Large Scale Prediction and Dissection of Complex Traits: PrediXcan

Hae Kyung Im, PhD

THE UNIVERSITY OF CHICAGO

Successes and Challenges of Genome Studies

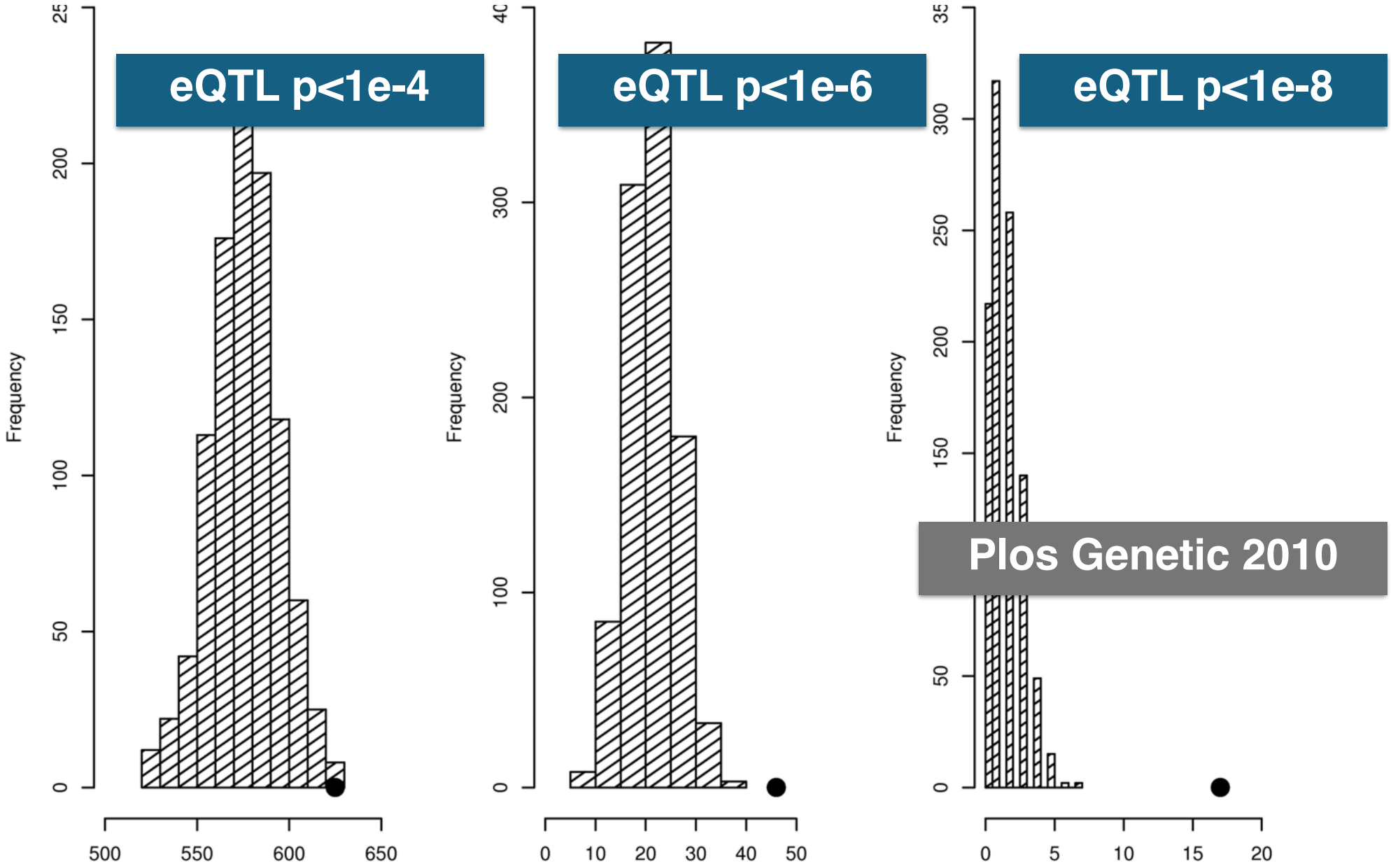
- GWAS/Sequencing
  - 10K robustly associated genetic variants
  - New insights into biology of many traits
- Biological understanding is still lacking
- Need to move beyond single variant paradigm
- Integrate more functional information into association studies

Gene Based Tests

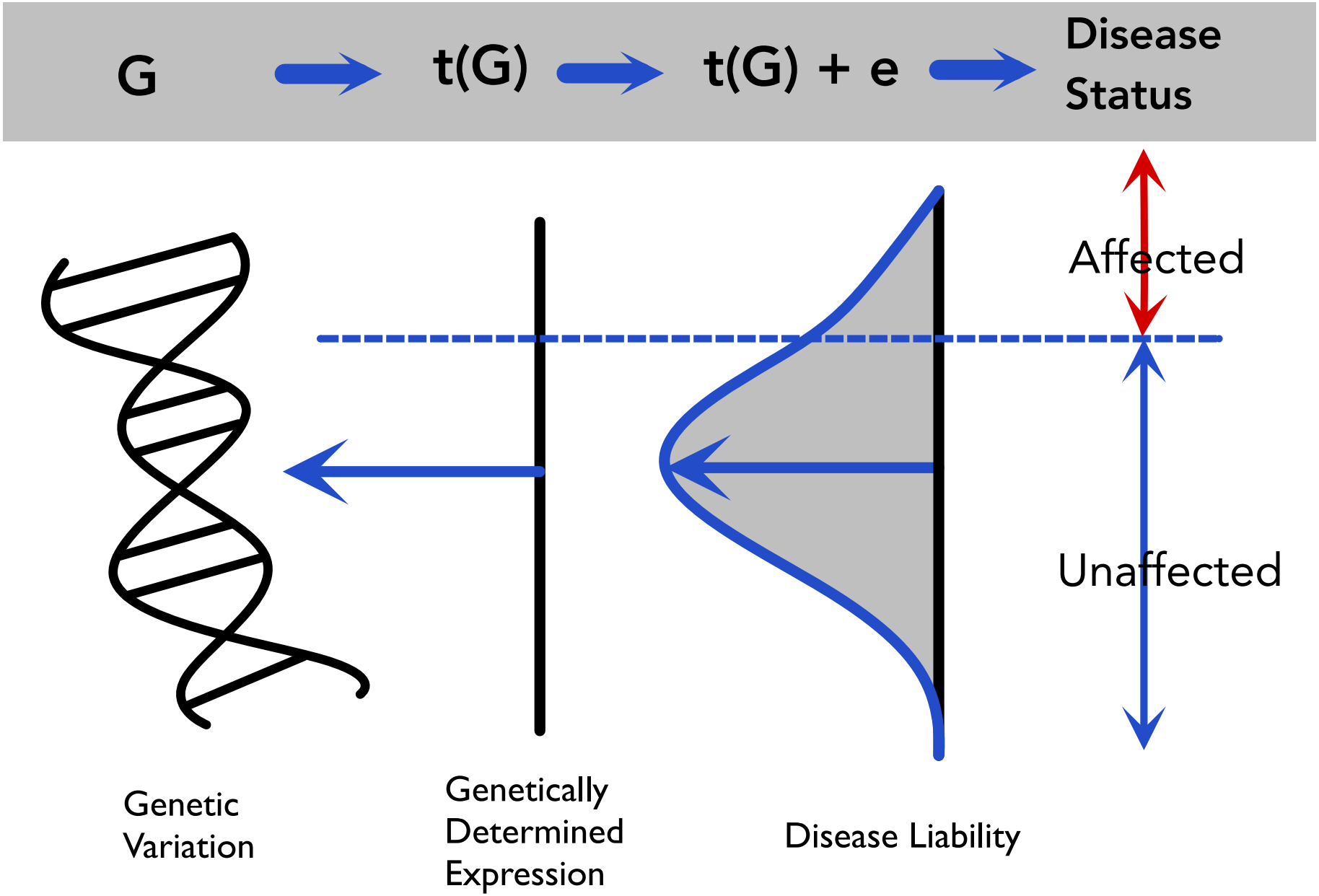
- Genes more attractive than genetic variants
  - A lot is know about their function
  - Follow up experiments can be easily devised
- Used extensively in whole exome studies to address low power of rare variants
- Limited success so far

Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS

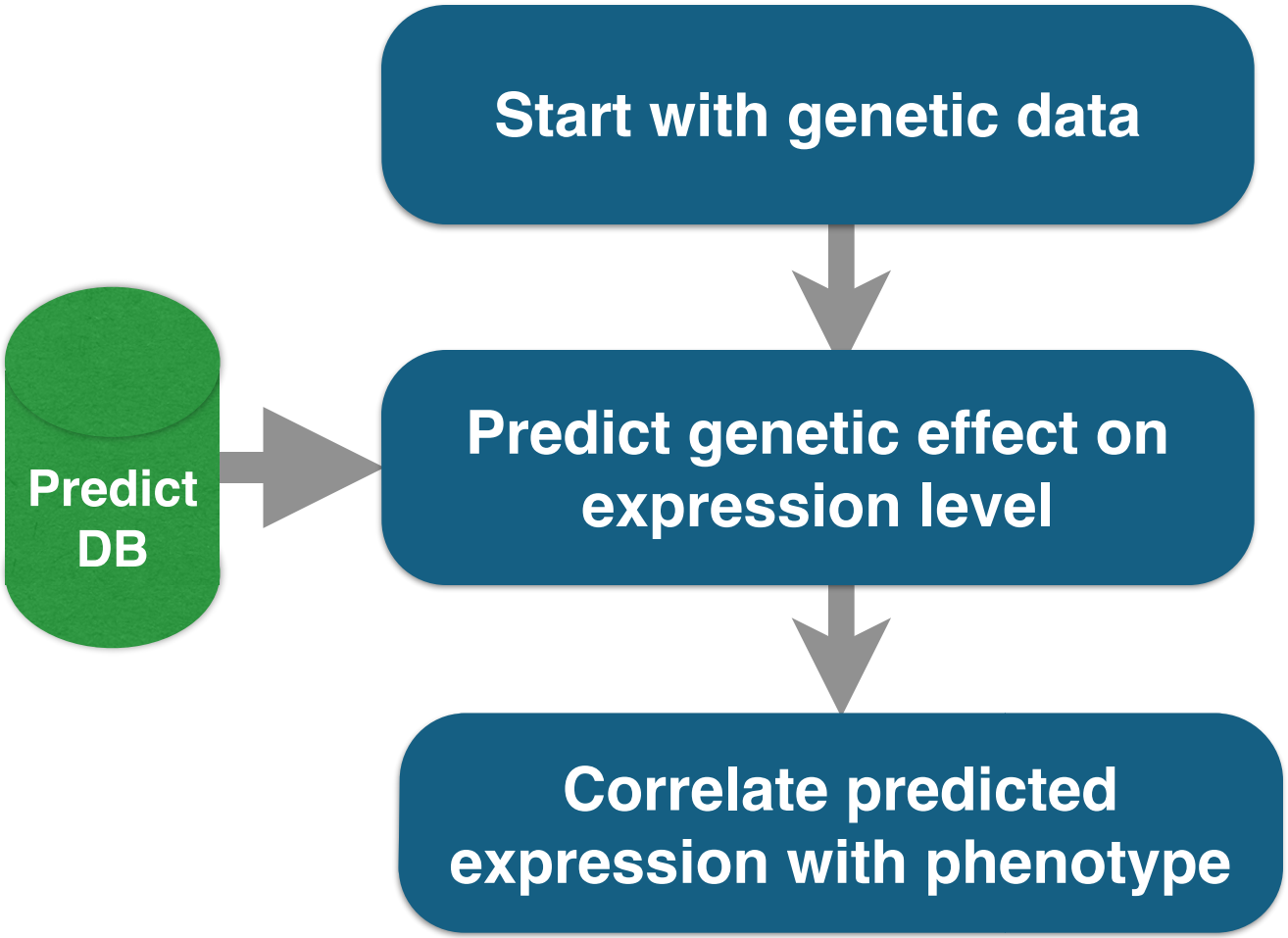
Dan L. Nicolae[1,2,3], Eric Gamazon[1], Wei Zhang[1], Shiwei Duan[1¤], M. Eileen Dolan[1,2], Nancy J. Cox[1,



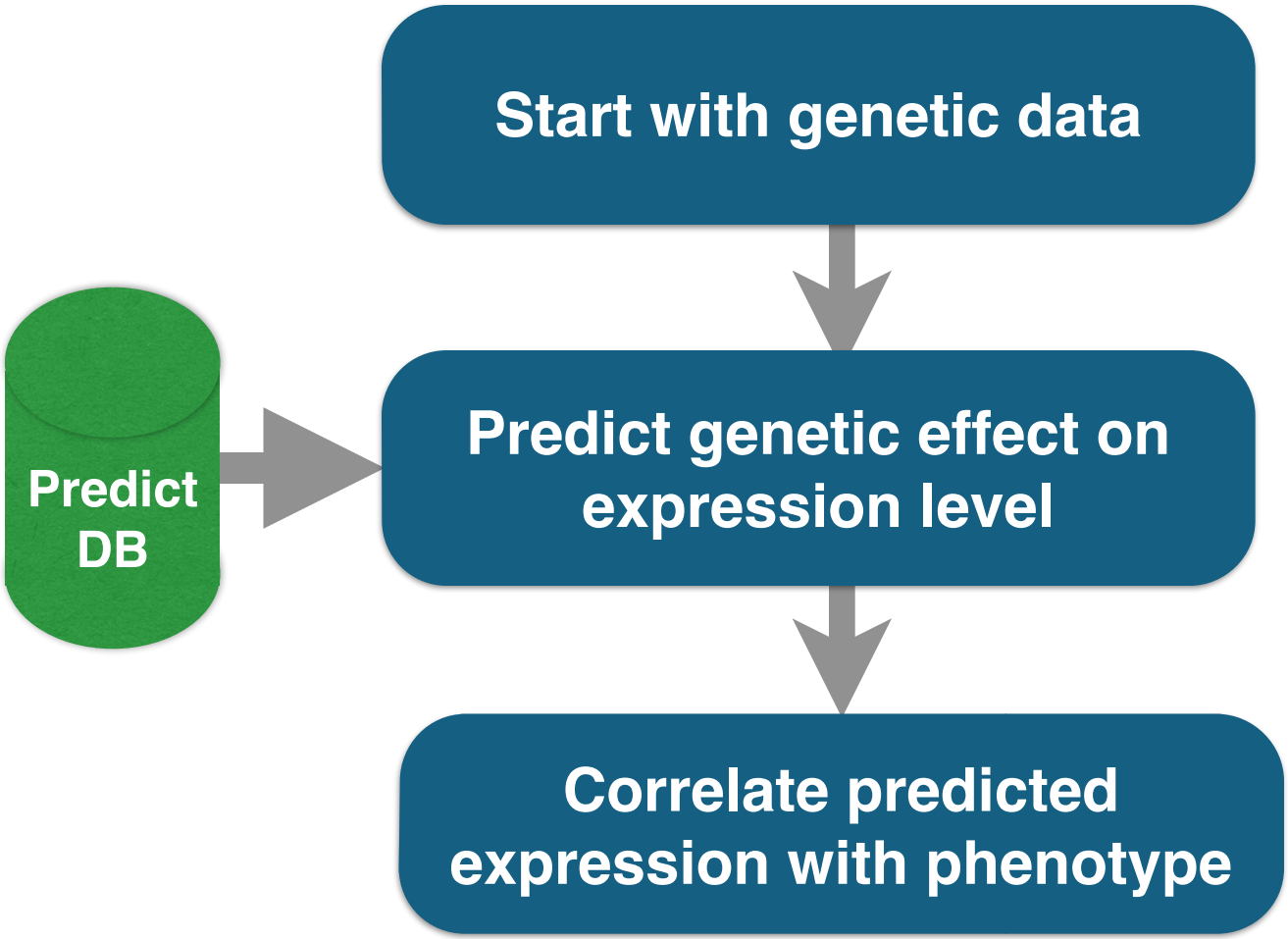
Genetic Control of Disease Through Gene Regulation



