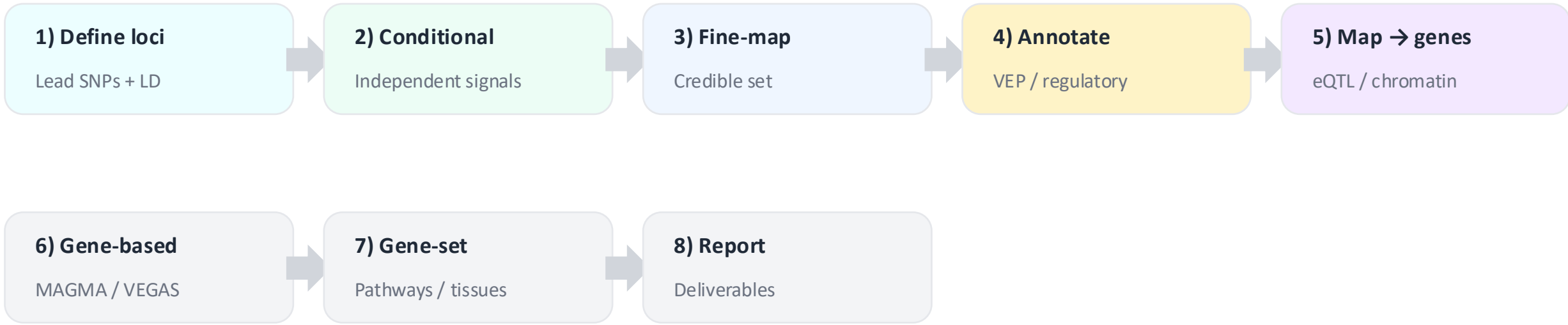
Post-GWAS analyses

Outline

- How to identify independent association signals (conditional analysis)
- How to move from variants → likely genes (annotation + mapping)
- How to summarize evidence at gene/pathway level (gene-based + enrichment)

Post-GWAS workflow – sequence of increasingly biological questions



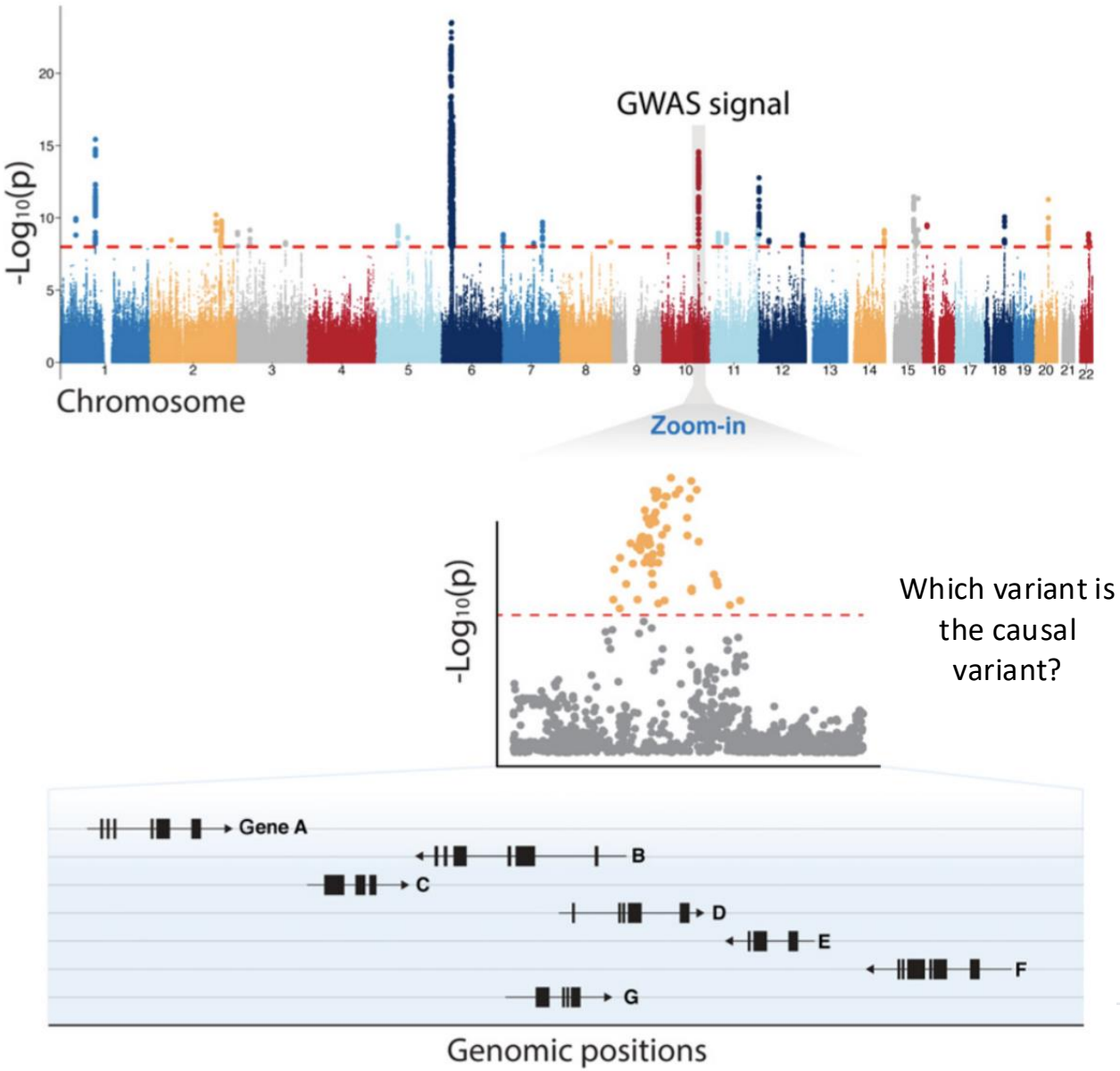
Why does it matter?

- Clarifies disease biology
- Informs therapeutic target discovery
- Enables biomarker identification

Signals & LD

Why “lead SNP → gene” is usually wrong

GWAS results: many signals, many questions



- A “hit” is a genomic region, not automatically a gene.

