Integrative Analysis

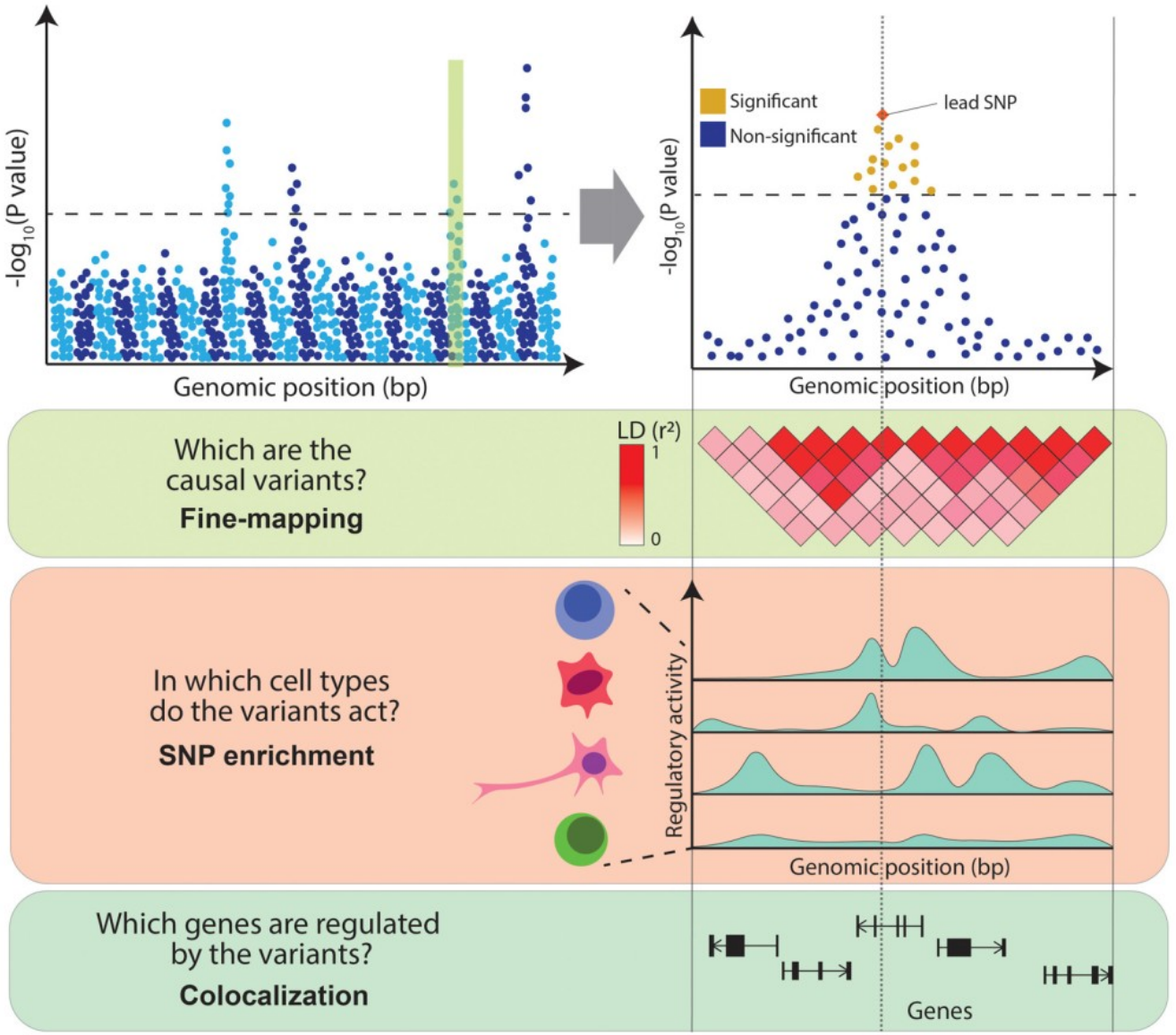
Beyond GWAS

- You have done a GWAS for your “favourite” trait

- Most of the identified genetic variants reside outside protein-coding regions, making it challenging to understand the biological mechanism underlying these identified associations
- Further complicated by the genetic effects of SNPs on complex traits likely act through a tissue-specific fashion

- What can we do to understand our trait?

