The Role of Genomic Prediction in Precision Medicine

Hae Kyung Im, PhD

THE UNIVERSITY OF CHICAGO

March 6, 2015

Overview

- Precision Medicine Initiative
- Monogenic vs. polygenic traits
- Review of prediction methods
- Poly-Omic integration: OmicKriging
- Role of regulatory variants in complex traits
- PrediXcan
- Prediction of gene expression traits

Precision Medicine

- Obama: Precision Medicine Initiative $215M for 2016 Budget
- Instead of “one-size fits-all-approach”
- “Right treatment, at the right time to the right person”
- Innovative approach to disease prevention and treatment based on individual differences in genes, environments, and lifestyles

http://www.whitehouse.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative

Precision Medicine Implementation

Prediction

- Disease
  - risk stratification
  - intervention strategies
- Adverse events
- Efficacy of treatment

Dissection

- Etiology of complex traits
- Mechanism by which genetic variation drives phenotypic variation
- Druggable targets

The Promise of the Human Genome Sequencing Project

- In year 2000, president Clinton announced the completion of the first draft of the human genome, which would "revolutionize the diagnosis, prevention, and treatment of most, if not all, human diseases.
- Francis Collins predicted that diagnosis of genetic diseases would be accomplished by 2010 and that treatments would start to roll out perhaps by 2015.
- Why are we not there yet?

The Promise of the Human Genome Sequencing Project

- In year 2000, president Clinton announced the completion of the first draft of the human genome, which would "revolutionize the diagnosis, prevention, and treatment of most, if not all, human diseases.
- Francis Collins predicted that diagnosis of genetic diseases would be routinely established by 2010 and that treatments would start to roll out perhaps by 2015.
- Why are we not there yet?

Prediction is hard !

The Promise of the Human Genome Sequencing Project

- In year 2000, president Clinton announced th ion of the first draft of the human genom utionize the diagnosis, preventio ll, human disease
- F sis of genetic diseases w 2010 and that treatments would start to ps by 2015.
- Why are we not there yet?

Genetic architecture is much more complex than anticipated

Monogenic vs. Polygenic Architecture

Genetic Architecture of Complex Traits

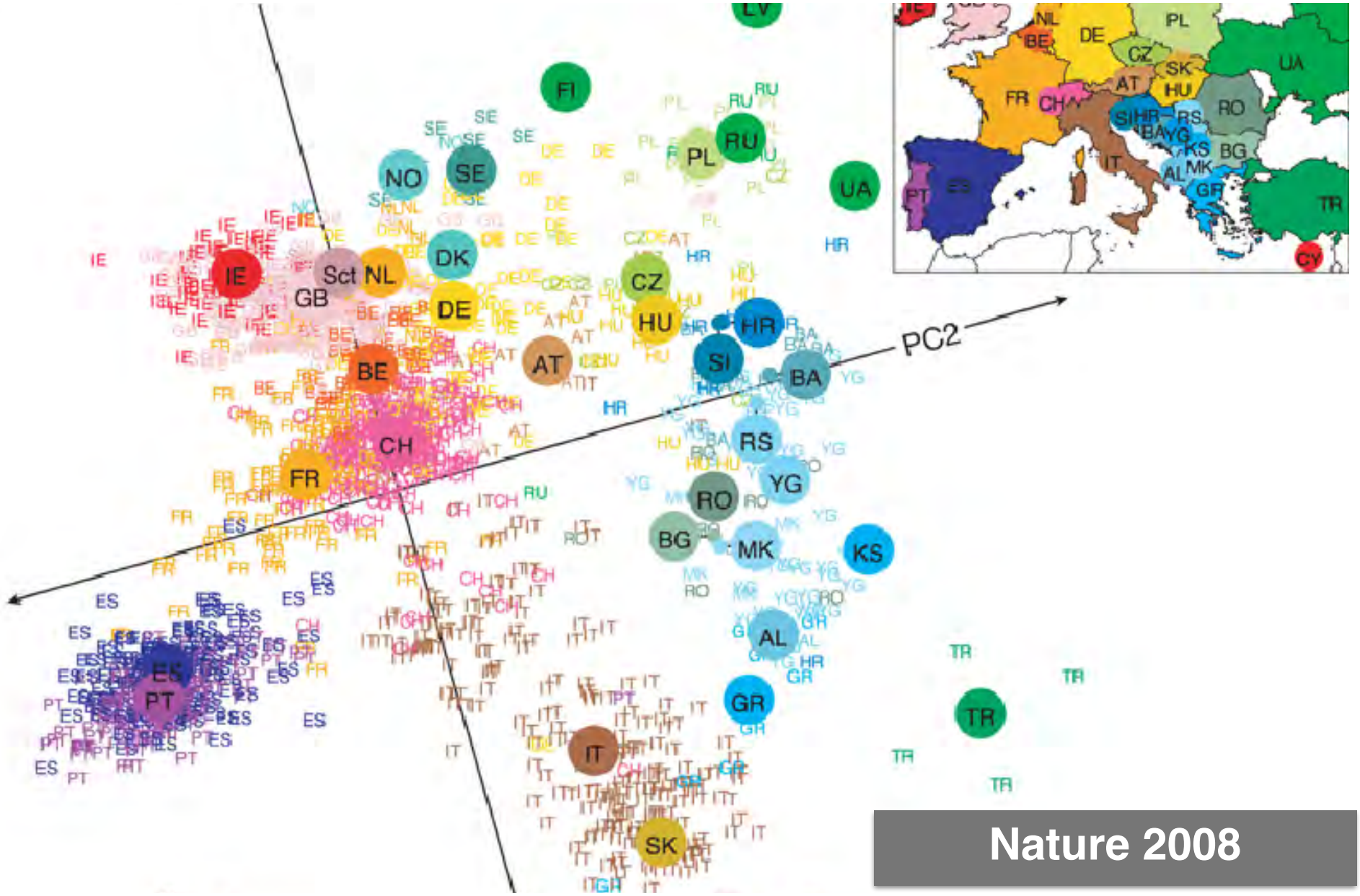
Genetic Architecture of Complex Traits

Genetic Architecture of Complex Traits

Single Variants Not Relevant for Highly Polygenic Traits

Genes mirror geography within Europe

John Novembre[1,2], Toby Johnson[4,5,6], Katarzyna Bryc[7], Zoltán Kutalik[4,6], Adam R. Boyko[7], Adam Auton[7], Amit Indap[7], Karen S. King[8], Sven Bergmann[4,6], Matthew R. Nelson[8], Matthew Stephens[2,3] & Carlos D. Bustamante[7]

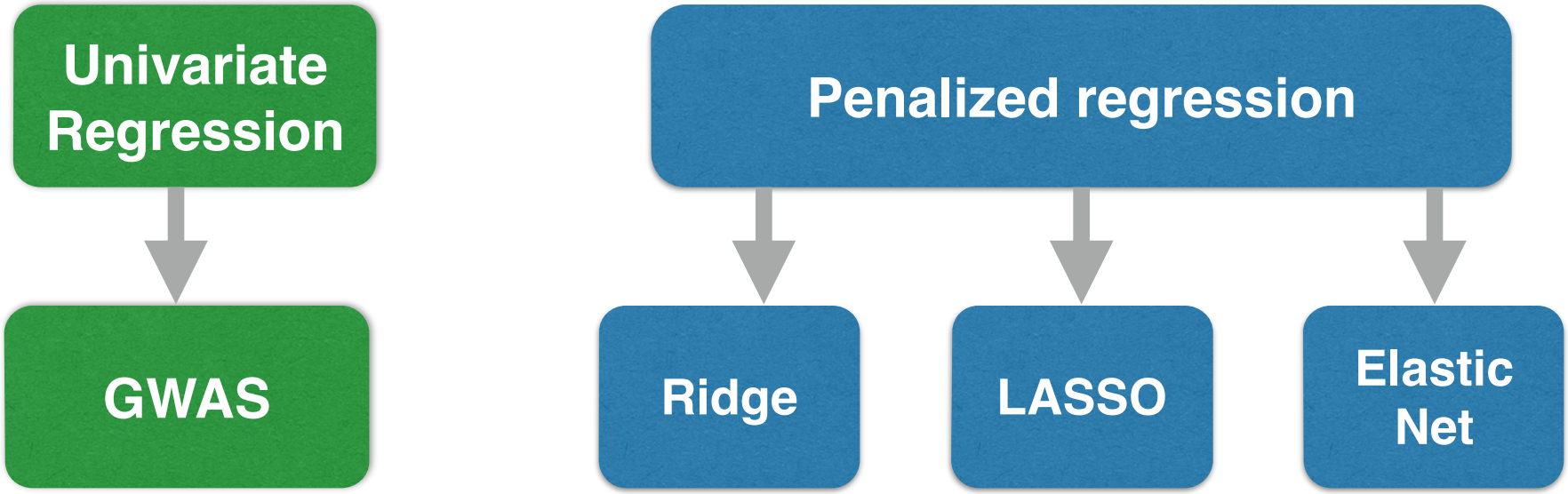


Nature 2008

Whole Genome Prediction Methods

Additive Genetic Model

$$Y = \sum_{k=1}^{M} \beta_k X_k + \epsilon$$



$$\|Y - X_k\beta_k\|_2$$

$$\|Y - \sum_k X_k\beta_k\|_2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta_2\|_2$$

Simple Polygenic Score

LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Nature 2009

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\mathrm{GWAS}} X_k$$

Univariate Regression

GWAS

Best Linear Unbiased Prediction (BLUP)/Ridge

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

AJHG 2011

$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\text{Ridge}} X_k$$

Penalized regression

Ridge

$$\|Y - \sum_k X_k \beta_k\|_2 + \lambda_2 \|\beta_2\|_2$$

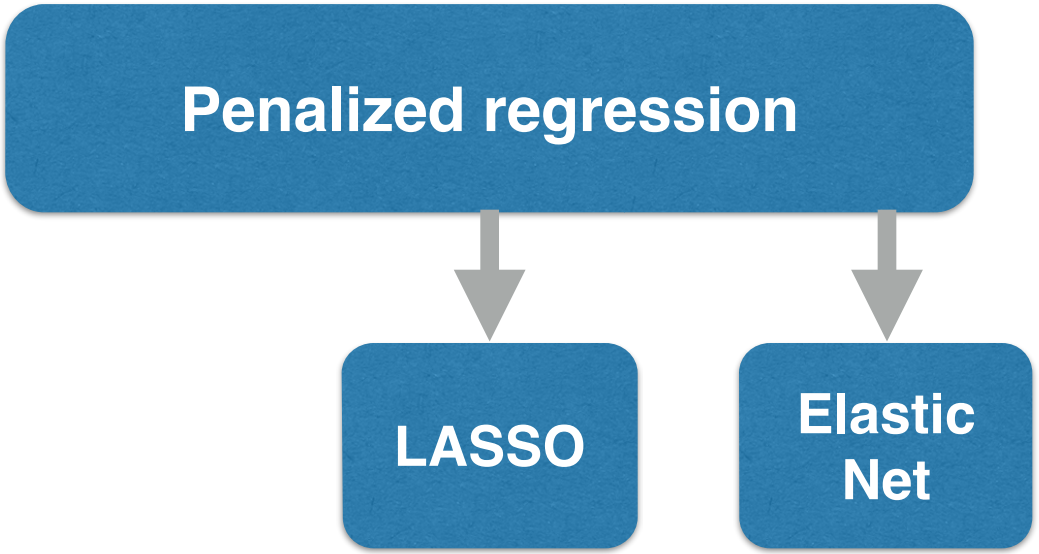
LASSO/Elastic Net Prediction

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie

Stanford University, USA



$$Y = \sum_{k=1}^{M} \hat{\beta}_k^{\text{E-N}} X_k$$

$$\|Y - \sum_k X_k \beta_k\|_2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta_2\|_2$$

Whole Genome Prediction Approaches

OPEN ACCESS Freely available online

PLOS | GENETICS

Polygenic Modeling with Bayesian Sparse Linear Mixed Models

Xiang Zhou[1]*, Peter Carbonetto[1], Matthew Stephens[1,2]*

$$Y = \sum_{k=1}^{M} \beta_k^L X_k + \sum_{k=1}^{M} \beta_k^S X_k + \epsilon$$

$$\beta_k^L \sim N(0, \sigma_L^2)$$

$$\beta_k^S \sim N(0, \sigma_S^2)$$

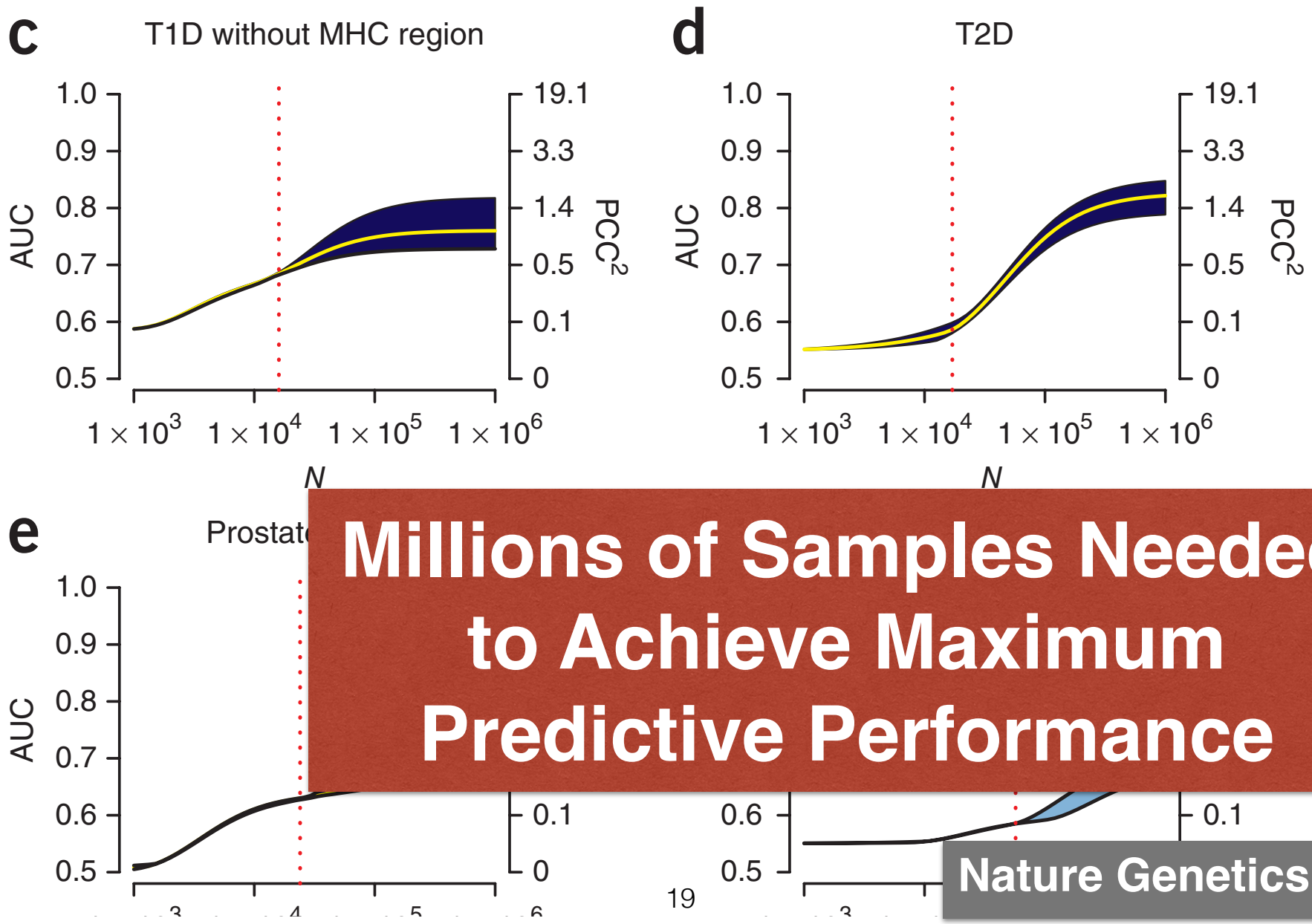
MultiBLUP: improved SNP-based prediction for complex traits

Doug Speed and David J Balding

Genome Res. published online June 24, 2014
Access the most recent version at doi:10.1101/gr.169375.113

Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies

Nilanjan Chatterjee[1], Bill Wheeler[2], Joshua Sampson[1], Patricia Hartge[1], Stephen J Chanock[1] & Ju-Hyun Park[1,3]



Millions of Samples Needed to Achieve Maximum Predictive Performance

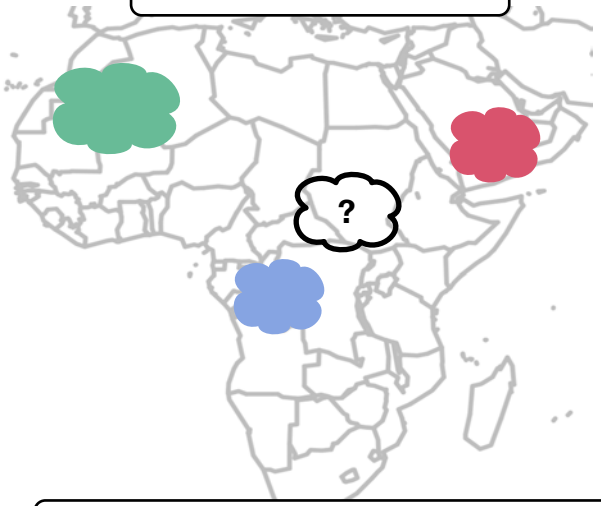
Nature Genetics 2013

OmicKriging: Integration of Multiple Omics Data

Genetic Epidemiology 2014

What is Kriging?

