype when conditioned on the covariates in a linear model and test if they are significant diffetent than 0

$$gene\ expressoin = \beta_0 + \beta_1 genotype + \beta_2 covaraites$$

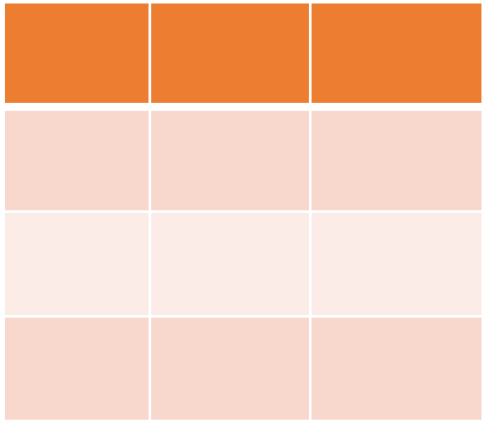
Expression



Genotype

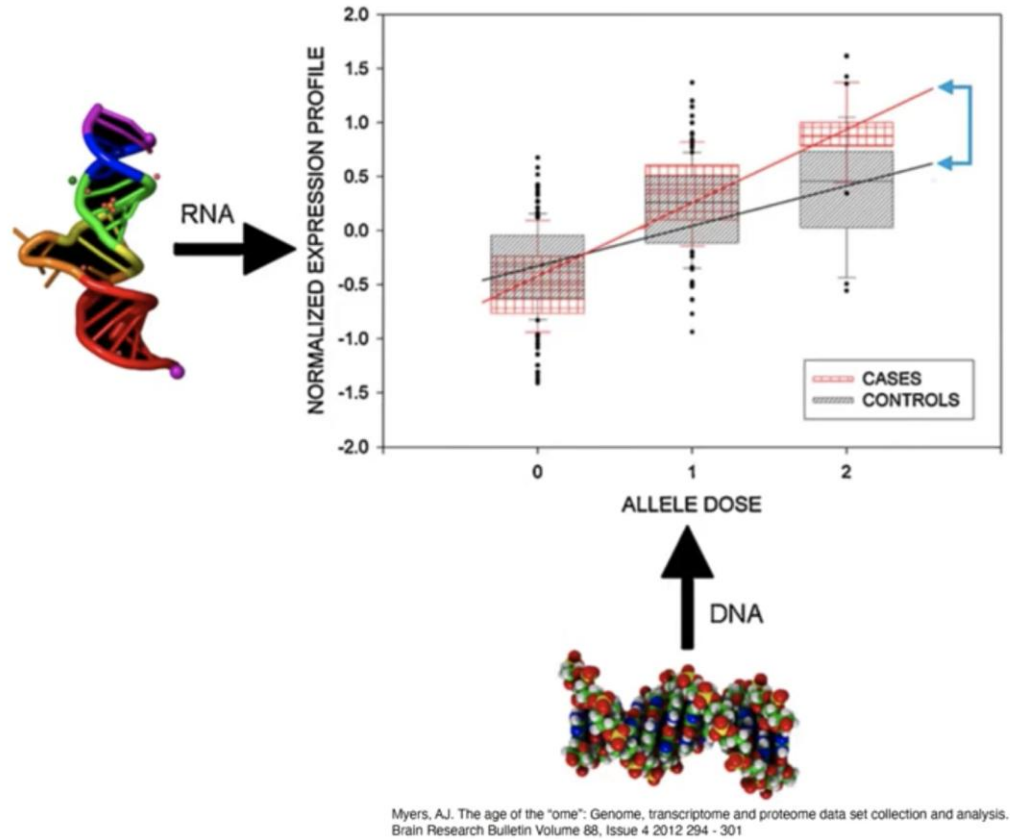


Covaraite

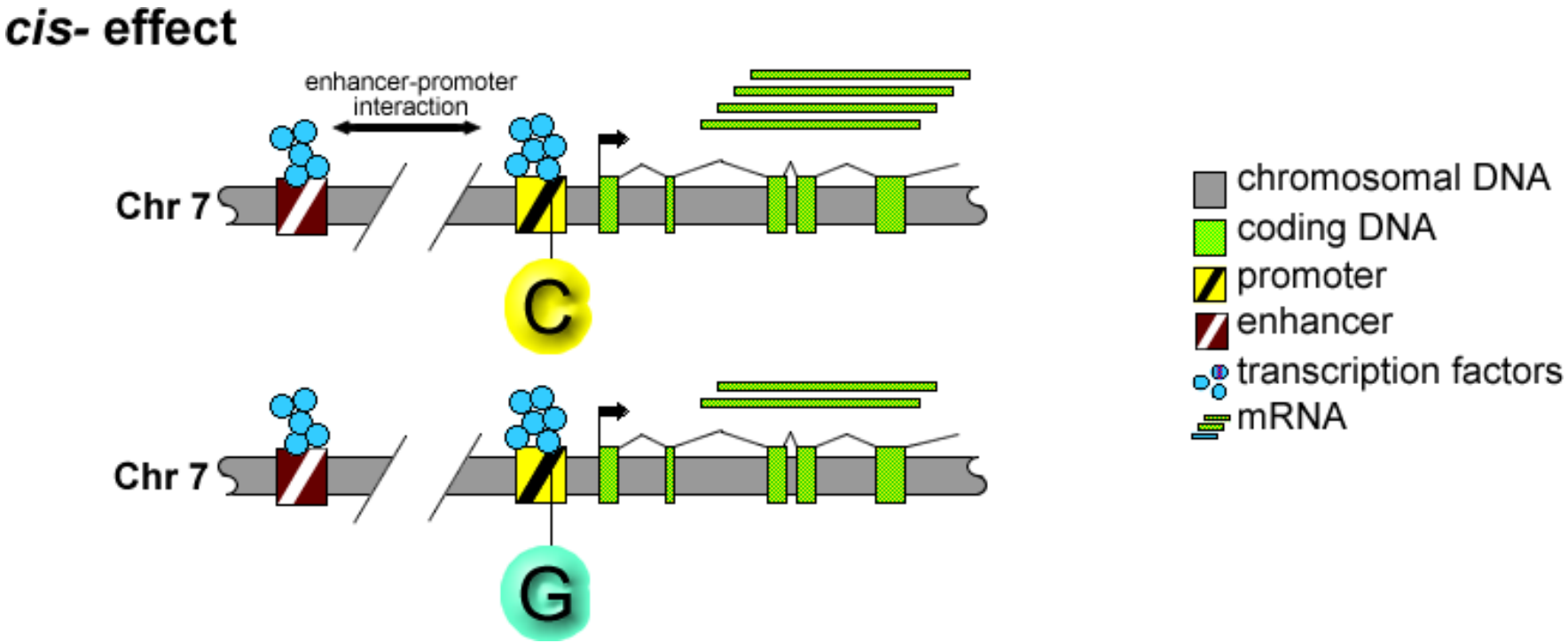


Mapping expression (e)QTL

- RNA expression levels can be treated like any other quantitative trait in QTL mapping.
- 30,000 genes by 10,000 SNPs = 300,000,000 comparisons!
- eQTL studies are sometimes called genetical genomics

