Post-GWAS Analyses II

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

What You Will Learn Today

- How to integrate GWAS with eQTL and other functional data
- Key tools and workflows: COLOC, Simple Sum, LocusFocus.
- The concept and basic science of polygenic risk scores (PRS)
- PheWAS, pleiotropy, and biobanks

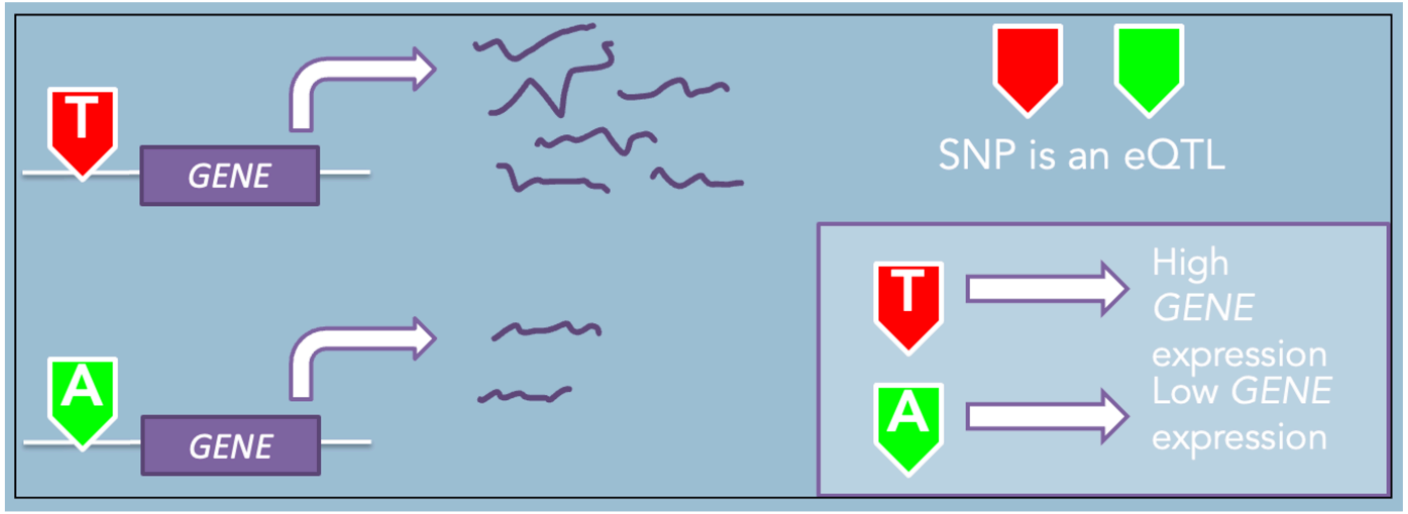
Colocalization Analyses

Genetic Association Analysis - Review

$$g\left\{E\left[\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}\right]\right\} = \begin{pmatrix} g_{1,j} \\ g_{2,j} \\ \vdots \\ g_{n,j} \end{pmatrix} \beta_j + \begin{pmatrix} x_{1,1} & \cdots & x_{1,q} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,q} \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_q \end{pmatrix}$$

- y_i : phenotype for i^{th} individual
- $g_{i,j}$: genotype for i^{th} individual at j^{th} SNP; $g_{ij} = 0, 1$ or 2
- $x_{i,j}$: other covariates
- Repeat the regression analysis ($H_0 : \beta_j = 0$) for $j = 1, 2, \ldots, M$.
  $\rightarrow$ GWAS summary statistics: $Z = (Z_1, Z_2, \ldots Z_m)$.

Expression Quantitative Trait Loci (eQTL)



