

Transcriptome-based Methods to Assign Function to GWAS Loci

Hae Kyung Im, PhD



THE UNIVERSITY OF
CHICAGO

Quantitative Genomic Training
June 12, 2020

The Promise of the Human Genome Project



During the announcement of the first draft of the human genome in 2000 Bill Clinton said that

"This profound new knowledge [...] will revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases.

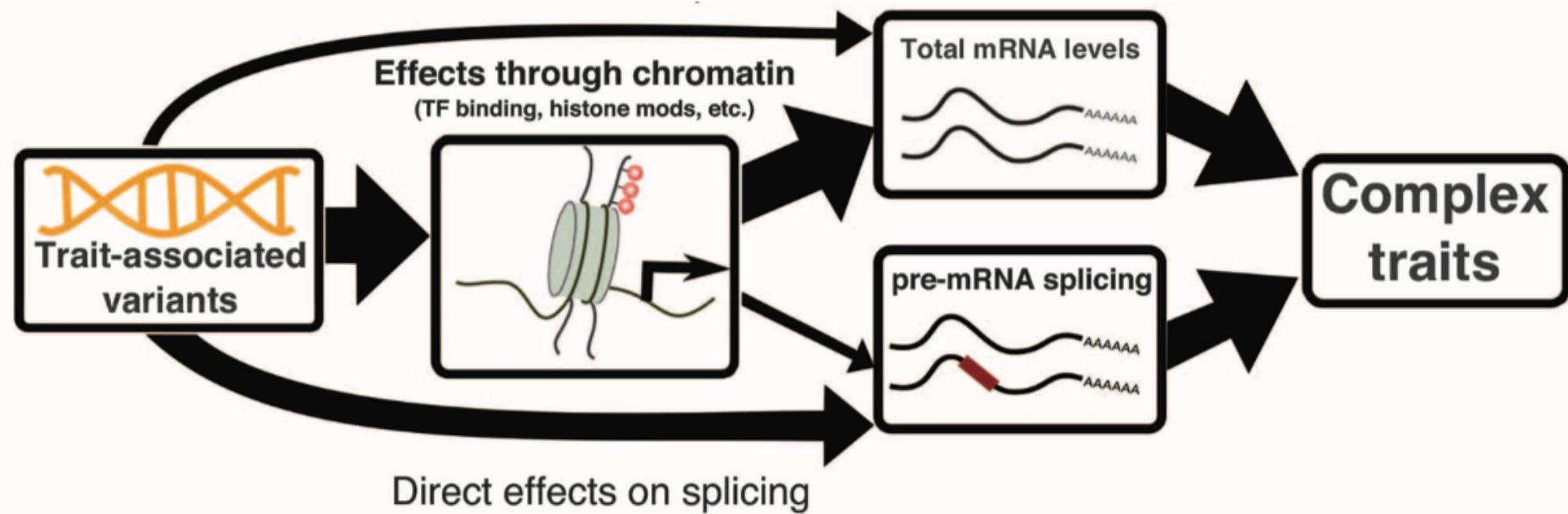
[...] it is now conceivable that our children's children will know the term 'cancer' only as a constellation of stars."

Why are we not there yet

- Prediction of the future is hard
- Polygenicity/Ominigenicity
 - Genetic architecture of common diseases and cancers is much more complex than anticipated
- Unknown function
 - Most loci associated with common diseases do not change protein coding

Most GWAS catalog
variants are non-coding

GWAS Variants Alter Complex Traits via Regulation



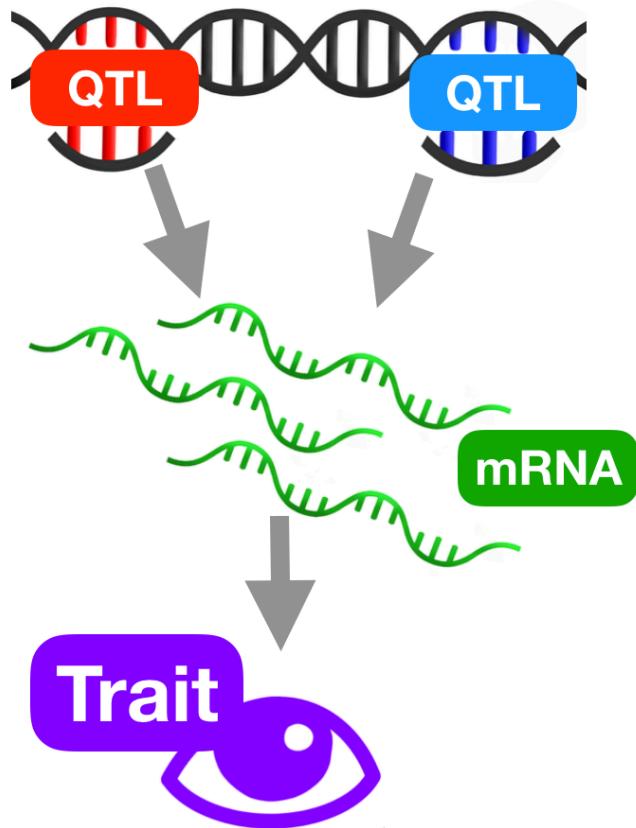
**RNA splicing is a primary link
between genetic variation and disease**

Yang I. Li,¹ Bryce van de Geijn,² Anil Raj,¹ David A. Knowles,^{3,4} Allegra A. Petti,⁵
David Golan,¹ Yoav Gilad,^{2,*} Jonathan K. Pritchard^{1,6,7*}

2016

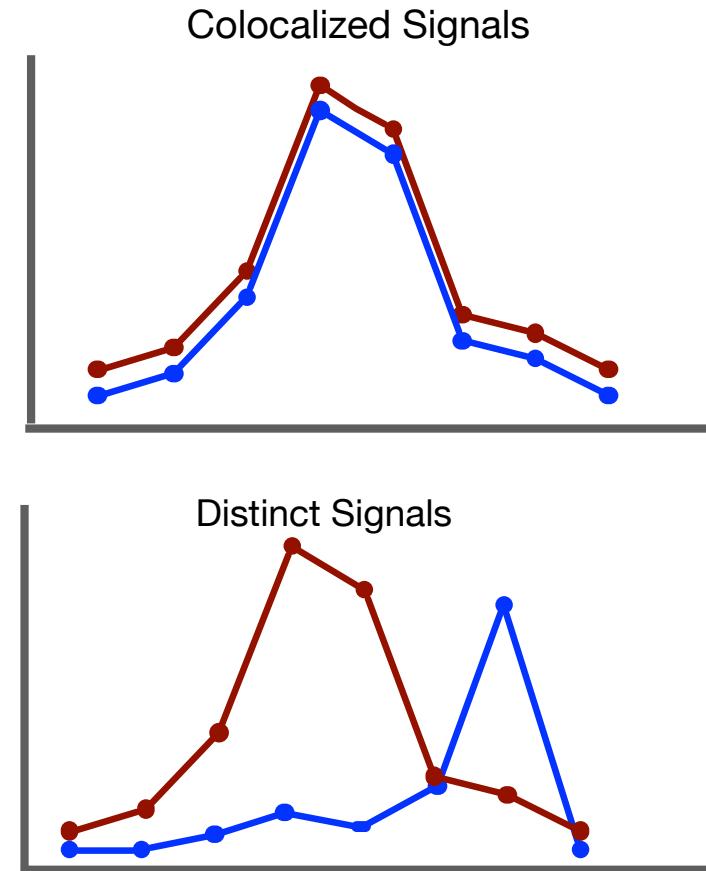
Association vs. Colocalization Methods

Association



PrediXcan, SMR,
FUSION (formerly TWAS)