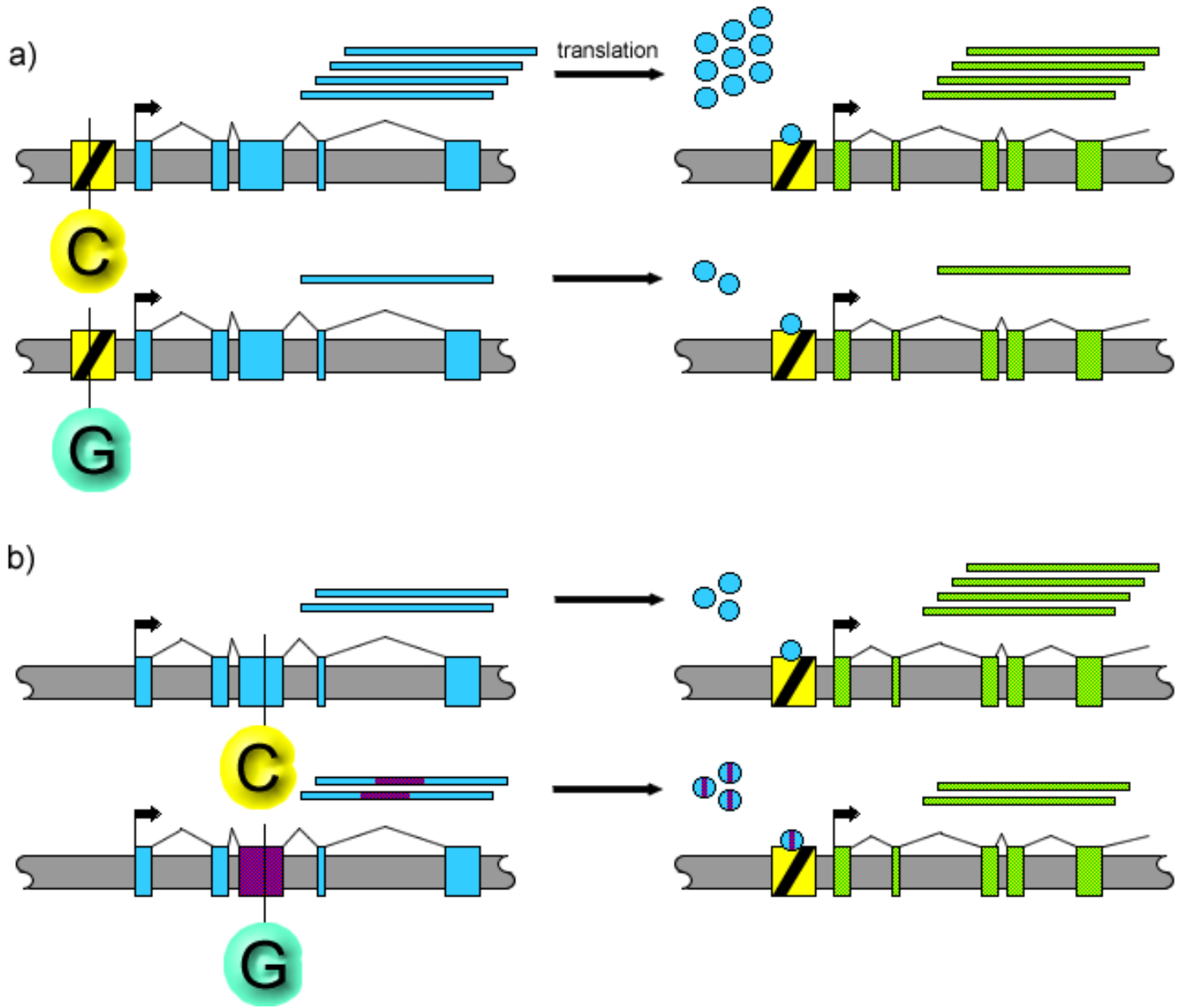


Myers, AJ. The age of the "ome": Genome, transcriptome and proteome data set collection and analysis. Brain Research Bulletin Volume 88, Issue 4 2012 294 - 301



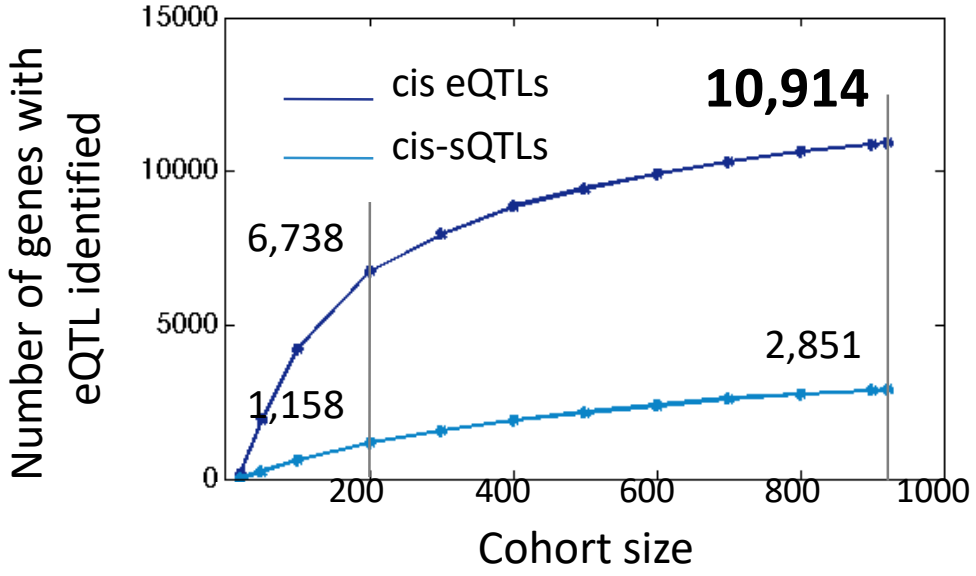
Canonical model

trans- effect



Prevalence of eQTLs

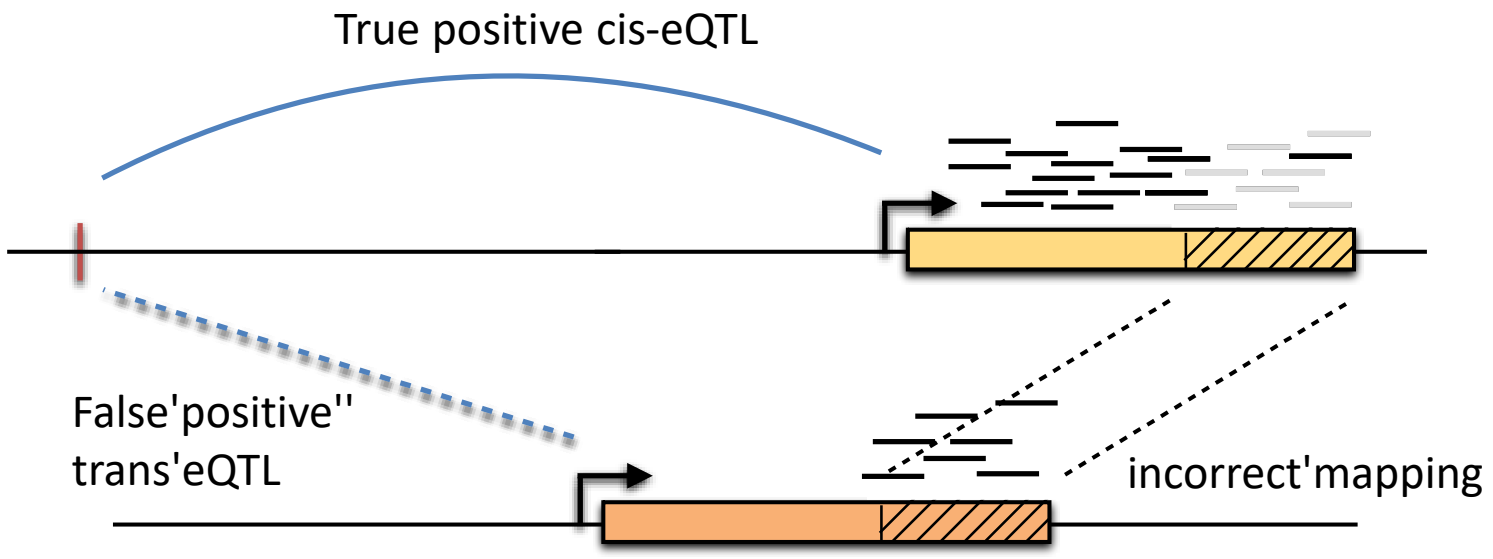
Cis-eQTLs have now been identified for nearly every human gene, with numerous large studies available



Battle, Genome Research, 2014

Challenges for trans-eQTL detection

- Power
- False positives from many sources e.g. over and under correcting confounders (Dahl et al, 2017)
- Mapping error (similar to probe cross-hybrid.)

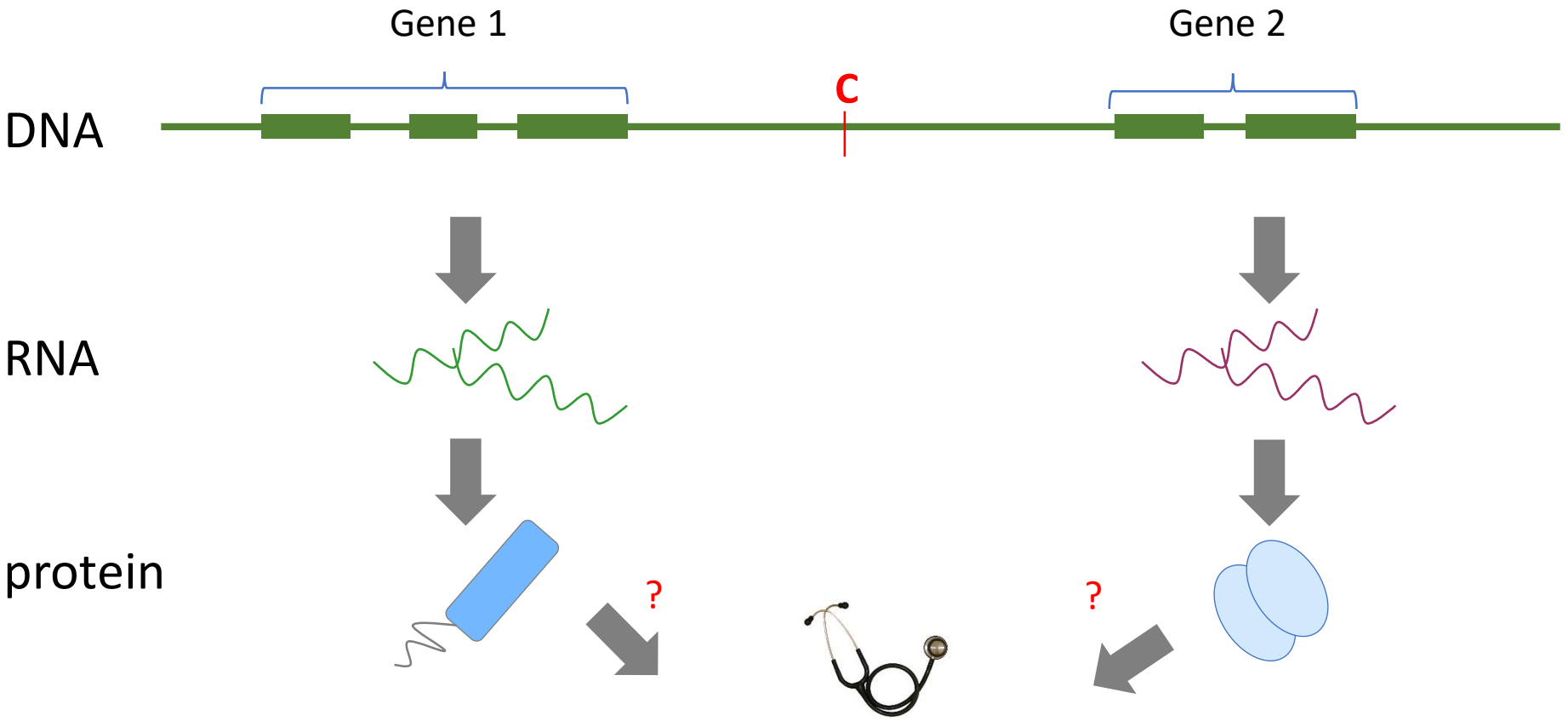


Slide adapted from Yuan He

eQTLs and complex disease genetics

Help interpret GWAS variants (especially non-coding):