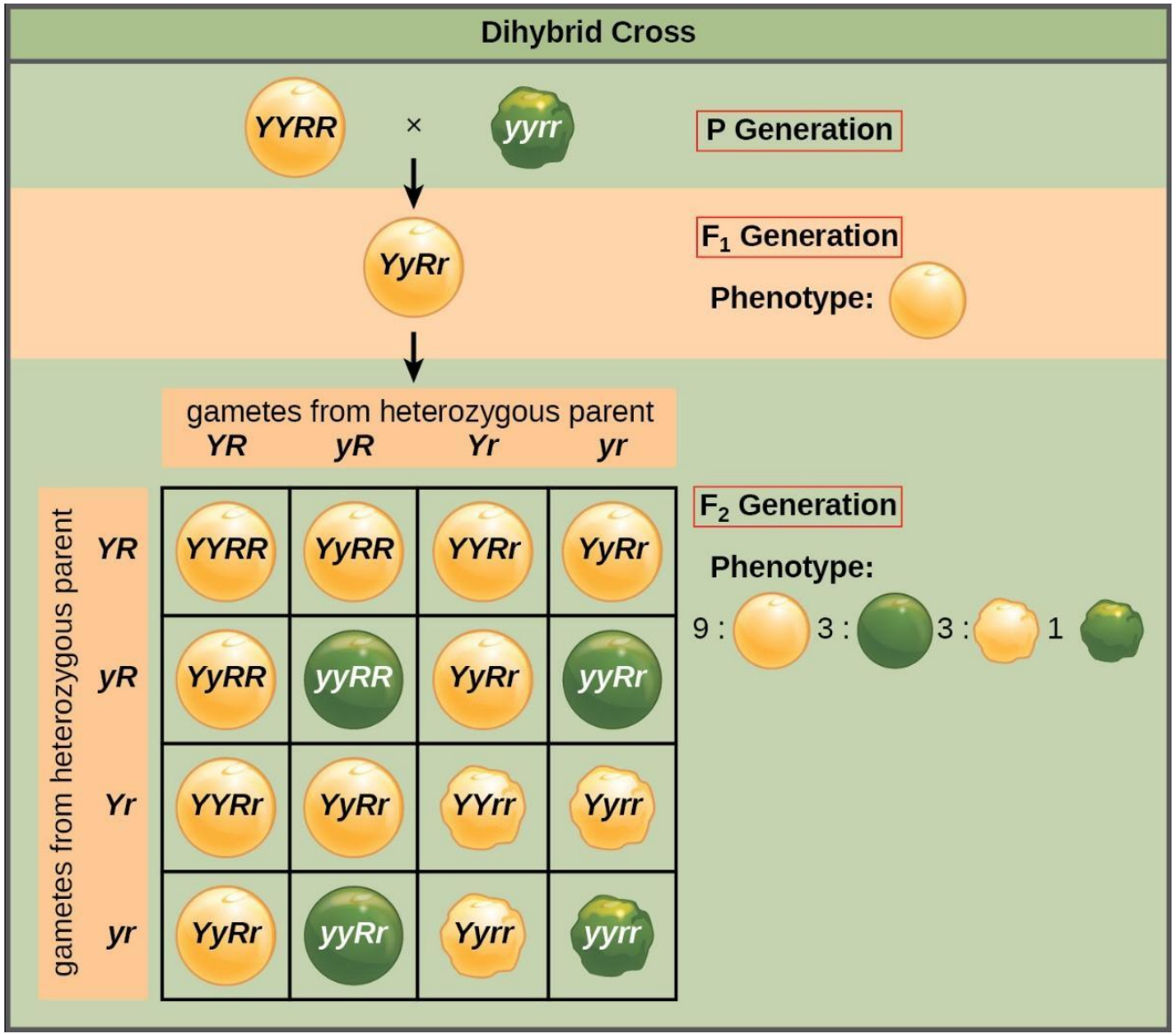
Region is a set of correlated variants that show a statistically significant association with the trait of interest, but linkage disequilibrium prevents pinpointing causal variants without further analysis

What is Linkage Disequilibrium?

Nearby genetic variants are inherited together

Mendel's 2nd law (~1850): law of independent assortment

Genes do not influence each other with regard to sorting alleles into gametes, and every possible combination of alleles for every gene is equally likely to occur.



Early 1900s: Bateson & Punnet discover genetic linkage

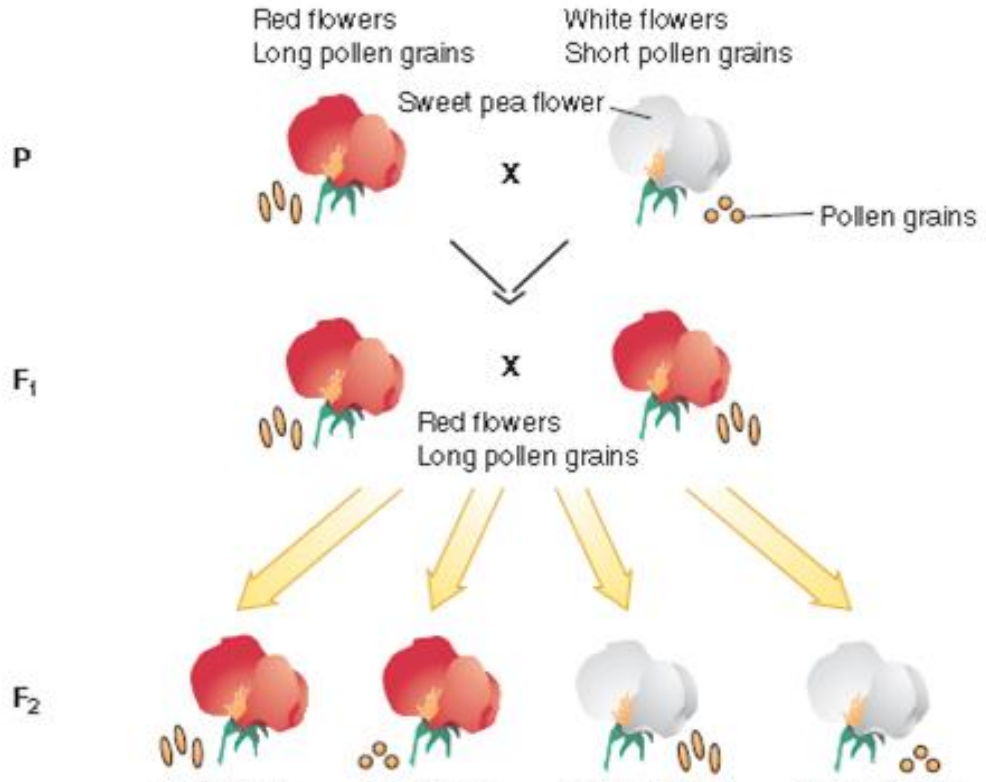
Expected 9:3:3:1 ratio but
Observed 24.3:1.7:1.1:7.1 ratio instead

Excess of parental phenotypes (i.e. red long & white short)

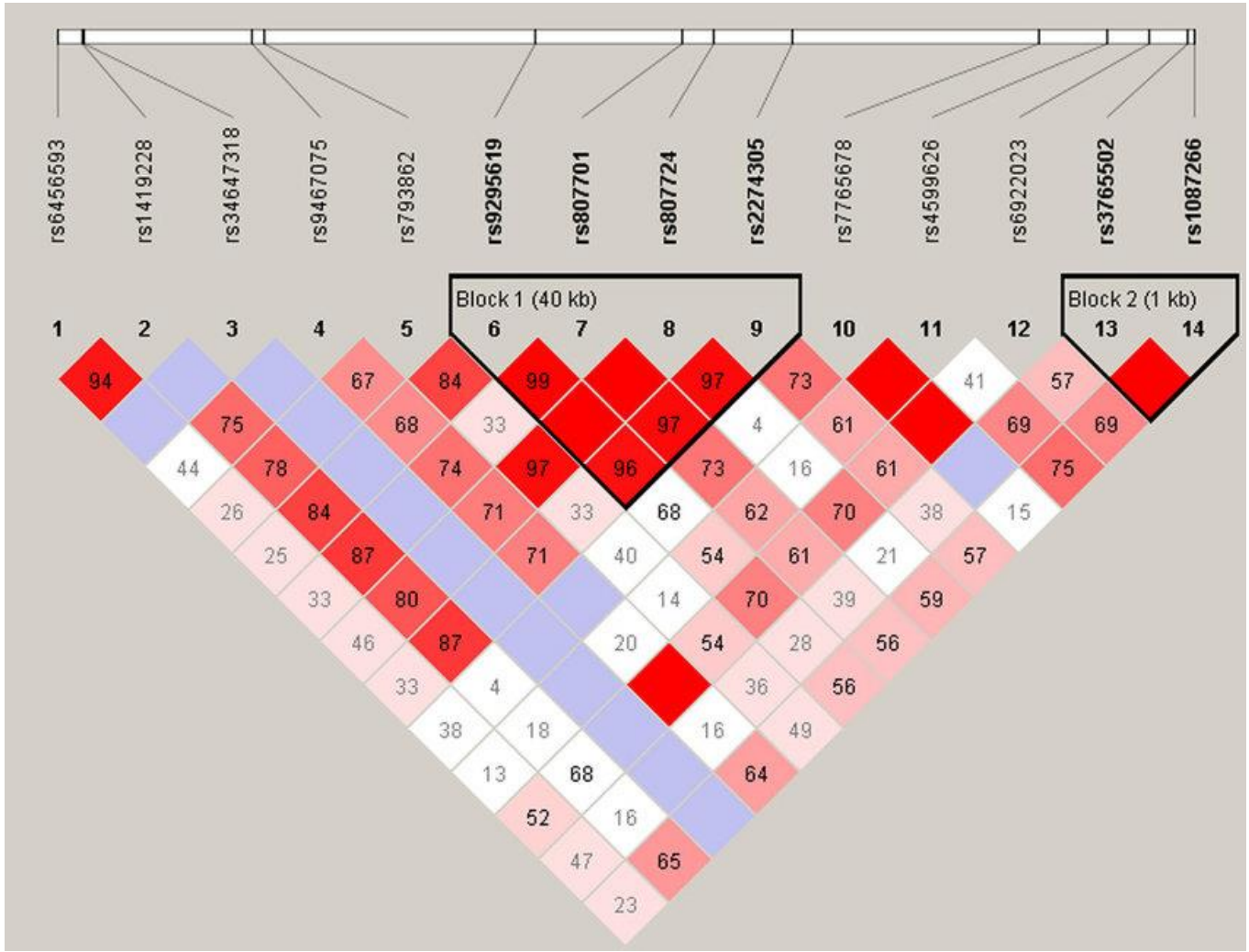
Explanation?

