Big Data and New Paradigms for Genome Discovery and Translation

Nancy J. Cox, Ph.D. http://genemed.bsd.uchicago.edu

Section of Genetic Medicine

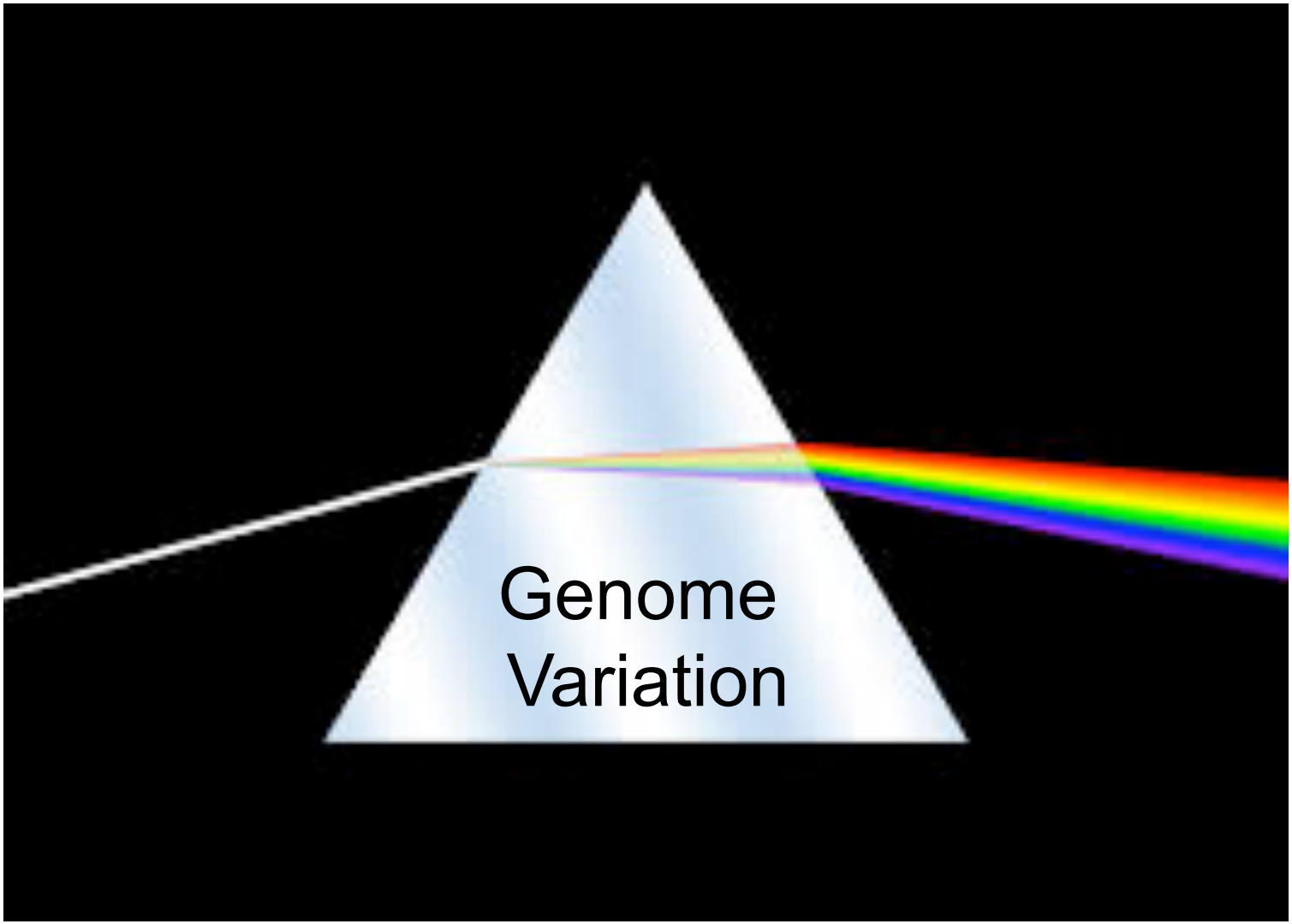
Depts. of Medicine and Human Genetics

The University of Chicago

Disclosures

(ok, maybe more like a confession)

I AM A GENETICIST


Genome
Variation

Context

- How does genome variation affect our risk of common diseases and our response to therapies for these diseases
  - What variants?
  - What mechanisms?
- Translating those discoveries to patient care
  - Paradigm for translation will involve “pre-emptive” genotyping and sequencing

Discovery and Translation

- In discovery research, we seek individual variants and aim to learn the driving biology behind the associations we detect
- For translation, we are often interested in prediction
  - Who will benefit from a particular drug therapy?
  - Who is at risk for an adverse event?
  - Who is at risk for a disease we can postpone, prevent, or alter risk for?

New in Genome Discovery

- **Key variants**
  - **Identifying classes of functional variation with strong enrichment among top GWAS signals**
  - **Identifying gene sets for which functional variants enriched**
- **Integration**
  - **Genome, transcriptome, SV, …**
- **Key genes**
  - **Mendelian disease genes may contribute to more than just Mendelian disease**

New in Genome Translation

- **Large-scale prediction**
  - **Polygenic prediction**
  - **Other –omics; poly-omic prediction**
- **EMR event monitoring**
  - **Patterns of care usage**
- **Crossing –omics prediction with EMR event monitoring**

Premise …

Paradigms developed for Mendelian diseases and rare adverse events are inadequate for translation of genome discoveries for common diseases and common adverse event and efficacy pharmaco-phenotypes

New in Genome Discovery

- **Key variants**
  - **Identifying classes of functional variation with strong enrichment among top GWAS signals**
  - **Identifying gene sets for which functional variants enriched**
- **Integration**
  - **Genome, transcriptome, SV, metabolome**
- **Key genes**
  - **Mendelian disease genes may contribute to more than just Mendelian disease**

Classes of Functional Variants Enriched in SNPs Associated with Common Disease and Complex Human Traits

- eQTLs – SNPs associated with mRNA transcript levels
- mQTLs – SNPs associated with methylation status at sites that are variably methylated
- pQTLs – SNPs that are associated with protein levels
- miRNA QTLs – SNPs associated with levels of miRNAs
- ENCODE annotations
- …

WE ACCELERATE DISCOVERY

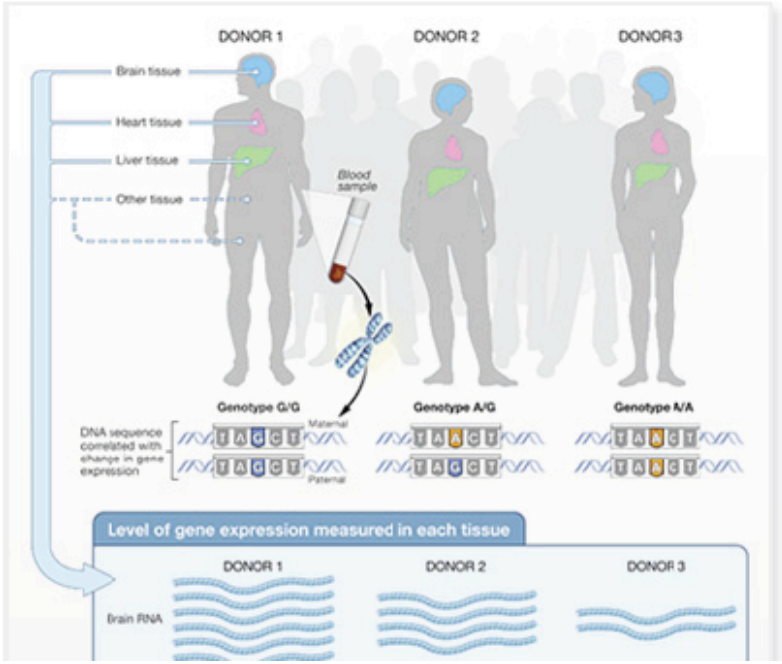
Genotype-Tissue Expression (GTEx)

