Post-GWAS Analyses I

```
$ echo "Data Sciences Institute"
```

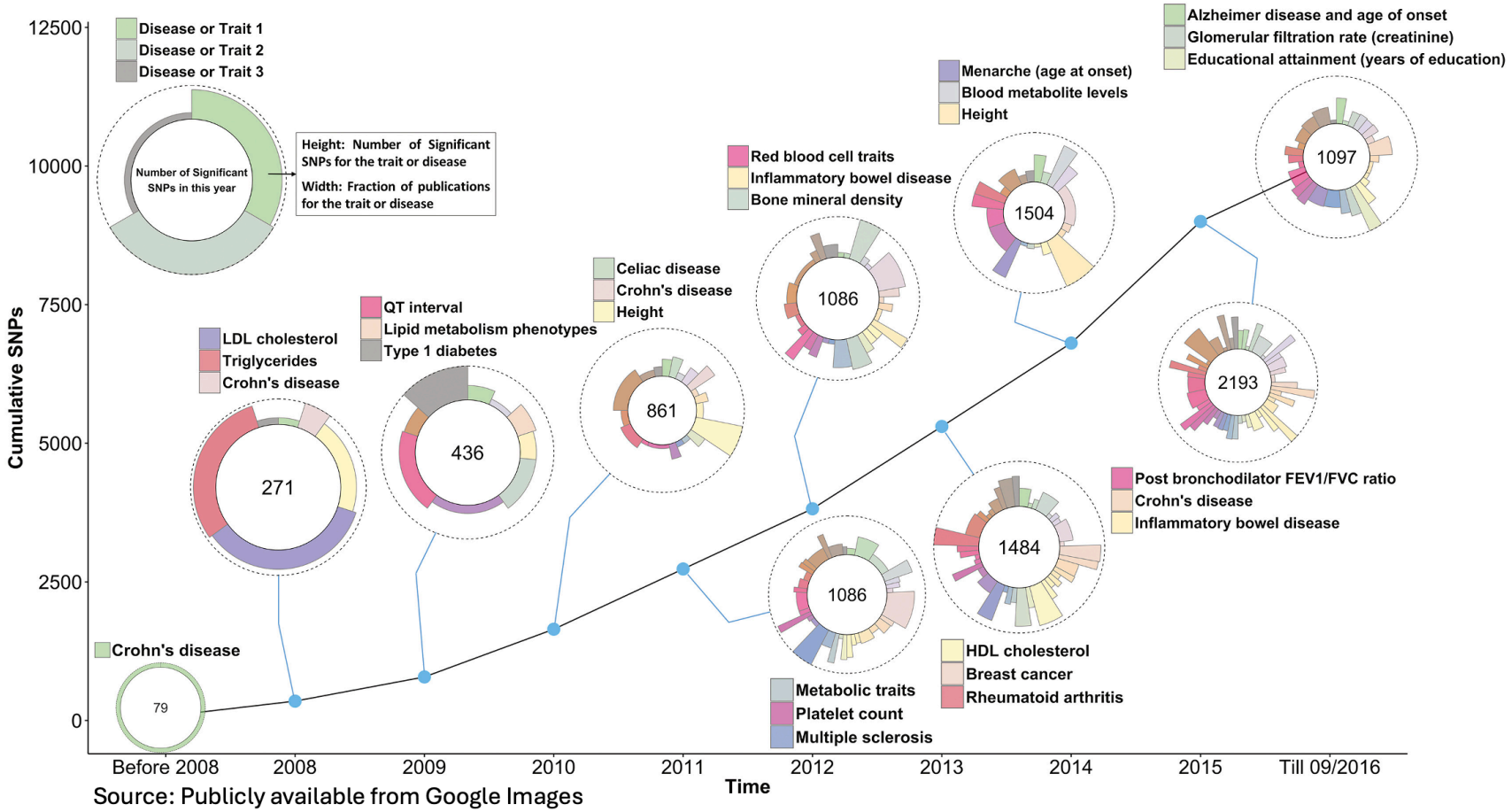
What You Will Learn Today

- **What GWAS teaches us**: polygenicity, effect sizes, and cross-ancestry considerations.
- **Post-GWAS fine-mapping basics**: Regular approaches and Bayesian approaches.
- **Functional annotation primer**: ENCODE, Roadmap, GTEx and molecular QTLs.

What have we learned so far from GWAS?

- Large-scale GWAS have mapped hundreds of thousands to millions of SNP-phenotype associations.
- Example (2021): The GWAS Catalog listed 4,865 publications and 247,051 associations.
    - These reported SNPs help illuminate molecular mechanisms of common diseases and the biological pathways underlying traits of interest.
- Associations for some SNPs linked to rare diseases have been tested intensively.
- Yet classic GWAS alone have not yielded solid, mechanistic insight into how variants drive phenotypes.

GWAS SNP-Trait Discovery Timeline



Source: Publicly available from Google Images

Practical Lessons from GWAS

- Complex traits are highly polygenic, with thousands of variants contributing small increments of risk or trait change.
- For common variants, single-variant effect sizes are typically modest; odds ratios frequently fall in the 1.05-1.20 range.
- For height, a well-powered model trait, the effect sizes are on the order of ~ 1 millimeter.
- Because effects are small and testing is genome-wide, large cohorts are needed to reach stringent significance (e.g., $\mathrm{p} < 5 \times 10^{-8}$).

Practical Lessons from GWAS

- A GWAS peak typically marks a cluster of nearby variants (about 10–100 kb) that move together because of linkage disequilibrium.
- Multiple independent association signals can reside within the same locus.
- The majority of GWAS associations lie outside protein-coding exons; these variants are believed to act mainly by regulating gene expression (e.g., enhancer or promoter activity) rather than by changing amino-acid sequence.
- Allele frequencies, LD structure, and effect sizes at disease loci can vary across ancestries.
- Pleiotropy (the same variant or locus is associated with multiple traits) is ubiquitous.