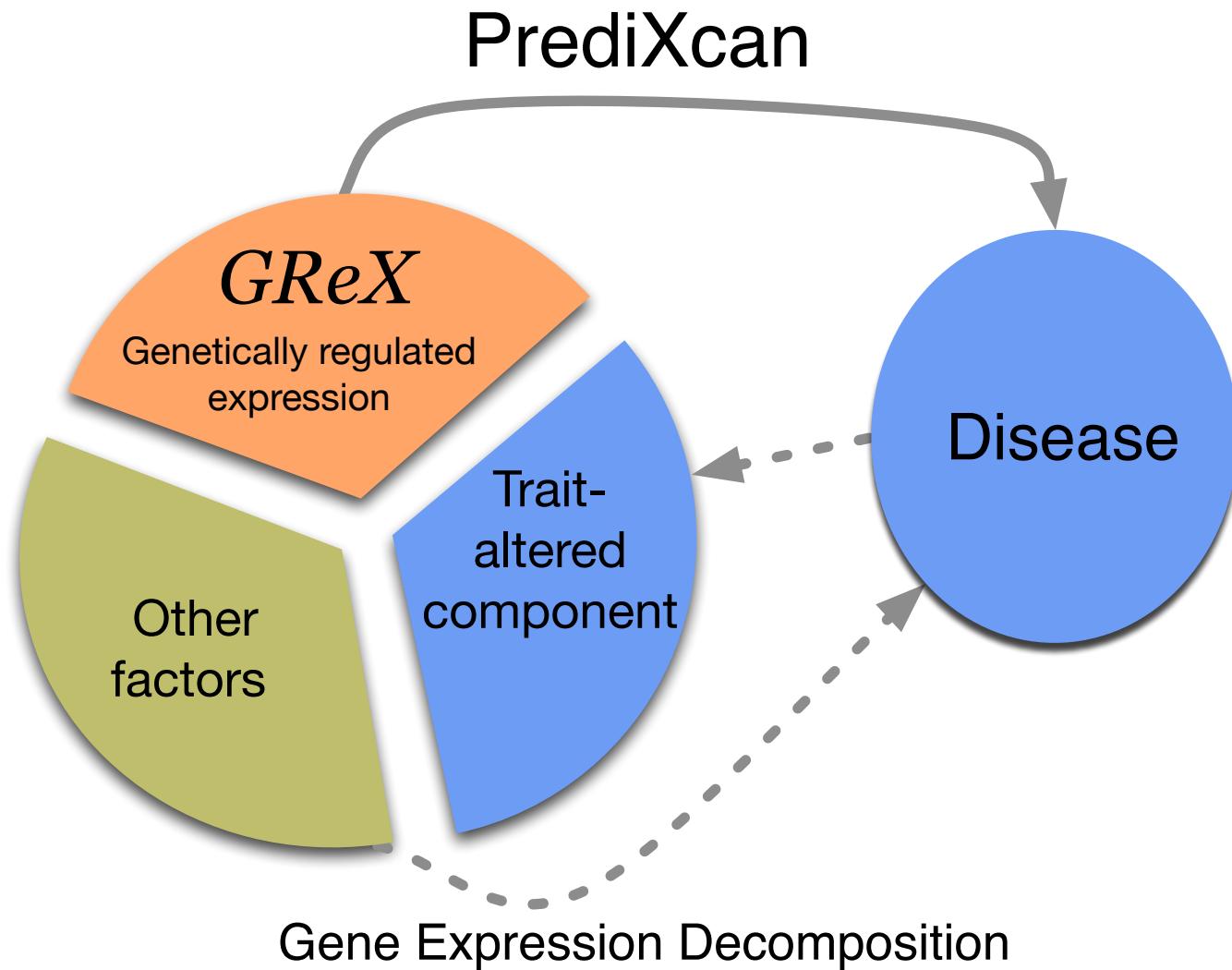
Colocalization



Coloc, Enloc, eCAVIAR, fastENLOC

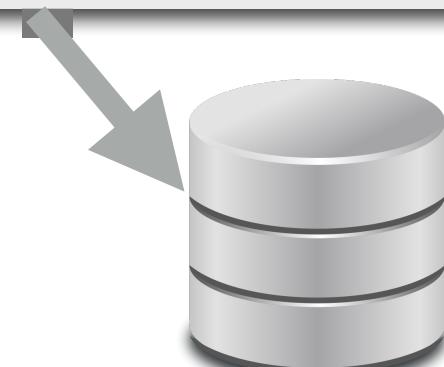
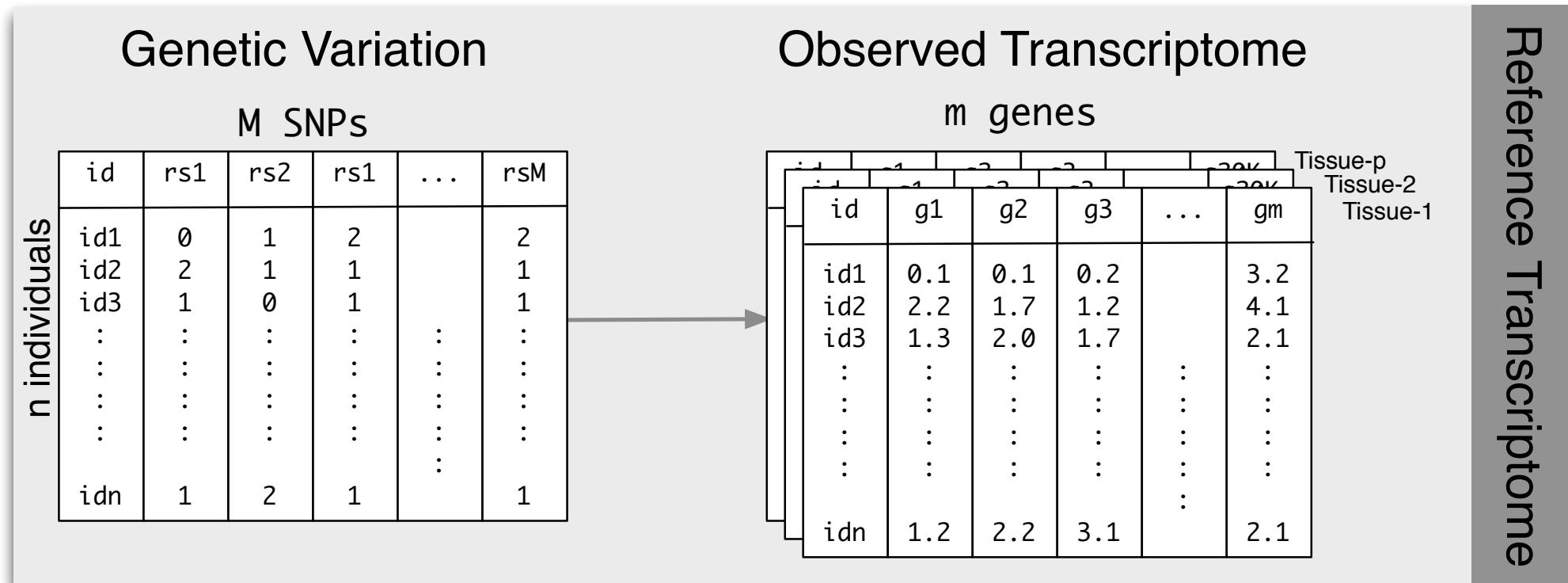
Association Approach

