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

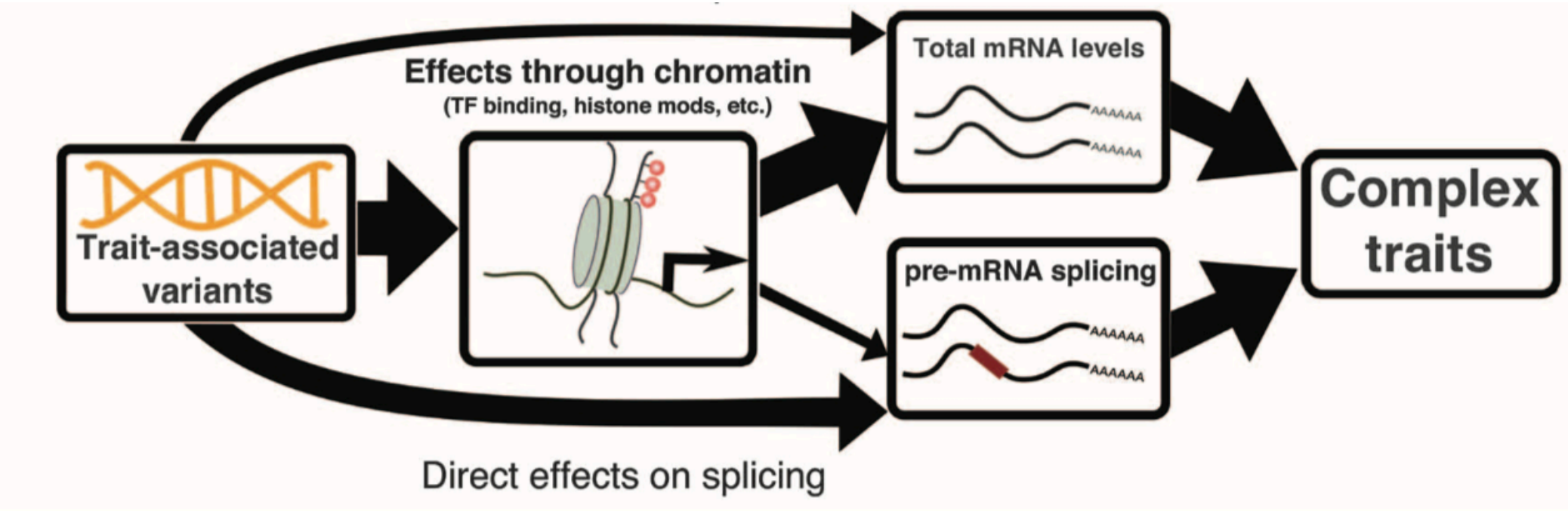
"This profound new knowledge [...] will revolutionize the diagnosis, prevention and treatment of most, if not all, human diseases.
[...] it is now conceivable that our children's children will know the term 'cancer' only as a constellation of stars."

Why are we not there yet

- Prediction of the future is hard
- Genetic architecture of common diseases and cancers is much more complex than anticipated
- 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



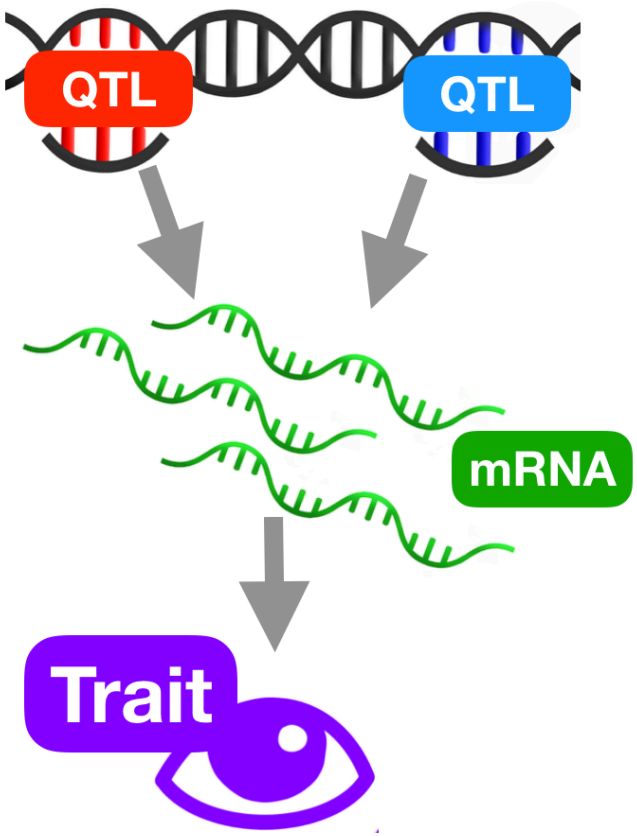
RNA splicing is a primary link between genetic variation and disease

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

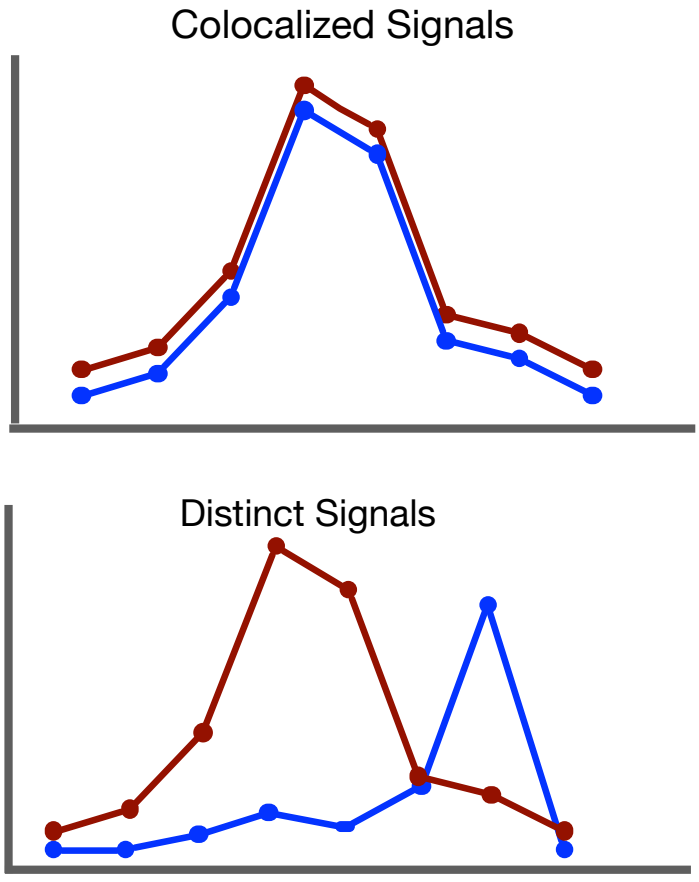
Association vs. Colocalization Methods

Association



PrediXcan, SMR,
FUSION (formerly TWAS)