Beyond GWAS


-log10(P value)
Genomic position (bp)
Significant
Non-significant
lead SNP
-log10(P value)
Genomic position (bp)
Which are the causal variants?
Fine-mapping
LD (r²)
1
0
In which cell types do the variants act?
SNP enrichment
Regulatory activity
Genomic position (bp)
Which genes are regulated by the variants?
Colocalization
Genes

Genomic Data Resources

- Many genomic resources have been developed in parallel to GWAS

- ENCODE
- Roadmap Epigenomics Project
- GTEx
- CommonMind
- ...

ENCODE

- **Encyclopedia of DNA Elements** Consortium
- Aims to build a comprehensive parts list of functional elements in the human genome
- Including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.
- The discovery and annotation of gene elements is accomplished primarily by sequencing a diverse range of RNA sources, comparative genomics, integrative bioinformatic methods, and human curation.
- Regulatory elements are typically investigated through DNA hypersensitivity assays, assays of DNA methylation, and immunoprecipitation (IP) of proteins that interact with DNA and RNA, i.e., modified histones, transcription factors, chromatin regulators, and RNA-binding proteins, followed by sequencing.

Roadmap Epigenomics Project

- Epigenetics the study of changes in the regulation of gene activity and expression that are not dependent on gene sequence
- Goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research
- Leverages experimental pipelines built around next-generation sequencing technologies to map DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts in stem cells and primary ex vivo tissues selected to represent the normal counterparts of tissues and organ systems frequently involved in human disease

GTEx

- **Genotype-Tissue Expression** Program
- Data resource and tissue bank to study the relationship between genetic variants and gene expression in multiple human tissues and across individuals
- DNA data from 948 postmortem donors and 17,382 RNA-seq across 54 tissue sites and two cell lines
- Established a comprehensive catalog of genetics variants that effect gene expression across multiple tissue
-

CommonMind

- Goal to generate and analyze large-scale genomic data from human subjects with neuropsychiatric disease
- Generating RNA and DNA sequencing, genotyping, epigenetics data across multiple brain regions from individuals with schizophrenia, bipolar disorder, and unaffected controls on a collection of more than 1,000 individuals

Many ‘omics QTL studies

- Gene expression (eQTLs
- Protein expression (pQTLs)
- Exon splicing (sQTLs)
- DNA methylation (mQTLs)
- Chromatin accessibility (caQTLs)
- ….

- Usually a summary statistics from these studies are published alongside research paper for use by the wider community

Using Genomic Data to Gain Biological Insight

- Many possibly analyses that may be done to integrate genomic data into GWAS results

- Colocalisation – identify functional gene underlying GWAS peak
- Identify target tissues
- ...

Colocalisation

- Colocalisation analysis is used to test whether two independent association signals at a locus are consistent with having a shared casual variant.

- A large number of available methods
- Have already been introduced to SMR

Colocalisation

Method	Publication	Approach	Input data
Regulatory trait concordance (RTC)	Nica et al., 2010	Conditional regression	Individual genotypes
Proportionality test	Wallace et al., 2012	Test for concordance of effects	Individual genotypes
Sherlock	He et al., 2013	Genome-wide comparison of association “signatures”	Summary statistics
COLOC	Giambartolomei et al., 2014	Bayesian test	Summary statistics
gwas-pw	Pickrell et al., 2016	Bayesian test	Summary statistics
eCAVIAR	Hormozdiari et al., 2016	Bayesian fine-mapping and colocalization	Summary statistics
enloc	Wen et al., 2017	Bayesian test for enrichment, fine-mapping and colocalization	Summary statistics
MOLOC	Giambartolomei et al., 2018	Bayesian test for multiple traits	Summary statistics

Coloc

- Given two genetic association studies both showing some association signal at a locus, how likely is it that the same variant is responsible for both associations?
- A shared causal variant is likely to imply an etiological link between the traits being considered
- But, correlation does not imply causality...