Physical variable
(ex. rainfall)



Prediction by kriging

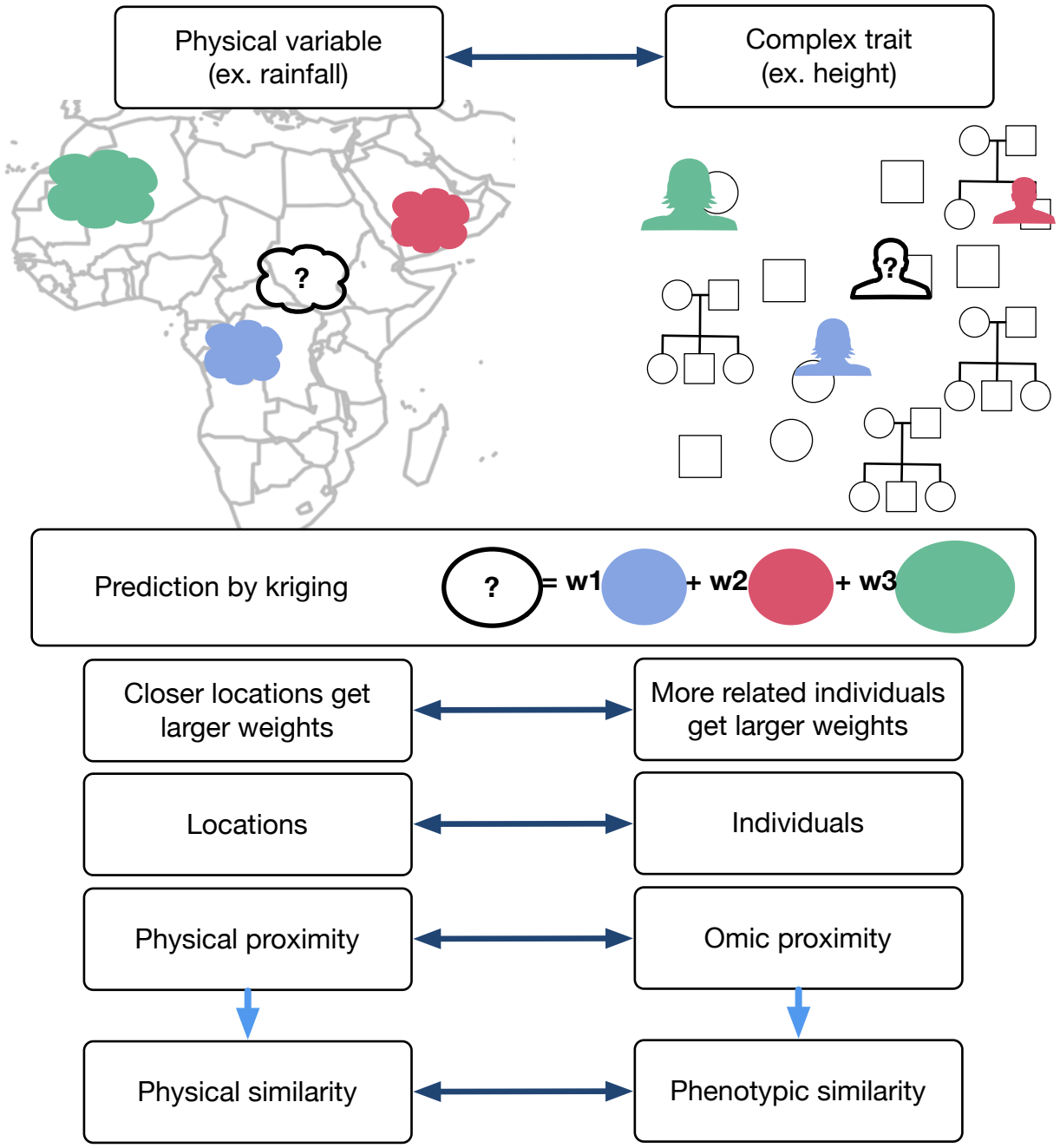
Closer locations get
larger weights

Locations

Physical proximity

Physical similarity

What is Kriging?



Kriging

Predicted Y is the weighted average of the observations

$$\text{Prediction}(Y_{\text{new}}) = \omega_1 Y_1 + \omega_2 Y_2 + \cdots + \omega_n Y_n$$

$$\omega_i = \text{function}(\text{all } n(n+1)/2 \text{ pairs of correlations})$$

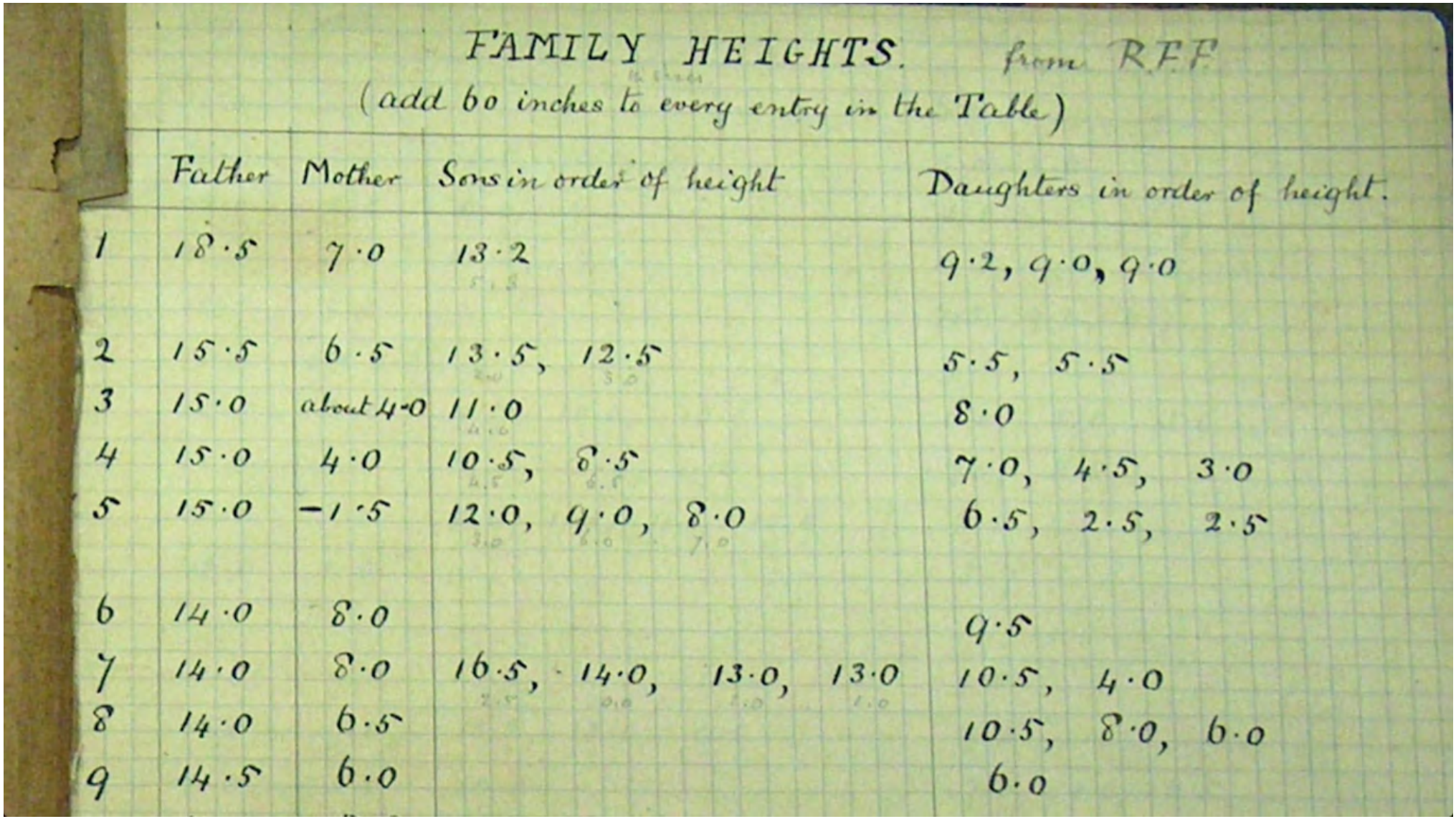
Without covariates

$$\boldsymbol{\omega}' = \boldsymbol{\rho}' \boldsymbol{\Sigma}^{-1}$$

$\boldsymbol{\rho}$ the correlation between the new value and the observed values and

$\boldsymbol{\Sigma}$ the correlation matrix of the observations.

Galton's Height Data



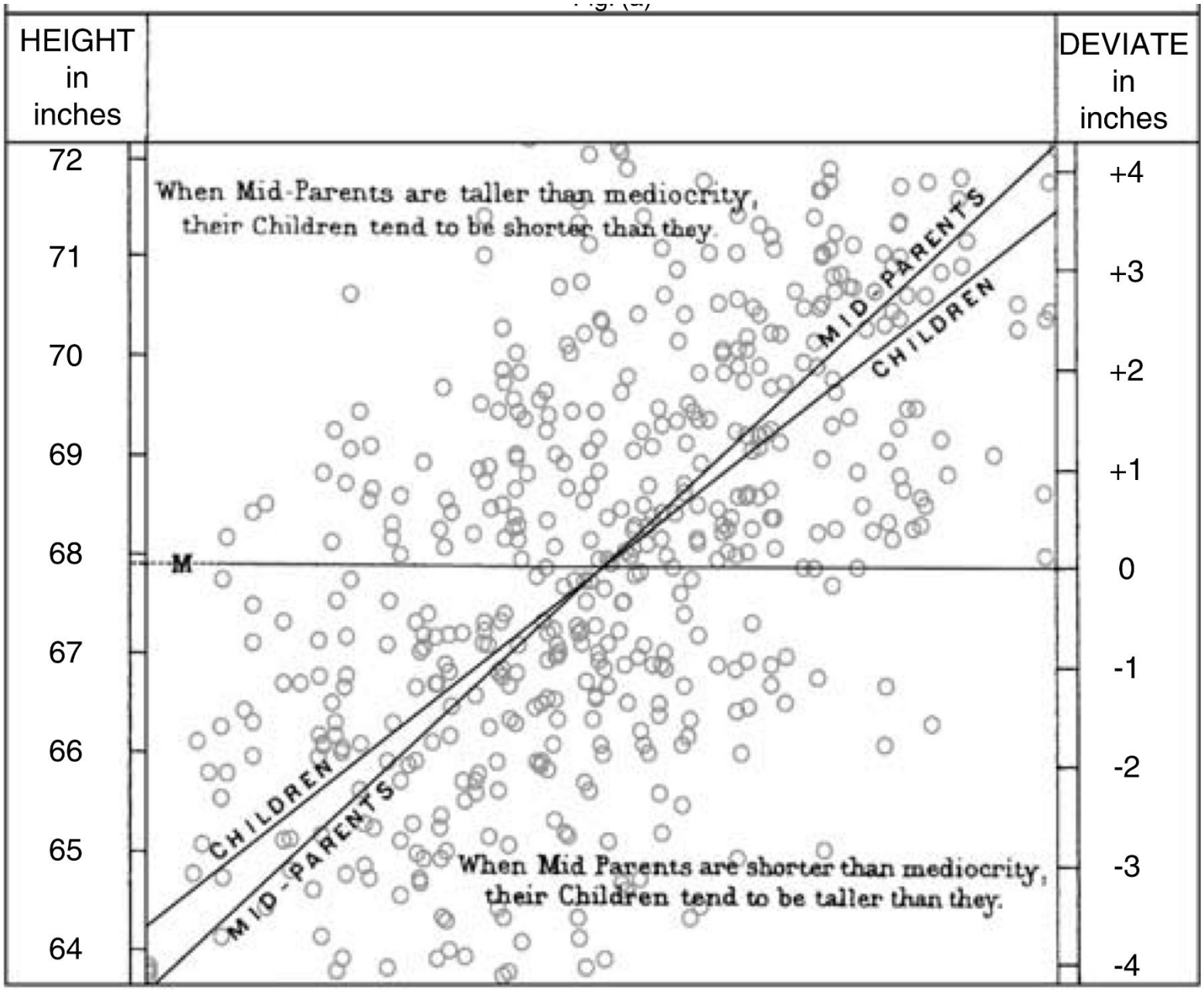
FAMILY HEIGHTS. from R.F.F

(add 60 inches to every entry in the Table)

	Father	Mother	Sons in order of height	Daughters in order of height.
1	18·5	7·0	13·2	9·2, 9·0, 9·0
2	15·5	6·5	13·5, 12·5	5·5, 5·5
3	15·0	about 4·0	11·0	8·0
4	15·0	4·0	10·5, 8·5	7·0, 4·5, 3·0
5	15·0	−1·5	12·0, 9·0, 8·0	6·5, 2·5, 2·5
6	14·0	8·0		9·5
7	14·0	8·0	16·5, 14·0, 13·0, 13·0	10·5, 4·0
8	14·0	6·5		10·5, 8·0, 6·0
9	14·5	6·0		6·0

Hanley JA: Transmuting Women into Men. The American Statistician 2004, 58:237243.

Galton Was Kriging with Kinship Matrix (1885)



Kriging = BLUP (Best Linear Unbiased Prediction)

- Galton (1885): parent to offspring
- Fisher (1918) and Wright (1921): pedigree
- Formalized by Henderson (1950,1975) and Goldberger (1962)
- G-BLUP: genetic relatedness estimated using genotype
- BLUP/Kriging can be interpreted as the posterior mean of the genetic component given observations (Y = G + error)

Kriging = BLUP (Best Linear Unbiased Prediction)

- Galton (1885): parent to offspring
- Fisher (1918) and Wright (1921): pedigree
- Formalized by Henderson (1950,1975) and Goldberger (1962)
- G-BLUP: genetic relatedness estimated using genotype
- BLUP/Kriging can be interpreted as the posterior mean of the genetic component given observations (Y = G + error)

BLUP/Kriging translates genetic similarity into phenotypic prediction

Polyomic Model

$$Y_i = a + G_i + T_i + O_i + \epsilon_i$$

$$G_i = \sum_{l=1}^{M} \beta_l^G X_{il}^G \qquad \text{genetic component}$$

$$T_i = \sum_{l=1}^{L} \beta_l^T X_{il}^T \qquad \text{transcriptomic component}$$

$$O_i = \sum_{l=1}^{L'} \beta_l^O X_{il}^O \qquad \text{other omic component}$$

$$(\beta_G, \beta_T, \beta_O)' \sim \mathrm{N}(0, \Sigma_\beta)$$

Optimal Similarity Matrix

$$Y_i = a + G_i + T_i + O_i + \epsilon_i$$

Assuming independence of β's

$$\Sigma_{i,j} = \theta_G \sum_{l=1}^{M} X_{il}^G X_{jl}^G + \theta_T \sum_{l=1}^{L} X_{il}^T X_{jl}^T + \theta_O \sum_{l=1}^{L'} X_{il}^O X_{jl}^O + \theta_\epsilon \delta_{ij}$$

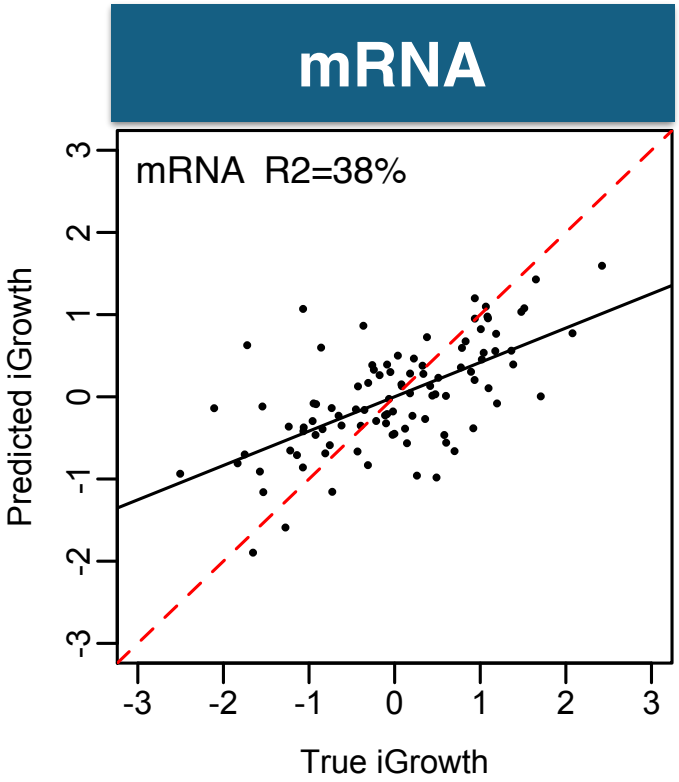
More generally

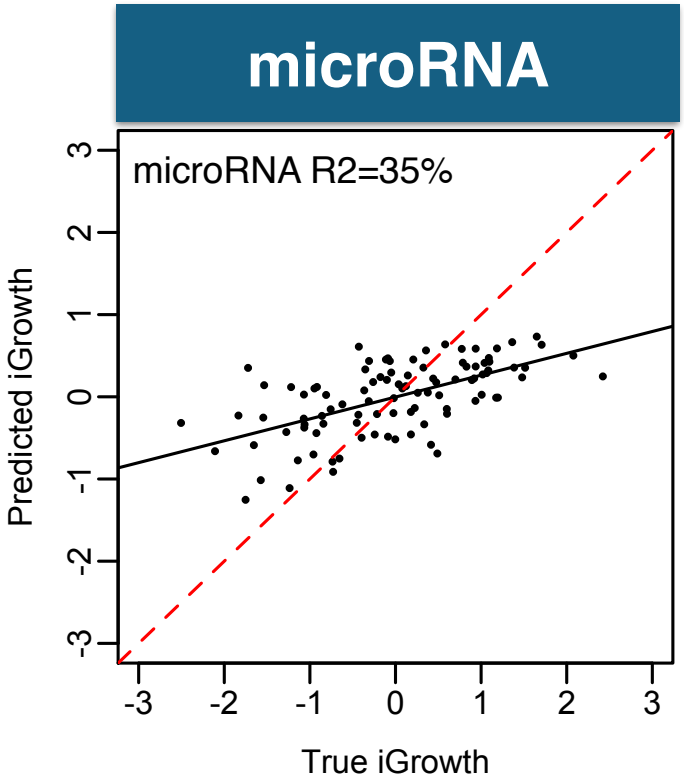
$$\Sigma_{i,j} = \theta_G \sum_{l=1}^{M} X_{ik}^G X_{jk}^G + \theta_T \sum_{l=1}^{L} X_{ik}^T X_{jk}^T + \theta_O \sum_{k=1}^{L'} X_{ik}^O X_{jk}^O + \theta_\epsilon \delta_{ij} + \sum_{k \neq l} \mathrm{cov}(\beta_k, \beta_l) X_{ik} X_{jl}$$

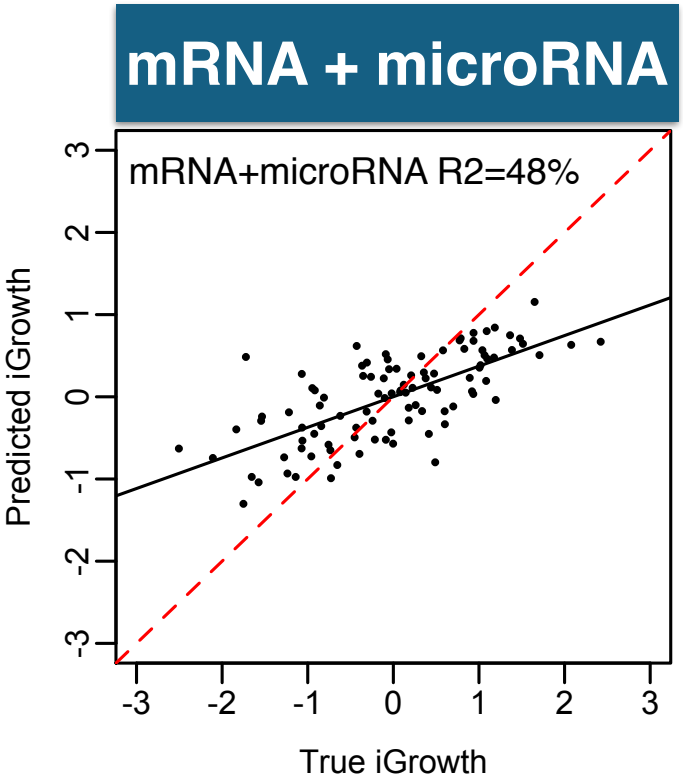
Application of OmicKriging to Cellular Growth

