Genome, Phenome, and What Happens in Between

Hae Kyung Im, PhD

THE UNIVERSITY OF CHICAGO

GSK, Philadelphia
June 8, 2017

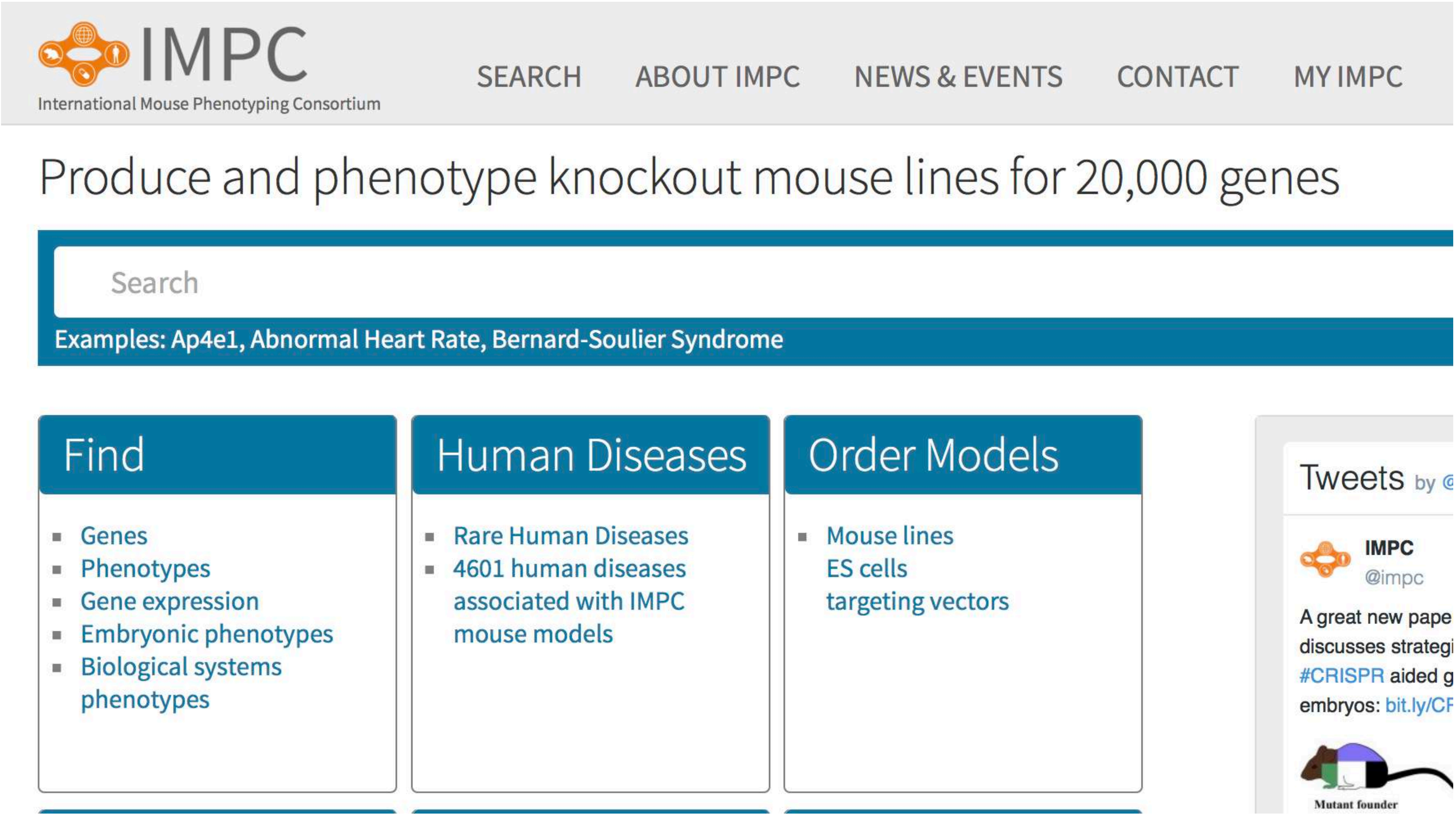
Goal

To develop statistical and computational methods to sift through large amounts of genomic and other high dimensional data to make discoveries that can be translated to improve human health.

Specific Goal

To catalog the phenotypic consequences of gene expression variation in humans

Model Organism Knock Out



IMPC

International Mouse Phenotyping Consortium

Produce and phenotype knockout mouse lines for 20,000 genes

Examples: Ap4e1, Abnormal Heart Rate, Bernard-Soulier Syndrome

Find

- Genes
- Phenotypes
- Gene expression
- Embryonic phenotypes
- Biological systems phenotypes

Human Diseases

- Rare Human Diseases
- 4601 human diseases associated with IMPC mouse models

Order Models

- Mouse lines

ES cells

targeting vectors

Tweets by @

IMPC

@impc

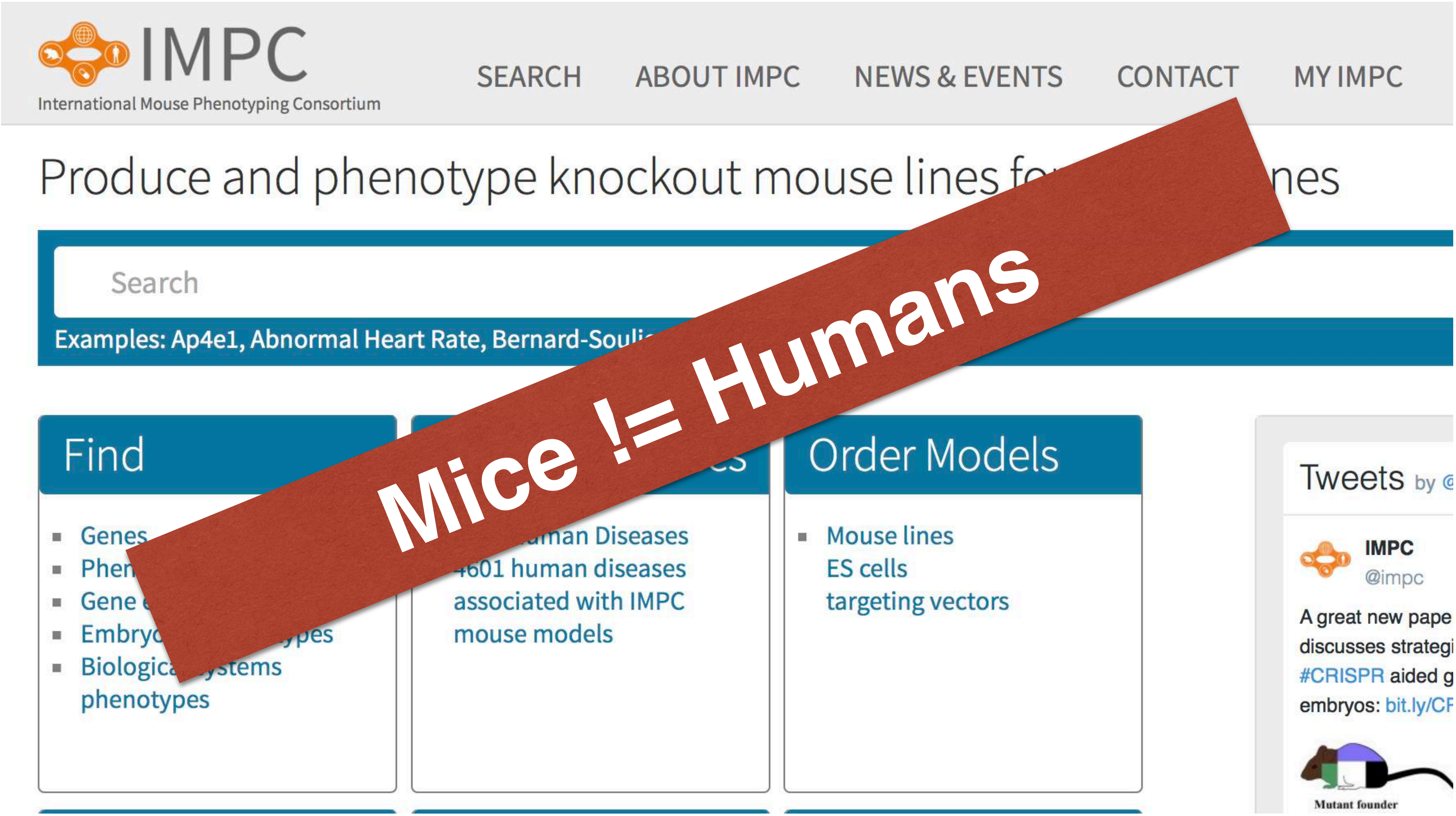
A great new pape

discusses strategi

#CRISPR aided g

embryos: bit.ly/CF

Model Organism Knockouts



IMPC
International Mouse Phenotyping Consortium

SEARCH ABOUT IMPC NEWS & EVENTS CONTACT MY IMPC

Produce and phenotype knockout mouse lines fo nes

Search

Examples: Ap4e1, Abnormal Heart Rate, Bernard-Souli

Mice != Humans

Find

- Genes
- Phen
- Gene e
- Embryo ypes
- Biologica ystems phenotypes

es

- uman Diseases
  4601 human diseases associated with IMPC mouse models

Order Models

- Mouse lines
  ES cells
  targeting vectors

Tweets by @

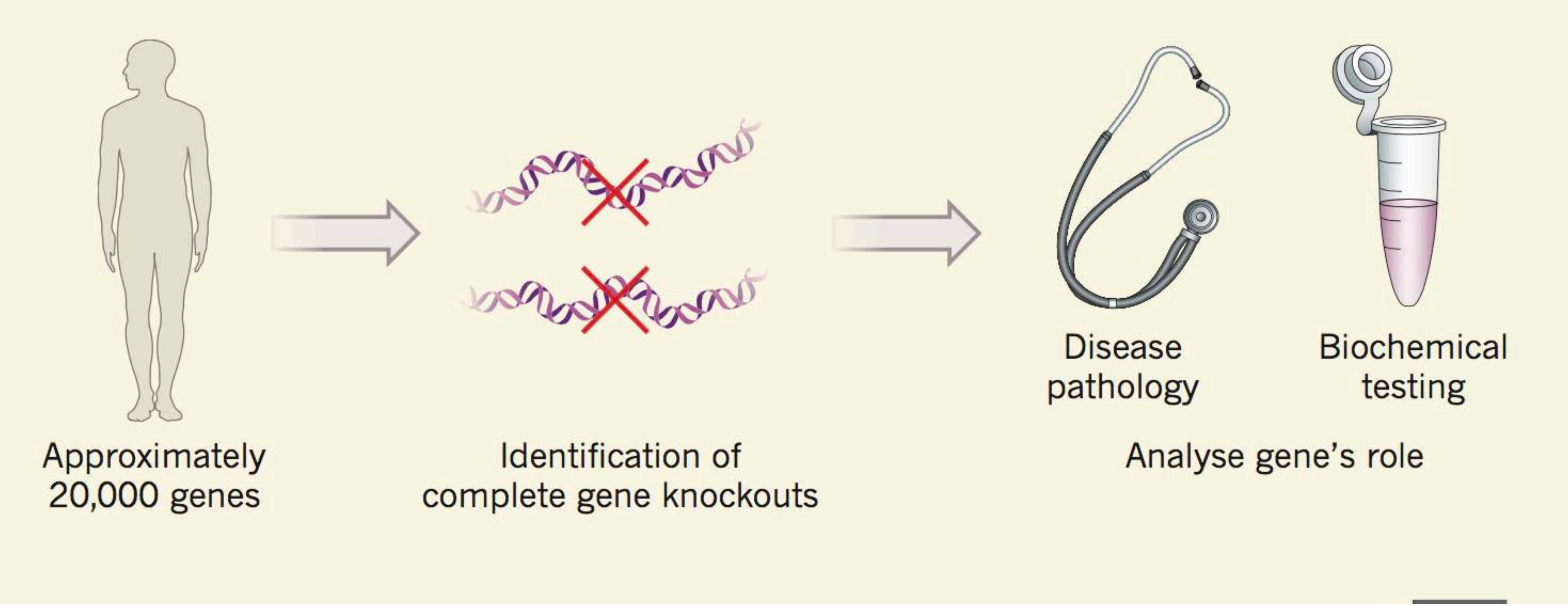
IMPC
@impc

A great new pape discusses strategi #CRISPR aided g embryos: bit.ly/CF

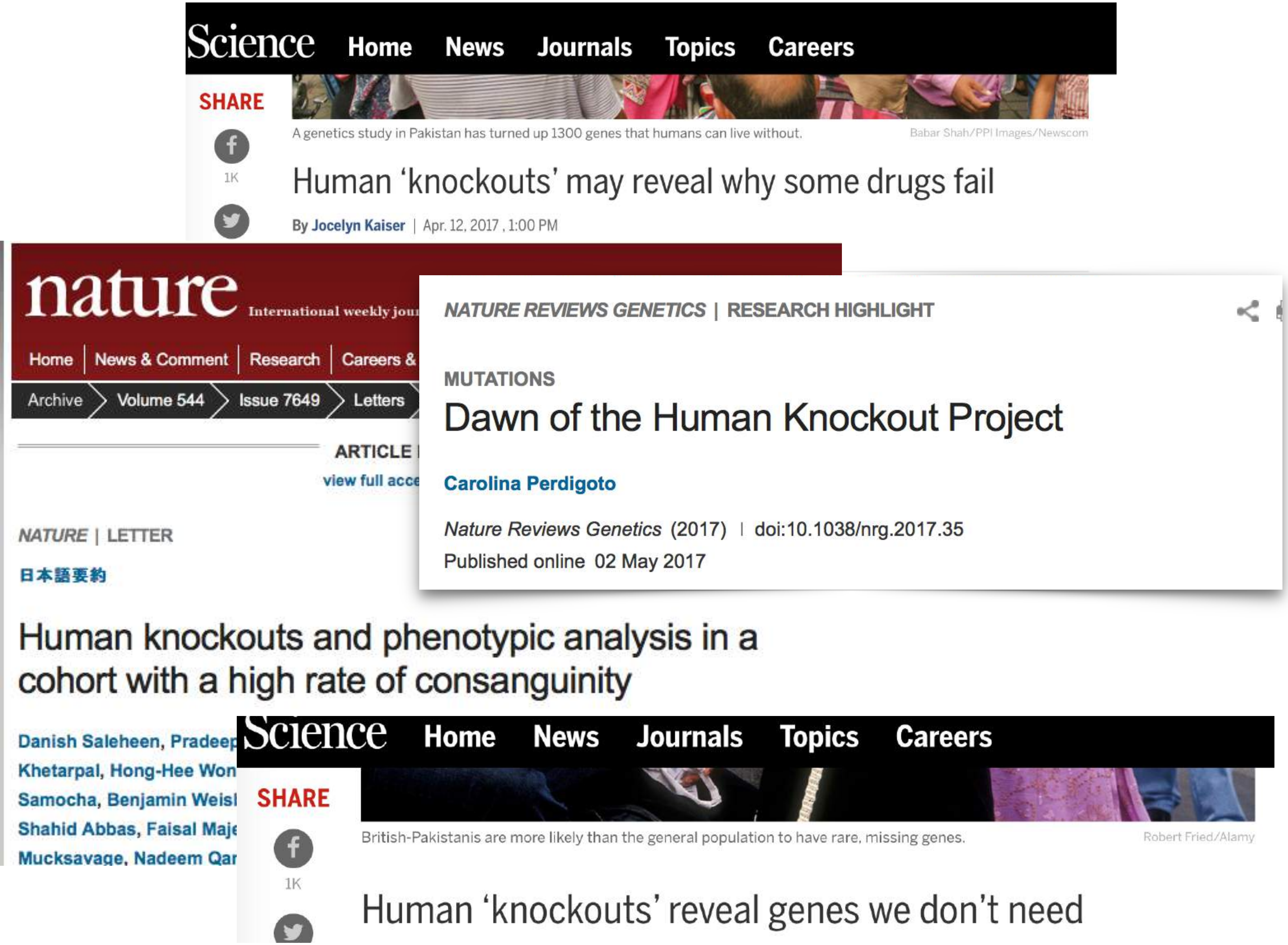
Mutant founder

Natural Human Knockouts

- People with loss of function mutations in both copies of the gene
- Natural experiments
- We can measure phenotypes to learn function of the gene


Approximately 20,000 genes
Identification of complete gene knockouts
Disease pathology
Biochemical testing
Analyse gene's role

R. M. Plenge, "Human genes lost and their functions found," Nature, 2017.



Science Home News Journals Topics Careers

SHARE

1K

A genetics study in Pakistan has turned up 1300 genes that humans can live without.