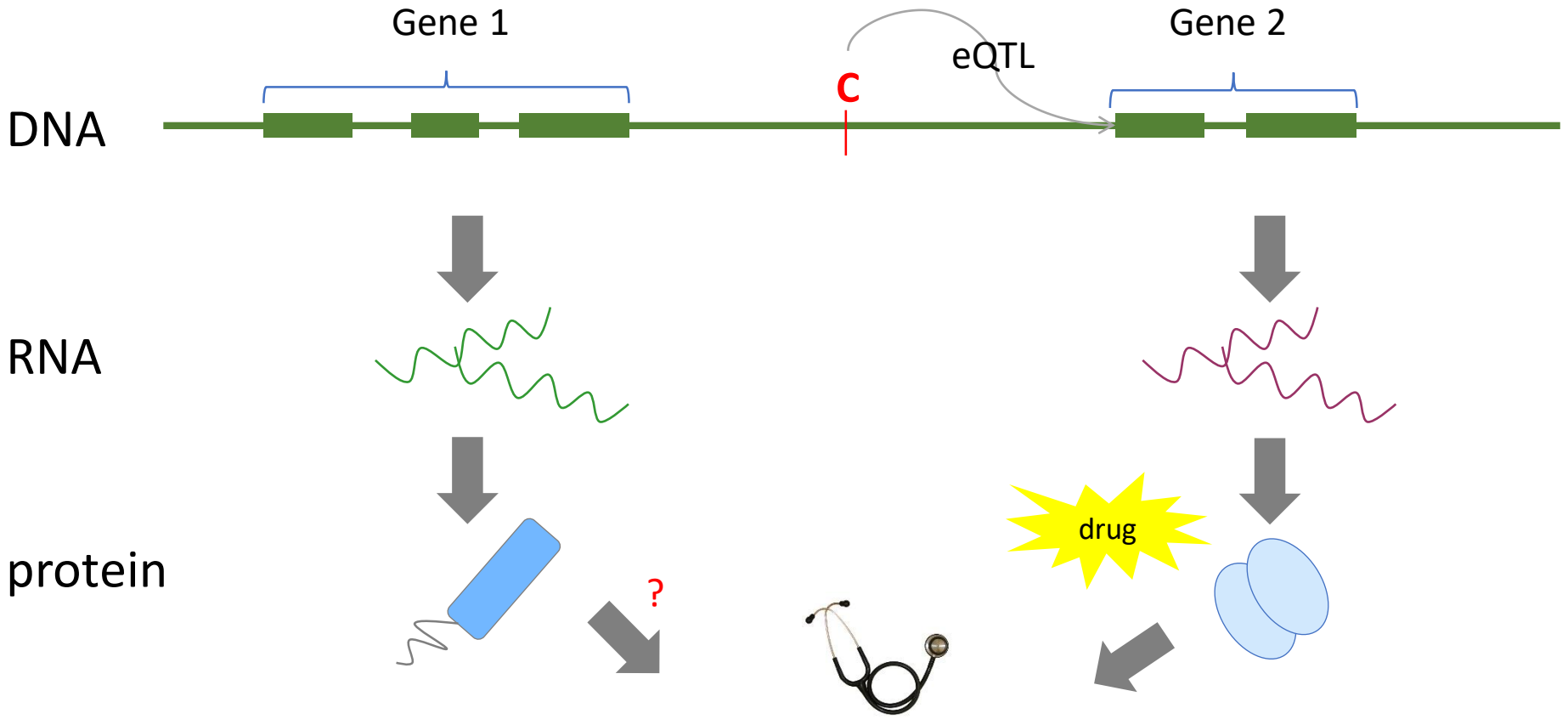
- understand mechanism
- guide interventions



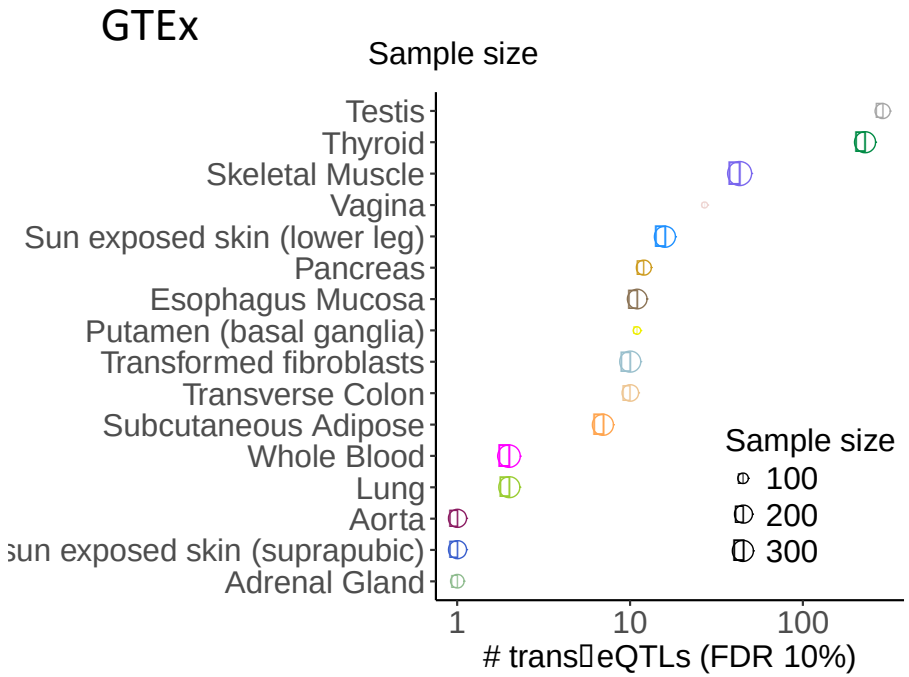
eQTLs and complex disease genetics

Help interpret GWAS variants (especially non-coding):

- understand mechanism
- guide interventions



Trans-eQTLs



Large studies:

Westra et al (N=5,311, using WAS variants only)
ALSPAC (N=869)
MUTHER (N=850)
DGN (N=922)
Framingham (N=5257)

Studies report wildly different genes

Replication and validation remains poor

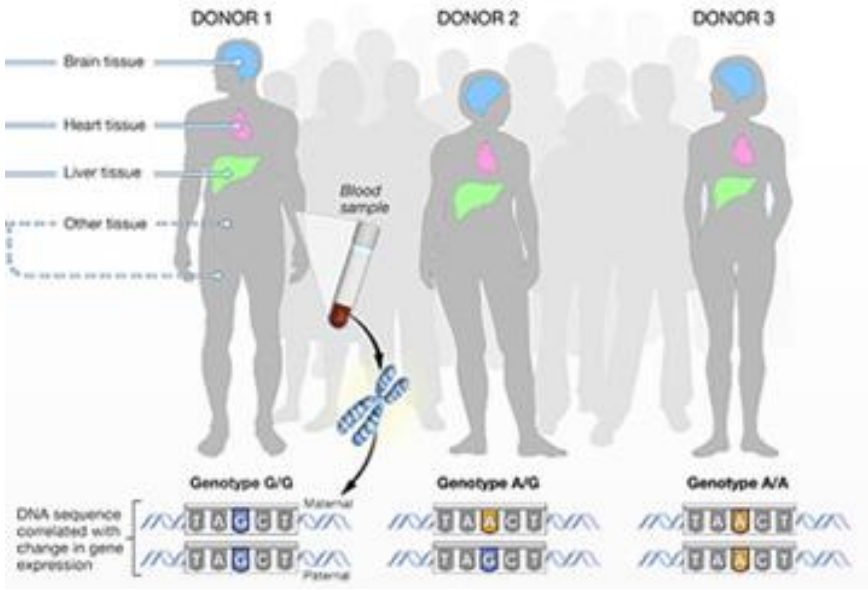
We remain underpowered at current sample sizes