- Intrinsic cellular growth phenotype (Im et al 2012 PLoS Genetics)
- Genes associated with iGrowth are prognostic of survival in cancer patients
- Multiple omics data
  - 99 HapMap cell lines (CEU and YRI)
  - Genotype, mRNA, microRNA

Application of OmicKriging to Cellular Growth



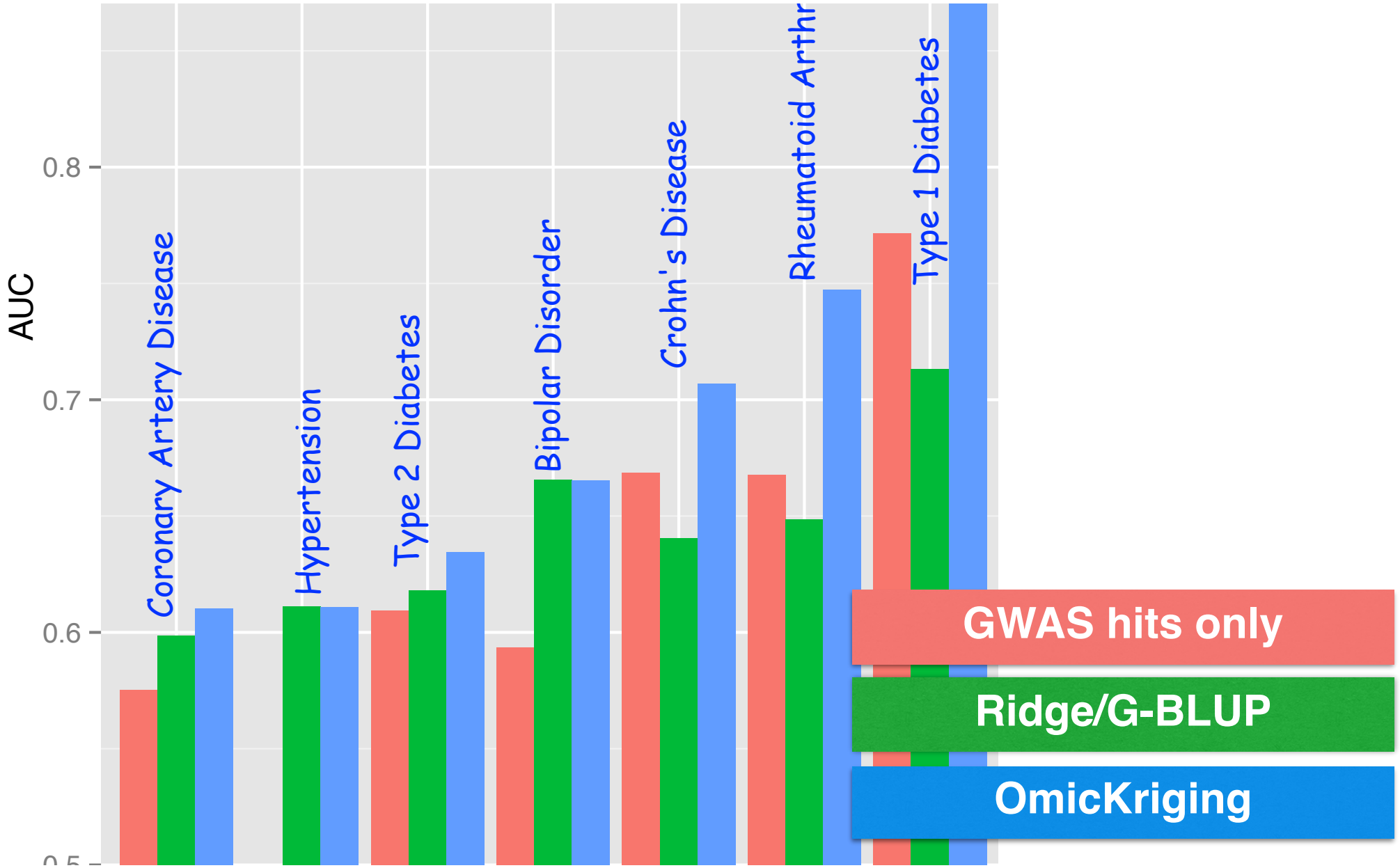




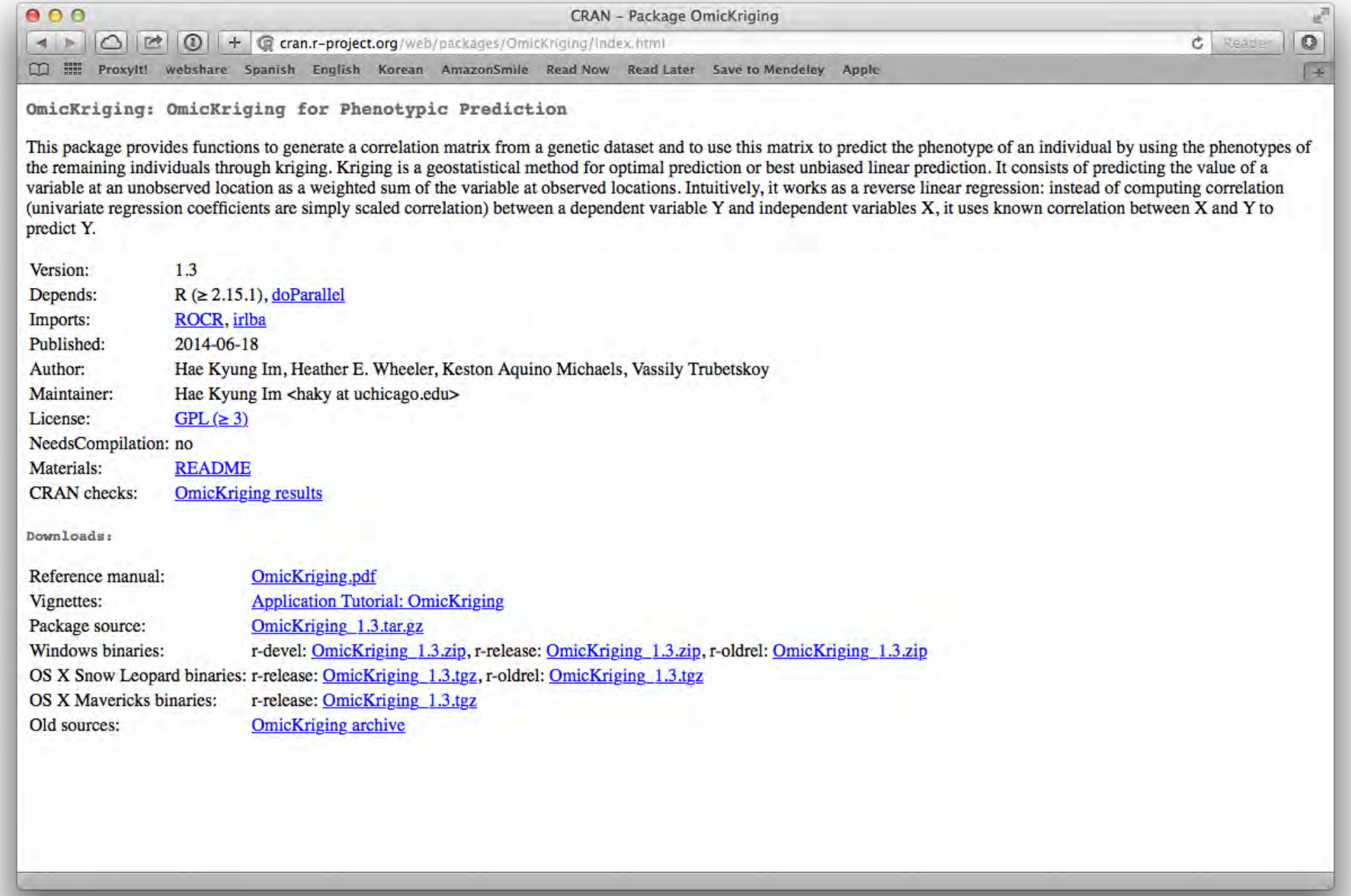
Application to Wellcome Trust Case Control Consortium

- WTCCC
- 7 disease cases and 2 control sets:
  - Coronary Artery Disease (2000)
  - Hypertension (2000)
  - Type 2 Diabetes (2000)
  - Bipolar Disorder (2000)
  - Crohn's Disease (2000)
  - Rheumatoid Arthritis (2000)
  - Type 1 Diabetes (2000)
  - 1958 Birth Cohort (1500)
  - UK National Blood Services (1500)

GWAS hits vs. Whole Genome Prediction (OmicKriging)



OmicKriging R Package



OmicKriging: OmicKriging for Phenotypic Prediction

This package provides functions to generate a correlation matrix from a genetic dataset and to use this matrix to predict the phenotype of an individual by using the phenotypes of the remaining individuals through kriging. Kriging is a geostatistical method for optimal prediction or best unbiased linear prediction. It consists of predicting the value of a variable at an unobserved location as a weighted sum of the variable at observed locations. Intuitively, it works as a reverse linear regression: instead of computing correlation (univariate regression coefficients are simply scaled correlation) between a dependent variable Y and independent variables X, it uses known correlation between X and Y to predict Y.

Version:	1.3
Depends:	R (≥ 2.15.1), doParallel
Imports:	ROCR, irlba
Published:	2014-06-18
Author:	Hae Kyung Im, Heather E. Wheeler, Keston Aquino Michaels, Vassily Trubetskoy
Maintainer:	Hae Kyung Im <haky at uchicago.edu>
License:	GPL (≥ 3)
NeedsCompilation:	no
Materials:	README
CRAN checks:	OmicKriging results

Downloads:

Reference manual:	OmicKriging.pdf
Vignettes:	Application Tutorial: OmicKriging
Package source:	OmicKriging_1.3.tar.gz
Windows binaries:	r-devel: OmicKriging_1.3.zip, r-release: OmicKriging_1.3.zip, r-oldrel: OmicKriging_1.3.zip
OS X Snow Leopard binaries:	r-release: OmicKriging_1.3.tgz, r-oldrel: OmicKriging_1.3.tgz
OS X Mavericks binaries:	r-release: OmicKriging_1.3.tgz
Old sources:	OmicKriging archive

OmicKriging

RESEARCH ARTICLE

Genetic Epidemiolog

OFFICIAL JOURNAL
INTERNATIONAL
EPIDEMIOLOGY SOC
www.geneticepi.org

Poly-Omic Prediction of Complex Traits: OmicKriging

Heather E. Wheeler,[1] Keston Aquino-Michaels,[2] Eric R. Gamazon,[2] Vassily V. Trubetskoy,[2] M. Eileen Dolan,[1] R. Stephanie Huang,[1] Nancy J. Cox,[2] and Hae Kyung Im[3]*

[1]Section of Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; [2]Section Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America; [3]Department of Health Studies, U Chicago, Chicago, Illinois, United States of America

Received 26 November 2013; Revised 11 March 2014; accepted revised manuscript 12 March 2014.
Published online 2 May 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21808

ABSTRACT: High-confidence prediction of complex traits such as disease risk or drug response is an ultimate goal sonalized medicine. Although genome-wide association studies have discovered thousands of well-replicated polymor

Summary OmicKriging

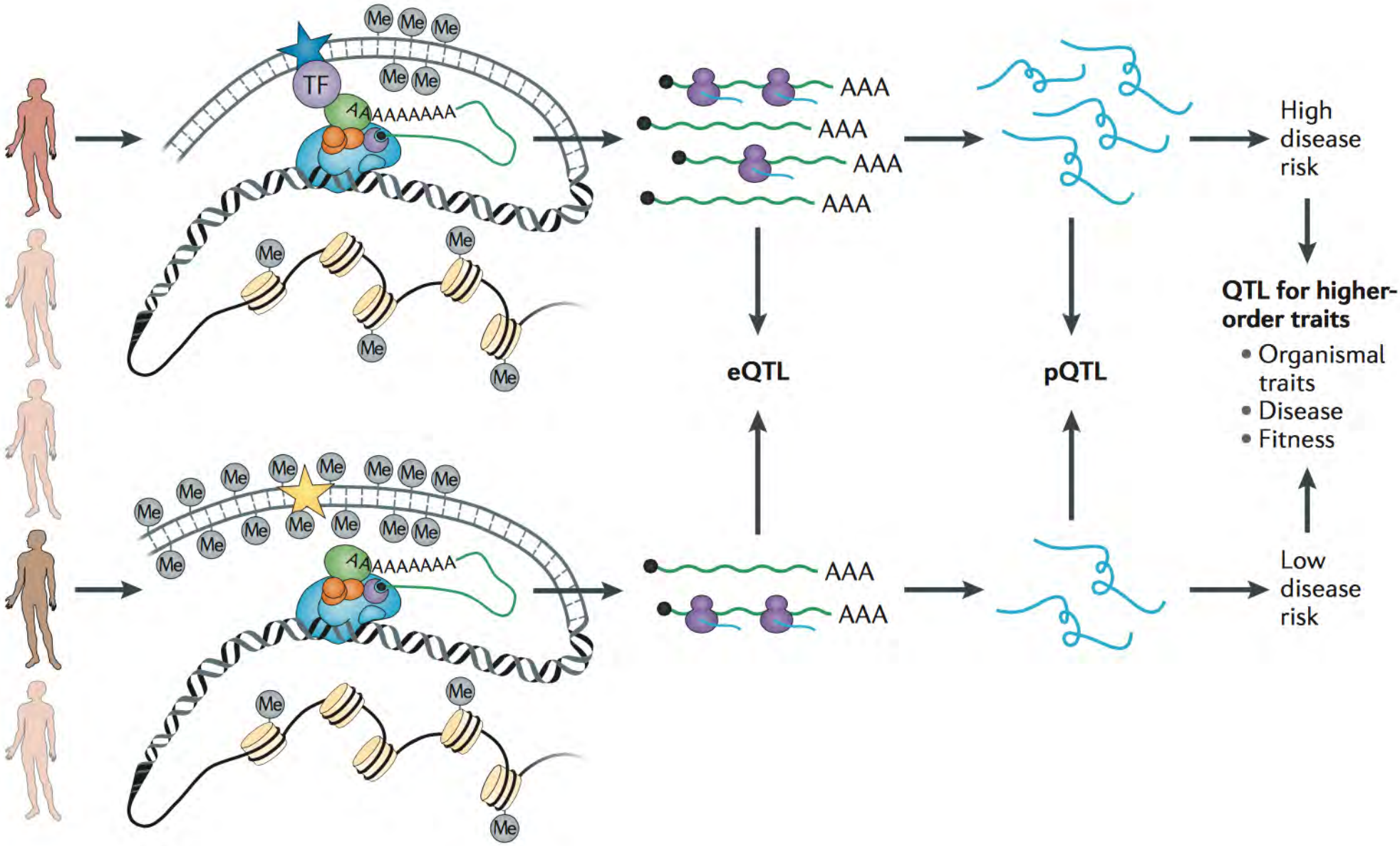
- OmicKriging is a systems approach to complex trait prediction that leverages and integrates multiple omic data
- We can attain relevant prediction even if we do not know the individual variant's contribution
- Important tool for integrating the vast amounts of data to be generated with the precision medicine initiative

Role of Regulatory Variation in Complex Traits

Mechanism of Genotype to Phenotype Link

- Most trait-associated SNPs are not coding
- Mechanism via regulation of gene expression levels

Altered Protein Levels Influences Disease Risk

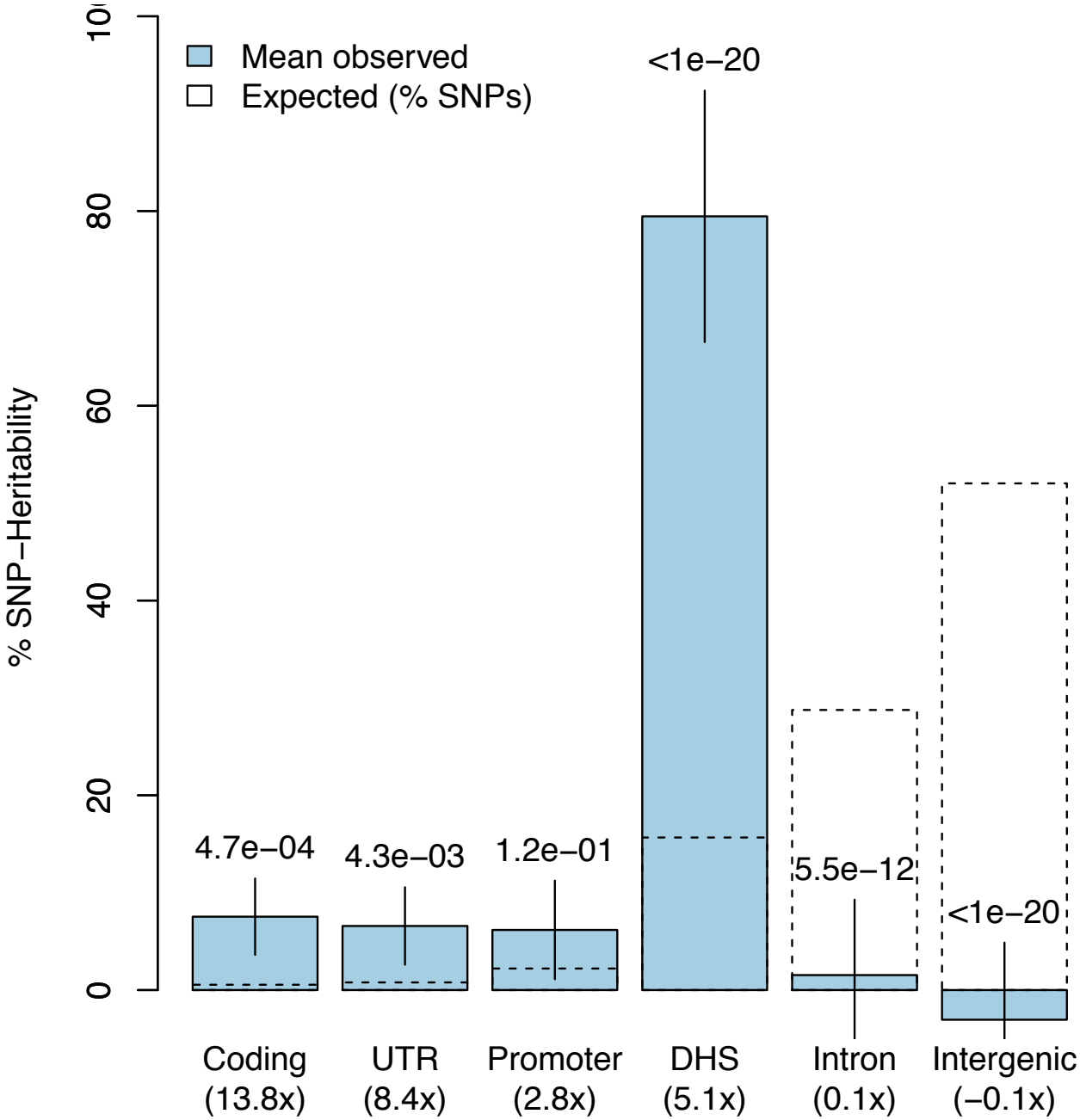


Albert & Kruglyak 2015 NGReviews

Regulatory variants explain much more heritability than coding variants across 11 common diseases

AJHG 2014

Alexander Gusev, S Hong Lee, Benjamin M Neale, et al.