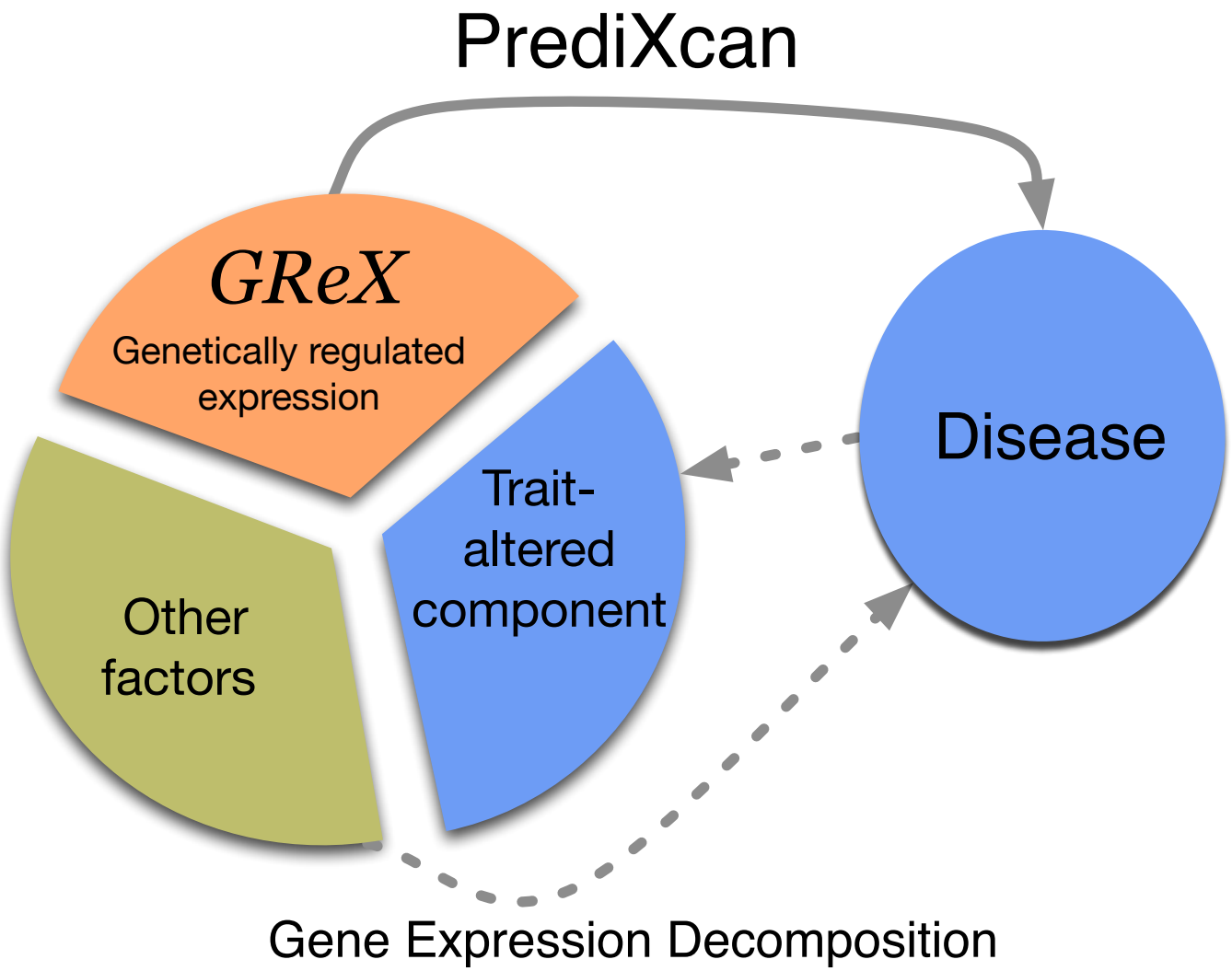
Colocalization



Coloc, Enloc, eCAVIAR, fastENLOC

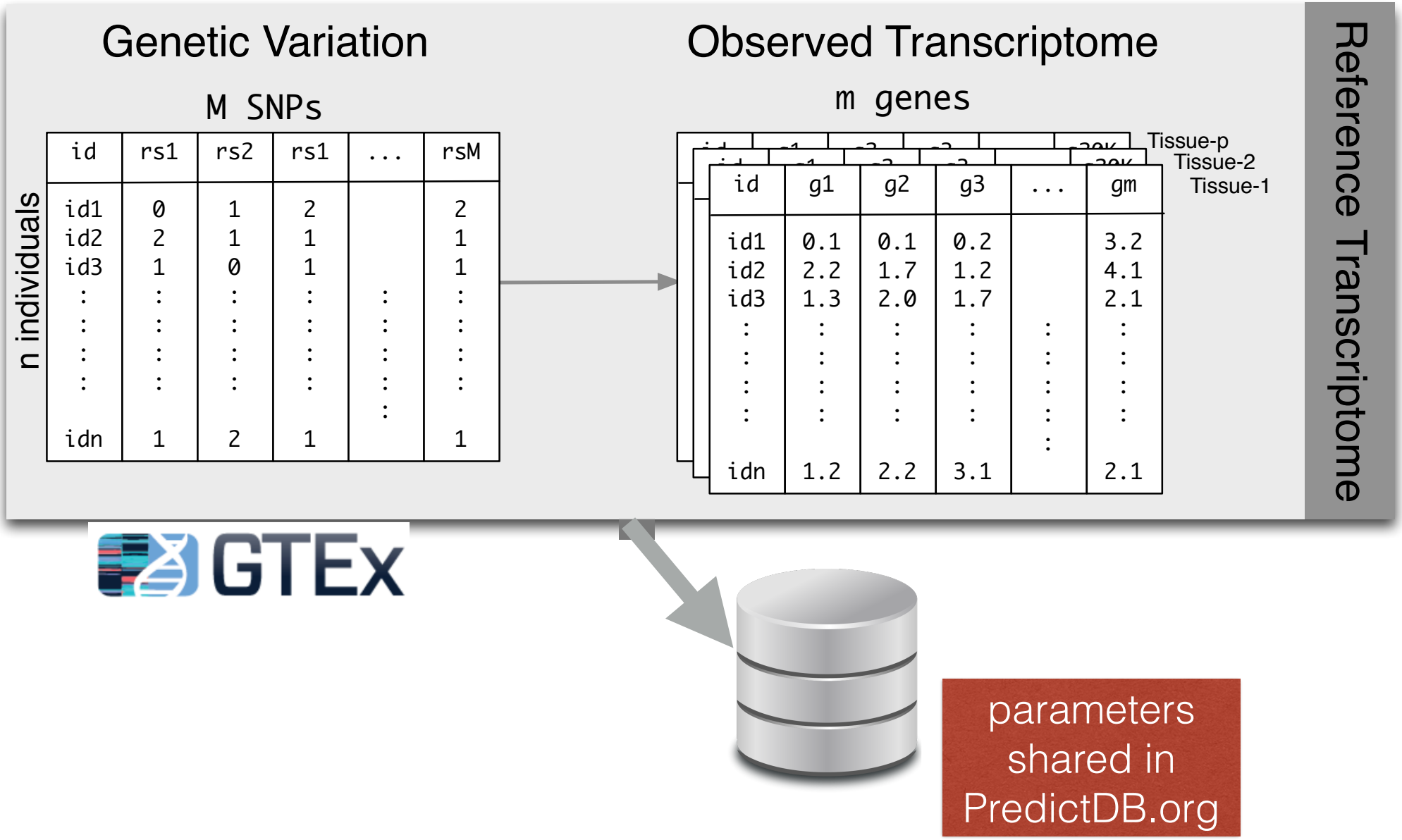
Association Approach

PrediXcan Uses Association Between GReX and Disease To Identify Causal Genes

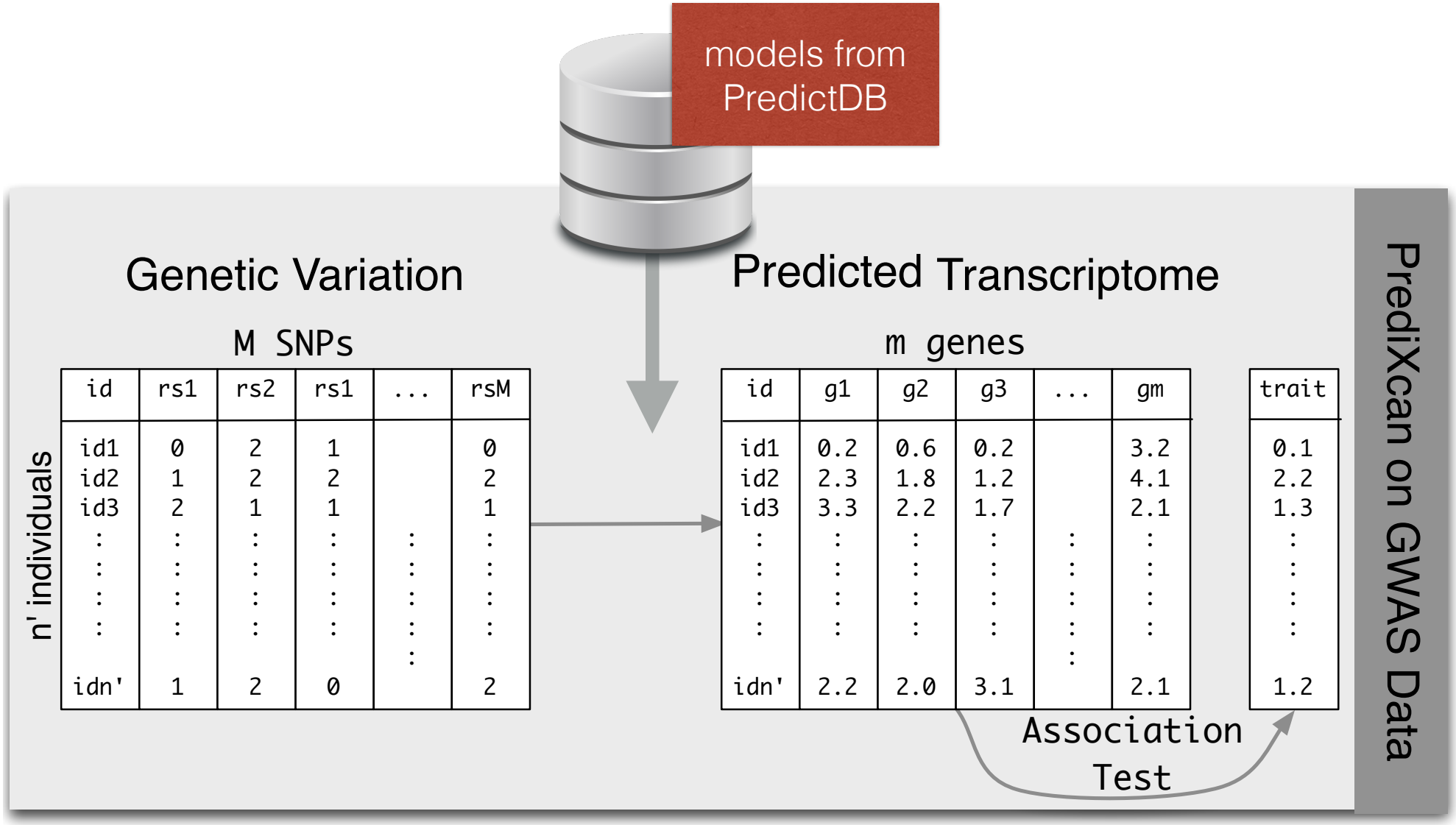