PrediXcan Uses Association Between GReX and Disease To Identify Causal Genes



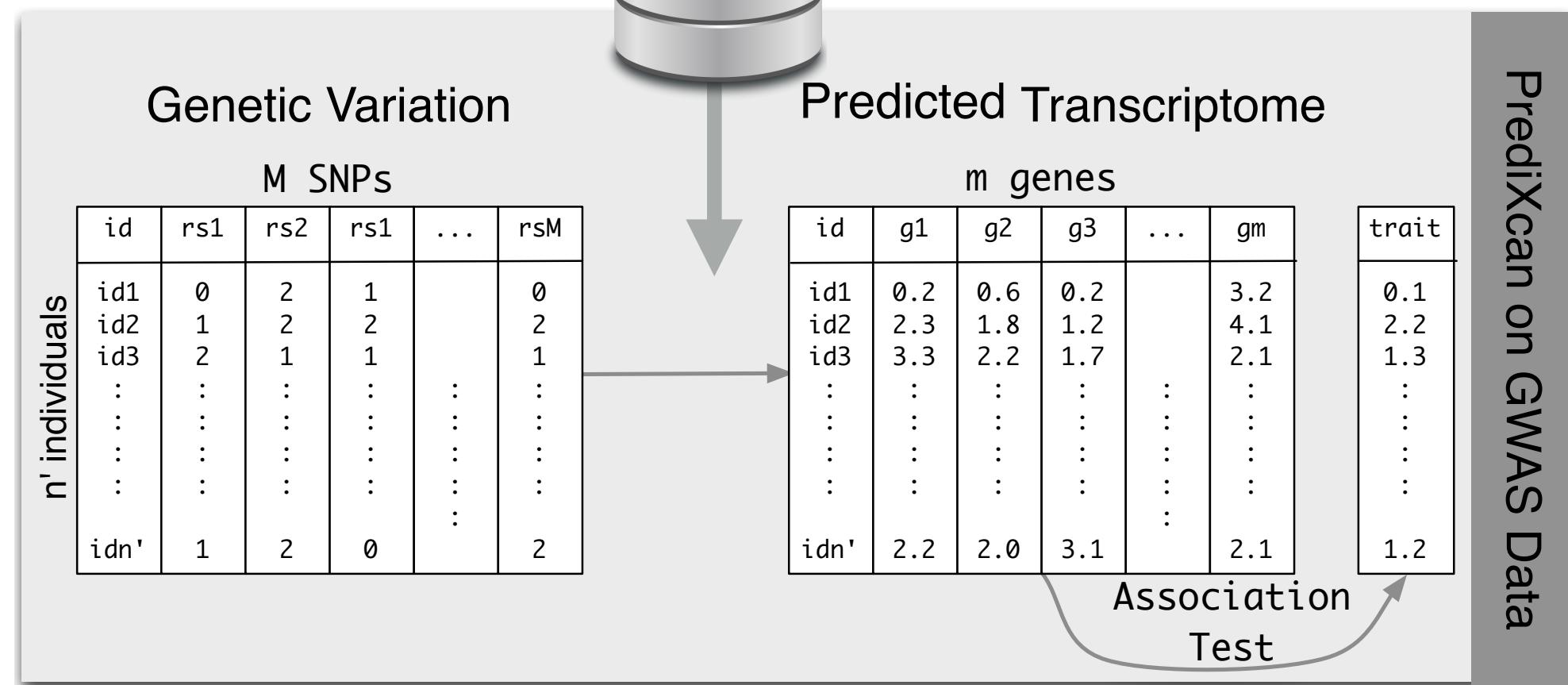
Gamazon, Wheeler, Shah et al 2015 Nature Genetics