Babar Shah/PPI Images/Newscom

Human ‘knockouts’ may reveal why some drugs fail

By Jocelyn Kaiser | Apr. 12, 2017 , 1:00 PM

nature International weekly jou

Home | News & Comment | Research | Careers &

Archive > Volume 544 > Issue 7649 > Letters >

ARTICLE

view full acce

NATURE | LETTER

日本語要約

Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity

Danish Saleheen, Pradeep
Khetarpal, Hong-Hee Won
Samocha, Benjamin Weis
Shahid Abbas, Faisal Maje
Mucksavage, Nadeem Qar

NATURE REVIEWS GENETICS | RESEARCH HIGHLIGHT

MUTATIONS

Dawn of the Human Knockout Project

Carolina Perdigoto

Nature Reviews Genetics (2017) | doi:10.1038/nrg.2017.35

Published online 02 May 2017

Science Home News Journals Topics Careers

SHARE

1K

British-Pakistanis are more likely than the general population to have rare, missing genes.

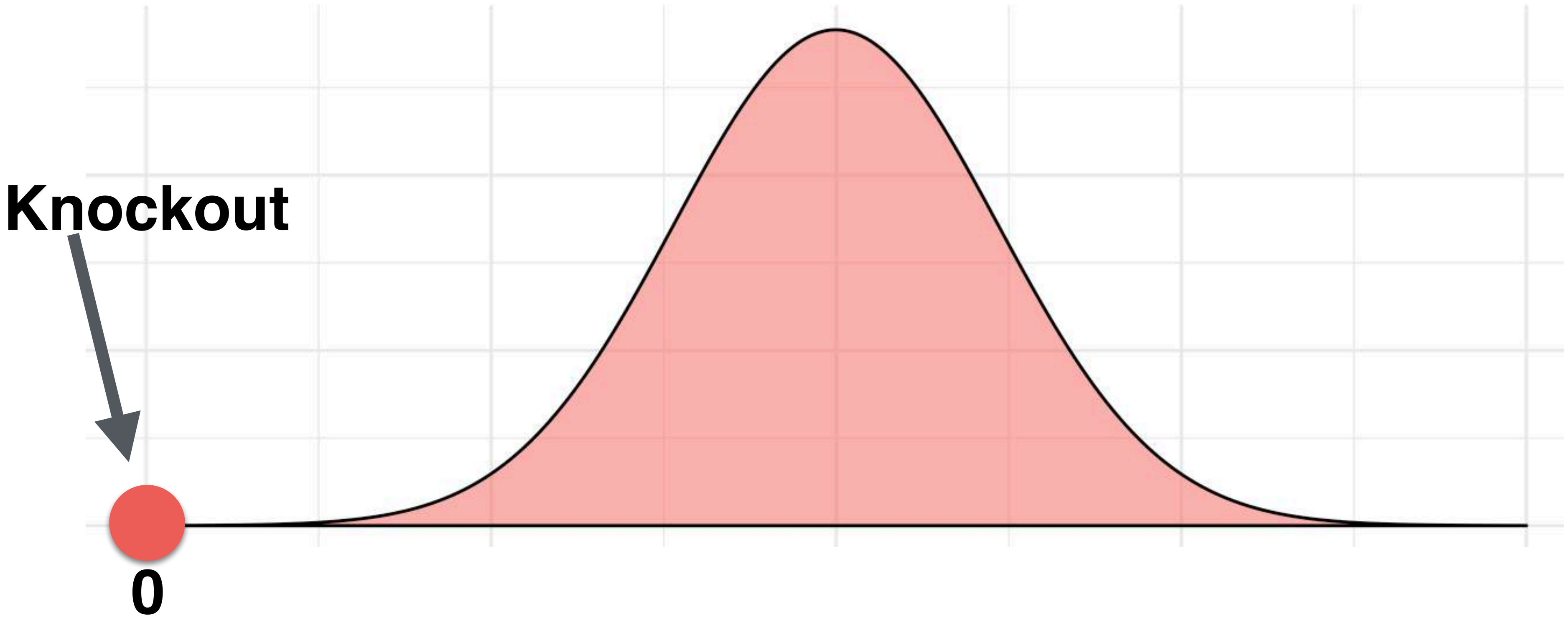
Robert Fried/Alamy

Human ‘knockouts’ reveal genes we don’t need

Human Knockout Project

Systematic effort to understand the consequence of complete disruption of every human gene

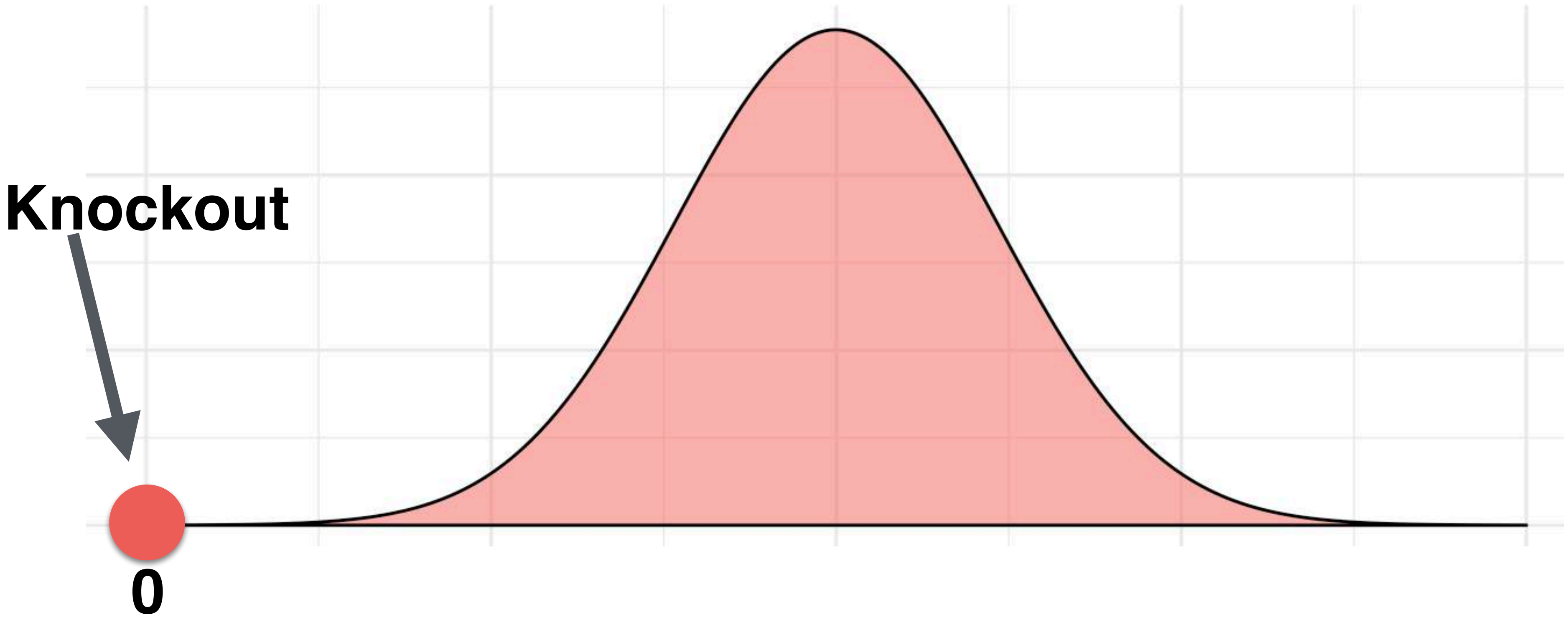
Human Knockout Project



Systematic effort to understand the consequence of complete disruption of every human gene

Human ~~Knockout~~ Project

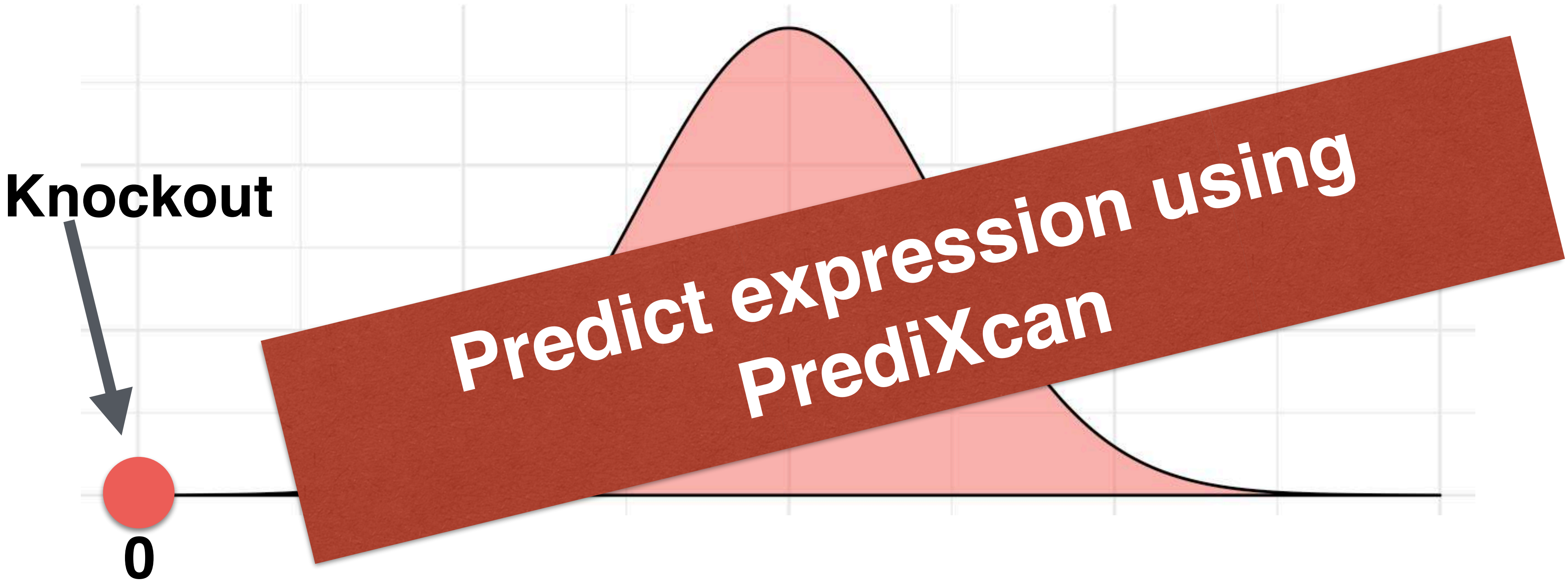
“Knockdown”



Systematic effort to understand the consequence of ~~complete~~ **partial** disruption of every human gene

Human ~~Knockout~~ Project

“Knockdown”



Predict expression using PrediXcan

Systematic effort to understand the consequence of ~~complete~~ **partial** disruption of every human gene

Human Knockout vs “Knockdown”

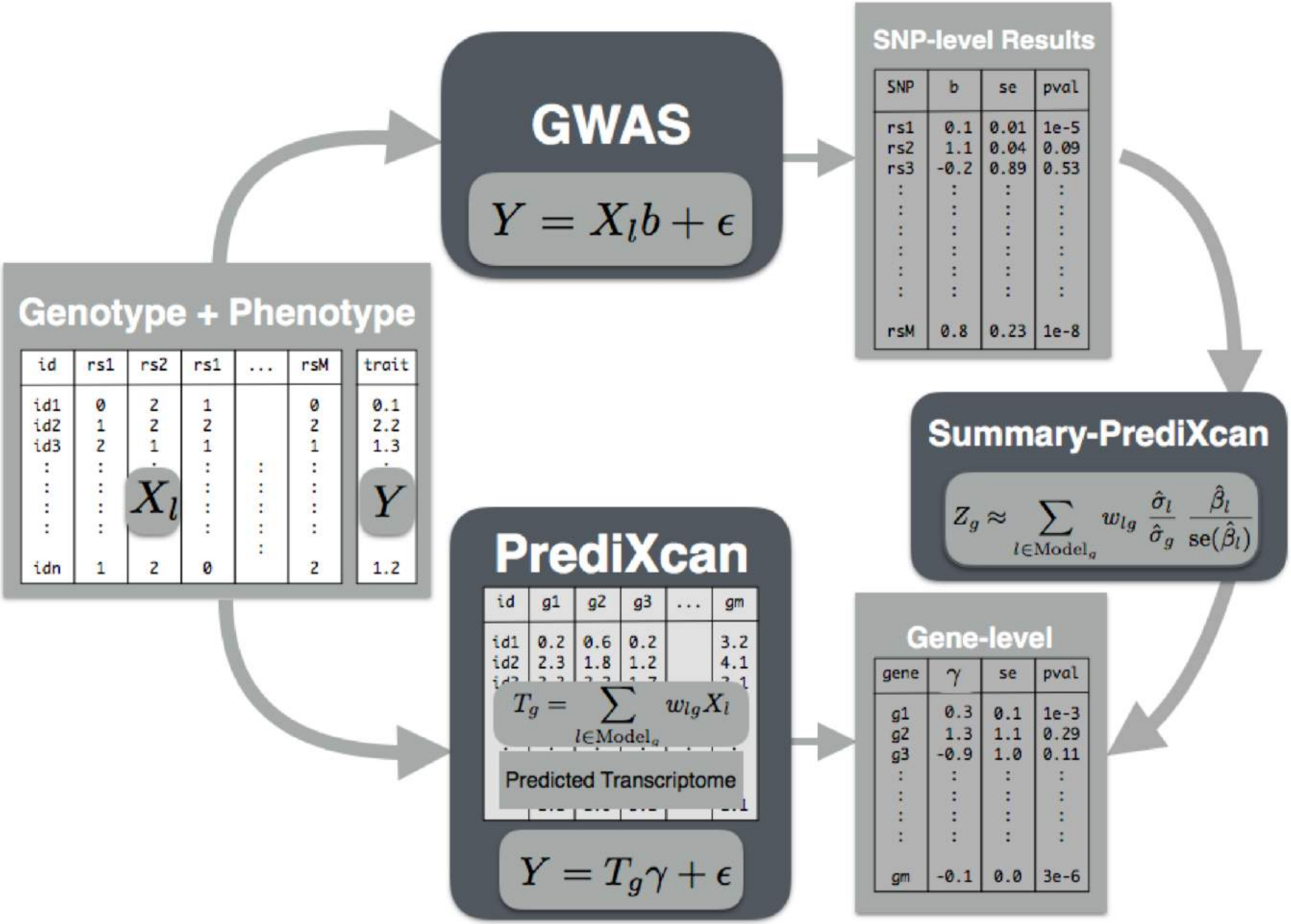
Knockout	Knockdown
• Large effect sizes	• Small effect sizes • LD-contamination • Pleiotropy
• Small sample size • Need to sequence large number of individuals	• Large sample size • Cheaper genotyping may be enough • Sequence data can be used