Genes for flower color and pollen length are located on the same chromosome.

They are **linked** and are more likely to be inherited together

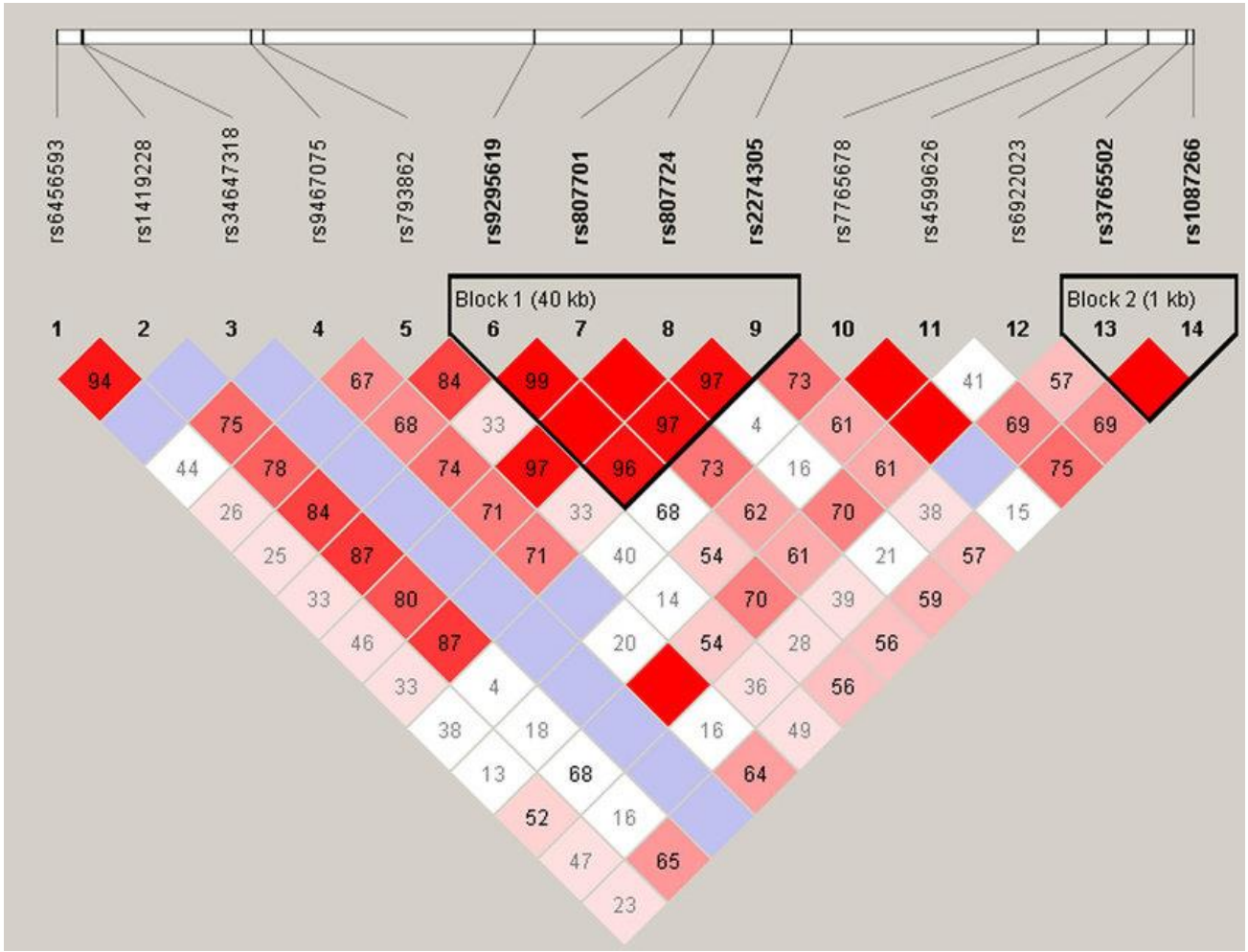


Linkage disequilibrium: nearby genetic variants are inherited together



- Recombination doesn't break up DNA evenly
- As a result, the genome is organized into **blocks of correlated variants**
- If you inherit one variant, you're likely to inherit its neighbors too.
- LD is a population-level property: it tells us about correlation, not causation.

The top GWAS hits are usually not functional/causal



- When one SNP is significant, **many nearby SNPs are also associated**
- This is because they are **correlated through LD**
- The strongest signal is often:

  The *best tag*, not the *true causal variant*

GWAS identifies an **associated region (locus)**
Not a single gene
Not a single variant
Not a mechanism

Linkage disequilibrium: the African context is special

- African ancestry populations often have shorter LD (smaller blocks) due to older population history this is a scientific advantage for fine-mapping:
    - GWAS signals are more precise
    - Fewer variants are correlated
    - Easier to narrow down causal variants
- But LD mismatch can break methods that need an LD reference.
    - Always describe ancestry + LD panel used.

Conditional analysis

aka “COJO = Conditional and Joint analysis”

Goal: Finding independent signals within a locus

Asks: “Is this GWAS signal still there once we account for/adjust for the strongest nearby variant?”

COJO **(1) removes the effect of the top SNP**
(2) then asks whether **any other SNPs in the region still show an independent association**

It helps you figure out whether a locus contains: **one signal**, or **multiple independent signals**

Why do conditional analysis?

- Question: “How many independent association signals are here in this region?”
- Motivation: allelic heterogeneity is common (≥2 signals per locus).
- Output: a set of approximately independent lead variants + their joint effects.
- This guides fine-mapping, functional follow-up, and polygenic risk score construction.

When it helps

- Two peaks in a locus plot
- Known locus with multiple signals (e.g., kidney transport genes)
- Meta-analysis with large N

When to be careful

- LD reference not matched
- Rare variants / imputation issues
- Highly heterogeneous cohorts

Yang et al. 2012 (GCTA-COJO)

Conditional analysis

> “Use GWAS summary statistics + an LD reference panel to approximate what would happen if we ran a multiple-SNP regression in the locus.”