PrediXcan Trains Predictors using Reference Transcriptome Data

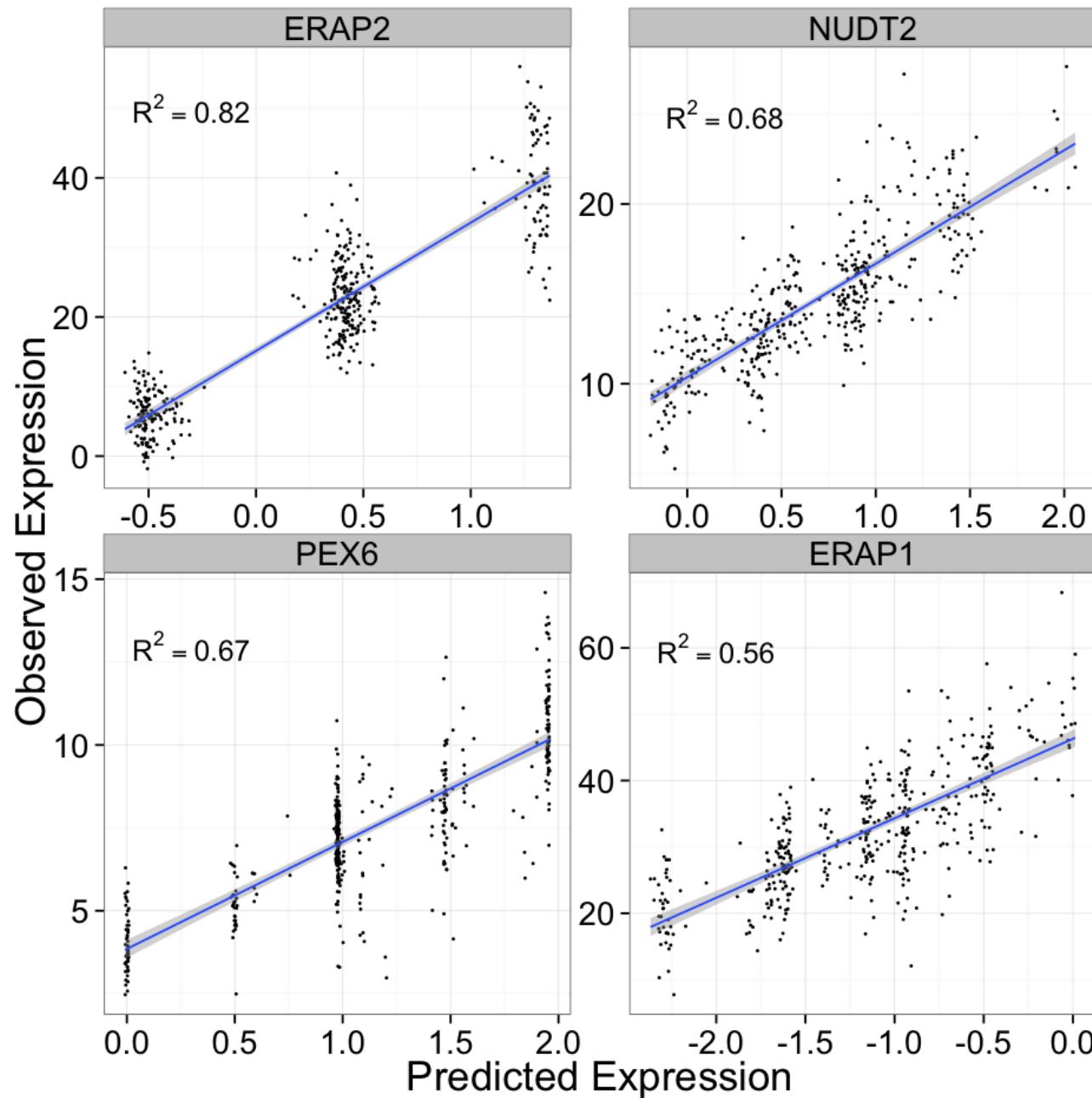


parameters
shared in
PredictDB.org

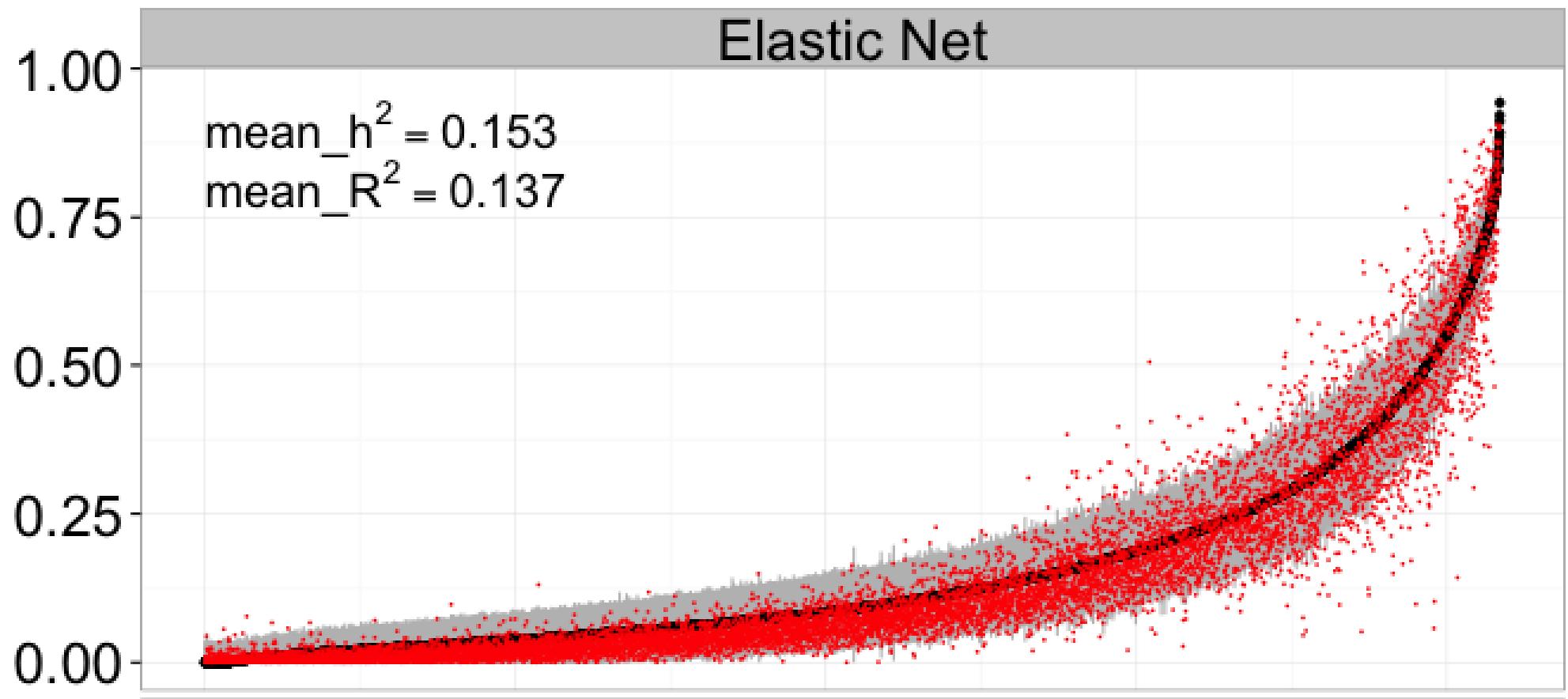
PrediXcan Predicts the Transcriptome & Tests Assoc.



Examples of Well Predicted Genes



Accuracy of Prediction Depends on Heritability



Advantages of Gene Level Associations

- Reduced multiple testing (from 10e6 to 10e4)
- Genes functions are much better annotated
- Validation in other model systems is possible
- Direction of effects can inform on protective or deleterious effects of gene knock down
- Prioritization of drug targets is more straightforward

Genotype + Phenotype

id	rs1	rs2	rs1	...	rsM	trait
id1	0	2	1		0	0.1
id2	1	2	2		2	2.2
id3	2	1	1		1	1.3
:	:	.	:		:	.
:	:		:		:	:
:	:		:		:	:
idn	1	2	0		2	1.2

GWAS

$$Y = X_l b + \epsilon$$

SNP-level Results

SNP	b	se	pval
rs1	0.1	0.01	1e-5
rs2	1.1	0.04	0.09
rs3	-0.2	0.89	0.53
:	:	:	:
:	:	:	:
:	:	:	:
:	:	:	:
rsM	0.8	0.23	1e-8

PrediXcan

id	g1	g2	g3	...	gm
id1	0.2	0.6	0.2		3.2
id2	2.3	1.8	1.2		4.1
id3	2.2	2.2	1.7		3.1

$$T_g = \sum_{l \in \text{Model}_g} w_{lg} X_l$$

Predicted Transcriptome

$$Y = T_g \gamma + \epsilon$$

S-PrediXcan

$$Z_g \approx \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\text{se}(\hat{\beta}_l)}$$

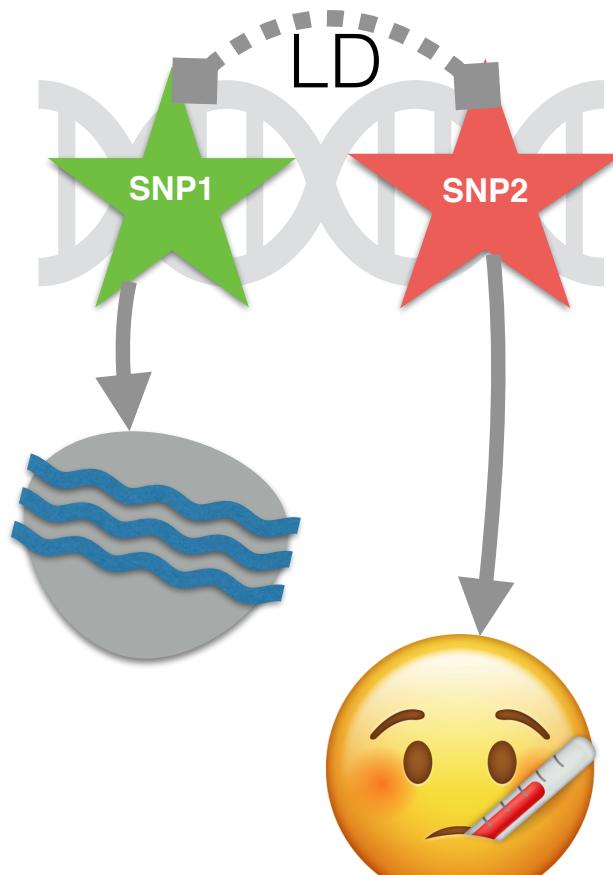
Gene-level Results

gene	γ	se	pval
g1	0.3	0.1	1e-3
g2	1.3	1.1	0.29
g3	-0.9	1.0	0.11
:	:	:	:
:	:	:	:
:	:	:	:
gm	-0.1	0.0	3e-6

Limitations of Current Association Methods

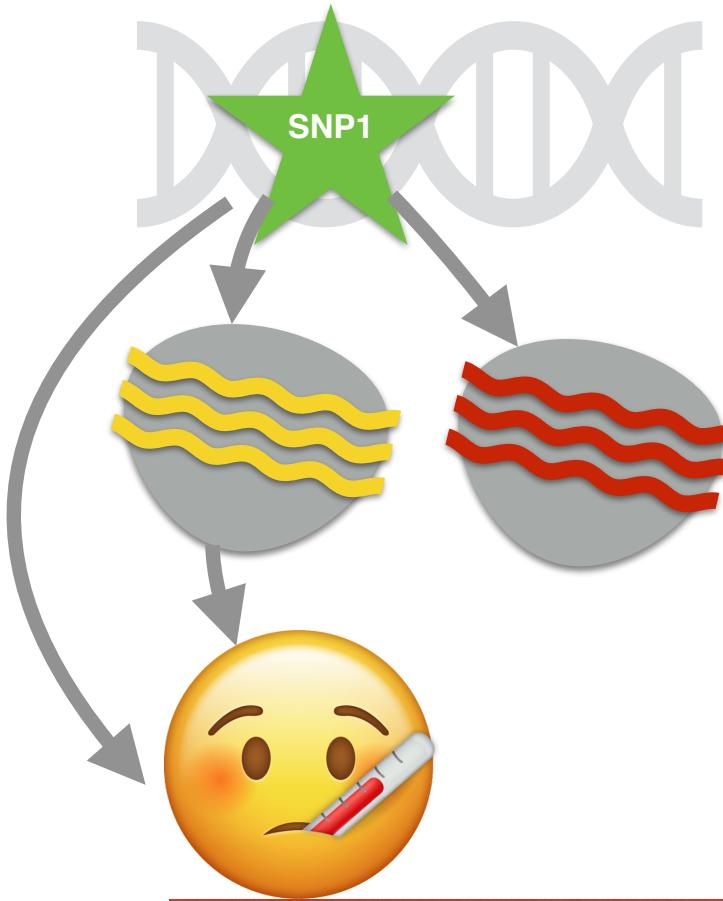
Limitations of Downstream Association Methods