Gamazon, Wheeler, Shah et al 2015 Nature Genetics

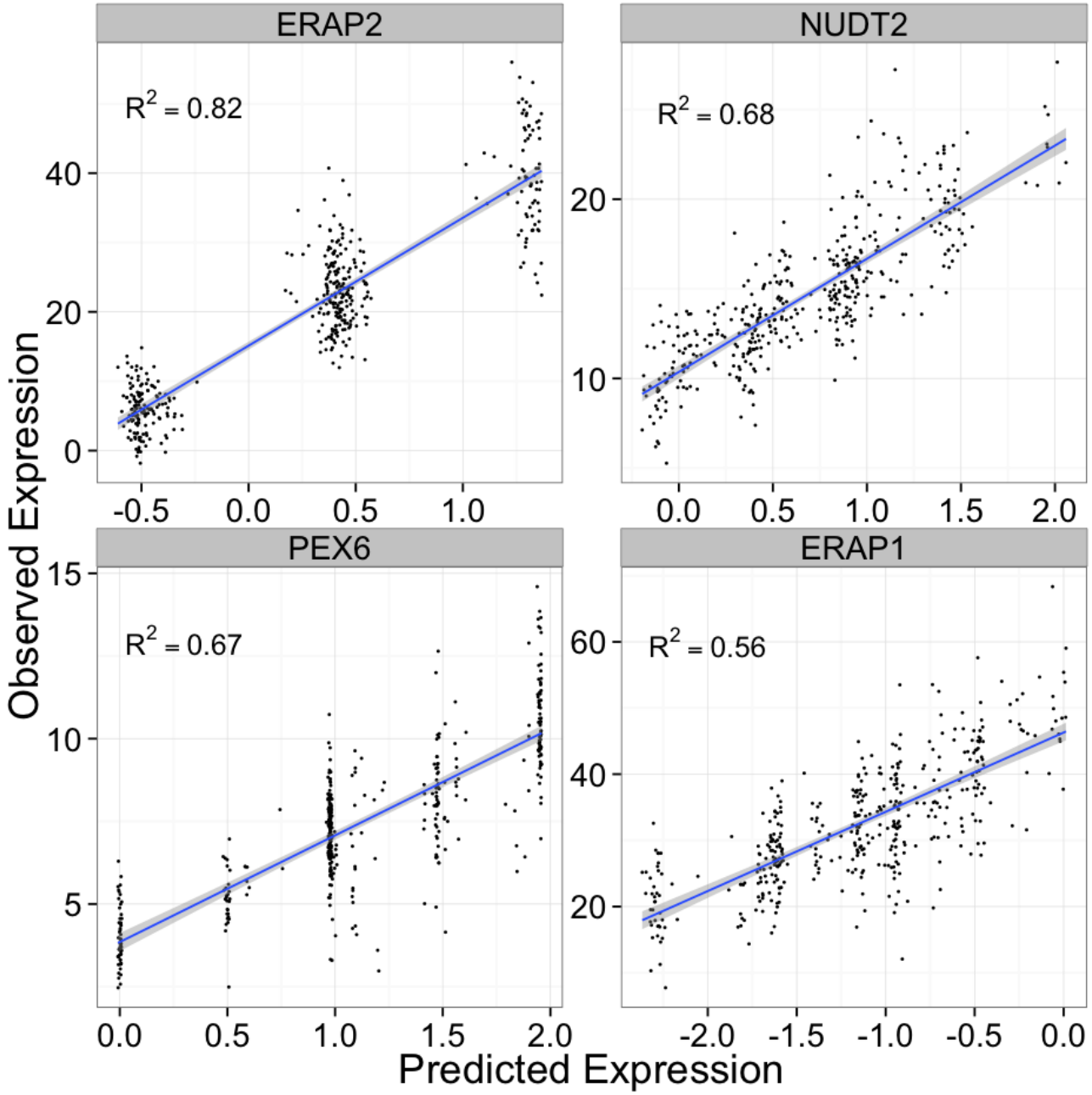
PrediXcan Trains Predictors using Reference Transcriptome Data



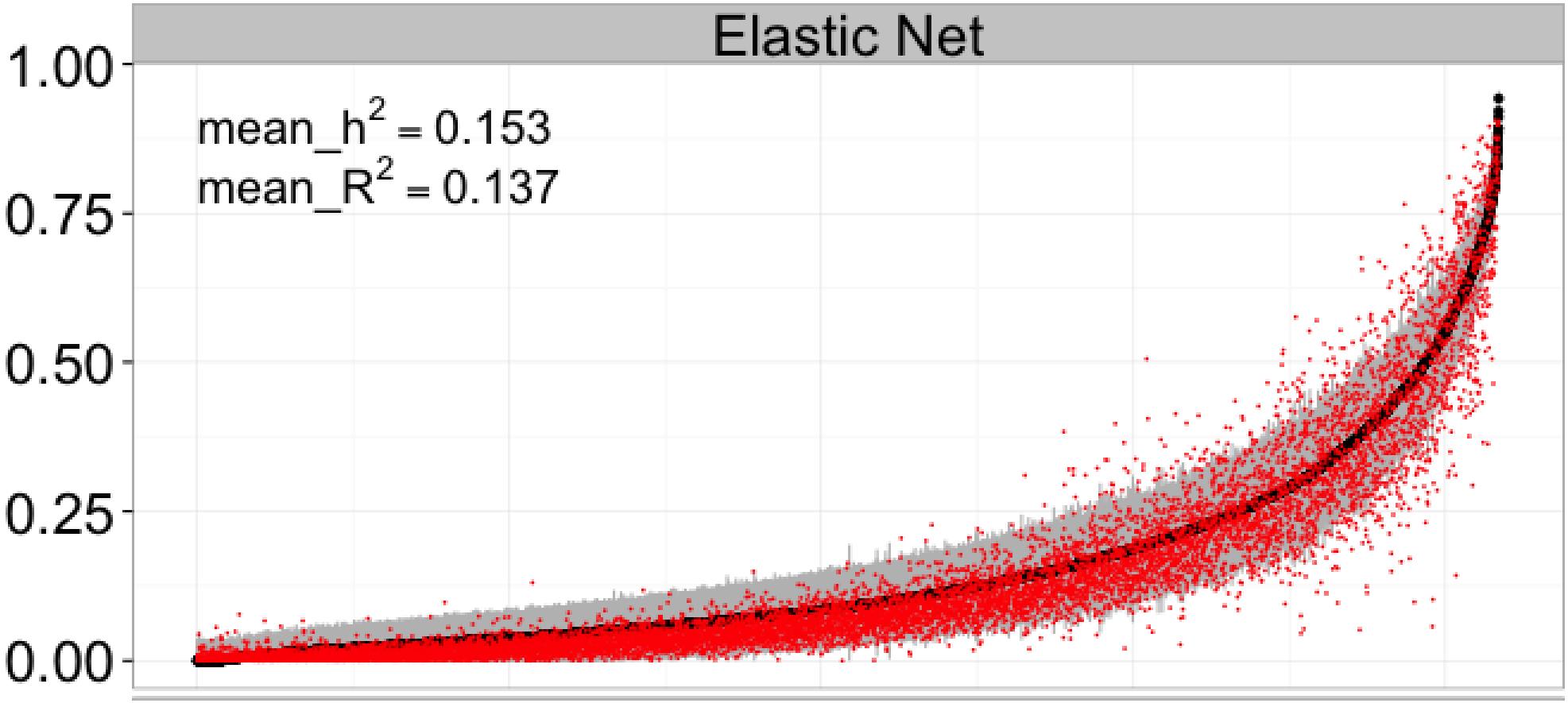
PrediXcan Predicts the Transcriptome & Tests Assoc.