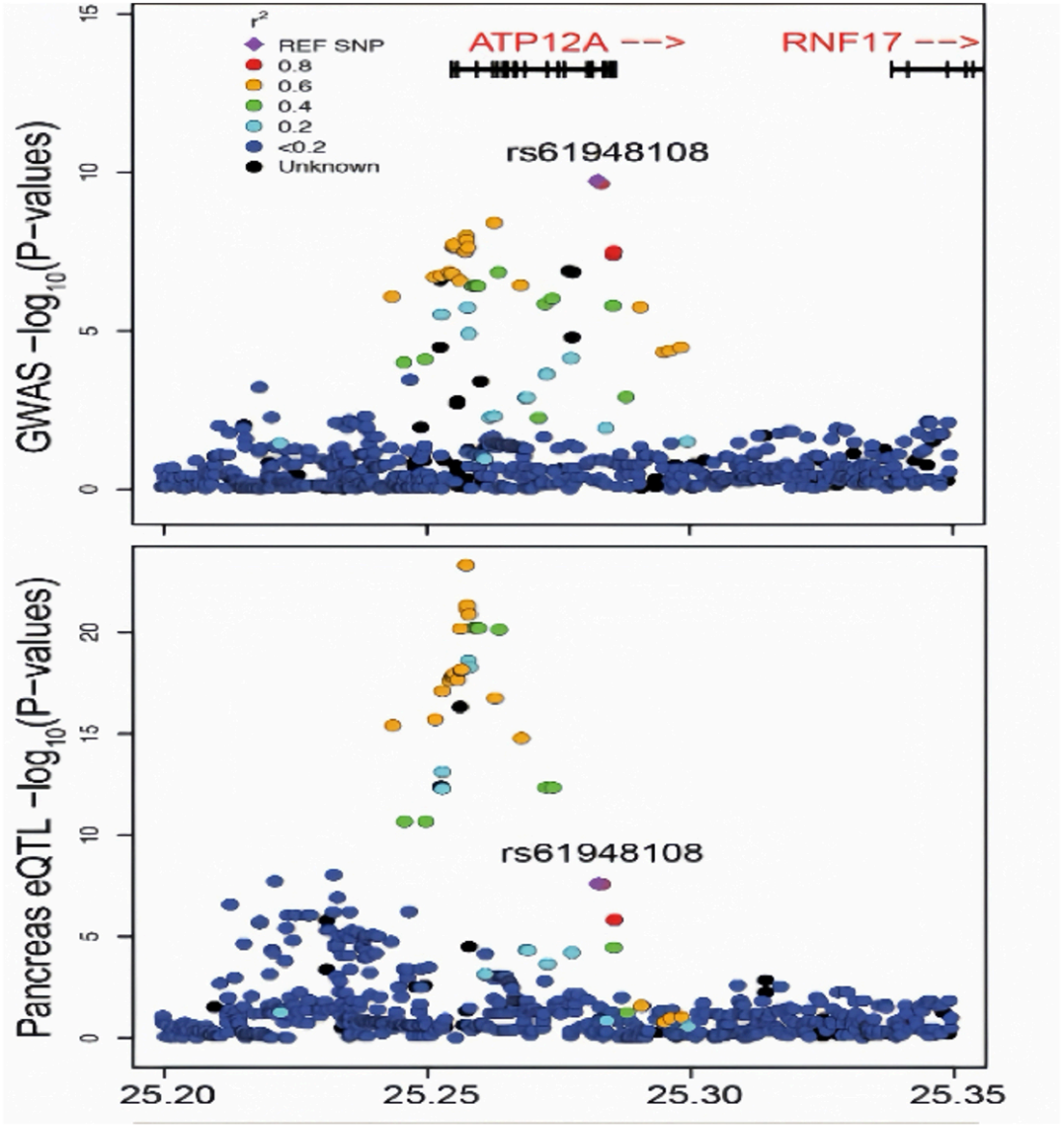
Source: Image created by Fan Wang

eQTL study:

- y_i : (normalized) gene expression (for a particular gene and tissue)
- eQTL summary statistics: $T = (T_1, T_2, \ldots T_m)$

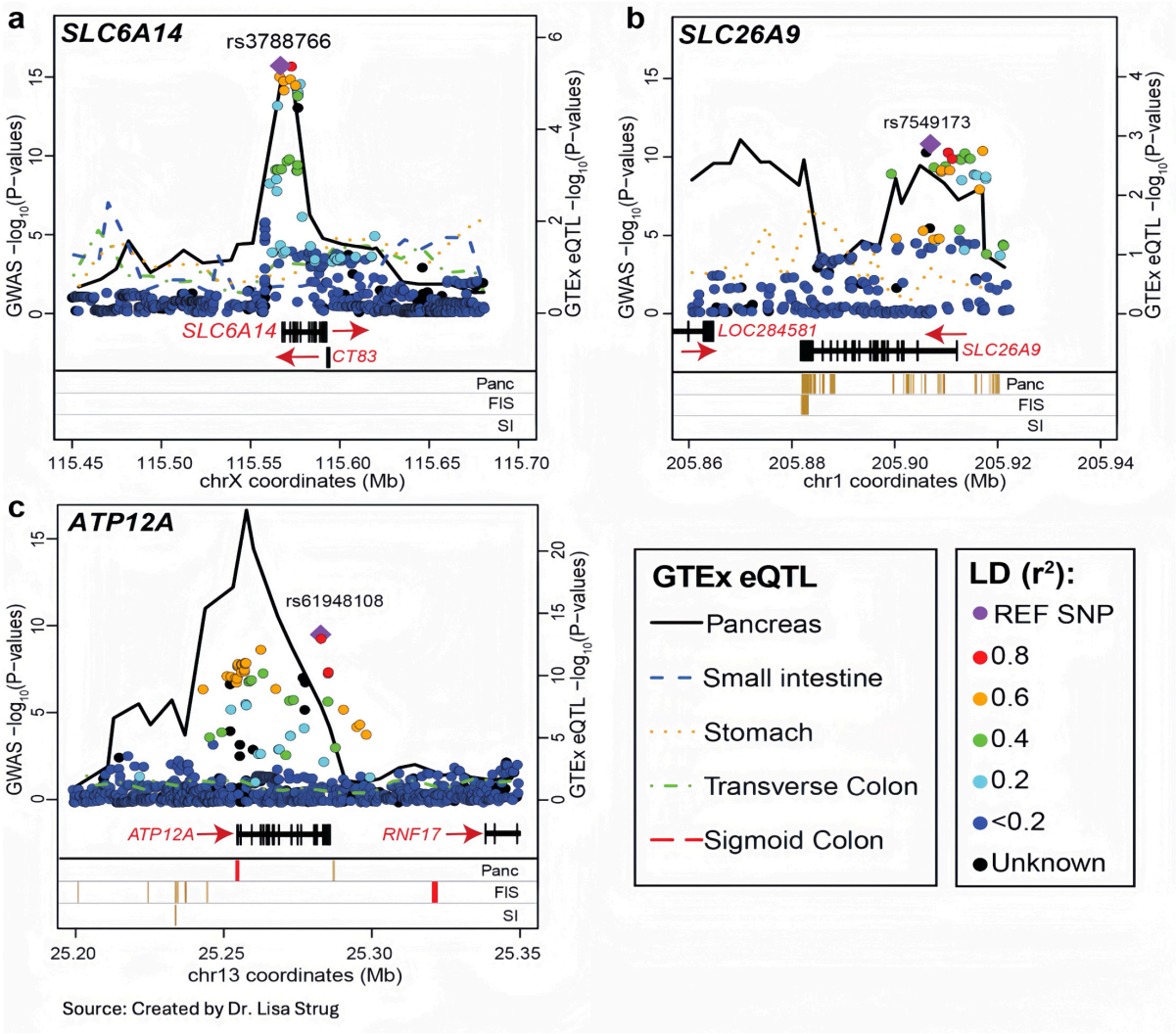
Visualizing GWAS and eQTLs

- We want to test whether eQTL p-values and GWAS p-values have similar/overlapping pattern at the set of same SNPs → Colocalization analysis

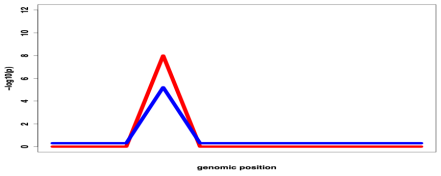
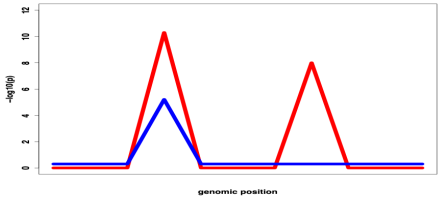
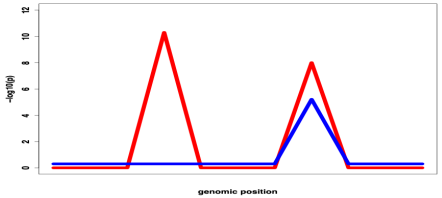