Program Snapshot

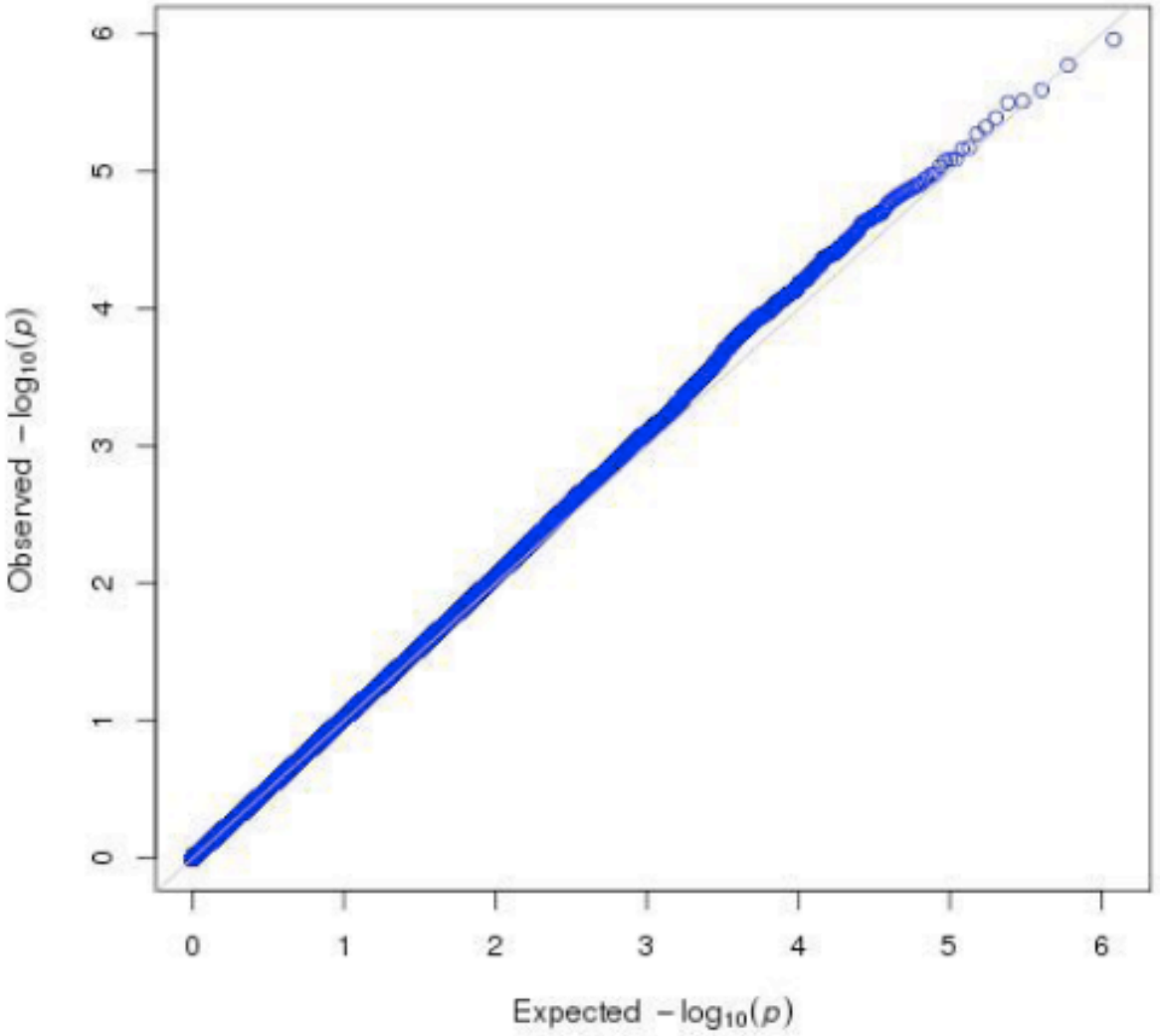
The Common Fund's **Genotype-Tissue Expression (GTEx)** program aims to study human gene expression and regulation in multiple tissues, providing valuable insights into the mechanisms of gene regulation and, in the future, its disease-related perturbations. Genetic variation between individuals will be examined for correlation with differences in gene expression level to identify regions of the genome that influence whether and how much a gene is expressed. The GTEx project includes the following initiatives:

- Novel Statistical Methods for Human Gene Expression Quantitative Trait Loci (eQTL) Analysis
- Laboratory, Data Analysis, and Coordinating Center (LDACC)
- caHUB Acquisition of Normal Tissues in Support of the GTEx Project

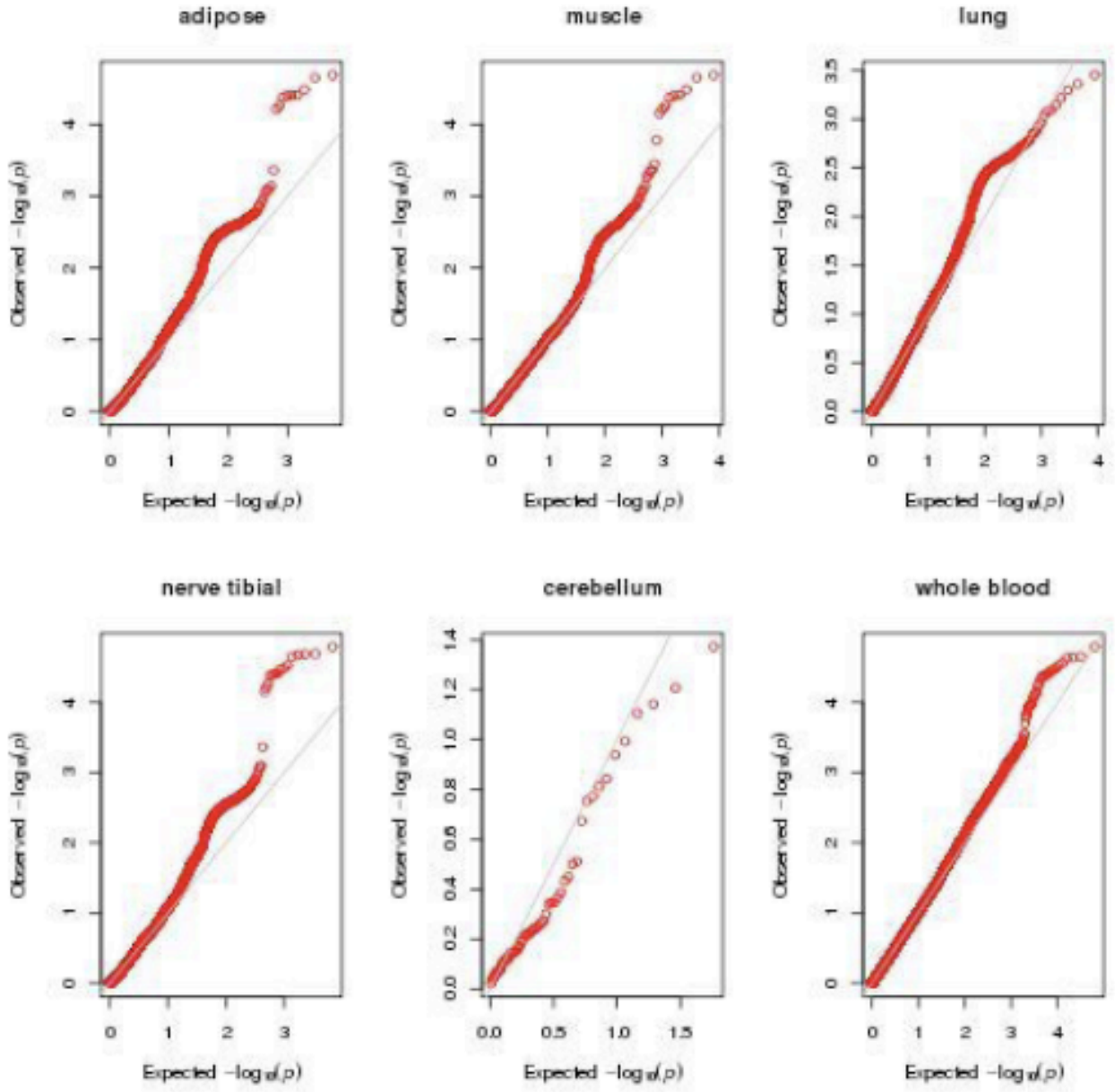
Read more...