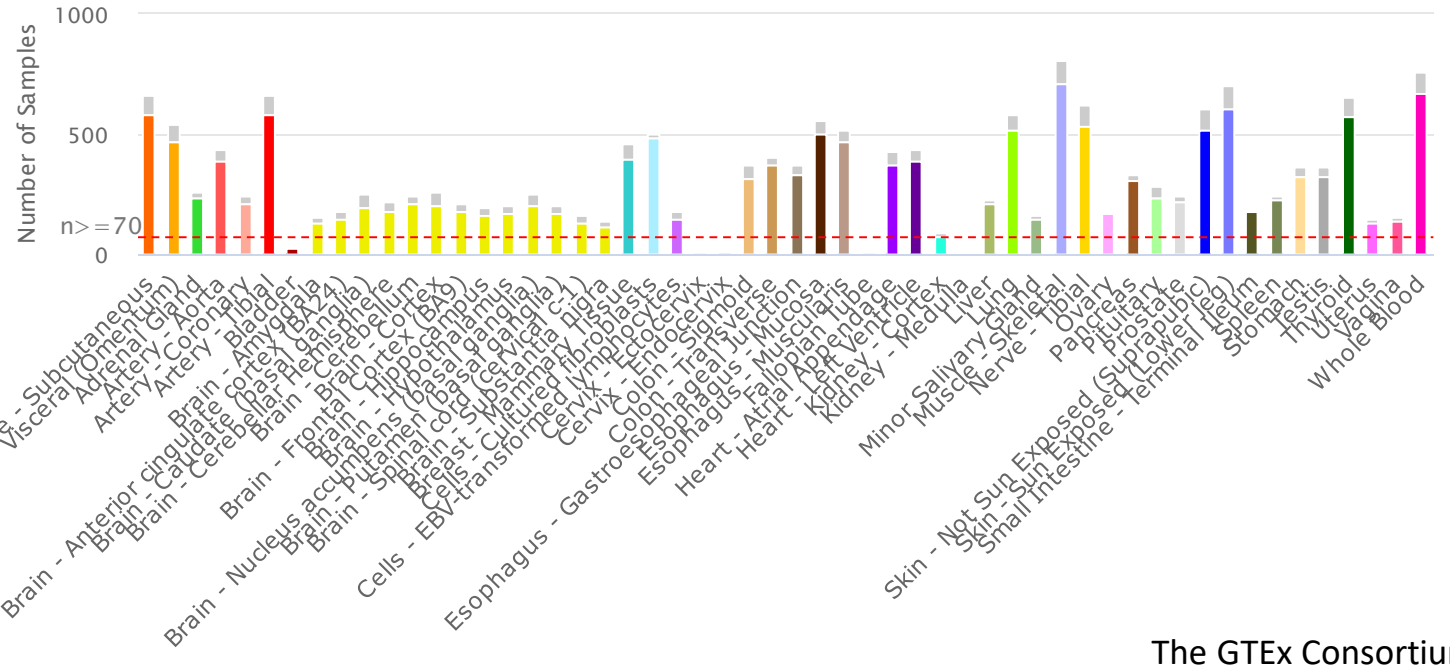
GTEx Project

GTEx Consortium v11 data

- 838 genotyped donors
- 15,201 RNA-sequencing samples

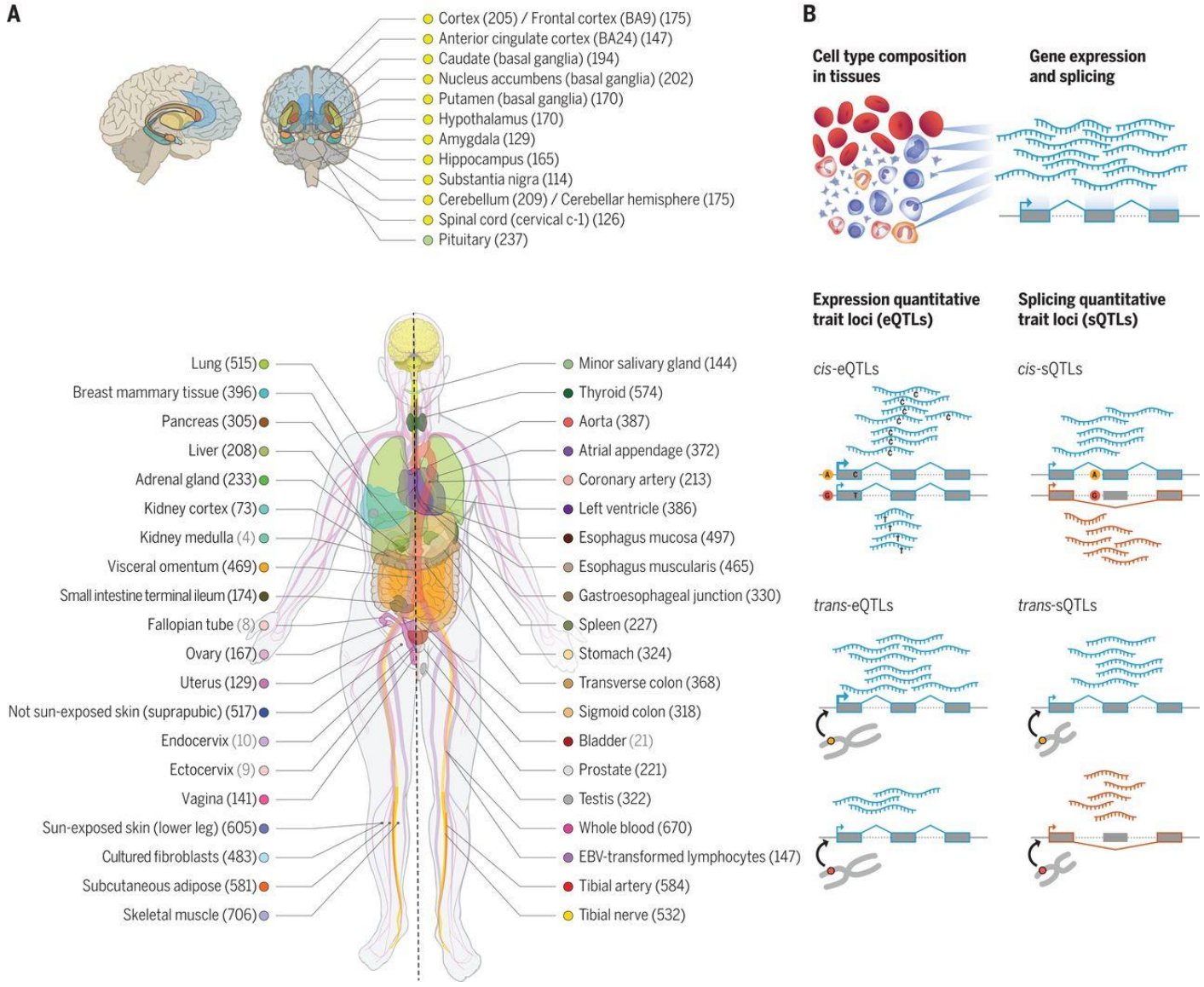
https://gtexportal.org/home/





The GTEx Consortium

Genetic effects across human tissues



The GTEx Consortium, Nature 2020

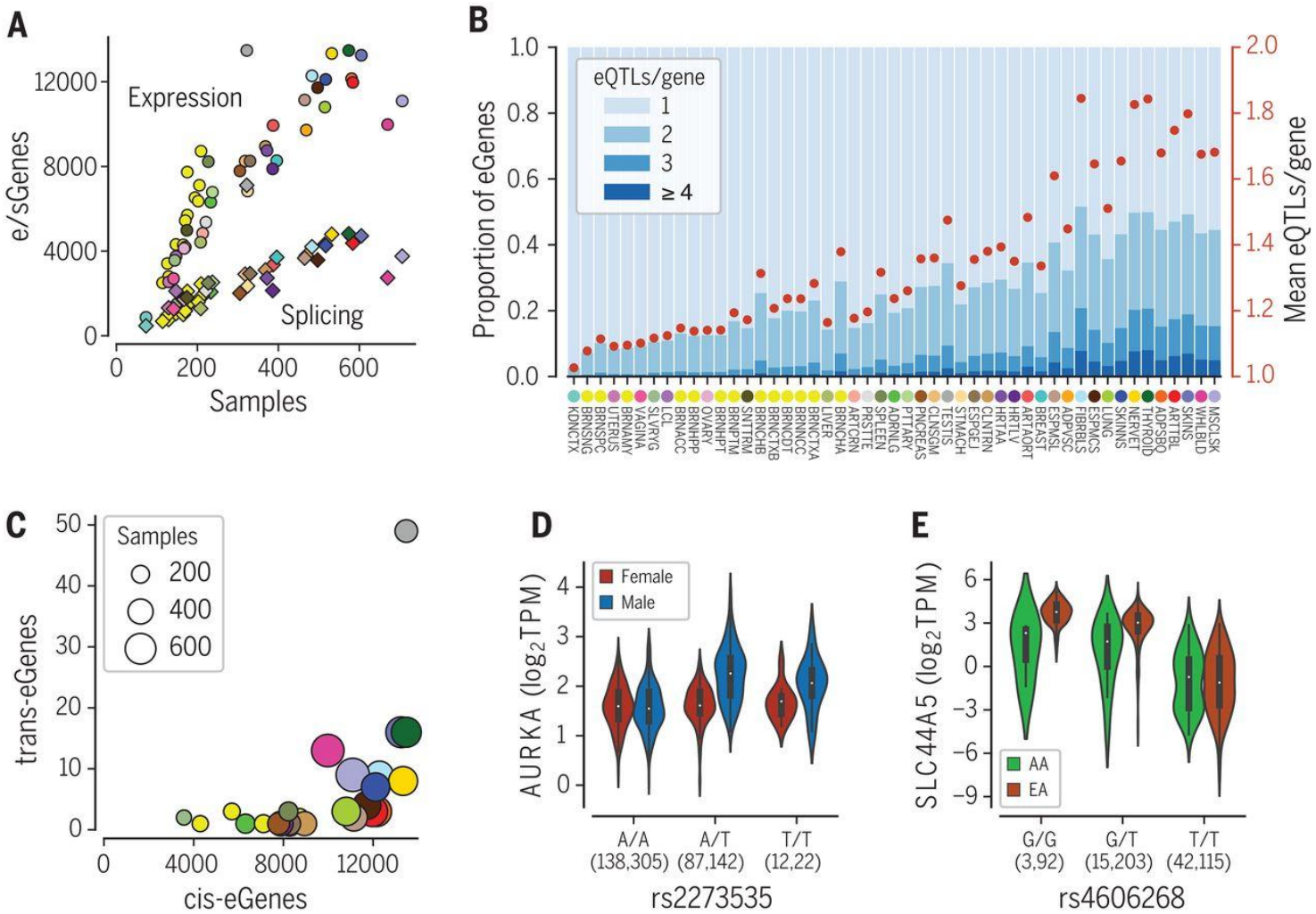
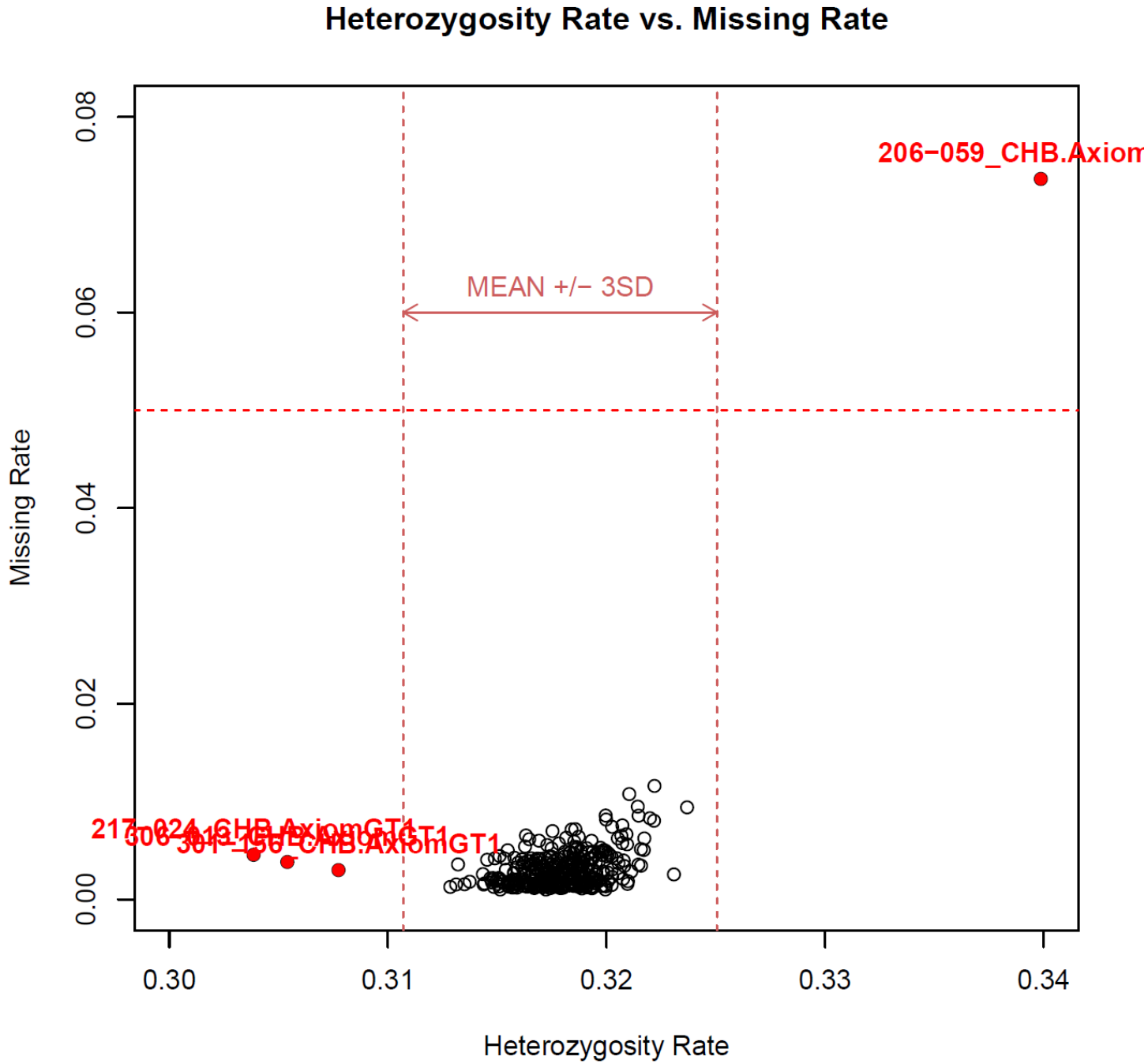
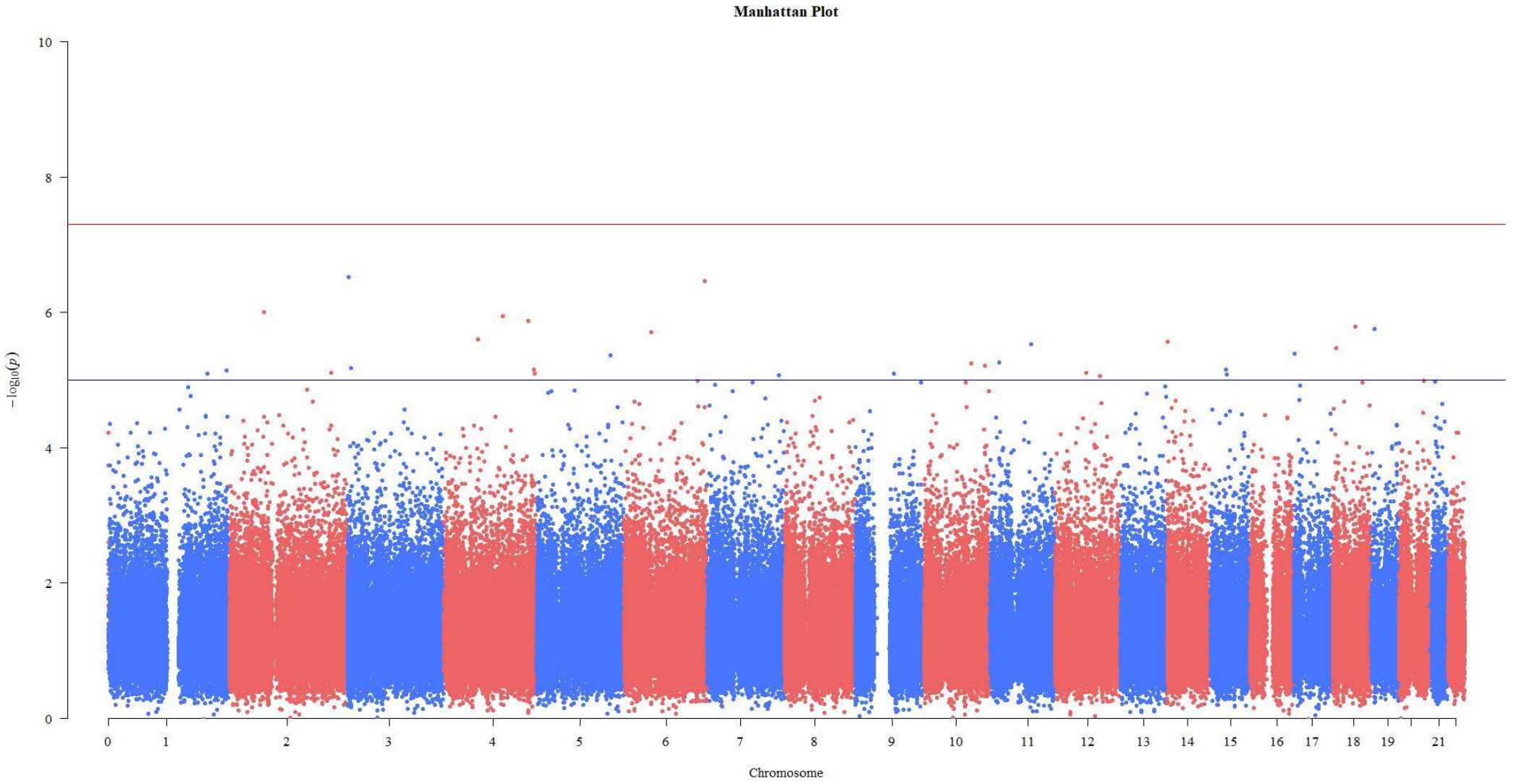


Fig. 2 QTL discovery.
(**A**) The number of genes with a cis-eQTL (eGenes) or cis-sQTL (sGenes) per tissue, as a function of sample size. See Fig. 1A for the legend of tissue colors. (**B**) Allelic heterogeneity of cis-eQTLs depicted as proportion of eGenes with one or more independent cis-eQTLs (blue stacked bars; left *y* axis) and as a mean number of cis-eQTLs per gene (red dots; right *y* axis). The tissues are ordered by sample size. (**C**) The number of genes with a trans-eQTL as a function of the number of cis-eGenes. (**D**) Sex-biased cis-eQTL for *AURKA* in skeletal muscle, where rs2273535-T is associated with increased *AURKA* expression in males ($P = 9.02 \times 10^{-27}$) but not in females ($P = 0.75$). (**E**) Population-biased cis-eQTL for *SLC44A5* in esophagus mucosa [aFC = −2.85 and −4.82 and in African Americans (AA) and European Americans (EA), respectively; permutation *P* value = 1.2×10^{-3}]. TPM, transcripts per million.


Heterozygosity Rate vs. Missing Rate
206−059_CHB.Axiom
MEAN +/− 3SD
Missing Rate
0.00
0.02
0.04
0.06
0.08
0.30
0.31
0.32
0.33
0.34
Heterozygosity Rate

Manhattan plot of demo study



Colocalization Analysis

GWAS signal (s_eddp, MMT cohort, n = 334) x whole-blood eQTLs (GTEx v8)

Bayesian coloc.abf (coloc 5.2.3) | GRCh38 | cis window 1 Mb | 17 2026

Result in one line: 12 candidate genes tested, 0 reach PP4 >= 0.5, highest PP4 = 0.26. A negative but informative control.