- Coloc is a Bayesian method which, for two traits, integrates evidence over all variants at locus to evaluate the following hypotheses
    - H0: No association with either trait
    - H1: Association with trait 1, not with trait 2
    - H2: Association with trait 2, not with trait 1
    - H3: Association with trait 1 and trait 2, two independent SNPs
    - H4: Association with trait 1 and trait 2, one shared SNP

Coloc

- Become a reference method for colocalization testing
- A limitation is that it only tests for two traits at a time
- A variant could increase DNA methylation, in turn reducing the expression of a nearby gene, impairing cell function and increasing disease
- Combining across multiple traits would give more biological insight
- MOLOC expanded the original formulation of COLOC to include multiple traits

eCAVIER

- A trait association signal can result from multiple causal variants
- Not accounting for this could lead to misclassifications of colocalisation

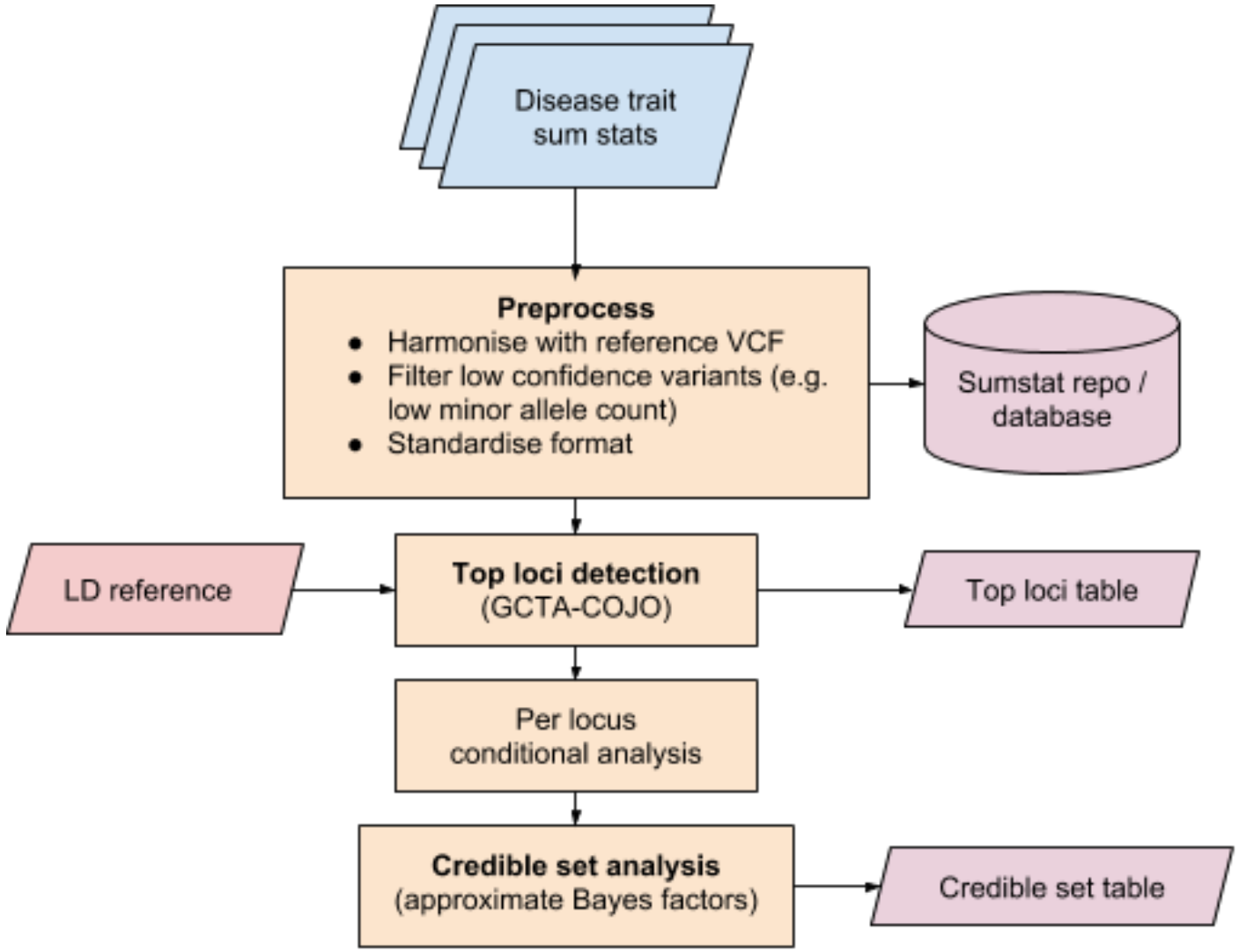
- eCAVIAR is a modified fine-mapping software that can account for multiple causal SNPs
- Estimates a probability every SNP in the region is causal and uses this information in colocalisation analysis