Small Effect Sizes

- Increase sample size

- to address burden of larger sample sizes
    - use Summary-PrediXcan

Summary-PrediXcan



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

X_l Y

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

Summary-PrediXcan

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

PrediXcan

id	g1	g2	g3	...	gm
id1	0.2	0.6	0.2		3.2
id2	2.3	1.8	1.2		4.1

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

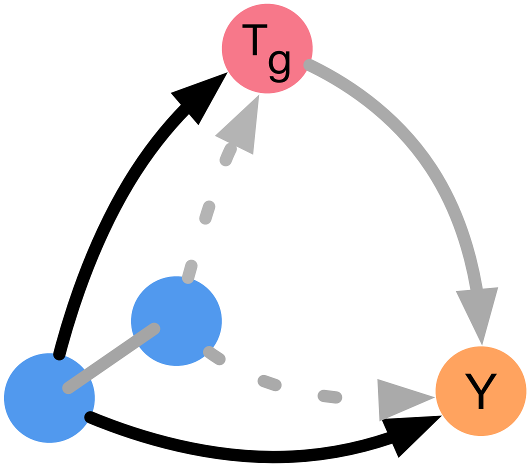
Predicted Transcriptome

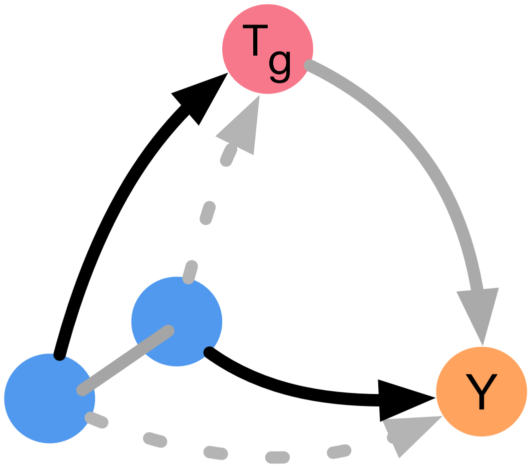
$$Y = T_g \gamma + \epsilon$$

Gene-level

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

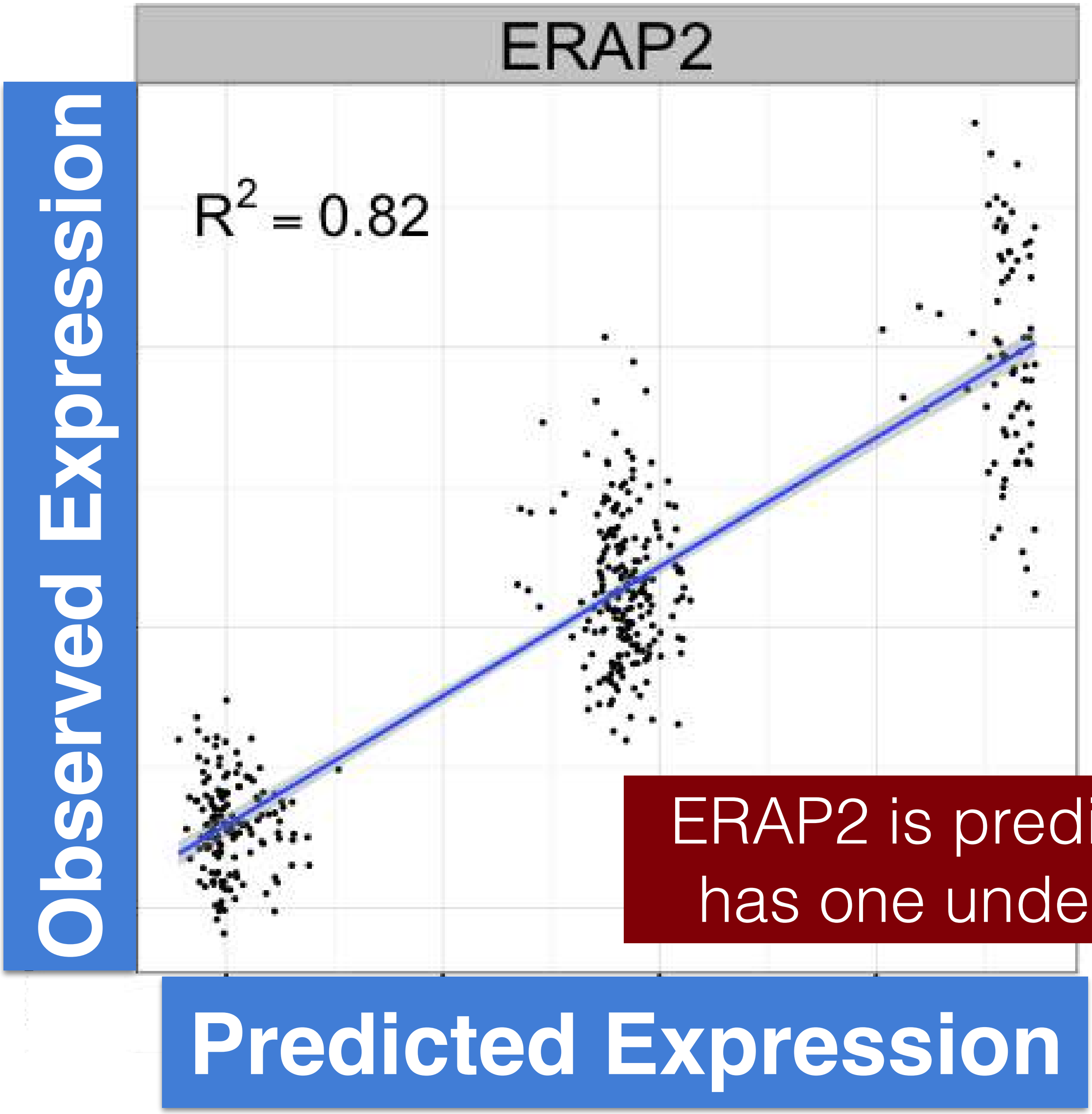
LD-Contamination





- Compute colocalization and discard if GWAS and eQTL signals are independent
  - COLOC (Giambartolomei et al, PLoS Genetics 2013)
  - RTC (Nica, …, Dermitzakis et al 2010)
  - eCAVIAR (Hormozdiari … Eskin et al, AJGH 2017)
  - ENLOC (Wen et al, PLoS Genetics 2017)
  - HEIDI (Zhu, …, Visscher, Yang, Nature Gen. 2016)

Use Causal rather than Associated SNPs



ERAP2 is predicted by 80 SNPs but has one underlying causal variant

How to Mitigate LD-Contamination

- Post filtering step with COLOC or other methods

- Current prediction models are purely statistical
  - Use causal predictors to reduce chance of LD
  - S-PrediXcan needs to efficiently impute GWAS results for causal variants

Pleiotropy

- Experiments in model systems

Best Practices Framework: MetaXcan

	Train Prediction Models	Filter-out Unreliable Models	Perform Gene-level Association	Adjust for Model Uncertainty	Filter-out Non-Colocalized Signals
PrediXcan	Elastic-Net	Correlation Observed vs Predicted FDR<0.05	Individual Level Data Y ~ Predicted X		Colocalization status (COLOC)
S-PrediXcan	Elastic-Net	Correlation Observed vs Predicted FDR<0.05	Summary Data $\sum_{l\in \text{Model}_{T_g}} w_{X_l,T_g} \frac{\hat{\sigma}_l}{\hat{\sigma}_{T_g}} Z_{X_l,Y}$		Colocalization status (COLOC)

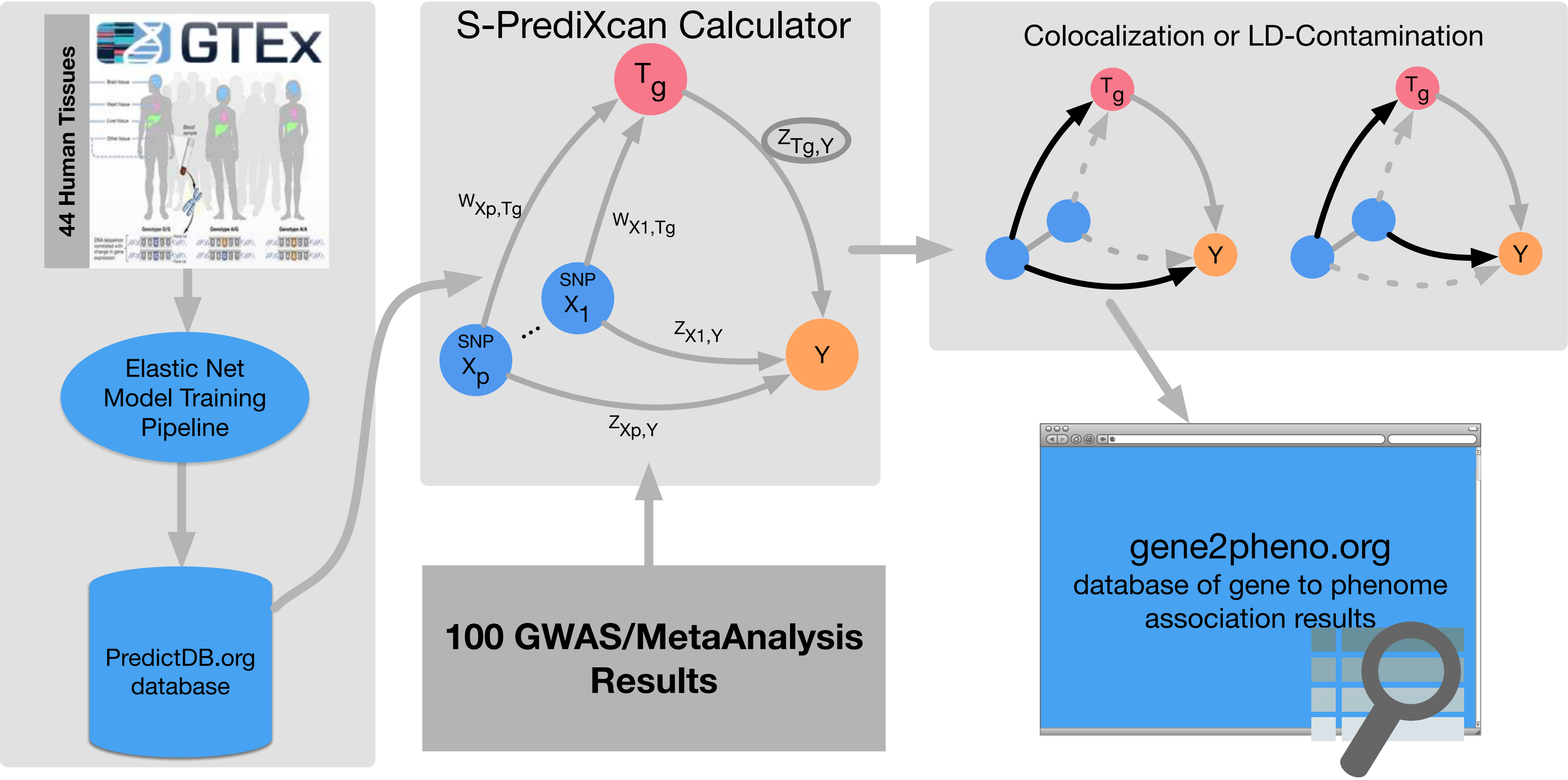
Causal models

Incorporate Model Uncertainty

Imputation of GWAS results

Allow for multiple causal variants in COLOC

Computing Phenotypic Consequences with Summary Stat



Elastic Net Model Training Pipeline

PredictDB.org database

S-PrediXcan Calculator

100 GWAS/MetaAnalysis Results

Colocalization or LD-Contamination

gene2pheno.org

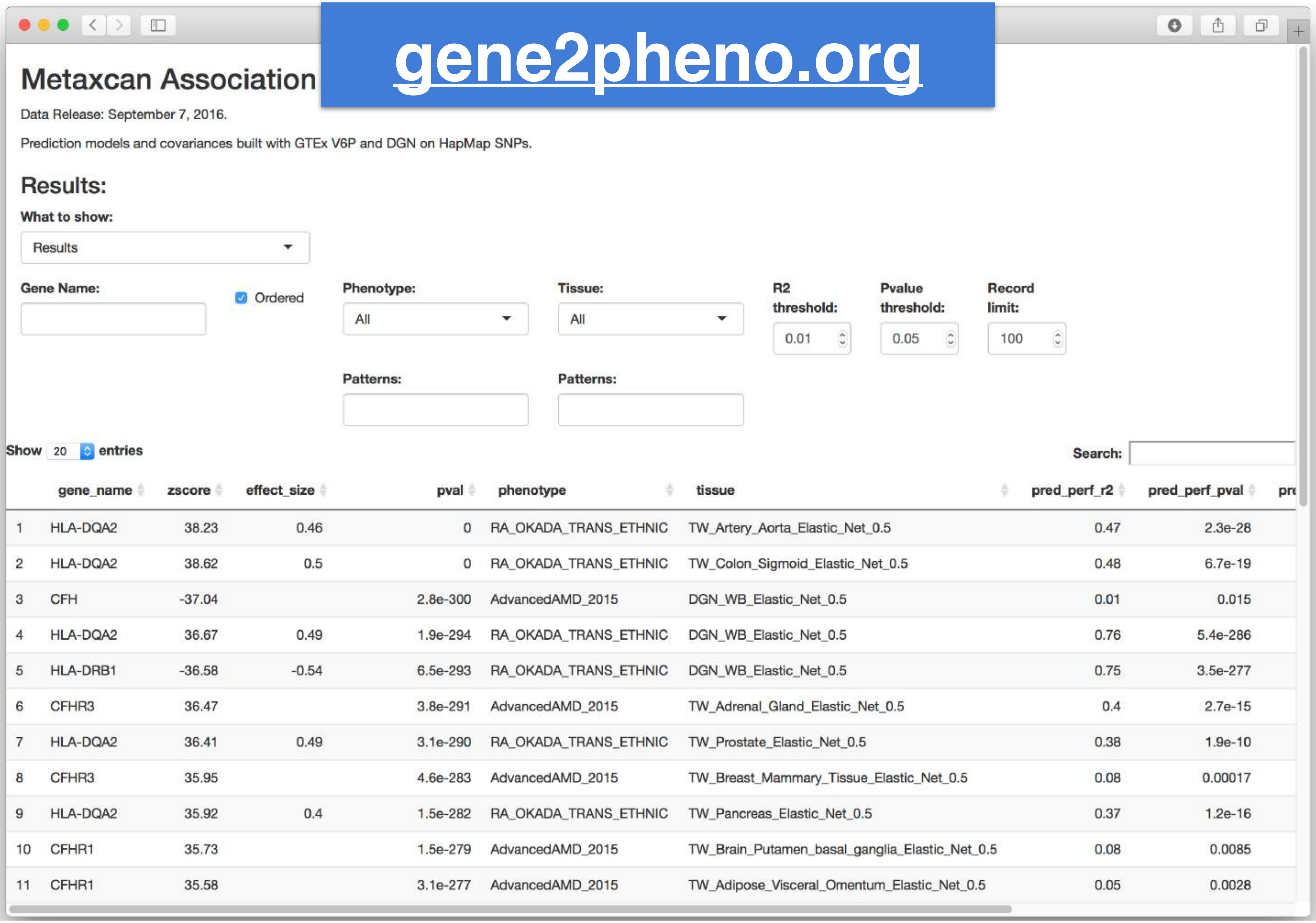
database of gene to phenome association results

https://github.com/hakyimlab/PredictDBPipeline

https://github.com/hakyimlab/MetaXcan

Catalog of Phenotypic Consequences

gene2pheno.org



Metaxcan Association

Data Release: September 7, 2016.