Pipeline: coloc/01-06 R scripts | outputs: coloc/results, coloc/figures, coloc/logs

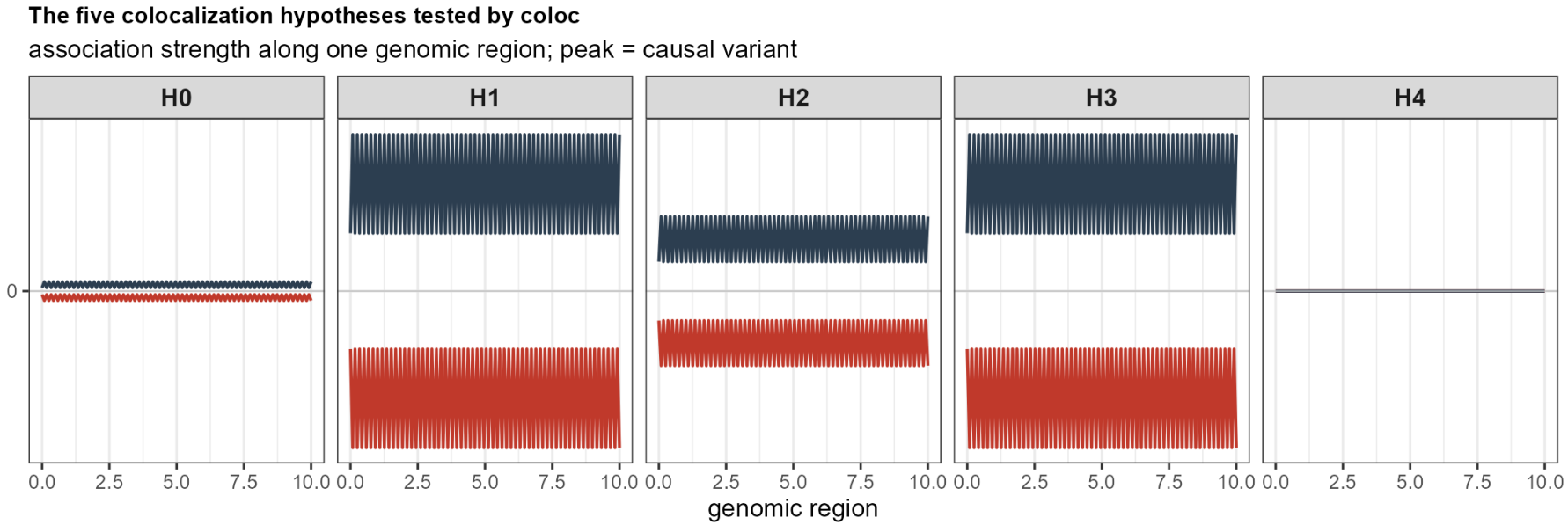
Why colocalization?

From an associated locus to a plausible causal gene

- A GWAS locus spans hundreds of kilobases and many correlated variants; it does not name a gene.
- A cis-eQTL tells which variants change the expression of a gene in a given tissue.
- Colocalization asks whether the trait signal and the expression signal in the same region are driven by ONE shared causal variant.
- High PP4 -> the gene is a credible mediator and the variant a credible causal SNP.
- High PP3 -> two distinct causal variants: the eQTL gene is probably not the mediator, even when both signals are strong.
- It is a statement about shared causal variants, not proof of mechanism.

The statistical model: five hypotheses

Giambartolomei et al., PLoS Genetics 2014 | R package coloc 5.2.3



association strength of variants along one region; a peak marks a causal variant

Per-variant evidence: Wakefield approximate Bayes factor computed from z = beta / SE, giving one log ABF per variant and per trait.

Hypothesis evidence: H1 = sum trait logABF, H2 = sum eQTL logABF, H4 = log sum of per-variant products, H3 = H1 + H2 - H4 (log scale).

Posteriors PP0-PP4 follow from the priors p1 = 1e-04, p2 = 1e-04, p12 = 1e-05 ; p12 << p1 x p2 encodes that true sharing is rare.

Convention used here: PP4 >= 0.8 strong, PP4 >= 0.5 supportive, PP3 dominant = two distinct causal variants.

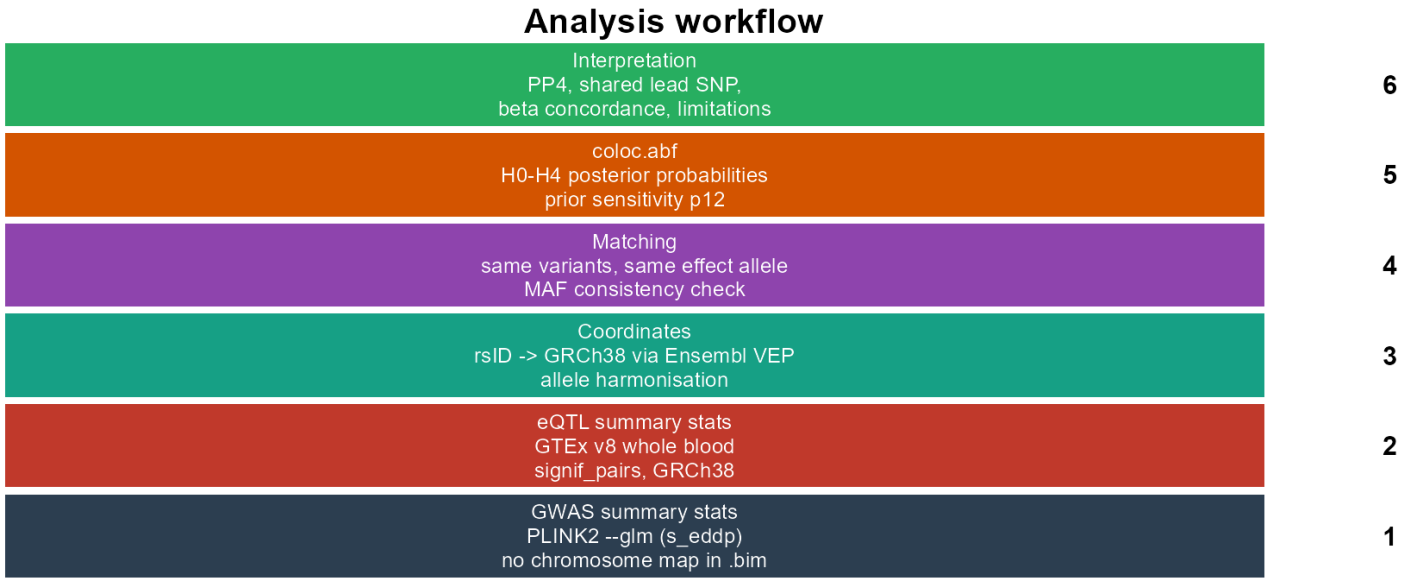
Input data actually used

Both sides are summary statistics: no individual-level data required

Item	Detail
GWAS file	mmt_atcg_rsid_cov_pc.s_eddp.glm.linear
Phenotype tested	s_eddp
Sample size / model	n = 334; additive linear model; covariates age, BMI, sex, PC1-PC5
Variants after QC	612,512 retained of 615,275
Genomic inflation	1.002
SNPs at P < 5e-8	0
eQTL file	Whole_Blood.v10.eQTLs.signif_pairs.parquet
eQTL content	2,985,690 significant gene-variant pairs; 13,903 genes; slope, slope_se, P, AF
Coordinates	GTEx: chr_pos_ref_alt_b38 (GRCh38); GWAS .bim had CHR = 0 and POS = 0
Matching route	rsID, resolved with Ensembl VEP in both directions

Harmonisation was the hard part

The GWAS .bim files carry CHR = 0 and POS = 0: coordinates were lost when probes were converted to rsIDs



The most significant GWAS rsIDs were mapped to GRCh38 with Ensembl VEP; only primary-assembly mappings were accepted, alt-contig hits were discarded.

GTEx positions were mapped in the opposite direction (position to rsID) so both studies meet on one variant list; 96% of cis positions resolved.

Of 20 candidate genes, 12 kept at least 12 variants shared with the GWAS array.

Effect alleles were aligned to the GTEx ALT allele (GWAS beta flipped when A1 equals REF); strand-ambiguous A/T and C/G SNPs were flagged.

Variants whose GTEx and cohort frequencies differ by more than 0.35 were dropped.

Locus and gene selection

Top 8 GWAS loci; up to 3 genes with the strongest whole-blood eQTL within +/- 1 Mb

Locus	Lead_SNP	GRCh38	P	Beta	MAF	Shared_variants
L1	rs5748896	chr22: 17,143,830	2.9e-06	0.553	0.118	2
L2	rs2286028	chr12: 906,303	4.0e-06	-0.673	0.072	1
L3	rs2541896	chr7: 35,488,405	4.5e-06	0.538	0.133	7
L4	rs16897045	chr8: 98,489,090	1.0e-05	0.637	0.075	3
L5	rs1344817	chr3:149,584,743	1.2e-05	-0.360	0.373	5
L6	rs1575944	chr9: 75,758,404	1.8e-05	0.720	0.058	3
L7	rs696236	chr3: 64,223,542	2.3e-05	-0.609	0.085	2
L8	rs4444765	chr4: 4,746,546	2.5e-05	-0.357	0.288	3