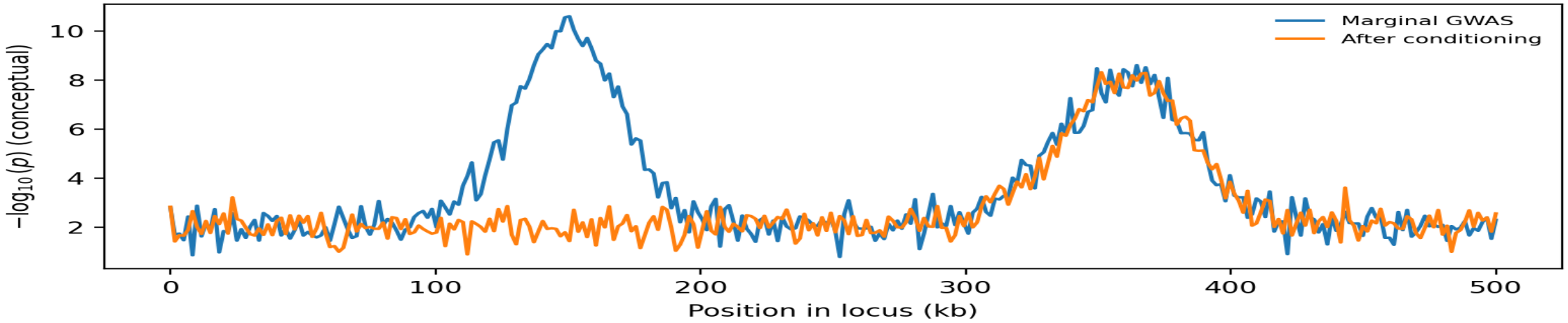
- Inputs: (1) summary stats, (2) LD matrix (from a reference), (3) parameters: window size, p-thresholds.
- Outputs: independent variants, joint betas/SEs, residual signals.
- Interpretation: “approximately independent given the LD reference.”

Example: GCTA-COJO

```
# 1) Prepare summary stats: SNP A1 A2 freq beta se p n
# 2) Choose an ancestry-matched LD reference (e.g., 1000G AFR, or your cohort)

# Stepwise selection of independent signals

./gcta64 \
+  --bfile LD_REFERENCE_1000G_AFR \
+  --cojo-file eGFR_GWAS.sumstats.gz \
+  --cojo-slct \
+  --cojo-p 5e-8 \
+  --cojo-wind 1000 \
+  --out eGFR_COJO

# Optional: condition on known SNP(s)
./gcta64 --bfile LD_REFERENCE_PREFIX --cojo-file eGFR_GWAS.sumstats.gz \
+  --cojo-cond rs12345 rs67890 --out eGFR_COJO_cond
```

Yang et al. 2012 Nat Genet

How to read a conditional result



Conceptual locus plot (illustration)

- If a second peak remains after conditioning → likely a second independent signal.
- If everything collapses → one main signal explains the locus (given the LD reference).
- Your “independent signals” table becomes the starting point for downstream steps.

Common pitfalls in conditional analysis

Pitfall

- LD reference mismatched to the study population
- Different imputation panels / QC across cohorts
- Long-range LD or structural variation

Practical mitigation

- Use ancestry-matched LD; sensitivity analysis with 2 panels
- Remove poorly imputed SNPs; harmonize allele/position
- Document filtering thresholds and window sizes
- Cross-check with regional plots + replication

Fine-mapping

From “independent signal” → “credible set of variants”

Fine-mapping: what question does it answer?

- Question: “Which variants are most likely causal (or tagging causal) within a signal?”
- Output: a *credible set* (a small list of variants) with probabilities (PIPs) associated with each one.
- Why it helps: you can focus functional assays on a smaller set.
- Caution: fine-mapping is only as good as LD reference + model assumptions.
- GWAS tells you *where* in the genome something is happening.

Fine-mapping tries to identify *which exact variant(s)* are responsible.

Reports which subset of SNPs best explains the observed vector of GWAS effect estimates, given their correlation structure?

Common outputs

- 95% credible set size
- Posterior inclusion probability (PIP)
- Sometimes multiple credible sets per locus

Popular tools

- FINEMAP (Bayesian, summary stats)
- SuSiE / SuSiE-RSS (credible sets)

…plus many others

Benner et al. 2016 (FINEMAP); Wang et al. 2020 (SuSiE); Zou et al. 2022 (SuSiE-RSS).

Fine-mapping terminology

Credible sets vs. lead SNPs

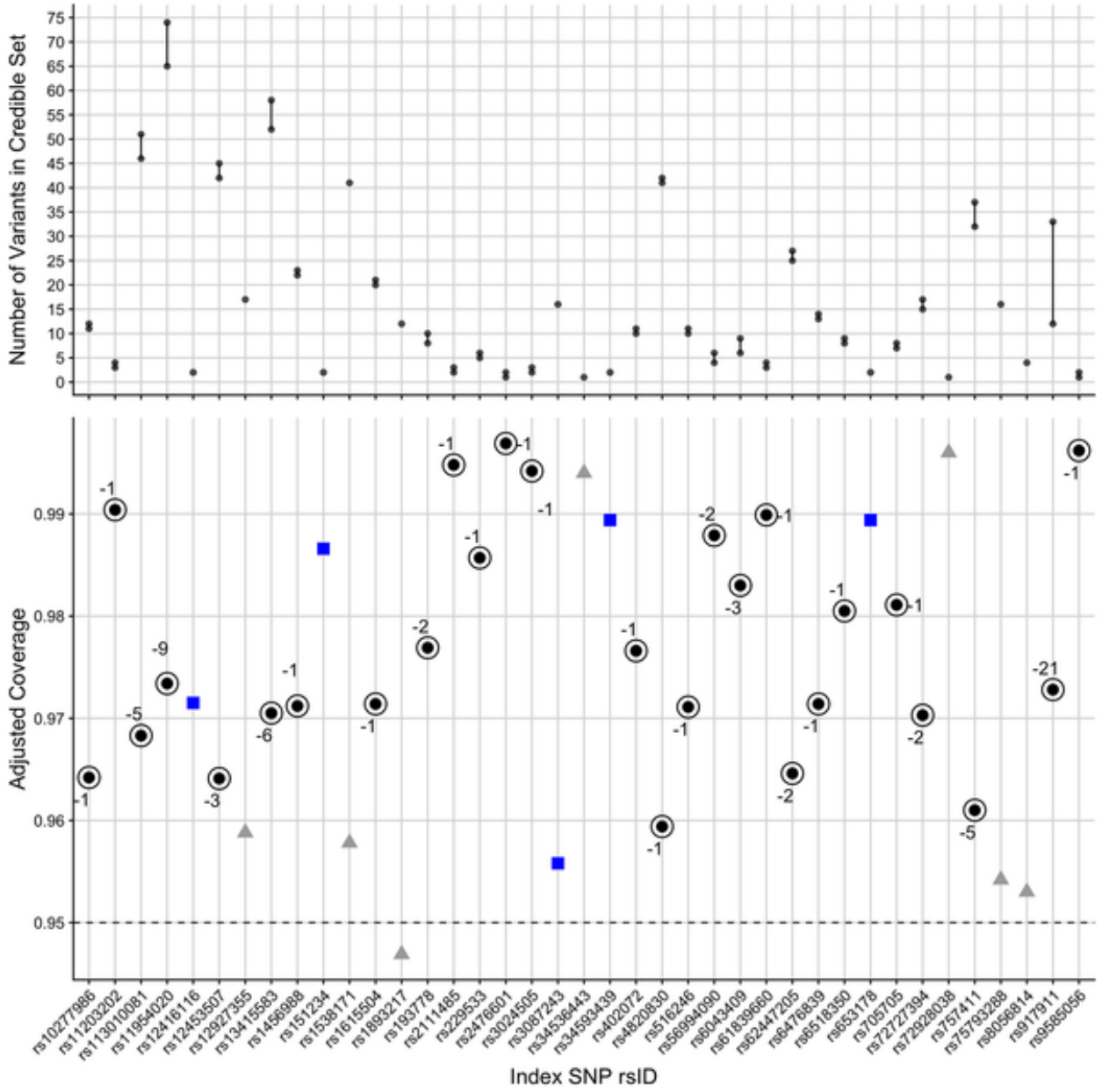
Lead SNP

- The single SNP with the **smallest p-value** at a locus
- Easy to report
- Often **not causal**
- Highly dependent on LD and sample size

Credible set

- A **group of SNPs** that together have a high probability (e.g. 95%) of containing the true causal variant
- Each SNP has a **posterior probability of causality**
- Accounts for LD and uncertainty