Examples of Well Predicted Genes

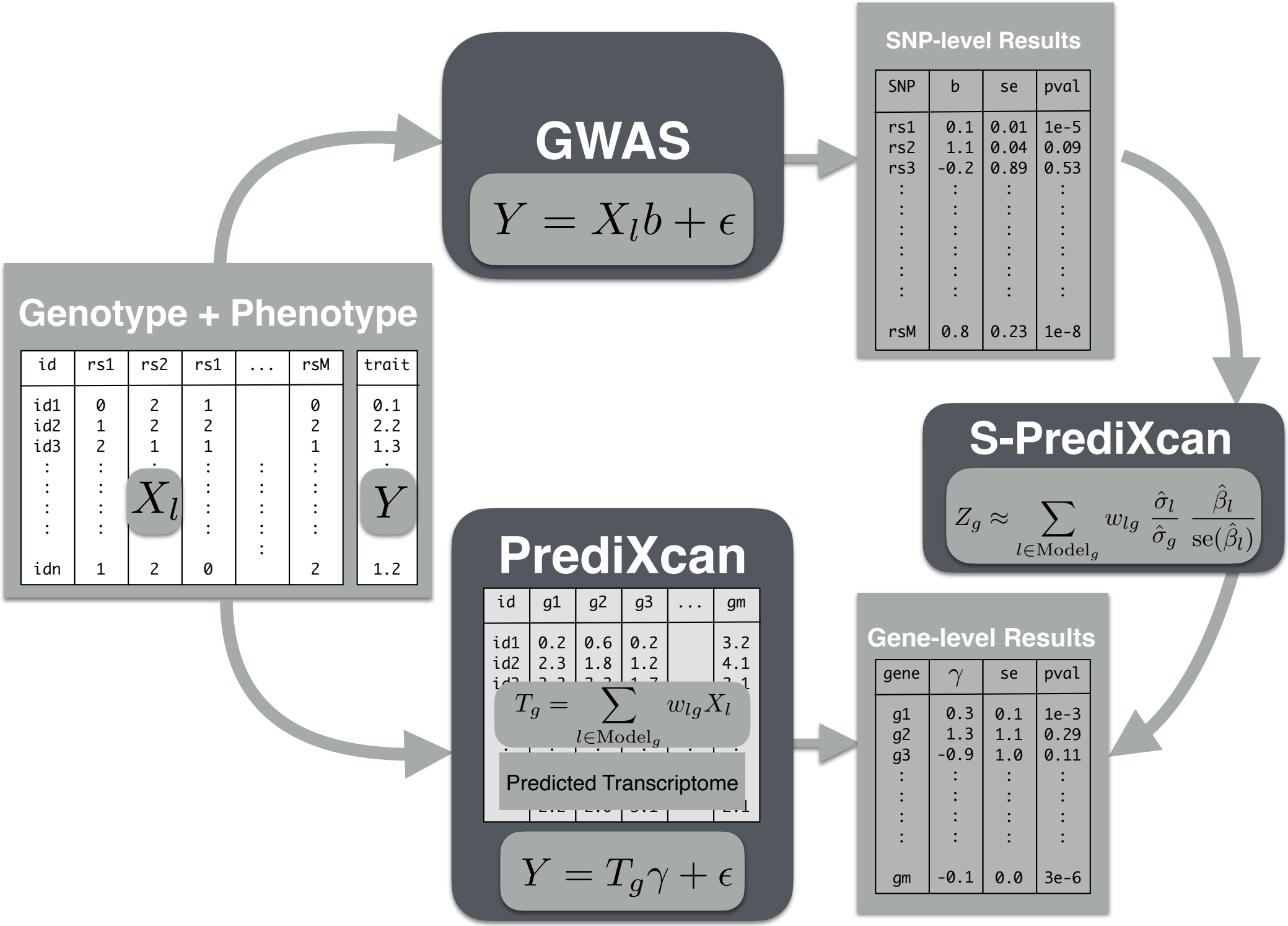


Accuracy of Prediction Depends on Heritability



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

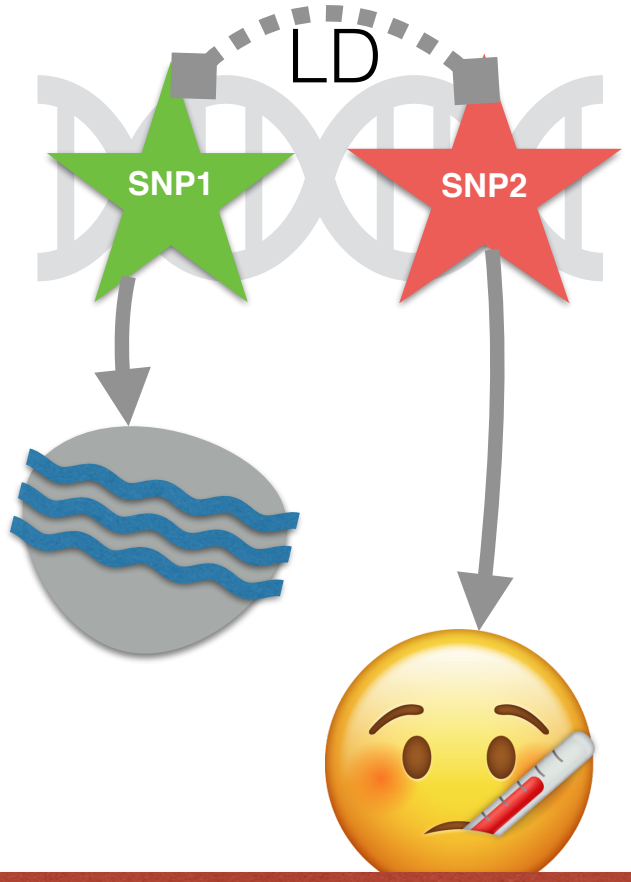


Barbeira et al 2018 https://www.nature.com/articles/s41467-018-03621-

Limitations of Current Association Methods

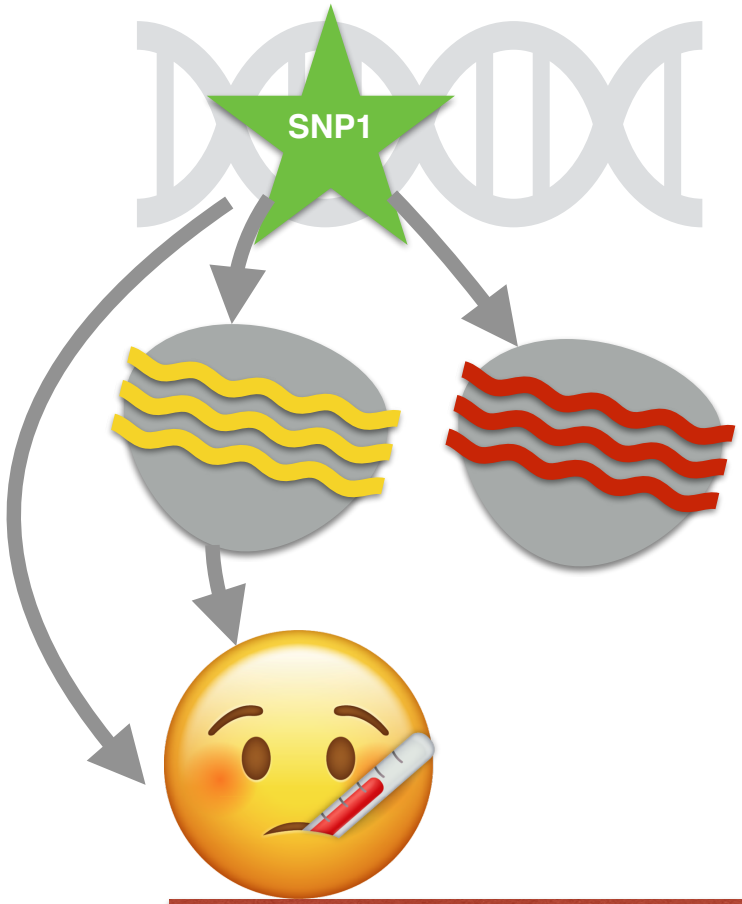
Limitations of Downstream Association Methods