PrediXcan Flow



PrediXcan Flow



No transcriptome data is needed

Only genotype and phenotype data needed

Additive Model for Genetic Effect Prediction

Predicted Expression Trait

$$t_i = \sum_{k=1}^{M} w_k G_{ki}$$

t_i is predicted effect on gene expression level for individual i
G_{ki} number of reference alleles for SNP k and individual i
w_k weight for SNP k

PredictDB hosts prediction models

Simple Polygenic Model

- w_k = single variant regression coefficient (Matrix eQTL output)
- w_k set to zero if p value > 0.05 for cis SNPs (1Mb TSS)
- w_k set to zero if p value $> 10^{-6}$ for trans SNPs

Expression Data for Prediction Model Building

Predicted Expression Trait

- GTEx - Genotype of Tissue Expression
  - Large scale Common Fund project
  - 900 organ donors
  - 45 tissues
  - RNAseq, whole exome seq, whole genome seq

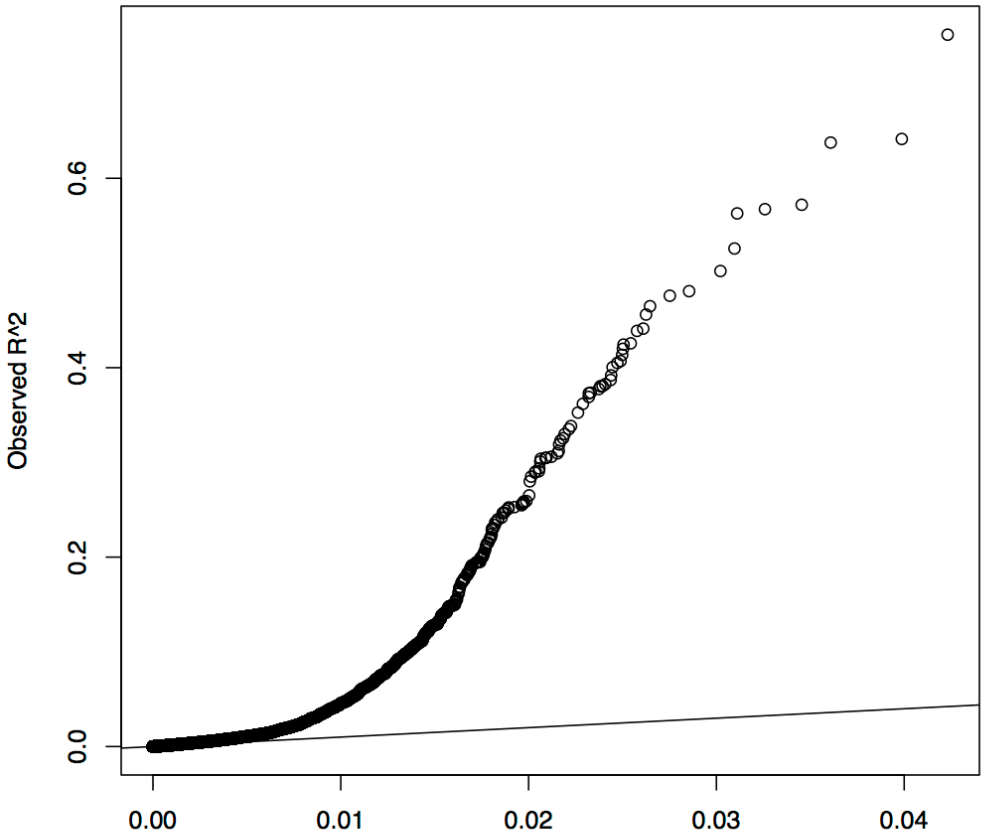
gEUVADIS

- RNAseq 462 individuals from the 1000 Genomes Project

Other array data

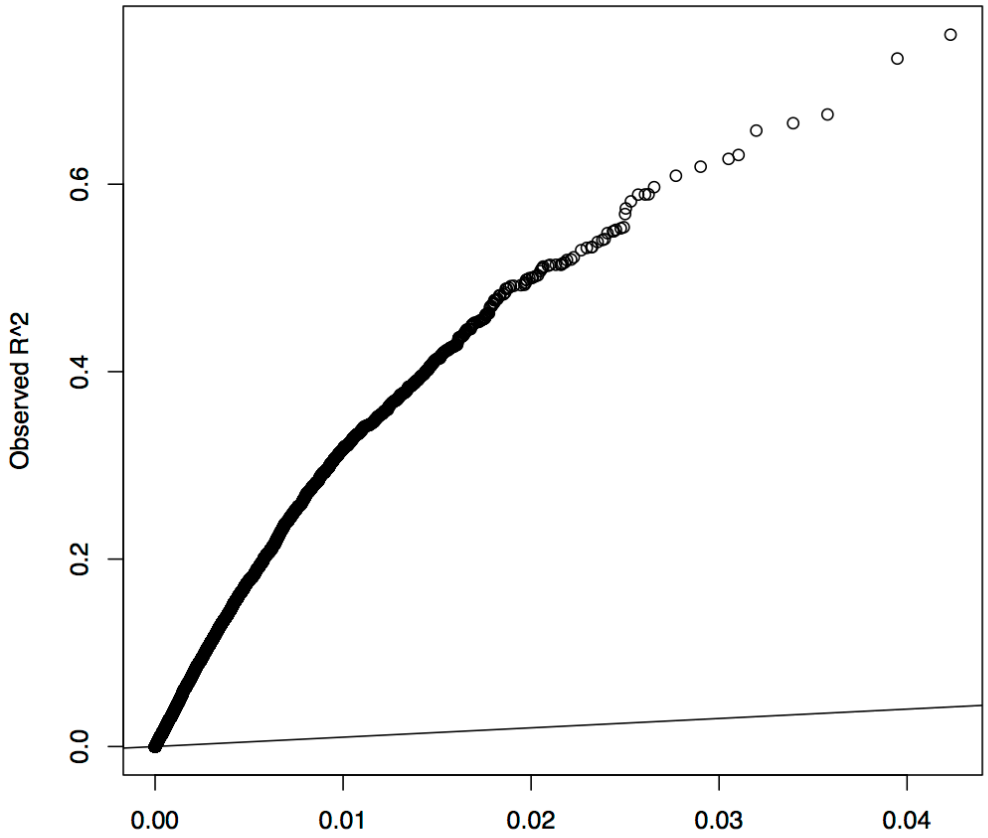
Good Prediction Performance

Prediction R^2



Training with GTEx Testing in 1K Genomes

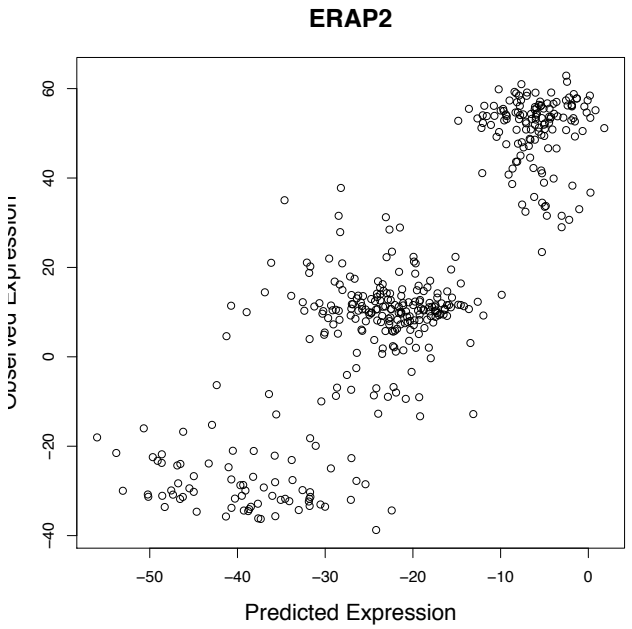
Replication R^2

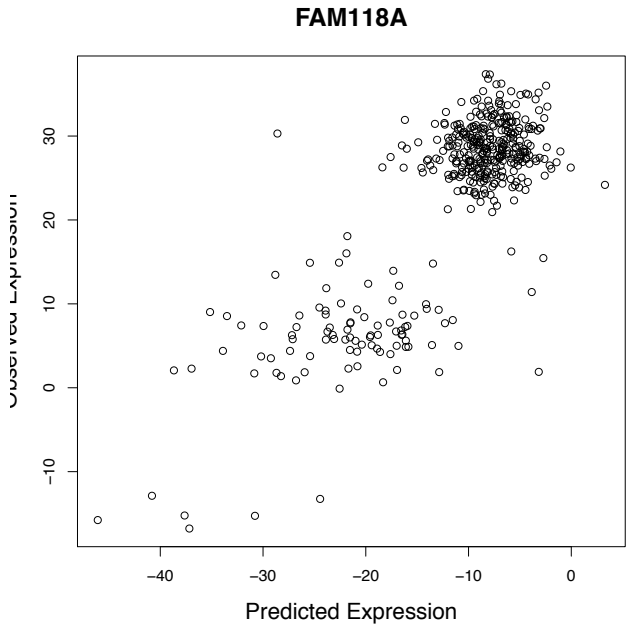


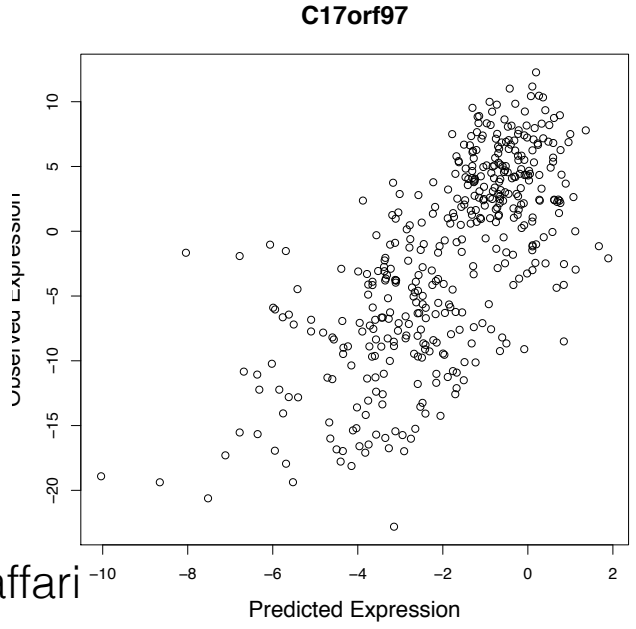
Replicate RNAseq Pickrell et al 2010 vs. 1K Genomes 2013

