

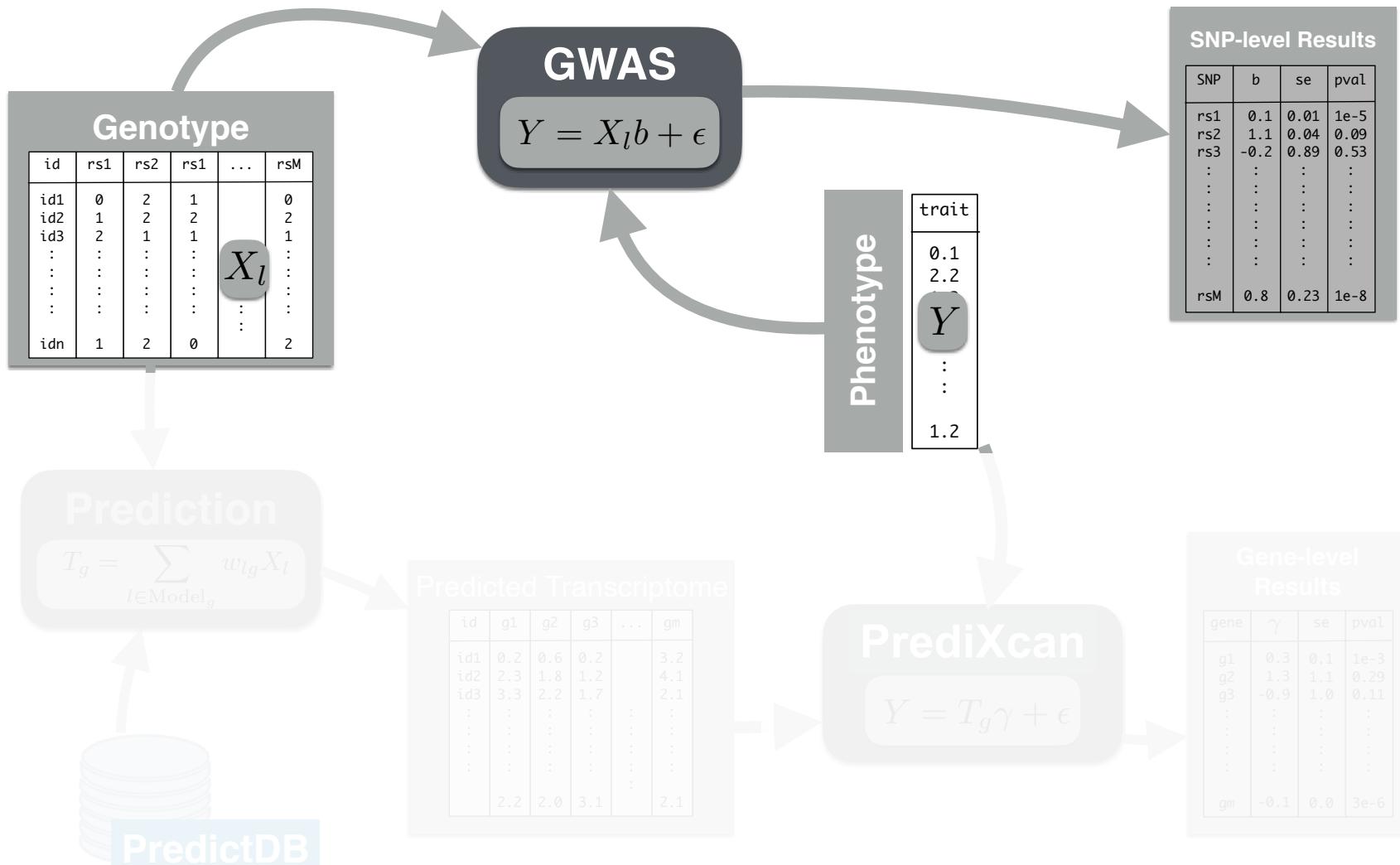
# Transcriptome-based Methods to Assign Function to GWAS Loci

Hae Kyung Im, PhD

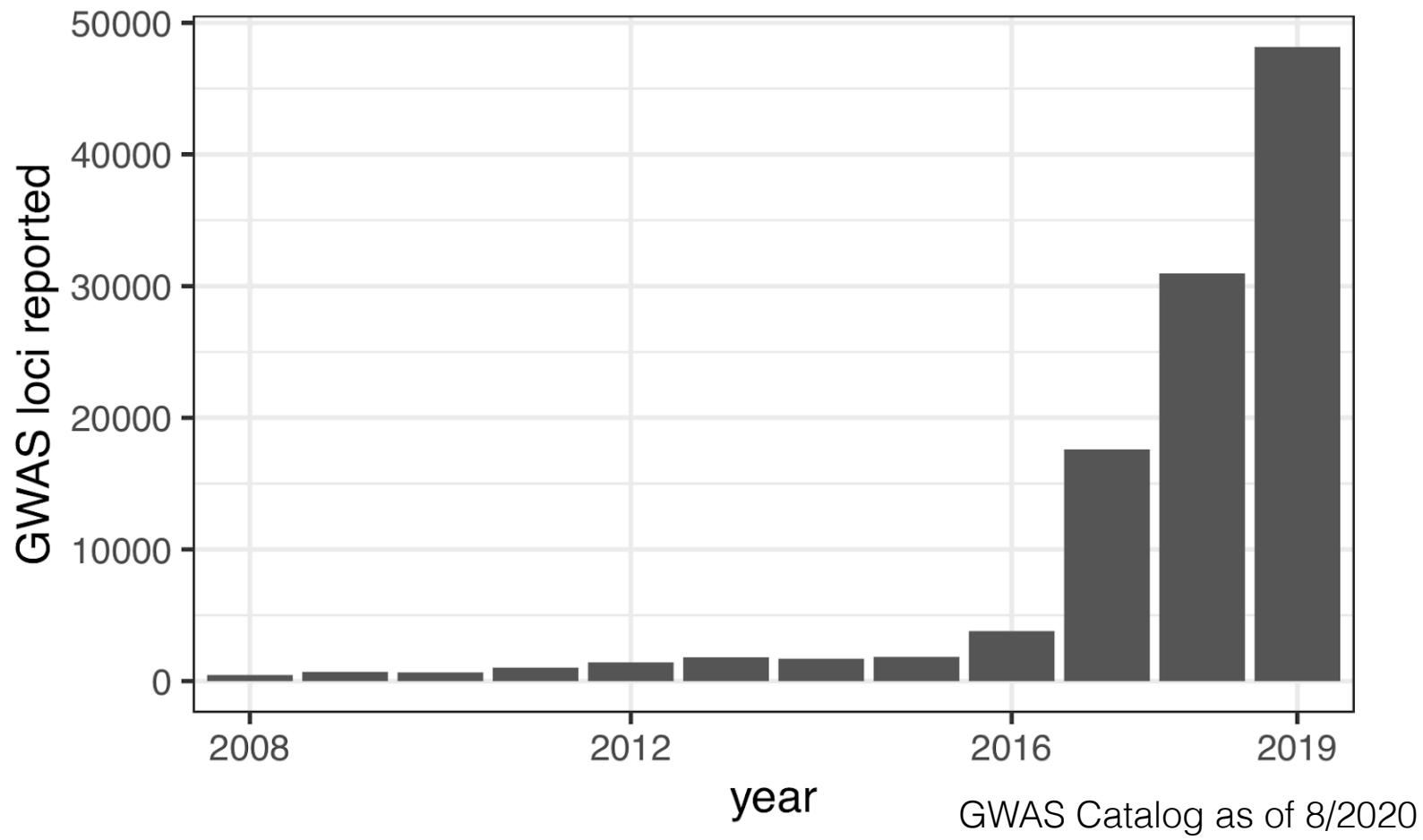


THE UNIVERSITY OF  
**CHICAGO**

Quantitative Genomic Training  
June 25, 2021

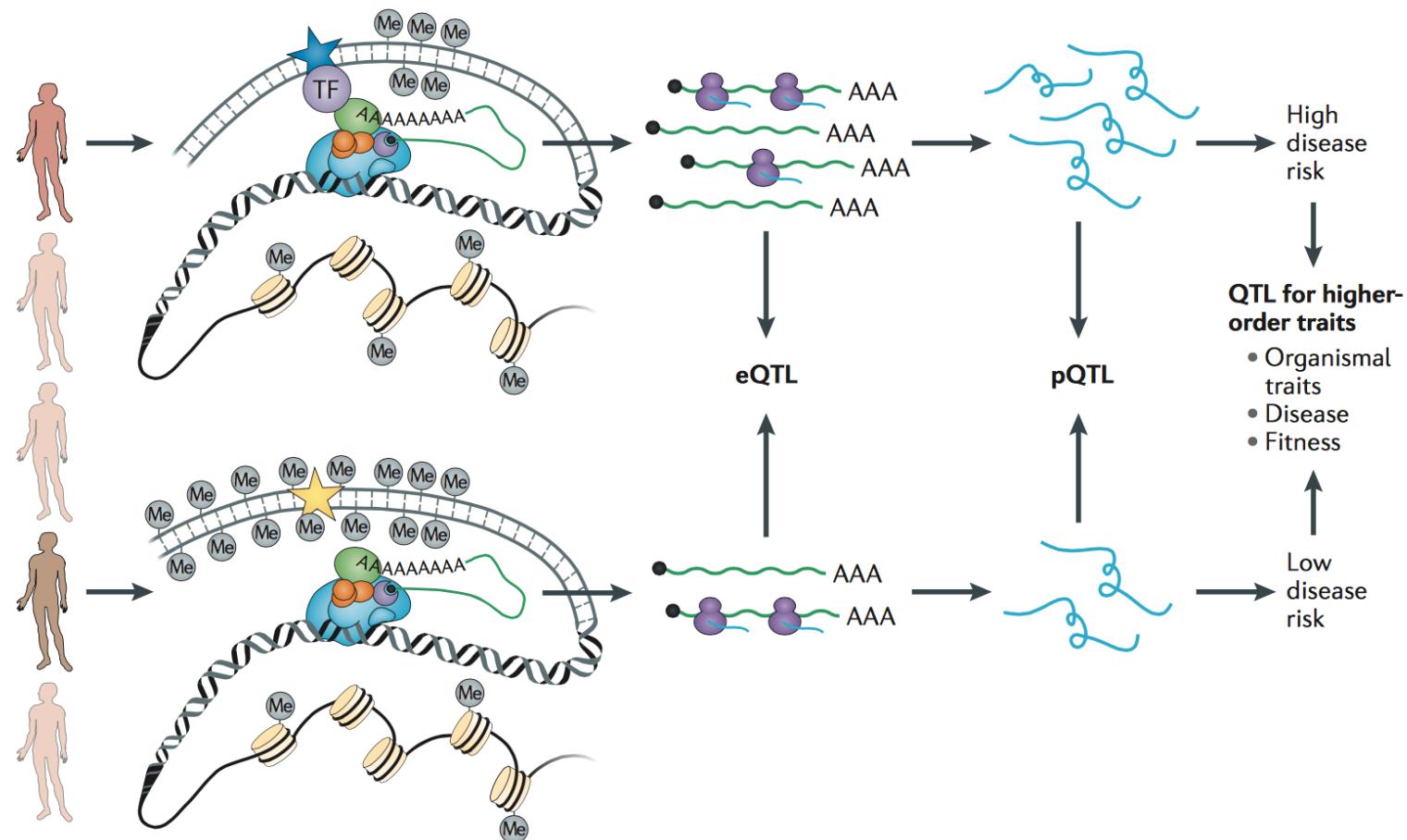


## GWAS Discovered 100K+ Loci



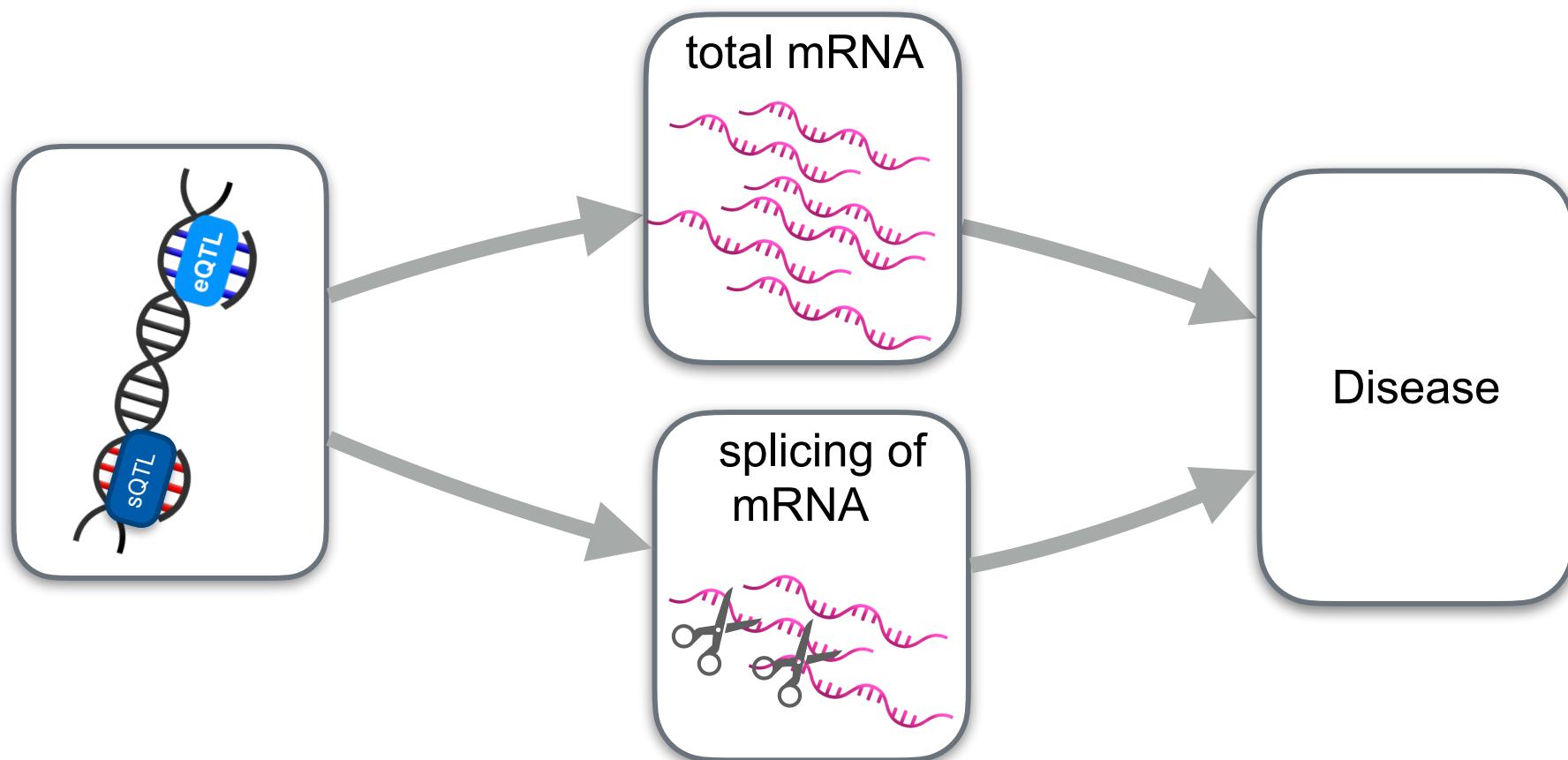
Most GWAS Variants Do  
Not Change Coding of  
Proteins