LD Contamination



Post Filtering with colocalization methods

Co-Regulation

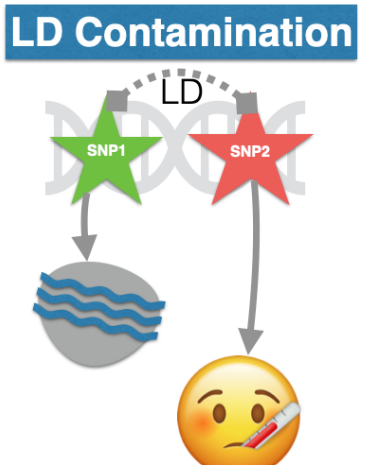


More difficult to tackle

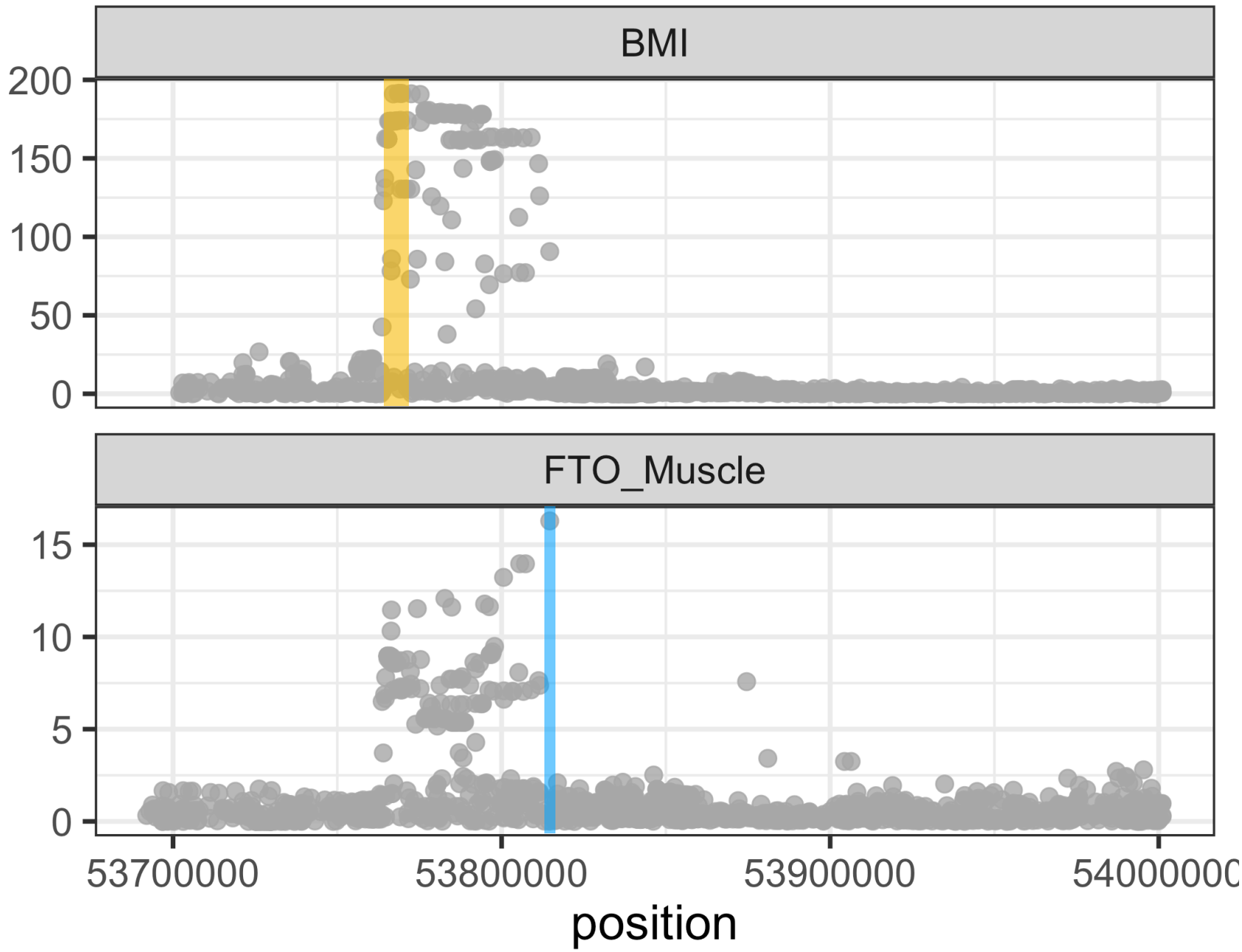
Take Home Message

Association methods can get confused with LD contamination

Colocalization

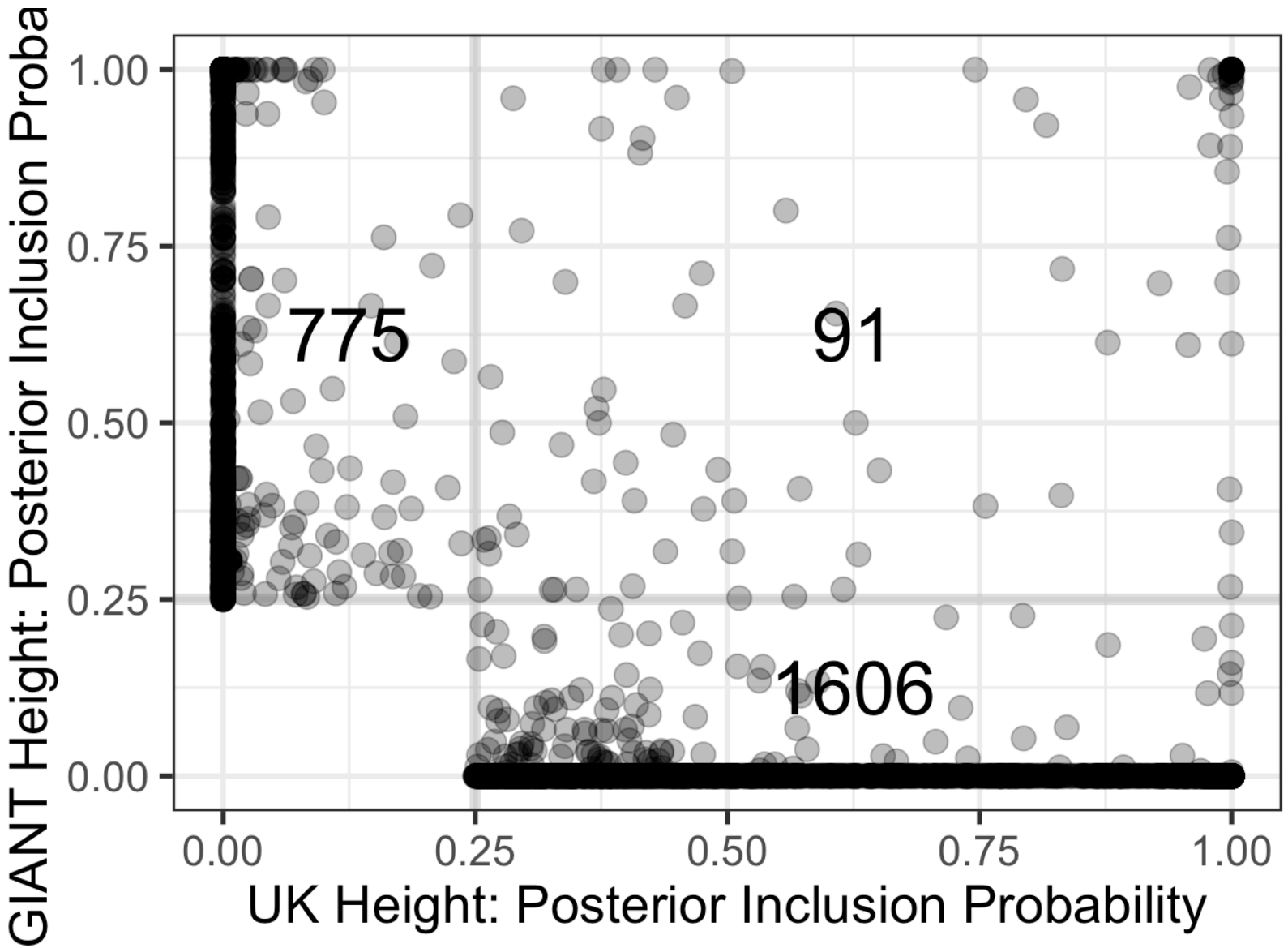


Colcalization: Are Fine-mapped Variants the Same?



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

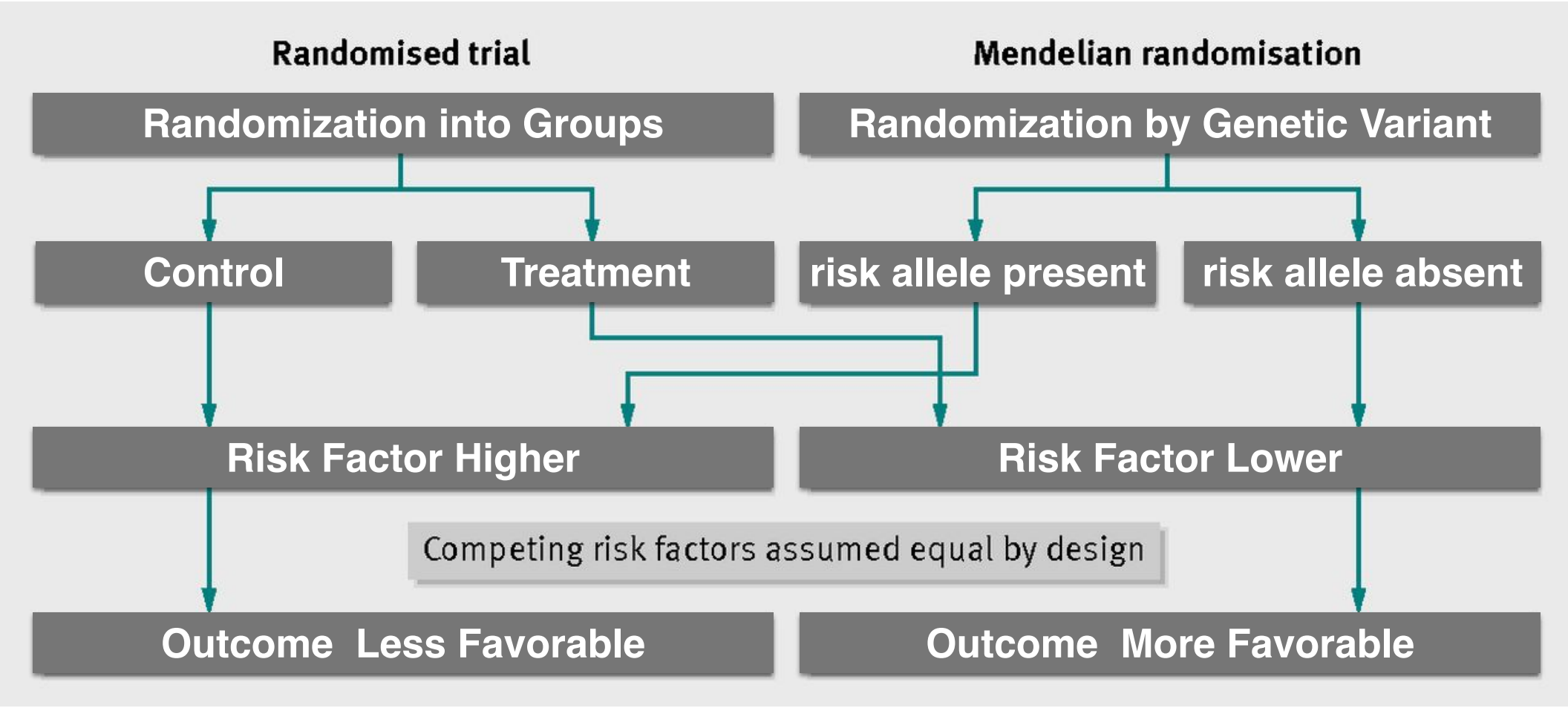
Colocalization probability can be severely underestimated