Prediction models and covariances built with GTEx V6P and DGN on HapMap SNPs.

Results:

What to show: Results

Gene Name: ☑ Ordered

Phenotype: All

Tissue: All

R2 threshold: 0.01

Pvalue threshold: 0.05

Record limit: 100

Patterns:

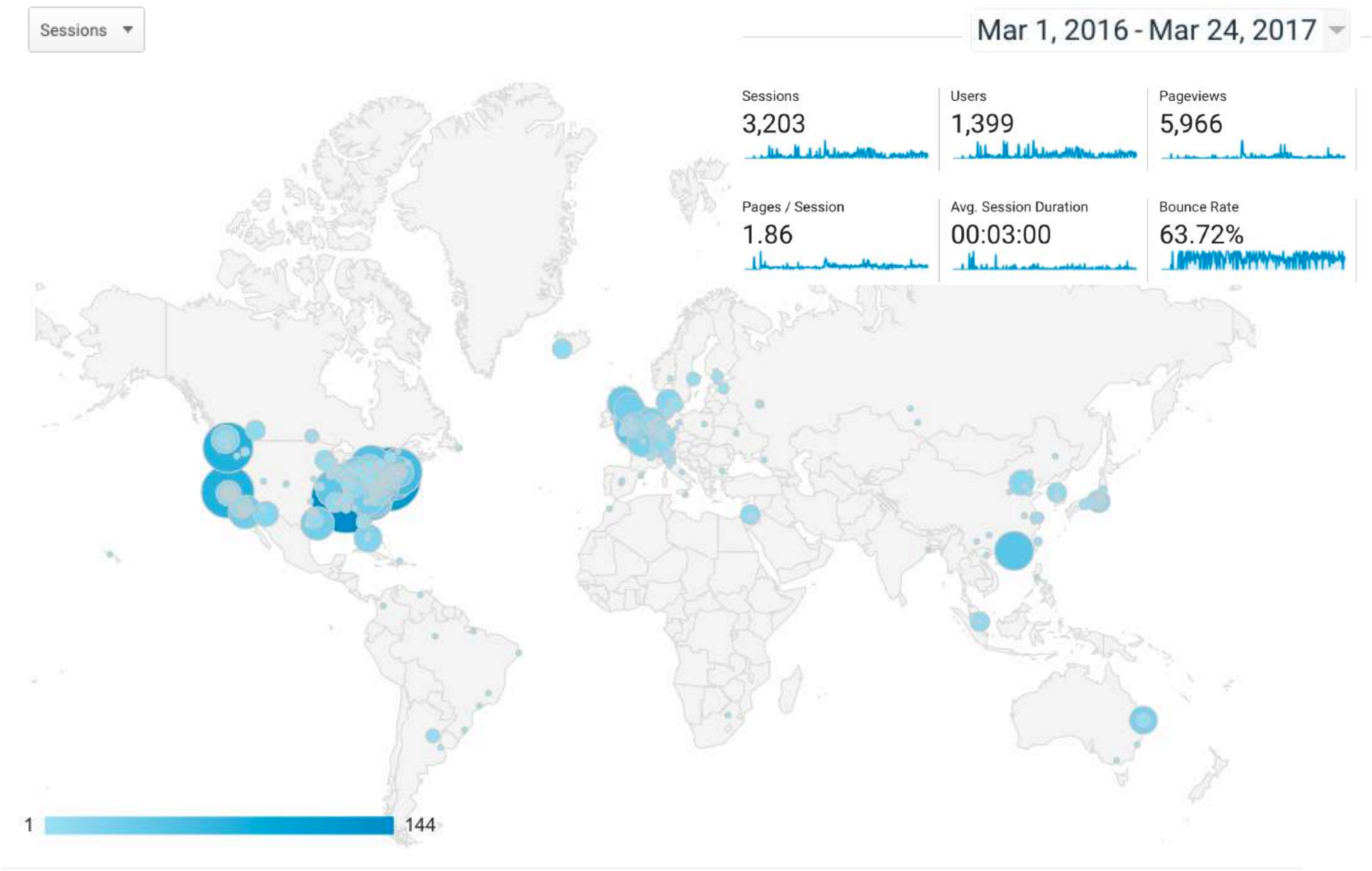
Patterns:

Show 20 entries

Search:

	gene_name	zscore	effect_size	pval	phenotype	tissue	pred_perf_r2	pred_perf_pval	pre
1	HLA-DQA2	38.23	0.46	0	RA_OKADA_TRANS_ETHNIC	TW_Artery_Aorta_Elastic_Net_0.5	0.47	2.3e-28	
2	HLA-DQA2	38.62	0.5	0	RA_OKADA_TRANS_ETHNIC	TW_Colon_Sigmoid_Elastic_Net_0.5	0.48	6.7e-19	
3	CFH	-37.04		2.8e-300	AdvancedAMD_2015	DGN_WB_Elastic_Net_0.5	0.01	0.015	
4	HLA-DQA2	36.67	0.49	1.9e-294	RA_OKADA_TRANS_ETHNIC	DGN_WB_Elastic_Net_0.5	0.76	5.4e-286	
5	HLA-DRB1	-36.58	-0.54	6.5e-293	RA_OKADA_TRANS_ETHNIC	DGN_WB_Elastic_Net_0.5	0.75	3.5e-277	
6	CFHR3	36.47		3.8e-291	AdvancedAMD_2015	TW_Adrenal_Gland_Elastic_Net_0.5	0.4	2.7e-15	
7	HLA-DQA2	36.41	0.49	3.1e-290	RA_OKADA_TRANS_ETHNIC	TW_Prostate_Elastic_Net_0.5	0.38	1.9e-10	
8	CFHR3	35.95		4.6e-283	AdvancedAMD_2015	TW_Breast_Mammary_Tissue_Elastic_Net_0.5	0.08	0.00017	
9	HLA-DQA2	35.92	0.4	1.5e-282	RA_OKADA_TRANS_ETHNIC	TW_Pancreas_Elastic_Net_0.5	0.37	1.2e-16	
10	CFHR1	35.73		1.5e-279	AdvancedAMD_2015	TW_Brain_Putamen_basal_ganglia_Elastic_Net_0.5	0.08	0.0085	
11	CFHR1	35.58		3.1e-277	AdvancedAMD_2015	TW_Adipose_Visceral_Omentum_Elastic_Net_0.5	0.05	0.0028	

PredictDB.org Users



Mar 1, 2016 - Mar 24, 2017

Sessions
3,203

Users
1,399

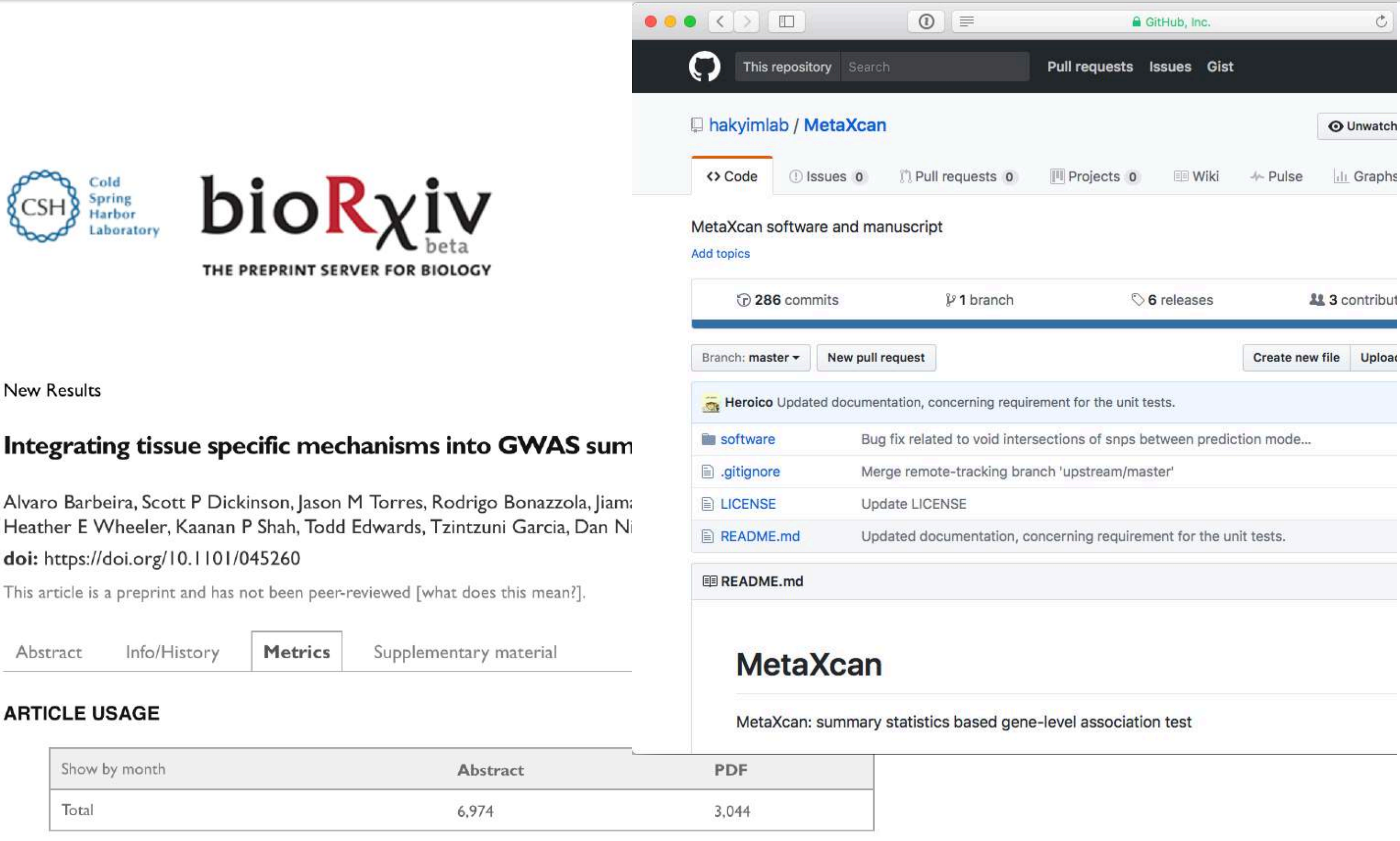
Pageviews
5,966

Pages / Session
1.86

Avg. Session Duration
00:03:00

Bounce Rate
63.72%

MetaXcan manuscript in BioRxiv - Revision Under Review



Cold Spring Harbor Laboratory

bioRxiv beta

THE PREPRINT SERVER FOR BIOLOGY

New Results

Integrating tissue specific mechanisms into GWAS sum

Alvaro Barbeira, Scott P Dickinson, Jason M Torres, Rodrigo Bonazzola, Jiam
Heather E Wheeler, Kaanan P Shah, Todd Edwards, Tzintzuni Garcia, Dan Ni

doi: https://doi.org/10.1101/045260

This article is a preprint and has not been peer-reviewed [what does this mean?].

ARTICLE USAGE

Show by month	Abstract	PDF
Total	6,974	3,044

hakyimlab / MetaXcan

MetaXcan software and manuscript

286 commits | 1 branch | 6 releases | 3 contribut

Heroico Updated documentation, concerning requirement for the unit tests.

software	Bug fix related to void intersections of snps between prediction mode...
.gitignore	Merge remote-tracking branch 'upstream/master'
LICENSE	Update LICENSE
README.md	Updated documentation, concerning requirement for the unit tests.