Source: Created by Dr. Lisa Strug

Visualizing GWAS and eQTLs

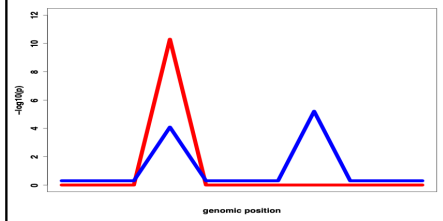
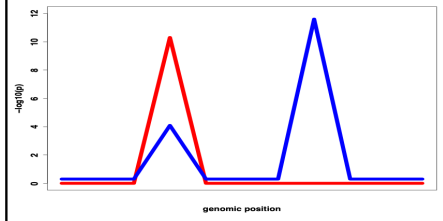
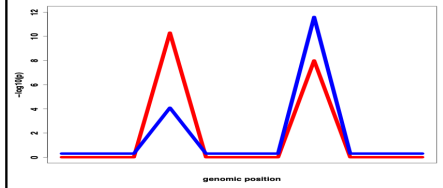


Composite Null Hypothesis

Different GWAS-eQTL Colocalization Scenarios	**Illustration (red: GWAS, blue: eQTL)**
Alter1: ONE GWAS SNP, and ONE eQTL	
Alter2: The eQTL peak overlapped with the higher GWAS peak	
Alter3: The eQTL peak overlapped with the lower GWAS peak	

Source: Image created by Fan Wang

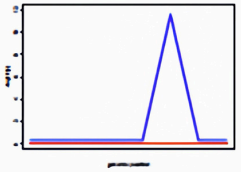
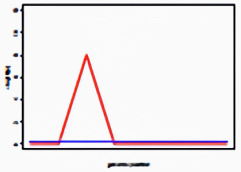
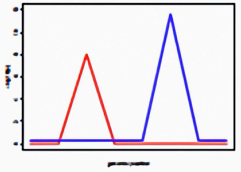
Composite Null Hypothesis

Different GWAS-eQTL Colocalization Scenarios	**Illustration (red: GWAS, blue: eQTL)**
Alter4: The non-overlapped eQTL peak is lower than the GWAS peak	
Alter5: The non-overlapped eQTL peak is higher than the GWAS peak	
Alter6: TWO overlapping GWAS SNPs and eQTLs	

Source: Image created by Fan Wang

Challenges

1. Composite null hypothesis

The null cases considered	Illustration (red:GWAS, blue:eQTL)
H_{01}: NO GWAS causal variants, and NO eQTLs	
H_{02}: NO GWAS causal variants, and YES eQTLs	
H_{03}: YES GWAS causal variants, and NO eQTLs	
H_{04}: YES GWAS causal variants, and YES eQTLs, but signal occurring at independent SNPs	

Source: Image created by Fan Wang

2. High linkage disequilibrium (LD; correlation between SNPs)
3. Allelic heterogeneity (multiple causal variants)
4. Meta-analysis with/without Related individuals within sub-studies
5. Overlapping/Related samples between studies
6. Individual-level data is not available

Exercise: Colocalization vs LD

Consider a locus where:

- GWAS identifies a strong association with type 2 diabetes (lead SNP: A).
- An eQTL study in pancreatic islets finds a strong association with expression of gene X in the same region, but the lead eQTL SNP is different (lead SNP: B).

Question:

1. Why might the lead GWAS SNP and the lead eQTL SNP be different even if there is a shared causal variant?

Methods to Integrate GWAS and eQTL Data

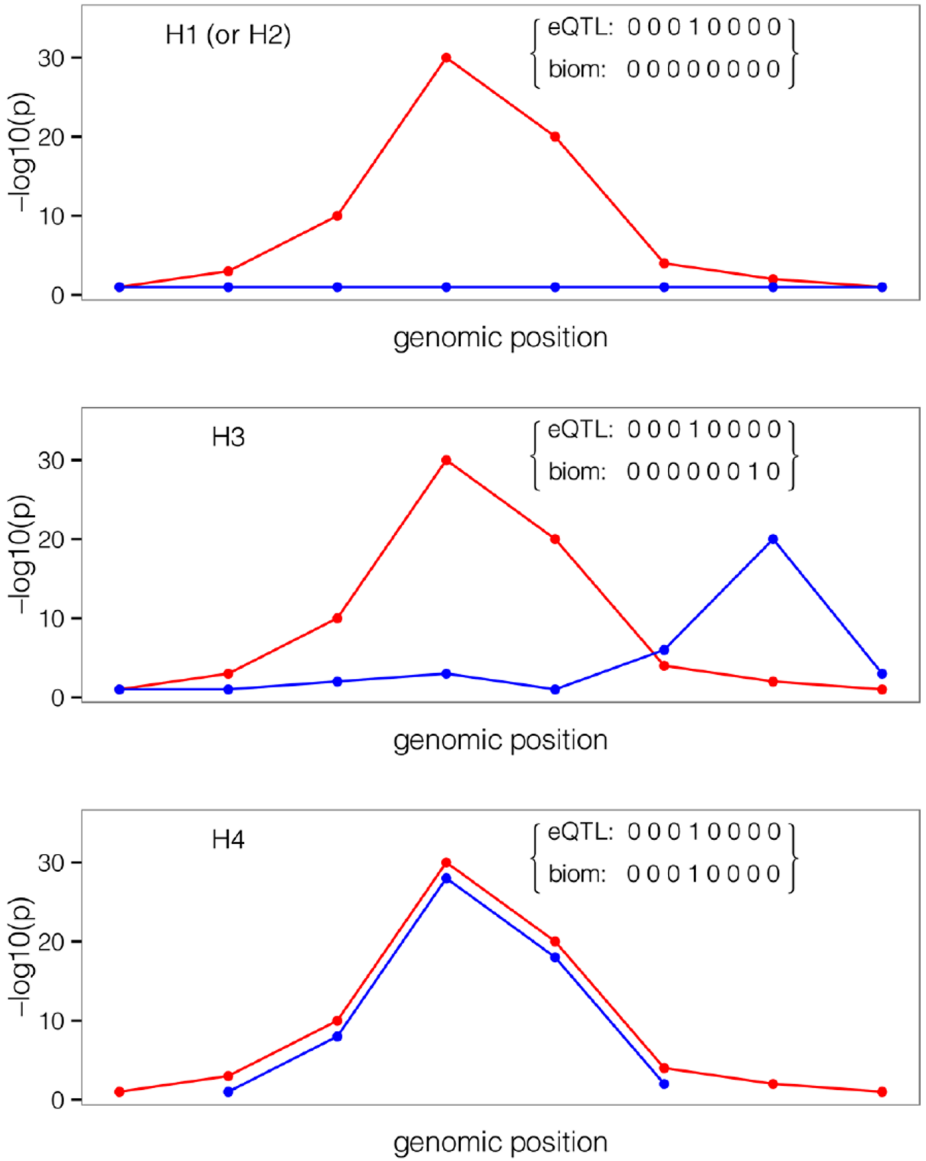
- **Bayesian approaches** aim to identify shared causal variants contributes to both the disease outcome and gene expression variation.
    - **COLOC , eCAVIAR, GWAS-PW**.etc.
- **Frequentist-based methods:**
- Methods that impute gene expression based on a reference, and then associate imputed expression with the trait:
    - **PrediXcan and TWAS**
- Integration methods based on Mendelian randomization:
    - **SMR, SMR-multi**, etc.
- Overlapping pattern:
    - **Simple Sum2**