# Regulation of Gene Expression Levels May Link Genotype with Phenotype

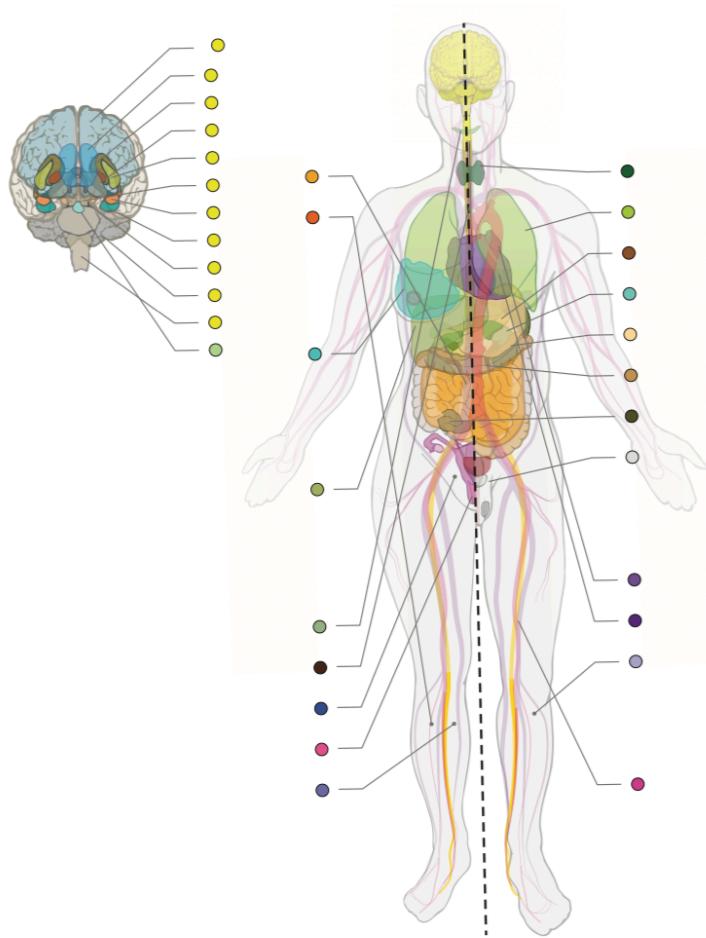


Albert & Kruglyak 2015 NGReviews

## Regulation of mRNA Levels and Alternative Splicing May Explain Genotype to Phenotype Association

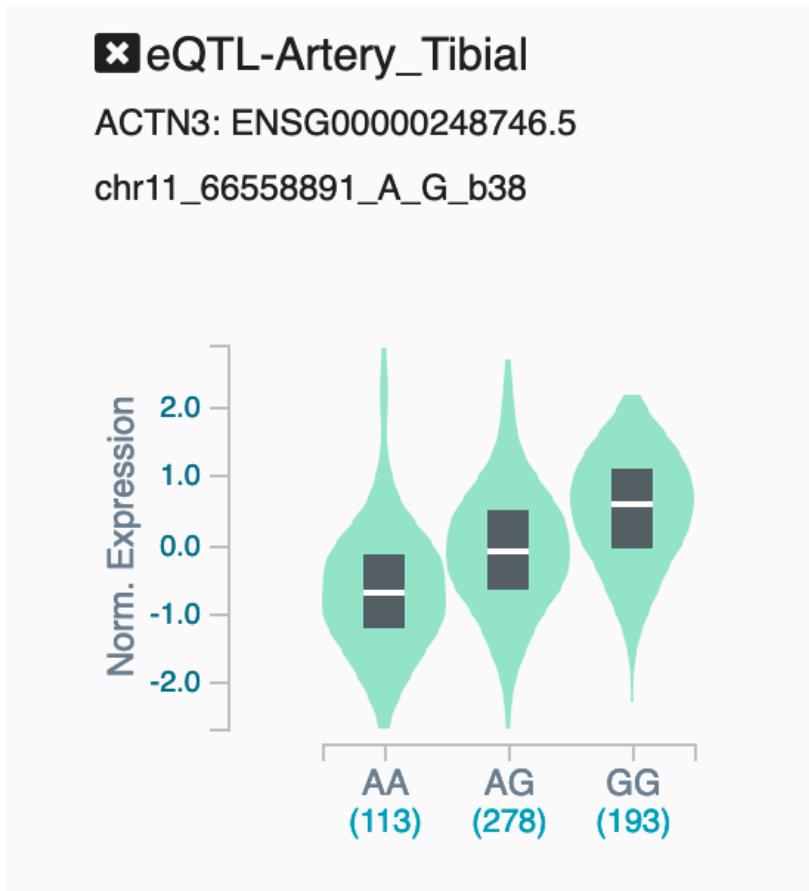


# GTEX Consortium Compiled Atlas of Genetic Effects on the Transcriptome



- 900+ organ donors
- 54 tissues
- 17K+ tissue samples
- mRNA-seq
- WGS 30x

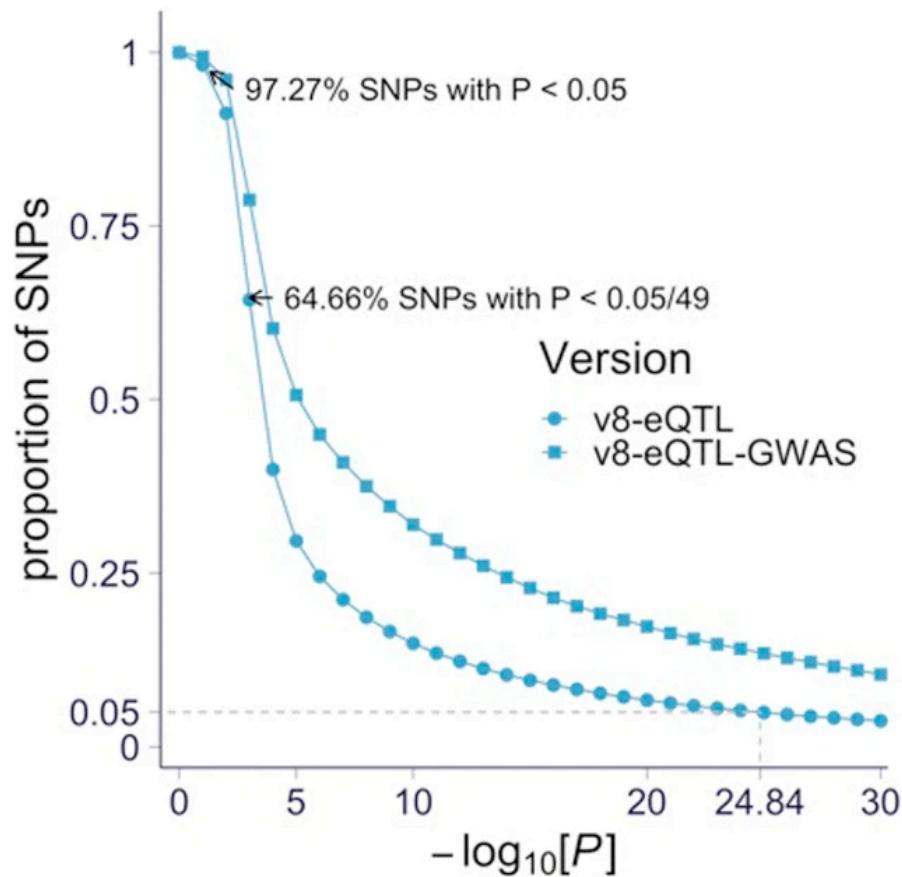
# eQTL: Expression Quantitative Trait Loci



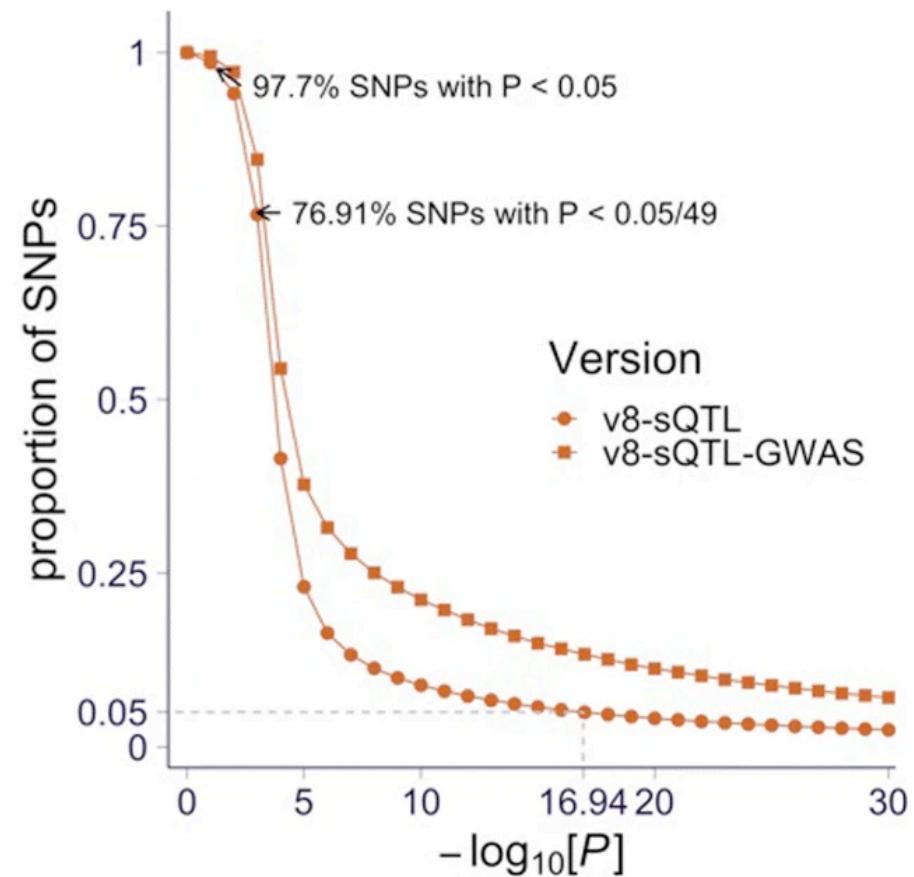
<https://gtexportal.org/home/locusBrowserPage/ACTN3>