README.md

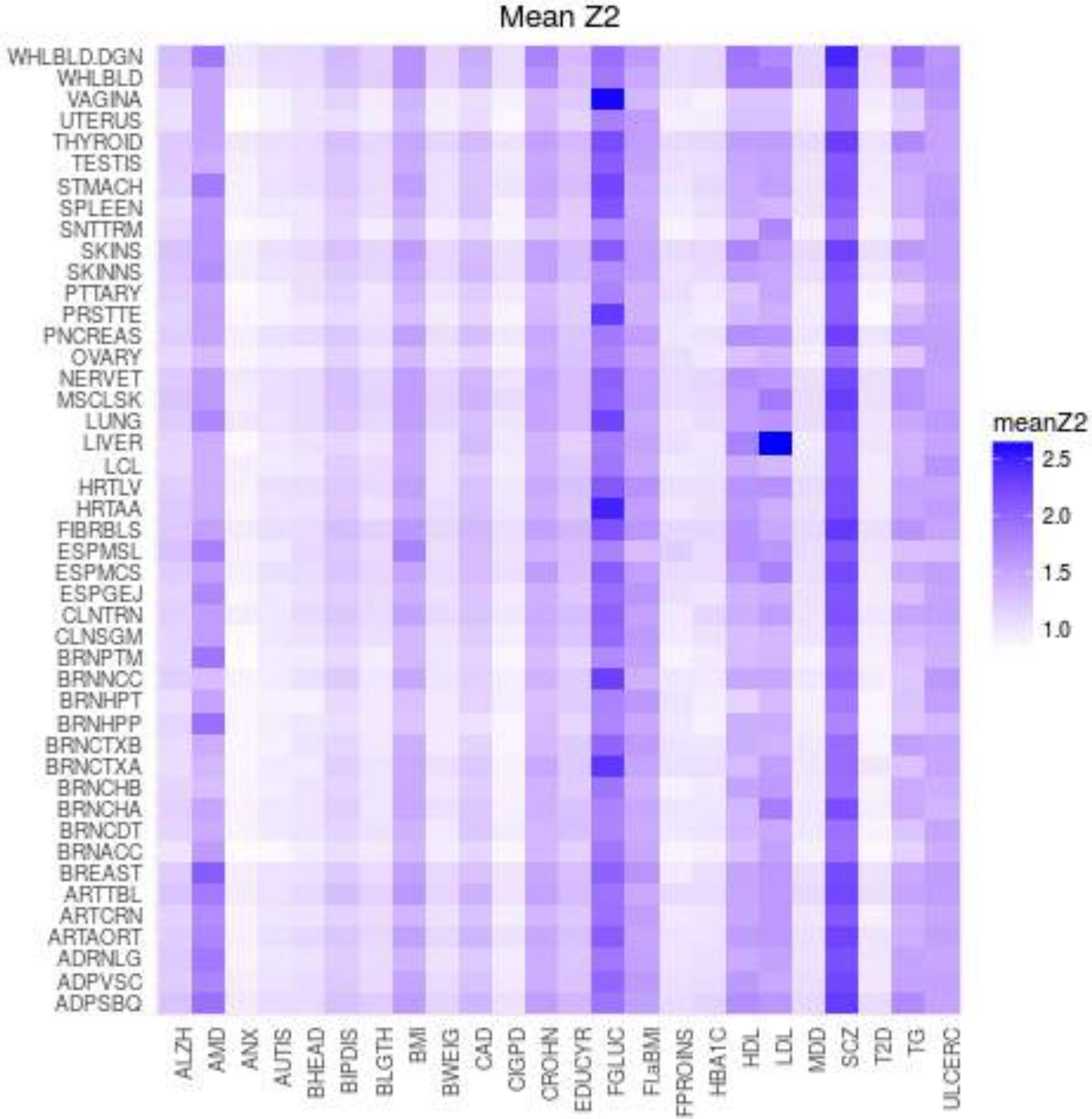
MetaXcan

MetaXcan: summary statistics based gene-level association test

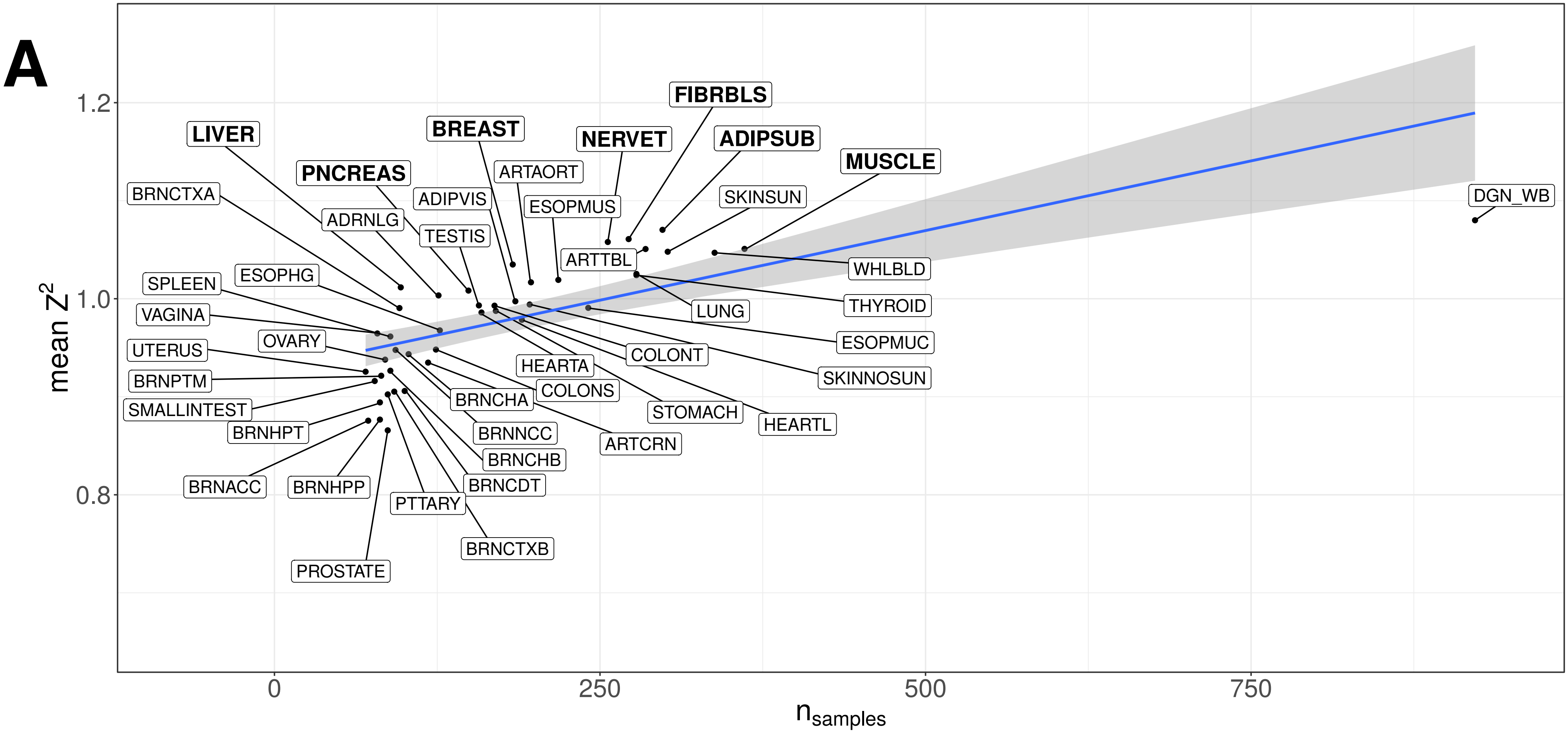
https://github.com/hakyimlab/MetaXcan

Tissue Specificity

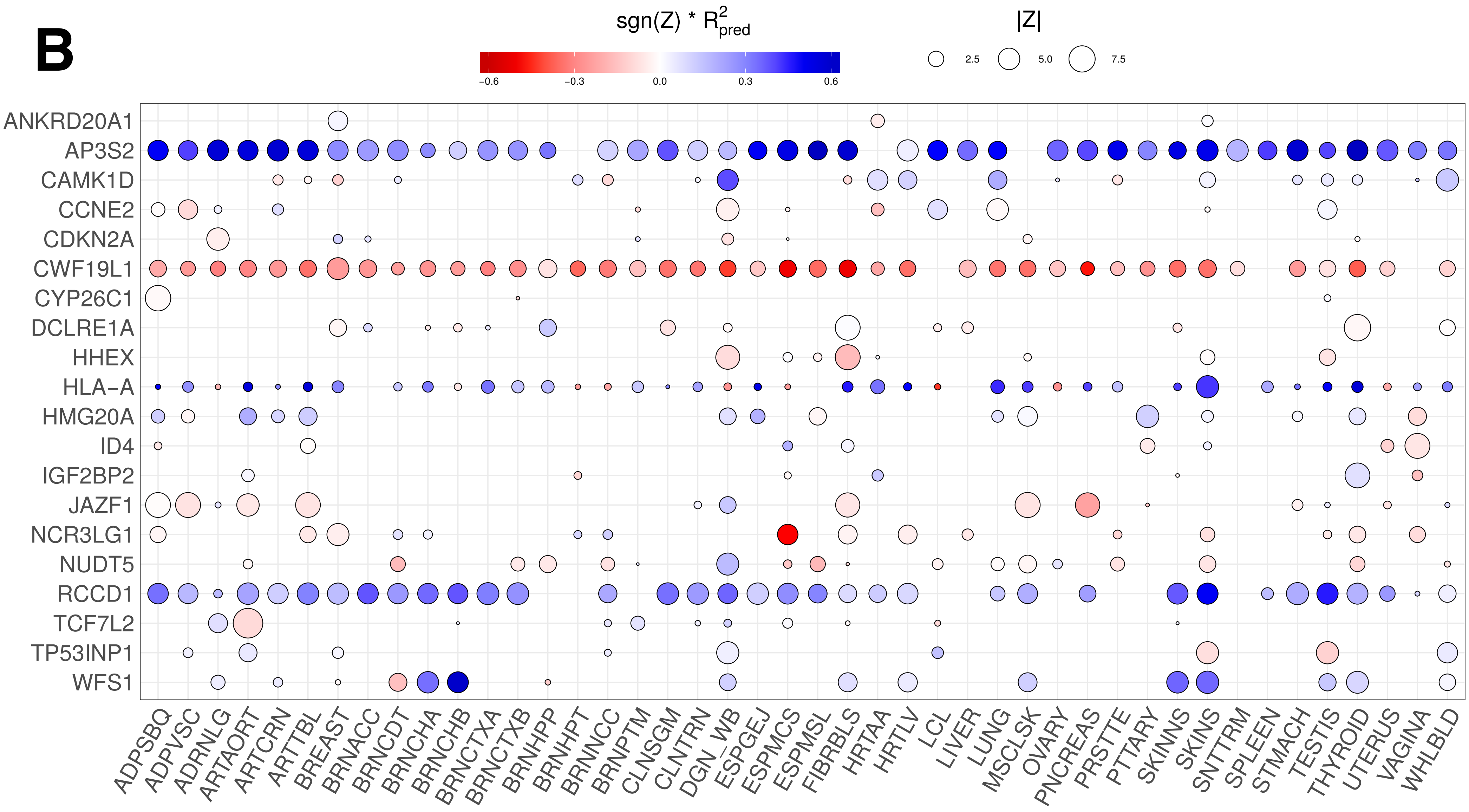
Target Genes Found across Multiple Tissues