COLOC

- Calculates posterior probability for 5 cases (from H0 to H4)

$$P\left(H_h \mid D\right) \propto \sum_{S \in S_h} P(D \mid S) P(S)$$

- Prior probabilities: Set at SNP level - typically $\mathrm{P}_1 = P_2 = 10^{-4}$ for association with one trait, and $P_{12} = 10^{-6}$ for colocalization (shared association).



Source: Image created by Fan Wang

Extension of COLOC

- GWAS-PW (Pickrell et al. 2016):
  - Extends COLOC by empirically estimating priors from the genome-wide data for the five hypotheses.
  - Incorporates sample relatedness
- COLOC2 (Dobbyn et al. 2017):
  - Uses estimated proportions in GWAS-PW as priors (or optionally, coloc default or user-specified priors) in the calculation of the posterior probability.
  - For a locus with multiple independent eQTL signals, adopting **a forward stepwise conditional analysis for the eQTL study**.
  - For a gene with k independent eQTLs, they run k colocalization models.

Application

- GWAS signature for IREB2 Colocalizes with the Conditional eQTL signature.

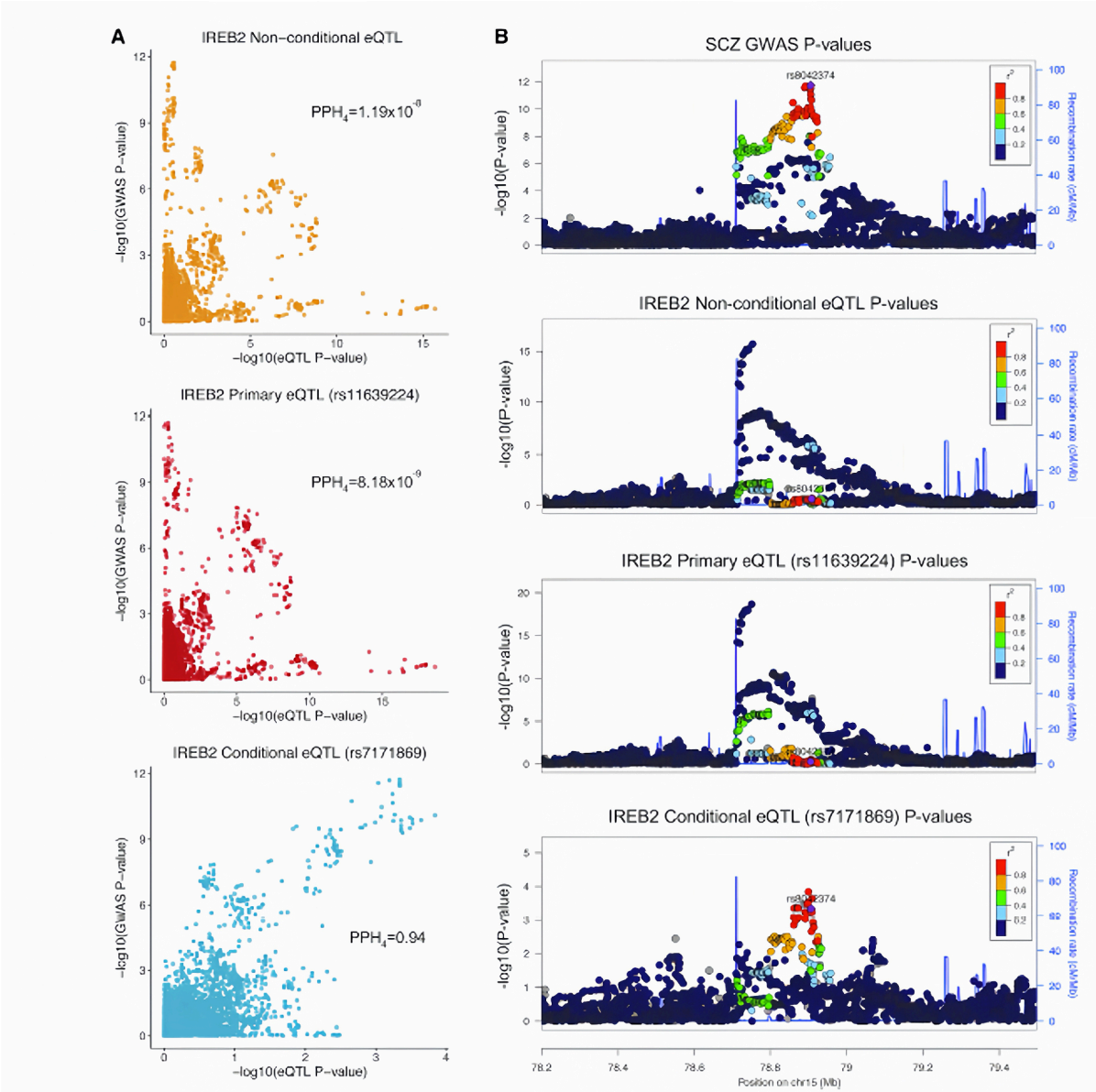


Figure 4. GWAS Signature for *IREB2* Co-localizes with the Conditional eQTL Signature
Source: Publicly available from Google Images

Pros and Cons of COLOC2

- Advantages:
  - Quick computation by Approximate Bayes Factor without iterative computation scheme (such as Markov Chain Monte Carlo) is required.
  - Distinguished evidence for non-colocalization.(H0-H4)
- Limitations:
  - Single-causal-variant assumption per trait within the region (can mislead in polygenic/allelic-heterogeneity settings).
  - Conditional analysis can be costly when extending to multiple signals or larger regions.