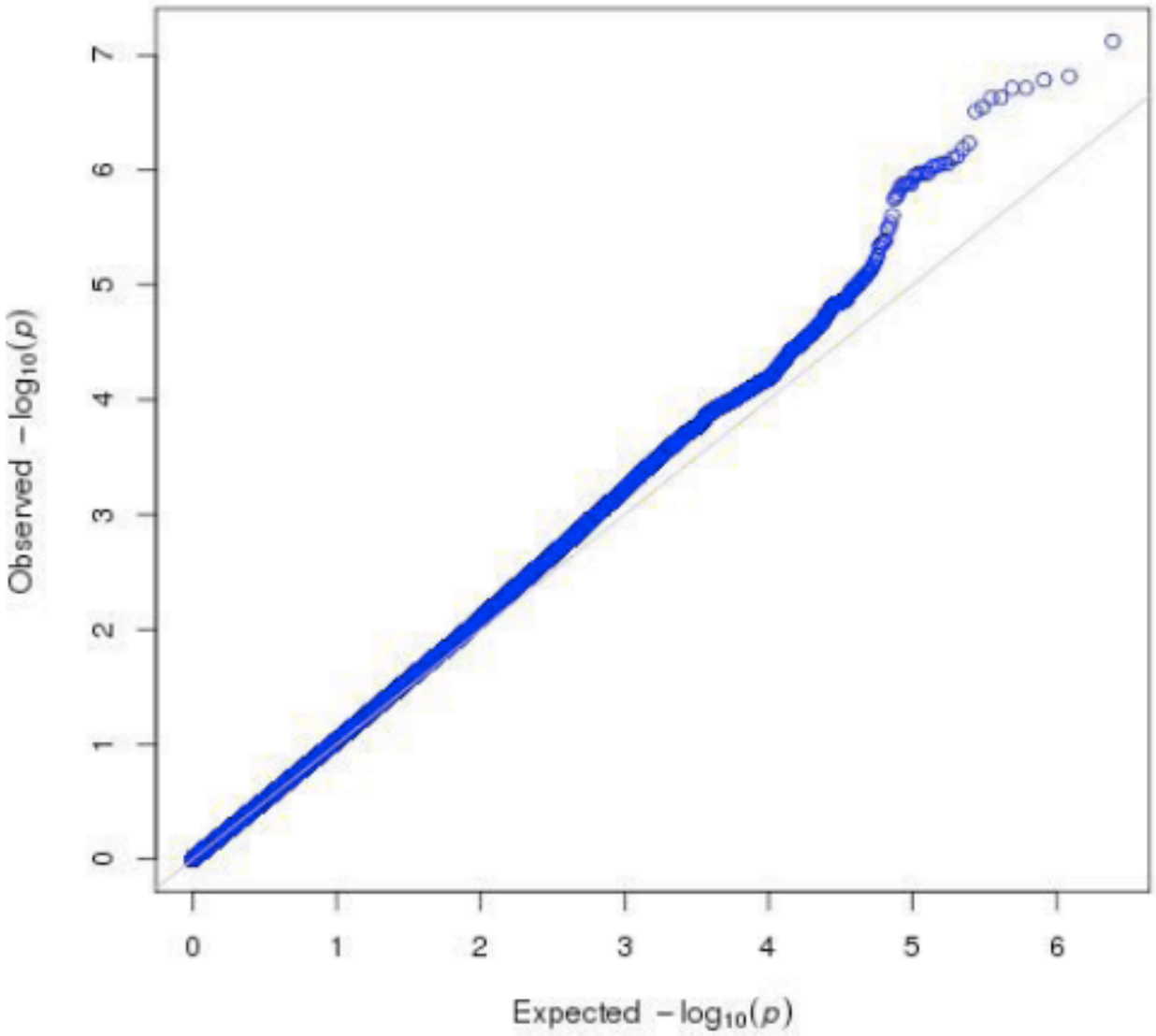
PGC: ADHD (all SNPs)



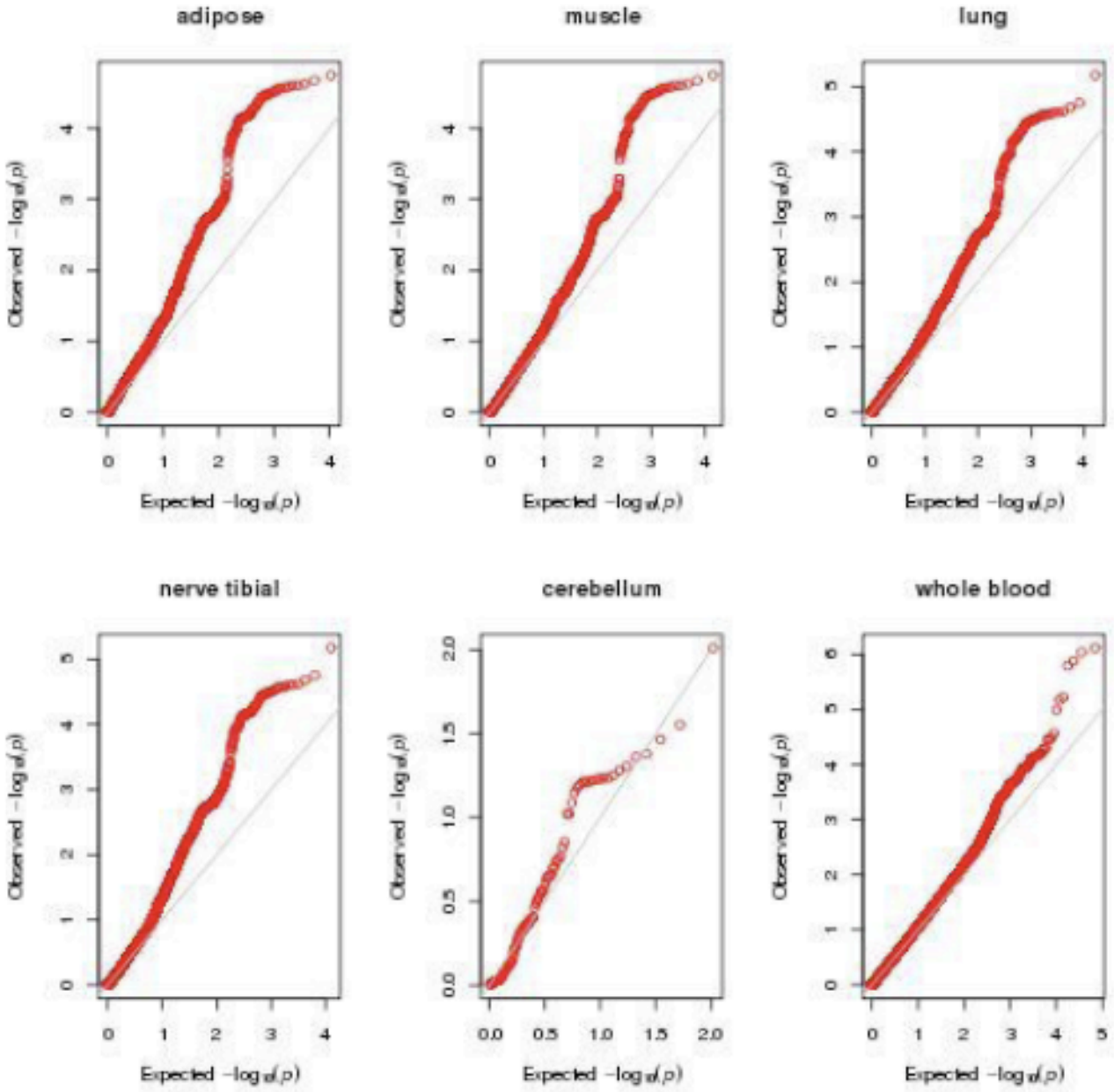
PGC: ADHD


adipose
muscle
lung
nerve tibial
cerebellum
whole blood
Observed −log10(p)
Expected −log10(p)

MAGIC: HOMA-IR (all SNPs)