Type 2 Diabetes Tissue Enrichment by Sample Size

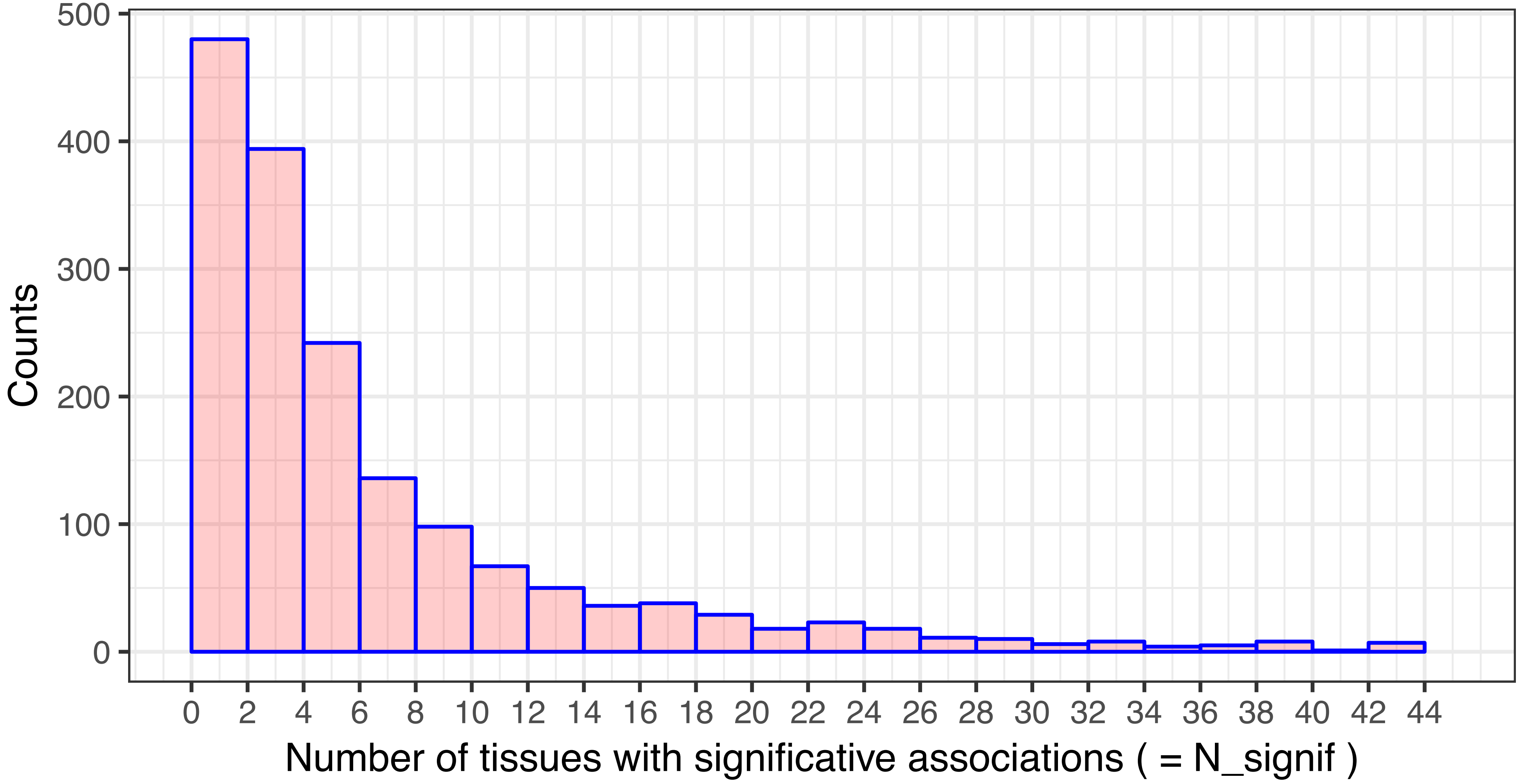


Top Type 2 Diabetes Gene Associations by Tissue

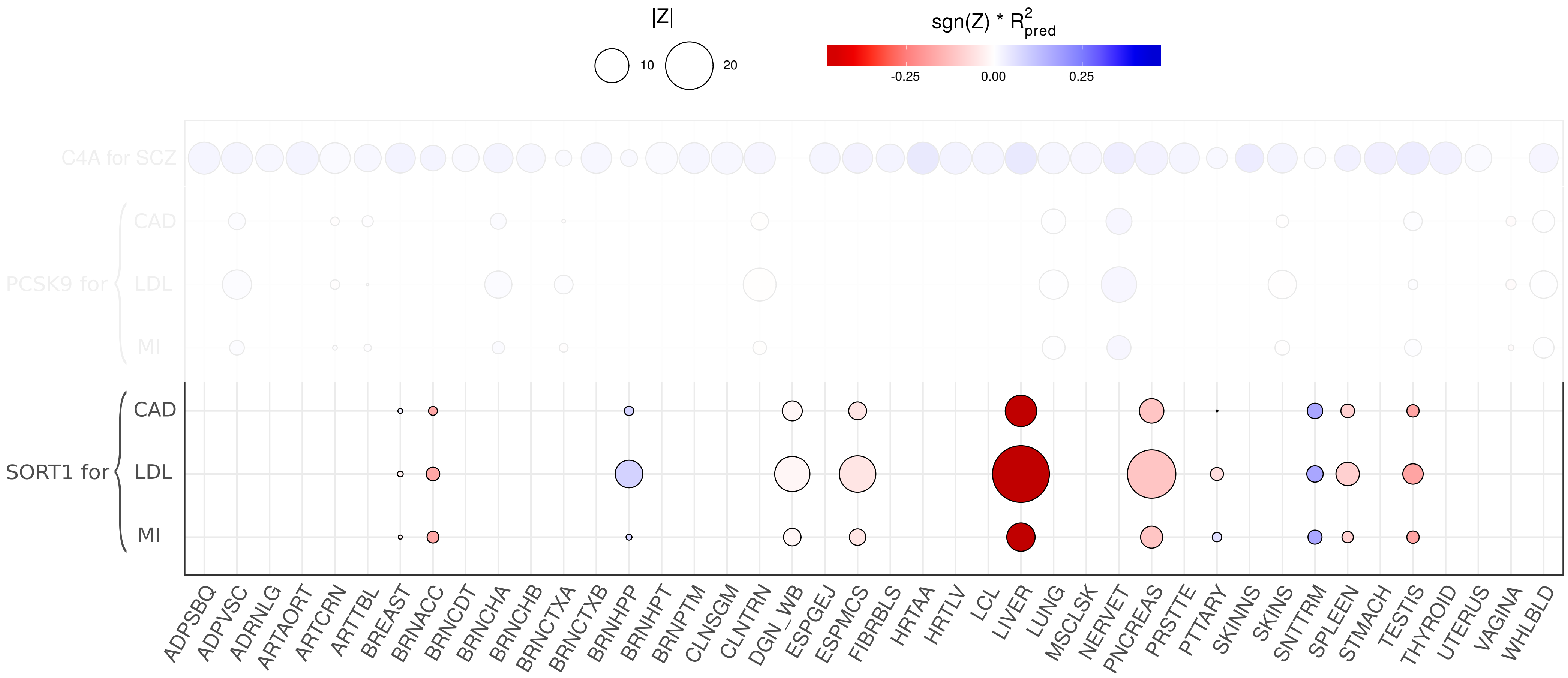


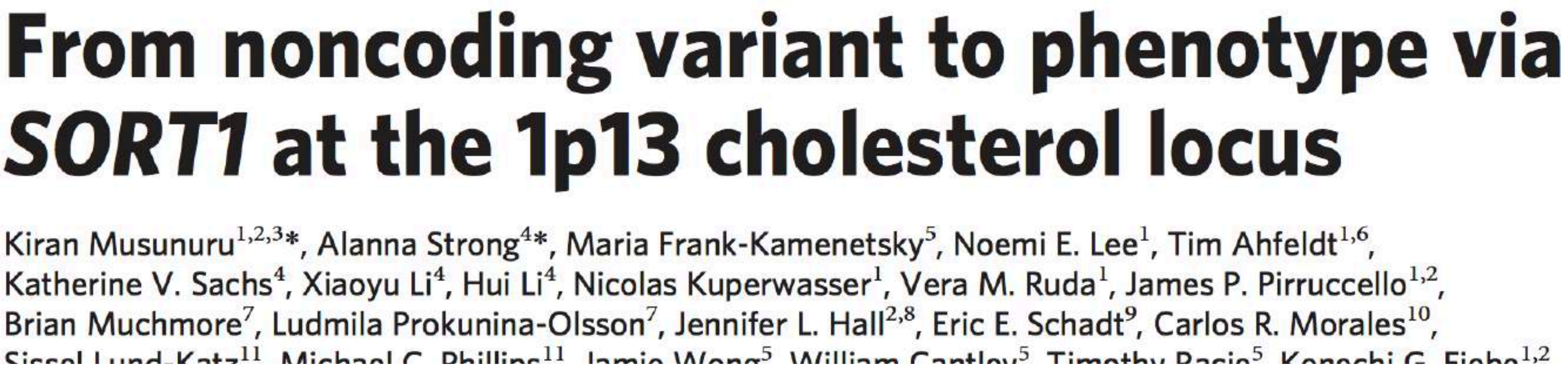
Most Associations Are Tissue Specific

Histogram for tissue−specificity
Phenotype: GIANT_HEIGHT
cutoff1 = 2.5e−07
cutoff2 = 0.05 / N_models, 1690 genes with more than 0 models.



SORT1 Strong Signal in Liver

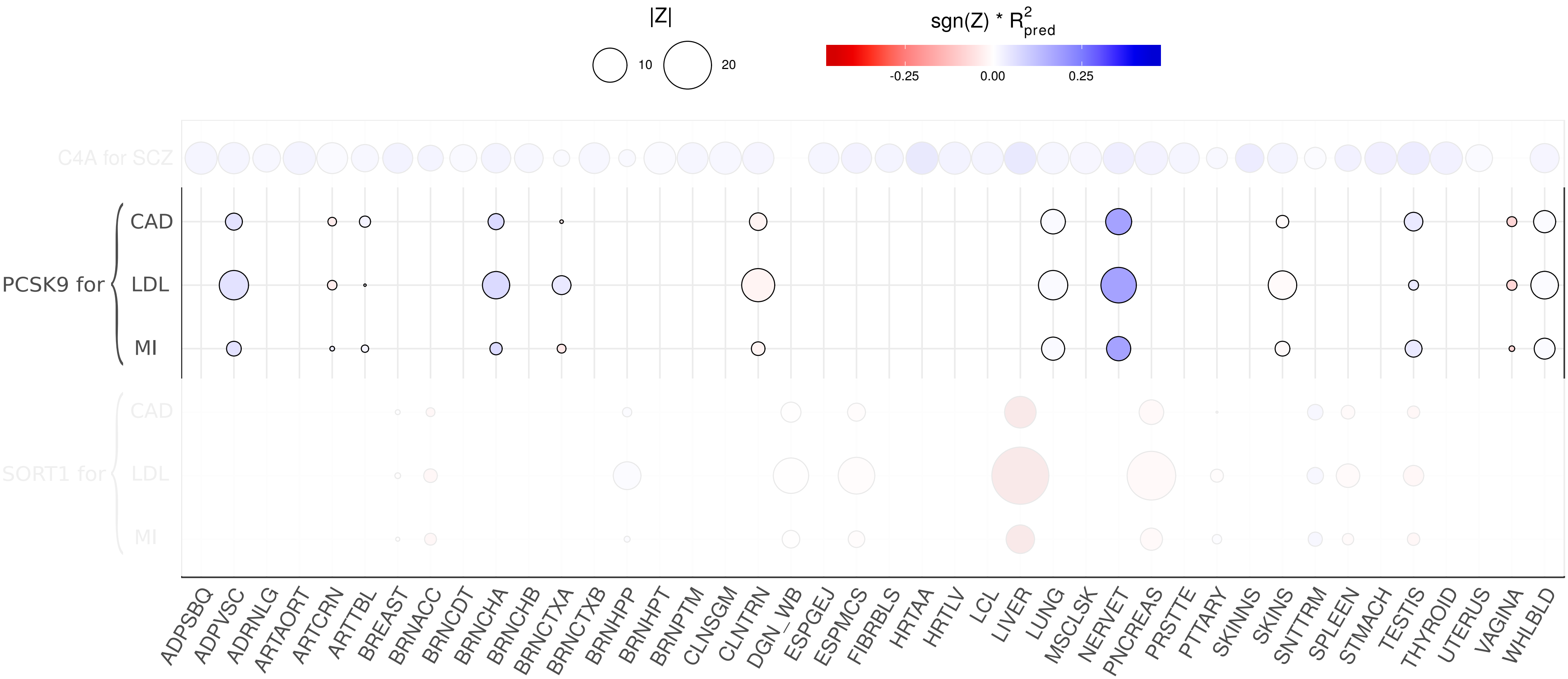


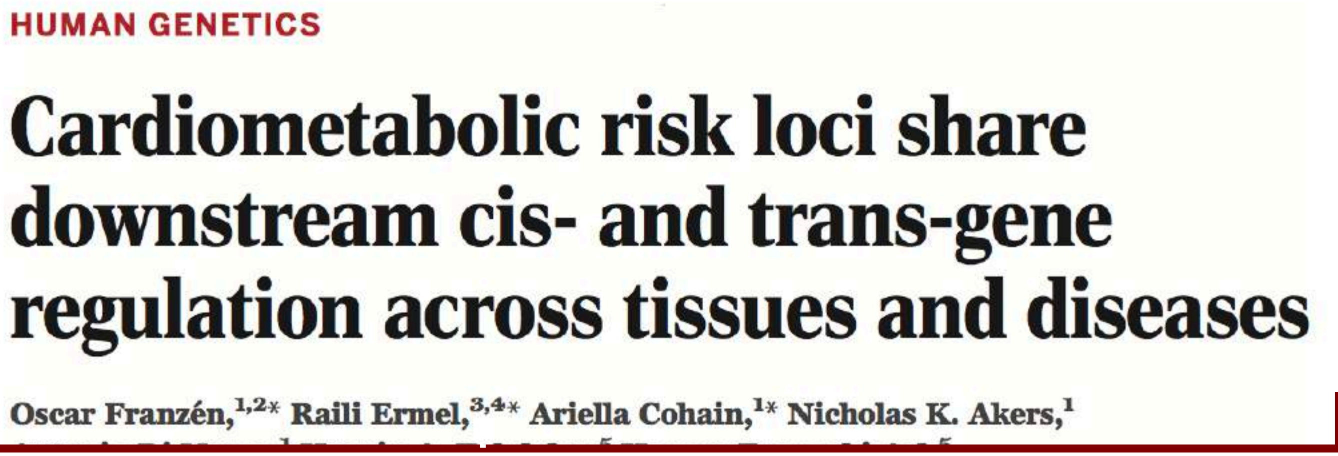


From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus

Kiran Musunuru[1,2,3]*, Alanna Strong[4]*, Maria Frank-Kamenetsky[5], Noemi E. Lee[1], Tim Ahfeldt[1,6], Katherine V. Sachs[4], Xiaoyu Li[4], Hui Li[4], Nicolas Kuperwasser[1], Vera M. Ruda[1], James P. Pirruccello[1,2], Brian Muchmore[7], Ludmila Prokunina-Olsson[7], Jennifer L. Hall[2,8], Eric E. Schadt[9], Carlos R. Morales[10], Sissel Lund-Katz[11], Michael C. Phillips[11], Jamie Wong[5], William Cantley[5], Timothy Racie[5], Kenechi G. Ejebe[1,2]

PCSK9 Colocalized in Tibial Nerve



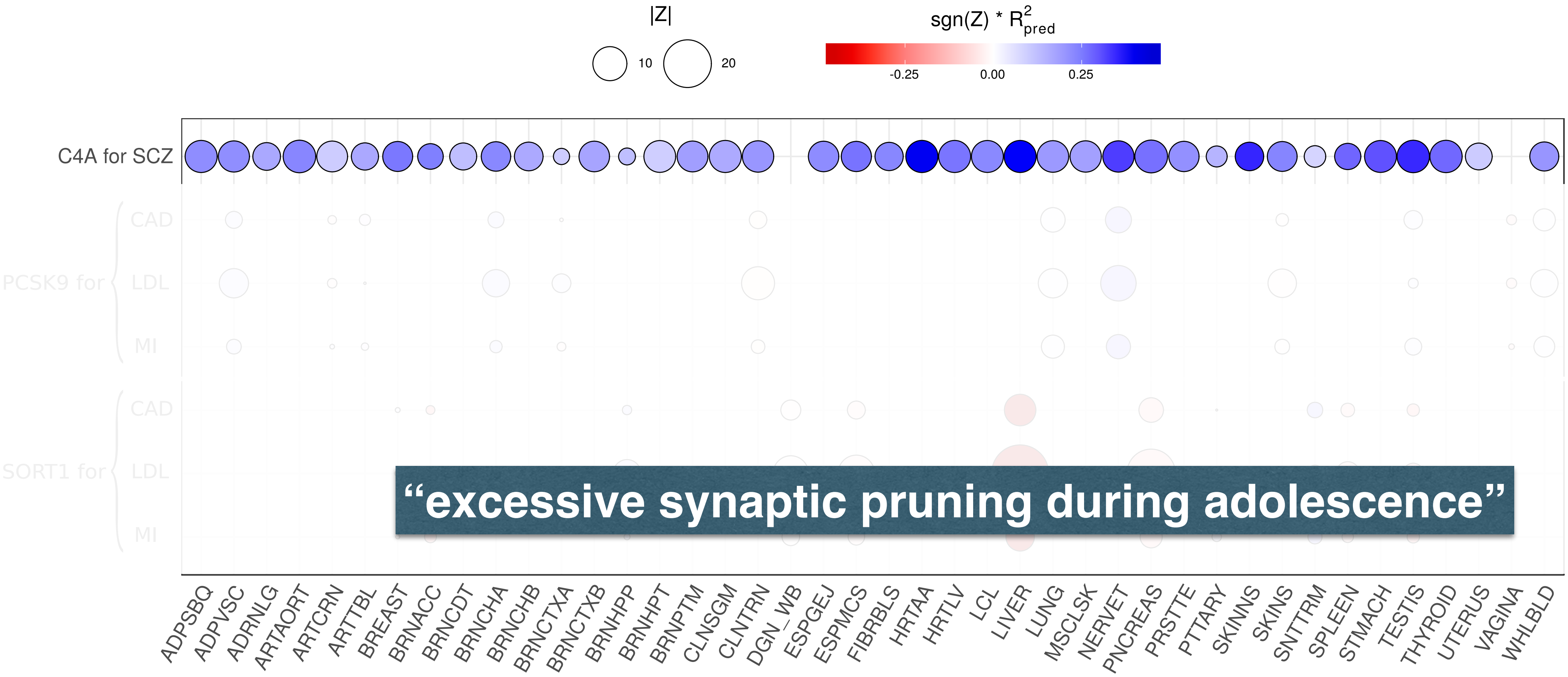


HUMAN GENETICS

Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases

Oscar Franzén,[1,2*] Raili Ermel,[3,4*] Ariella Cohain,[1*] Nicholas K. Akers,[1]

C4A Causal Gene

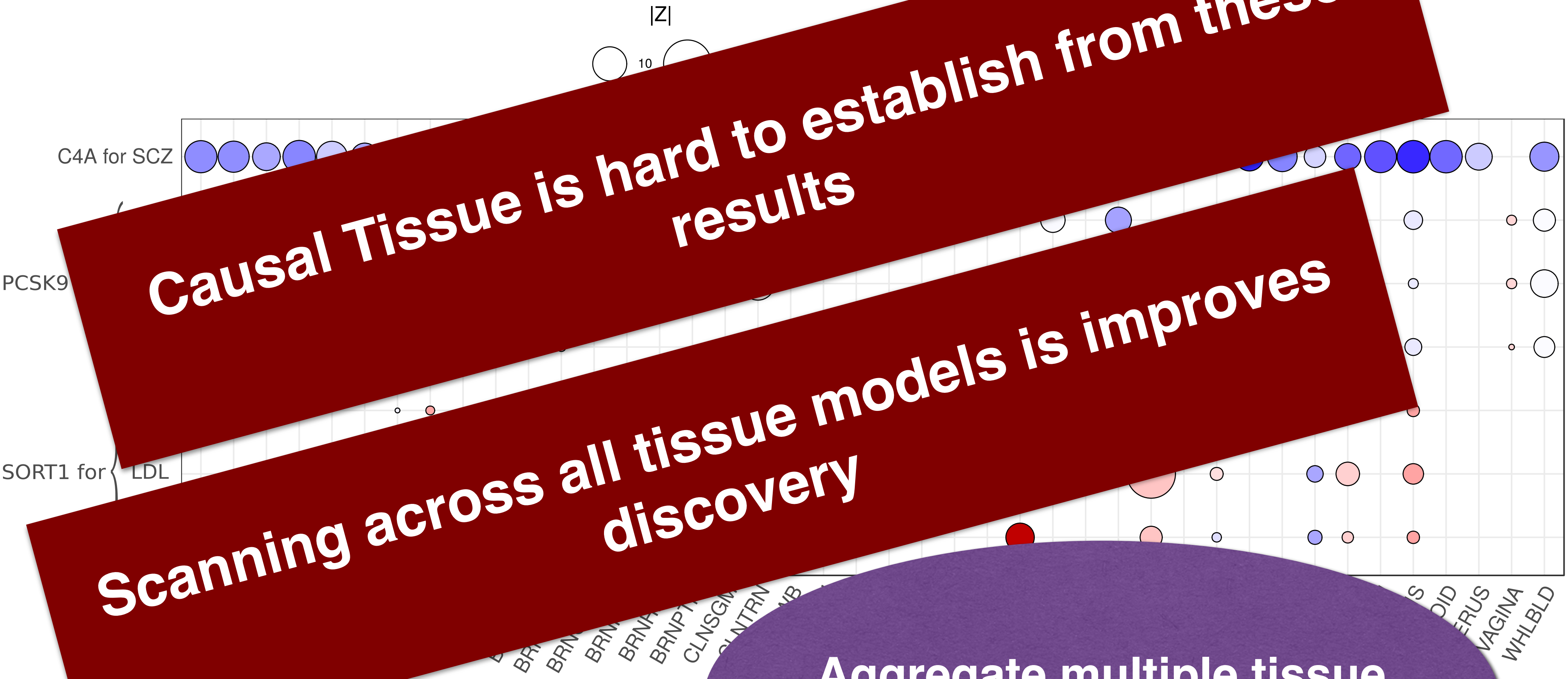


“excessive synaptic pruning during adolescence”