DHS: DNAse hypersensitivity sites, control accessibility of the region thus levels of transcription

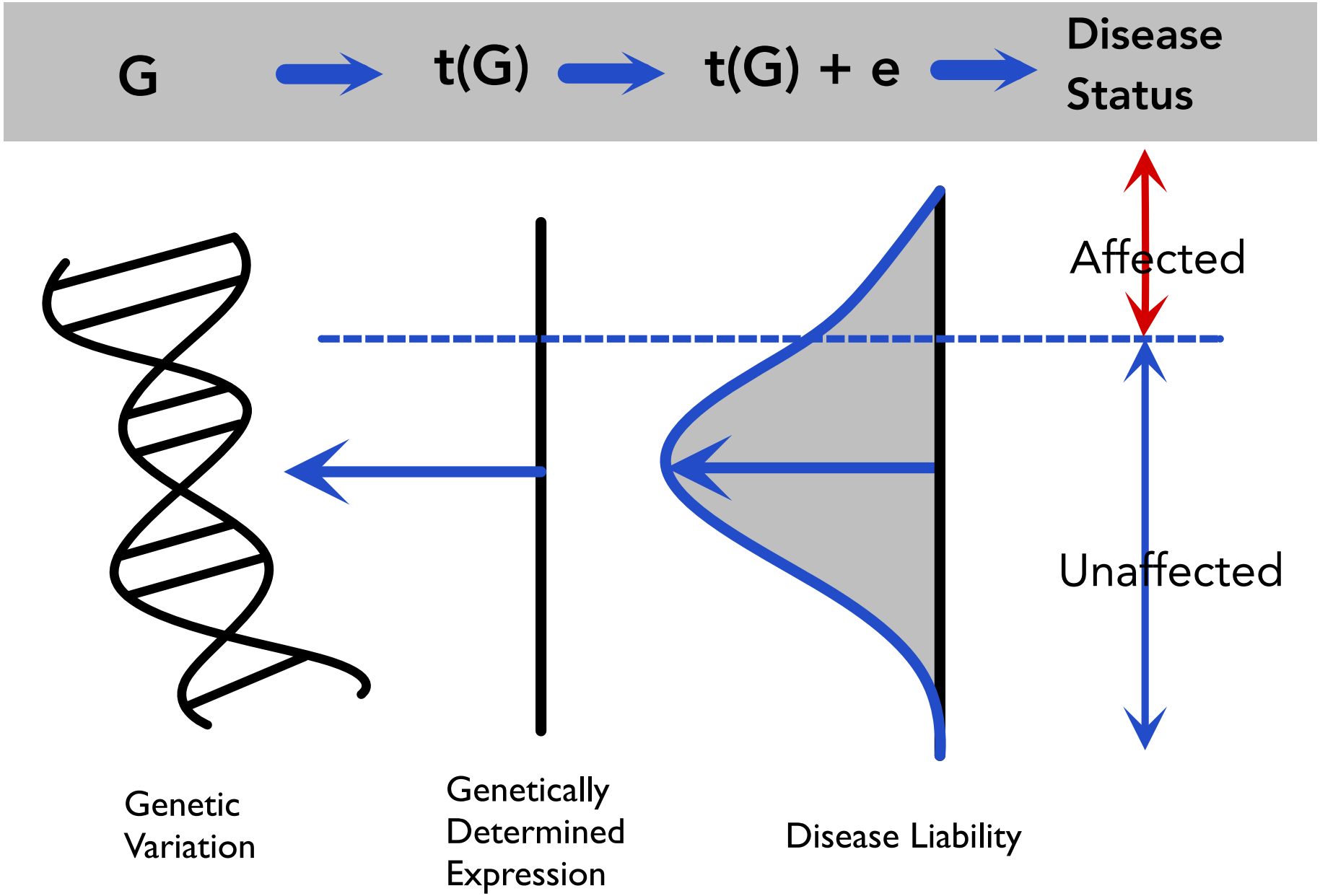
PrediXcan

Nature Genetics - under revision

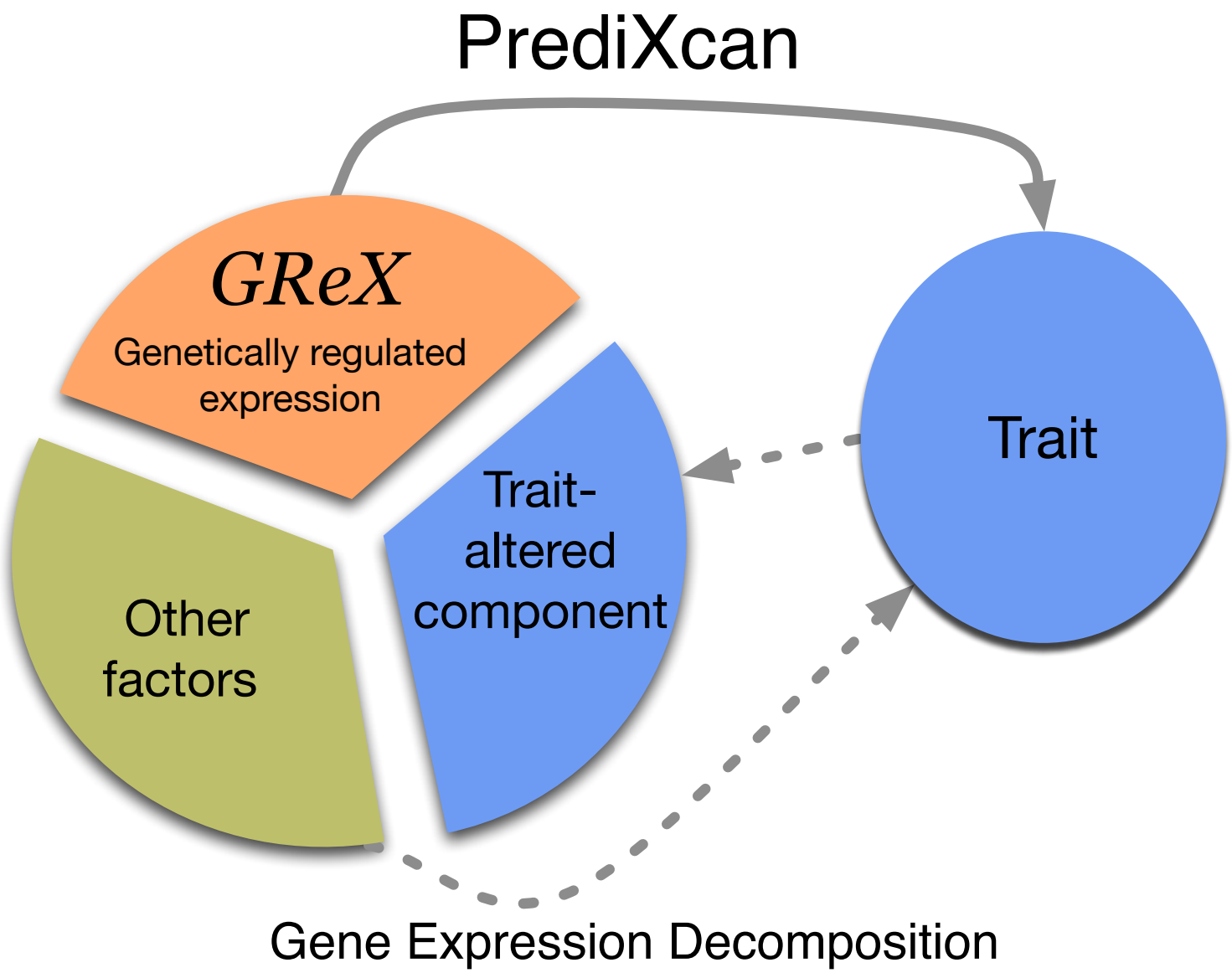
Motivation for PrediXcan

- Lack of mechanistic understanding of most GWAS discoveries
- Large proportion of variation explained by regulatory variants
- We propose PrediXcan that tests the proposed mechanism

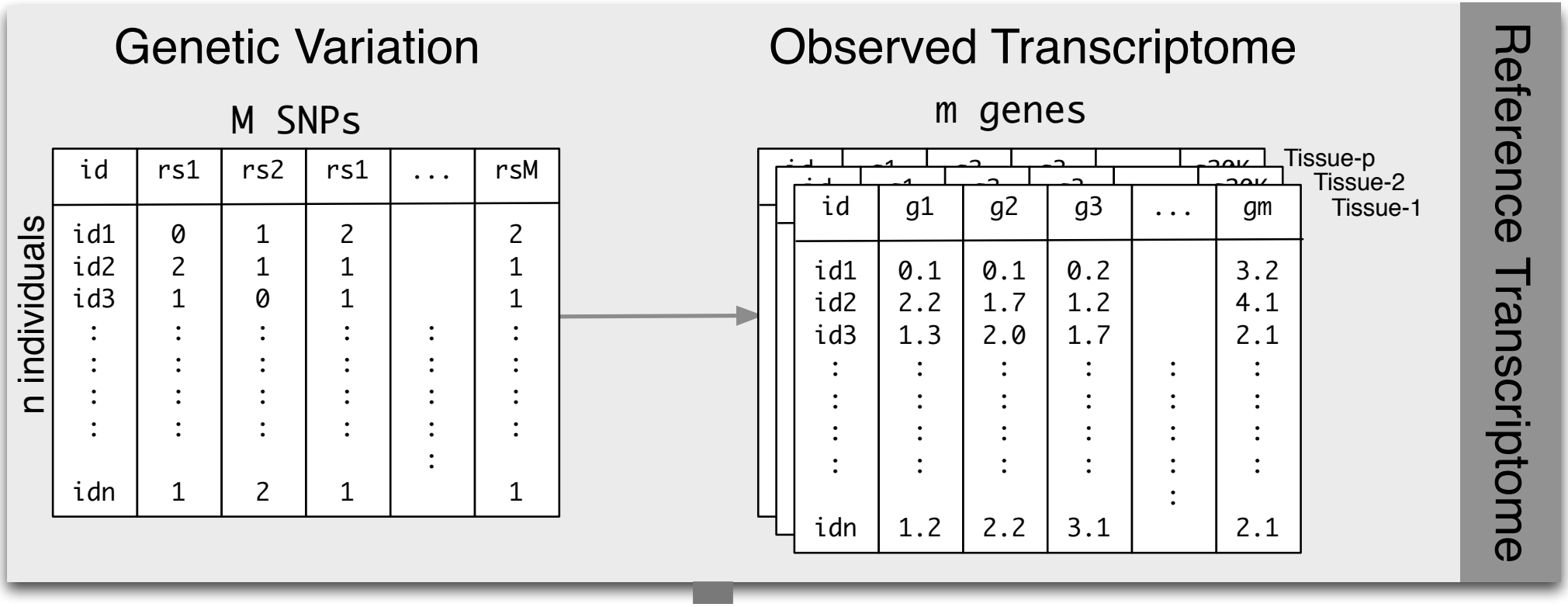
Genetic Control of Disease Through Gene Regulation



Mechanisms Tested by PrediXcan

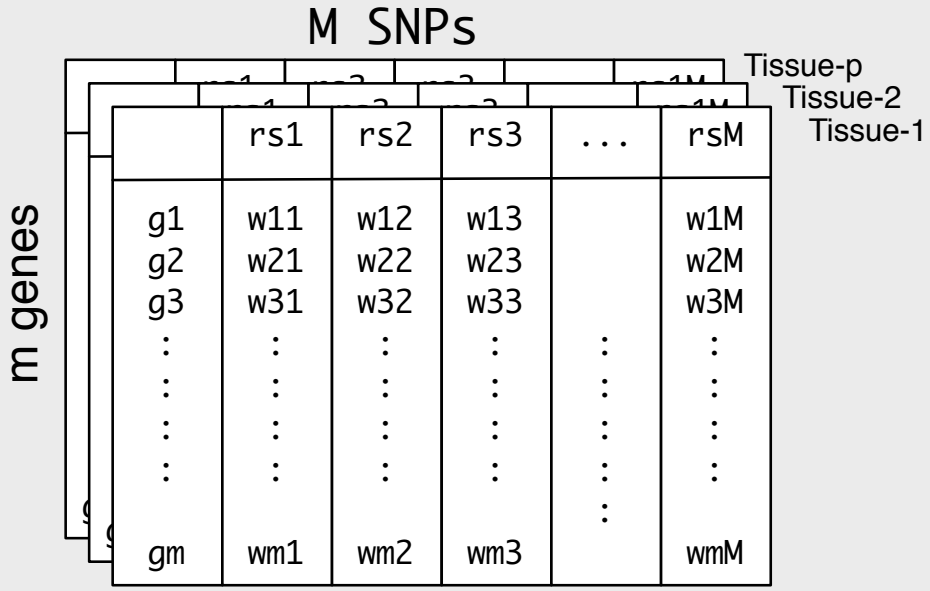


PrediXcan uses Reference Transcriptome



PredictDB: Public Database of Weights for GReX

PredictDB: Database of Prediction Models

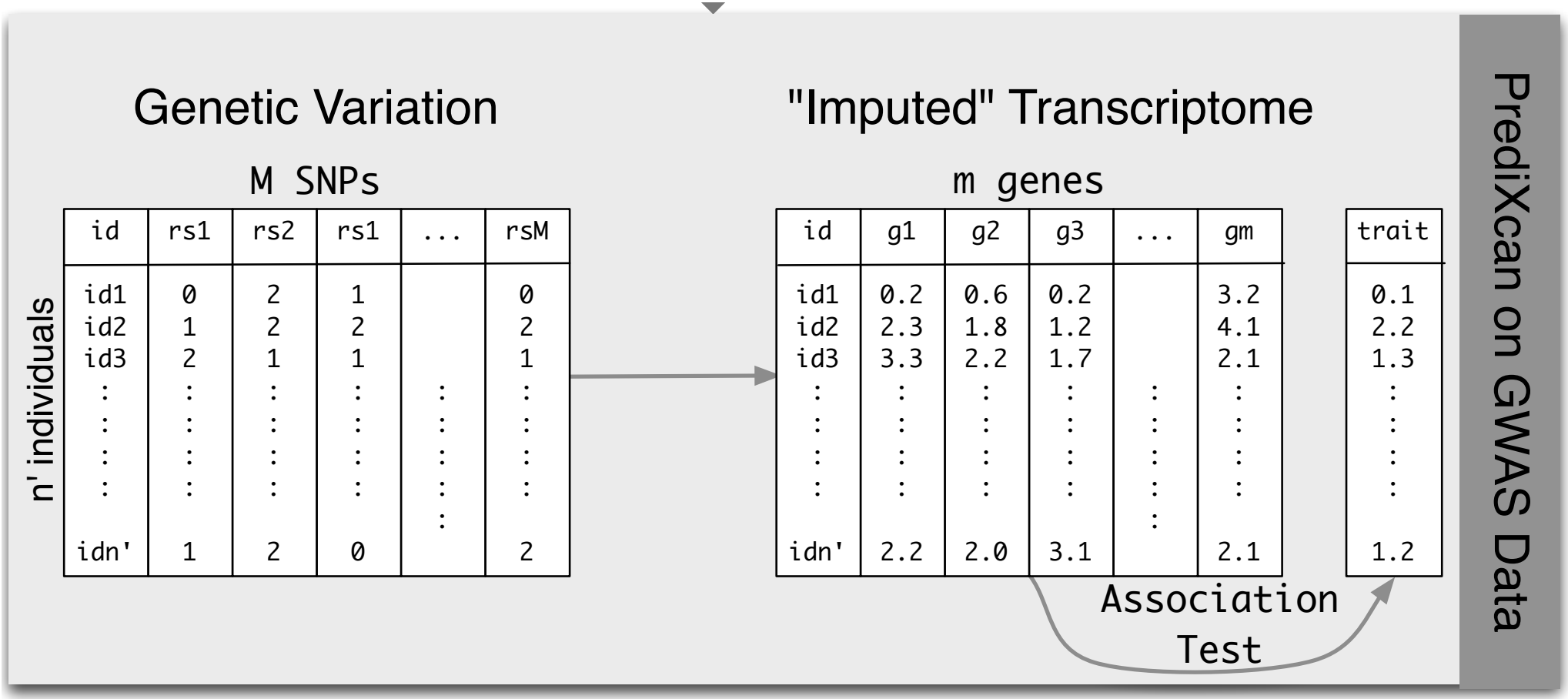


Additive model of gene expression trait trained in reference transcriptome datasets

$$T = \underbrace{\sum_k w_k X_k}_{GReX} + \epsilon$$

Weights stored in PredictDB

PrediXcan Imputes Transcriptome & Tests Assoc.