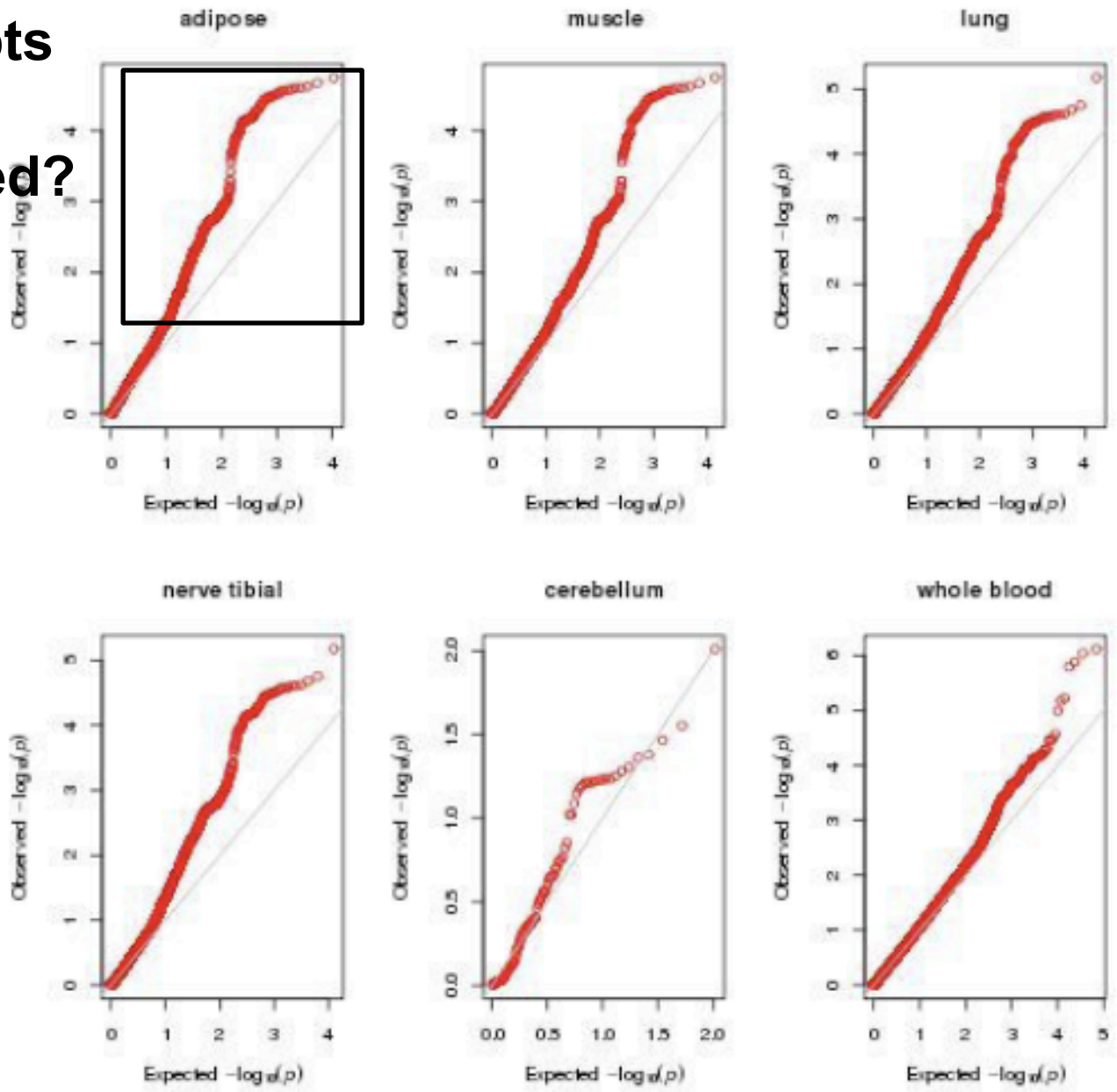


MAGIC: HOMA-IR



MAGIC: HOMA-IR

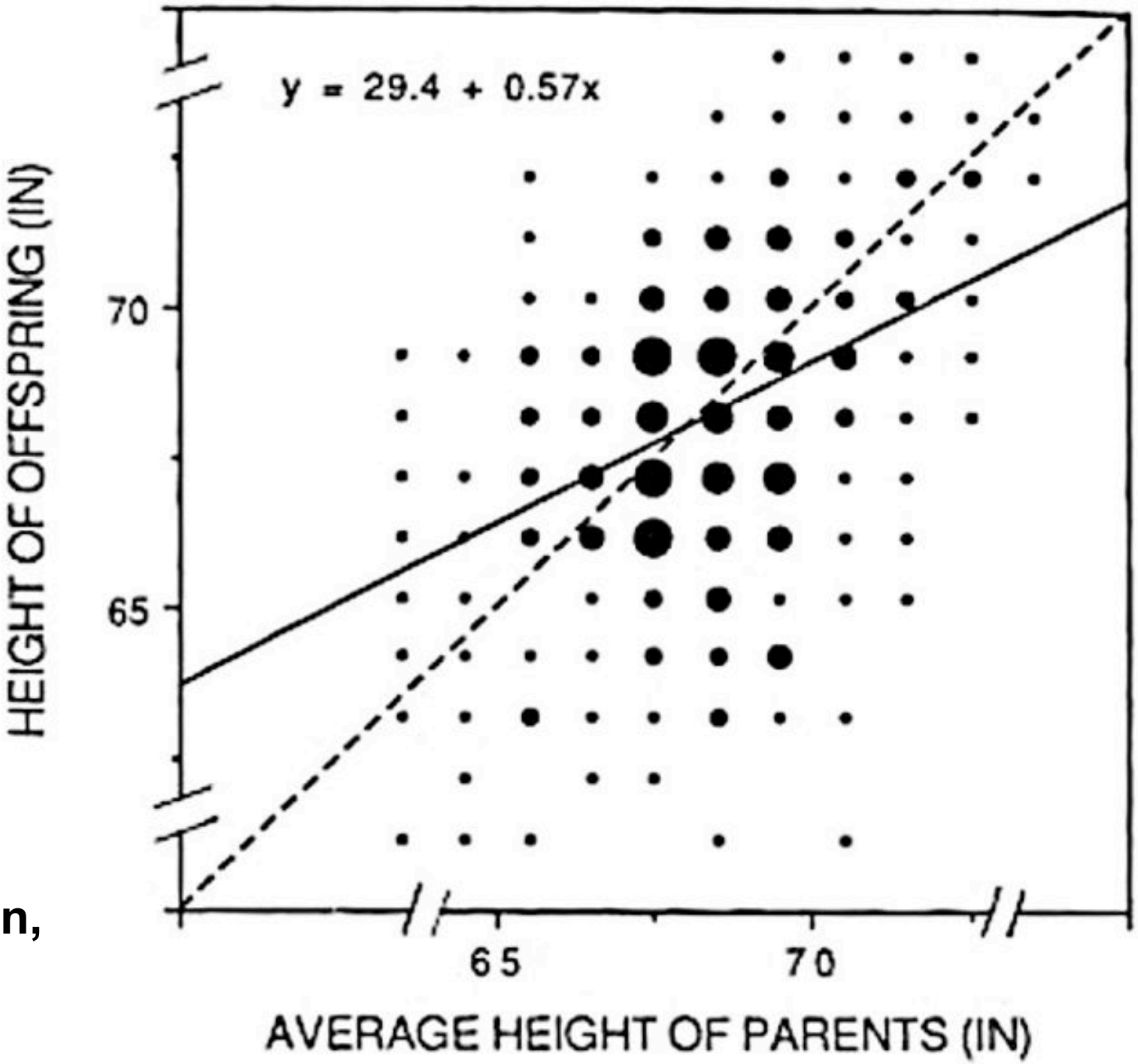
What transcripts are implicated?



Only a minority of GTEx eQTLs target the local or nearest gene

New in Genome Discovery

- Key variants
  - Identifying classes of functional variation with strong enrichment among top GWAS signals
  - Identifying gene sets for which functional variants enriched
- **Integration**
  - **Genome, transcriptome, SV, …**
- Key genes
  - Mendelian disease genes may contribute to more than just Mendelian disease


y = 29.4 + 0.57x
HEIGHT OF OFFSPRING (IN)
70
65
65
70
AVERAGE HEIGHT OF PARENTS (IN)

Galton, 1889

Concentrating Heritability

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]

ARTICLE

Estimating Missing Heritability for Disease from Genome-wide Association Studies

Sang Hong Lee,[1] Naomi R. Wray,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1,*]

Table 2 Comparison of results of different polygenic methods across diseases

			Caused by common GWAS SNPs		
				Polygenic modeling and Bayesian inference	
Disease	Prevalence (%)	Family based heritability[a]	LMM-based heritability (s.e.)	Total variance explained (50% CI)	*N* SNPs (50% CI)
Rheumatoid arthritis	1	0.53–0.68 (–0.13 MHC)[b]	0.32 (0.037)	0.18 (0.15–0.20) (+0.04 known non-MHC)[b]	2,231 (1,588–2,740)
Celiac disease	1	0.5–0.87 (–0.35 MHC)[b]	0.33 (0.042)	0.44 (0.40–0.47)	2,550 (1,907–3,061)
MI/CAD	6	0.3–0.63	0.41 (0.067)	0.48 (0.43–0.54)	1,766 (1,215–2,125)
T2D mellitus	8	0.26–0.69	0.51 (0.065)	0.49 (0.46–0.53)	2,919 (2,335–3,442)

[a]Family based heritability estimates were taken from previous data for rheumatoid arthritis[27,28], celiac disease[18,30], MI/CAD[31,32] and T2D[33,34]. [b]We excluded some loci in certain analyses: although the family based heritability estimates are based on the whole genome, the extended MHC region was removed from the common GWAS SNP analyses for rheumatoid arthritis and celiac disease, and validated non-MHC loci were further removed from the polygenic modeling analysis of the rheumatoid arthritis GWAS data. 50% CI, 50% credible interval; s.e., standard error.