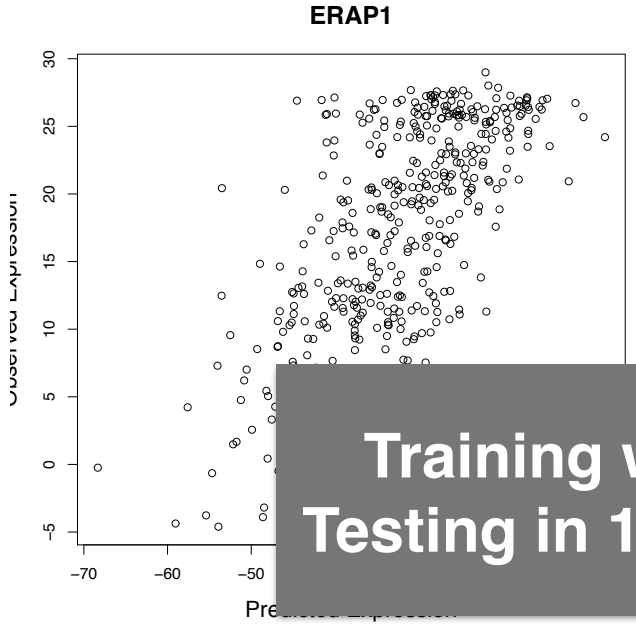
Sahar Mozaffari

Examples of Well Predicted Genes





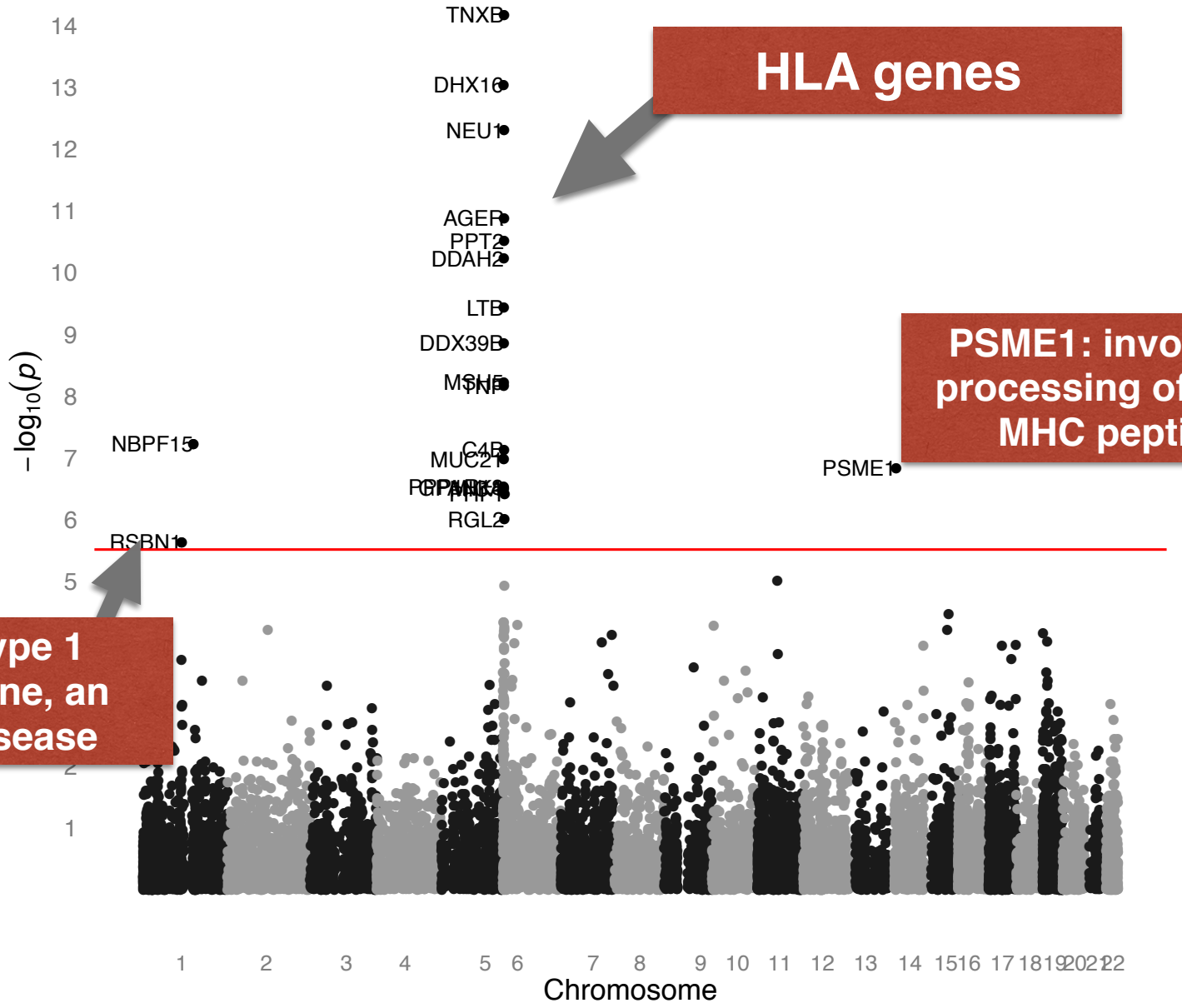




Sahar Mozaffari

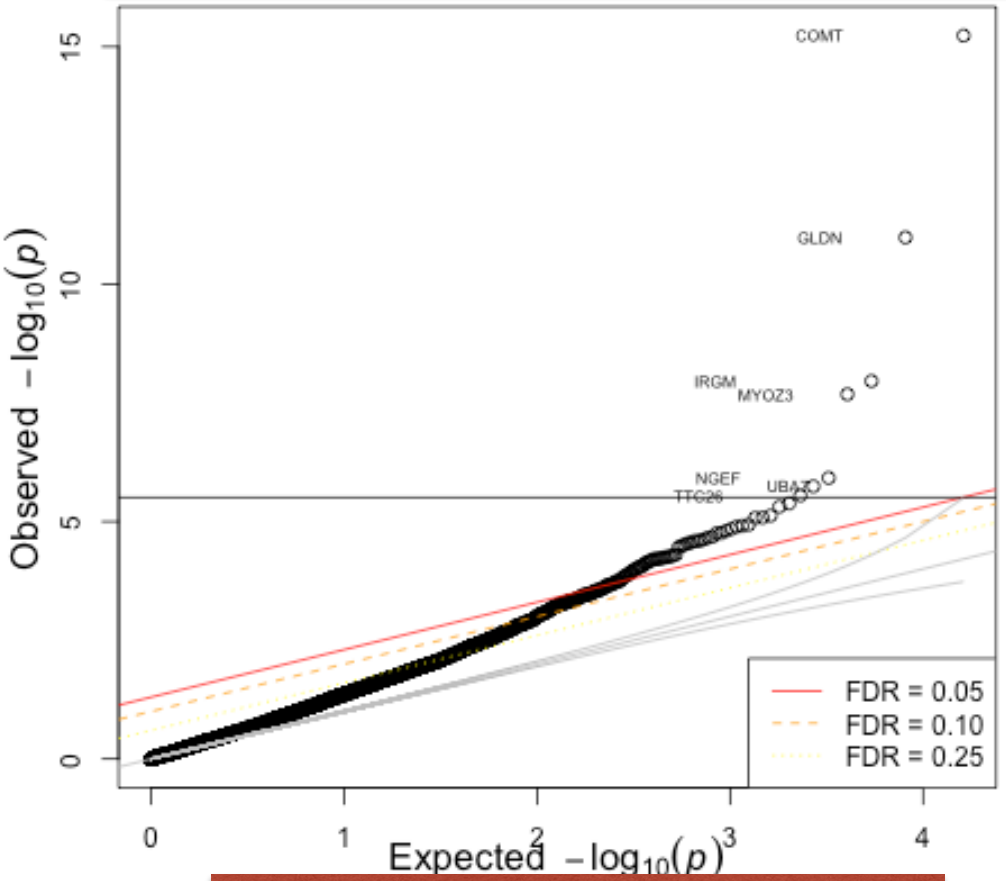
Training with GTEx
Testing in 1K Genomes

Genes Associated with Rheumatoid Arthritis



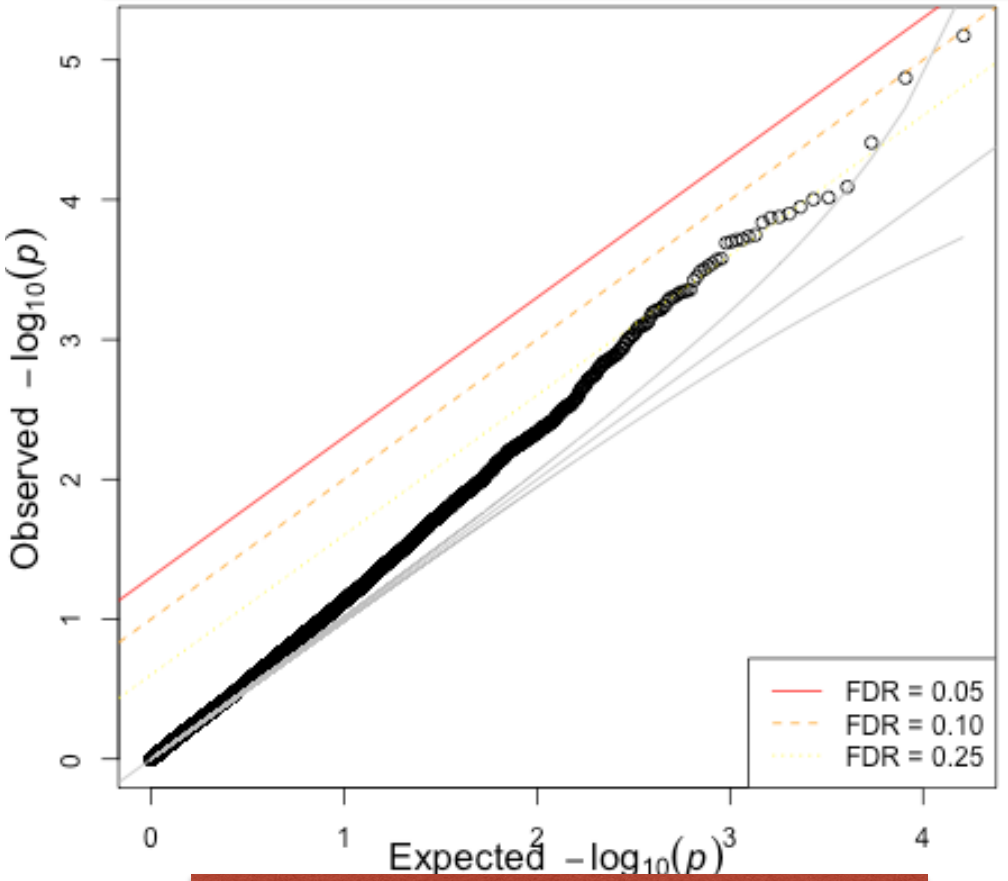
PrediXcan Results for Crohn's Disease and Hypertension