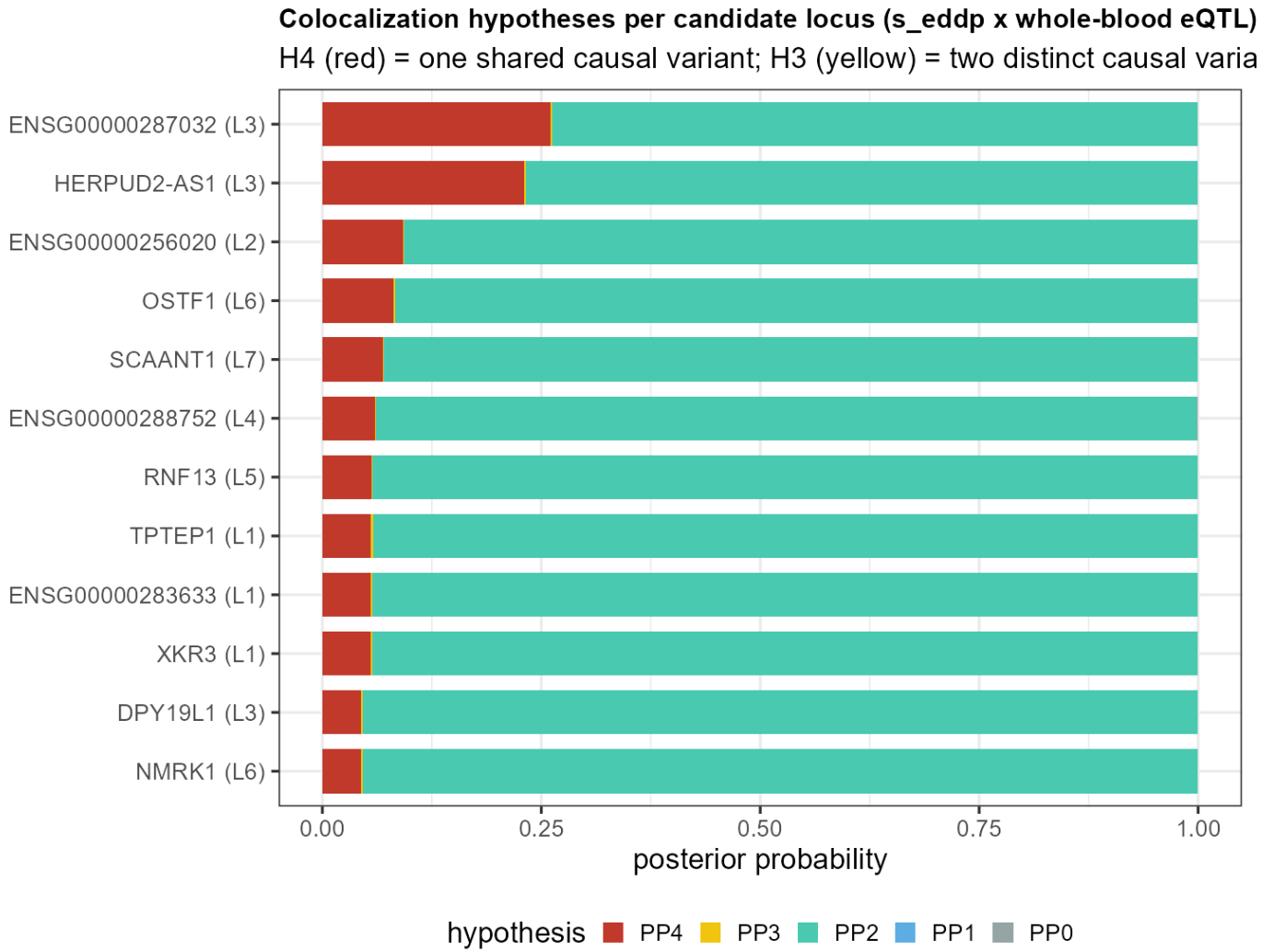
No variant reaches P < 5e-8: with n = 334 the cohort is underpowered for genome-wide significance, so this ranking is nominal and carries a winner's-curse risk.

GTEx whole blood is the opposite: eQTL P values inside these same windows fall below 1e-100.

Every asymmetry in the results that follow is a consequence of this power contrast.

Results: which hypothesis wins?

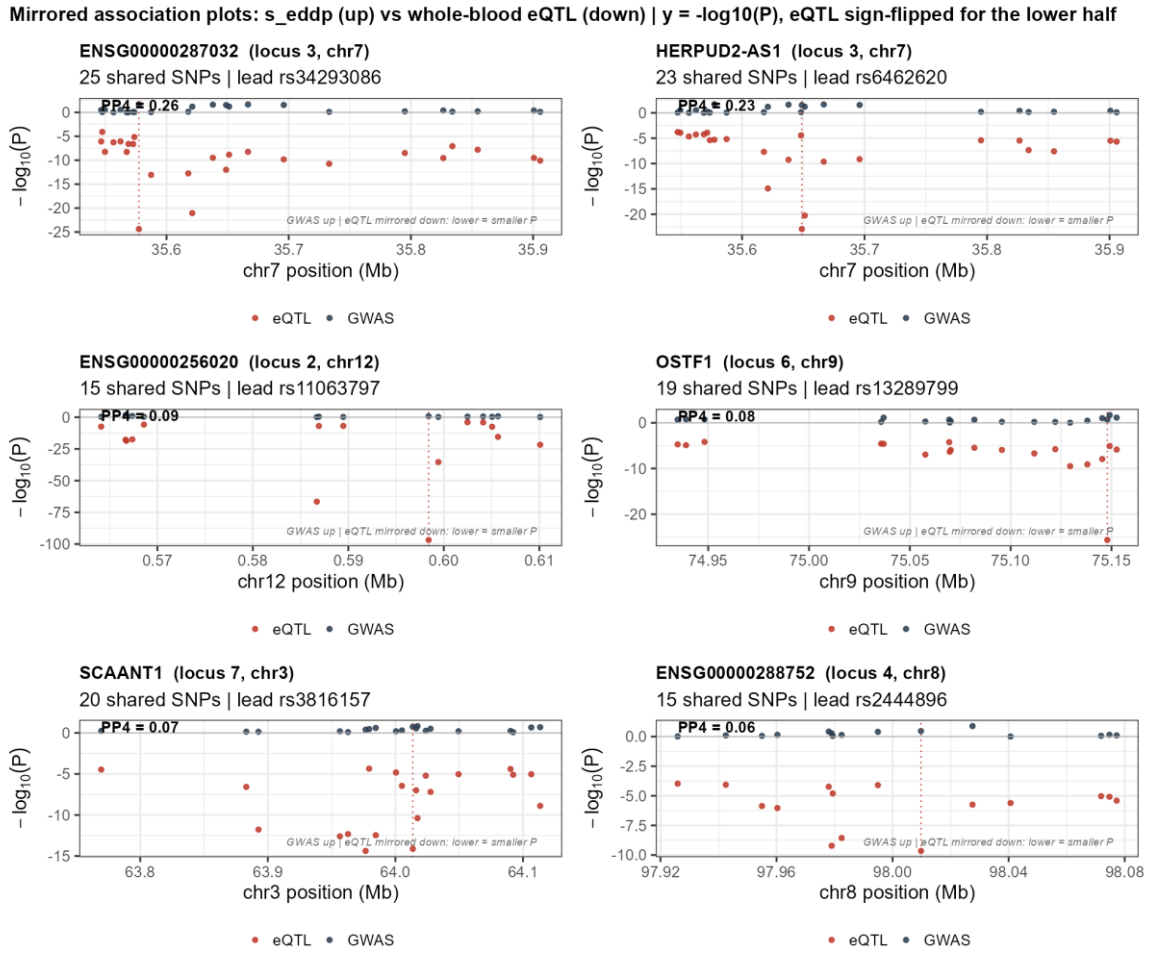
0 of 12 genes reach PP4 >= 0.5; the highest PP4 is 0.26 (novel/unnamed gene)



Gene	Loc…	Shar…	PP2	PP4	Lead_…	Call
novel/unnam…	L3	25	0.74	0.26	diffe…	single trait …
HERPUD2-AS1	L3	23	0.77	0.23	diffe…	single trait …
novel/unnam…	L2	15	0.91	0.09	diffe…	single trait …
OSTF1	L6	19	0.92	0.08	diffe…	single trait …
SCAANT1	L7	20	0.93	0.07	diffe…	single trait …
novel/unnam…	L4	15	0.94	0.06	diffe…	single trait …
RNF13	L5	17	0.94	0.06	diffe…	single trait …
TPTEP1	L1	34	0.94	0.06	diffe…	single trait …
novel/unnam…	L1	31	0.94	0.06	diffe…	single trait …