LD Contamination



Post Filtering with
colocalization methods

Co-Regulation

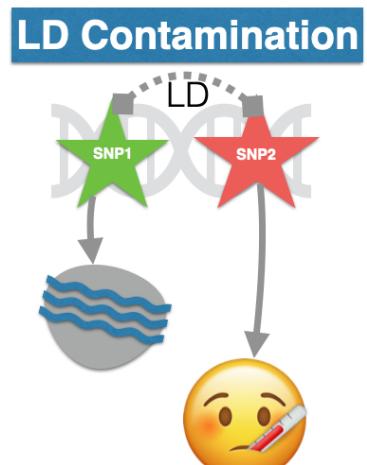


More difficult to tackle

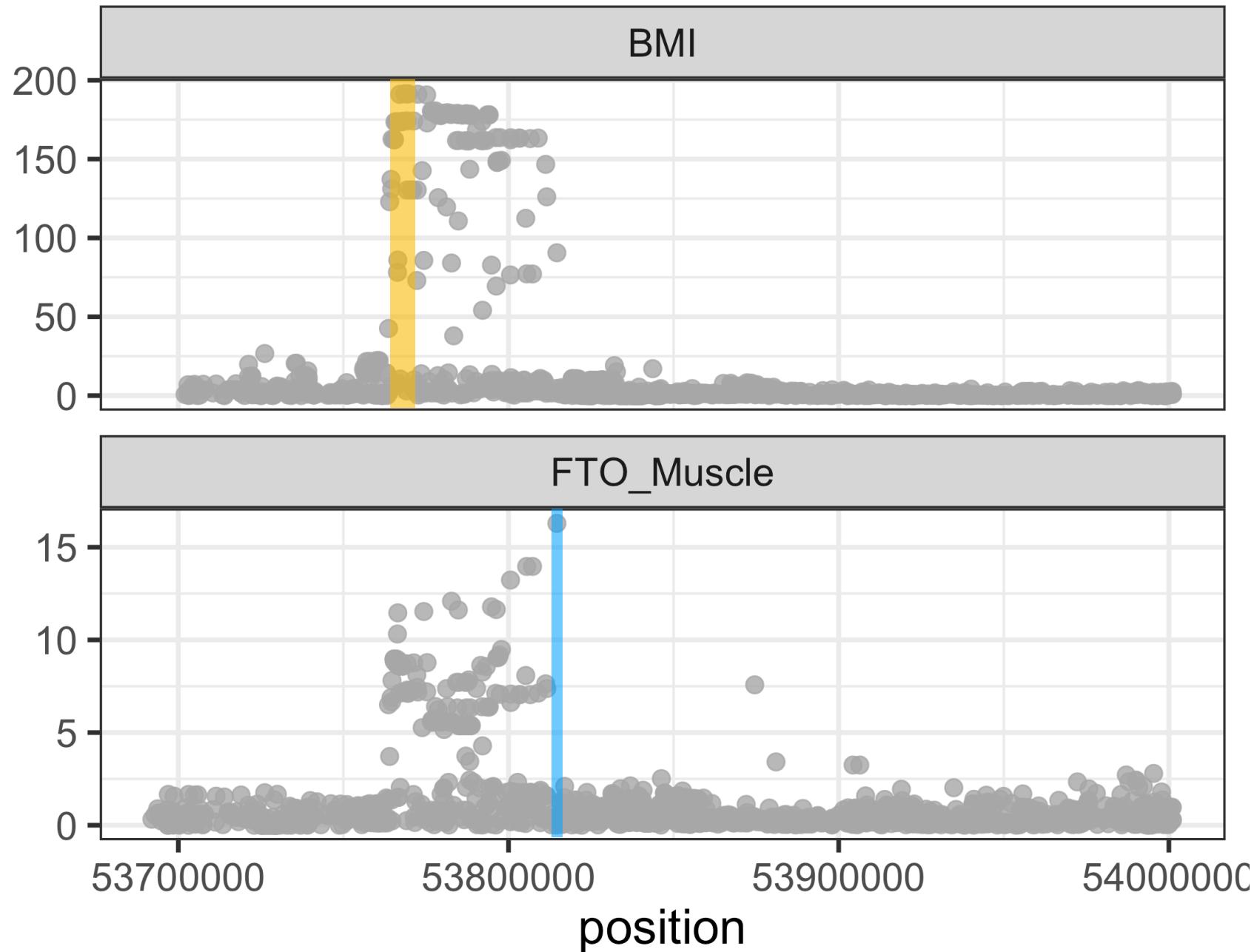
Take Home Message

Association methods can get confused with LD contamination

Colocalization

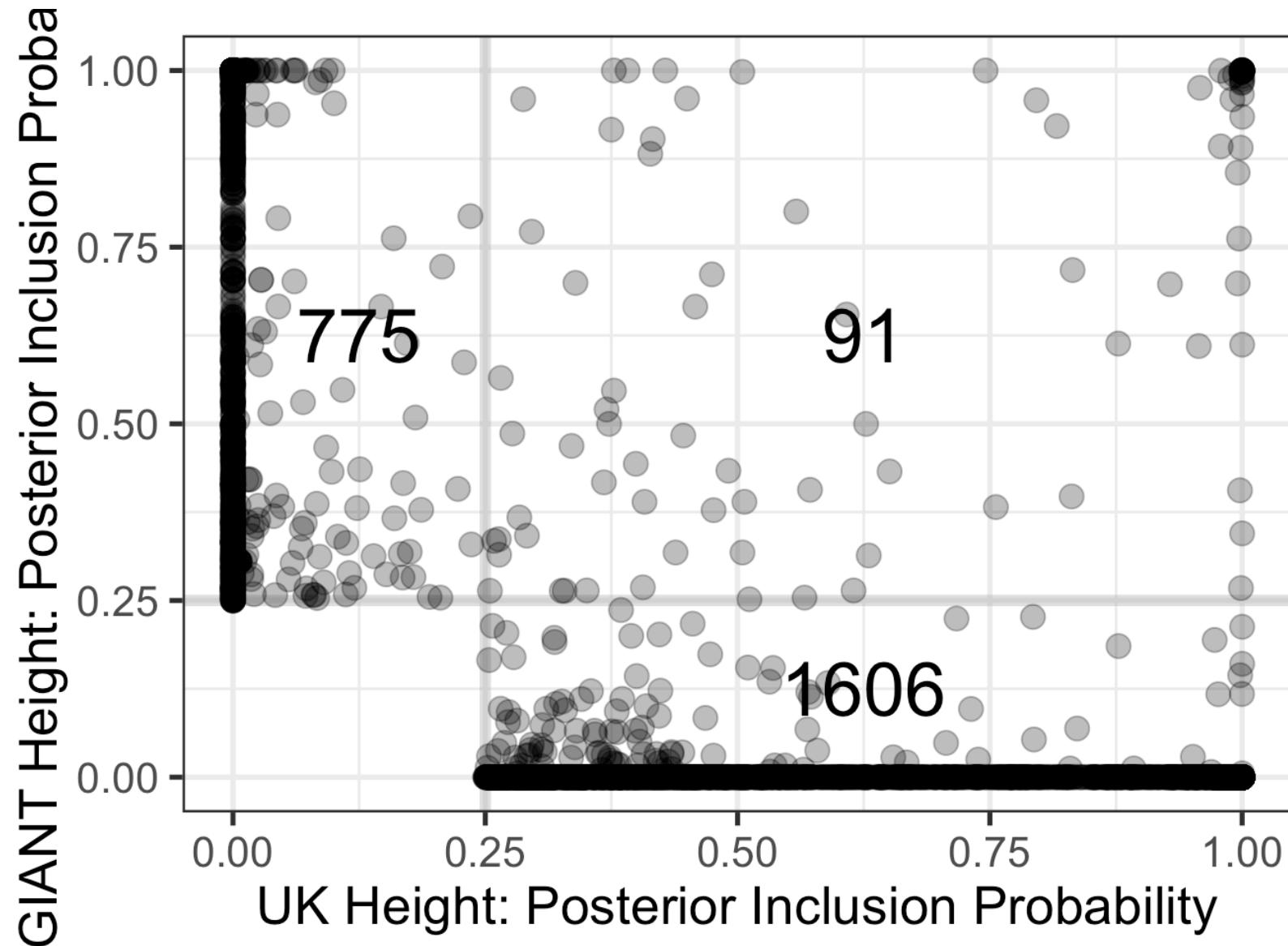


Colocalization: Are Fine-mapped Variants the Same?