# Trait Associated Variants Are Enriched Among eQTLs

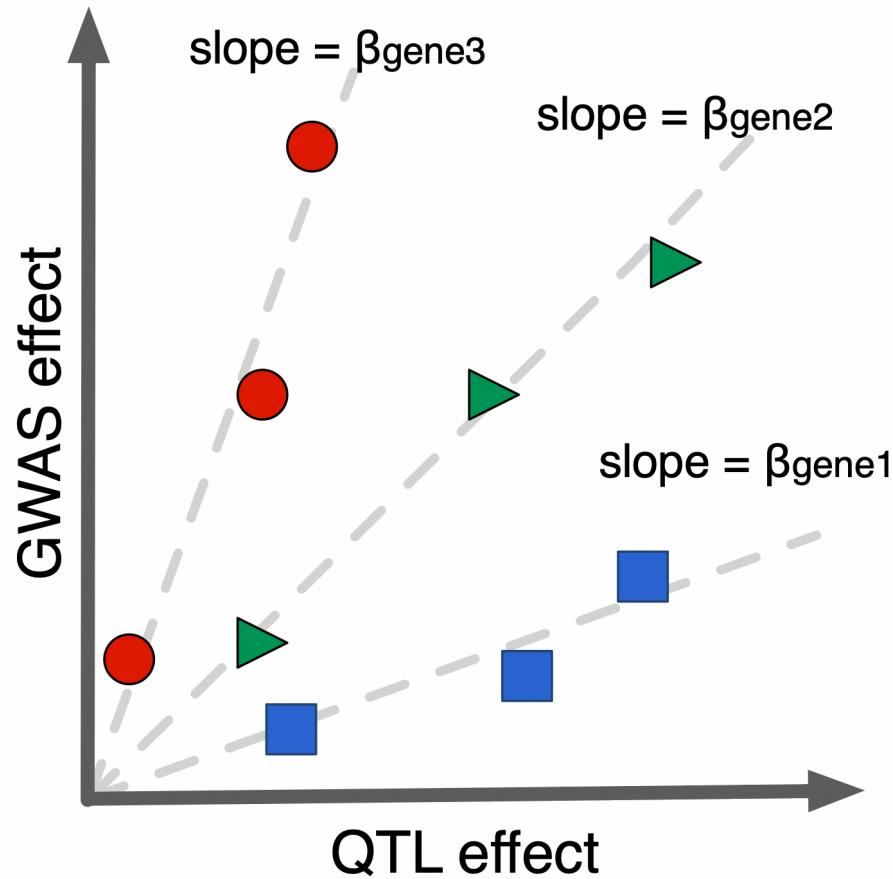
A



B

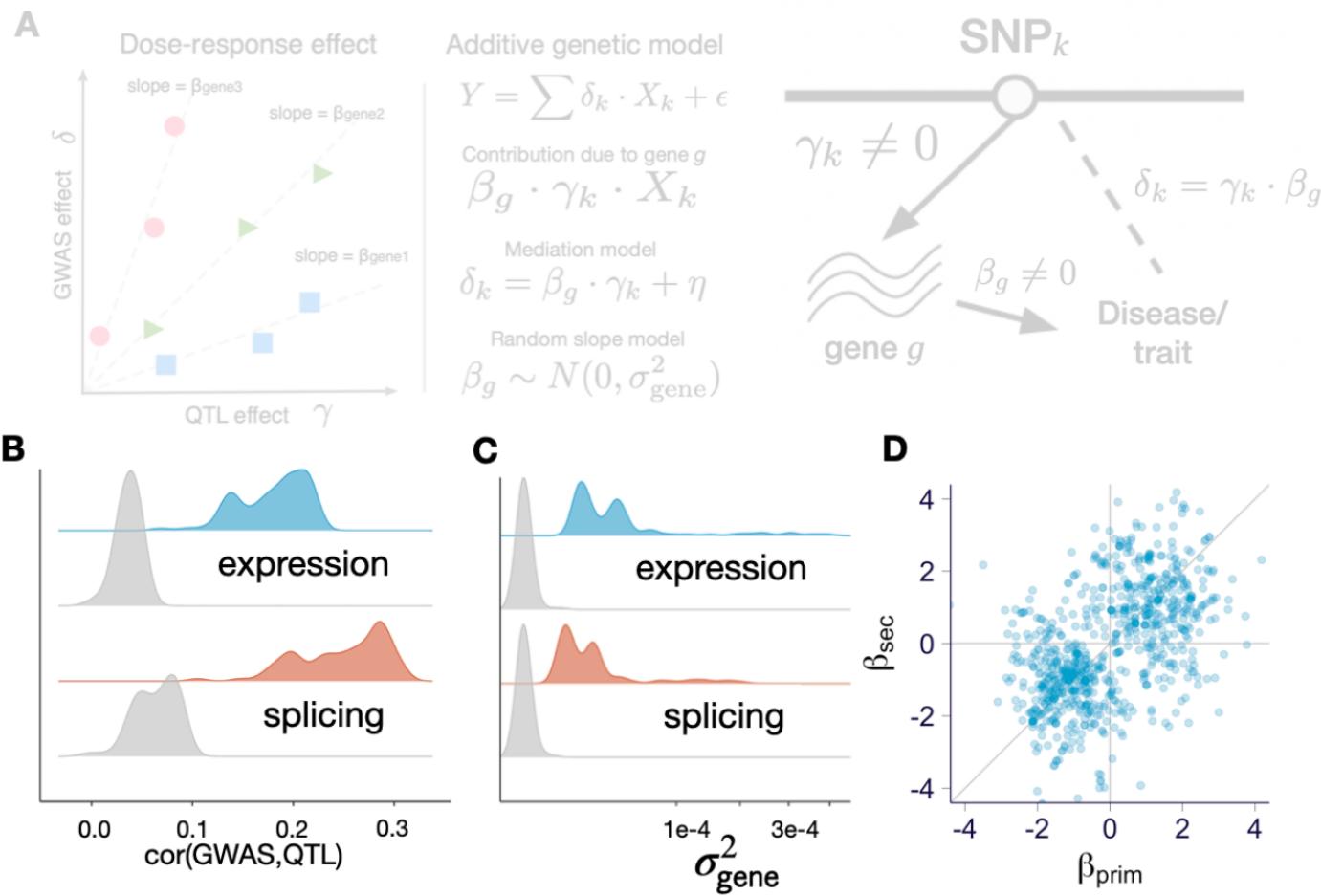


## Do We See a Dose-Response Effect?



A dose-response effect would be a convincing evidence of a causal link

# Consistent Dose-Response Effects of QTLs on Complex Traits Indicate Causal Link

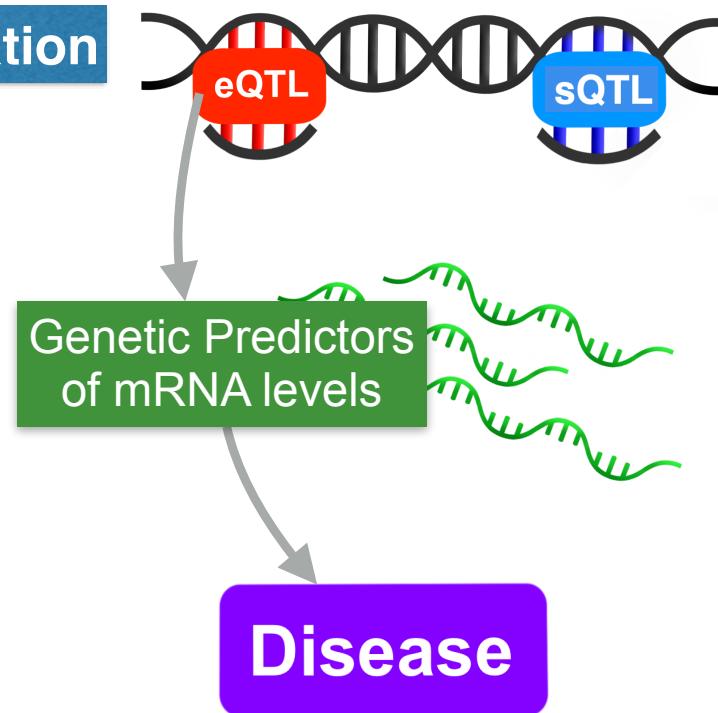


Dose-Response Effects Provide  
Strong Evidence of Causal Link  
between Gene Expression Traits  
and Complex Diseases

# Finding Causal Genes

# Association vs. Colocalization Methods

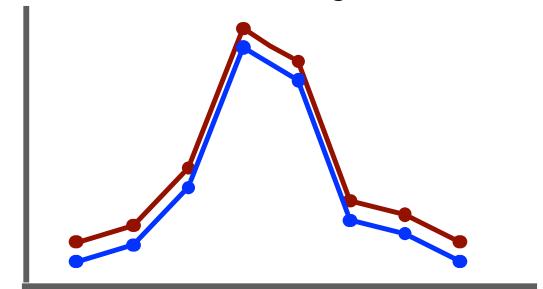
## Association



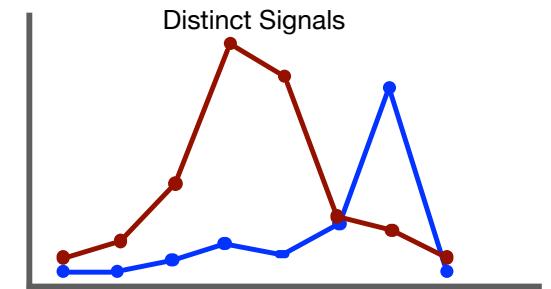
PrediXcan, SMR,  
FUSION

## Colocalization

Colocalized Signals



Distinct Signals



Coloc, Enloc, eCAVIAR, fastENLOC

# Association Approach

# Use Reference Transcriptome Data to Train Genetic Predictors

id	rs1	rs2	rs3	...	rsM
id1	0	2	1		0
id2	1	2	2		2
id3	2	1	1		1
:	:			:	:
:	:	:	:	:	:
:	:	:	:	:	:
:	:	:	:	:	:
idn	1	2	0		2

Genotype

49 tissues

id	q1	...	qm
id1	0.2		3.2
id2	2.3		4.1
id3	1.1		2.1
:	:		:
:	:		:
:	:		:
idn	2.2		2.1

Transcriptome



# Transcriptome Prediction Model Training

Reference Genotype and Transcriptome

id	rs1	rs2	rs3	...	rsM
id1	0	2	1		0
id2	1	2	2		2
id3	2	1	1		1
:	:				
⋮	⋮	⋮	⋮	⋮	⋮
idn	1	2	0		