Open Targets Genetics

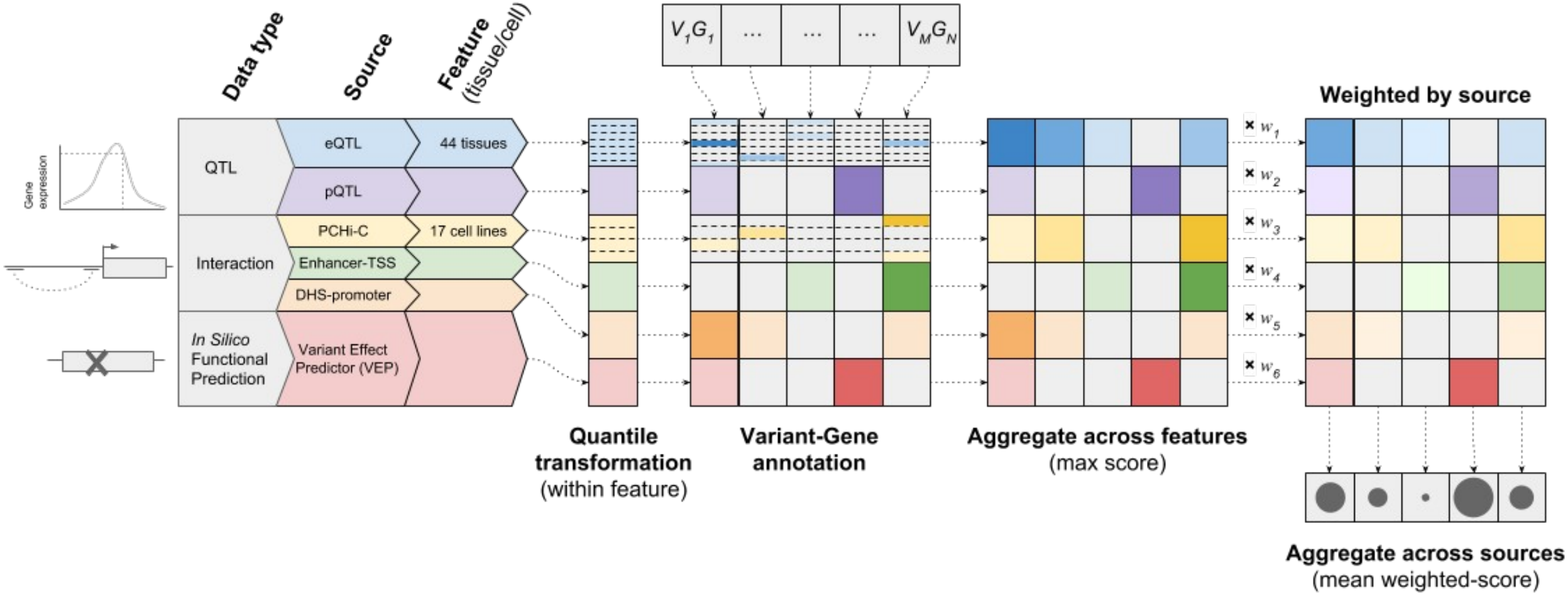
- A to aggregate evidence linking (i) variants to disease, and (ii) variants to genes, so that for a specific disease potential drug targets can be prioritised based on robust genetic information.