Limitations of Colocalization

Fine-mapping Can Be Unreliable



Two thirds of the GIANT height loci do not colocalize with the UK Biobank height loci

SusieR ran by Yanyu Liang

Take Home Message

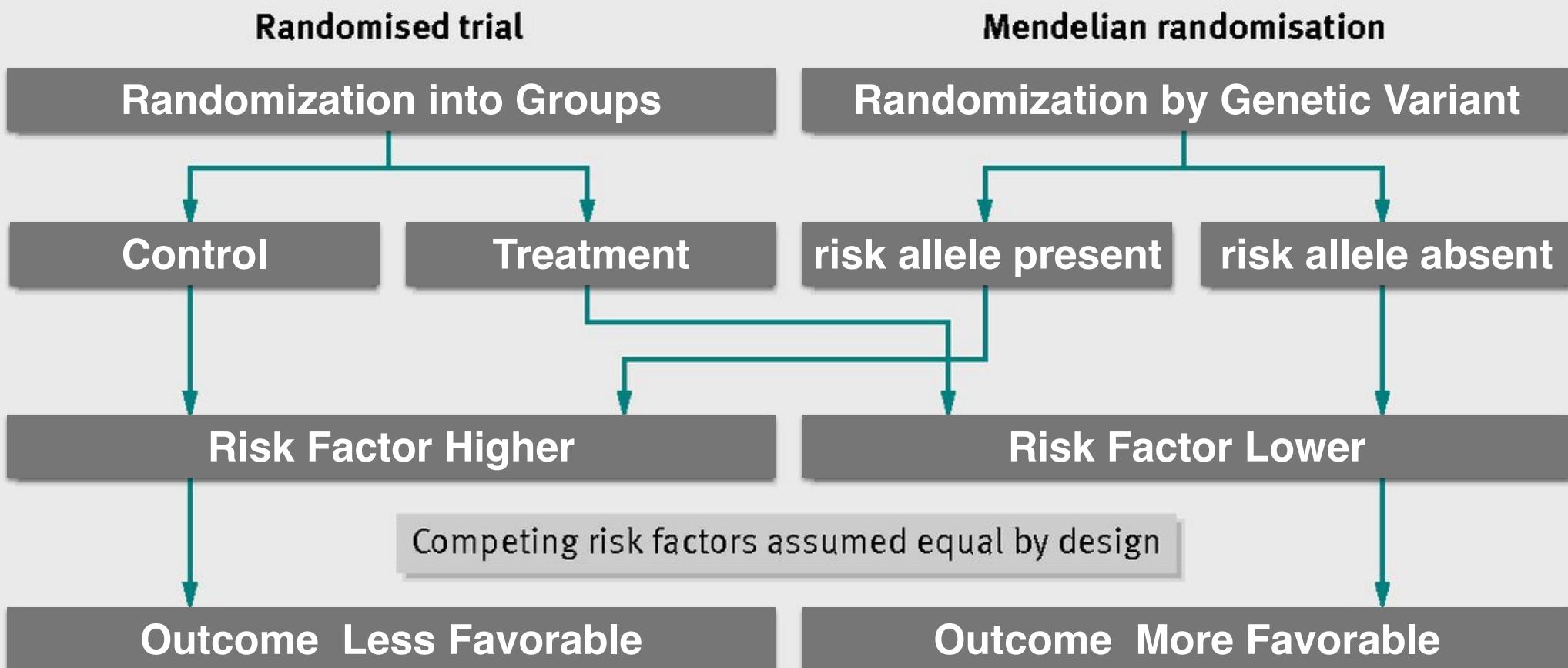
Colocalization probability can
be severely underestimated

Recommendation for Post GWAS Analysis

- Start with an association method to find list of candidate causal genes
- Use colocalization to filter out LD contamination
- Be aware that this is quite conservative and real signals may be tossed out
- As usual, try to get multiple lines of evidence

Mendelian Randomization

Randomized Trial vs. Mendelian Randomization



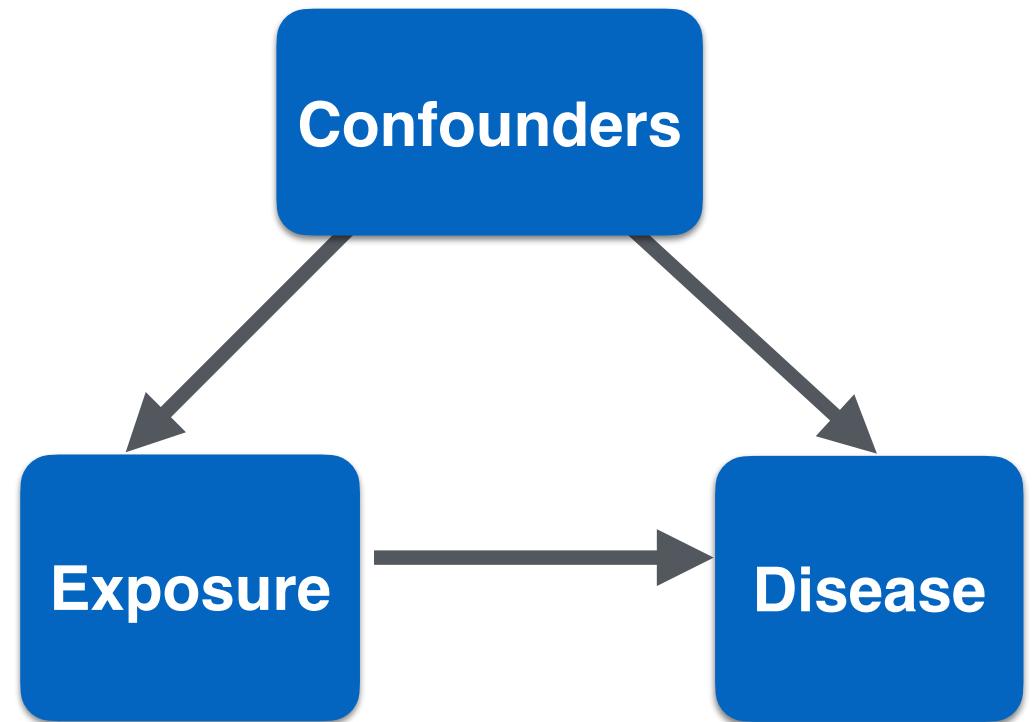
Burgess et al, Use of Mendelian randomisation to assess potential benefit of clinical intervention, BMJ 2012

Mendelian Randomization Question



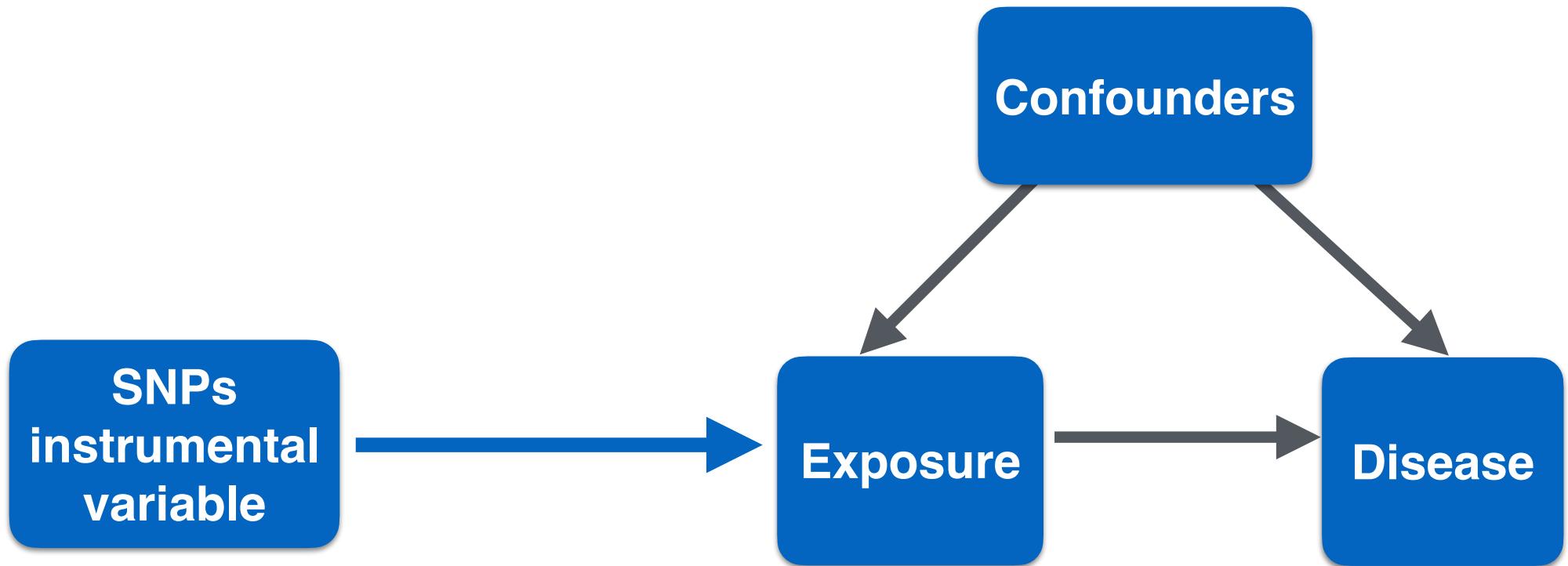
the goal is to test association between a modifiable exposure and disease. Smoking, HDL cholesterol levels, etc.