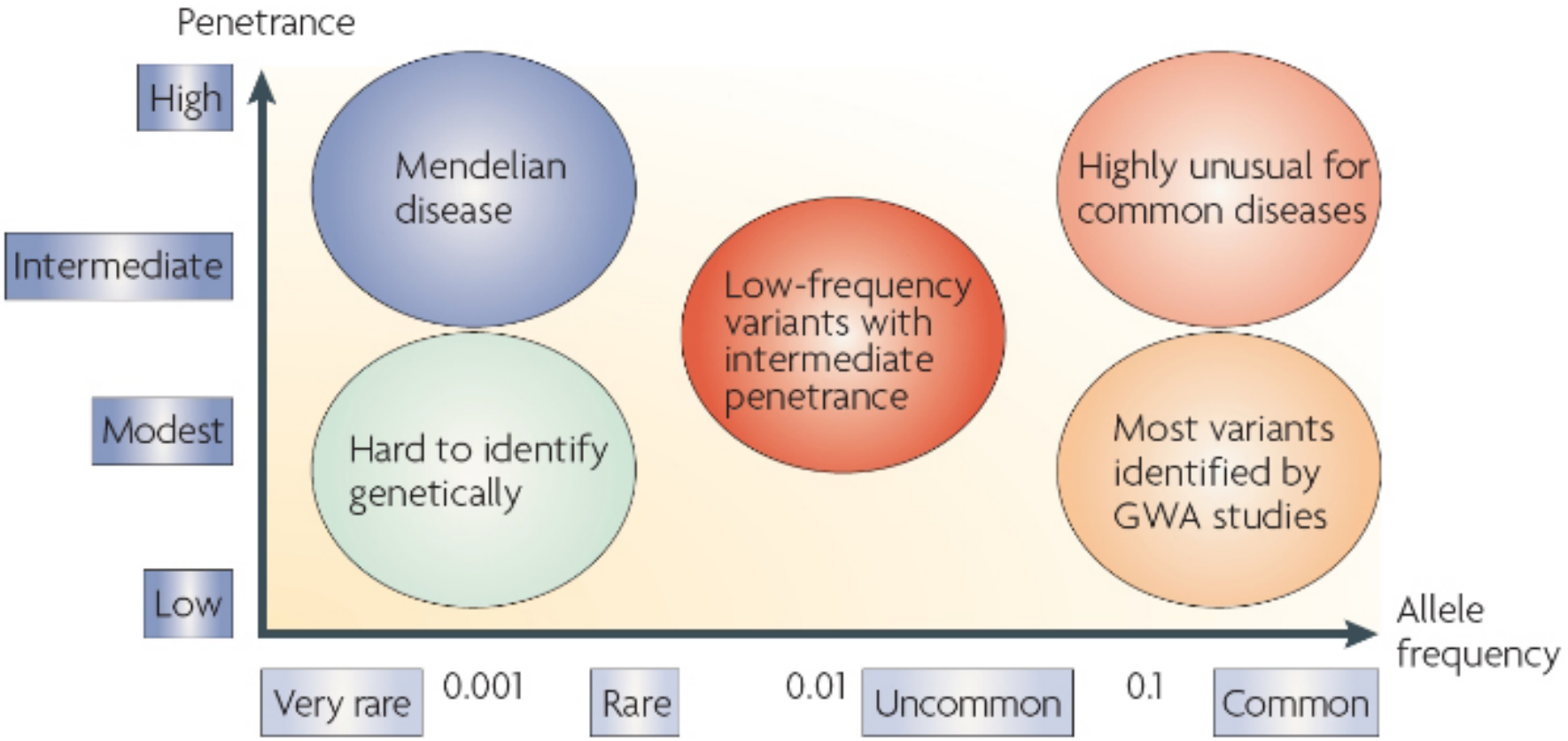
Stahl et al, Nat Gen

	Type 1 Diabetes			Crohns Disease	
	V(G)/V(P)	SE		V(G)/V(P)	SE
adipose	0.21	0.019		0.03	0.008
heart	0.199	0.02		0.017	0.006
lung	0.192	0.018		0.02	0.007
muscle	0.188	0.018		0.028	0.008
nerve	0.191	0.018		0.025	0.008
whole blood	0.187	0.023		0.17	0.024
Overall	0.48	0.06		0.50	0.07

Concentration of Heritability

- Smaller numbers of eQTLs (3-30K) account for 30-60% of heritability estimated for all variants after QC (150-600K)
- Observed across autoimmune and inflammatory diseases, neuropsychiatric, metabolic, etc.
- Partitioning by cross vs. single tissues, cis- and trans-, common and rare

Relationship Between Risk and MAF



Lobo, I. (2008) Multifactorial inheritance and genetic disease. Nature Education 1(1):5

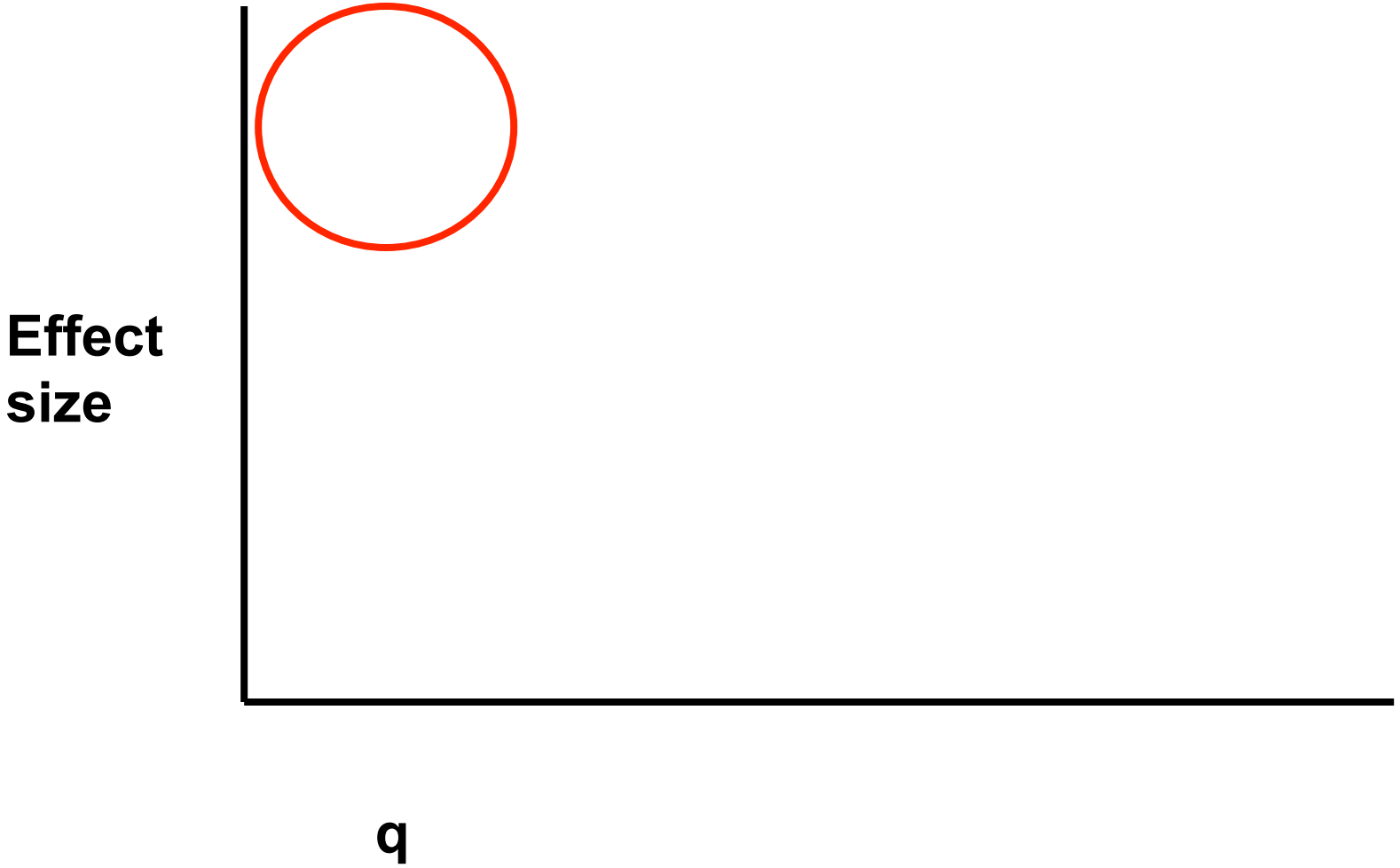
Expected Relationship Between MAF and Effect Size?