- Given a set of potentially causal tag variants, assign variants to genes using a pipeline:
    - Molecular phenotype quantitative trait loci experiments (e.g. eQTLs and pQTLs)
    - Chromatin interaction experiments (e.g. Promoter Capture Hi-C)
    - *In silico* functional predictions (e.g. Variant Effect Predictor from Ensembl)
    - Distance from the canonical transcript start site (TSS)

Open Targets Genetics



Open Target Genetics

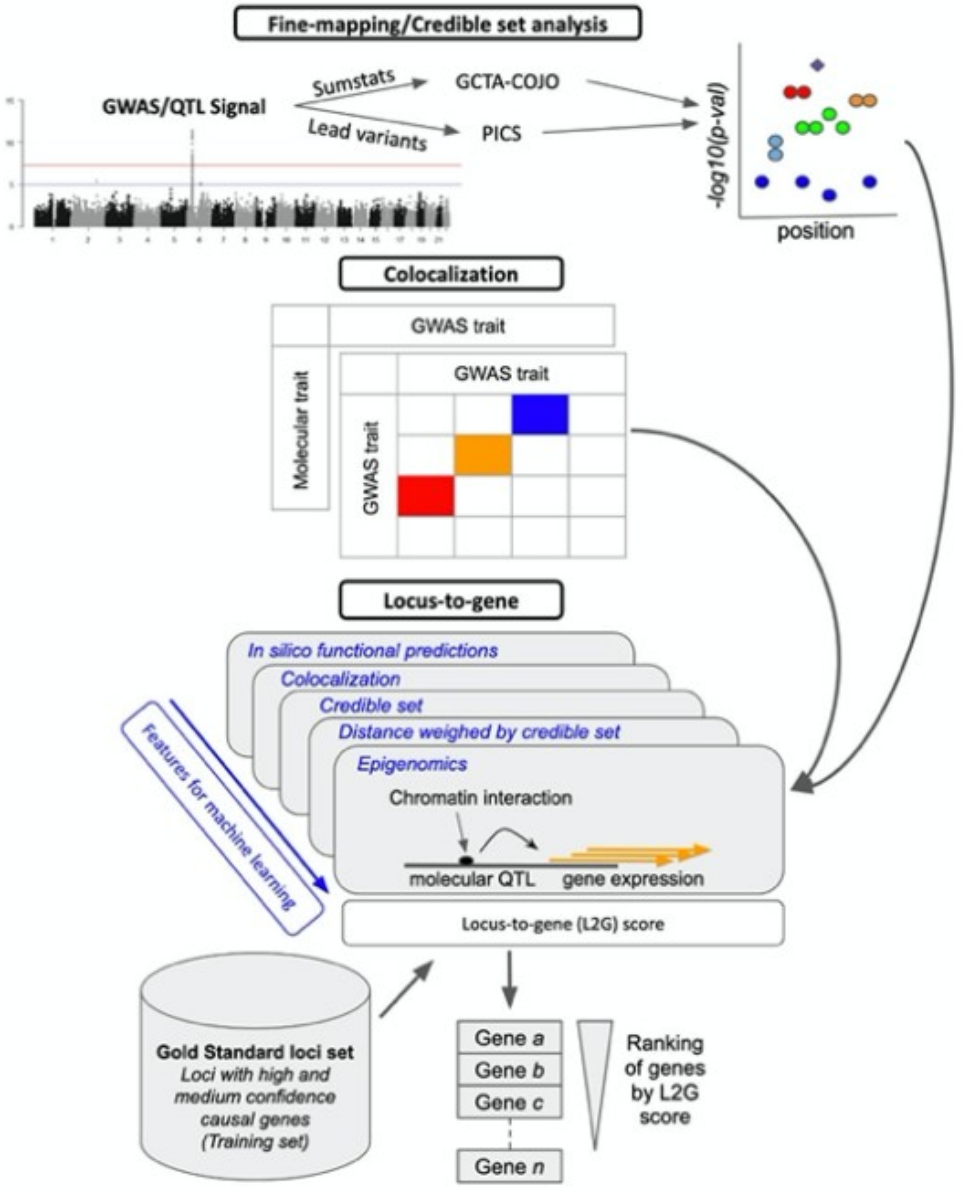


Open Target Genetics

Data type	Experiment type	Source	Weighting
In silico functional prediction	Transcript consequence	VEP	1.0
QTL	eQTL	*many*	0.66
QTL	pQTL	Sun *et al.* (Nature, 2018)	0.66
Interaction	PCHi-C	Javierre *et al.* (Cell, 2016)	0.33
Interaction	Enhancer-TSS correlation	Andersson *et al.* (Nature, 2014)	0.33