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

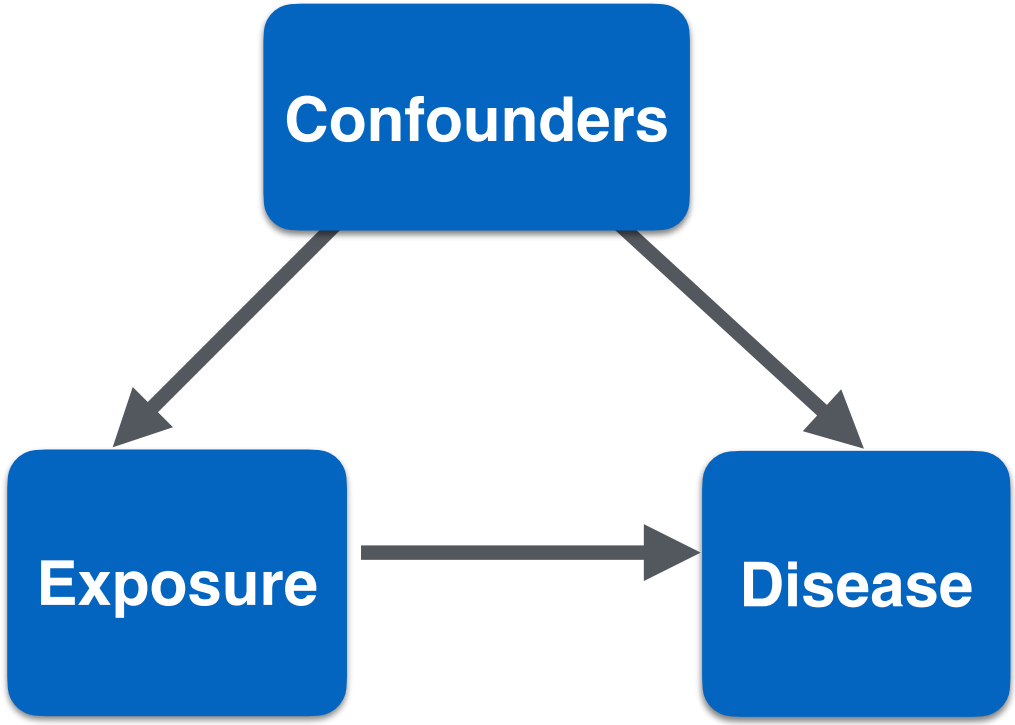


Burgess et al, Use of Mendelian randomisation to assess potential benefit of clinical intervention, BMJ 2012

Mendelian Randomization Question

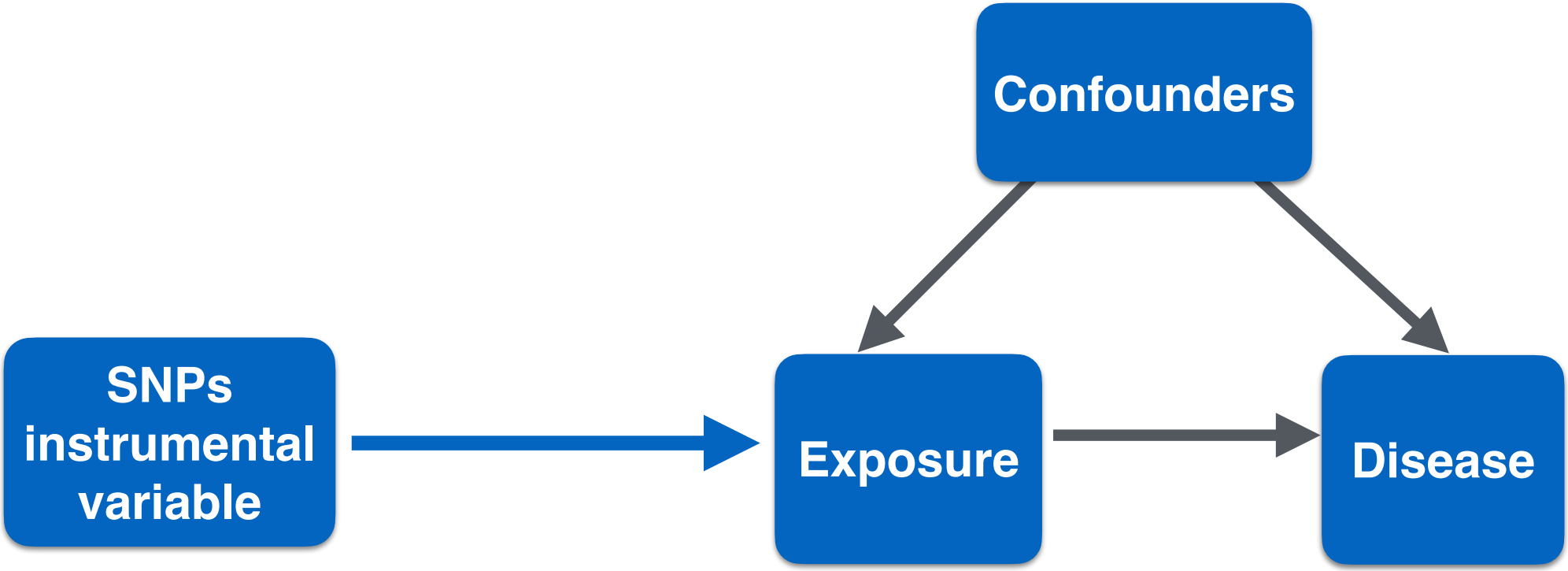
the goal is to test association between a modifiable exposure and disease. Smoking, HDL cholesterol levels, etc.

Why We Need Mendelian Randomization?