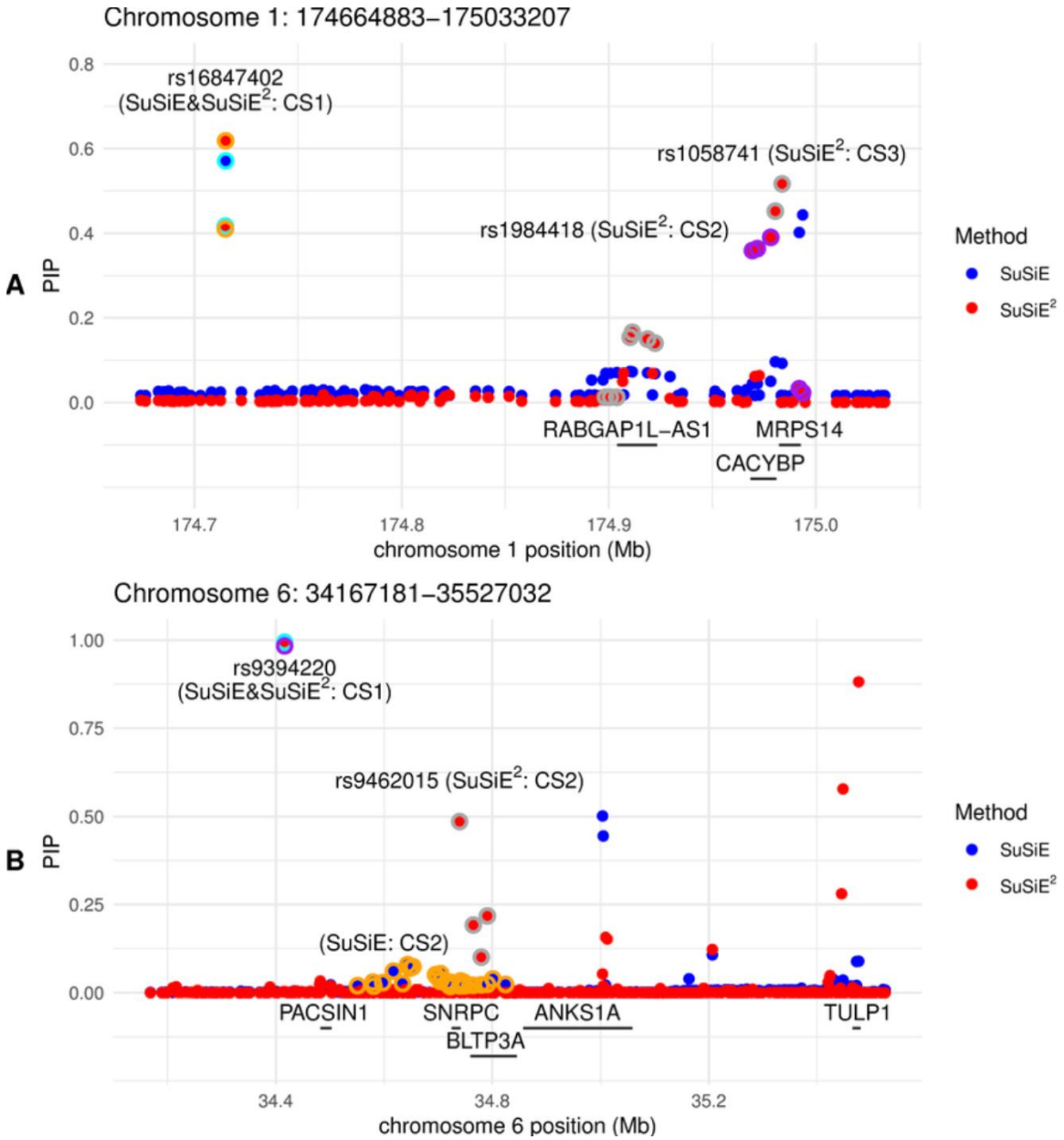
Fine Mapping results example



95% credible set
This is the **smallest set of SNPs whose PIPs add up to ≥ 95%**.

PIP = probability this SNP is truly causal, given the GWAS data and LD

Adjusted coverage:
How confident you should be that the credible set really contains the true causal variant, after accounting for model uncertainty and LD structure.



Posterior inclusion probability (PIP):

the probability that a specific genetic variant is truly causing the GWAS signal

Lets you:
Rank variants by likelihood of causality
Accept uncertainty instead of forcing a single “lead SNP”
Decide which variants to study first in functional work

Rule of thumb:
- PIP > 0.8 → high priority
- 0.2–0.8 → plausible
- < 0.1 → unlikely

Fine-mapping software programs

Software	Best at	Why people use it
SuSiE	Multiple signals per locus	Clean, interpretable, modern
FINEMAP	Complex loci	Powerful and flexible
CAVIAR	Simple loci	Conservative and clear
PAINTOR	Functional integration	Adds annotations
DAP-G	Large datasets	Scales well
ABF (Wakefield)	Quick screening	Fast and simple

Variant annotation

For each SNP in a credible set, the next question to ask is what does this variant “do” biologically?

Variant annotation

- Goal: add biological context to lead / credible variants.
- Coding consequences (missense, stop-gain) are easiest to interpret — but less common in GWAS hits.
- Non-coding variants may act through regulation (enhancers, promoters, splicing).
- Annotation ≠ proof of causality. It is prioritization.

Typical annotations to report

- Nearest gene(s) + distance
- Predicted consequence per transcript
- Population allele frequencies
- Known trait associations

Kidney-relevant additions

- Kidney eQTL / sQTL evidence
- Cell-type expression (tubule, podocyte)
- Chromatin marks in kidney tissue

Basic Annotations

For each SNP in a credible set:

Is the SNP inside a gene or between genes (intergenic) ?

If it is inside a gene, is it in an exon (aka coding region of gene) or intron (aka noncoding region)?

If it is a coding variant, is it:

- Synonymous: no amino acid change (usually low priority)
- Missense: amino acid change
- Nonsense / frameshift: truncates protein (high impact)

Does it overlap regulatory elements?

- Promoters
- Enhancers
- Open chromatin
- Histone marks (H3K27ac, H3K4me1, etc.)

What gene(s) does it touch or regulate?

Basic variant annotations come from public reference databases
FUMA conveniently bundles many of them in one place