Genotype

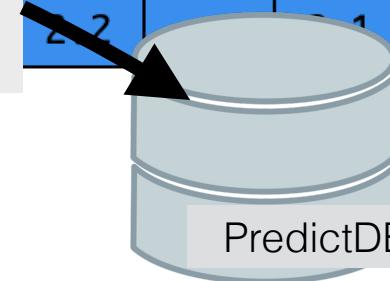
$$T = \sum_k w_k X_k + \epsilon$$

*GReX*

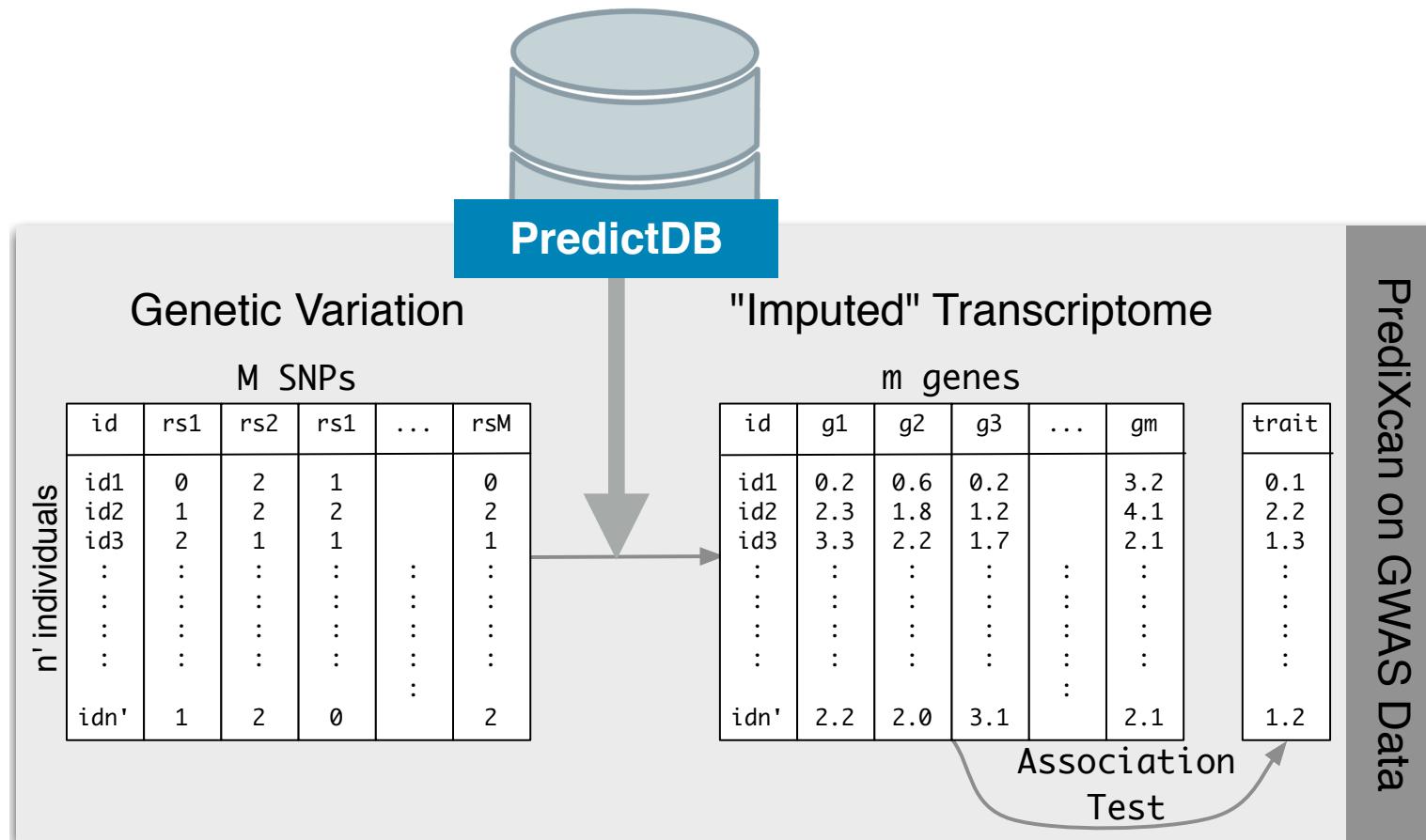
Weights stored in PredictDB

id	g1	...	gm
id1	0.2		3.2
id2	2.3		4.1
⋮	⋮	⋮	⋮
idn	2.2		4.1

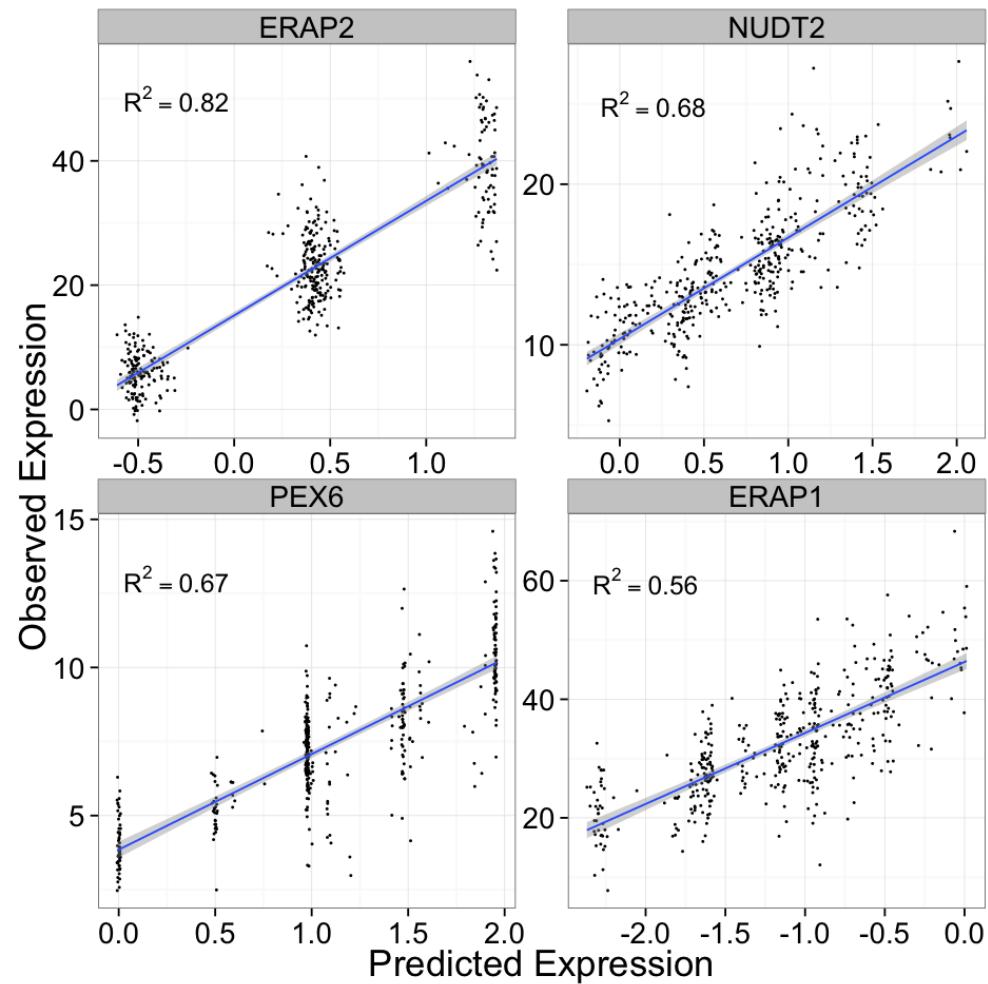
Transcriptome



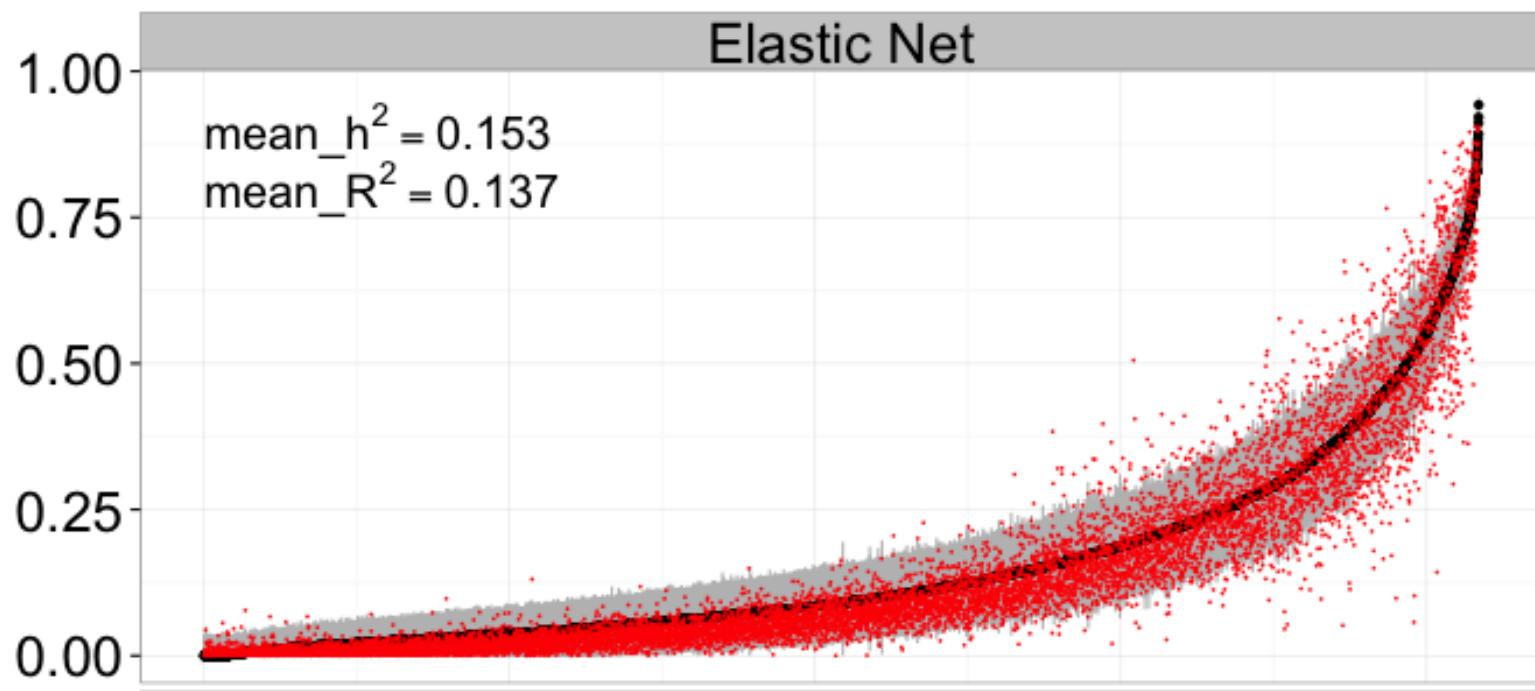
# Impose Transcriptome & Tests Association



## Examples of Well Predicted Genes



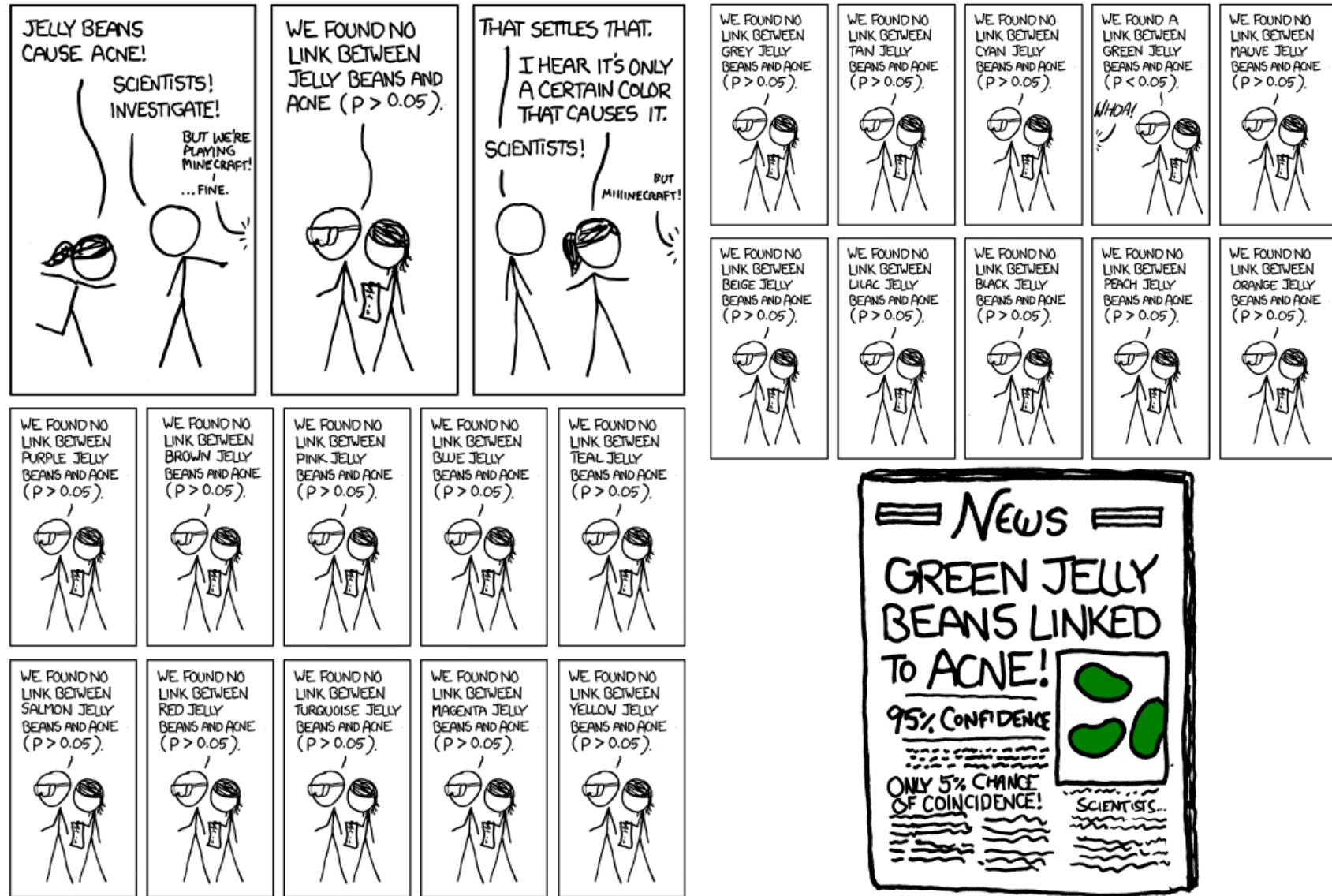
## Accuracy of Prediction Depends on Heritability