SuSiE-COLOC

- SuSiE is a fine-mapping framework to distinguish multiple signals for a given trait, and is more computationally efficient than the forward stepwise conditional analysis.
- SuSiE outputs a 95% credible set selecting a subset of L signals with inferred causal SNPs.
- If L_1 and L_2 signals are selected for trait1 and trait2, respectively, we run COLOC $L_1 \times L_2$ times on all possible pairs of signals between the traits.

ECAVIAR

(Hormozdiari et al. 2016)

- Simultaneously performing fine-mapping for GWAS and eQTL studies by considering almost all combinations of causal status between SNPs
- Estimate the posterior probability that the same variant is causal in both studies :

$$\mathrm{P}\left(c_i^{(\mathrm{p})}=1, c_i^{(\mathrm{e})}=1 \mid S^{(\mathrm{p})}, S^{(\mathrm{e})}\right)=\mathrm{P}\left(c_i^{(\mathrm{p})}=1 \mid S^{(\mathrm{p})}\right) \times \mathrm{P}\left(c_i^{(\mathrm{e})}=1 \mid S^{(\mathrm{e})}\right).$$

- $c_i^{(p)}$ and $c_i^{(e)}$: indicator that SNP i is causal in the GWAS (phenotype) study and eQTL study, respectively.
- $S^{(p)}$ and $S^{(e)}$: vector of Z-scores from the GWAS and eQTL study, respectively.

Pros and Cons of ECAVIAR

Advantage:

- Higher true positive rate compared to COLOC when multiple variants are causal in a locus with low LD.

Limitations:

- Computation can be very long when number of SNPs is huge
- Imposes an assumption on the maximum number of causal SNPs in a locus (six)
- The posterior probability of colocalization will be averaged across several variants in high LD, resulting in ambiguous colocalization conclusions.

Extensions to Multi-Trait Colocalization

- **Moloc** evaluates all possible configurations of shared vs. distinct causal variants across traits.
  - Computationally intense: number of configurations $= 2^T - 1$ (for T traits)
  - Good when you want explicit posterior probabilities for all trait-sharing scenarios
- **Hyprcoloc** uses a greedy algorithm to group traits that share a causal variant.
  - Prioritizes traits that colocalize strongly first, then expands to others
  - Hyprcoloc is suited for large numbers of traits (e.g., >10).

Further Issues

- eCAVIAR evaluates multiple causal configurations but is computationally intensive with many SNPs.
- SuSiE-COLOC reduces complexity but still requires $\mathrm{L}_1 \times \mathrm{L}_2$ coloc runs.
- Reduced power when there is high LD
- Few formal corrections for multiple hypotheses across loci.
- Posterior probabilities (e.g., PPH4) must be interpreted with caution in genome-wide scans.
- Adjusting priors using false positive report probability (FPRP) to control for multiple testing.

Simple Sum

- A frequentist integration method that combines GWAS and eQTL signals within a locus.
- Particularly powerful in regions with high LD and allelic heterogeneity (multiple causal variants).
- The extension Simple Sum 2 (SS2) is designed to control type I error under a composite null, correct for multiple testing, and handle meta-analysis / sample relatedness.

Chi-square (χ^2) Distribution: Quick Review

- Fact from probability: if $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$.
- Consider **independent** $Z_1, \ldots, Z_k \sim N(0,1)$.
  - Define a statistic $S = \sum_{j=1}^{k} Z_j^2$.
  - Then $S \sim \chi^2_k$, a chi-square distribution with k **degrees of freedom**.
- In practice, SNP Z-scores within a locus are **correlated** because of LD.

Weighted χ^2 Distributions: Quick Review

- Let $Z = (Z_1, \ldots, Z_k)^\top \sim N(0, \Sigma)$, where Σ is the LD (correlation) matrix.
- Consider the statistic $S = \sum_{j=1}^{k} Z_j^2$.
- Using the eigenvalues $f_1, \ldots, f_k$ of the LD matrix Σ, we can write

$$S \overset{d}{\approx} \sum_{j=1}^{k} f_j \chi^2_{1,j},$$

  a **weighted sum of independent χ^2_1 variables**.

- Intuition: each eigenvalue f_j represents how much variation comes from one "independent direction" of the LD structure; larger f_j give more weight in the sum.

Quadratic Forms: Quick Review

- Many region- or gene-level test statistics can be written as a **quadratic form** $S = \sum_i a_{i,j} Z_i Z_j = Z^\top A Z$, for some symmetric matrix A (e.g. weights on SNPs).
  - For example, $A = I$.
- In practice, software computes the eigenvalues and then uses this weighted χ^2 distribution to obtain the **p-value** under the null (i.e., $Z \sim N(0, \Sigma)$):

$$S = Z^\top A Z \overset{d}{\approx} \sum_{j=1}^{m} d_j \chi^2_{1,j}.$$

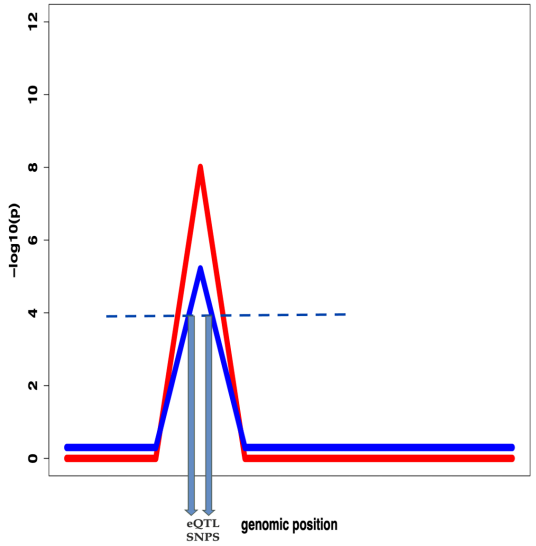
- T behaves like a **weighted sum of independent χ^2_1 variables**, with weights are the eigenvalues d_j of matrix $\Sigma^{1/2} A \Sigma^{1/2}$.

Simple Sum

[Gong*, Wang*, Xiao*, et al., 2019. PLOS Genetics.]