PrediXcan: Mechanism-driven Gene-Based Test

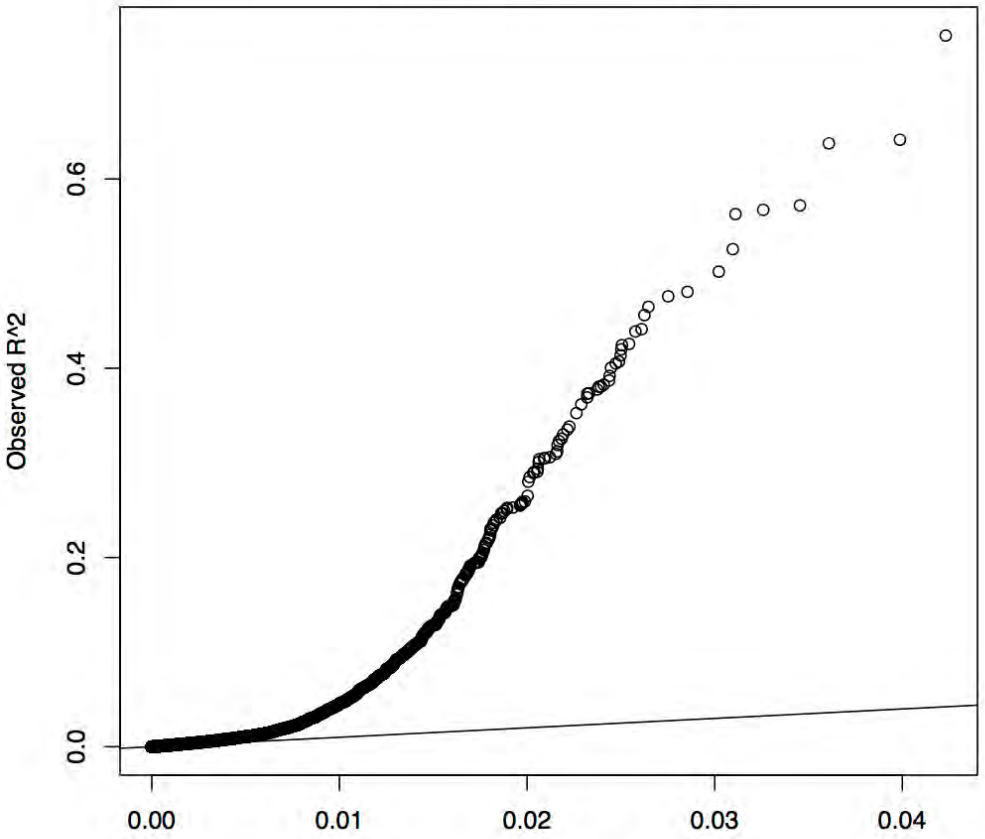
- Directly tests the molecular mechanism through which genetic variants affect phenotype
- Genes more attractive than genetic variants
  - A lot is known about their function
  - Follow up experiments can be easily devised
  - Reduced multiple testing burden
- Direction of effects
  - Positive effects: down regulation is therapeutic option
  - Negative effects: more likely to harbor deleterious rare variants
- No reverse causality issues
- Can be systematically applied to existing GWAS studies
- Tissue-specificity can be inferred

Reference Transcriptome Data

- 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
- Cerebellum expression (Array GSE35974)
- Framingham, n>5000m, whole blood
- Depression Genes & Networks, n>900, whole blood

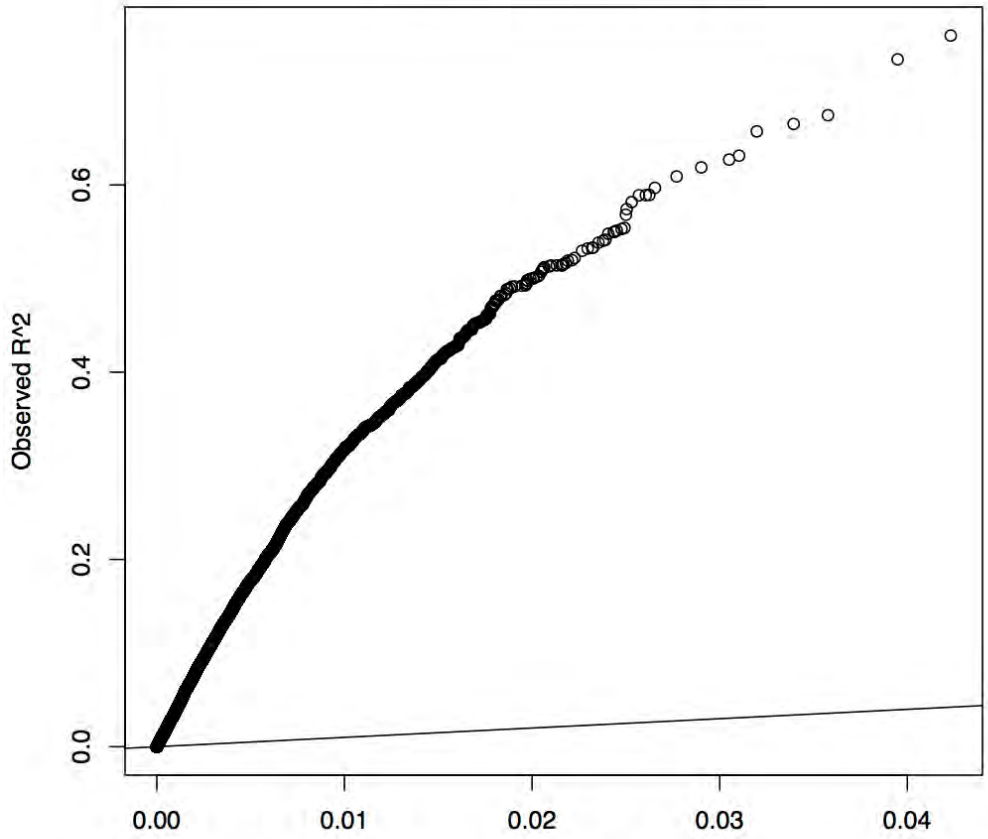
Good Prediction Performance

Prediction R^2



Training with GTEx Testing in 1K Genomes

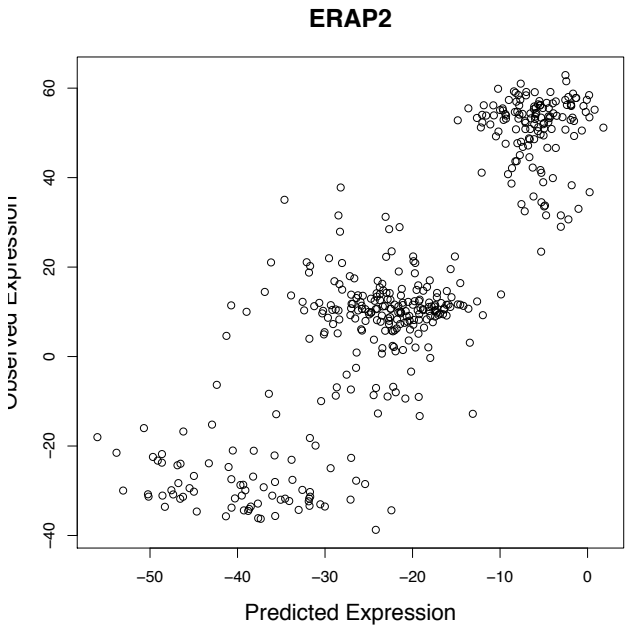
Replication R^2

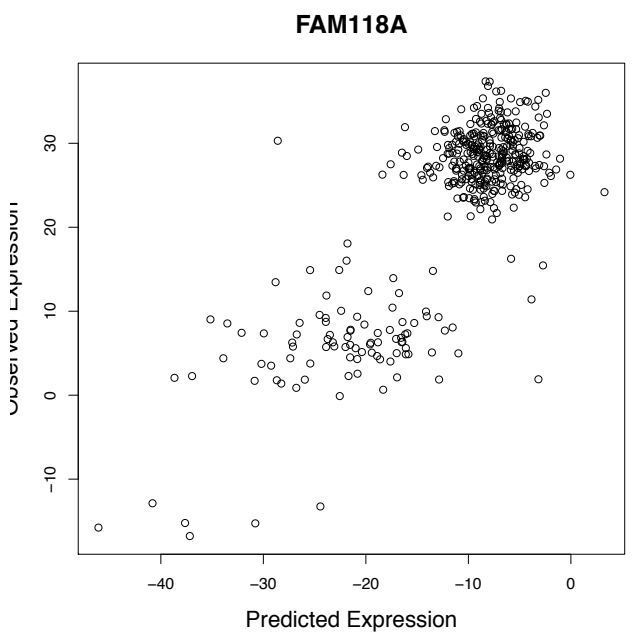


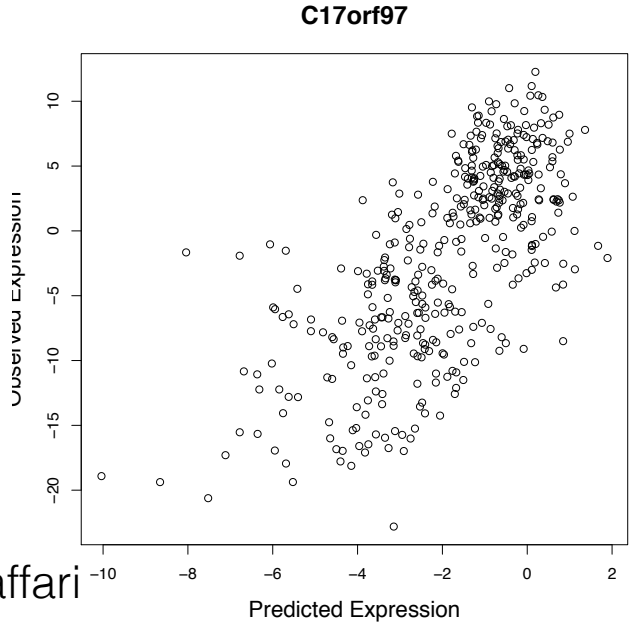
Replicate RNAseq Pickrell et al 2010 vs. 1K Genomes 2013

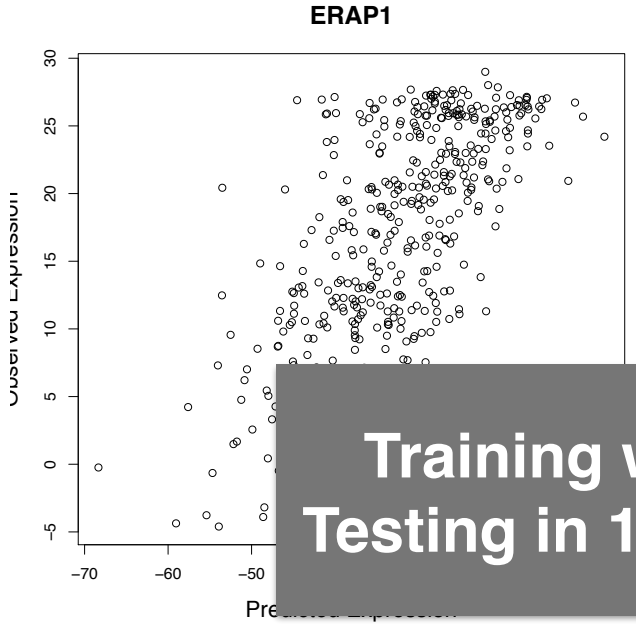
Sahar Mozaffari

Examples of Well Predicted Genes





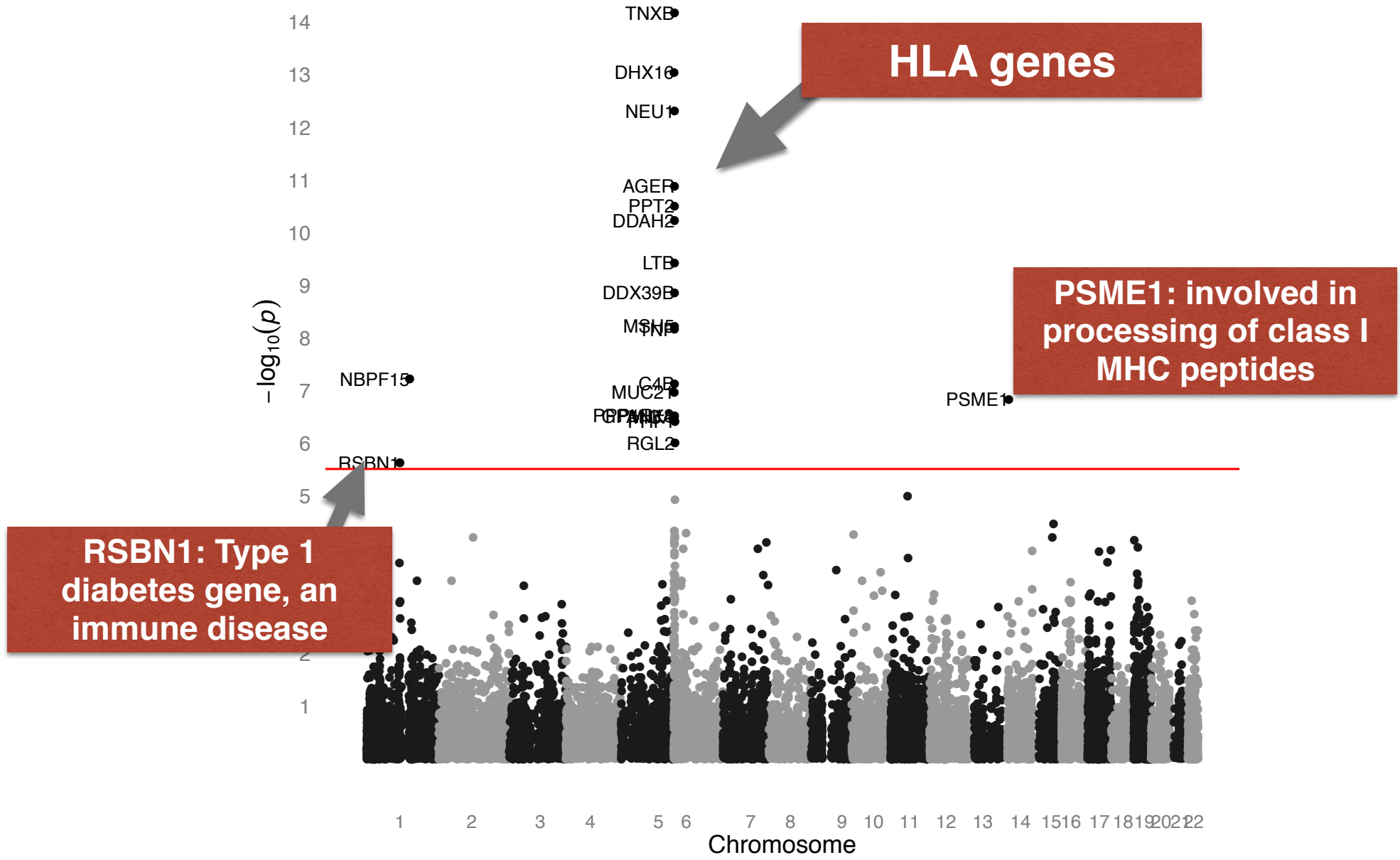




Training with GTEx
Testing in 1K Genomes

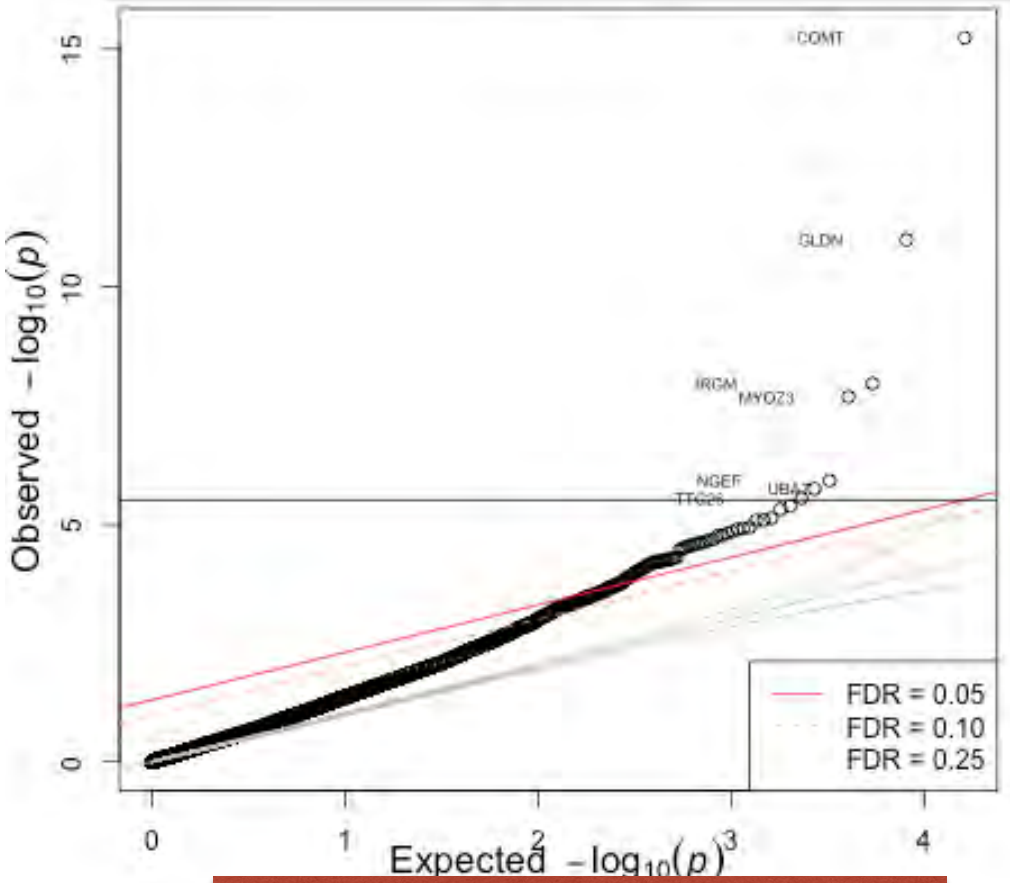
Sahar Mozaffari

Genes Associated with Rheumatoid Arthritis



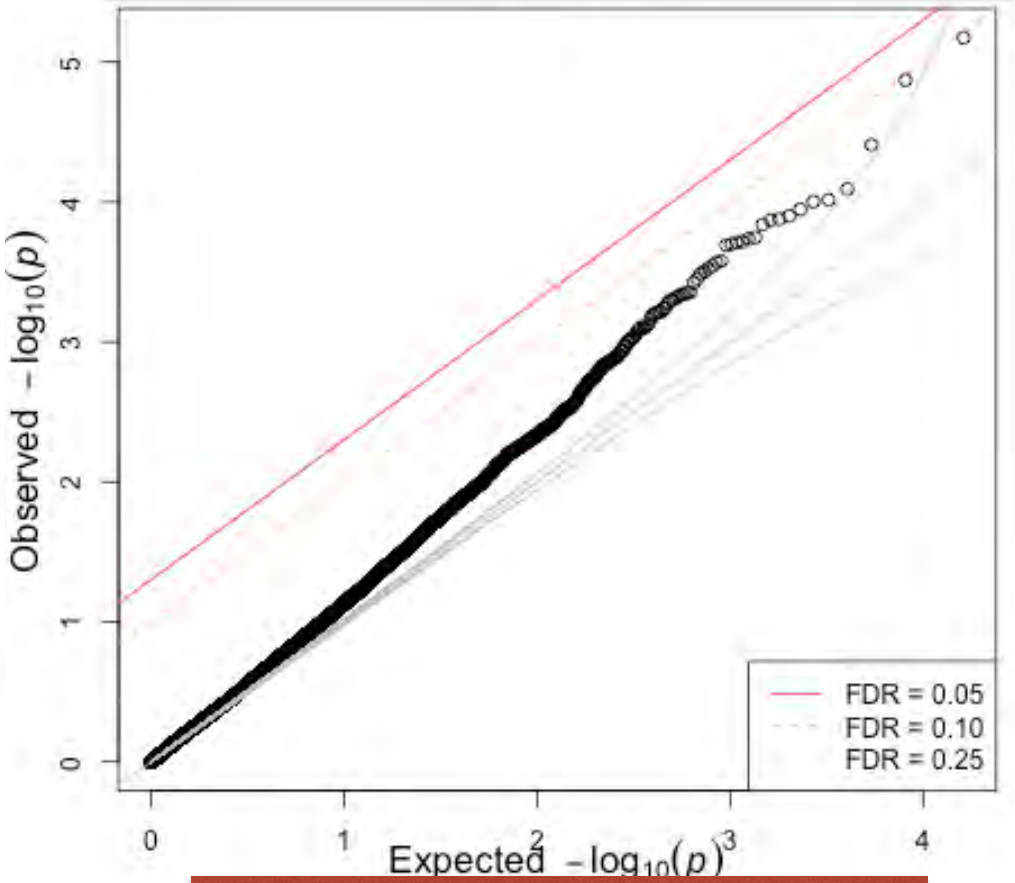
PrediXcan Results for Crohn's Disease and Hypertension