Crohn's Disease



IRGM is a known Crohn's gene

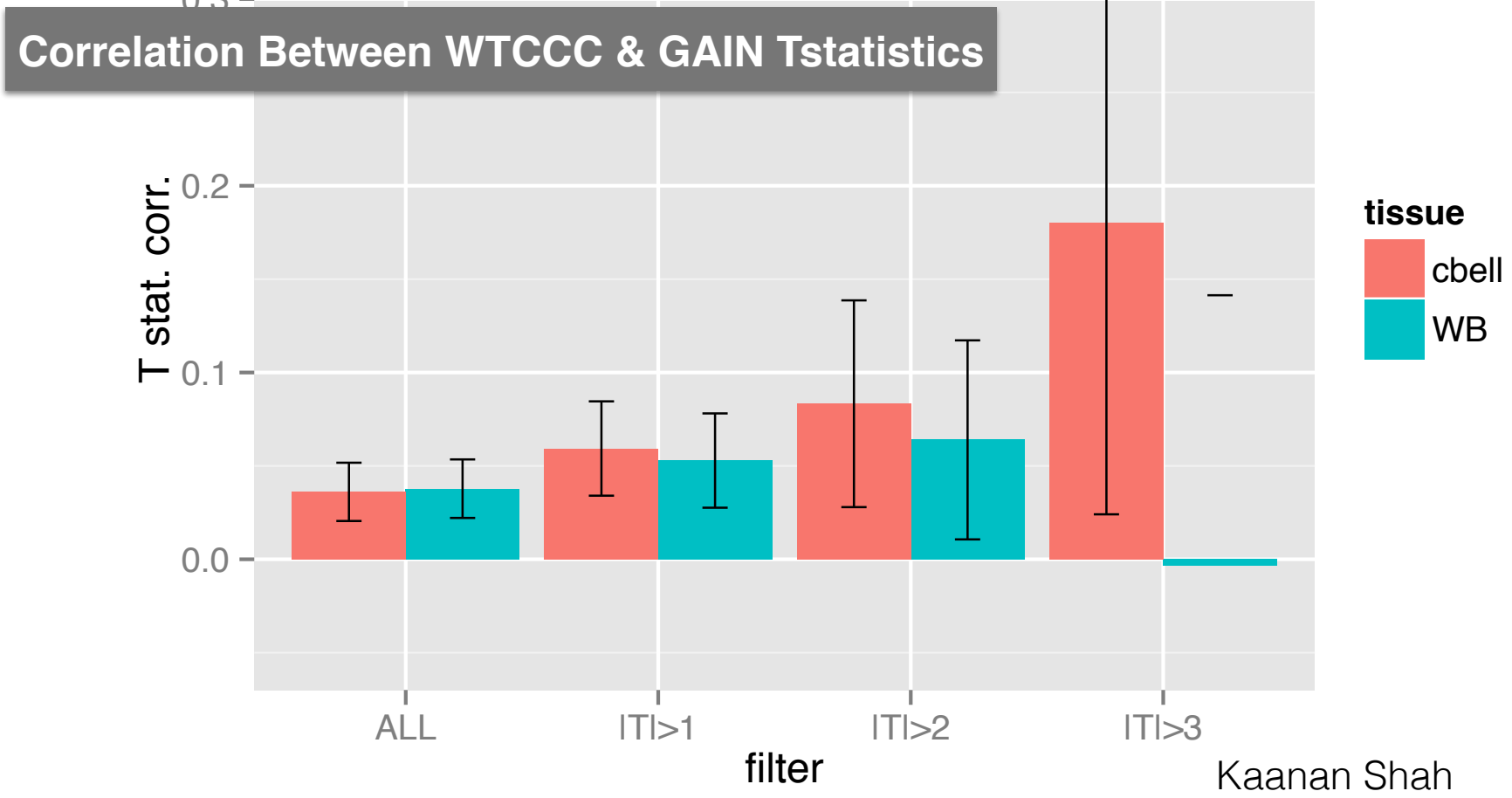
Hypertension



Whole blood may not be relevant tissue

Significant Concordance Between Independent Bipolar Studies

Higher correlation for cerebellum based predictions than whole blood based



PrediXcan: a Gene Discovery Approach

- PrediXcan is a powerful gene based association test
- It directly tests the molecular mechanism through which genetic variants affect phenotype
- Reduced multiple testing burden compared to single variant approach
- Unlike other gene based tests, it provides direction of effects
- Advantages relative to gene expression studies
  - Applicable to any GWAS datasets
    gene expression levels are predicted from genotype data
  - No reverse causality
    disease status does not affect germline DNA
  - Multiple Tissues can be evaluated
    tissue expressions are only needed to build prediction models

Extension to Other Molecular Phenotypes

- microRNA levels
- lncRNA levels
- Methylation status

Acknowledgements

Thank you!

Contributors

- Keston Aquino Michaels
- Heather Wheeler
- Nancy Cox
- Kaanan P. Shah
- Sahar Mozaffari
- Eric Gamazon
- GTEx Consortium