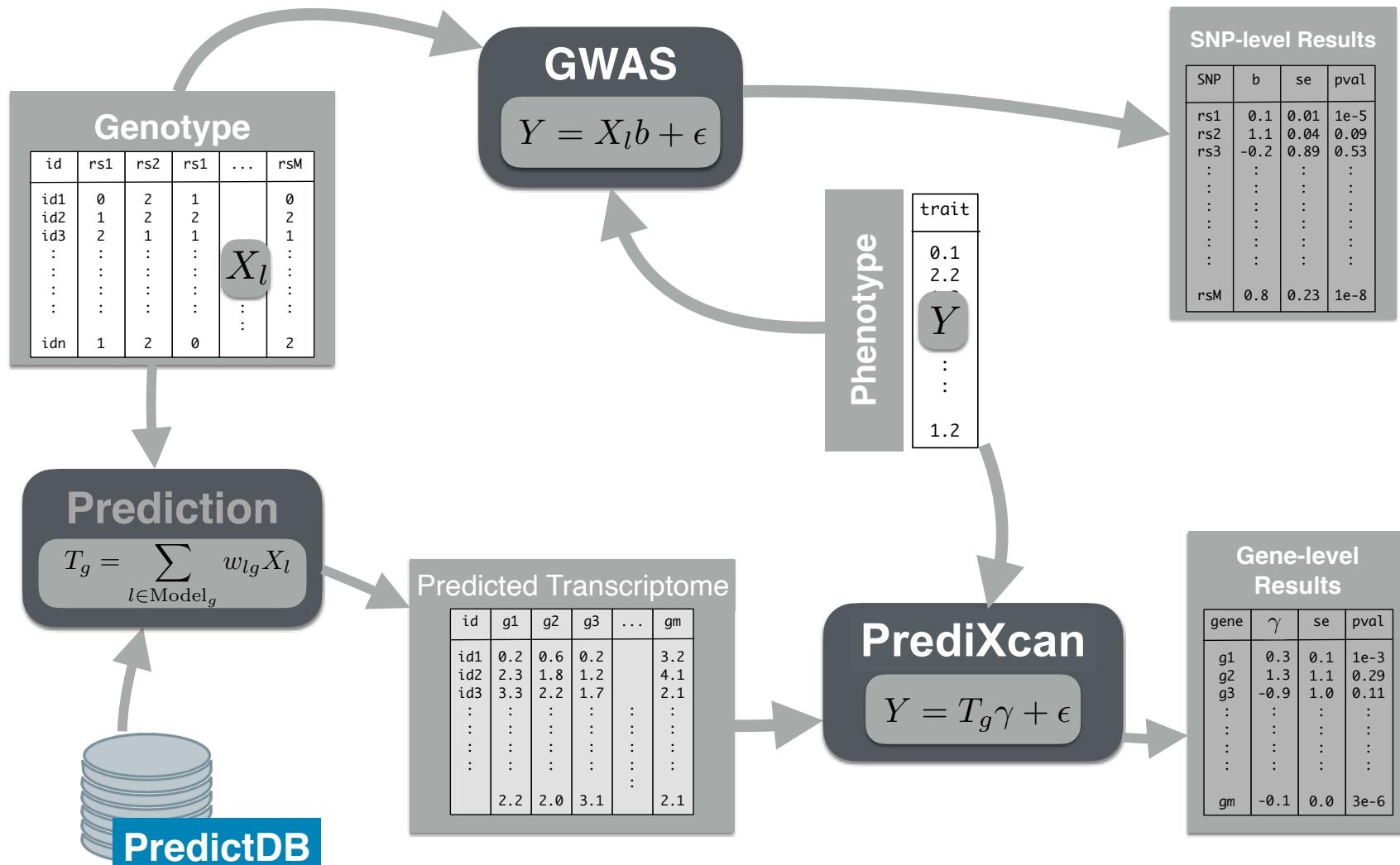


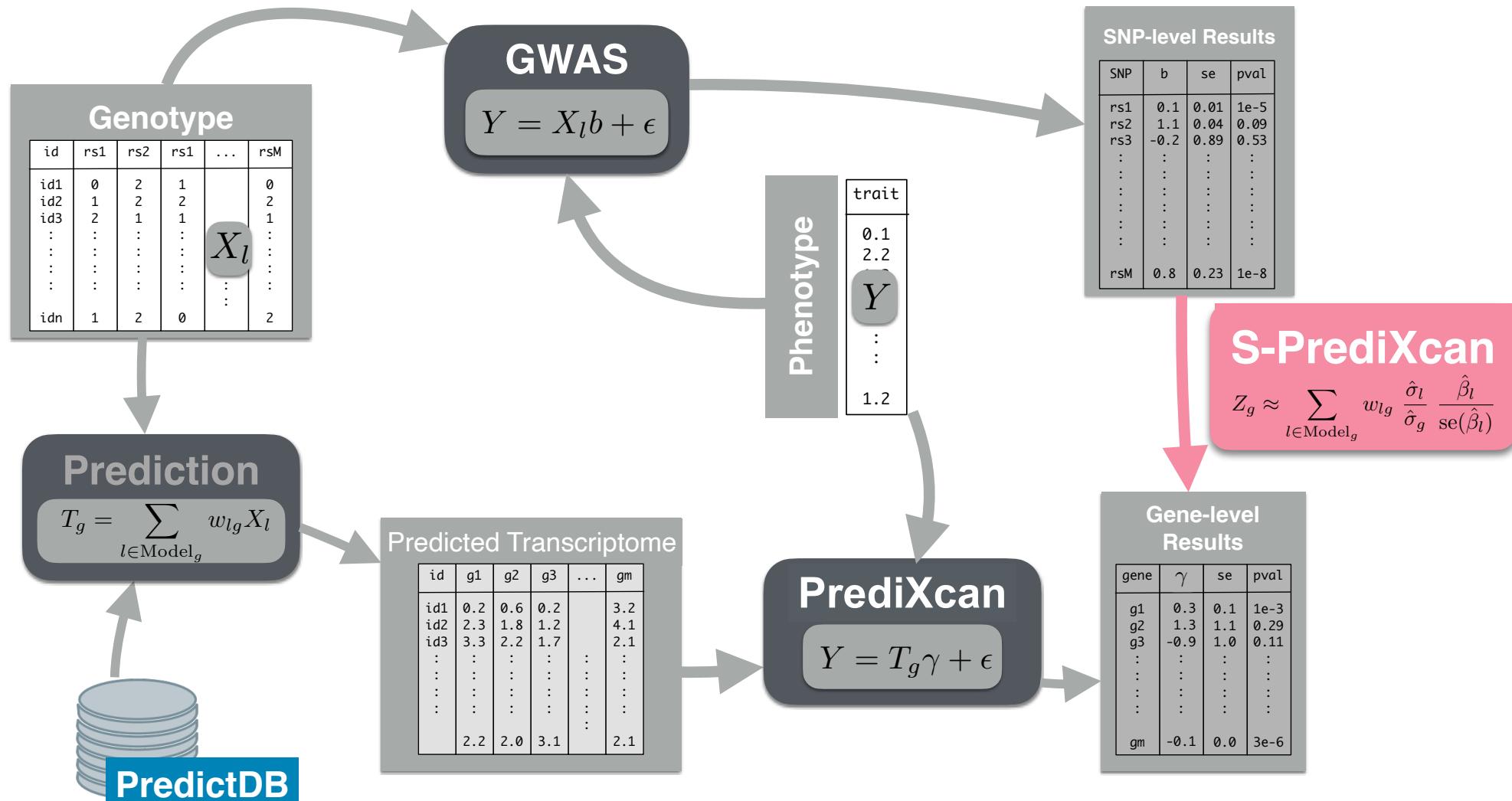
# Advantages of Gene Level Associations

- Reduced multiple testing burden (from 10e6 to 10e4)
- The function of genes are much better annotated than SNPs
- Validation in other model systems is possible
- Reverse causality issues is less of an issue
- It provides the direction of effects, i.e. whether up or down regulation of a gene increases the risk of a disease
- The candidate causal gene is a good target for drug development

On why we  
need to be  
more  
stringent  
when there  
are multiple  
testing



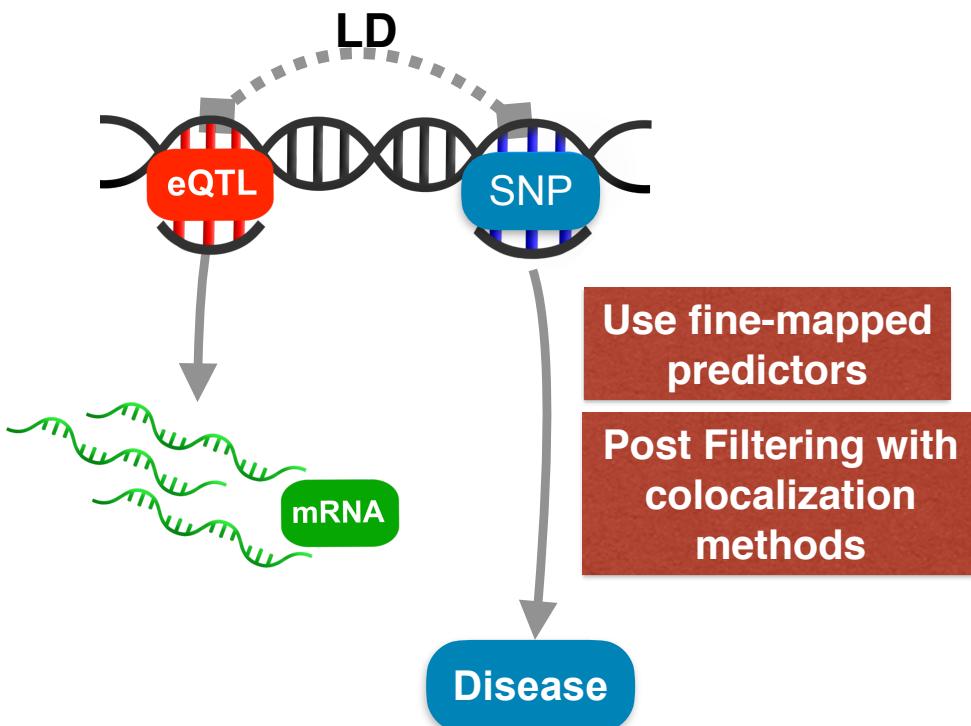




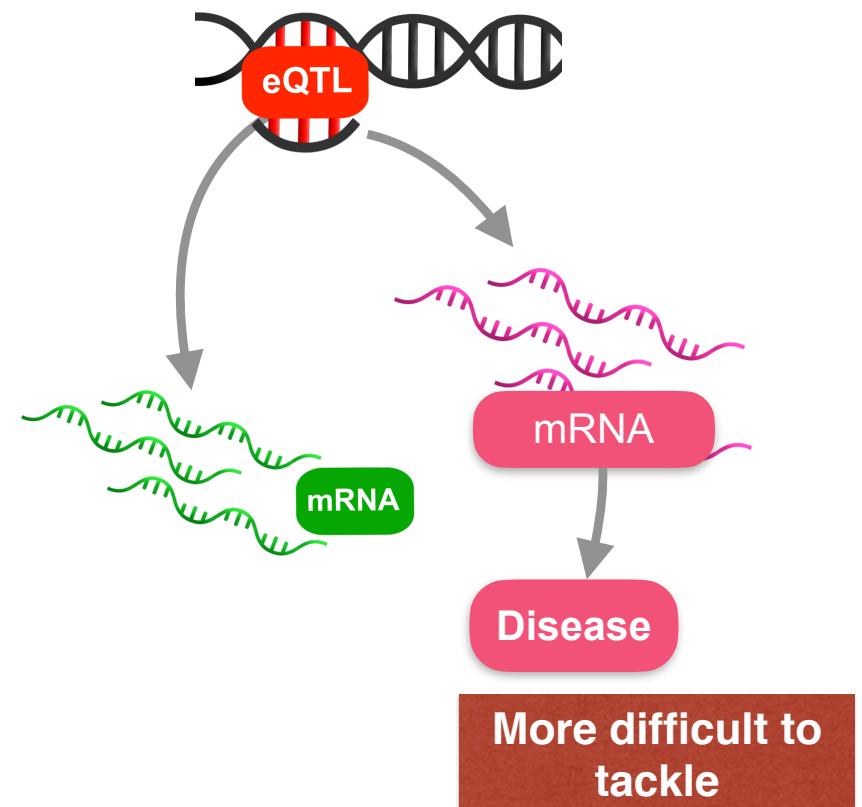
# Limitations of Current Association Methods

# Limitations of Association Methods

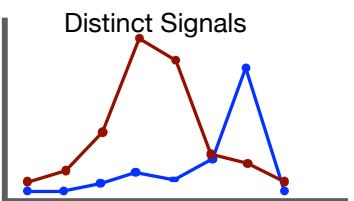
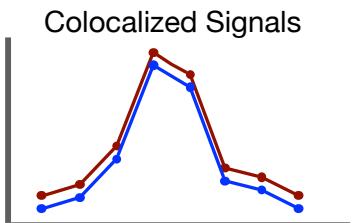
## LD Contamination



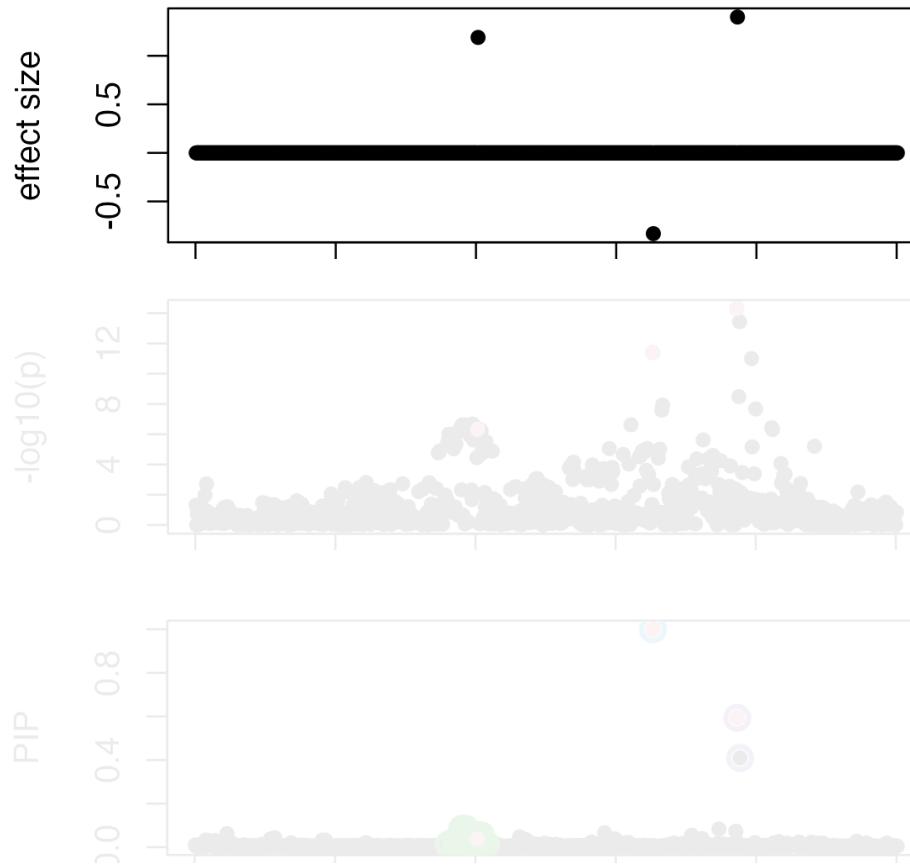
## Co-regulation



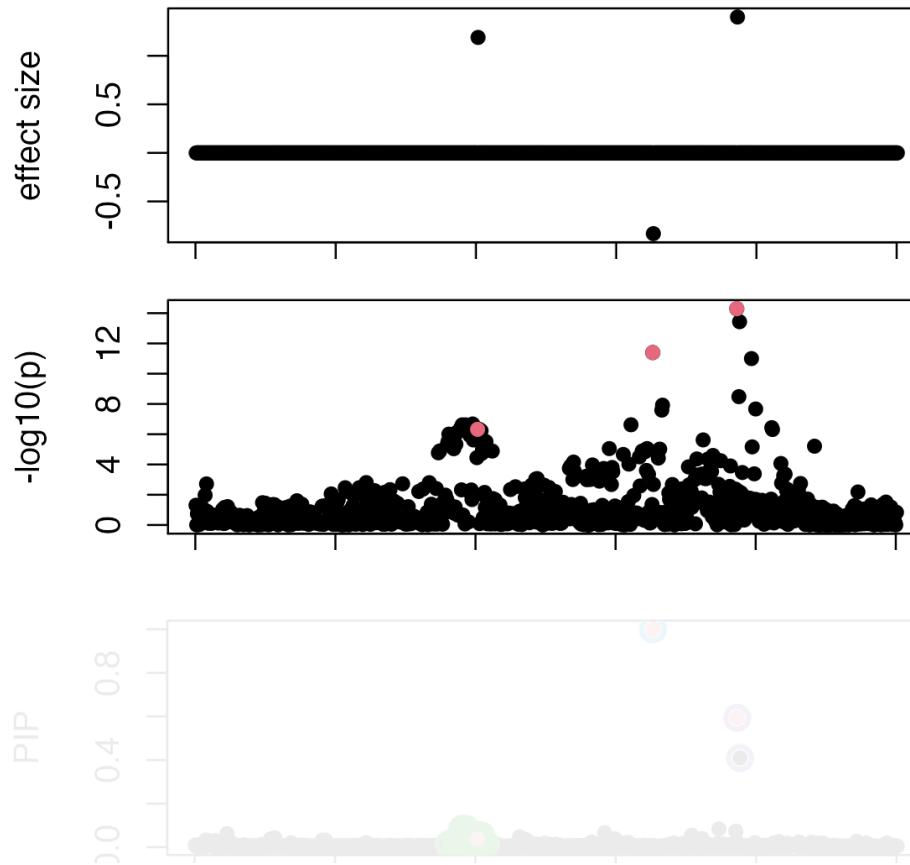
# Colocalization



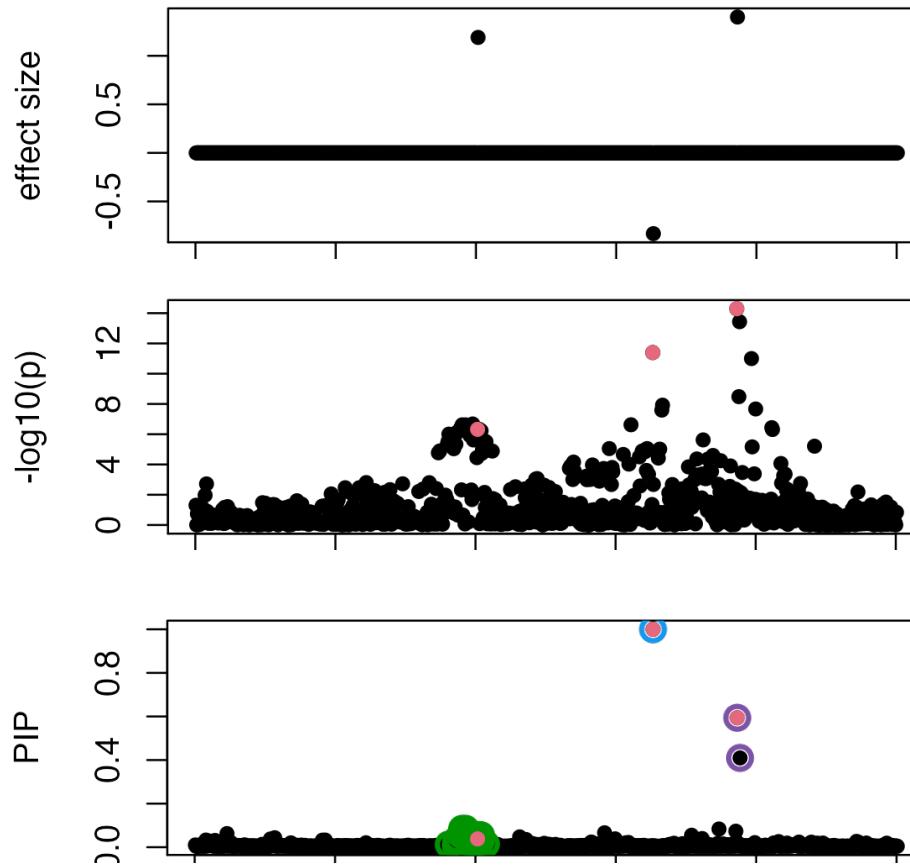
# Fine-Mapping Attempts to Find Causal Variants