Schizophrenia risk from complex variation of complement component 4

Aswin Sekar[1,2,3], Allison R. Bialas[4,5], Heather de Rivera[1,2], Avery Davis[1,2], Timothy R. Hammond[4], Nolan Kamitaki[1,2], Katherine Tooley[1,2], Jessy Presumey[5], Matthew Baum[1,2,3,4], Vanessa Van Doren[1], Giulio Genovese[1,2], Samuel A. Rose[2], Robert E. Handsaker[1,2], Schizophrenia Working Group of the Psychiatric Genomics Consortium*, Mark J. Daly[2,6], Michael C. Carroll[5], Beth Stevens[2,4] & Steven A. McCarroll[1,2]

Cross Tissue and Tissue Specific Associations



Causal Tissue is hard to establish from these results

Scanning across all tissue models is improves discovery

Aggregate multiple tissue results into one gene level one

Multi-Tissue PrediXcan

$$Y = a + b_1 X_g^{\text{tissue}_1} + b_2 X_g^{\text{tissue}_2} + \cdots + b_k X_g^{\text{tissue}_k} + \epsilon$$

Multi-Tissue PrediXcan

$$Y = a + b_1 X_g^{\text{tissue}_1} + b_2 X_g^{\text{tissue}_2} + \cdots + b_k X_g^{\text{tissue}_k} + \epsilon$$

What if we only have univariate regression coefficients?

Multi-Tissue PrediXcan

$$Y = a + b_1 X_g^{\text{tissue}_1} + b_2 X_g^{\text{tissue}_2} + \cdots + b_k X_g^{\text{tissue}_k} + \epsilon$$

What if we only have univariate regression coefficients?

$$\begin{aligned} Y &= a + \beta_1 X_g^{\text{tissue}_1} + \epsilon' \\ Y &= a + \beta_2 X_g^{\text{tissue}_2} + \epsilon'' \\ &\dots \\ Y &= a + \beta_k X_g^{\text{tissue}_k} + \epsilon''' \end{aligned}$$

Multi-Tissue PrediXcan

$$Y = a + b_1 X_g^{\mathrm{tissue}_1} + b_2 X_g^{\mathrm{tissue}_2} + \cdots + b_k X_g^{\mathrm{tissue}_k} + \epsilon$$

What if we only have univariate regression coefficients?

$$Y = a + \beta_1 X_g^{\mathrm{tissue}_1} + \epsilon'$$
$$Y = a + \beta_2 X_g^{\mathrm{tissue}_2} + \epsilon''$$
$$\dots$$
$$Y = a + \beta_k X_g^{\mathrm{tissue}_k} + \epsilon'''$$

$$\begin{aligned} \hat{\boldsymbol{b}} &= (\boldsymbol{X'X})^{-1}\boldsymbol{D}\widehat{\boldsymbol{\beta}} \\ \mathrm{var}(\widehat{\boldsymbol{b}}) &= \sigma_j(\boldsymbol{X'X})^{-1} \end{aligned}$$

$$\chi_k^2 = \hat{\boldsymbol{b}}'(\boldsymbol{X'X})^{-1}\hat{\boldsymbol{b}}$$

$$D_t = \sum_i X_{it}^2$$

Need to Estimate and Invert Covariance of Predicted Expression

$$\chi_t^2 = \hat{\boldsymbol{b}}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\hat{\boldsymbol{b}}$$

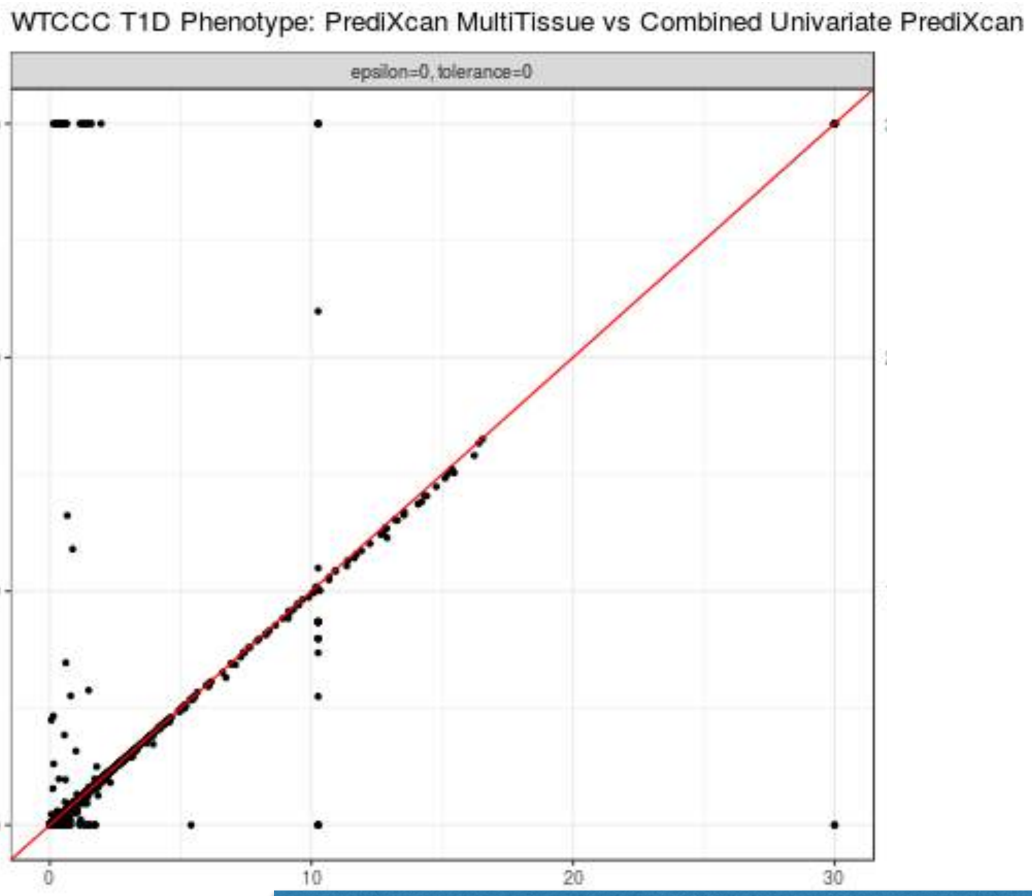
Predicted expression of a gene across tissues

We can predict expression in reference population (1000G, GTEx, etc)

Inverse may not be computable because of correlation between tissues

Combined Univariate vs. Multivariate PrediXcan

Combined Univeraite PrediXcan -log10 p



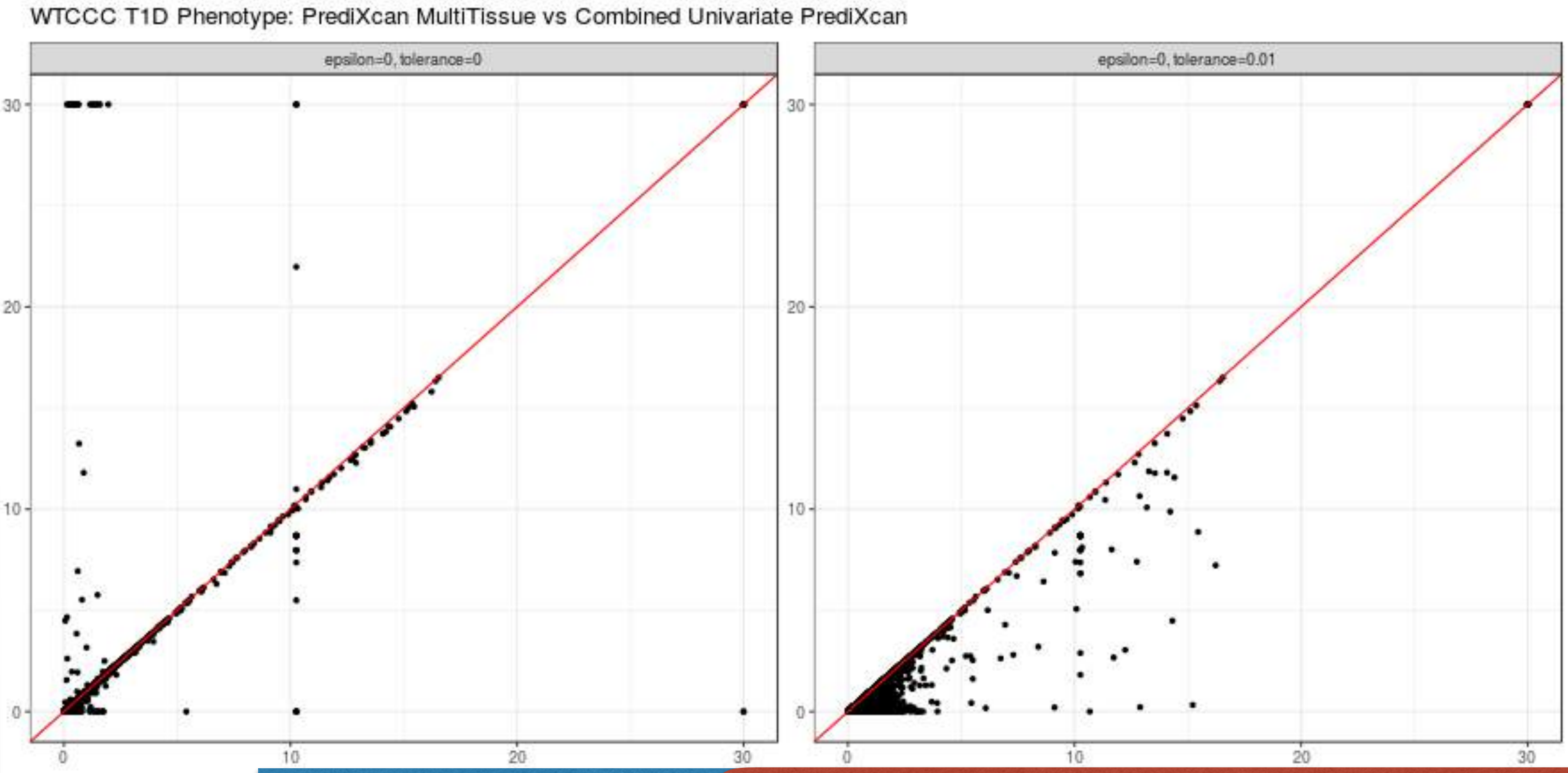
Multivariate PrediXcan -log10 p

$$\begin{aligned} \hat{\boldsymbol{b}} &= (\boldsymbol{X'X})^{-1}\boldsymbol{D}\widehat{\beta} \\ \mathrm{var}(\widehat{\boldsymbol{b}}) &= \sigma_j(\boldsymbol{X'X})^{-1} \end{aligned}$$

$$\chi^2_k = \hat{\boldsymbol{b}}'(\boldsymbol{X'X})^{-1}\hat{\boldsymbol{b}}$$

Combined Univariate vs. Multivariate PrediXcan

Combined Univariate -log10 p



WTCCC T1D Phenotype: PrediXcan MultiTissue vs Combined Univariate PrediXcan

epsilon=0, tolerance=0

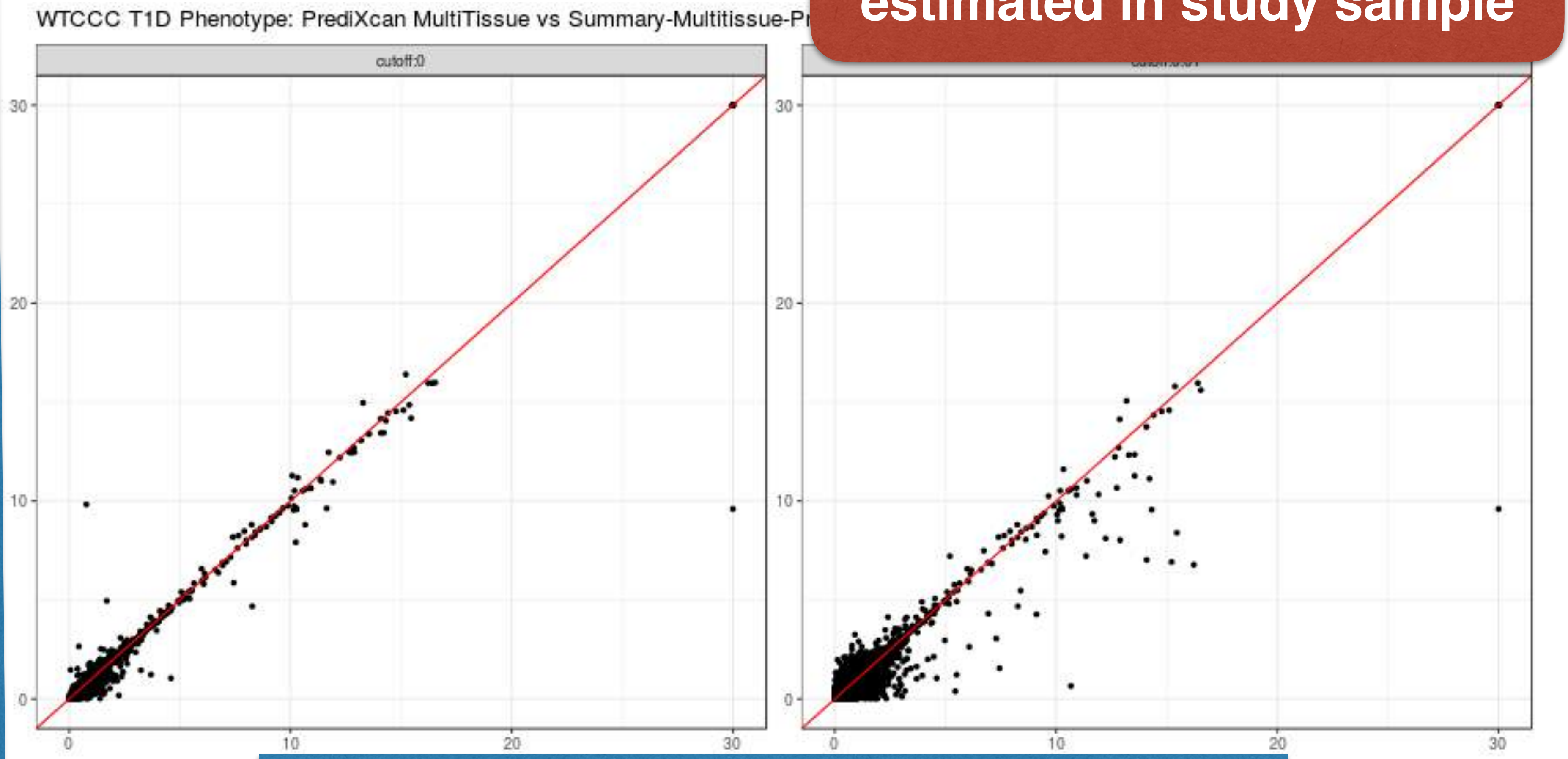
epsilon=0, tolerance=0.01

Multivariate

Avoid singularity of covariance eliminating axis of variations with small eigenvalues