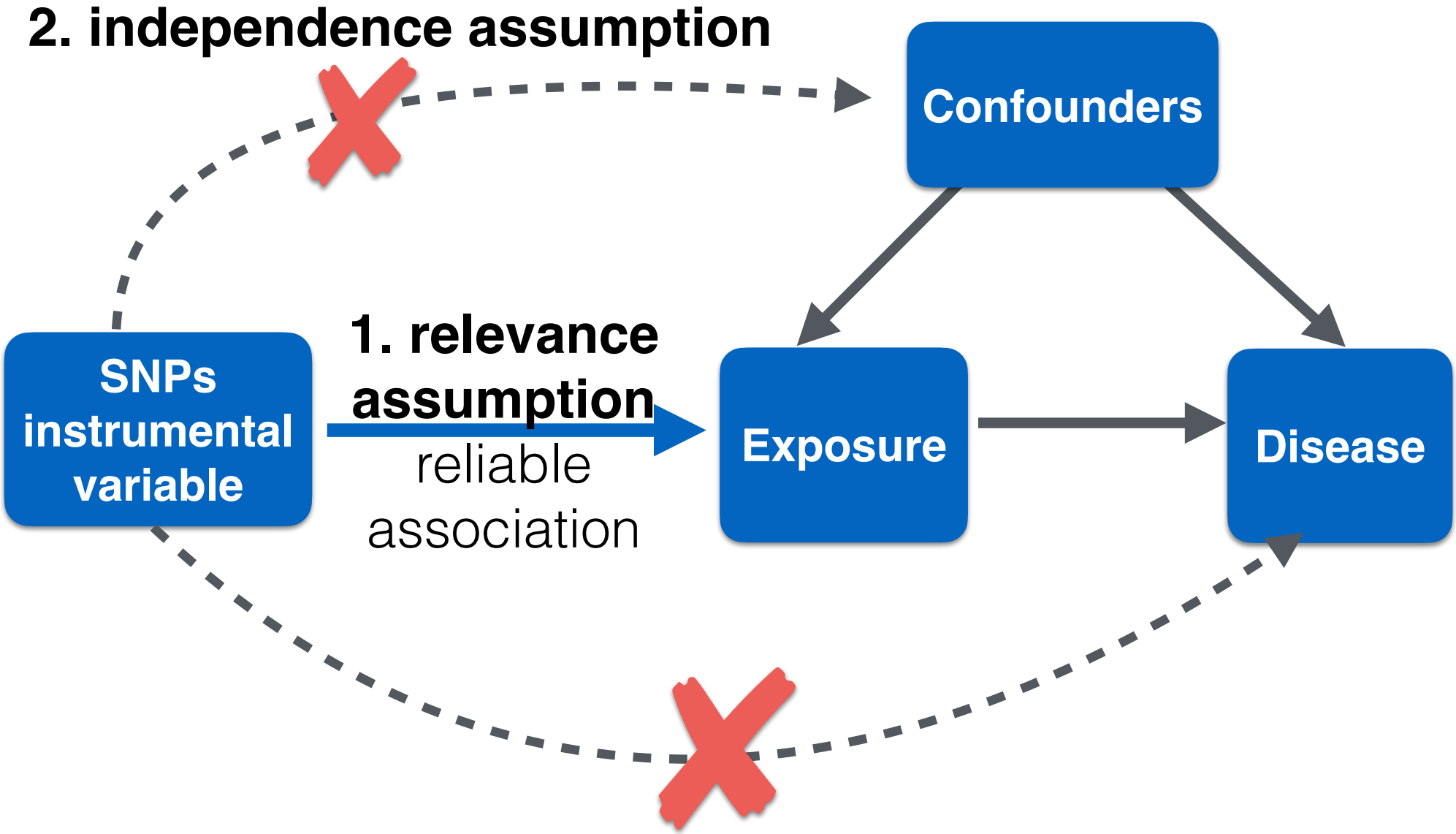
Confounders may cause misleading associations