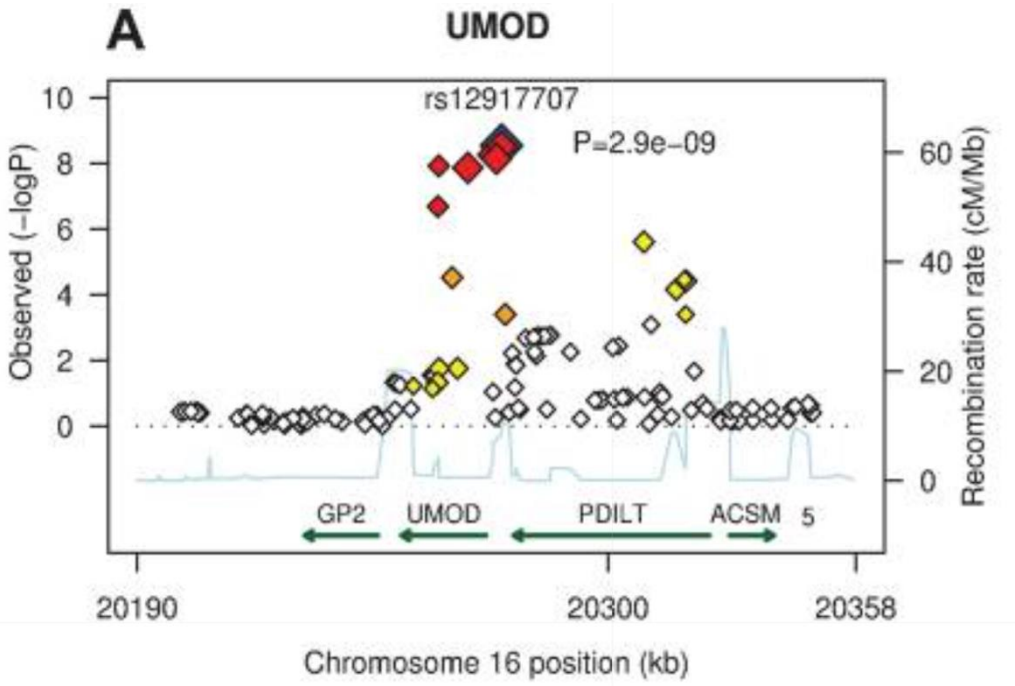
Basic Annotations, example

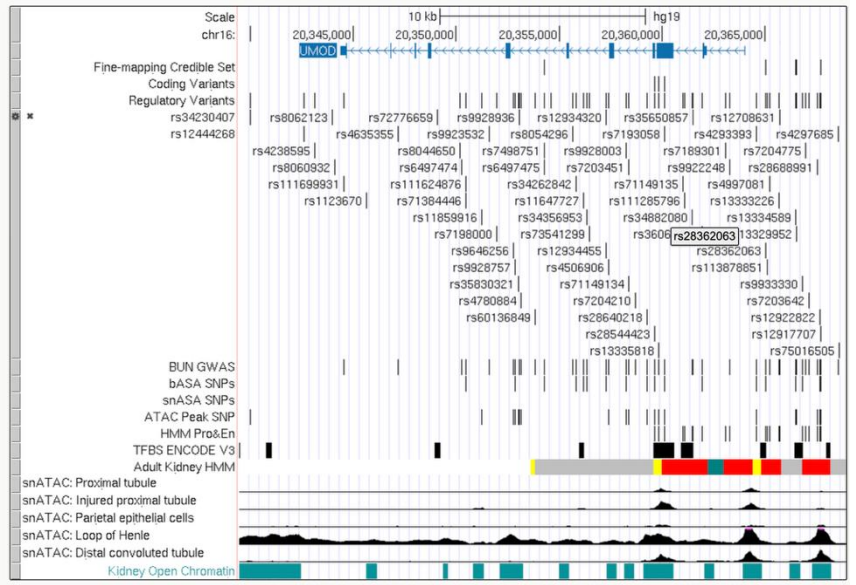
SNP: rs12917707
GWAS trait: eGFR
Paper: Kottgen et al. *Nature Genetics* 2009 (first identified)

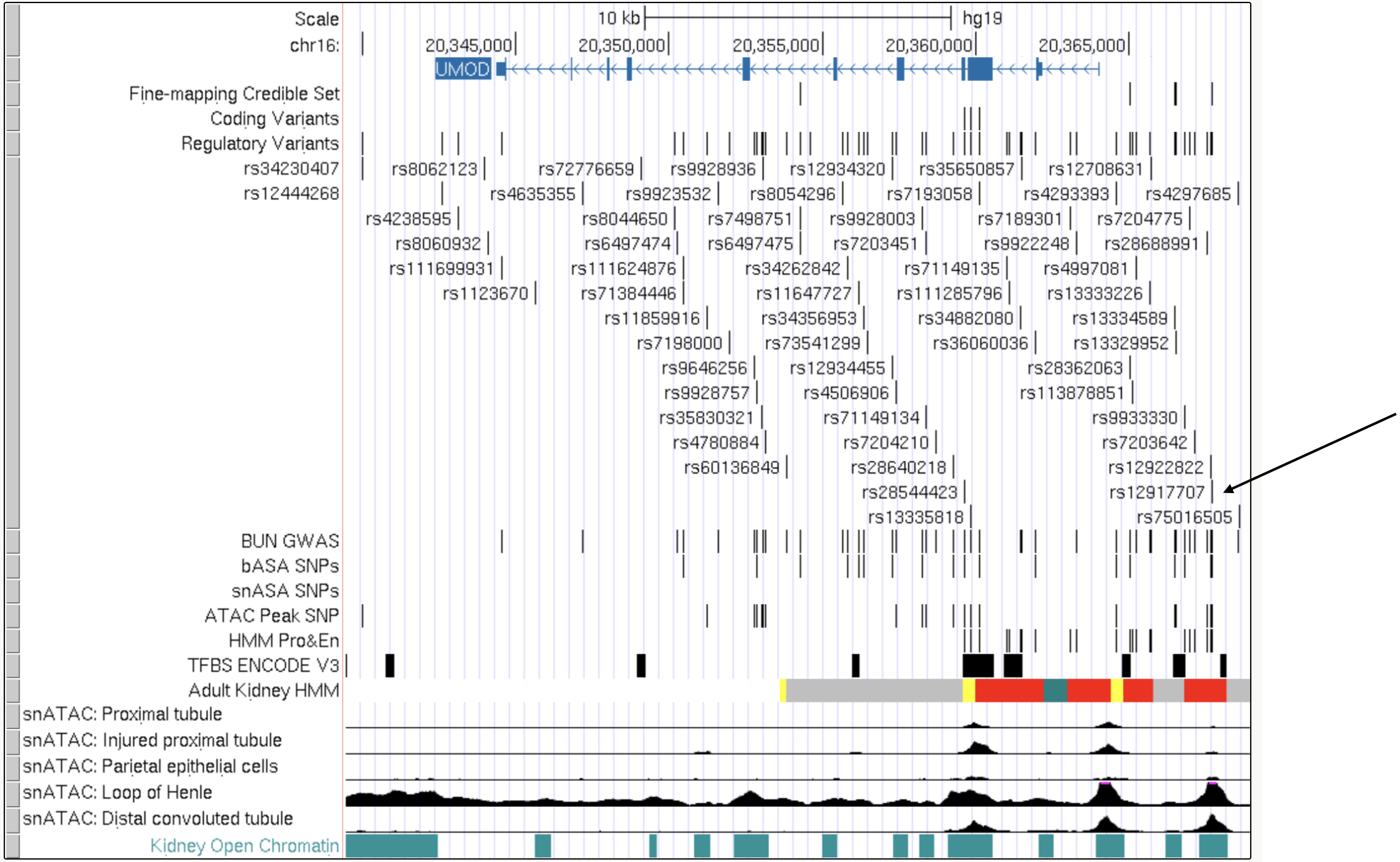
Basic GWAS context:
Location: Chromosome 16
Strong, replicated association with eGFR

FUMA:
Variant type: noncoding
Location: ~3 kb upstream of **UMOD**
This SNP does not alter the uromodulin protein sequence
Regulatory element?
This SNP overlaps promoter/enhancer marks and open chromatin
Activity seen in thick ascending limb

From kidney tissue studies:
rs12917707 is a **cis-eQTL for UMOD**
Risk allele → **higher UMOD expression**
This variant controls UMOD expression levels in the kidney.