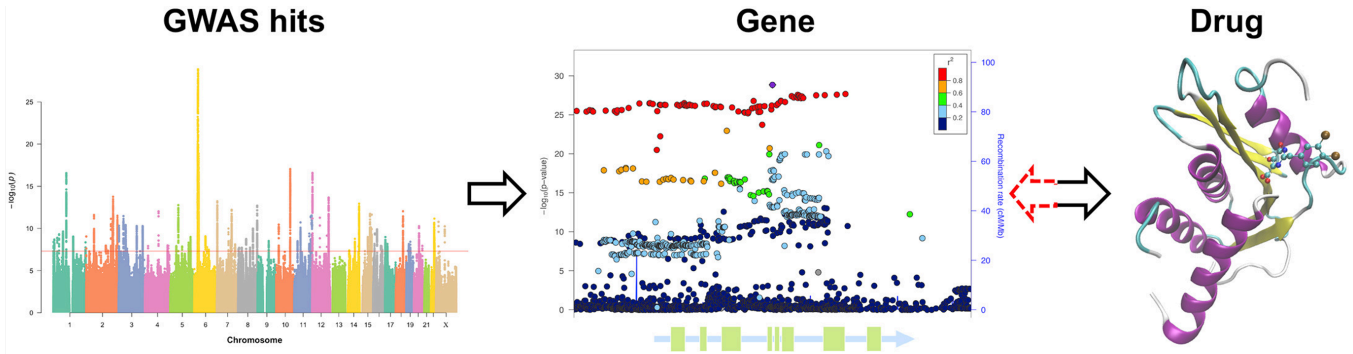
Practical Lessons from GWAS

- Many individual GWAS are underpowered to detect the smallest effects; nonetheless, the large number of contributing variants means genuine signals still emerge.
- Even when a variant's effect on a biomarker is small, the clinical impact can be substantial: the implicated gene may encode a tractable drug target, and modulating it can yield large therapeutic benefits.
- For example, common variants near HMGCR only have a small influence on LDL-cholesterol, but drugs targeting the encoded protein reduce LDL by ~30%.

From GWAS discovery to Medicines

- The overarching aim of human genetics is to enable translational medicine-turning genetic insights into better diagnostics, prevention, and therapies.
- Genetically supported targets are more likely to progress successfully through clinical development, including to phase III trials and eventual approval.
- E.g. People with loss-of-function mutations in SLC30A8 (the ZnT-8 transporter) have a lower risk of type 2 diabetes, leading to companies to develop ZnT-8 antagonists for diabetes therapies.

Success Stories: From GWAS to Clinical Impact



Trait	Gene with GWAS hits	Known or candidate drug
Type 2 Diabetes	*SLC30A8/KCNJ11*	ZnT-8 antagonists/Glyburide
Rheumatoid Arthritis	*PADI4/IL6R*	BB-Cl-amidine/Tocilizumab
Ankylosing Spondylitis(AS)	*TNFR1/PTGER4/TYK2*	TNF-inhibitors/NSAIDs/fostamatinib
Psoriasis(Ps)	*IL23A*	Risankizumab
Osteoporosis	*RANKL/ESR1*	Denosumab/Raloxifene and HRT
Schizophrenia	*DRD2*	Anti-psychotics
LDL cholesterol	*HMGCR*	Pravastatin
AS, Ps, Psoriatic Arthritis	*IL12B*	Ustekinumab

Source: Publicly available from Google Images

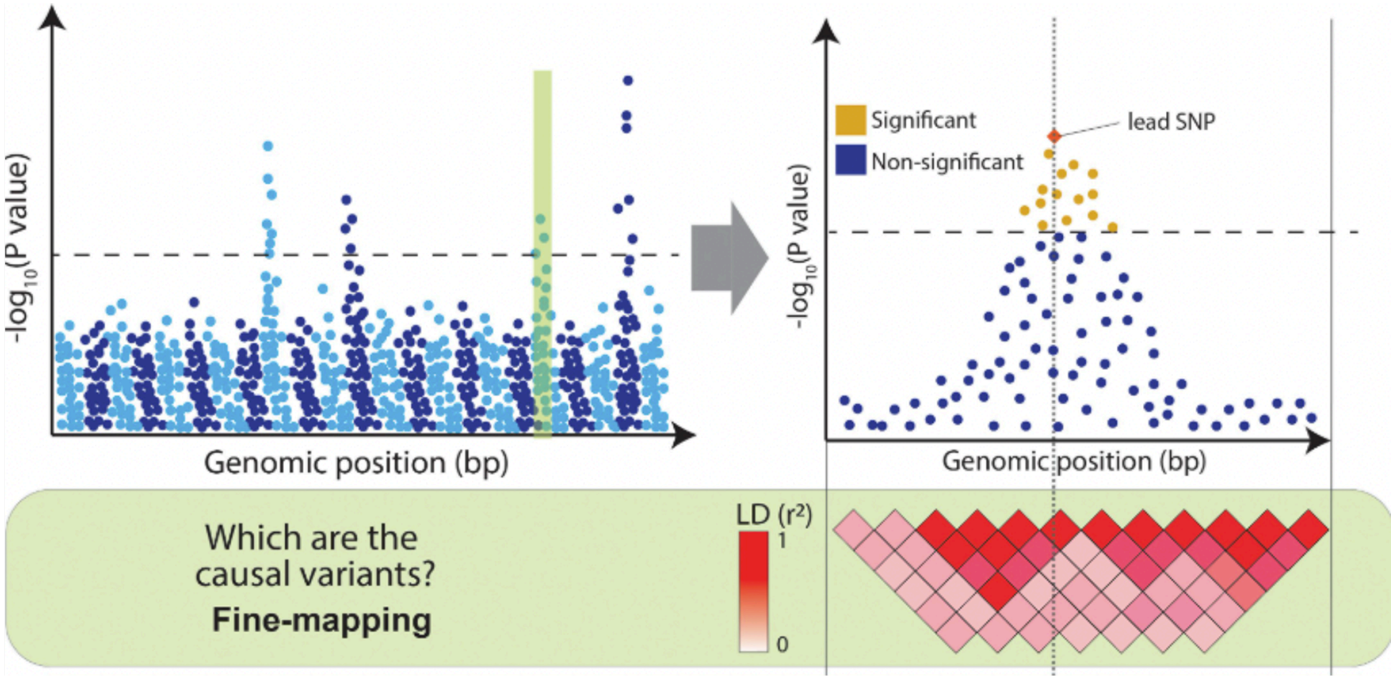
Post-GWAS Analyses

Motivation

- pGWAS (post-GWAS) is a crucial step to move beyond SNP-level associations toward biological mechanism.
- Beyond the basic task of identifying genetic associations, several post-GWAS analyses can be performed:
    i. Fine-mapping: statistical approach to identifying causal variants.
    ii. CRISPR experiments: experimental techniques
    iii. Polygenic Risk Score (PRS): predicting trait values based on genotype profiles.
    iv. Others: colocalization, TWAS, network inference, G×G and G×E analyses, etc.

Confounding in GWAS -LD

- One major issue is confounding caused by local correlation among sites (linkage disequilibrium), which makes it difficult to distinguish true signal from variants that are merely correlated.



Source: Publicly available from Google Images

What is fine-mapping?