Combined Summary PrediXcan vs Multivariate

Predicted expression is estimated in study sample



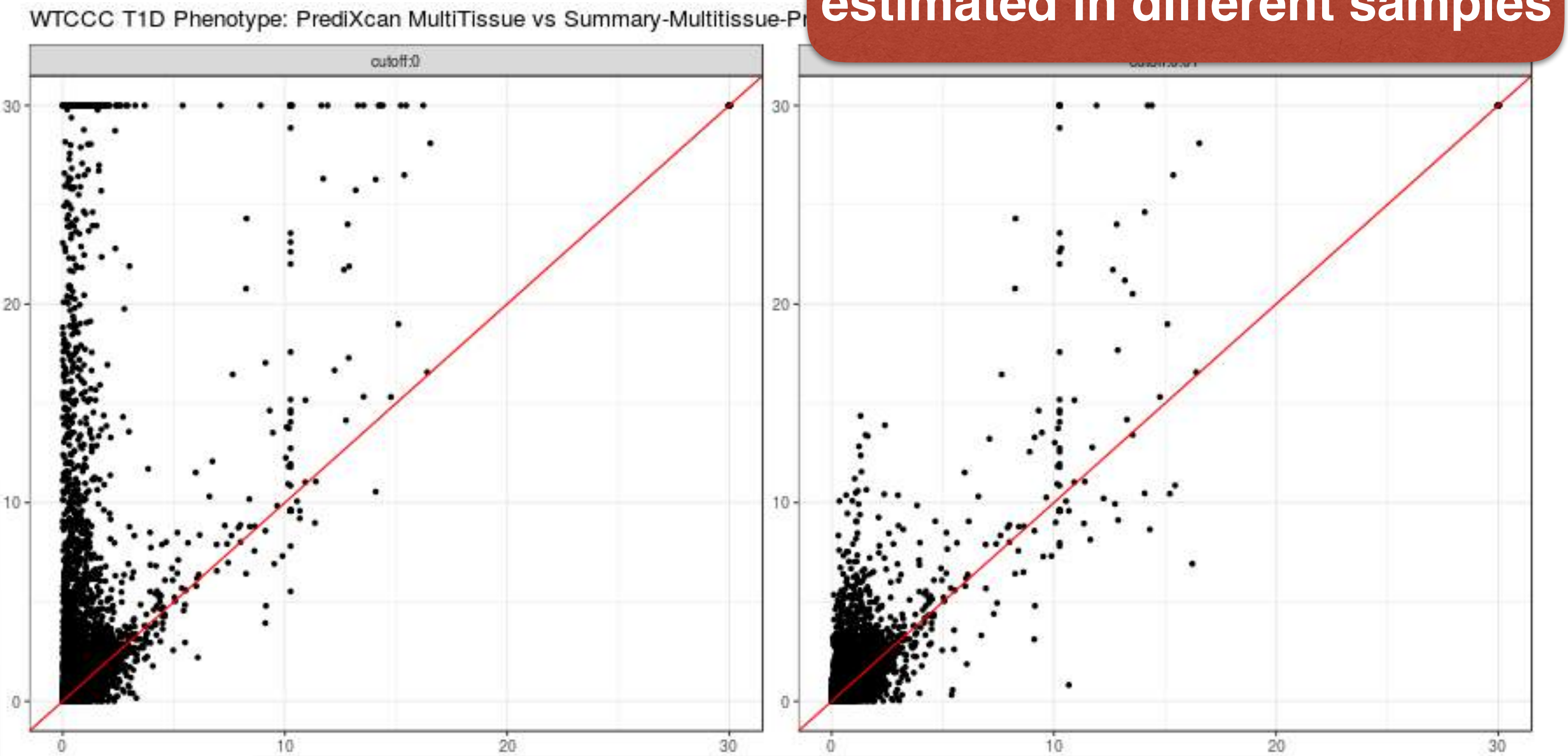
Combined Summary PrediXcan -log10 p

Multivariate PrediXcan -log10 p

$$\begin{aligned}\hat{\boldsymbol{b}} &= (\boldsymbol{X'X})^{-1}\boldsymbol{D}\widehat{\boldsymbol{\beta}} \\ \mathrm{var}(\widehat{\boldsymbol{b}}) &= \sigma_j(\boldsymbol{X'X})^{-1}\end{aligned}$$

$$\chi^2_k = \hat{\boldsymbol{b}}'(\boldsymbol{X'X})^{-1}\hat{\boldsymbol{b}}$$

Combined Summary PrediXcan vs Multivariate



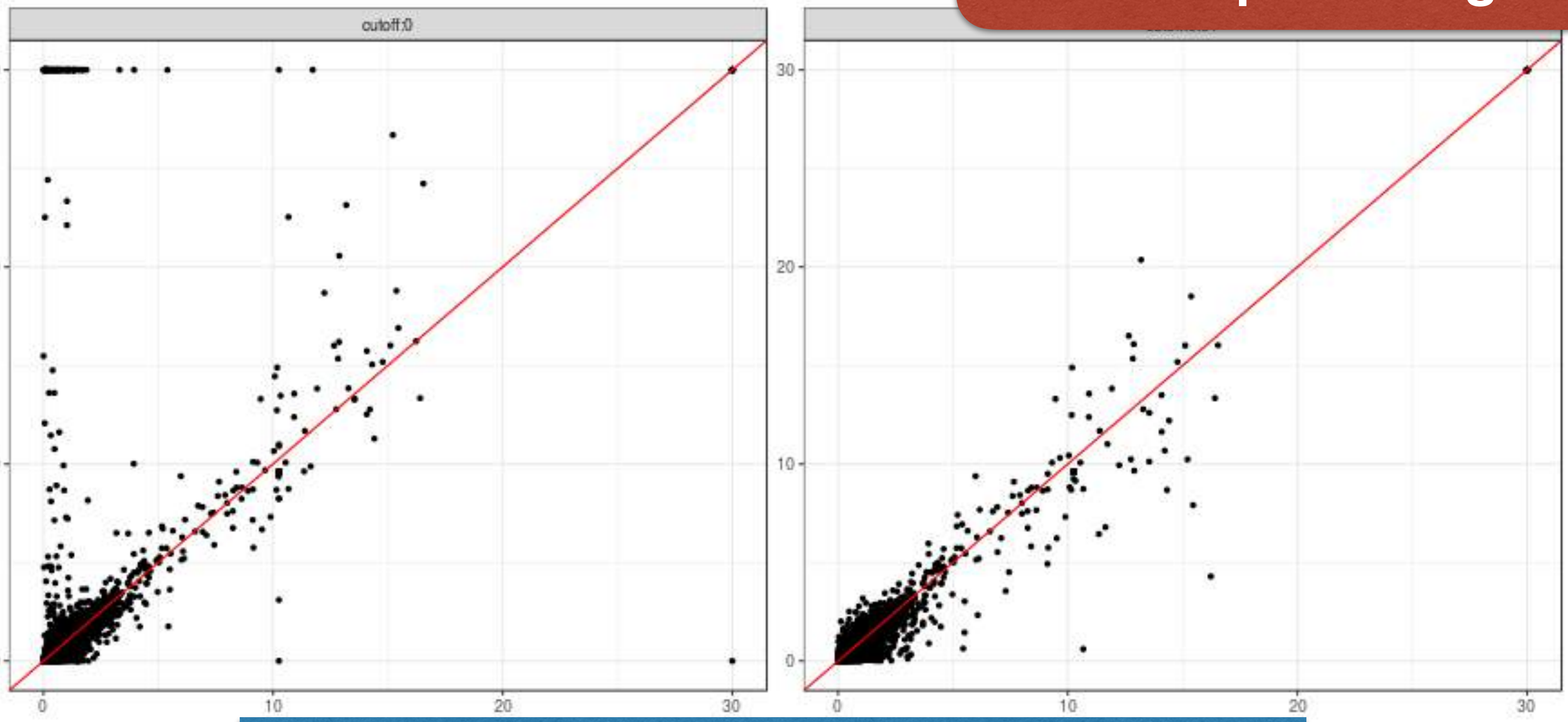
Predicted expression is estimated in different samples

Combined Univeraite PrediXcan -log10 p

Multivariate PrediXcan -log10 p

Combined Summary PrediXcan vs Multivariate

Covariance estimated directly from SNPs rather than predicting

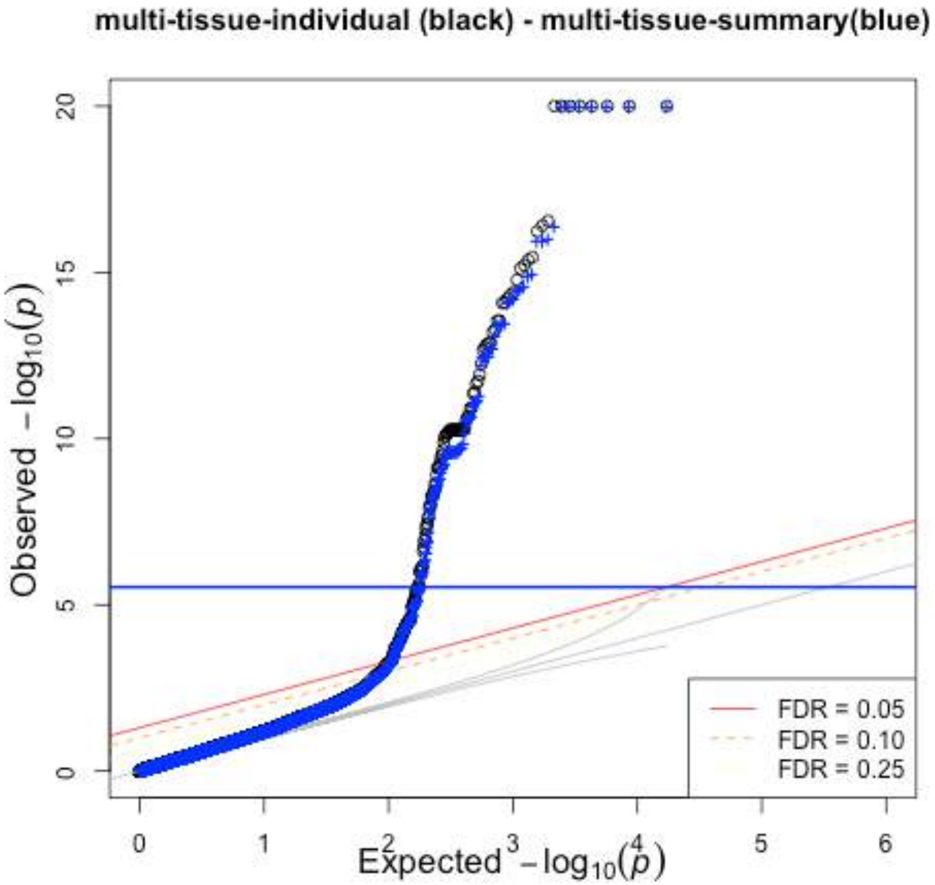


WTCCC T1D Phenotype: PrediXcan MultiTissue vs Summary-Multitissue-PrediXcan
(GTEx SNP covariance, Expression covariance from GTEX SNP intersection to GWAS)

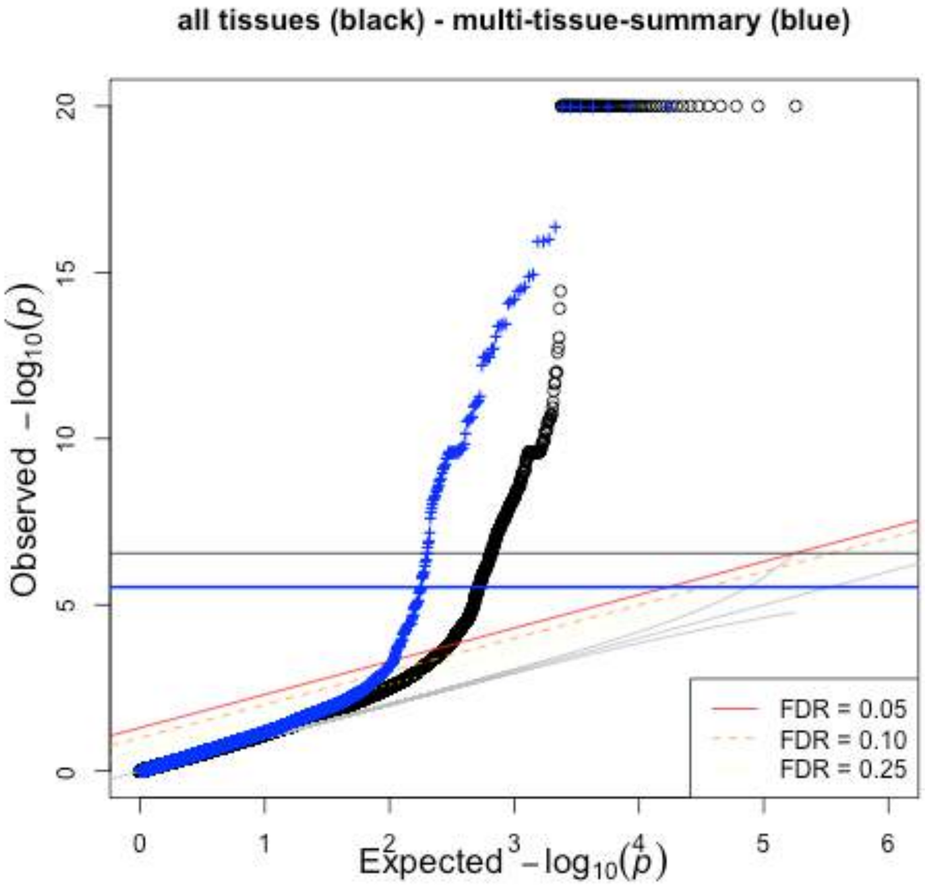
Combined Univeraite PrediXcan -log10 p

Multivariate PrediXcan -log10 p

Summary Multi Tissue Performs Similar to Multi Tissue



Summary Multi Tissue OutPerforms Using All Tissues



Summary

- Human knockouts are invaluable experiments of nature that provides information on function of genes
- Human “knockdown” gene2pheno.org, related and complementary
- Need to develop new methods to address challenges

- Summary Multi Tissue PrediXcan

Thank You

Haky Im Lab

- Alvaro Barbeira
- Jiamao Zheng
- Scott Dickinson
- Rodrigo Bonazzola
- Milton Pividori

Collaborators

- Jason Torres
- Kaanan Shah
- Heather Wheeler
- Eric Tortesson
- Tzintzuni Garcia
- Nancy Cox
- Dan Nicolae
- Graeme Bell

- Funding
  - R01MH107666 (HKI)
  - R01MH101820 (GTEx)
  - P30DK020595 (DRTC)

No conflicts of interests