Why We Need Mendelian Randomization?



Confounders may cause misleading associations

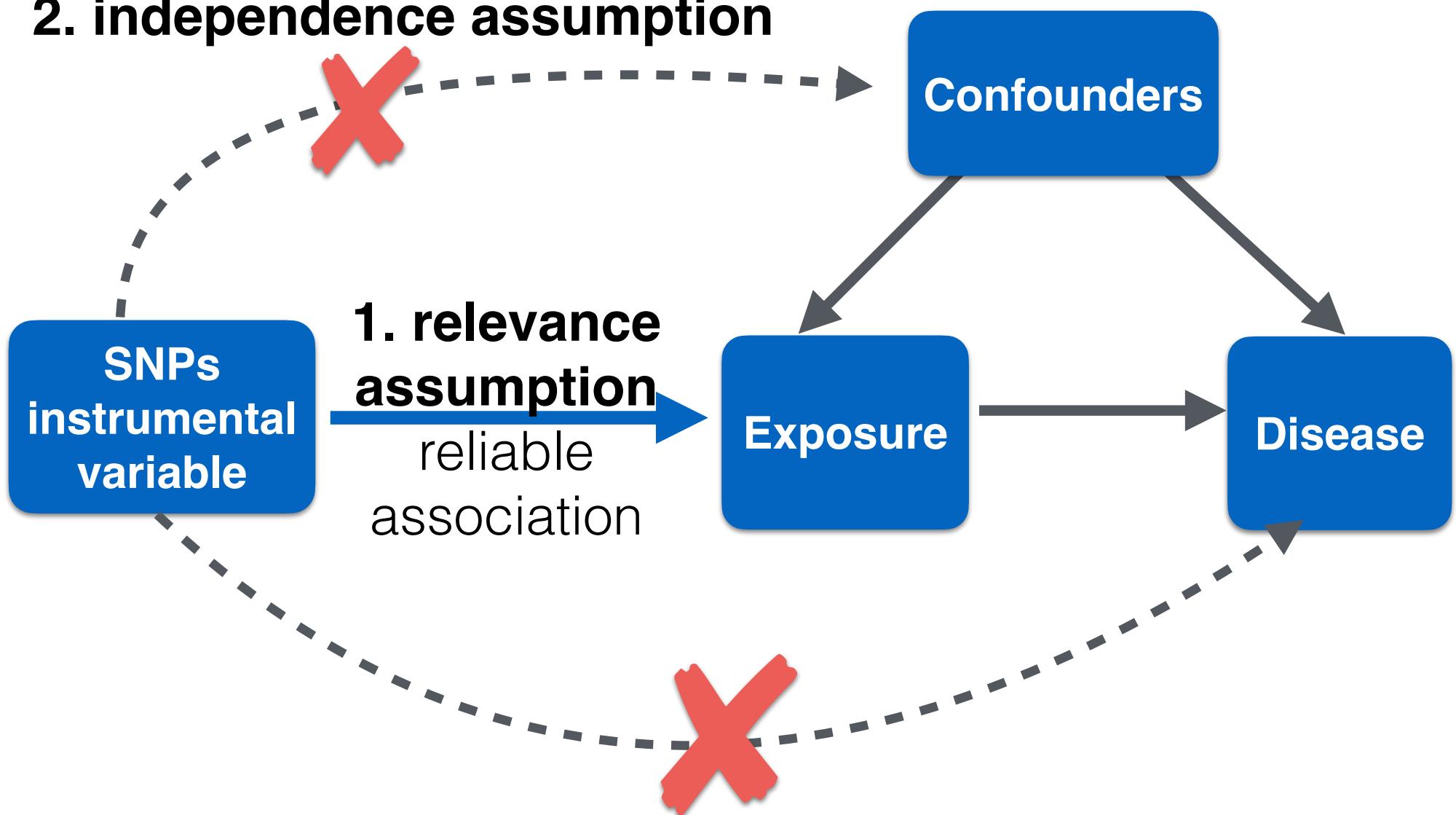
Mendelian Randomization: Instrumental Variable