- Human populations have been expanding super-exponentially
- Are rare variants largely functional and strongly selected against? Or largely neutral?
- What are the implications for this relationship when fitness affects variation at a gene through phenotype A, but some variants at the gene affect risk for disease B (unrelated to A)?

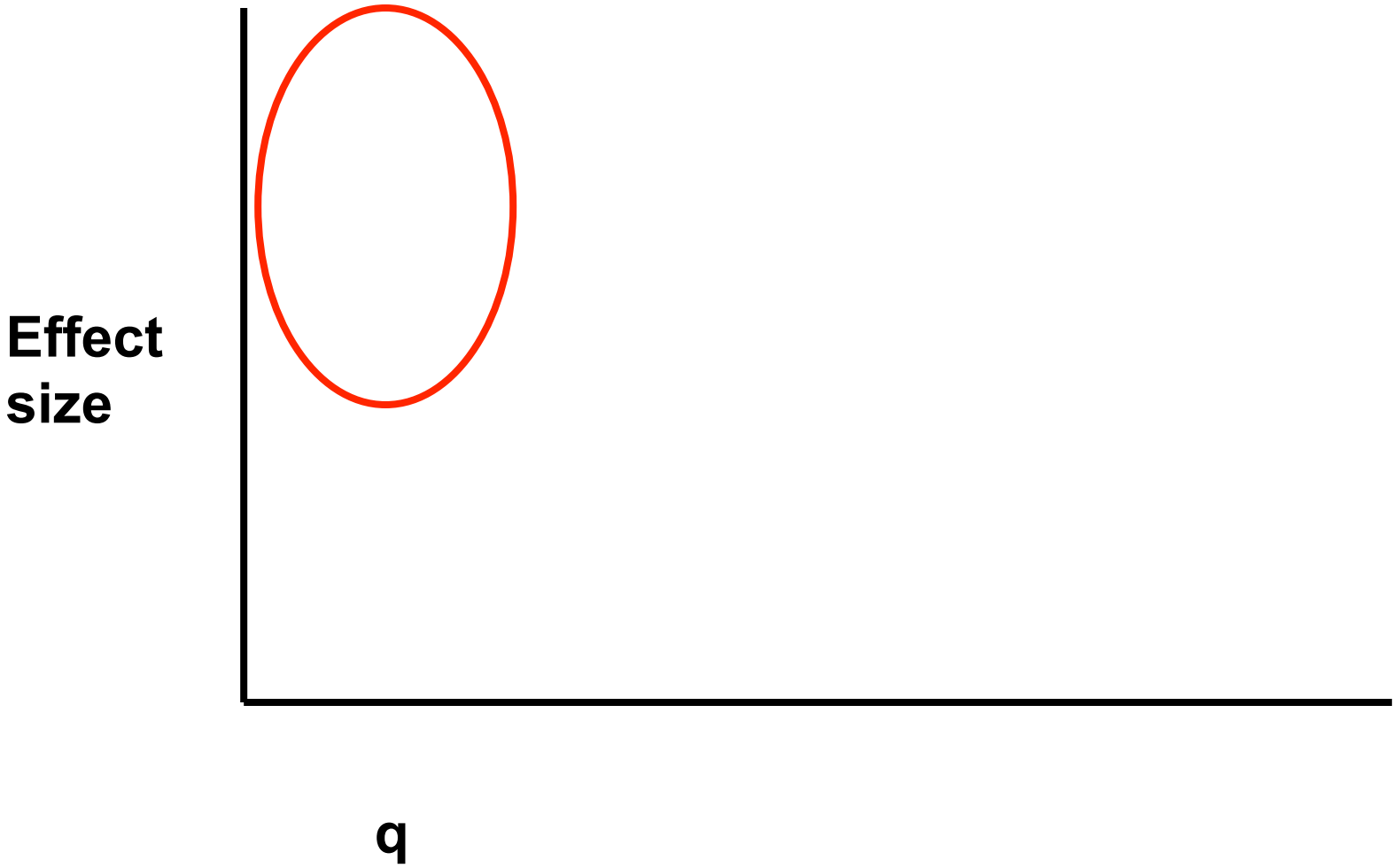
Improving Inference in Studies of Rare Variants

- Maximizing the information on rare variant associations will require considering new dimensions in analysis
- Current generation of studies have considered the contributions of rare and common variants in complete isolation

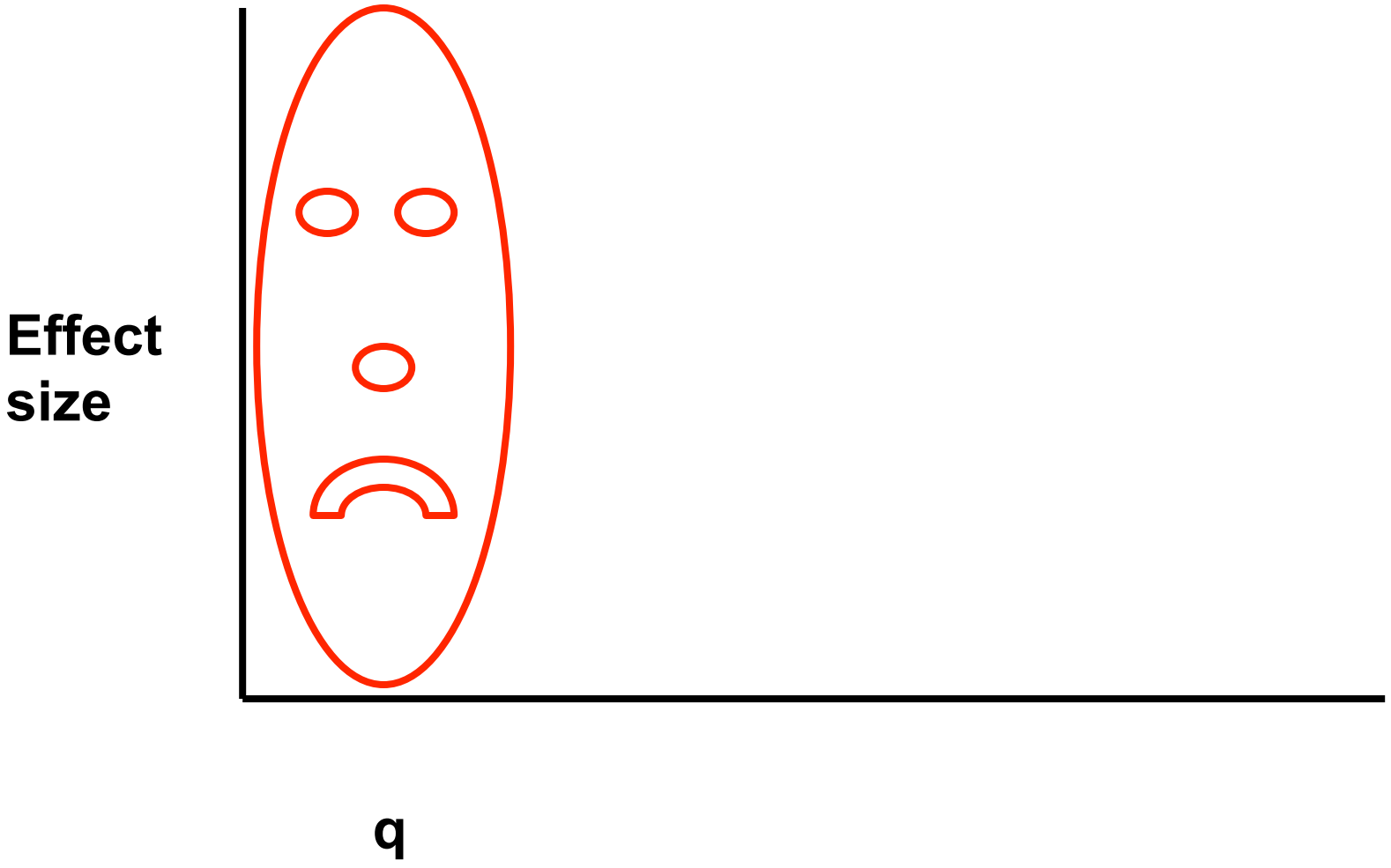
Dimensions in the Analysis of Rare Variants