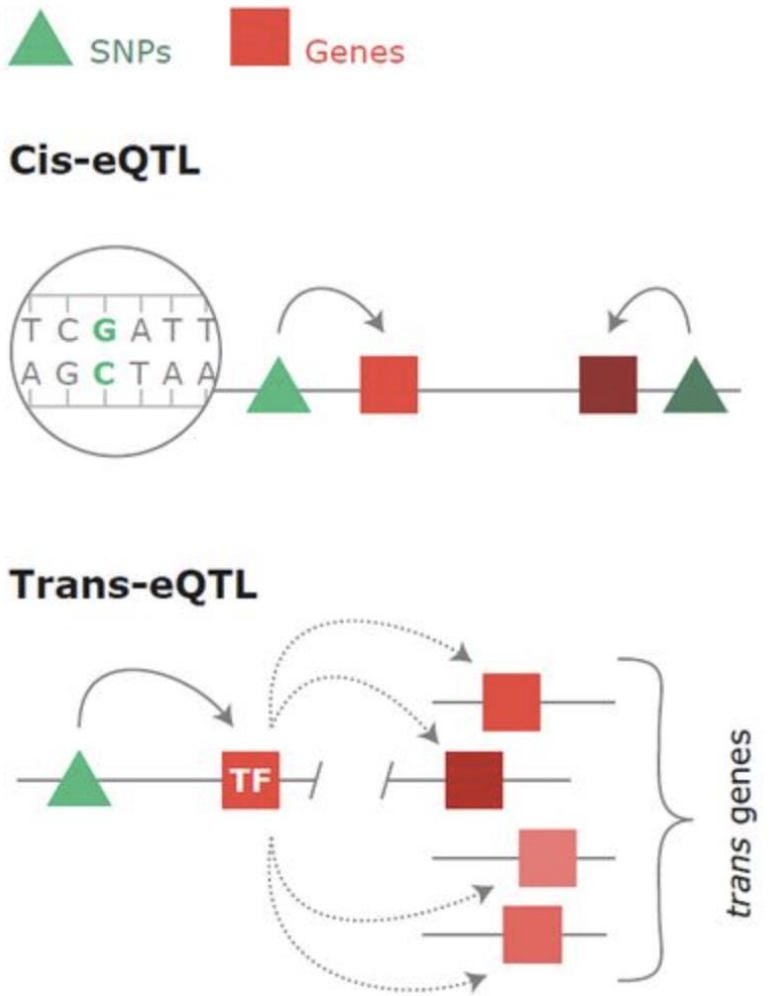





Scale
10 kb
hg19
chr16:
20,345,000
20,350,000
20,355,000
20,360,000
20,365,000
UMOD
Fine-mapping Credible Set
Coding Variants
Regulatory Variants
rs34230407
rs8062123
rs72776659
rs9928936
rs12934320
rs35650857
rs12708631
rs12444268
rs4635355
rs9923532
rs8054296
rs7193058
rs4293393
rs4297685
rs4238595
rs8044650
rs7498751
rs9928003
rs7189301
rs7204775
rs8060932
rs6497474
rs6497475
rs7203451
rs9922248
rs28688991
rs111699931
rs111624876
rs34262842
rs71149135
rs4997081
rs1123670
rs71384446
rs11647727
rs111285796
rs13333226
rs11859916
rs34356953
rs34882080
rs13334589
rs7198000
rs73541299
rs36060036
rs13329952
rs9646256
rs12934455
rs28362063
rs9928757
rs4506906
rs113878851
rs35830321
rs71149134
rs9933330
rs4780884
rs7204210
rs7203642
rs60136849
rs28640218
rs12922822
rs28544423
rs12917707
rs13335818
rs75016505
BUN GWAS
bASA SNPs
snASA SNPs
ATAC Peak SNP
HMM Pro&En
TFBS ENCODE V3
Adult Kidney HMM
snATAC: Proximal tubule
snATAC: Injured proximal tubule
snATAC: Parietal epithelial cells
snATAC: Loop of Henle
snATAC: Distal convoluted tubule
Kidney Open Chromatin

Expression quantitative trait loci (eQTL)

- A genetic variant associated with **gene expression levels**
- Links **DNA variation → RNA abundance**
- Can act **locally (cis)** or **at a distance (trans)**
- Strongly **tissue- and cell-type specific**



Colocalization (coloc)

Asks the question: Is the GWAS signal and the gene-expression signal driven by the same variant?

GWAS signal:
•rs12917707 associated with eGFR / CKD

Molecular signal:
•rs12917707 associated with UMOD expression in kidney tissue (eQTL)

Inferential statistical test giving the probability that the same variant is associated both with eGFR and with UMOD expression
•Under the hood: Compares the two signals (GWAS and molecular) across the locus
Accounts for linkage disequilibrium
Estimates the probability they share one causal variant

Coloc for rs12917707: **PP4 = 0.92**

Interpretation: There is a 92% probability that the GWAS signal and the molecular signal are driven by the *same causal variant*.

Even more annotations!

Especially for coding variants and rare variants, use Variant Effect Predictor (VEP)
VEP translates genomic coordinates into biological consequences

- Annotates genetic variants with predicted gene and protein consequences
- Identifies which gene(s) and transcript(s) the variant affects
- Classifies variant type:
  Intergenic, intronic, missense, nonsense, splice-site
- Reports protein-level changes (e.g., amino acid substitutions, truncations)
- Supports loss-of-function classification
- Provides variant-level detail, not locus-level prioritization
- Complements GWAS fine-mapping and regulatory annotation tools like FUMA

Even more annotations, example

SNP: rs2231142
GWAS traits: serum uric acid, gout
Gene: *ABCG2 (uric acid transporter)*

Input variant into VEP and get the following:

- Location: coding exon of *ABCG2*
- Missense variant (changes the protein sequence)
- Amino acid change: p.Gln141Lys (Q141K)
- Affects an ATP-binding cassette transporter
- Known to reduce transporter expression and activity
- *ABCG2* transports urate
- Risk allele → reduced urate efflux
- Leads to higher serum uric acid and gout risk

VEP directly links the GWAS signal to a specific amino acid change in a known urate transporter, making the causal mechanism immediately interpretable.

Gene-based testing

Summarize SNP evidence at the gene level

Why do gene-based tests?

- Question: “Is this gene associated when aggregating all SNP evidence in/near it?”
- Motivation: many SNPs have small effects; gene-based tests can improve interpretability.
- Important: gene-based tests still depend on LD and gene boundary definitions.

Common tools

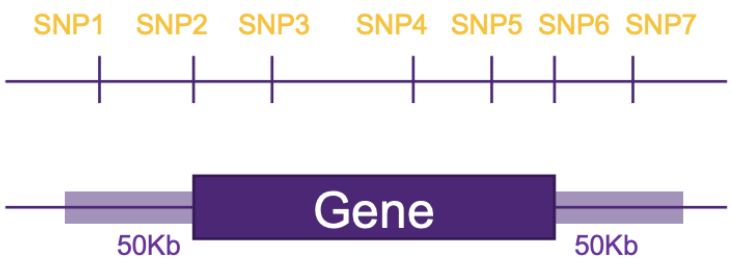
- MAGMA (popular; also does gene-set tests)
- VEGAS (simulation-based)
- PASCAL (fast gene + pathway scores)

Outputs

- Gene p-value (and often Z-score)
- Top genes per locus
- Input for gene-set enrichment

de Leeuw et al. 2015 (MAGMA); Liu et al. 2010 (VEGAS); Lamparter et al. 2016 (PASCAL).

Gene-based test



Reduce test burden

- **Combine the effects of many variants** in and around a gene
- Ask: *Does this gene, as a whole, show evidence of association with the trait?*
- This can **increase power** to discover relevant genes

Output can be used for gene-set enrichment or pathway analyses

Gene-set enrichment

From top genes → pathways / tissues / cell types

Gene-set enrichment: the simplest explanation

Question: “Are genes in a biological pathway *more associated* than other genes?”

- Inputs: gene-level statistics + a set of gene lists (GO, Reactome, KEGG…).
- Output: enriched pathways / tissues (hypotheses), with corrected p-values.
- Interpretation: “This biology is over-represented among associated genes.”

Two common test types

- Competitive: compares genes inside the set vs outside (MAGMA).
- Self-contained: tests only genes inside the set.