Regional evidence: a one-sided picture

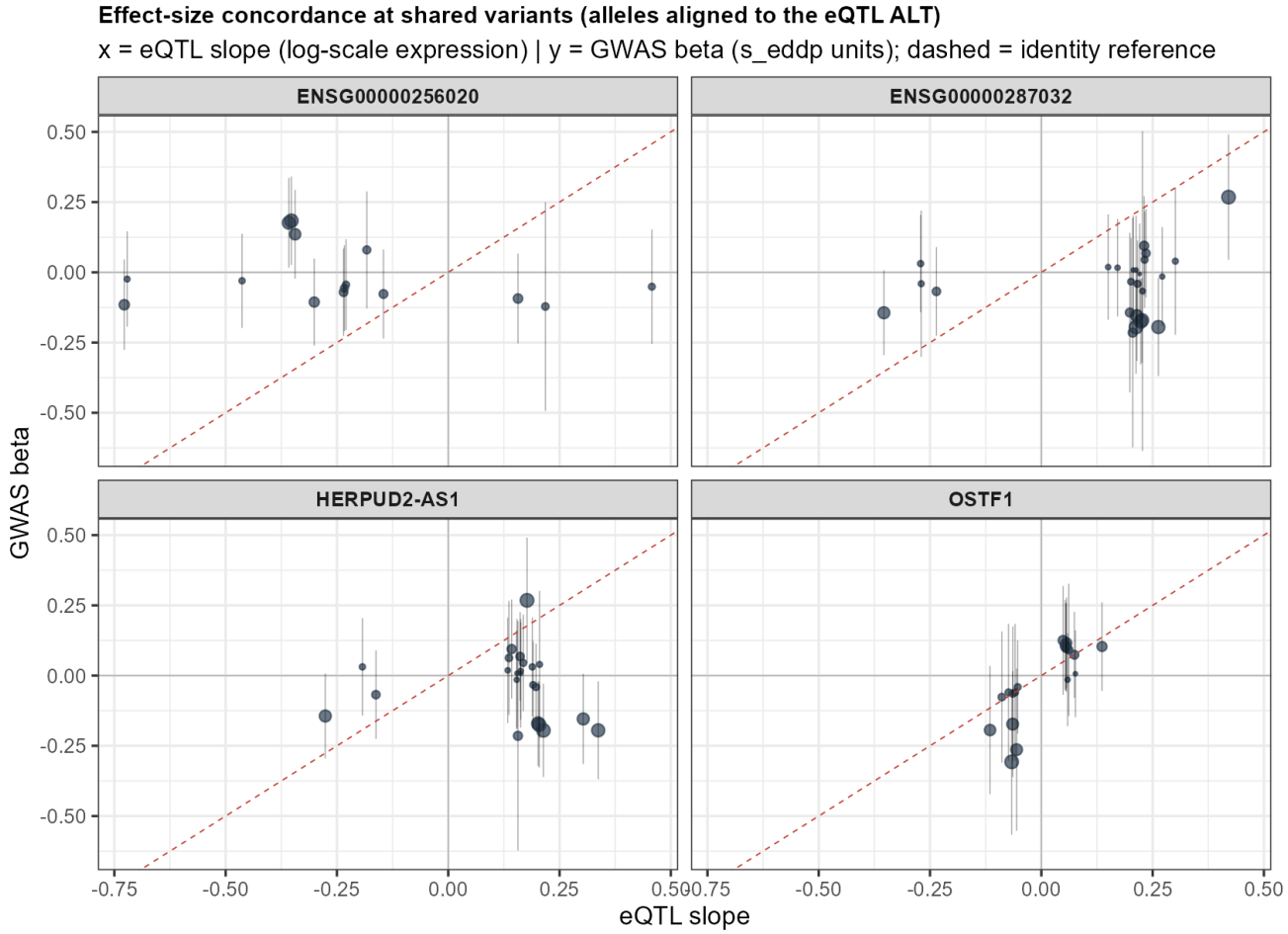
Mirrored: GWAS upward, eQTL downward (eQTL sign-flipped, so lower = smaller P); dotted line = variant with the highest H4 posterior



GWAS traces are essentially flat while eQTL traces exceed -log10(P) = 20; that is exactly why PP2 dominates over PP4

Effect-size concordance

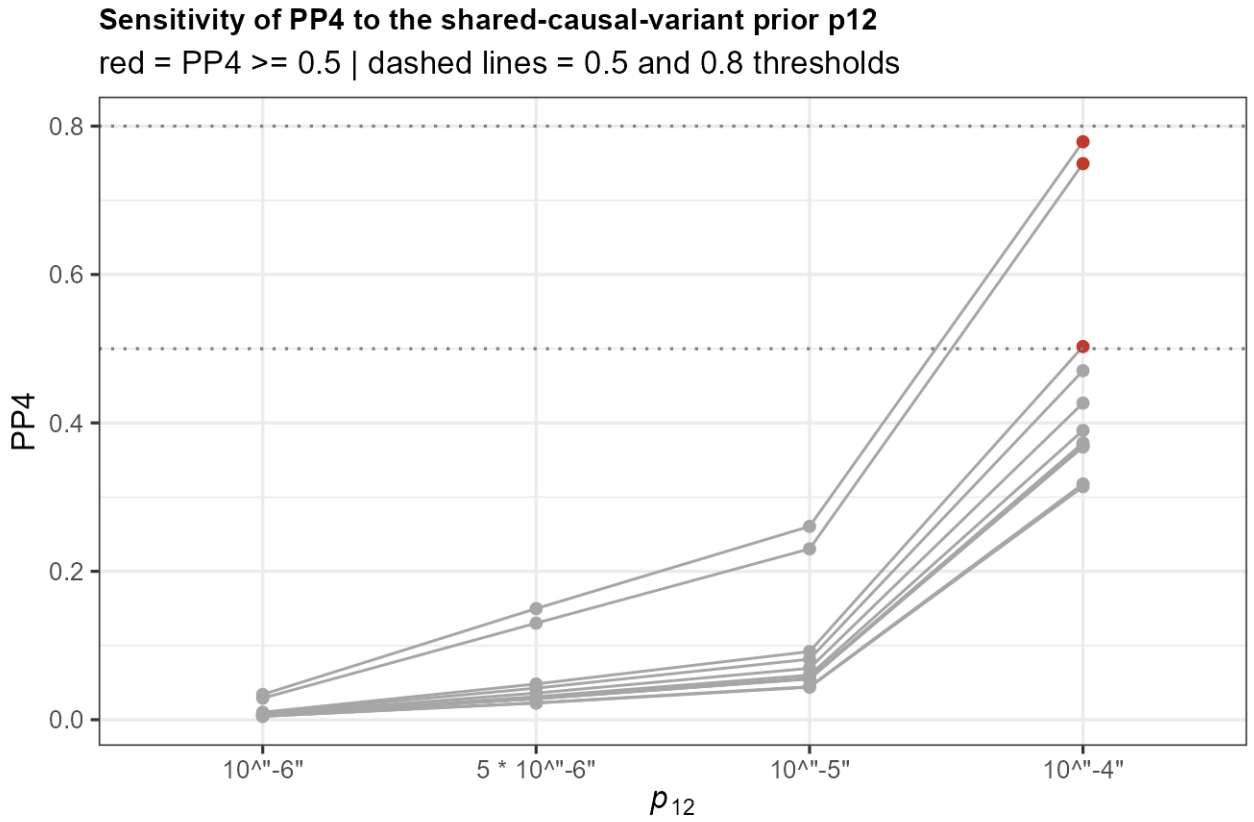
A shared causal variant must move both traits in the same direction



per-gene correlation and fraction of sign-concordant variants: results/coloc_results.tsv (beta_cor, frac_concordant)

How much can PP4 be trusted here?

PP4 depends on the sharing prior, on the variant set and on the relative power of the two studies



Read PP4 as an ordinal ranking, not a calibrated probability, when:

- only a subset of the true cis variants is available
- the studies differ in ancestry and therefore in LD
- one study is orders of magnitude better powered

All three conditions hold in this project. That is why the prior scan, the shared-variant counts and the regional plots are reported next to PP4, and why no gene is claimed as a mediator.

Limitations to state in any write-up

Ranked by how much they can move the conclusion

1. Power. n = 334 versus about 670 GTEx samples: PP4 is pushed down and PP2 up, so a genuinely shared variant can easily score PP4 < 0.5.

2. Summary-statistic source. GTEx signif_pairs holds only variants past a per-gene threshold, not the complete cis window, while coloc assumes the whole region; shared counts of 15-34 SNPs per gene follow from this and distort H3/H4.

3. Cross-ancestry LD. The cohort is East Asian (CHB array identifiers); GTEx v8 is European. Different LD means different peaks and inflates apparent H3.

4. No LD modelling. coloc.abf treats variants as independent although all listed variants are correlated, so effective information is far below the SNP count. coloc.susie with an LD matrix is the correct upgrade.

5. Single-causal-variant assumption per trait per region; loci with secondary signals are mis-modelled.

6. Nominal locus selection and 20 candidate genes tested without formal multiple-testing correction.

7. Units are not comparable: eQTL slopes are on normalised log expression, GWAS betas in native s_eddp units; only signs and z-scores transfer.

What would actually change the answer

Ordered by expected impact

1. Grow the GWAS: meta-analyse the other three phenotypes (r_eddp, r_met, s_met) and external cohorts. Highest impact by far.

2. Replace signif_pairs with the complete GTEx v8 cis association files for the candidate genes, so coloc sees the full cis window.

3. Run coloc.susie with LD from an East Asian reference panel (1000G CHB), and European LD as a comparator.

4. Test ancestry robustness with an East Asian eQTL resource and with large blood eQTL consortia (eQTLGen, eQTL Catalogue).

5. Add orthogonal statistics on the shared lead SNP: SMR/HEIDI and MR-style checks.

6. Re-run every phenotype through this pipeline: only GWAS_FILE and GWAS_PHENO change in coloc/coloc_config.R (four phenotypes are available: r_eddp, r_met, s_eddp, s_met).

7. Before any biological claim, check expression direction and tissue specificity of the top-PP4 genes.

Reproducibility

Six scripts regenerate every number and every figure

Script	Role
coloc/coloc_config.R	parameters, priors, QC thresholds, allele harmonisation utilities
coloc/ens_api.R	cached Ensembl VEP and lookup wrappers with retry and rate-limit handling
coloc/01_gwas_prepare.R	GWAS QC, lambda_GC, QQ plot, GRCh38 mapping, locus definition
coloc/02_select_genes.R	GTEx table, TSS reconstruction, candidate gene selection
coloc/03_prepare_coloc_inputs.R	position-to-rsID mapping, allele alignment, MAF checks, variant join
coloc/04_run_coloc.R	coloc.abf per gene plus the p12 prior sensitivity scan
coloc/05_figures.R	figures 04-09
coloc/06_make_slides.R	this deck

Results: coloc_results.tsv, coloc_sensitivity.tsv, candidate_genes.tsv, coloc_inputs_summary.tsv, gwas_top_loci.tsv, variant_mapping_summary.tsv, coloc_report.txt

Figures: fig01 QQ, fig02 locus positions, fig03 gene selection, fig04 posteriors, fig05 regional, fig06 concordance, fig07 prior scan, fig08-09 schematics

Logs: coloc/logs/coloc_run.log keeps the full provenance of every stage, including mapping rates and filtering counts.