Crohn's Disease



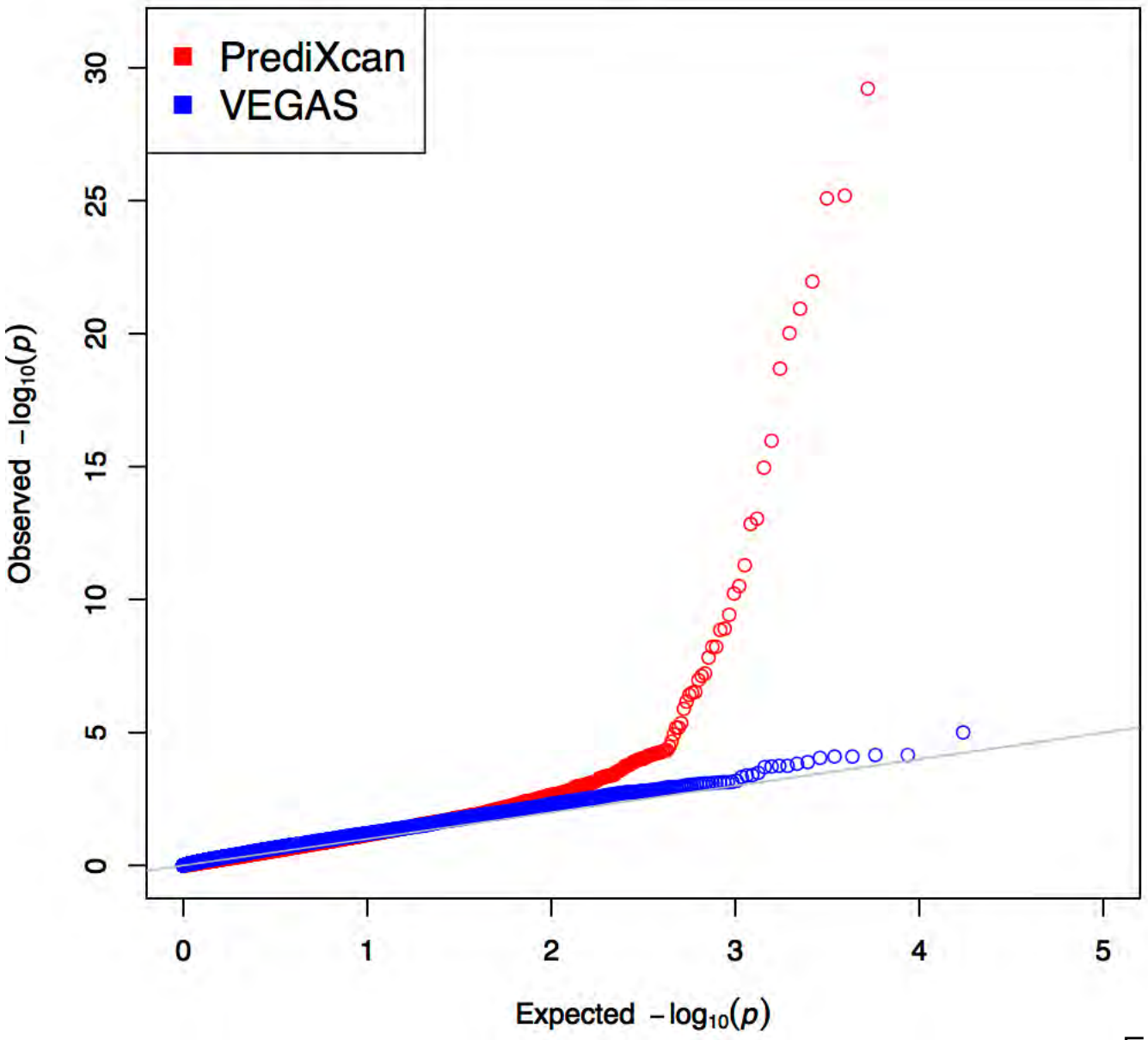
IRGM is a known Crohn's gene

Hypertension



Whole blood may not be relevant tissue

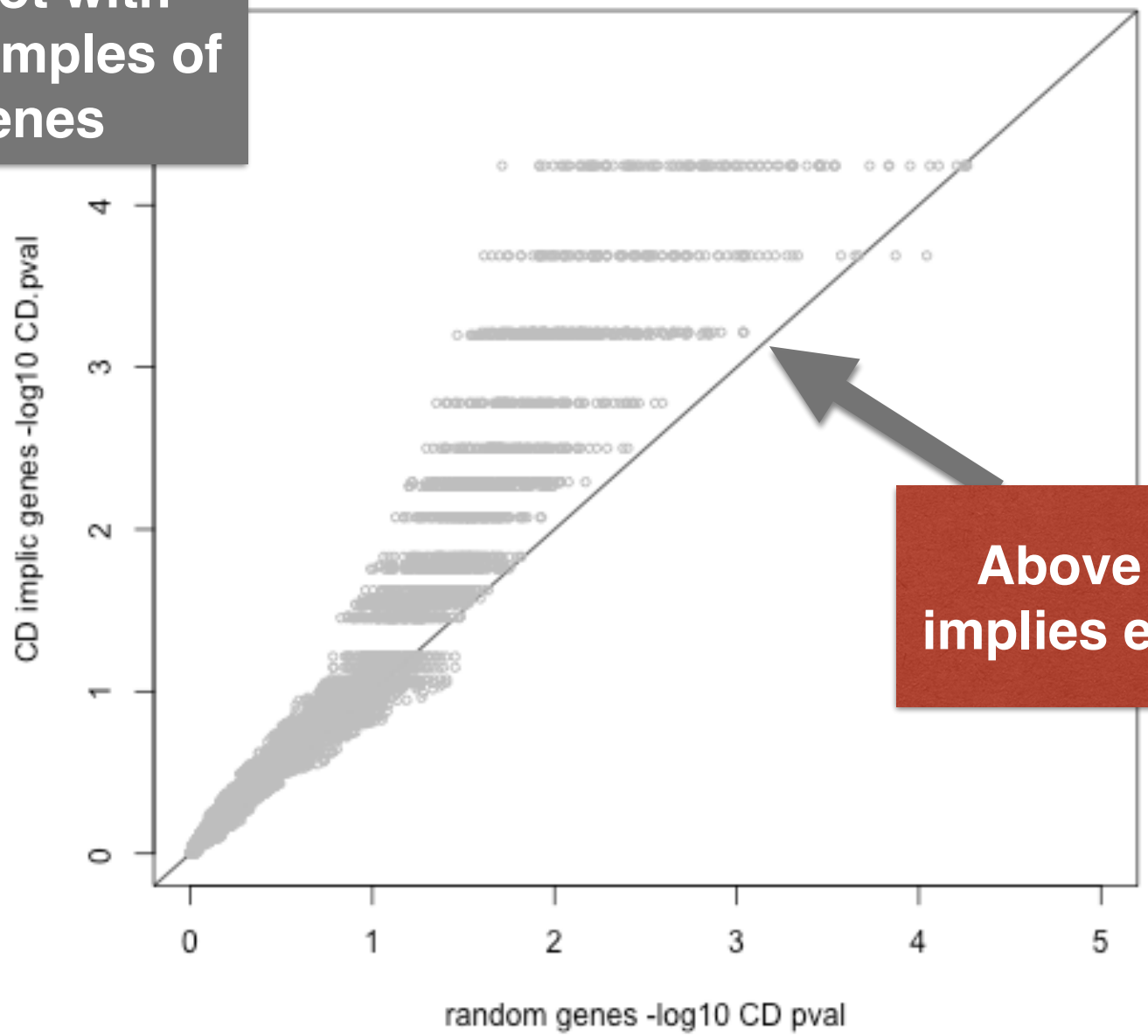
PrediXcan Outperforms VEGAS



Eric Gamazon

Enrichment of Known Crohn's Genes Among Findings

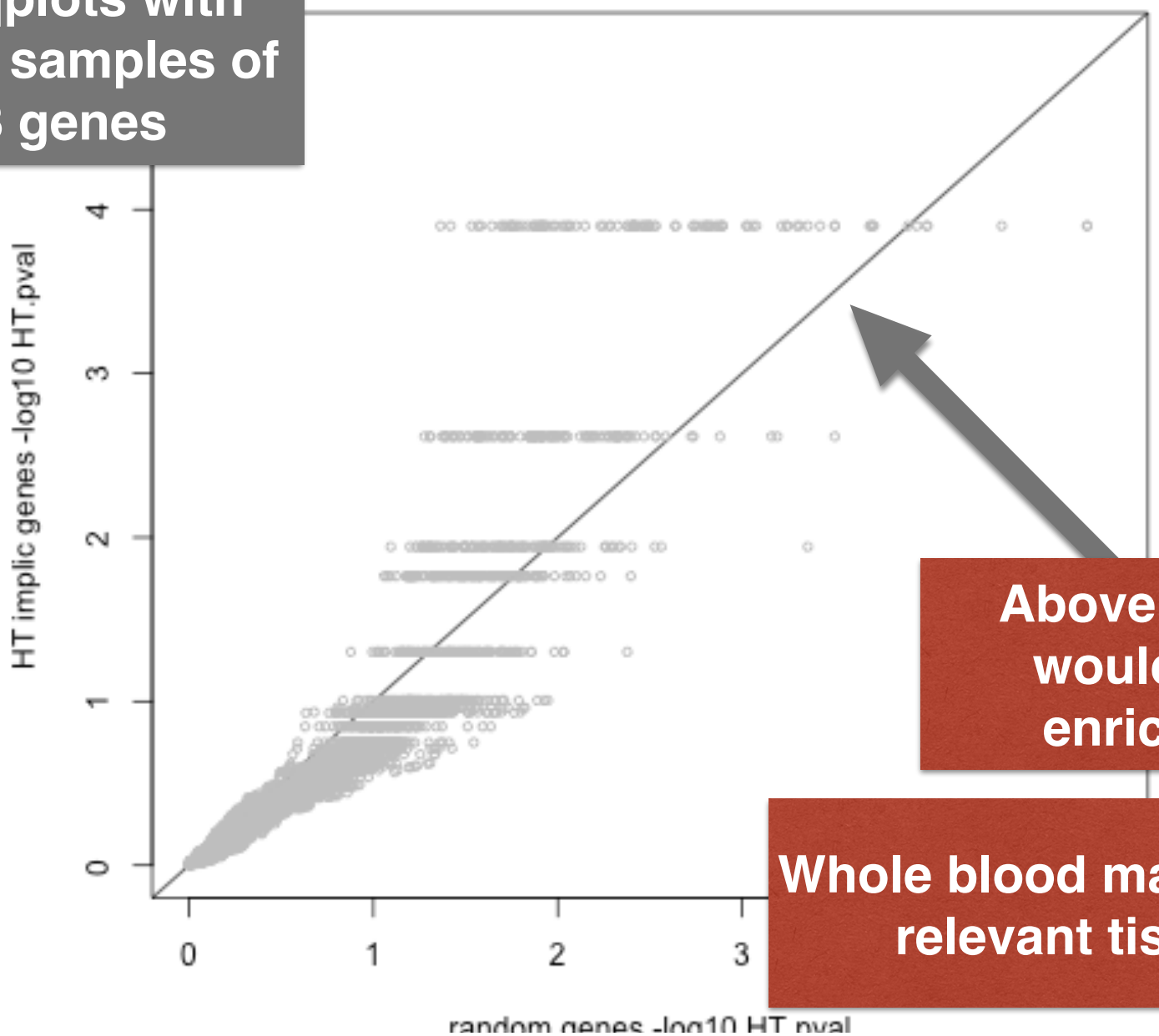
100 qqplot with random samples of 205 genes



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

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

Prediction of Gene Expression Traits

Genetic Architecture to Improve Prediction

- Local and distant regulation (heritability)

- Sparsity/Polygenicity

- This information guides us to improve prediction, i.e. estimates of GReX

Local/Distant Heritability Estimation

- Gene expression trait model

$$Y = \sum_{\text{local}} \beta_k^{\text{local}} X_k + \sum_{\text{distant}} \beta_k^{\text{distant}} X_k + \epsilon$$

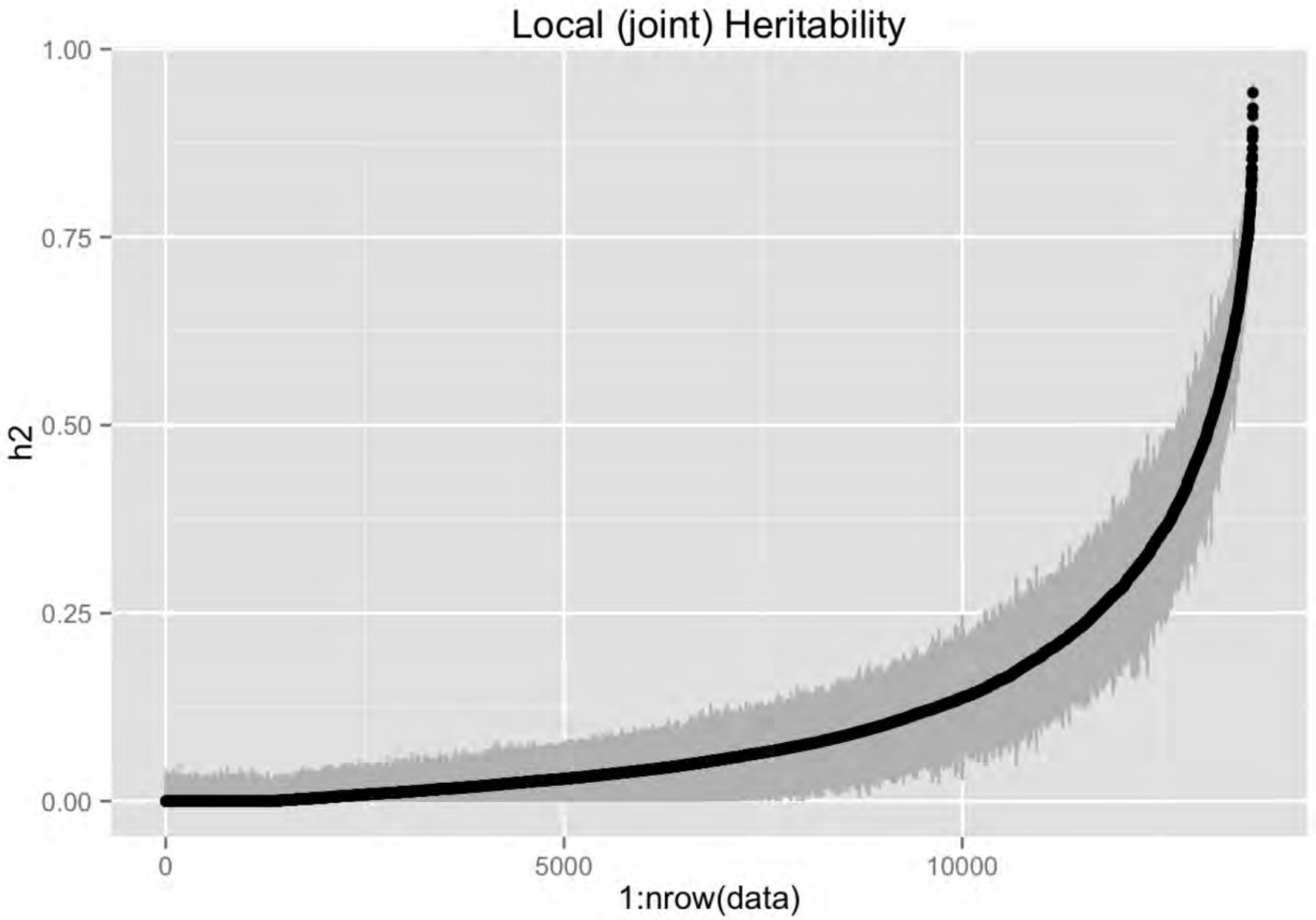
- REML to estimation of local and distant contributions jointly
- Covariance of local component: GRM using SNPs nearby
- Covariance of distant component: GRM using distant SNPs
- We use GCTA as REML calculator

Total Heritability = Local H2 + Distant H2

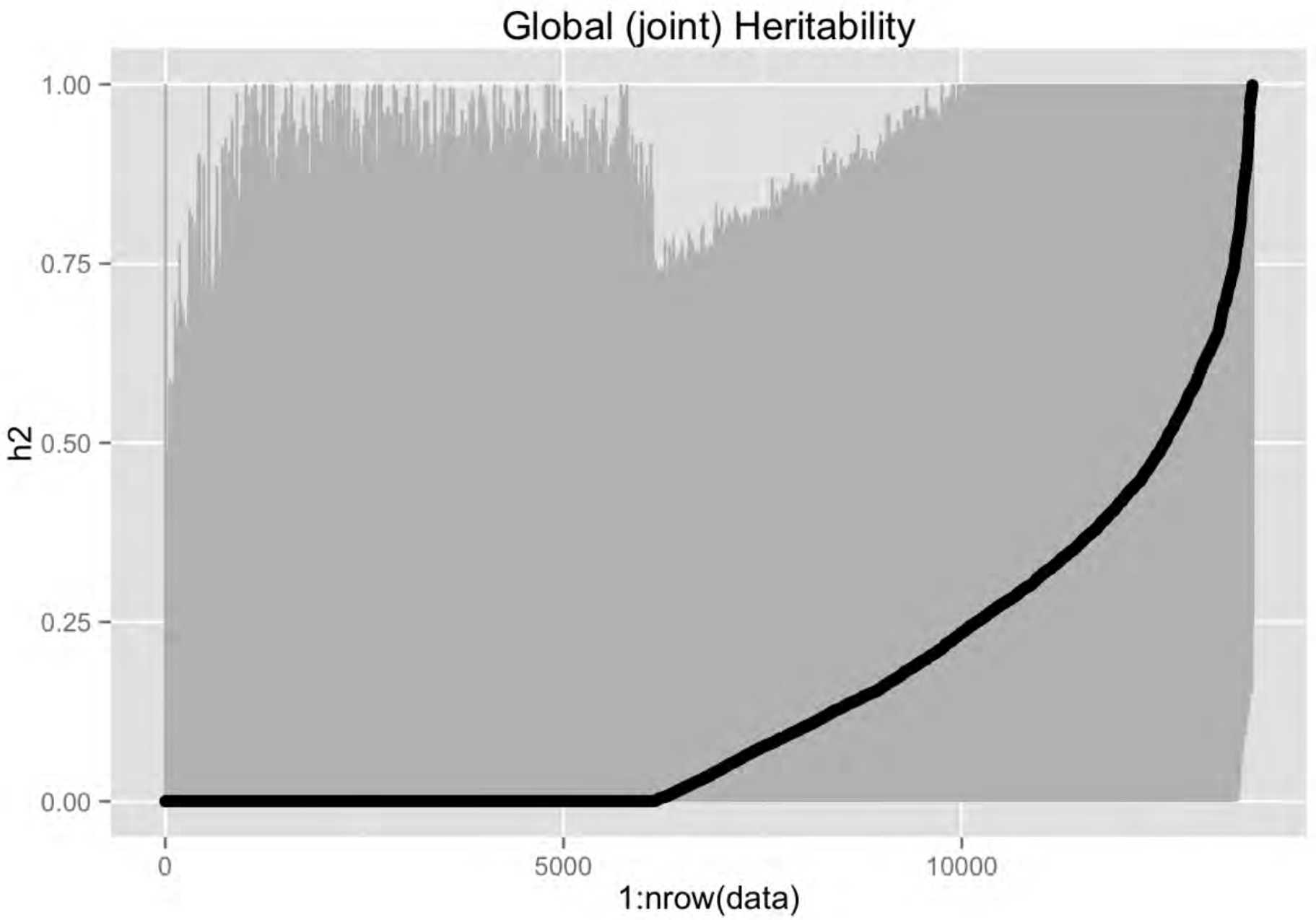
Whole Blood Expression Data: DGN

- Battle et al. “Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.” Genome Research 2014, 24(1):14-24
- Whole blood from Depression Genes and Networks study
- n = 922
- RNA-seq