Mendelian Randomization: Instrumental Variable

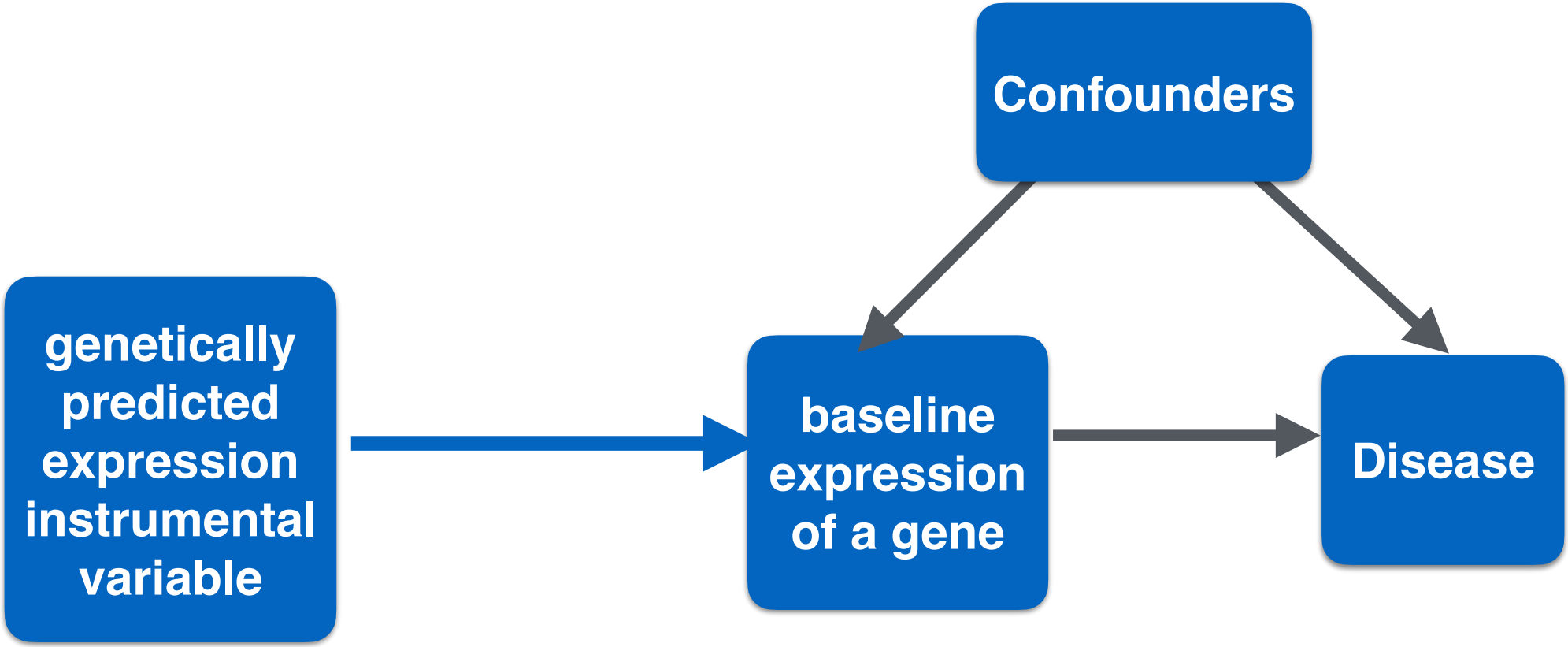


The idea is to use an "instrumental variable" without the confounding/noise

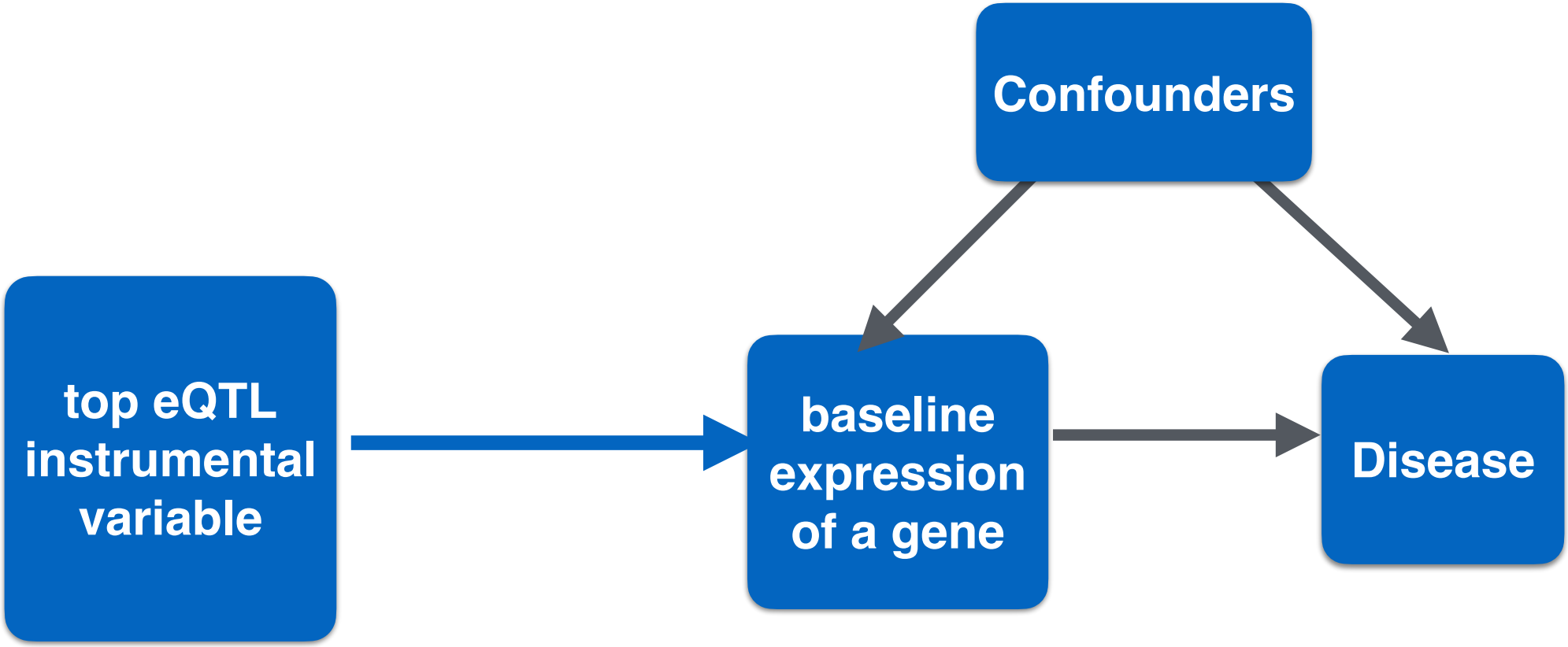
Assumptions of Mendelian Randomization



PrediXcan as a Mendelian Randomization Approach

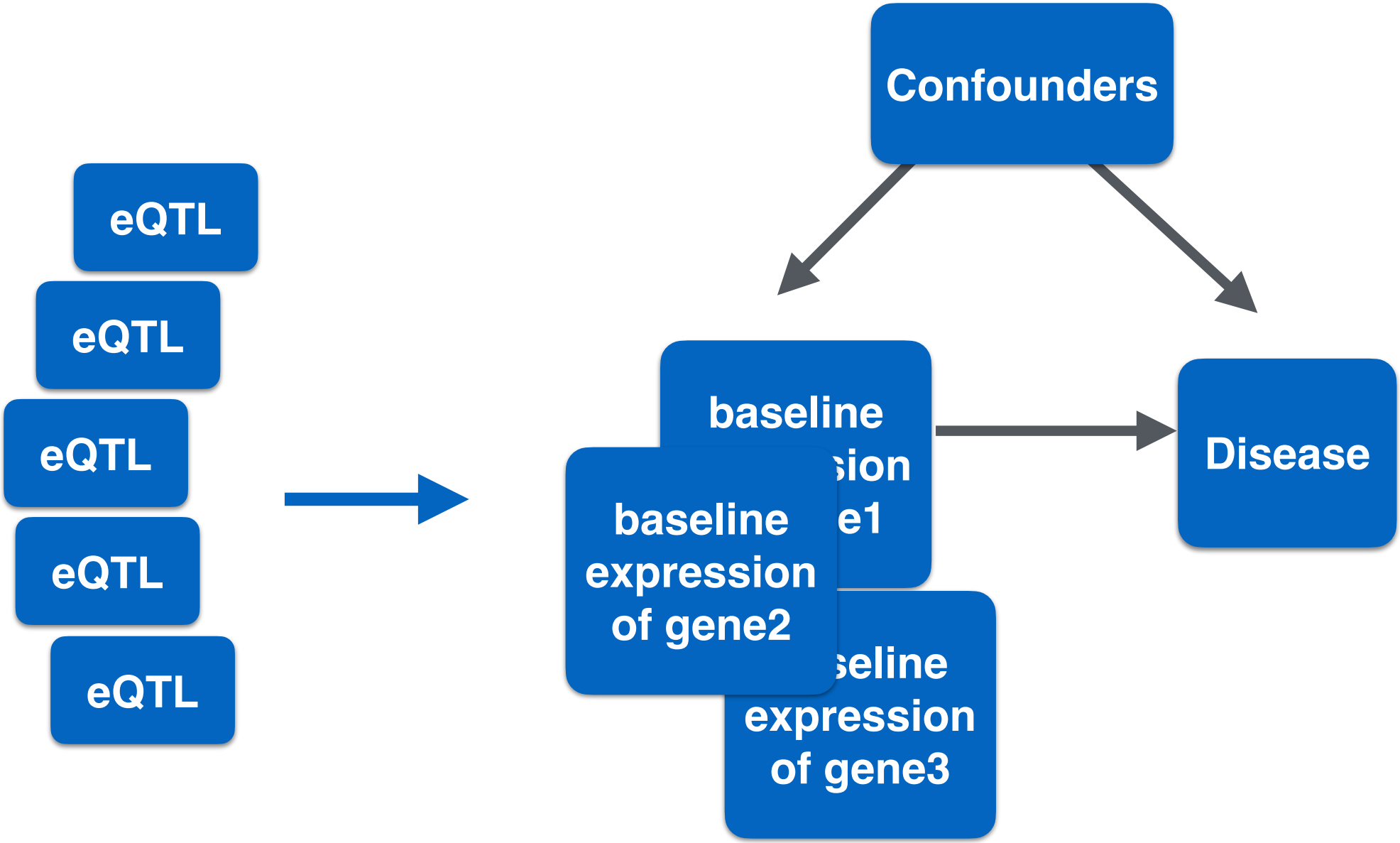


Summary data–based Mendelian randomization (SMR)



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 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)

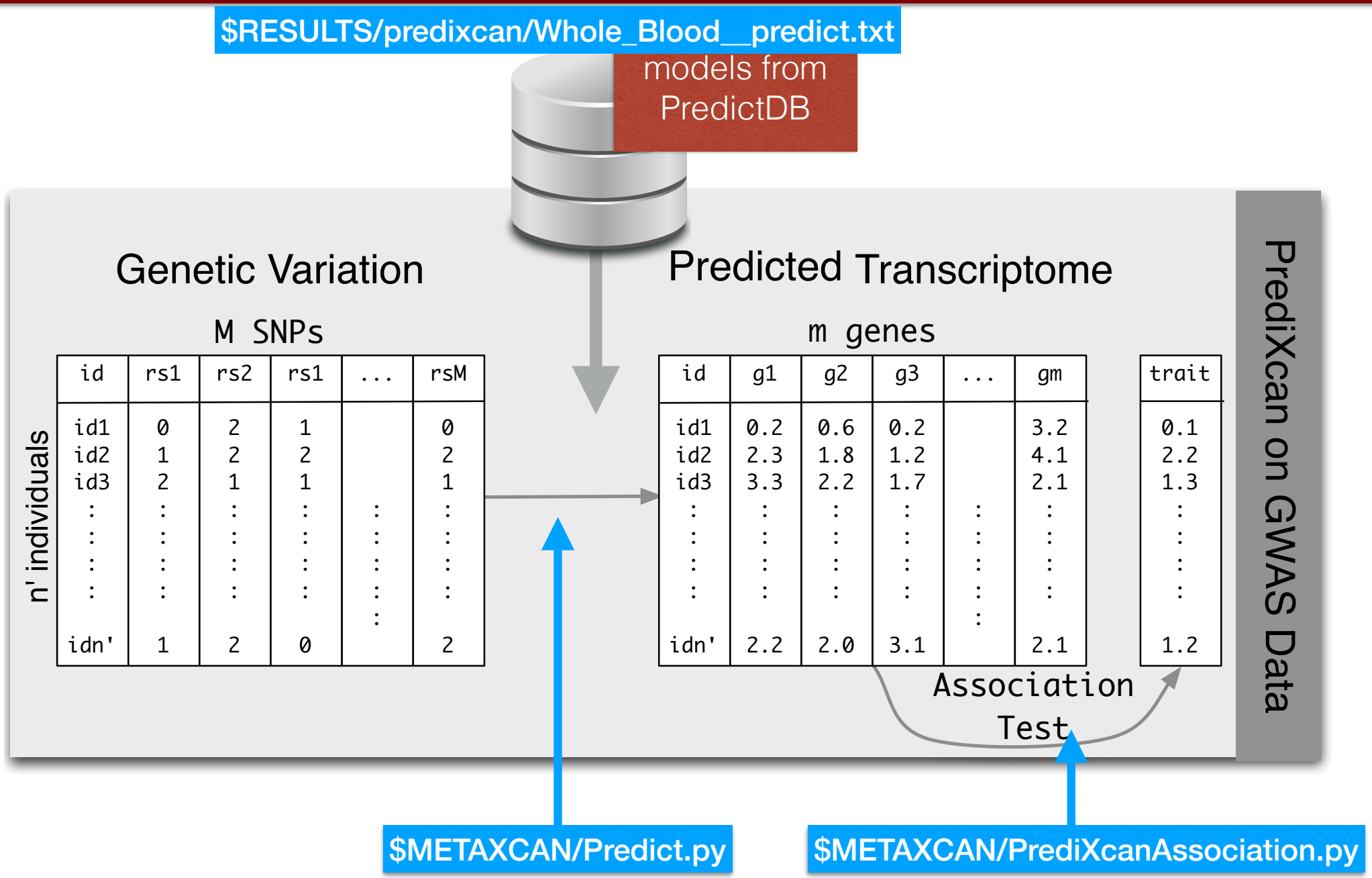
Hands On Exercises

Prerequisites

- Download data and software from here
  - https://uchicago.box.com/s/zhapf2zfxcpj7thvq4sjnqale3emleum
- Option 1 (full installation)
  - You need to have access to a Linux system and be quite familiar with it
  - Install Miniconda/Anaconda
  - Set up the imlabtools environment
    - https://github.com/hakyimlab/MetaXcan/blob/master/software/conda_env.yaml
  - Rstudio/R/tidyverse
- Option 2
  - RStudio server will be provided with all the software and environment set up
  - Claim one server here
    - https://docs.google.com/spreadsheets/d/1PKVfywvTu1RZuDCyHV4xt_cCngzhmjzw60_dHzLbLrc/edit#gid=145770066
- email haky@uchicago.edu with any questions about setup and data

```yaml
name: imlabtools
channels:
  - defaults
  - conda-forge
  - moble
  - bioconda
dependencies:
  - python=3.7
  - pandas=0.25.3
  - scipy=1.4.1
  - numpy=1.18.1
  - bgen_reader=3.0.2
  - cyvcf2=0.20.0
  - pyliftover=0.4
  - statsmodels=0.11.1
  - h5py=2.10.0
  - pyarrow=0.11.0
```

PrediXcan Run



Colcalization Run