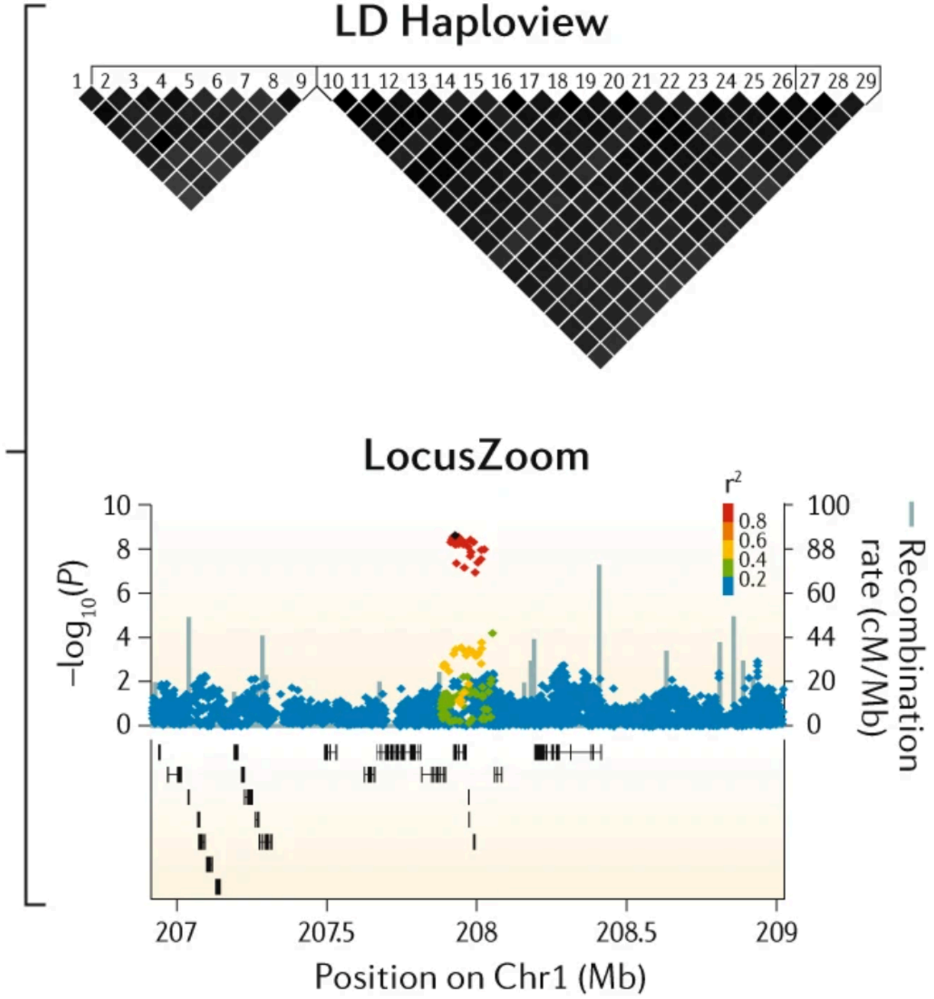
- GWAS often identifies a broad locus with many associated SNPs.
- Due to **linkage disequilibrium (LD)**, many SNPs in the locus are correlated and show similar p-values.
- Fine-mapping asks:
  - Which SNP(s) in this region are most likely to be **causal**?
  - How many **independent signals** are there?
- Goal: identify specific variants that are the best causal candidates.
- This is a key step before functional follow-up and experimental validation.

Heuristic fine-mapping: LD-based candidate selection

- Idea: use the **LD pattern around the lead SNP** to pick nearby SNPs that are likely to be causal.
- **LD thresholding:**
  - Compute pairwise LD (r^2) between the lead SNP and other SNPs.
  - Keep SNPs with LD above a threshold (e.g. $r^2 > 0.6$) as **candidate causal SNPs**.
- **LD clustering :**
  - Hierarchical clustering of all SNPs in a region based on their pairwise r^2 to create clusters.

Heuristic Fine-mapping: LD-based Candidate Selection

- Visualization tools such as **LocusZoom** or **Haploview**:
  - Combining the GWAS lead SNP with SNPs in the same LD block to select potential causal SNPs.



Source: Publicly available from Google Images

Heuristic Fine-mapping: LD-based Candidate Selection

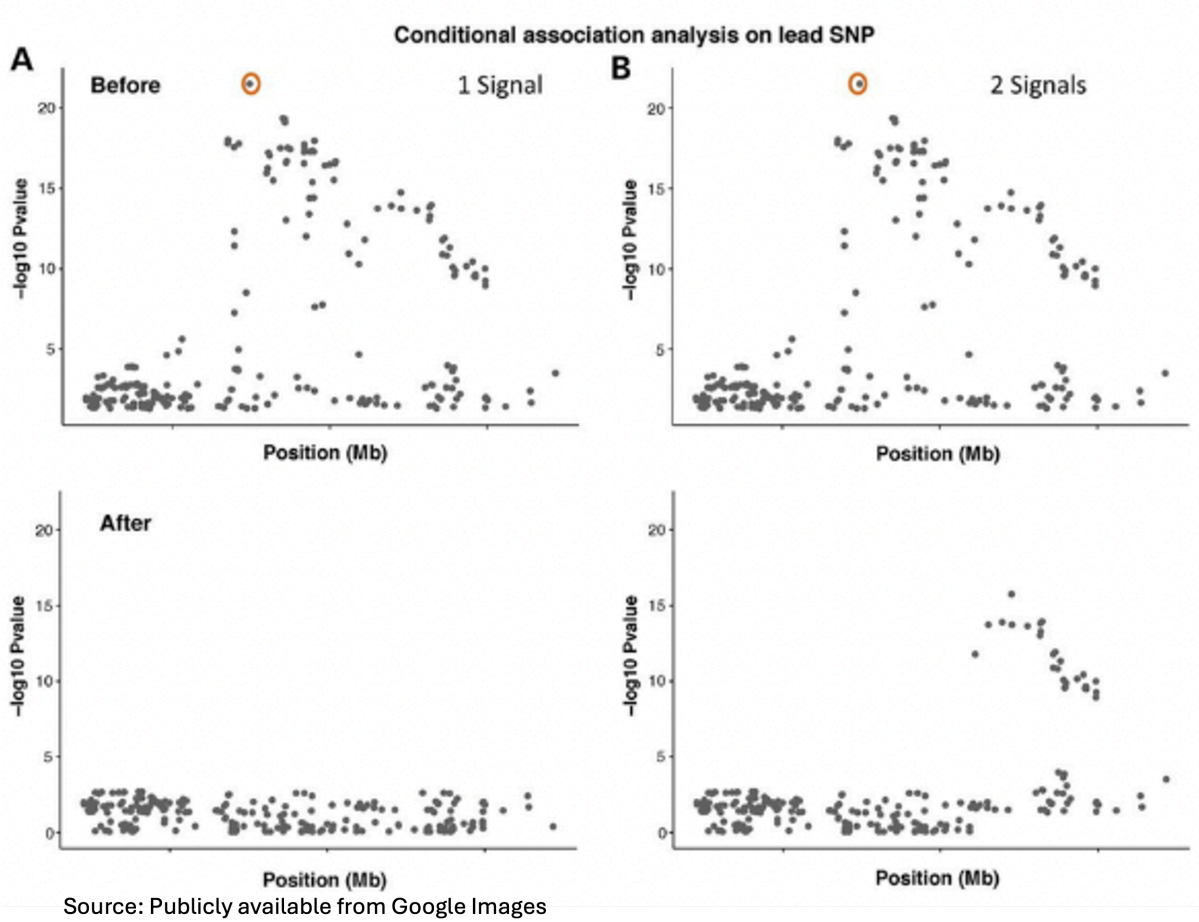
- Heuristic LD-based methods are useful for **initial candidate selection**, but not sufficient on their own to define causal variants.
- **Limitations:**
  - Relies on **arbitrary thresholds** for LD and window size.
  - Does **not** model the **joint effects** of multiple SNPs on the trait.
  - Does **not** provide an objective measure (e.g. probability) that a SNP is causal—interpretation is partly **subjective**.
  - More rigorous approaches (penalized regression, Bayesian fine-mapping) explicitly model multiple SNPs together and quantify uncertainty.