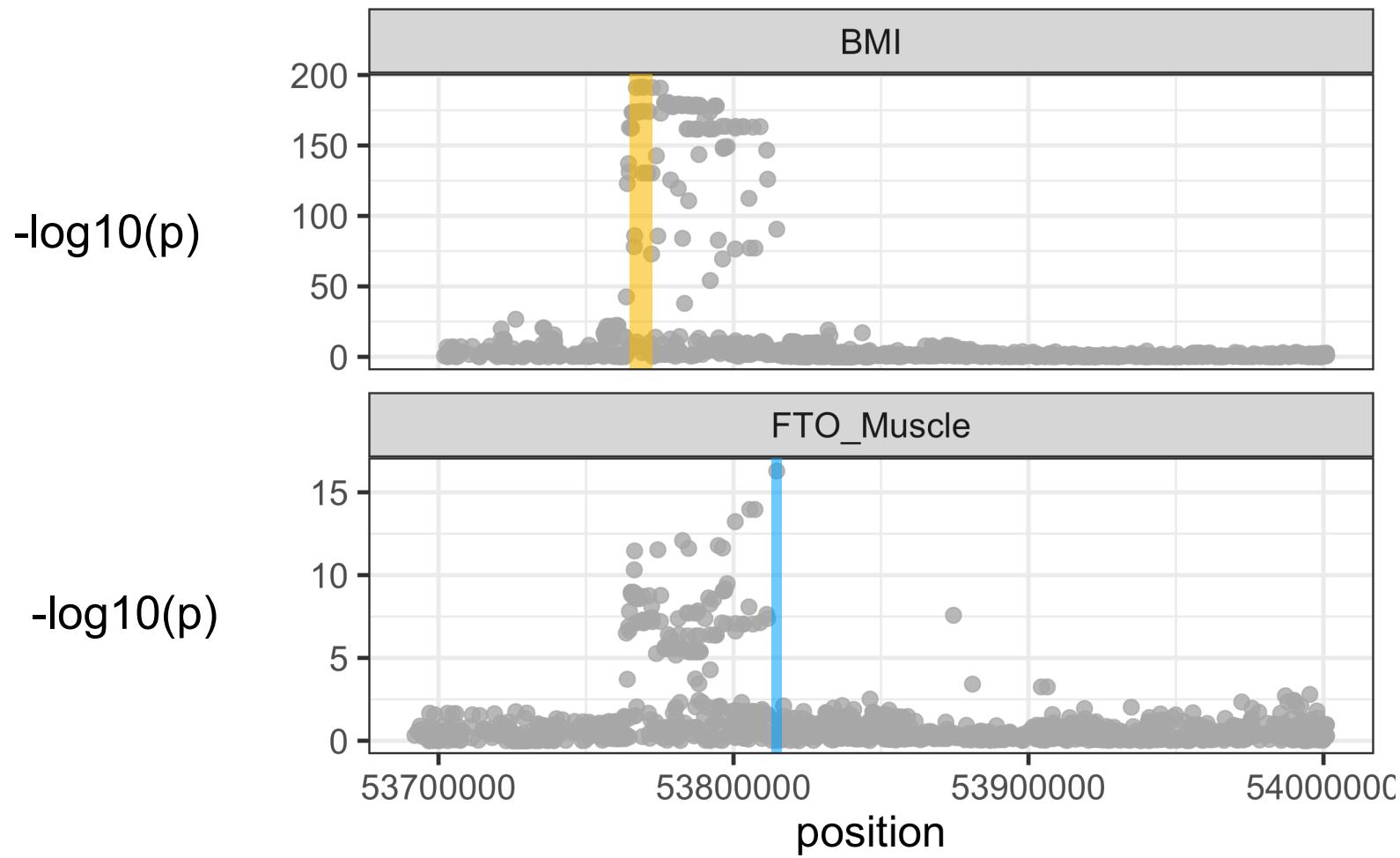
# Fine-Mapping Attempts to Find Causal Variants



# Fine-Mapping Attempts to Find Causal Variants



## Colocalization: Are causal variants = ?

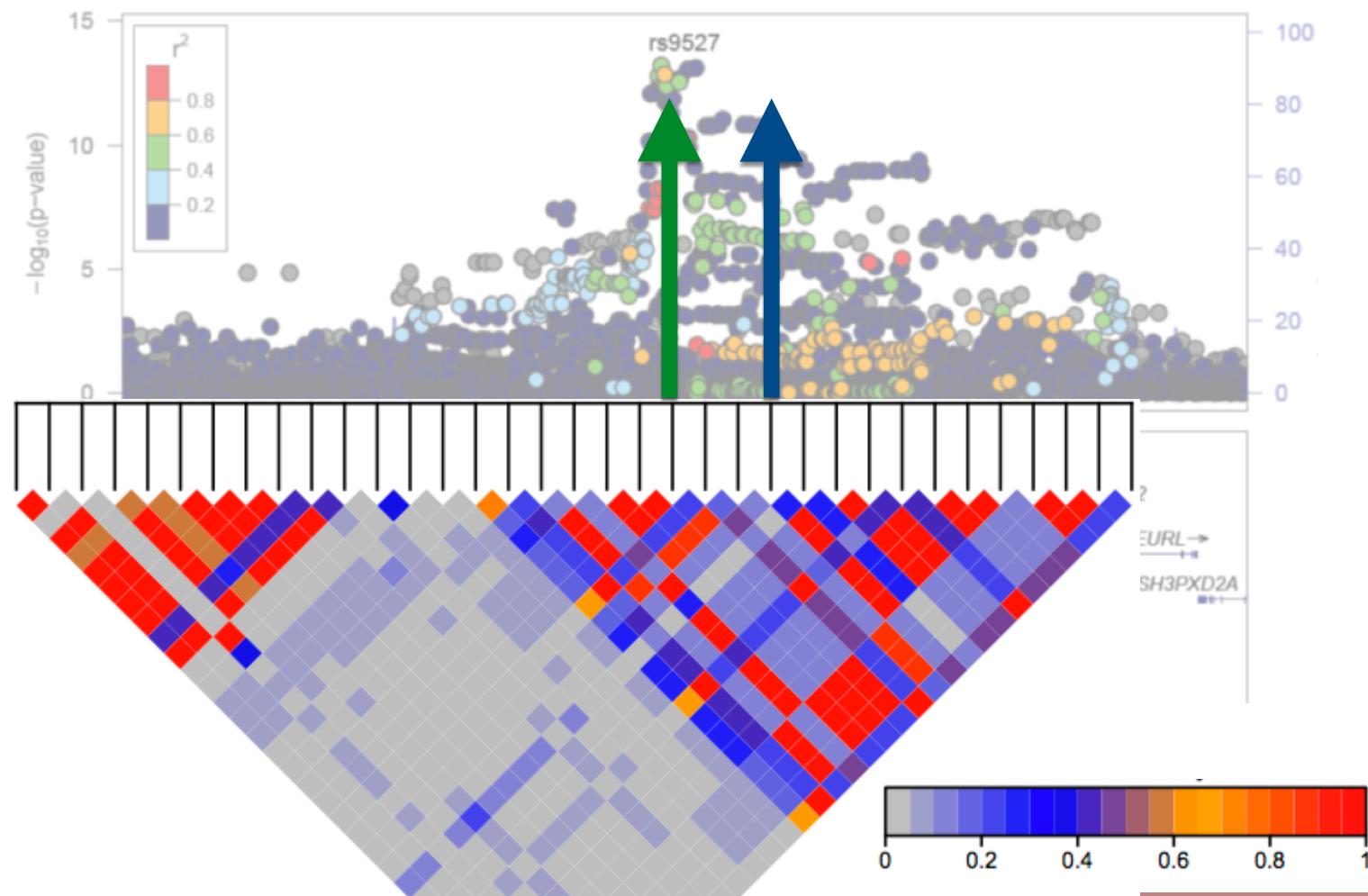


## Colocalization is not Causation

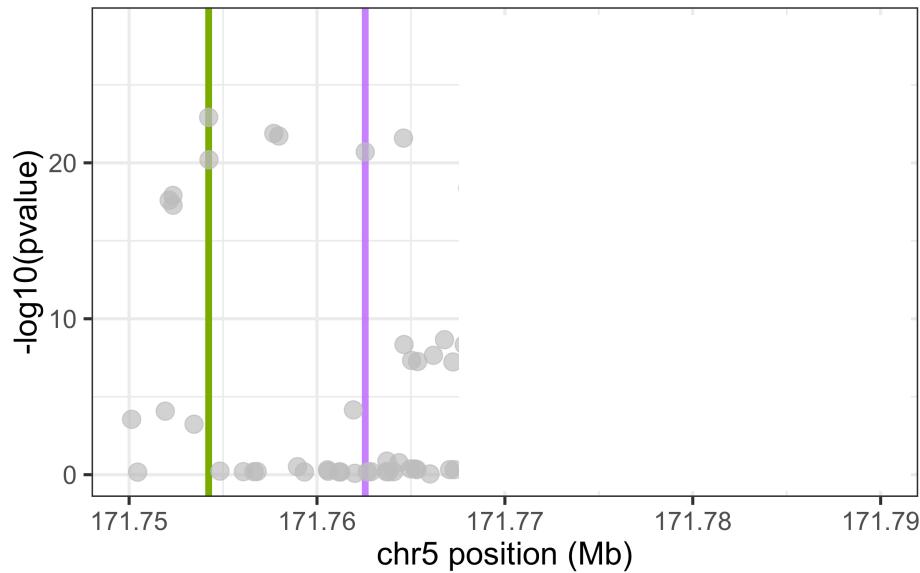


# Limitations of Colocalization

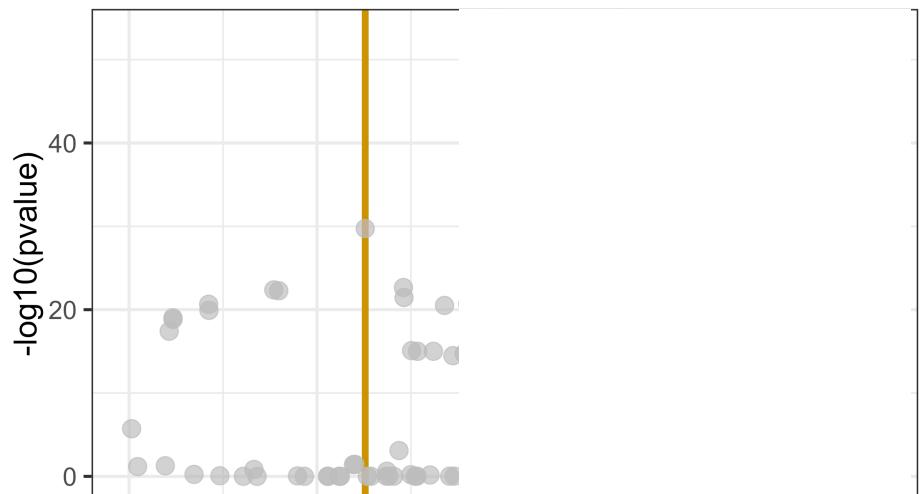
# Fine-Mapping Can Be Sensitive To LD Reference



GIANT Height



UK Height



In some loci,  
height does not  
colocalize with  
height!