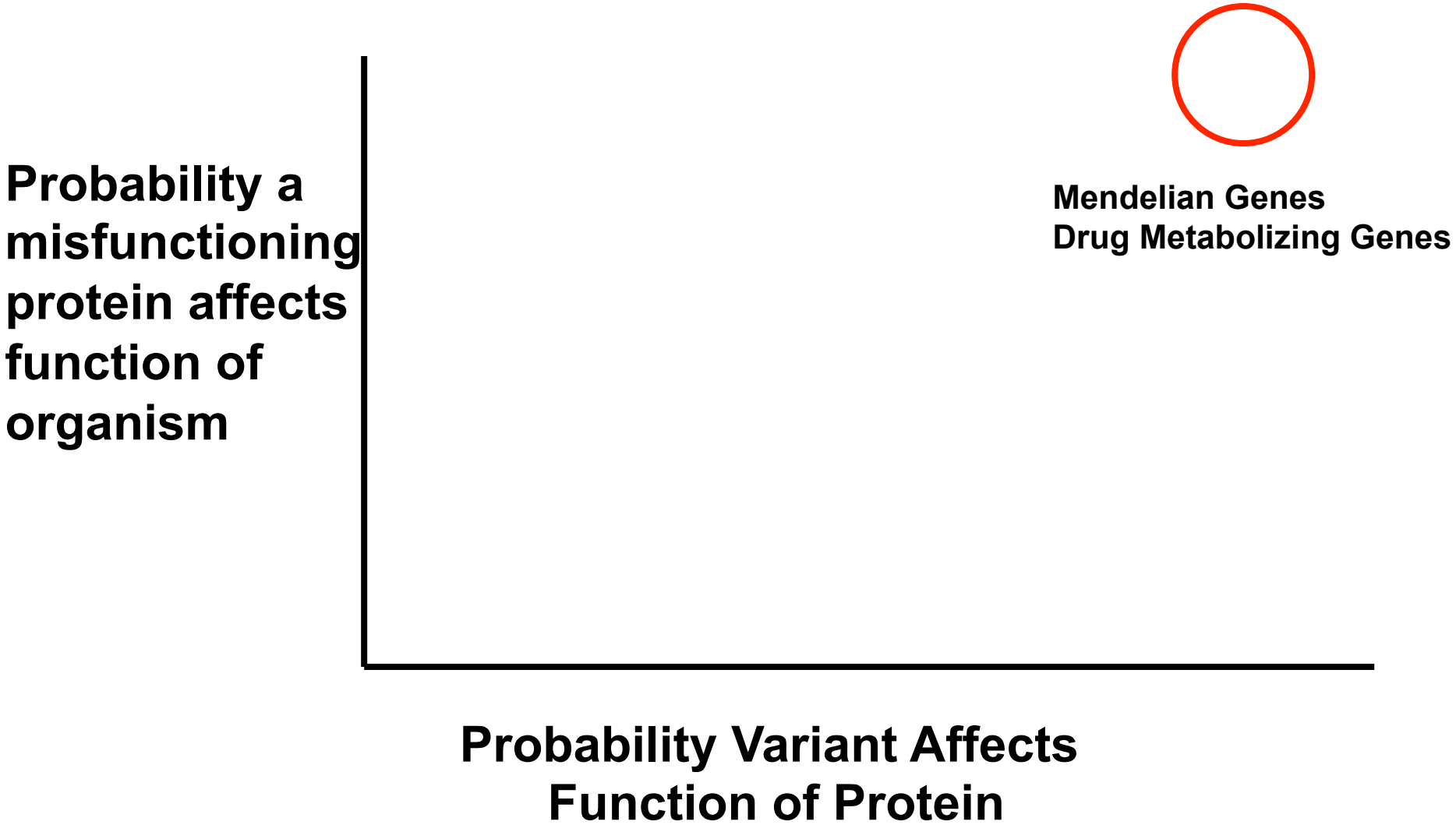
Dimensions in the Analysis of Rare Variants



Dimensions in the Analysis of Rare Variants



Dimensions in the Analysis of Rare Variants



New in Genome Discovery

- Key variants
  - Identifying classes of functional variation with strong enrichment among top GWAS signals
  - Identifying gene sets for which functional variants enriched
- Integration
  - Genome, transcriptome, SV, …
- **Key genes**
  - **Mendelian disease genes may contribute to more than just Mendelian disease**

Cell
Volume 155
Number 1
September 26, 2013
www.cell.com

A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk

Blair DR, Lyttle CS, Mortensen JM, Bearden CF, Jensen AB, Khiabanian H, Melamed R, Rabadan R, Bernstam EV, Brunak S, Jensen LJ, Nicolae D, Shah NH, Grossman RL, Cox NJ, White KP, Rzhetsky A

Mendelian Disease Genes…

- Have larger variation in expression than other genes
- Are more broadly expressed across tissues than other genes
- Are more likely to have at least one SNP highly significantly associated with its expression (an eQTL)
- eQTLs for Mendelian disease genes are more likely to be associated with common disease and complex traits

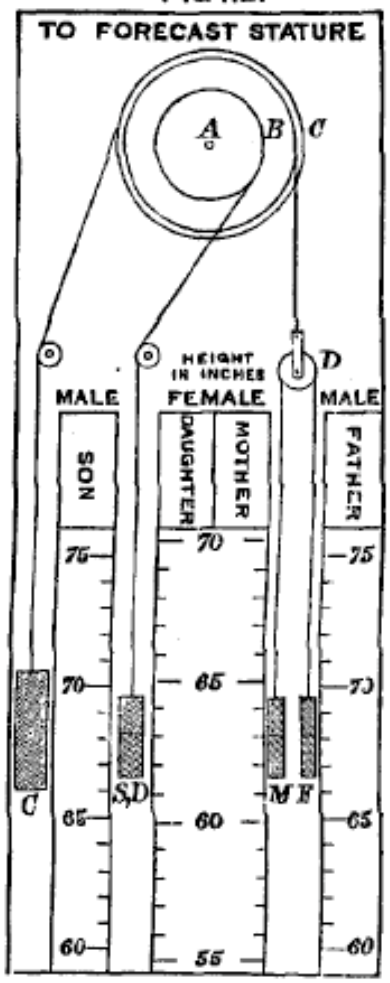
New in Genome Translation

- **Large-scale prediction**
  - **Polygenic prediction**
  - **Other -omics**
- **EMR event monitoring**
  - **Patterns of care usage**
- **Crossing –omics prediction with EMR event monitoring**

Prediction in the Era of Big Data

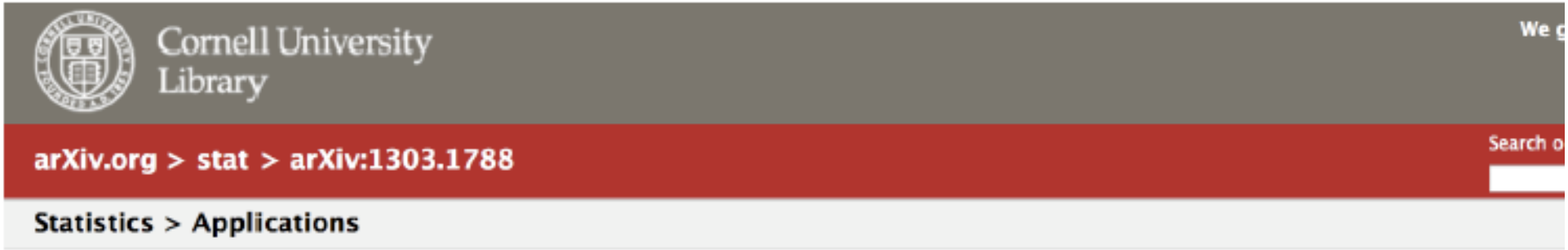
contrived more than one form of apparatus by which the probable stature of the children of known parents can be mechanically reckoned. Fig. 12 is a representation of one of them, that is worked with pulleys and weights. A, B, and C are three thin wheels with grooves round their edges. They are screwed together so as to form a single piece that turns easily on its axis. The weights M and F are attached to either end of a thread that passes over the movable pulley D. The pulley itself hangs from a thread which is wrapped two or three times round the groove of B and is then secured to the wheel. The weight SD hangs from a thread that is wrapped two or three times round the groove of A, and is then secured to the wheel. The diameter of A is to that of B as 2 to 3. Lastly, a thread is wrapped in the opposite direction round the wheel C, which may have any convenient diameter, and is

FIG. 12.