Local Heritability Can Be Well Estimated



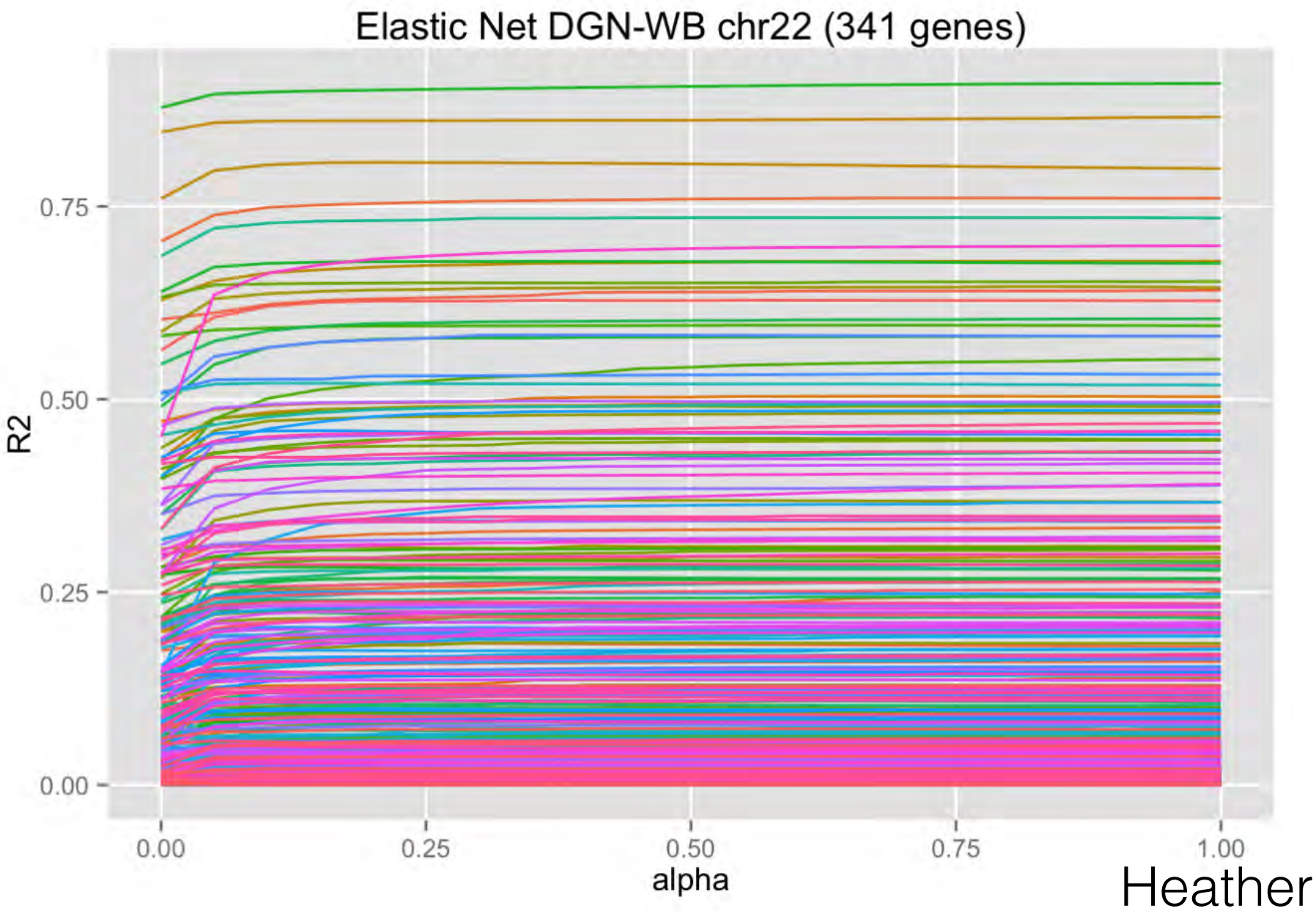
Distant Heritability Not Reliable



Proportion of LASSO to Ridge as Measure of Sparsity

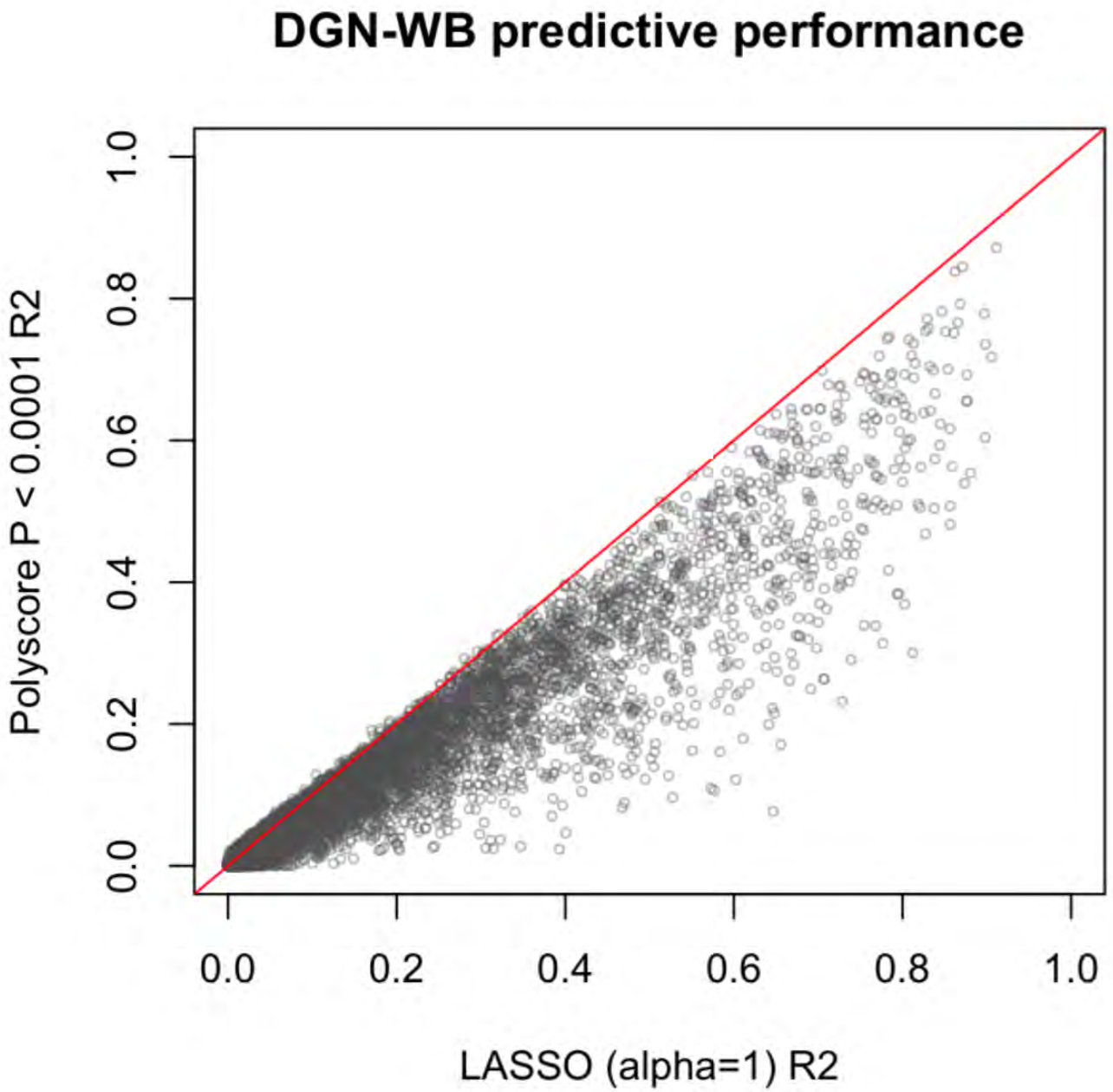
- Only local component can be assessed
- LASSO performs slightly better than E-N 0.50 in cross validated R2

Performance vs sparsity

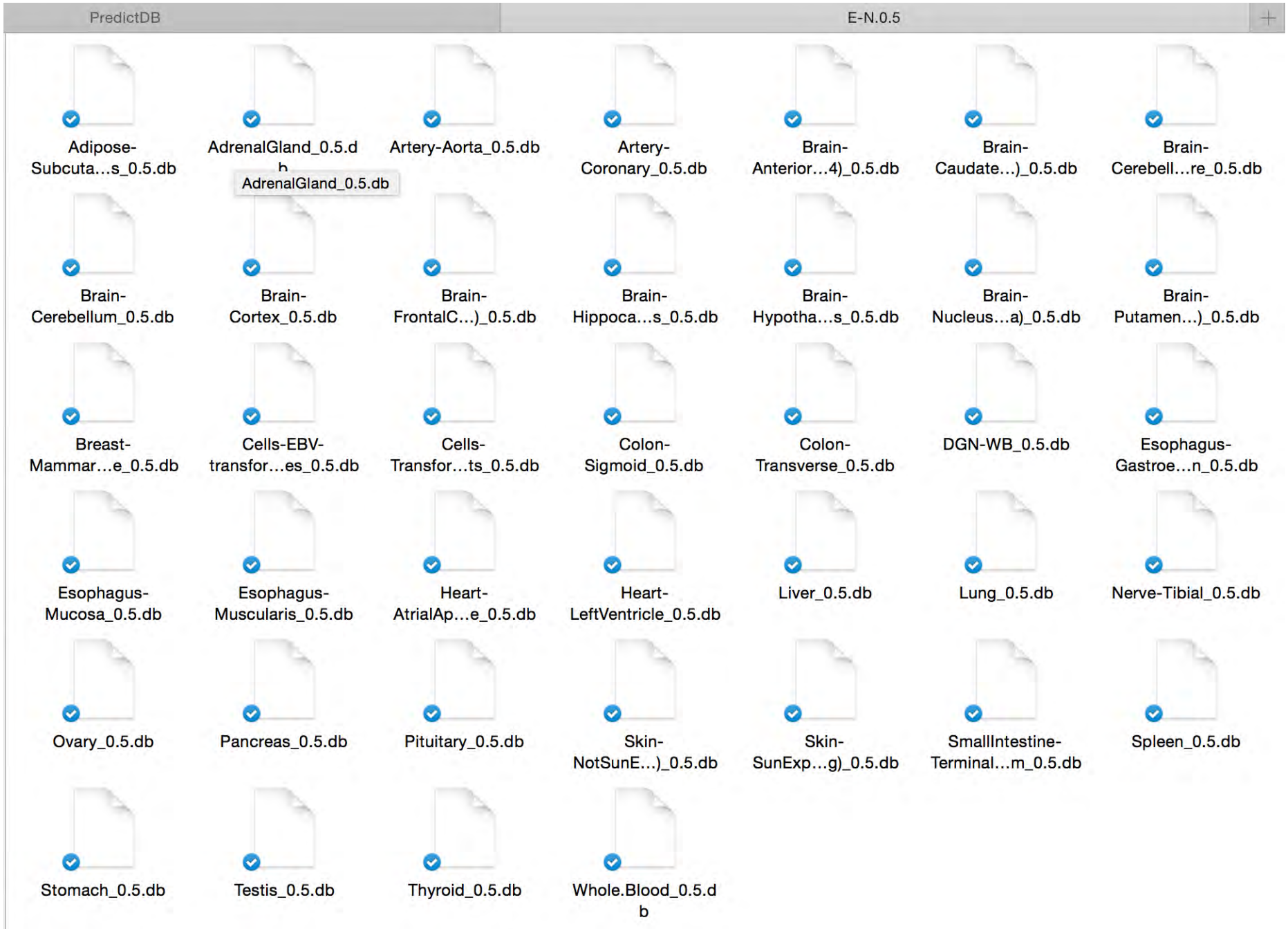


Heather

E-N & LASSO Outperform Polygenic Score



Whole Blood DGN (n=922) + 38 GTEx Tissue Models

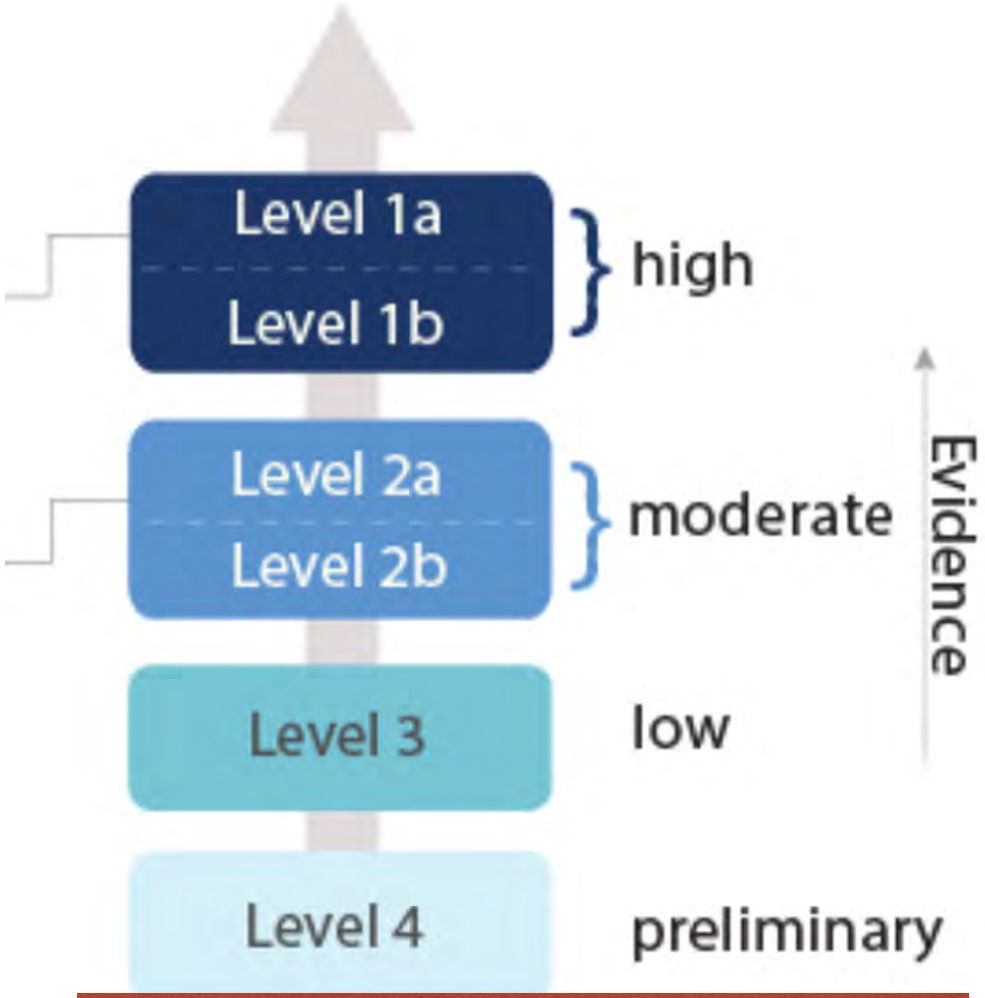


Challenges in Pharmacogenomic Predictions

Pharmacogenomic Findings

Evidence Level	Counts	%
1a	40	3
1b	17	1
2a	96	6
2b	74	5
3	1175	76
4	145	9
Total	1547	100

https://www.pharmgkb.org/



Only Level 1a findings have clinical guidelines

Challenges of Pharmacogenomic Studies

- Smaller sample size
- Even more important to integrate prior data
- Integrate other functional data
- Heritability estimates are harder
  - Limited family data
  - Usually samples greater than 1K are needed for GCTA

Bevacizumab Induced Hypertension

- Bevacizumab is a humanized monoclonal antibody that inhibits VEGF induced angiogenesis
- Hypertension is a common adverse event to bevacizumab treatment
- The incidence of hypertension with bevacizumab is 20-30%, while grade 3 or greater hypertension occurs in only 10-15% of patients.