- Dichotomized eQTL evidence (threshold τ):

$$SS = \frac{1}{\sum_{j=1}^{m} I\left(|T_j| \geq \tau\right)} \sum_{j=1}^{m} Z_j^2 I\left(|T_j| \geq \tau\right) - \frac{1}{\sum_{j=1}^{m} I\left(|T_j| < \tau\right)} \sum_{j=1}^{m} Z_j^2 I\left(|T_j| < \tau\right)$$

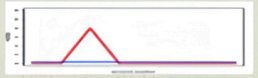
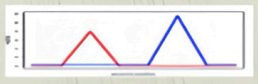


Source: Image created by Fan Wang

One-sided test is performed because only a significant positive value suggests colocalization of GWAS association and eQTL evidence.

Simple Sum

- $$SS = \sum_j Z_j^2 \left[\frac{T_j^2 - \frac{\sum_j T_j^2}{m}}{\left(\sum_j T_j^2 - \frac{\sum_j T_j^2}{m} \right)^2} \right] = Z' A Z,\ A = \mathrm{diag}(a_j), a_j = \left[\frac{T_j^2 - \frac{\sum_j T_j^2}{m}}{\left(\sum_j T_j^2 - \frac{\sum_j T_j^2}{m} \right)^2} \right].$$

Illustration (red: GWAS, blue: eQTL)	SS	Type I error Control
H_{01}: 	$SS \sim \sum_j d_j \chi_1^2$, where d_j' s are eigenvalues of $\left(\Sigma^{\frac{1}{2}}\right)' A \Sigma^{\frac{1}{2}}$.	✓
H_{02}: 		✓
H_{03}: 	***Require arbitrary eQTL threshold***	✓
H_{04}: 	the *SS* test statistic would have negative mean, resulting in a large p-value if a one-sided test was performed	✓

Source: Image created by Fan Wang

Simple Sum 2

- Stage 1: Formally test eQTLs by $\sum_{j=1}^{m} T_j^2$.
- Under the null of no eQTLs (H_{01} and H_{03}), $\sum_{j=1}^{m} T_j^2 \sim \sum_{j=1}^{m} f_j \chi_1^2$, where $f_j's$ are the eigenvalues of Σ (LD matrix).
- Stage 2: Perform the Simple Sum test $\sum_j Z_j^2 \left[\frac{T_j^2 - \frac{\sum_j T_j^2}{m}}{\left(\sum_j T_j^2 - \frac{\sum_j T_j^2}{m} \right)^2} \right]$.
- The type I error rate for a single test under composite null hypothesis is controlled at α .
- Bounded by the maximum type I error rates for the stage 1 and the stage 2 test.

Complex Data Scenarios

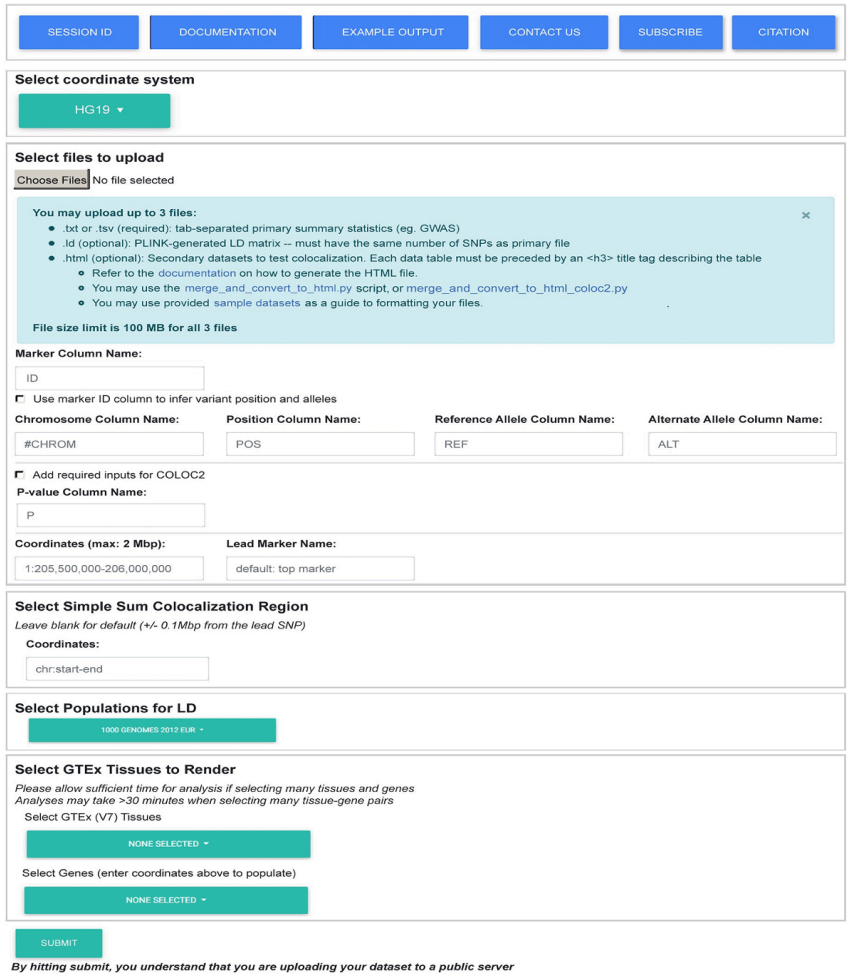
- Many gene-tissue tests: extend to multiple gene-tissue pairs and control the family-wise error rate (FWER) across tests (e.g., Bonferroni correction).
- Meta-analysis with related individuals: combine sub-studies by modeling the covariance of SS2 test statistics to account for relatedness.
- Sample overlap: explicitly account for overlap/relatedness between GWAS and eQTL cohorts, adjusting cross-study correlations when performing inference.

LocusFocus

(https://locusfocus.research.sickkids.ca/)

- SS2 is implemented in this web-based tool, which enables integration of GWAS summary statistics with any secondary SNP-level dataset such as eQTL, mQTL, or other phenotypic associations from GWAS.
- The tool is developed to **conduct set/gene-based testing, colocalization analysis** and **visualization of signals**.
- The eQTL summary statistics from GTEx V8 are made available for selection within the web server to test colocalization with tissues and genes.
- COLOC2 colocalization testing is also available.