Heuristic Fine-mapping: Conditional Analysis

- Start from the **lead SNP** in a locus (smallest p-value).
- Use **conditional analysis / forward stepwise regression**:
  - Fit a regression model (linear or logistic) with the current set of SNPs in the locus (initially just the lead SNP).
  - For each remaining SNP, test its effect **conditional on** the SNPs already in the model (add one candidate SNP at a time).
  - Add the SNP with the **smallest conditional p-value** if it is below a pre-specified threshold (e.g. 5×10^{-8} or a locus-specific threshold).
  - Repeat these steps until **no remaining SNP** has a significant conditional p-value → the SNPs in the final model are treated as **independent association signals** in the locus.

Heuristic Fine-mapping: Conditional Analysis

- Implemented in tools such as:
  - **PLINK** (e.g. `--condition`, `--condition-list`, stepwise conditional analysis)
  - **GCTA-COJO** (conditional and joint multiple-SNP analysis)



Source: Publicly available from Google Images

Fine-mapping: Penalized Regression Models

- **Jointly model many SNPs** in a region using regression.
- Let Y be the phenotype, X the genotype matrix, and β the SNP effects.
- Penalized regression estimates β while shrinking small effects towards zero:
  - Examples: **lasso**, **elastic net**, other sparse penalties.
- Objective (elastic net form):

$$\min_{\beta} \ \frac{1}{2n}\|Y - X\beta\|^2 + \lambda\left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right).$$

- Result: a sparse model where only a few SNPs have non-zero effects → candidate causal SNPs.

Fine-mapping: Penalized Regression Models

- Works best with **individual-level data** and many correlated SNPs.
- **Tuning parameter** λ (and α) chosen by cross-validation to minimize prediction error.
- Advantages over forward selection:
  - More stable when SNPs are highly correlated.
  - Simultaneously estimates effect sizes and performs variable selection.
- Limitations:
  - Aims to choose a good model for Y, not to quantify causal probabilities.
  - This motivates Bayesian variable selection / Bayesian fine-mapping.

Ingredients of Bayesian inference

- We have an unknown quantity θ:
  - e.g. effect size, or an indicator "SNP j is causal".
- **Prior** $p(\theta)$:
  - Our belief about θ *before* seeing the data.
- **Likelihood** $p(\text{data} \mid \theta)$:
  - How likely the observed data are, if θ had a given value.
- **Posterior** $p(\theta \mid \text{data})$:
  - Our updated belief about θ *after* seeing the data.
- Bayes' rule:

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta)\, p(\theta)}{p(\text{data})} \quad \propto \quad p(\text{data} \mid \theta)\, p(\theta).$$

Bayesian fine-mapping: big picture

- Same goal as penalized regression: decide **which SNPs have non-zero effects**.
- Key difference: Bayesian methods assign **probabilities to many models**, not just pick one best model.
- We specify a **prior** over which SNPs are causal (e.g. all equally likely, or a fixed expected number per region) and update it with the data using Bayes' rule.
- Output:
  - **Posterior probabilities** for different models (combinations of causal SNPs),
  - **Posterior inclusion probabilities (PIPs)** for each SNP.
- This gives a clear **probabilistic interpretation** of fine-mapping results.

Bayesian Fine-mapping

- For m SNPs, define an indicator vector $c = (c_1, \ldots, c_m)$:
  - $c_j = 1$ if SNP j is causal, $c_j = 0$ otherwise.
  - There are 2^m possible c vectors → 2^m possible causal models.
- Using Bayes' formula, for a specified model M_c:

$$P(M_c \mid D) = \frac{P(D \mid M_c) \cdot P(M_c)}{P(D)}$$

- The posterior probabilities for different models can be used to determine the posterior probability of including each SNP in any of the models (PIP).

Posterior Inclusion Probability (PIP)

- PIP for SNP i: sum of posteriors over all models that include SNP i as causal.

$$PIP_i = \sum_c I(\text{ model containing SNP i as causal })P(M_c \mid D)$$

- Use PIP ranks to prioritize putative causal variants.
- Caution in high-LD regions: probability spreads across correlated SNPs.
- Posterior expected number of causal SNPs $\approx \Sigma PIP_i$ over the region.