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

Assumptions of Mendelian Randomization

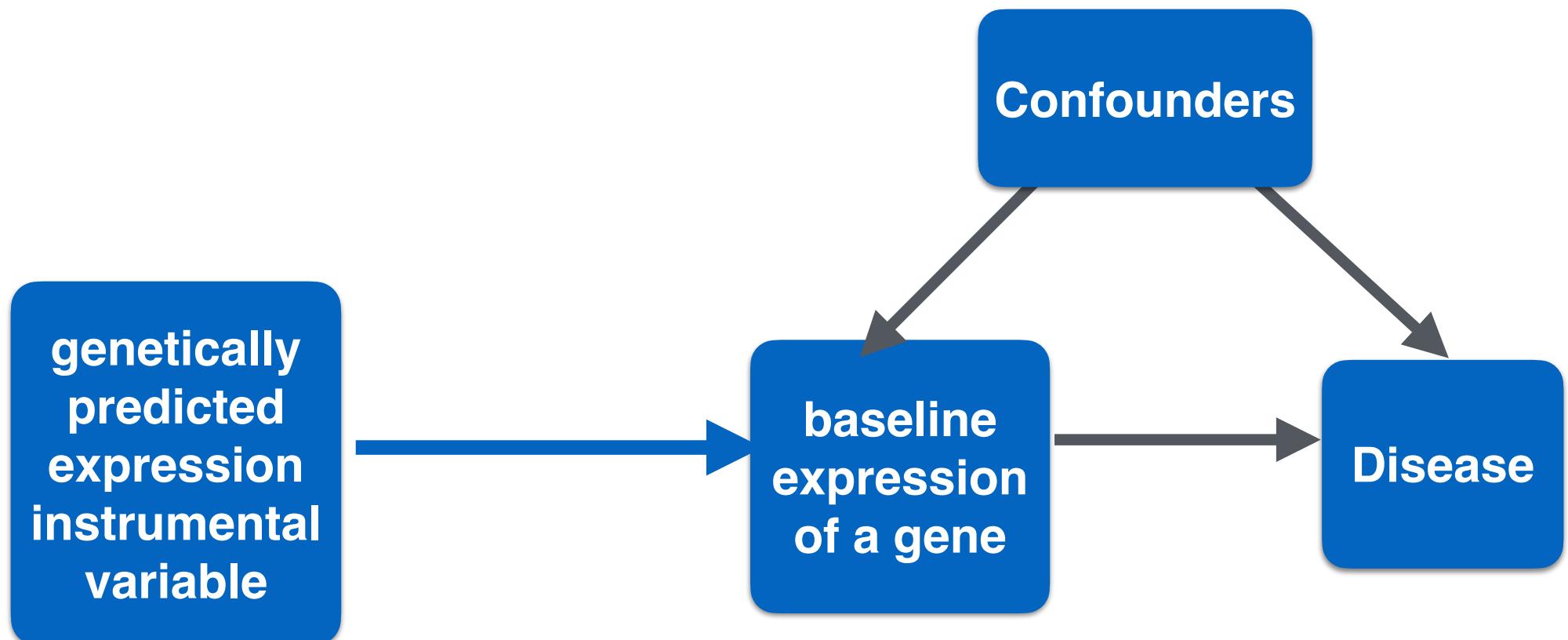
2. independence assumption



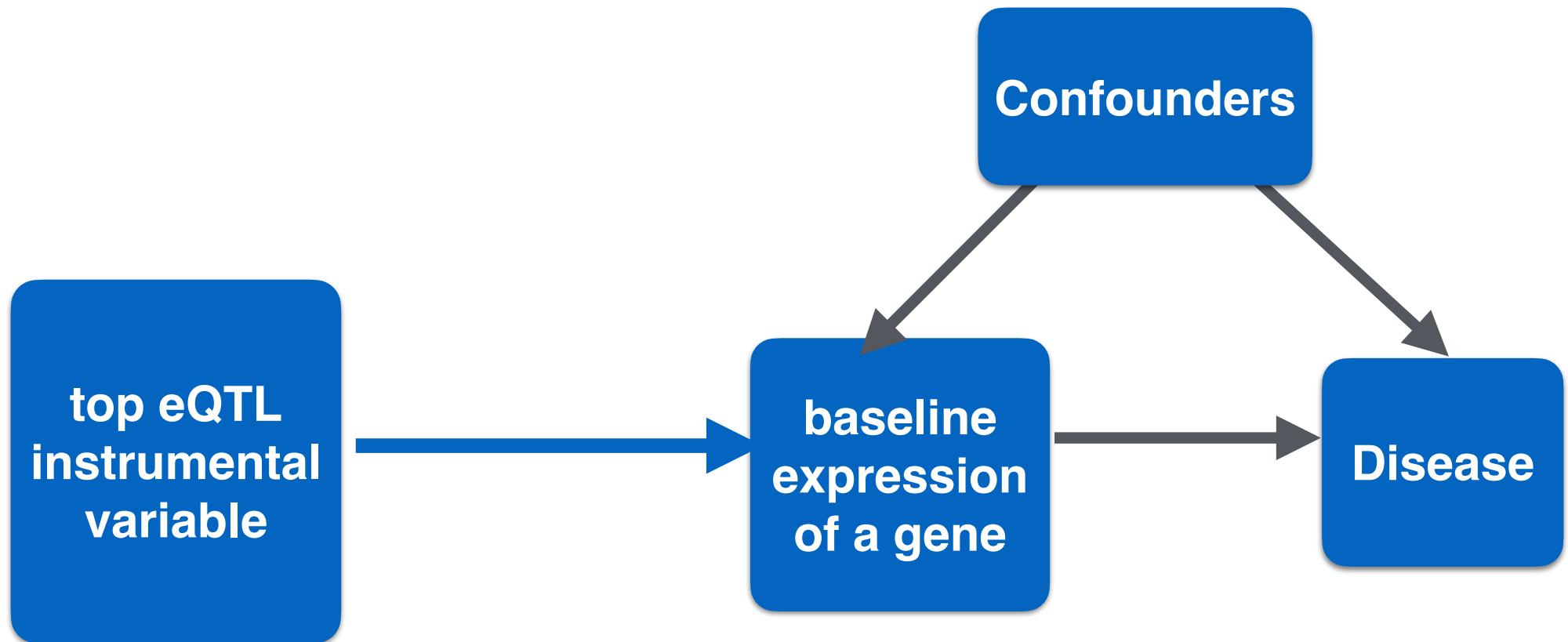
3. exclusion restriction assumption

no direct effect

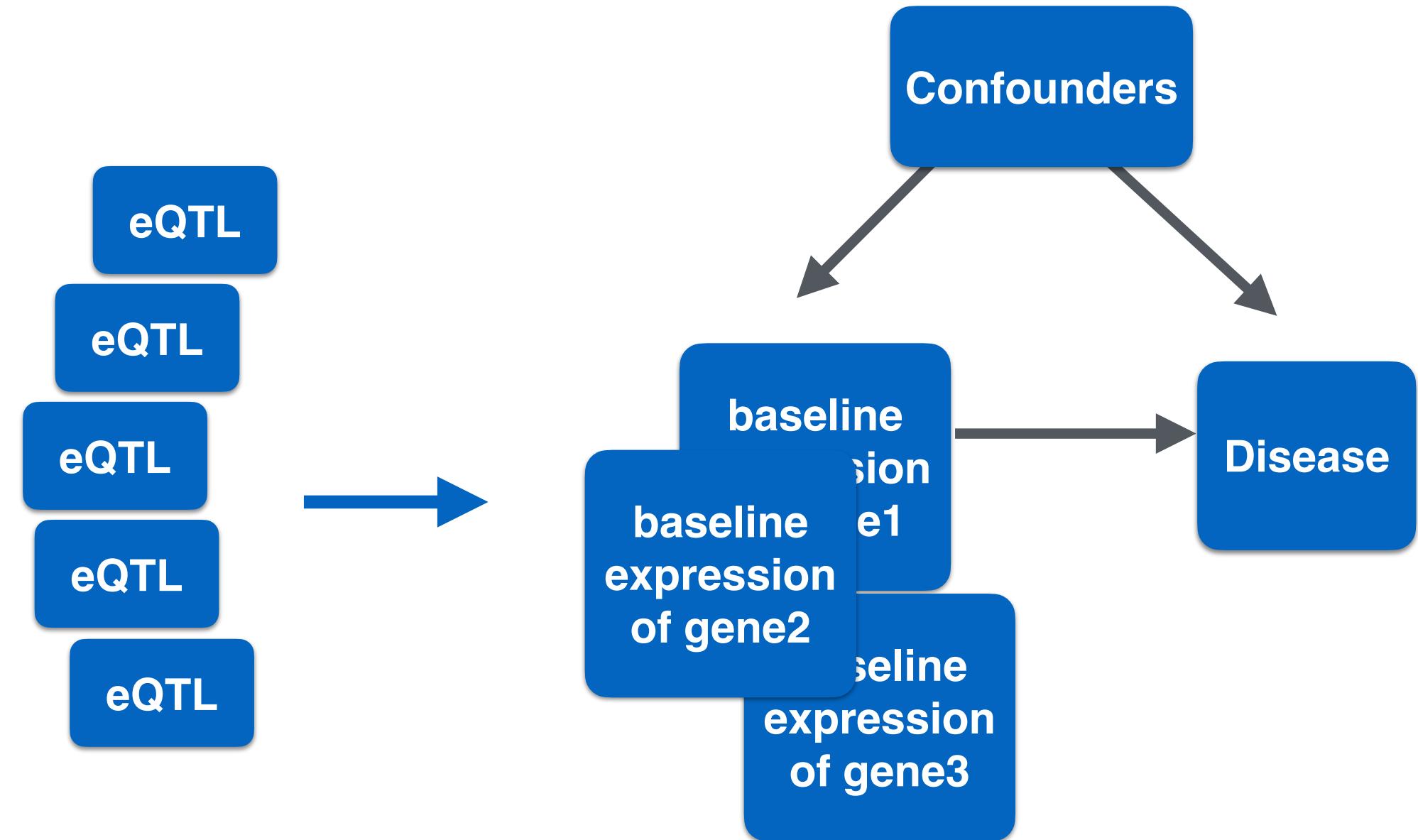
PrediXcan as a Mendelian Randomization Approach



Summary data-based Mendelian randomization (SMR)



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>

Hands On Exercises

Prerequisites

- Up to date information on data can be found here
 - https://hakyimlab.github.io/QGT-Columbia-HKI/analysis_plan.html
- Option 1 (full installation)
 - You need to have access to a Linux system and be quite familiar with it
 - Download data and software from here
 - <https://uchicago.box.com/s/zhapf2zfxcpj7thvq4sjnqale3emleum>
 - Follow additional steps from
 - https://hakyimlab.github.io/QGT-Columbia-HKI/analysis_plan.html
- Option 2
 - RStudio server will be provided with all the data, software, and environment set up
 - Claim one server here
 - https://docs.google.com/spreadsheets/d/1PKVfywvTu1RZuDcyHV4xt_cCngzhmjzw60_dHzLbLrc/edit#gid=145770066