LocusFocus Input



Source: Created by Fan Wang

→ Documentation and example output is provided

→ Enable selection of of the hg19 or hg38 coordinate systems

Upload buttons for SS2 analyses:

→ 1. Summary statistics for the primary dataset

2. The LD matrix (i.e. Plink)

3. A multi-sample dataset (in HTML format) with the secondary summary statistics

Other Options:

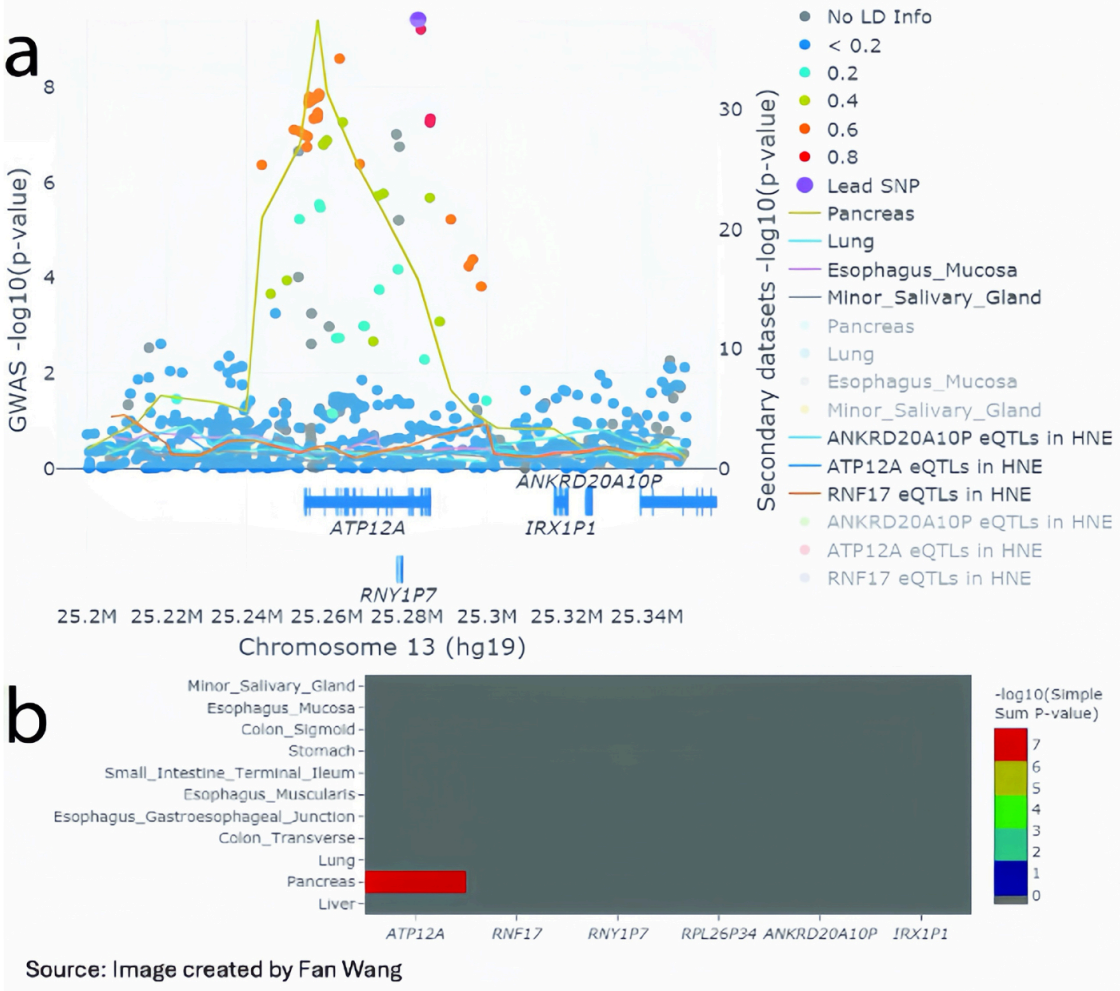
→ Select coordinates to view plot results and to conduct analyses

→ Select population for the external LD matrix

→ Select GTEX tissues and genes (all 48 tissues from GTEx (v8) are available)

LocusFocus Output

Sample interactive plot output from the LocusFocus web application.



Source: Image created by Fan Wang

Features:

- Zoom in plots
- Reset, rescale or shift of axes
- Tooltips for each data point
- Save image options in png or svg vector format

The Heatmap summarizes the stage 2 SS tests for all the genes in the user-defined region and across all the selected tissues.

Demo of LocusFocus

Discussion: What can colocalization tell us (and not tell us)?

- Suppose colocalization analysis suggests a shared signal between a GWAS trait and an eQTL for gene X in a relevant tissue.

Questions:

1. What are **limitations** of relying only on colocalization to infer causality?
2. What additional data or analyses would you want before declaring gene X causal?