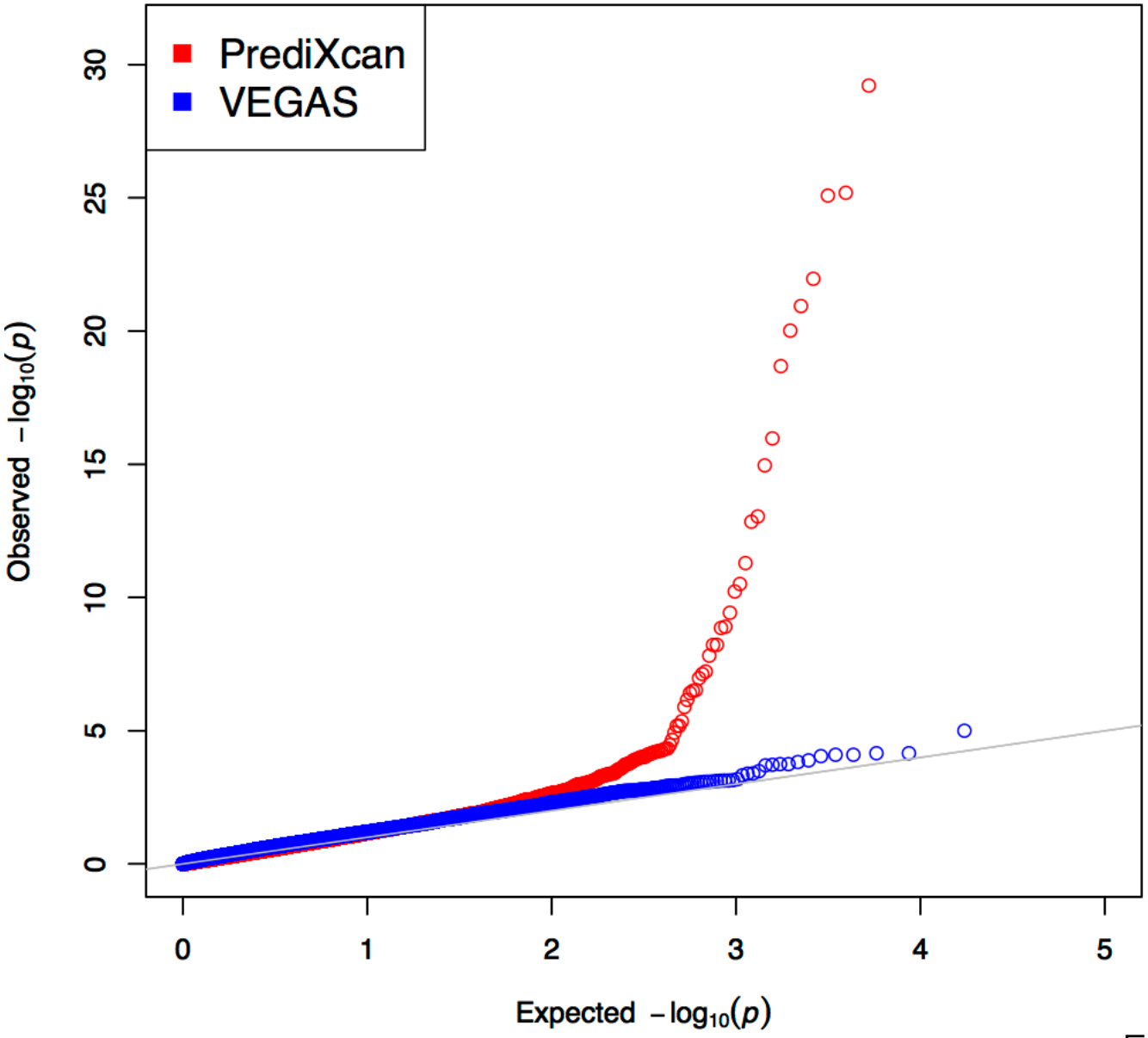
Data sources

- WTCCC
- GAIN Bipolar Disease

Funding

- HKI was funded in part by Uchicago CTSA NCI K12CA139160
- University of Chicago Diabetes Research and Training Center: P60 DK20595, P30 DK020595
- Genotype of Tissue Expression GTEx R01 MH090937 and R01 MH101820
- Pharmacogenomics of Anticancer Agents PAAR UO1GM61393
- Pharmacogenomics Research Network (PGRN) Statistical Analysis Resource (P-STAR) U19 HL065962
- Conte Center grant P50MH094267