Keston Aquino Michaels

Bevacizumab Trials

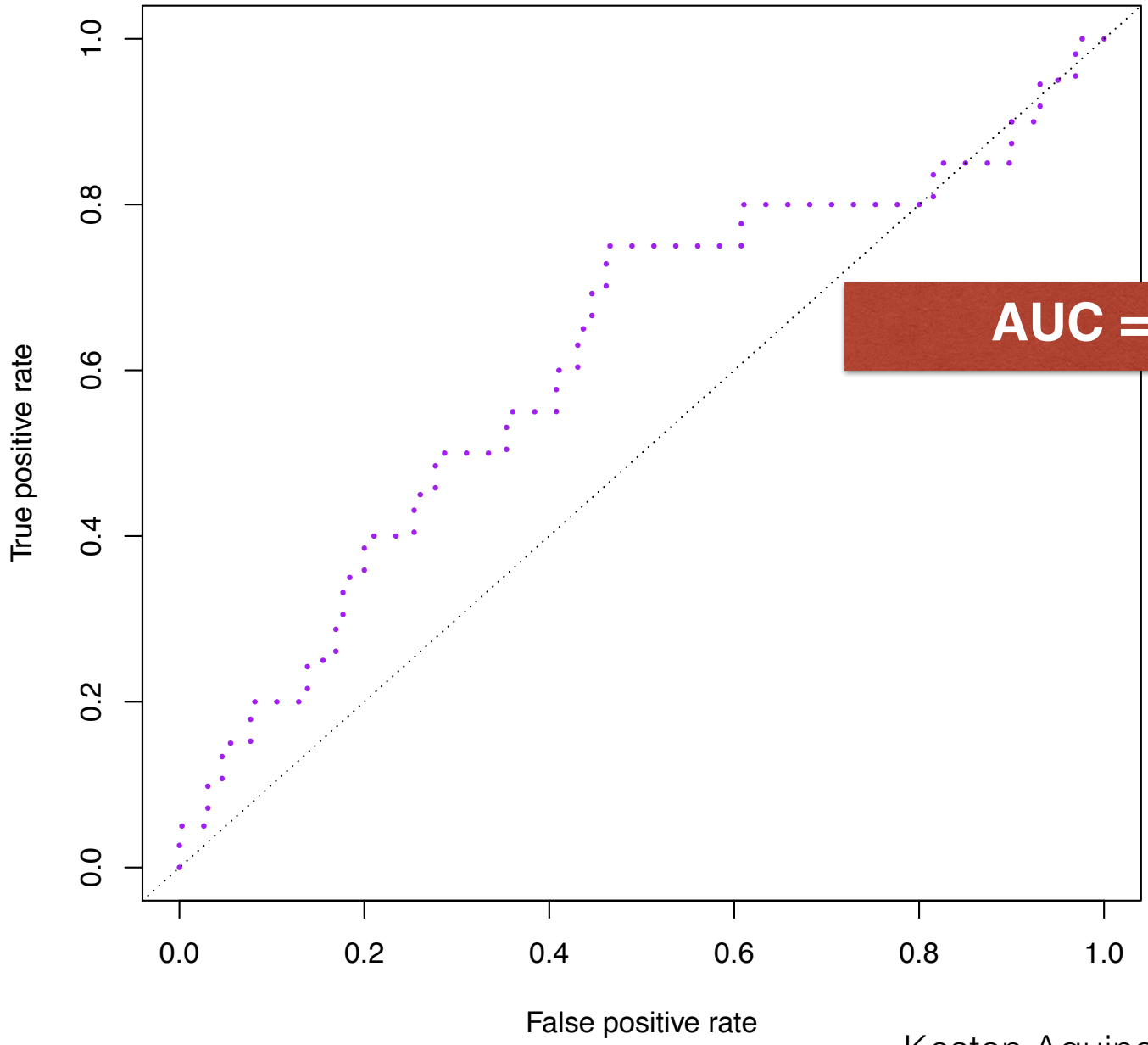
- CALGB 90401
  - a randomized double-blinded placebo controlled phase III trial comparing docetaxel and prednisone with and without bevacizumab in men with hormone refractory prostate cancer
  - n = 664 (with genotype data after QC)
  - PI: Howard McLeod
- CALGB 80303
  - a randomized phase III trial of gemcitabine plus bevacizumab versus gemcitabine plus placebo in patients with advanced pancreatic cancer
  - n = 152 (with genotype data after QC)
  - PI: Federico Innocenti

Bevacizumab Induced Hypertension

- Is primary hypertension risk score predictive of bevacizumab induced hypertension
  - Hypertension results from Cross Consortia Pleiotropy group (n~20K)
- Can we predict drug induced hypertension?
  - 90401 training set
  - 80303 test set

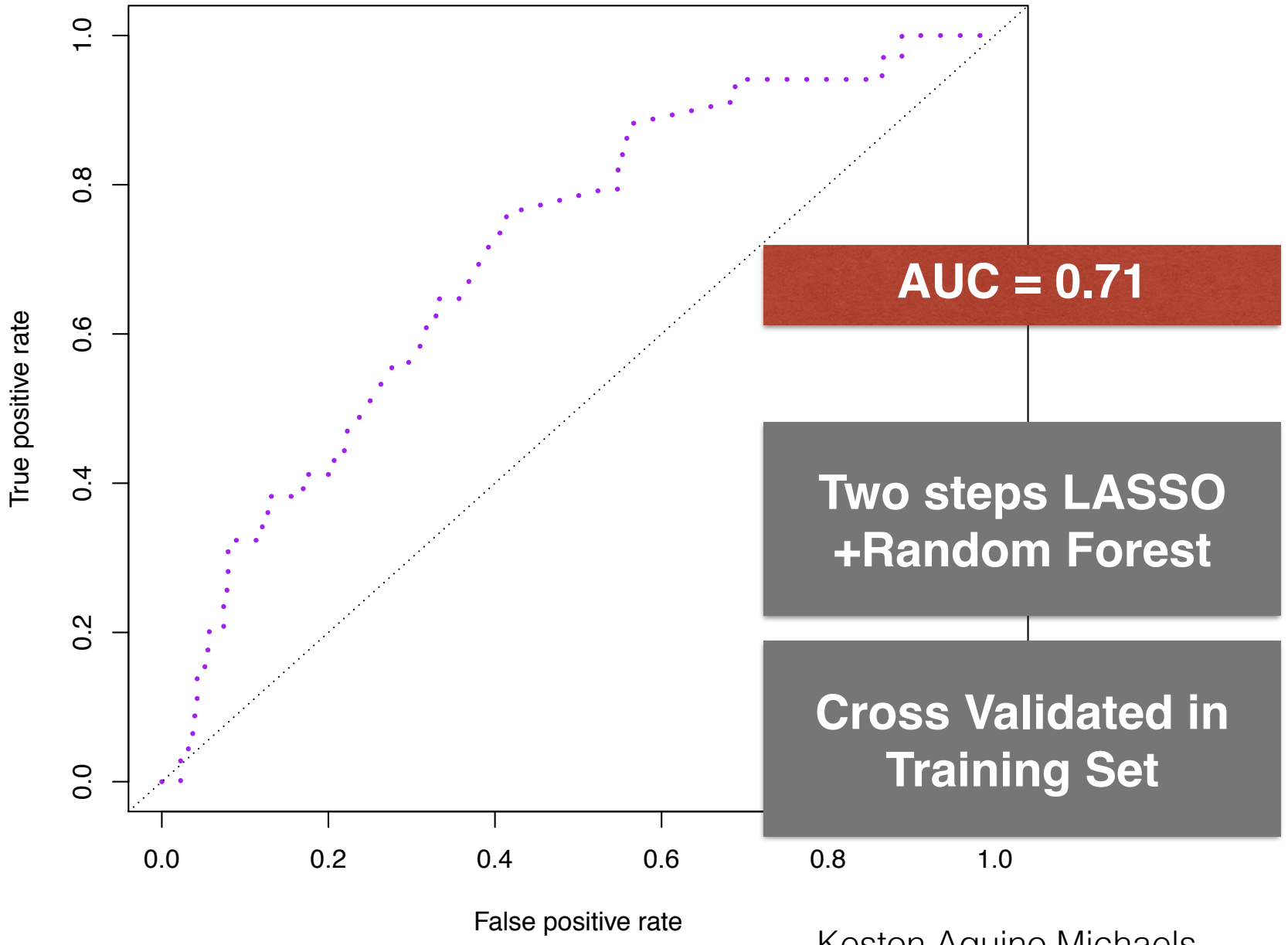
Keston Aquino Michaels & Heather Wheeler

Primary Hypertension Score Predicts Bev-induced HT



Keston Aquino Michaels

Bev-Hypertension Predicted Within Study



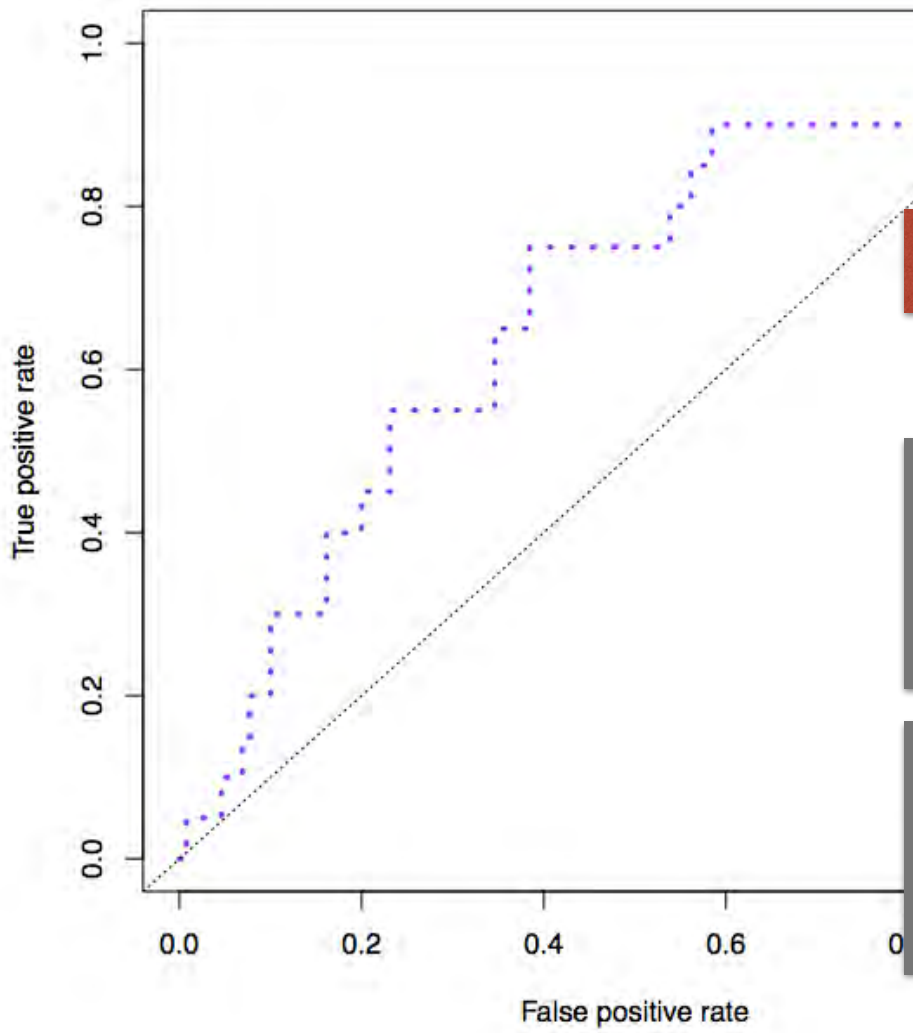
AUC = 0.71

Two steps LASSO +Random Forest

Cross Validated in Training Set

Keston Aquino Michaels

Bev-Hypertension Predicted in Independent Study



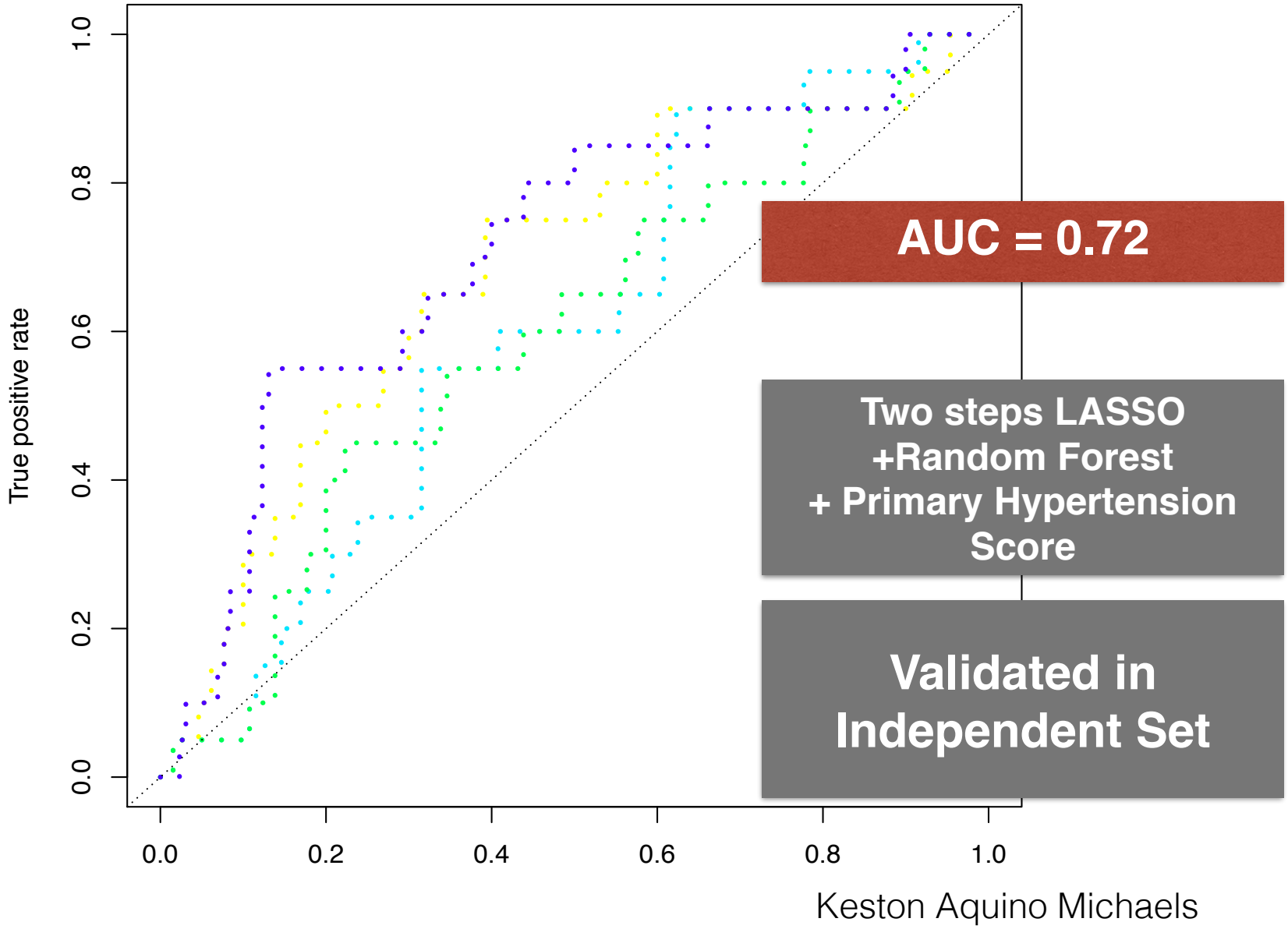
AUC = 0.68

Two steps LASSO +Random Forest

Validated in Independent Set AUC down to 0.68 from 0.71

Keston Aquino Michaels

Bev-Hypertension Predicted in Independent Study



Summary Pharmacogenomics

- Most single variant findings have limited clinical utility
- Whole genome approaches to prediction improves utility
- Bevacizumab induced hypertension example
  - primary hypertension results help in predicting drug induced hypertension
  - successfully predicted bevacizumab induced hypertension in independent study
  - combining primary + bevacizumab induced HT leads to improved prediction

Summary

- Shift from monogenic to polygenic paradigm
- Systems approach to genomics
  - Most single variant findings have limited clinical utility
  - Whole genome approaches to prediction improves utility
- Larger sample sizes will be needed, 1Million+
- OmicKriging: prediction method that integrates heterogeneous sources of data well suited for data from the Precision Medicine Initiative
- Large role of regulation variants in complex traits
- PrediXcan: novel gene based test that test mechanism
- Prediction of gene expression traits

Conclusion

- recognizing the complexity of the genetic architecture and mechanisms of genetic control,
- collecting deep phenotype data from large number of individuals,
- brodly sharing data and results, and
- integrating multiple sources of data
- using mechanism-driven tests

We will achieve the promise of precision medicine

Thank You!

Contributors

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

Data sources

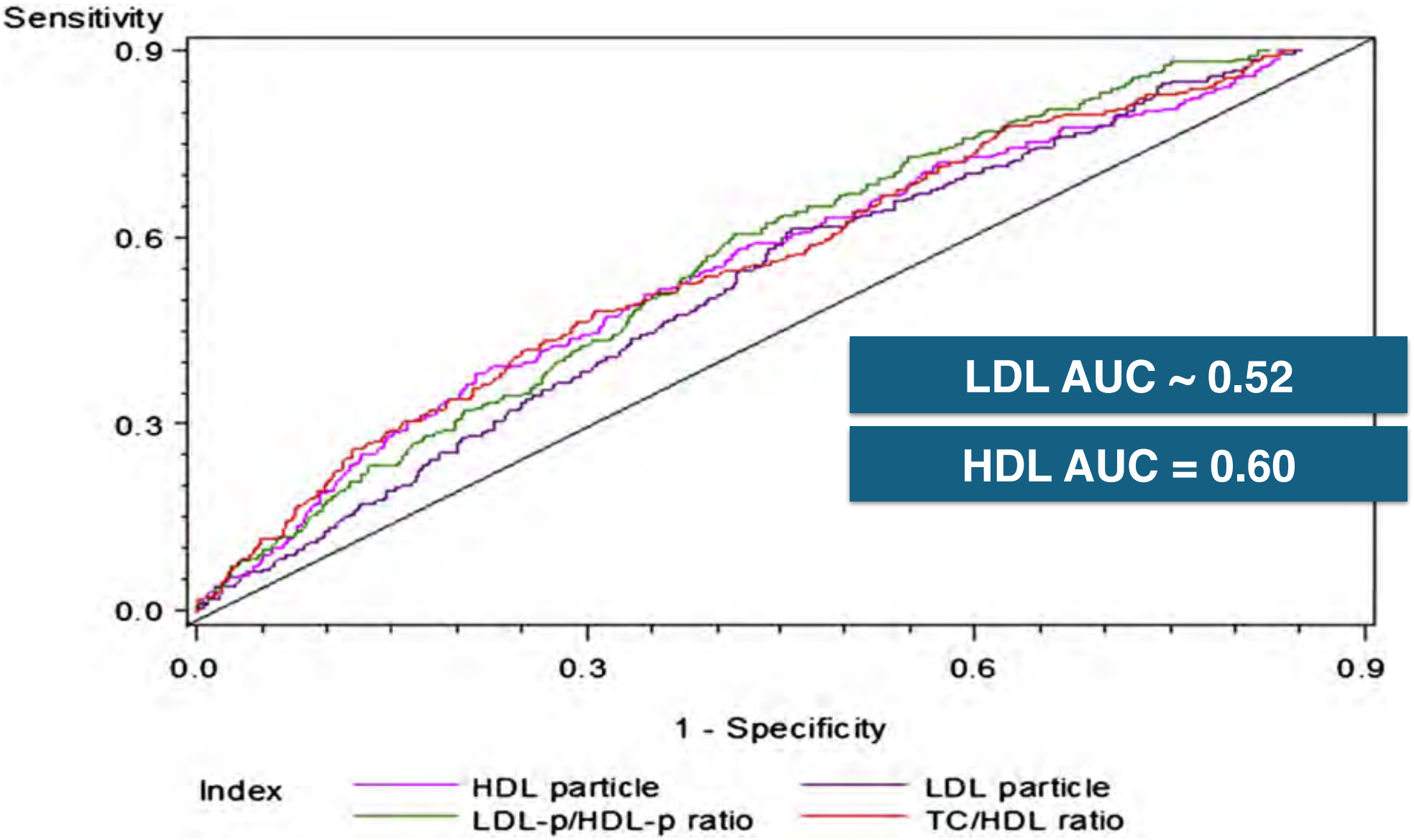
- WTCCC
- GAINS/Bipolar Disorder
- GoKinD
- Disease Genes & Networks

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
- P50DA037844 - Integrated GWAS of complex behavioral and gene expression traits in outbred rats
- Pharmacogenomics of Anticancer Agents PAAR UO1GM61393
- Pharmacogenomics Research Network (PGRN) Statistical Analysis Resource (P-STAR) U19 HL065962
- Conte Center grant P50MH094267

Lipid Markers AUC

Manickam et al 2011 J Clinical Lipidology



LDL AUC ~ 0.52

HDL AUC = 0.60

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,