Interaction	DHS-promoter correlation	Thurman *et al.* (Nature, 2012)	0.33
Distance	Canonical TSS		0.33

Open Target Genetics

Gold-standard source	Number of GSP loci in training data
ChEMBL III	37
ChEMBL IV	88
Eric Fauman Twitter	82
ProGeM	156
T2D Knowledge Portal	49
Open Targets Curated	33
Total	445



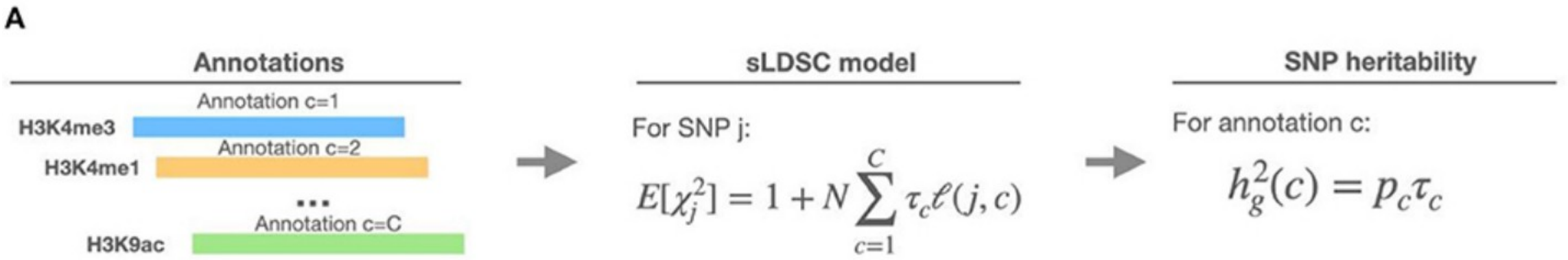
Which Cell Type is Relevant for Your Trait?

- The majority of disease-associated loci lie in non-coding regions of the genome
- It is unclear which genes they regulate and in which cell types or physiological contexts this regulation occurs.

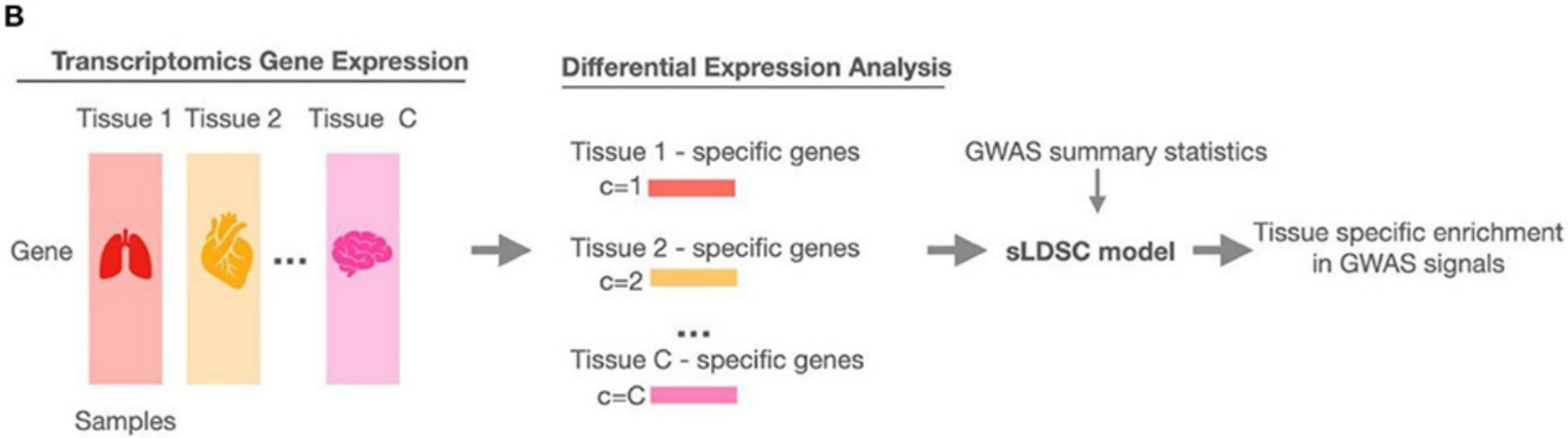
- We can use genomic data to identify relevant cell types
- Allows us to study the effect of genes in relevant cellular models
- Valuable for screening potential therapeutic compounds at high-throughput