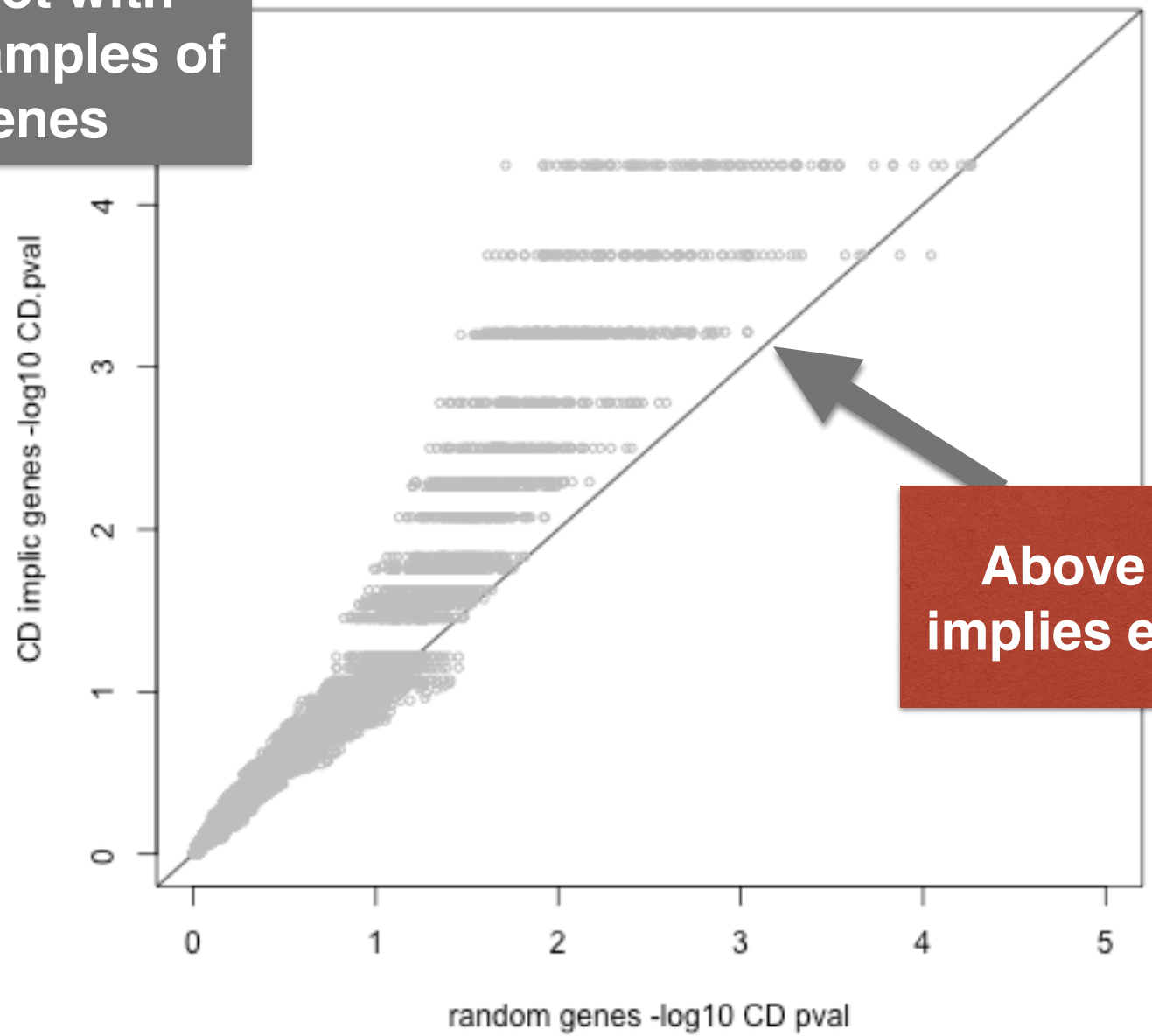
PrediXcan Outperforms VEGAS



Eric Gamazon

Enrichment of Known Crohn's Genes Among Findings

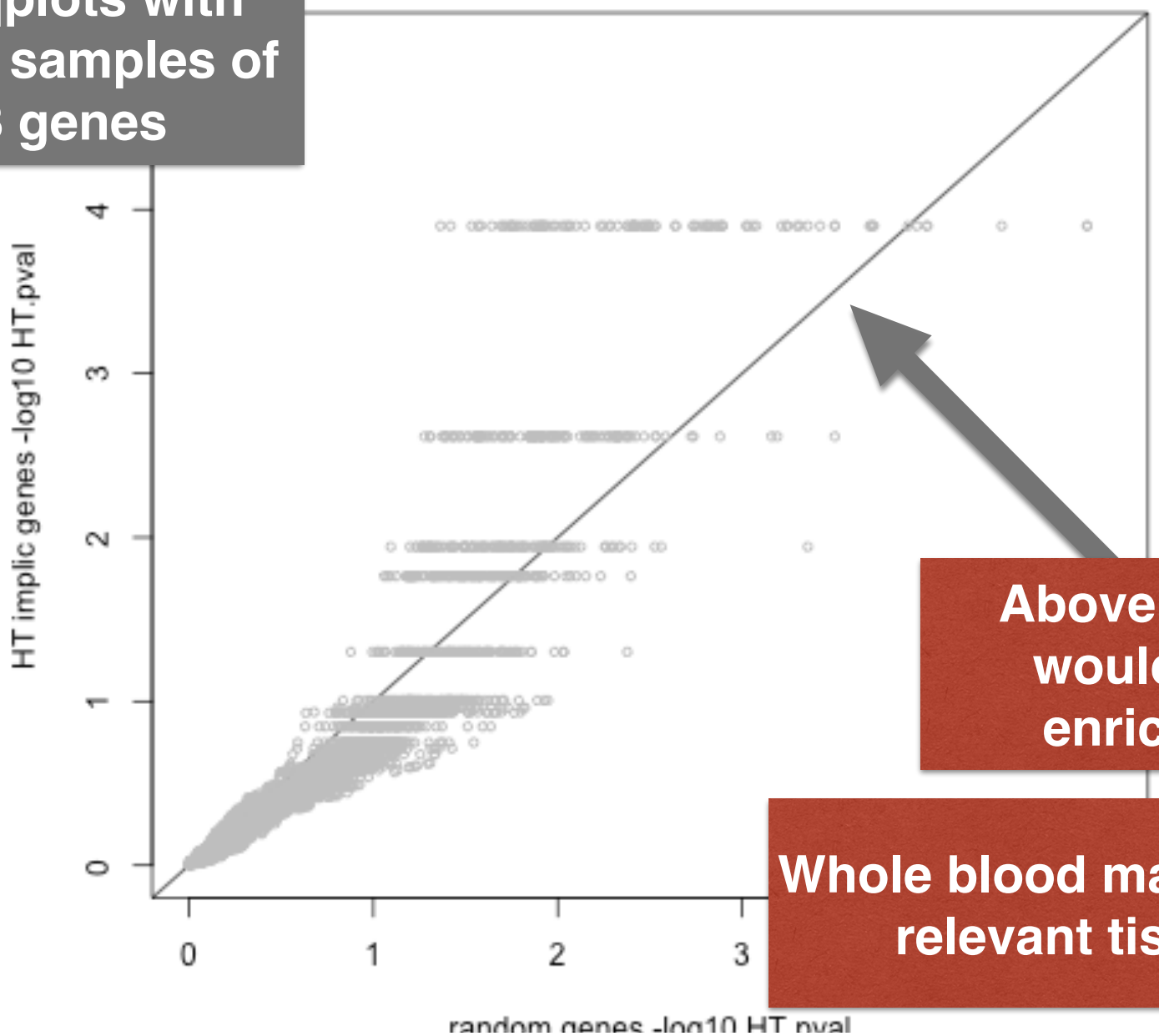
100 qqplot with random samples of 205 genes



Above this line implies enrichment

No Enrichment Among Hypertension Findings

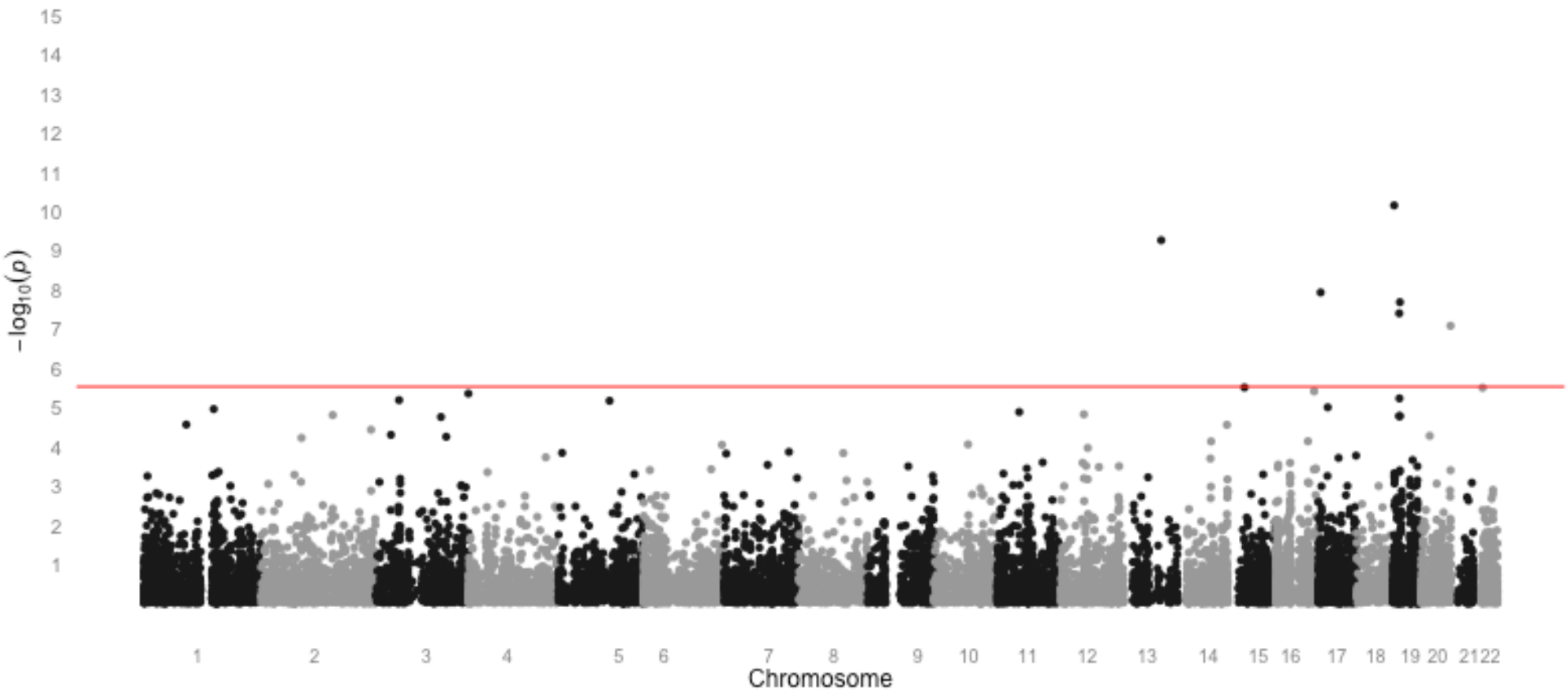
100 qqplots with random samples of 133 genes



Above this line would imply enrichment

Whole blood may not be relevant tissue

Bipolar Disorder WTCCC results



Bipolar Disorder Replication

- GAIN (n=2000) & WTCCC Bipolar Disorder (n=5000)
- Whole Blood

- Significant genes
  - RFNG (pmeta = 10–8, pGAIN = 2.5x10–6, pWTCCC = 0.00017) Modulator of Notch signaling
    Implicated in neurogenesis
  - LPHN1 (pmeta = 10–6, pGAIN = 0.36, pWTCCC = 2x10–8)
    Receptor for TENM2 that mediates heterophilic synaptic cell-cell contact and postsynaptic specialization
    Candidate gene for mental disorder based on mouse model phenotypes

Kaanan Shah