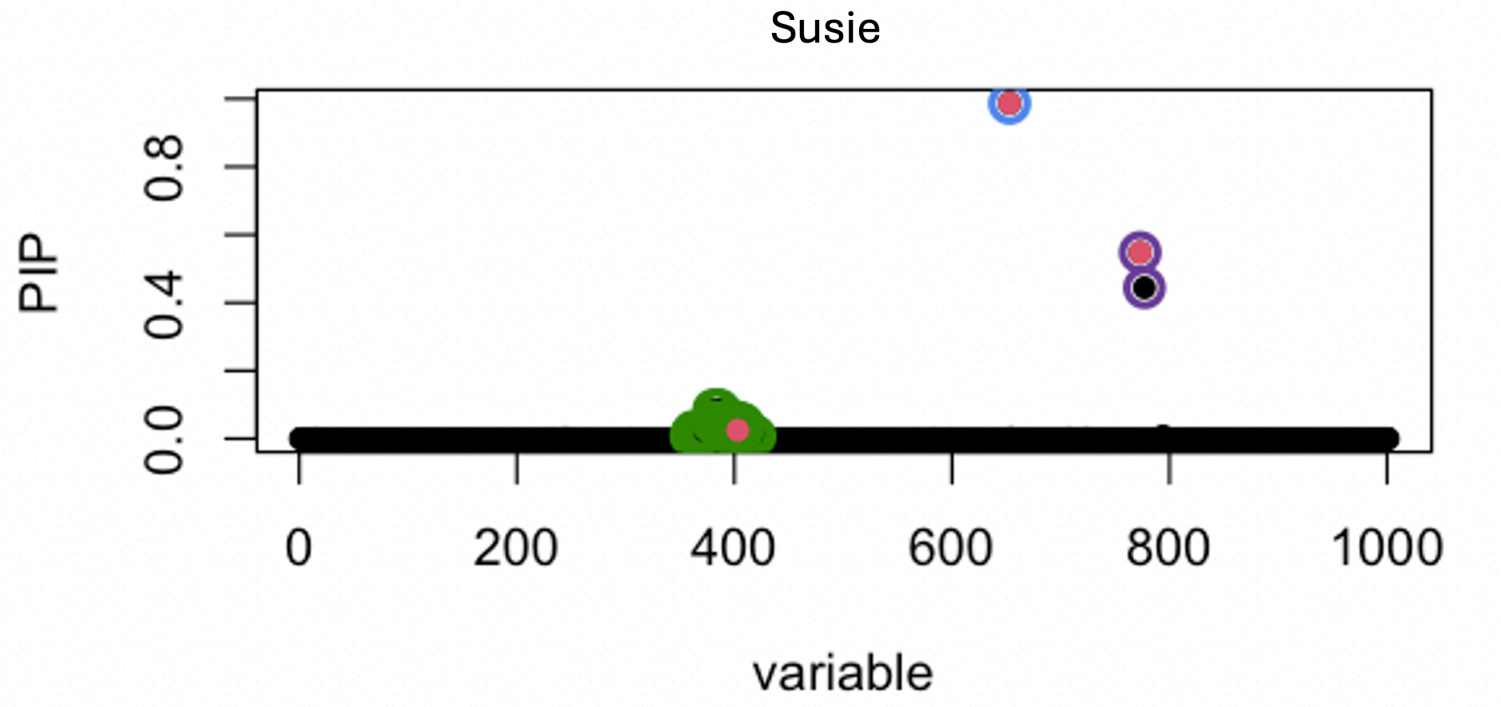
Credible Sets

- "A level ρ credible set is defined to be a subset of correlated variables (with correlation within the set greater than some threshold r) that has probability ρ or greater of containing at least one effect variable (i.e. causal SNP)."
- In short, it defines a set of variants likely to contain the causal SNP(s).
- Procedure:
    i. Rank SNPs by PIP (largest $\rightarrow$ smallest).
    ii. Accumulate PIPs until reaching coverage a (e.g., 95% or 99%).
    iii. Selected variants form the a credible set.

Question: Are credible sets unique for a given level ρ?

Discussion

- Answer: Not necessarily.
- Ties or near-ties in PIPs can yield multiple valid sets; software typically reports one (often the smallest) based on its ranking rules.

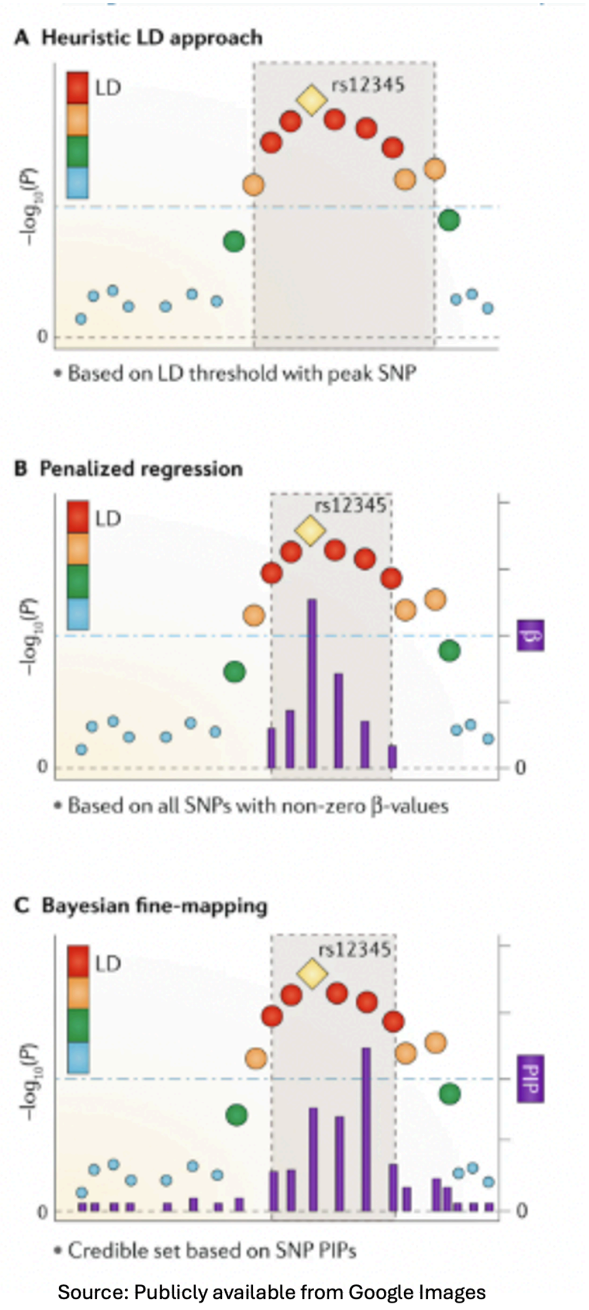


Practical Workflow & Tools

- Typical inputs: GWAS summary statistics, ancestry-matched LD reference panels, and (optionally) functional annotations.
- Steps:
  - Define regions (around lead SNPs; PLINK --clump for sentinel signals).
  - Run Bayesian fine-mapping to obtain PIPs & credible sets.
  - Outputs to show: top-PIP variants, credible set sizes, locus zoom-style plots.
- Tools: CAVIAR, FINEMAP, SuSiE, CARMA

Hypothetical Examples

- The purple bars represent additional variant-level statistics produced by fine-mapping.
  - β-values for penalized regression
  - PIPs for Bayesian methods
- The light grey boxes represent the regions selected by fine-mapping.