- There are many possible approaches….
- Review: Zhu et al., Front. Genet., 22 January 2021 | https://doi.org/10.3389/fgene.2020.587887

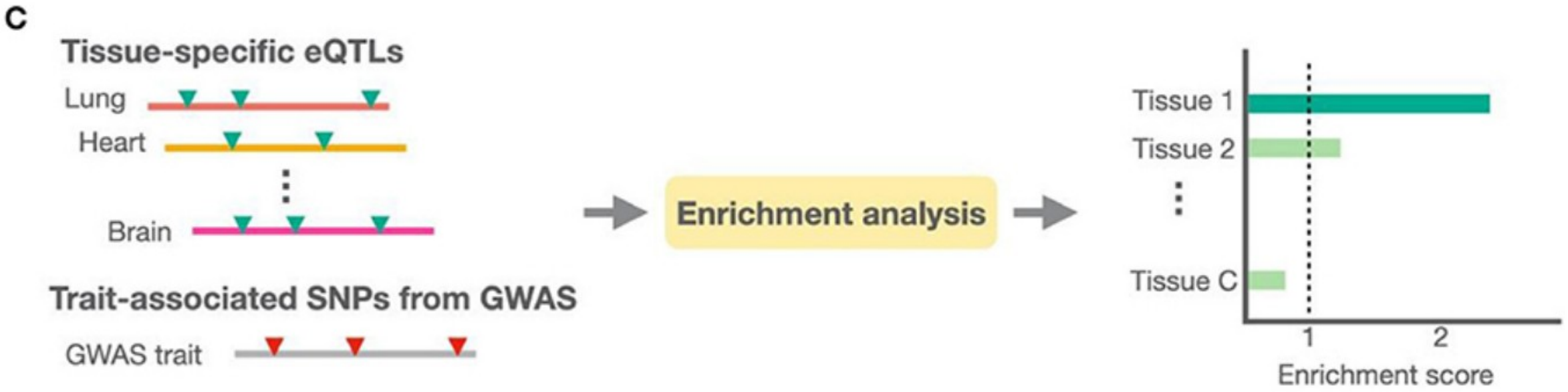
Using Epigenetic Information