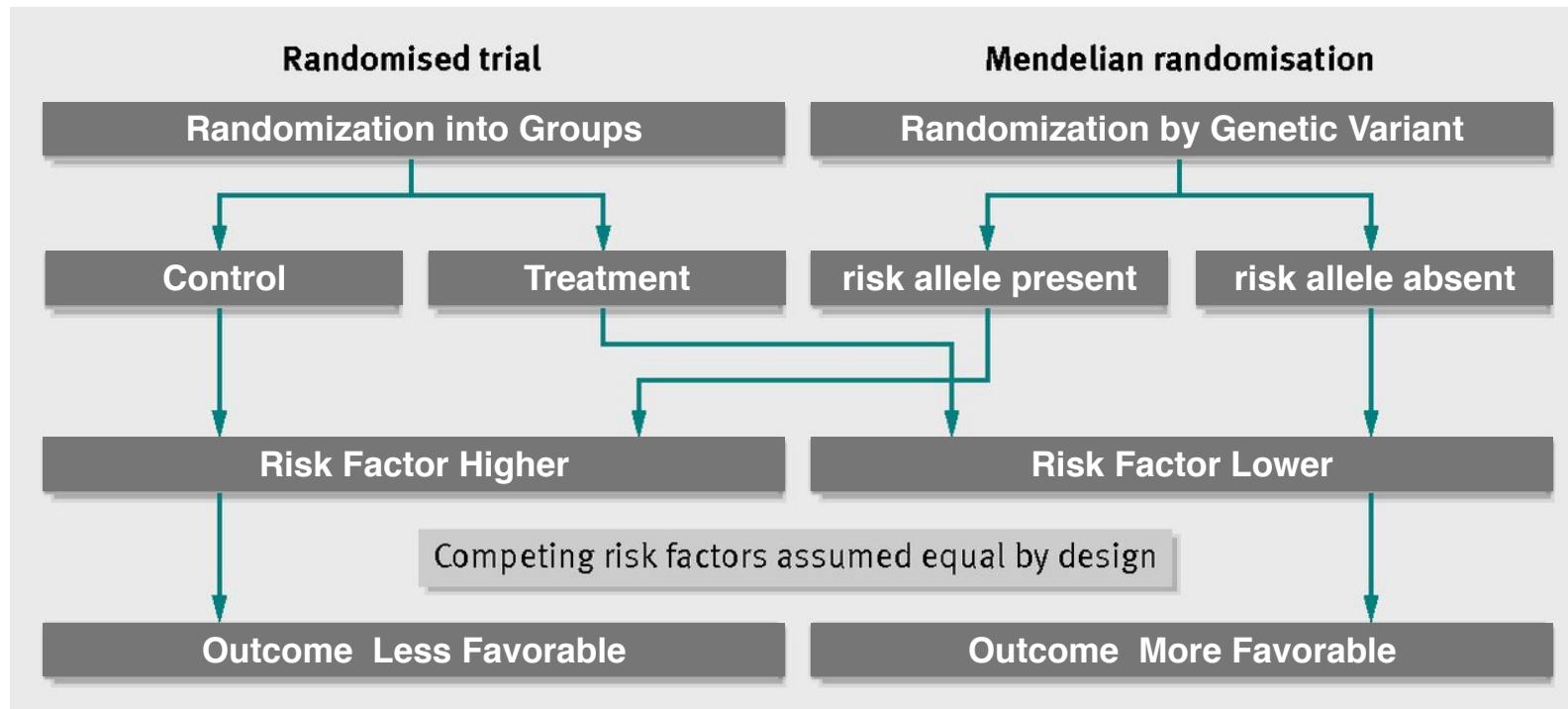
Fine-mapping results  
can be unreliable if we  
have a mismatch  
between  
GWAS summary results  
Reference LD

SusieR by Yanyu Liang

Colocalization can help  
finding causal genes but  
can be too conservative

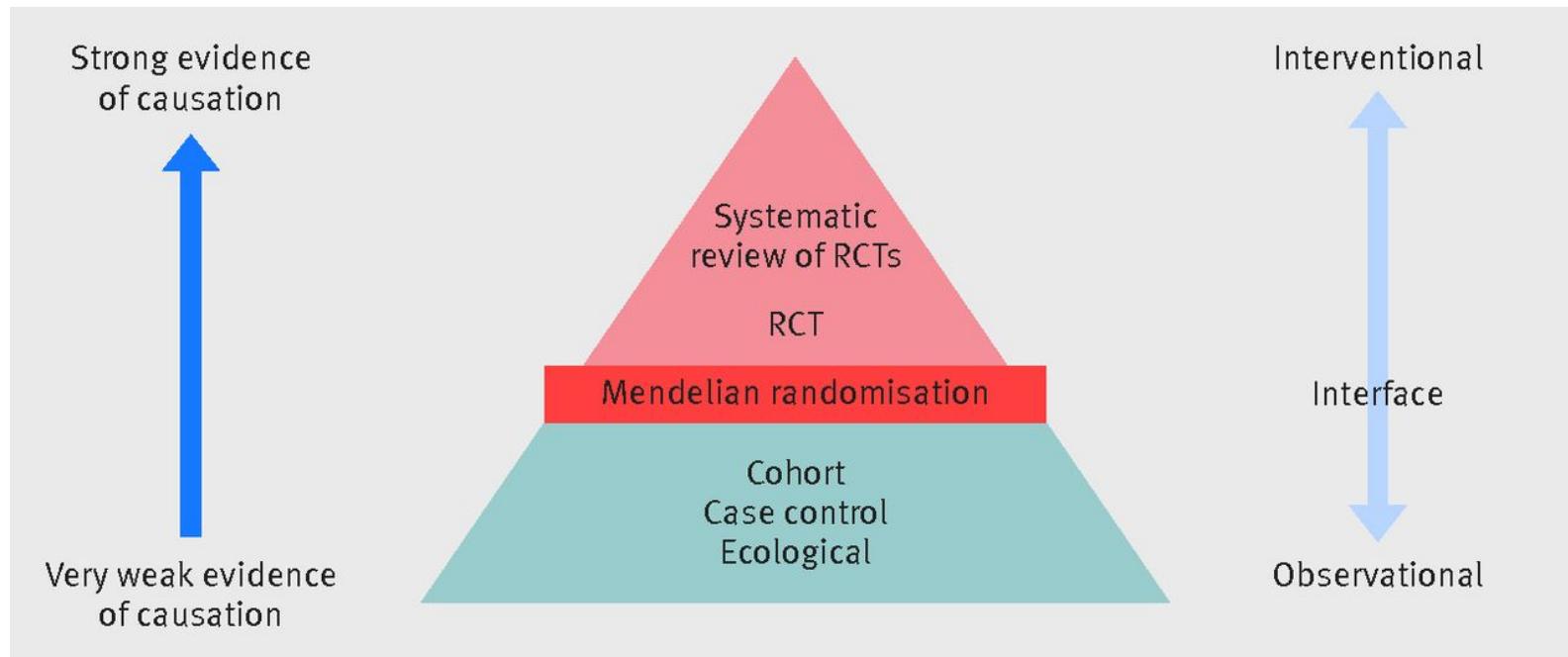
# Mendelian Randomization

# Randomized Trial vs. Mendelian Randomization



Burgess et al, Use of Mendelian randomisation to assess potential benefit of clinical intervention, BMJ 2012

# Evidence For Causal Link



RCT: Randomized Clinical Trials

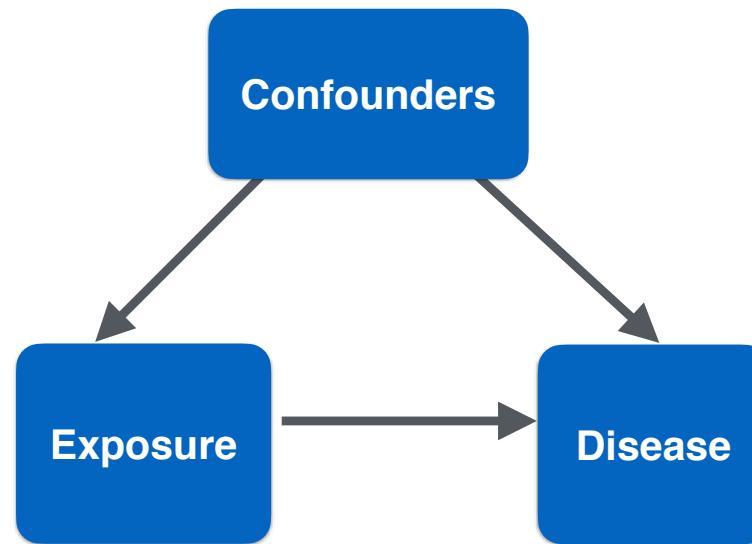
Neil M Davies, Michael V Holmes, George Davey Smith, Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians, BMJ 2018

# Mendelian Randomization Question



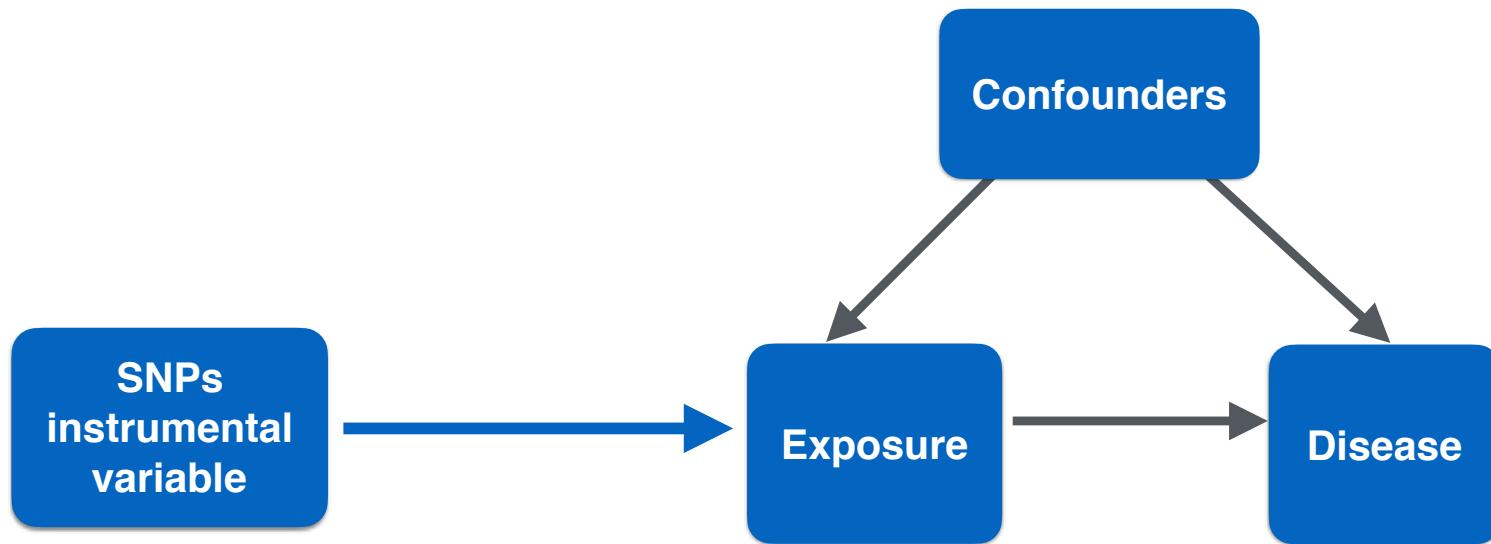
the goal is to test association between a modifiable exposure and disease. Smoking, HDL cholesterol levels, etc.

# Why We Need Mendelian Randomization?



Confounders may cause misleading associations

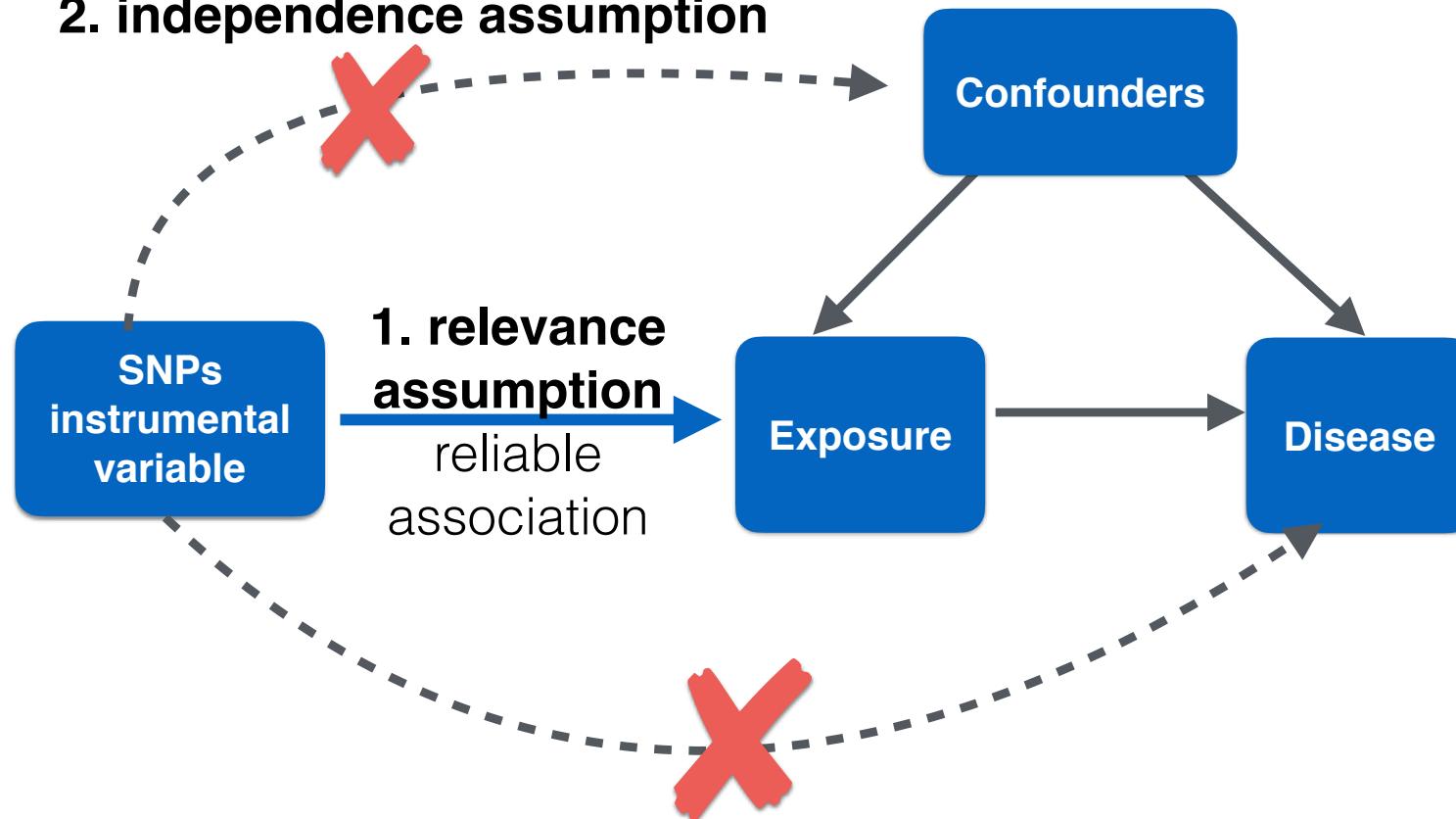
# Mendelian Randomization: Instrumental Variable