Source: Publicly available from Google Images

More Complex Issues

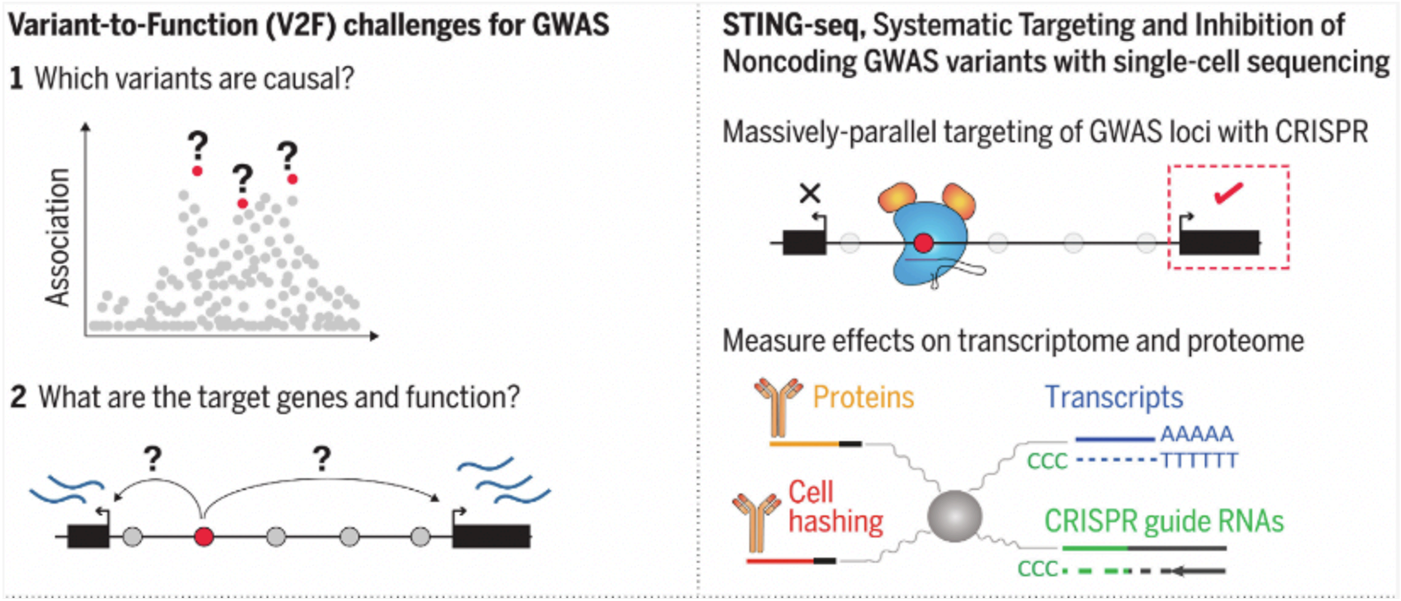
- Many GWAS results come from meta-analyses without individual-level data.
- Sample size varies by SNP in the meta-analysis, leading to inconsistencies.
- LD information is often taken from external reference panels.
- Mismatches between external LD and GWAS summary stats can invalidate fine-mapping.
- Leverage high-dimensional functional annotations to improve inference.

Fine-mapping methods: summary

- **Heuristic LD-based methods**
  - Use LD thresholds and visual inspection to pick SNPs near lead SNPs.
  - Fast and intuitive, but arbitrary and non-probabilistic.
- **Penalized regression**
  - Jointly models many SNPs, encourages sparse solutions.
  - Better than simple forward selection in high-LD regions.
- **Bayesian fine-mapping**
  - Models uncertainty over many possible causal configurations.
  - Produces PIPs and credible sets → probabilistic interpretation.

Functional Follow-up

- Fine-mapping prioritizes putative causal variants at GWAS loci that can be subjected to functional studies.
- Massively parallel CRISPR perturbation of GWAS loci:



Source: Publicly available from Google Images

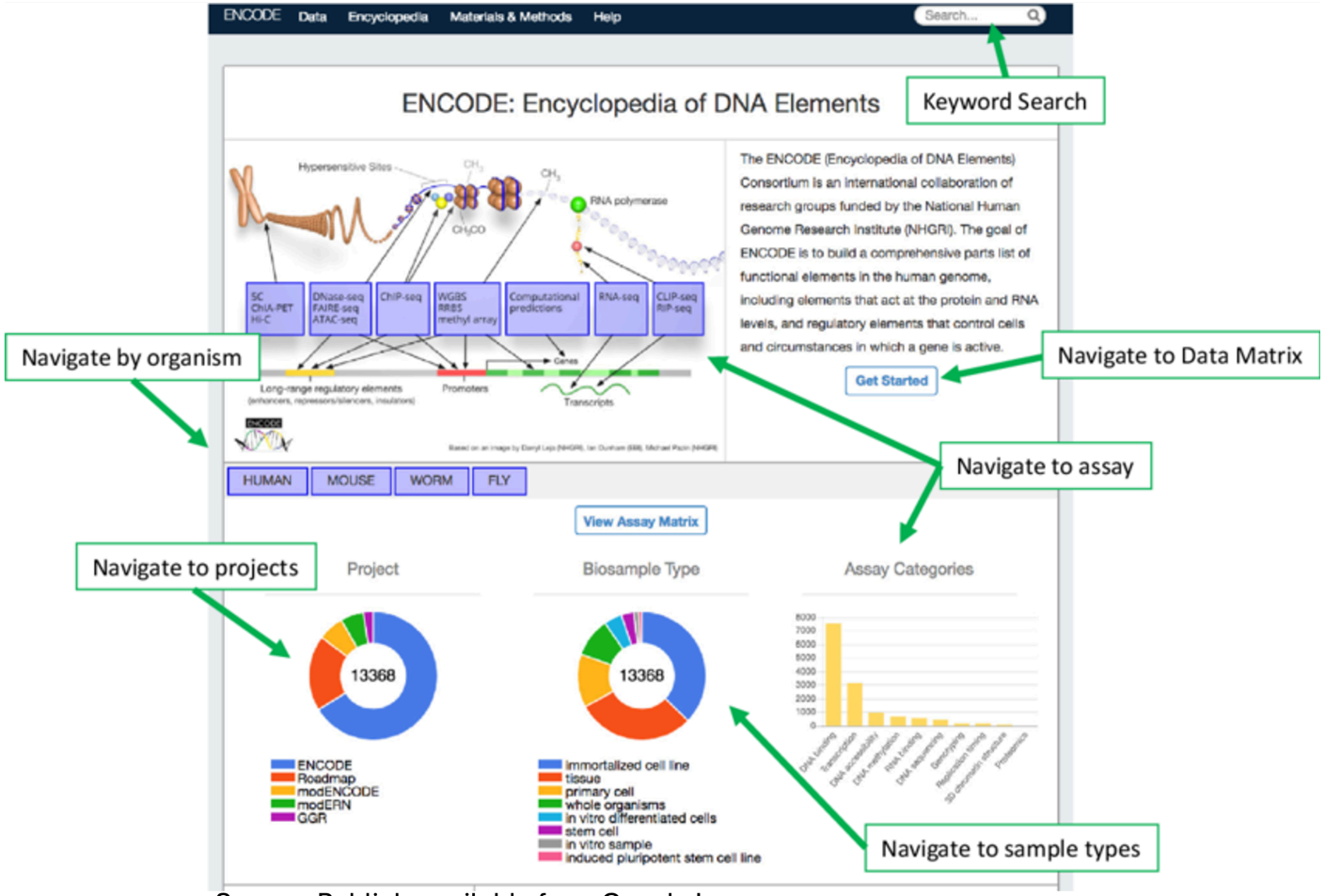
Functional Annotation

Functional Annotation of variants

- GWAS pinpoints where associations occur, but not how the implicated variants exert their effects.
- Functional annotation adds biological context to a variant to infer its potential impact on genes, regulatory programs, and molecular traits.
- Use diverse resources to answer specific questions:
    - Does the variant alter amino-acid sequence or function? (e.g. PolyPhen-2)
    - Does the variant fall within an enhancer, promoter, or other regulatory element? (ENCODE, Roadmap Epigenomics)
    - Is it in a conserved region? (evolutionary constraint)
    - Does it affect a molecular trait like gene expression or protein level? (eQTL / pQTL)

ENCODE

The Encyclopedia of DNA Elements (ENCODE) is a public research project that aims to build a comprehensive parts list of functional elements in the human genome.

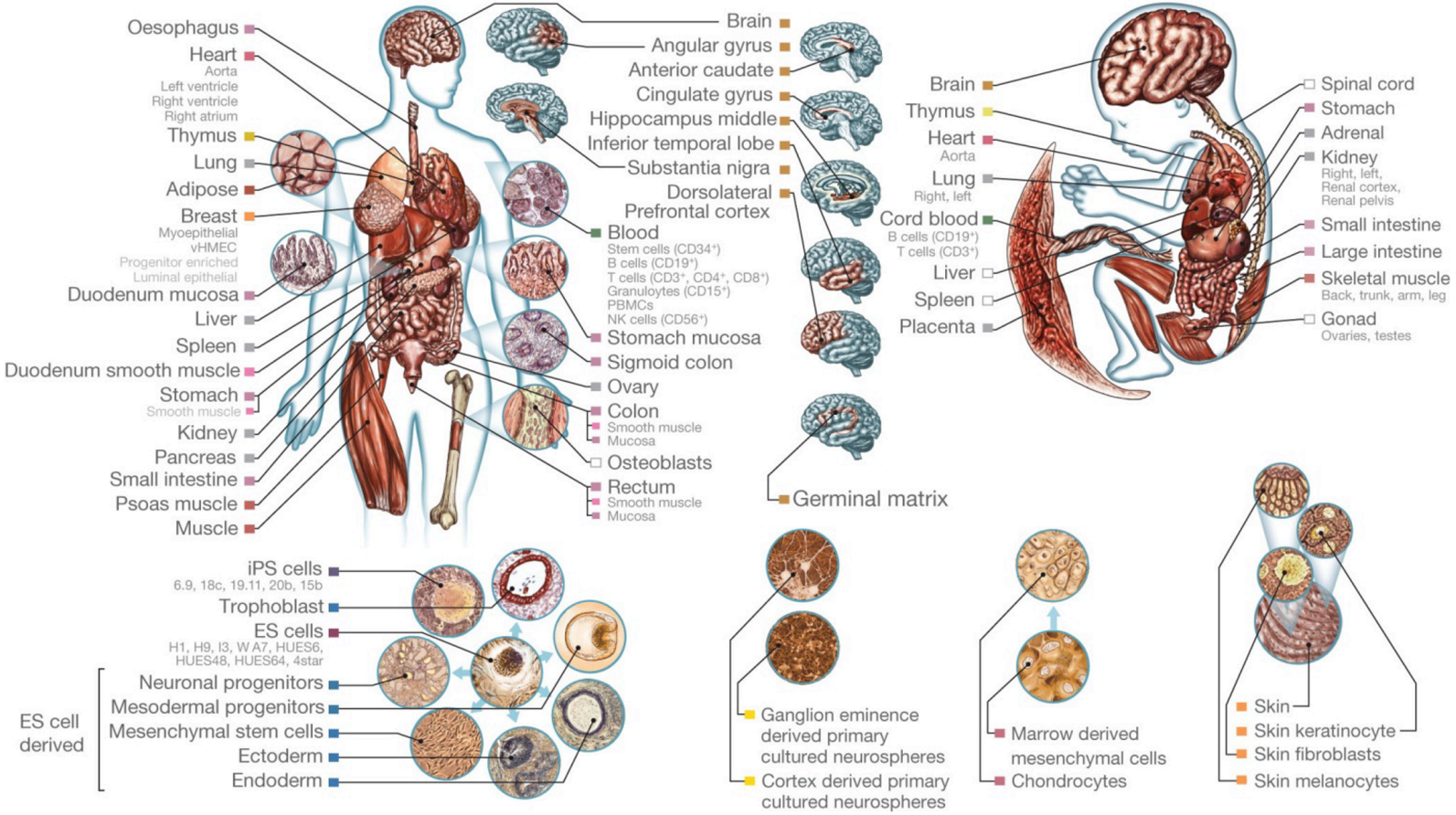


Source: Publicly available from Google Images

Roadmap Epigenomics

- "The NIH Roadmap Epigenomics Mapping Consortium was launched with the goal of producing a public resource of human epigenomic data to catalyze basic biology and disease-oriented research."
- Coverage: 127 tissues/cell types (Roadmap and ENCODE) with coordinated measurements of histone marks, DNA methylation, open chromatin, and TF binding.