Introduction to Rare Variation and Sequencing Studies

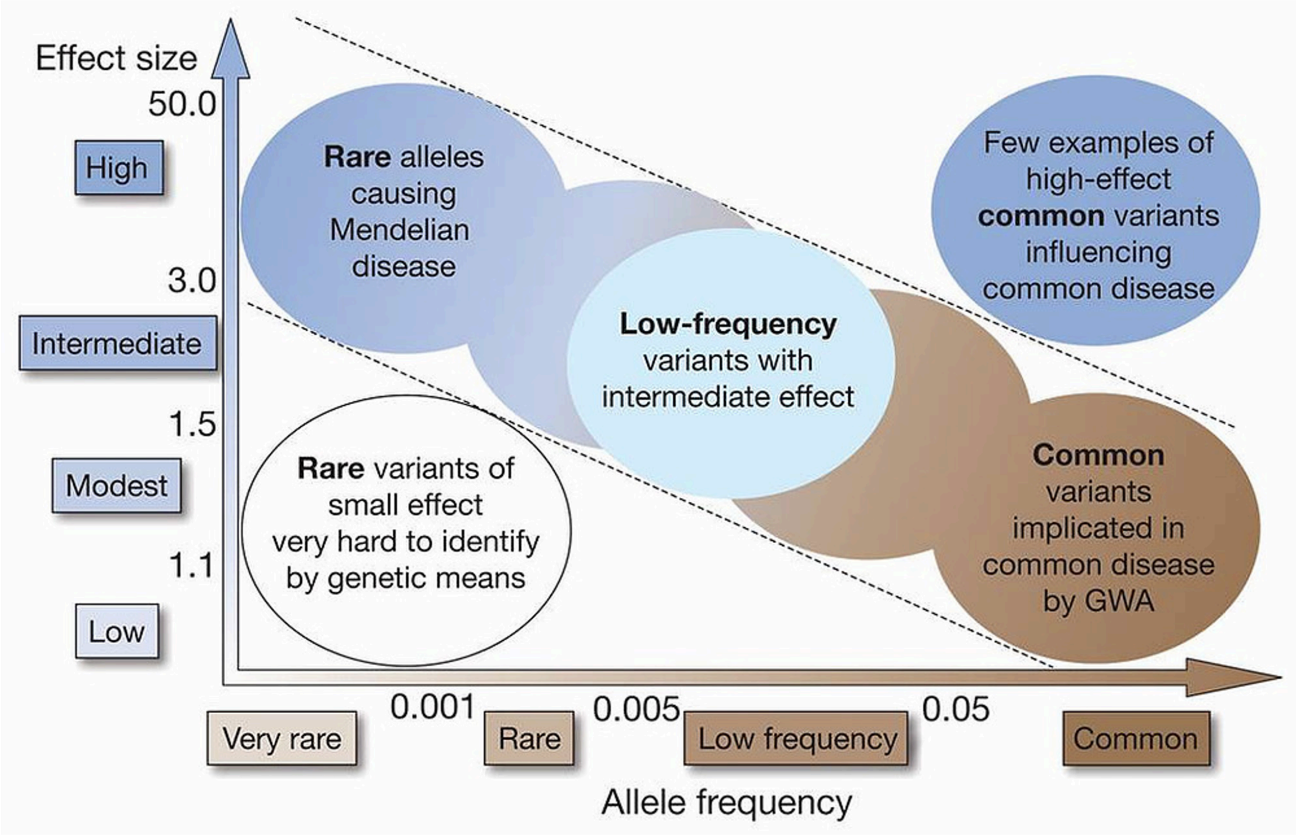
Rare Variation in the Genome

- GWAS typically focus on common variants (minor allele frequency $\geq 5\%$).
- **Sequencing studies** (including whole-exome and whole-genome sequencing) enable the discovery of **rare and low-frequency** variants.
- The vast majority of genetic variants in the human genome are **rare**.
- **De novo mutations** are the rarest type of genetic variation: germline variants present in a child but absent in both parents.

Why study rare variation?

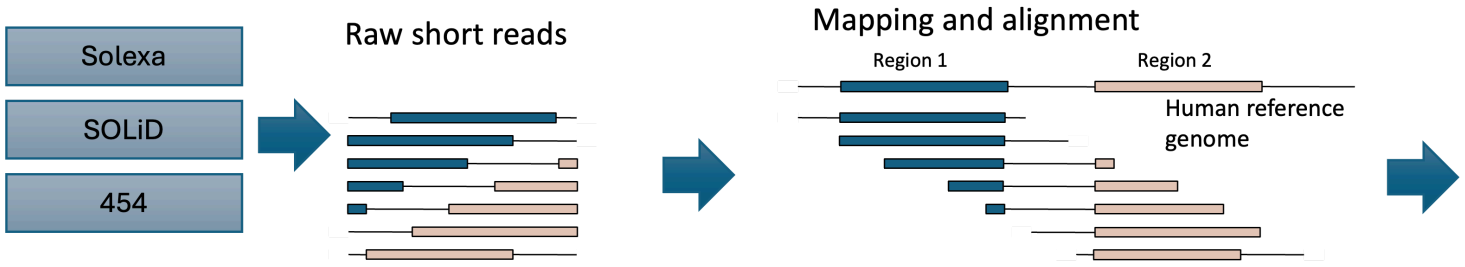
- Rare variants can play an important role in disease, especially for traits strongly influenced by natural selection.
- Mendelian diseases represent an extreme example of the impact of rare variants.
- In autism spectrum disorders (ASD), de novo mutations are estimated to contribute to about 25-30% of cases.
- Rare variants may have comparatively smaller effects on more late-onset diseases (e.g., type 2 diabetes, Alzheimer's disease) and on quantitative traits (e.g., cardiometabolic phenotypes).
- There is an inverse relationship between effect size and allele frequency, so rare variants can have large effects and may be particularly clinically actionable.

Variant Frequency and Effect Size



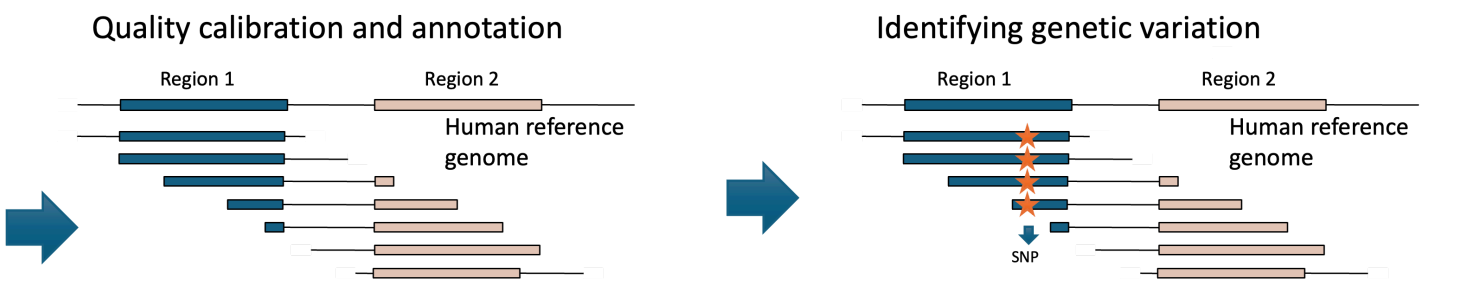
Source: Publicly available from Google Images

From raw short reads to genetic variants in next-generation sequencing studies



A single run of a sequencer generates ~50M ~75 – 400 bp short reads for analysis

The origin of each read from the human genome sequence is found



The quality of each read is calibrated and additional information annotated for downstream analyses

SNPs and indels from the reference are found where the reads collectively provide evidence of a variant

Source: Publicly available from Google Images

Advanced Topics/Looking Toward the Future

- Next-generation sequencing data analysis
- Multi-omic data: microarray, methylation, functional annotation, etc.
- Transcriptome-wide association studies (TWAS)
- Mendelian randomization
- Evaluation of PRS and advanced risk prediction methods
- Integrating WGS data, epigenomic context, and large language models for risk prediction
- ...and many more emerging approaches