The idea is to use an "instrumental variable" without the confounding/noise

# Assumptions of Mendelian Randomization

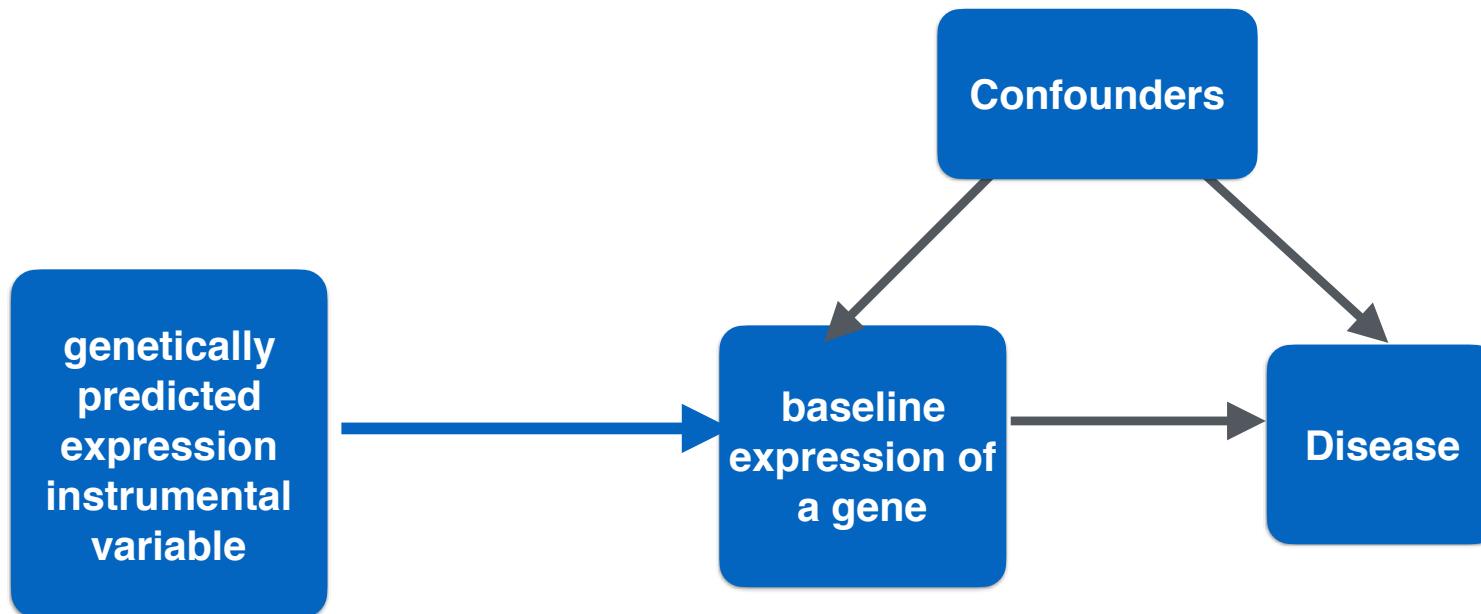
## 2. independence assumption



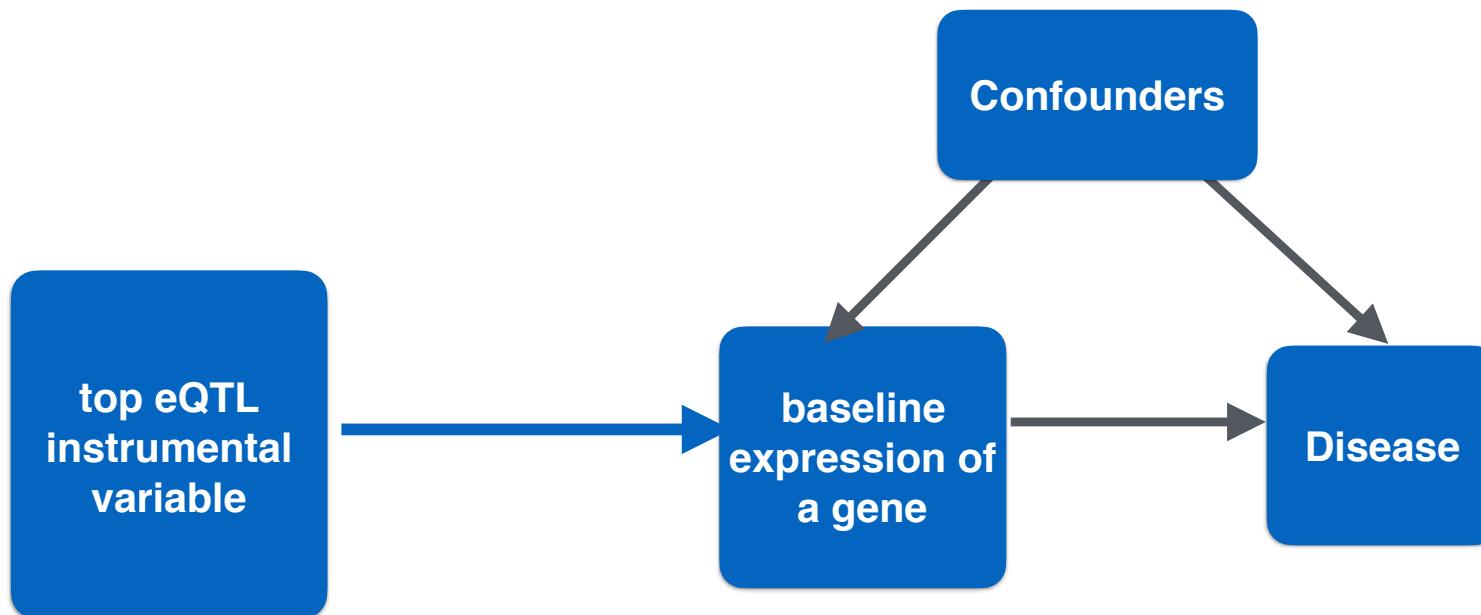
## 3. exclusion restriction assumption

no direct effect

# PrediXcan as a Mendelian Randomization Approach

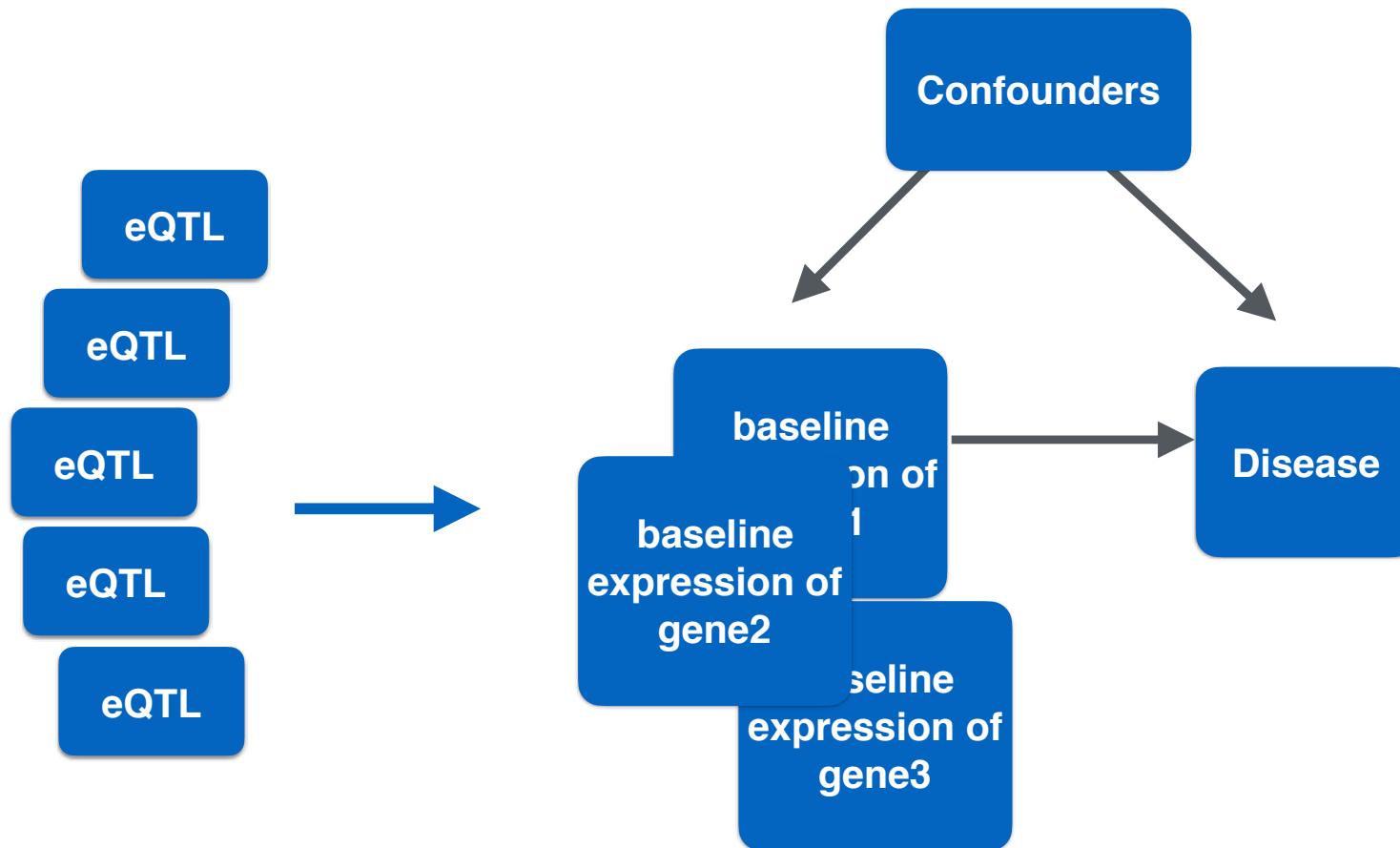


# Summary data–based Mendelian randomization (SMR)



Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. 48, 481–487 (2016).

# Transcriptome-wide Mendelian Randomization



Porcu, E., Rueger, S., Lepik, K., Agbessi, M., Ahsan, H., Alves, I., et al. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications*, 1–12. <http://doi.org/10.1038/s41467-019-10936-0>

# Other Association Methods

- TWAS/FUSION
  - TWAS originally used BLUP and BSLMM to predict expression
    - our analysis showed that these models with polygenic components may improve prediction of expression but can also lead more LD contamination (<http://dx.doi.org/10.1038/s41467-018-03621-1>)
  - FUSION uses BLUP, BSLMM, top eQTL, elastic net, LASSO for prediction and chooses the model that yields the best prediction
    - this sounds very reasonable but after extensive benchmarking we concluded that using more sparse prediction models leads to more reliable causal gene identification (<https://doi.org/10.1101/2020.03.19.997213>)
- UTMOST
  - uses group lasso to borrow information across multiple tissues and predict expression
  - our newest model based on prediction and smoothing across tissues (mashr models) have shown better performance for detecting causal genes than the utmost models (<https://doi.org/10.1101/2020.03.19.997213>)

# Other Association Methods

- FOCUS
  - similarly to TWMR focus tests multi-genes. The difference is that this tries to calculate the probability of causal role for each gene instead of assessing the significance of the association. It also allows for direct effects of genetic variants but uses predicted expression instead of individual genetic variants as instruments. The latter can be problematic if gene expression is not well predicted (rather common use case)

# Other Colocalization Methods

- COLOC
- ENLOC
- eCAVIAR

# References

- PrediXcan
  - Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. Nature Publishing Group, 47(9), 1091–1098. <http://doi.org/10.1038/ng.3367>
- S-PrediXcan
  - Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications*, 9(1), 1–20. <http://doi.org/10.1038/s41467-018-03621-1>
- GTEx GWAS subgroup
  - Barbeira, A. N., Bonazzola, R., Gamazon, E. R., Liang, Y., Park, Y., GTEx Consortium, et al. (n.d.). Exploiting the GTEx resources to decipher the mechanisms at GWAS loci. *bioRxiv*. <http://doi.org/10.1101/814350>
- fastENLOC
  - Pividori, M., Rajagopal, P. S., Barbeira, A. N., BioRxiv, Y. L., & Im, H. K. (n.d.). PhenomeXcan: Mapping the genome to the phenotype through the transcriptome. *Biorxiv.org*. <http://doi.org/10.1101/833210>
- MultiXcan
  - Barbeira, A. N., Pividori, M., Zheng, J., Wheeler, H. E., Nicolae, D. L., & Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet*, 15(1), e1007889–20. <http://doi.org/10.1371/journal.pgen.1007889>
- Finemapping improves causal gene detection
  - Barbeira, A. N., Liang, Y., Bonazzola, R., Wang, G., Wheeler, H. E., Melia, O. J., et al. (2020). Fine-mapping and QTL tissue-sharing information improve causal gene identification and transcriptome prediction performance, 33(1), 1–28. <http://doi.org/10.1101/2020.03.19.997213>
- GTEx v8 main paper
  - Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., et al. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, 7, 1860. <http://doi.org/10.1101/787903>
- TWMR
  - Porcu, E., Rueger, S., Lepik, K., Agbessi, M., Ahsan, H., Alves, I., et al. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications*, 10(1), 1–12. <http://doi.org/10.1038/s41467-019-10936-0>

# Recommendations for Post GWAS Analysis

- Start with an association method to determine a list of candidate causal genes (e.g. PrediXcan, FUSION, UTMOST)
  - use methods that aggregate across multiple tissues and contexts
  - you lose power when using a single tissue (even if it's the most relevant)
- Use colocalization (coloc, enloc, eCAVIAR) to filter out LD contamination
  - but keep in mind that colocalization can throw away real signals
- Higher resolution data (single cell, context, space, and time) and new methods that integrate them are emerging: incorporate them as additional sources of evidence
  - Our experience showed time and again that combining multiple sources of evidence leads the best causal gene identification

# Hands On Exercises

# Instructions to get started

- Up to date information on data can be found here
  - [https://hakyimlab.github.io/QGT-Columbia-HKI-repo/2021\\_analysis\\_plan.html](https://hakyimlab.github.io/QGT-Columbia-HKI-repo/2021_analysis_plan.html)