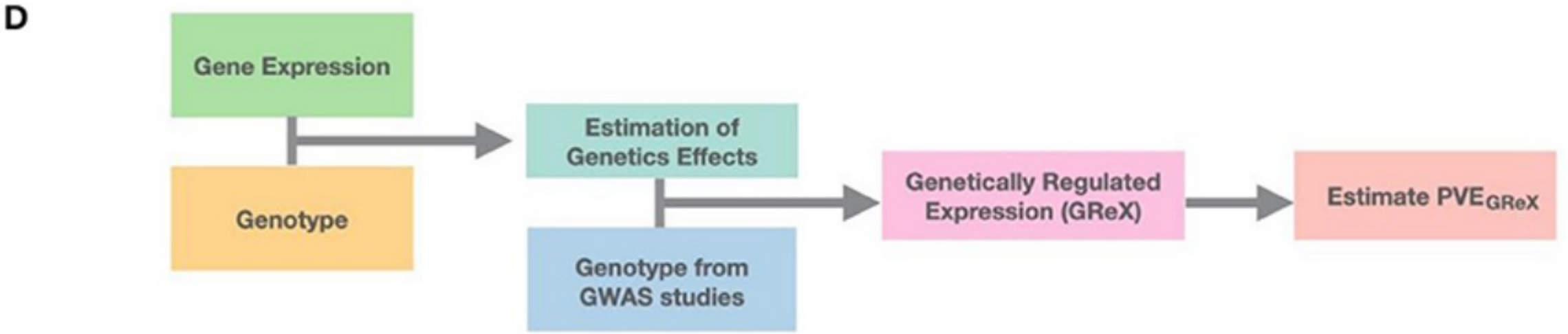
Using Transcriptomic Data



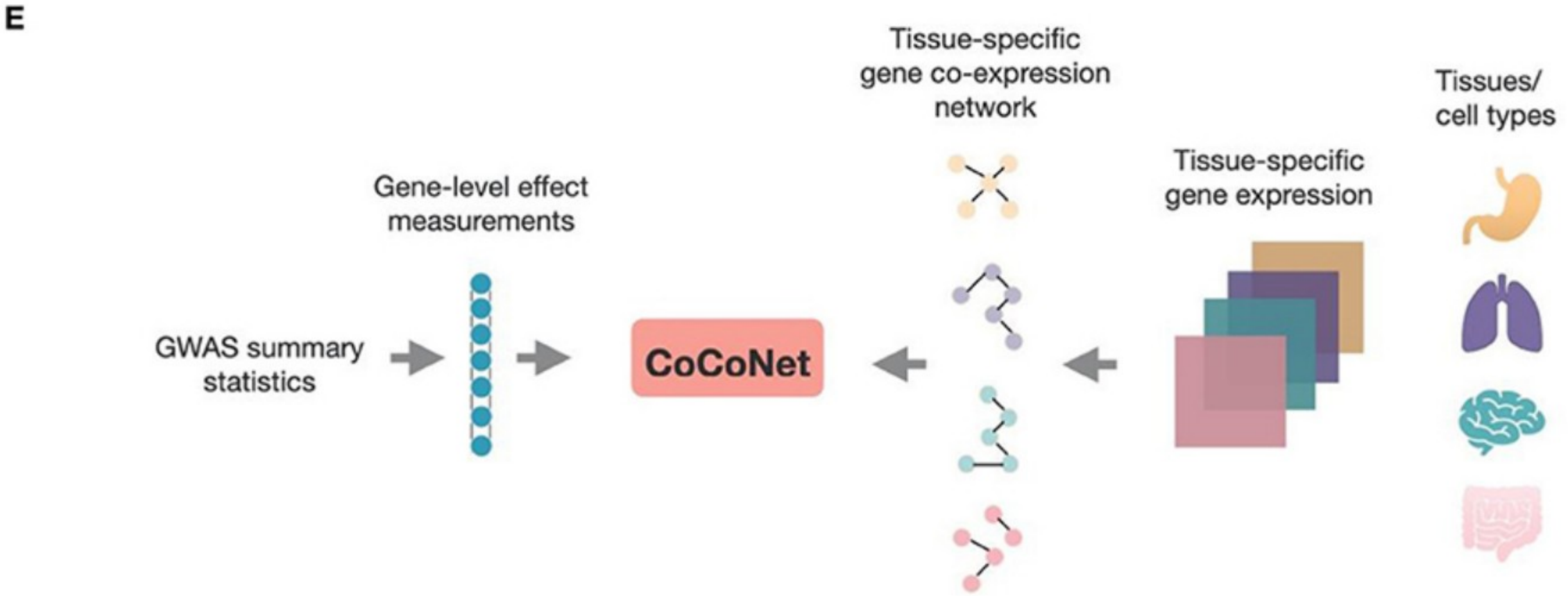
eQTL Enrichment



Genetically Regulated Expression (GReX)



Tissue-specific Gene Co-expression Networks



Which Method?