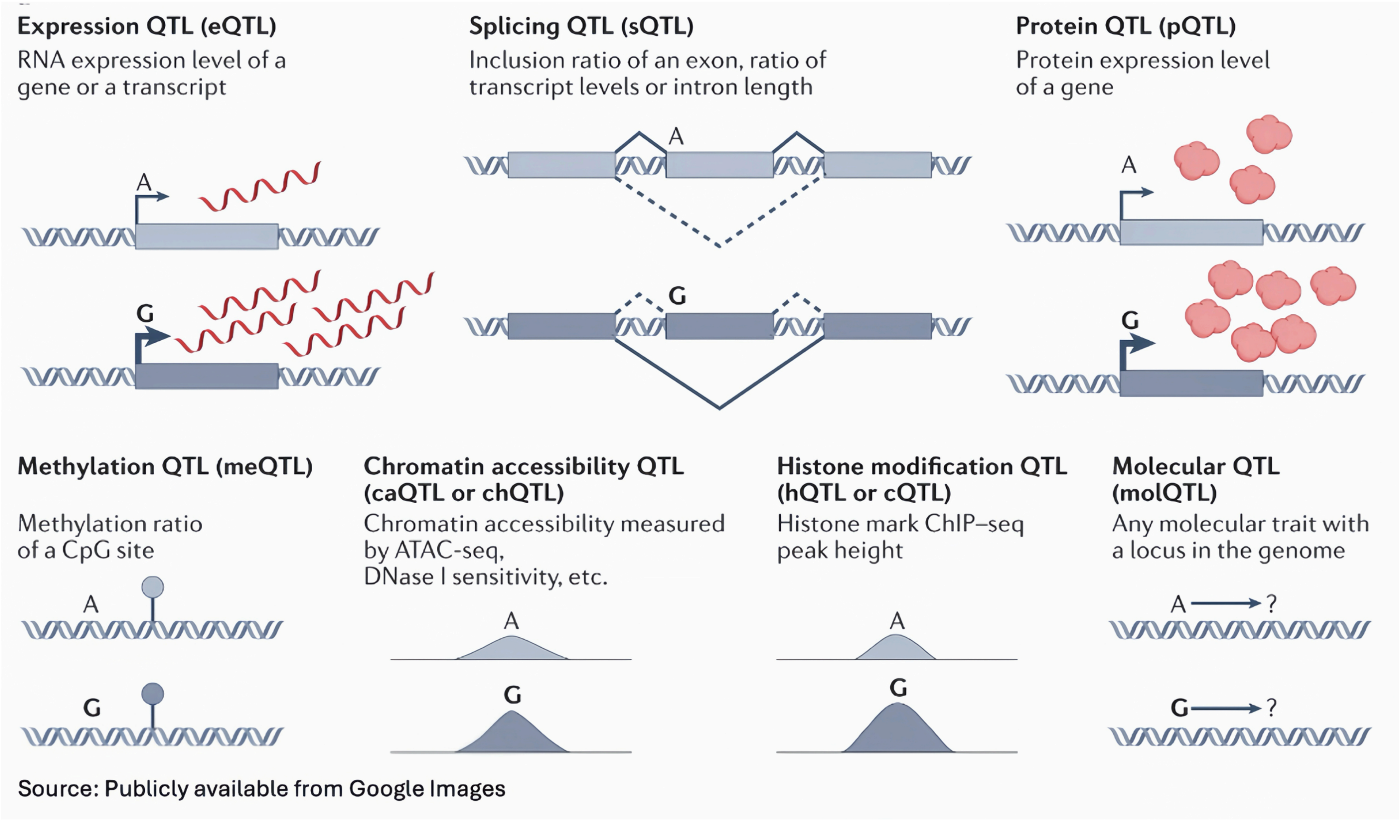
Roadmap Epigenomics



Roadmap Epigenomics Consortium *et al. Nature* **518**, 317-330 (2015) doi:10.1038/nature14248

Molecular QTLs

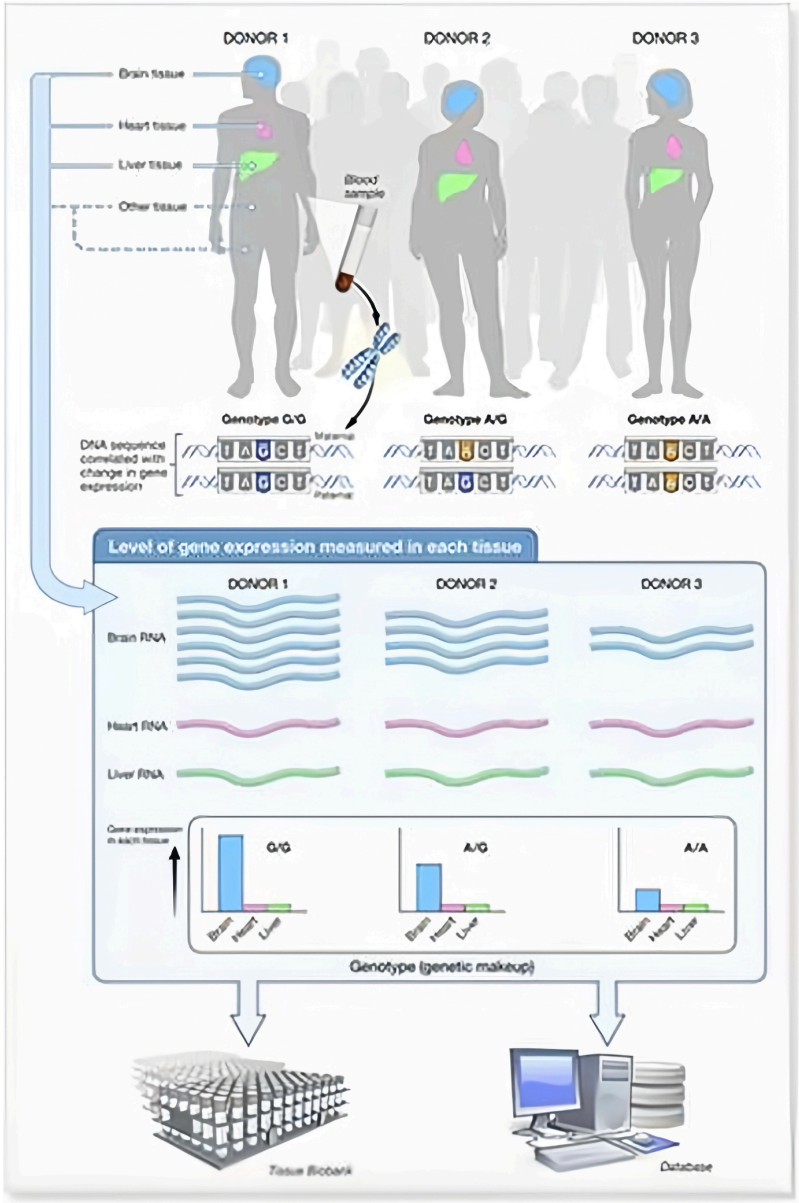
- Using molecular QTLs together lets us build a causal chain from variant → molecular effect → phenotype.



Source: Publicly available from Google Images

GTEx

- Genotype-Tissue Expression project (GTEx) links genotype to expression across tissues.
- Collected DNA and RNA from many human donors, multiple tissues per person.
- For each variant: test whether different genotypes show different gene-expression levels in a tissue.



Source: Publicly available from Google Images

What's Next

- How to integrate GWAS with eQTL and other functional data (colocalization frameworks).
- Web tools to visualize GWAS + eQTL signals and perform practical colocalization analysis.

What questions do you have about anything from today?