Why students get confused

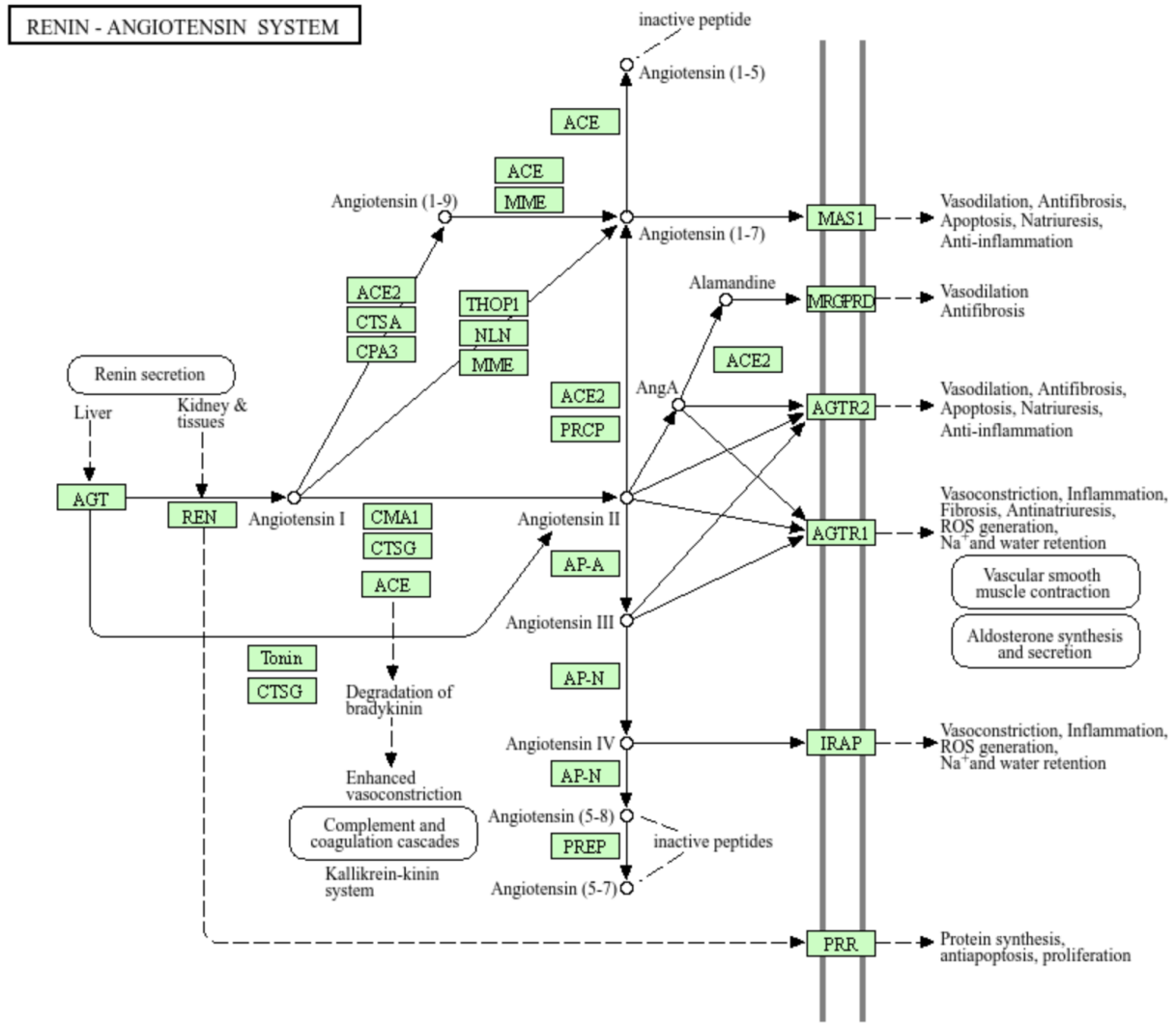
- Different tools use different null hypotheses.
- Gene set size and LD can drive results.
- Many gene sets overlap → interpret themes, not one pathway name.

What is a biological pathway?

A set of genes or proteins working together

Represents a biological process (e.g. cell cycle, metabolism

Curated in databases like KEGG, Reactome, or GO



Why use pathway analysis?

Reduces complexity of large gene-sets

Improves biological interpretation

Identifies affected biological processes

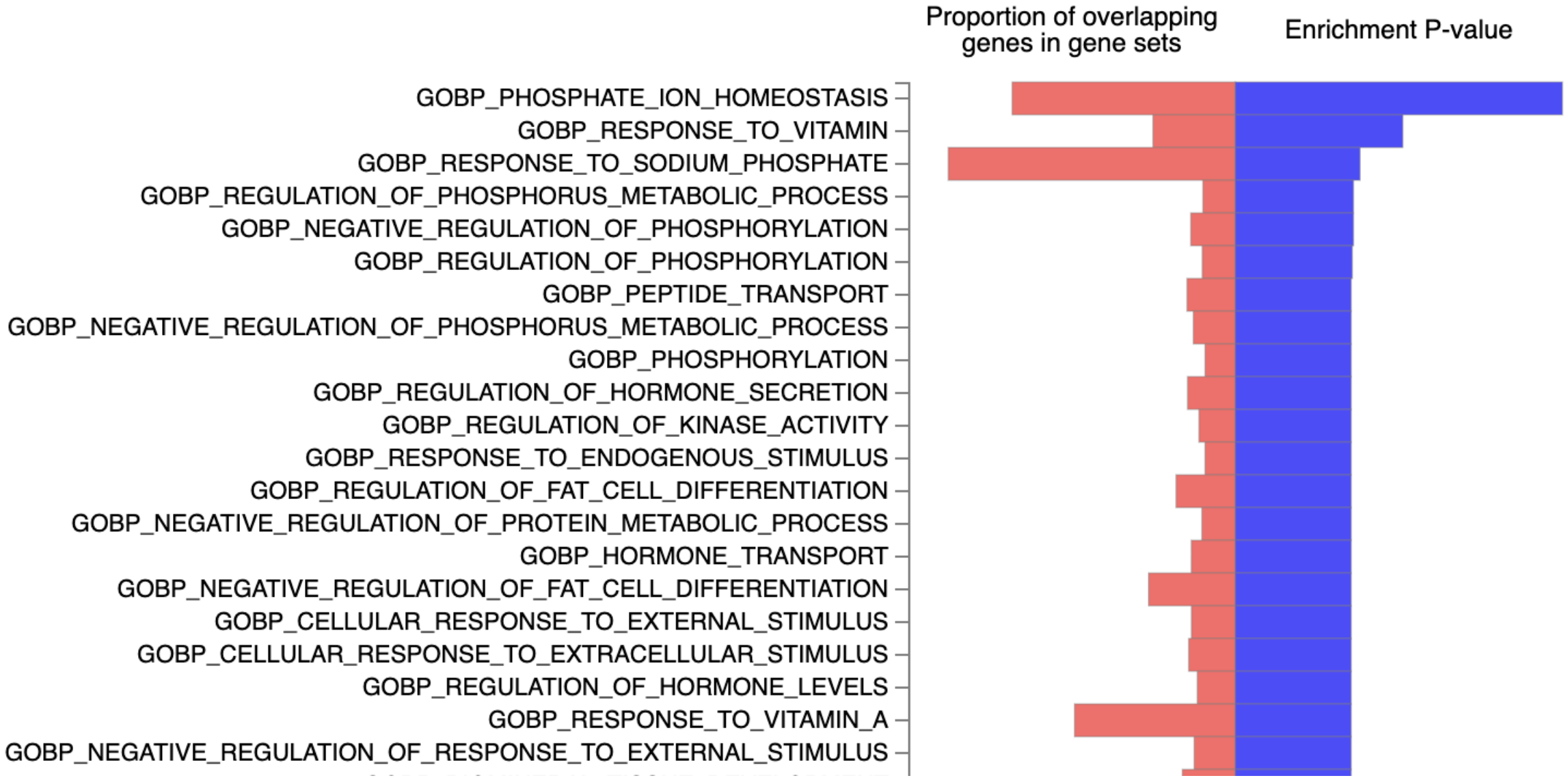
Typical workflow

Generate a list of genes (from gene set analysis)

Compare the list to known pathways

Identify pathways that are over-represented or enriched

Phosphate GWAS in UKBB



Deliverables checklist for a post-GWAS report

- 1) Locus table: lead SNPs + definitions (build, window, LD threshold).
- 2) Independent signal table: conditional analysis results (COJO).
- 3) Fine-mapping: credible sets + PIPs.
- 4) Variant annotation table: VEP consequences + key regulatory annotations.
- 5) Gene prioritization summary: evidence types used + top candidates.
- 6) Gene-based + gene-set results