Galton, 1889

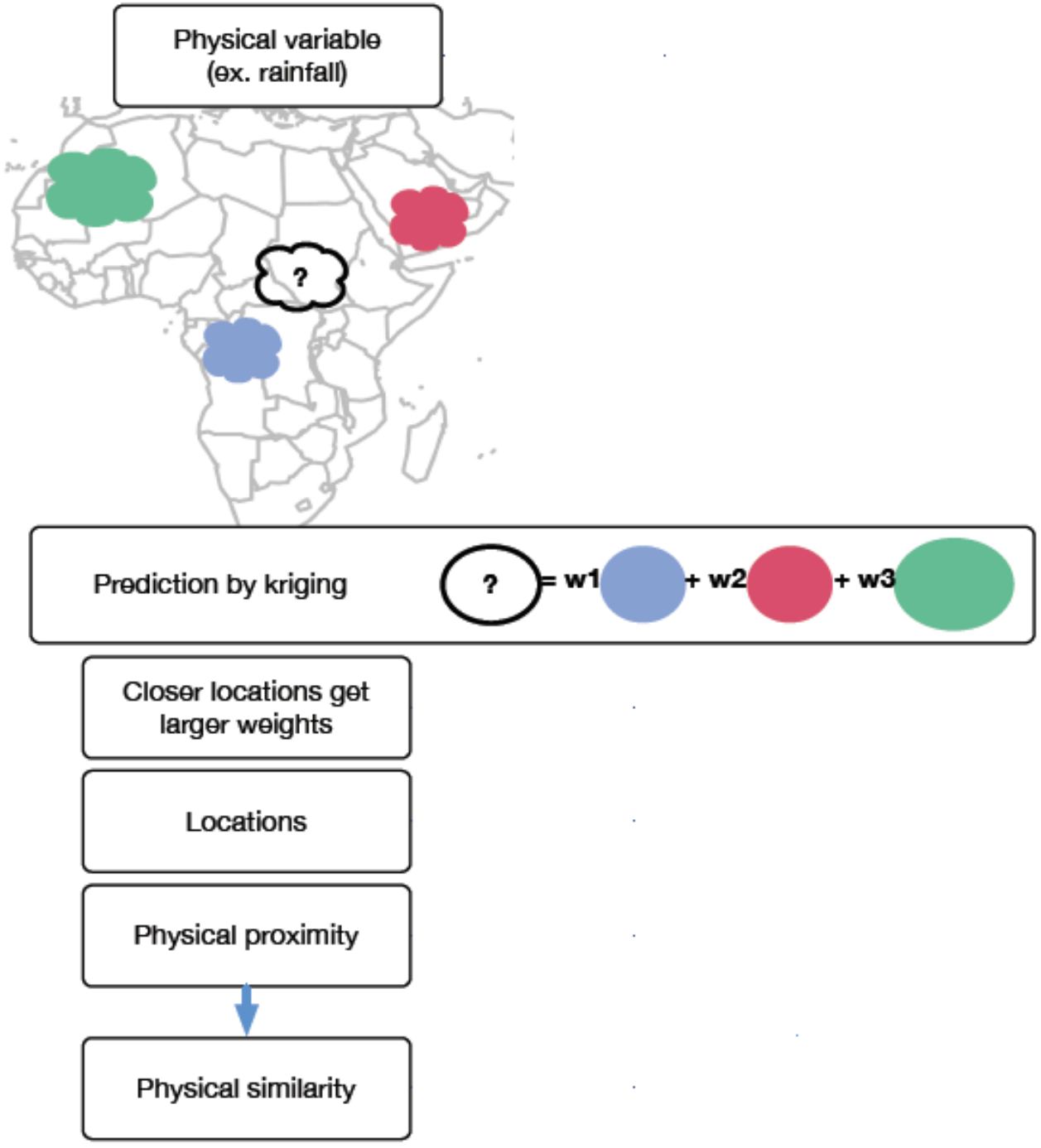
In Press Genetic Epidemiology

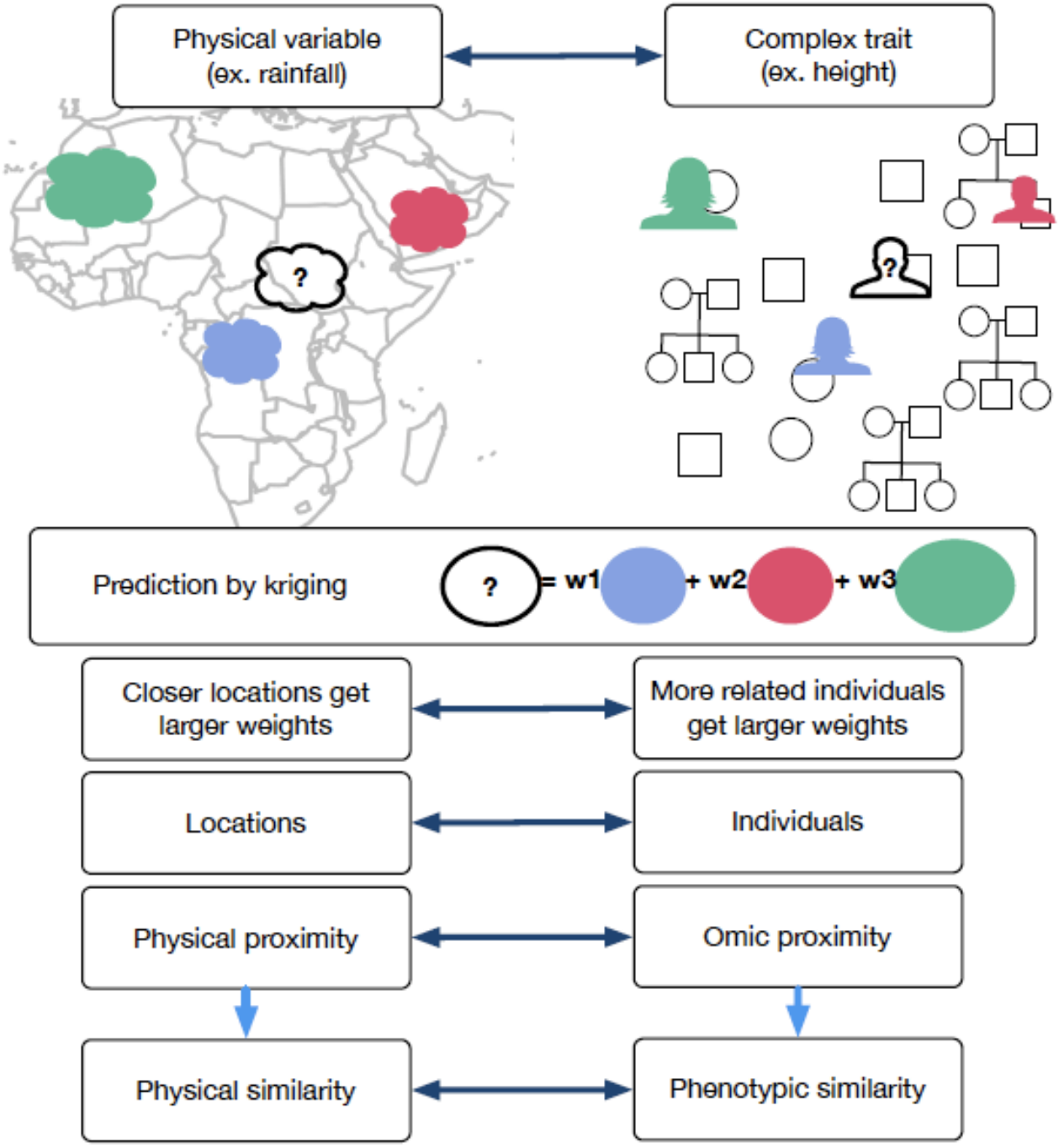


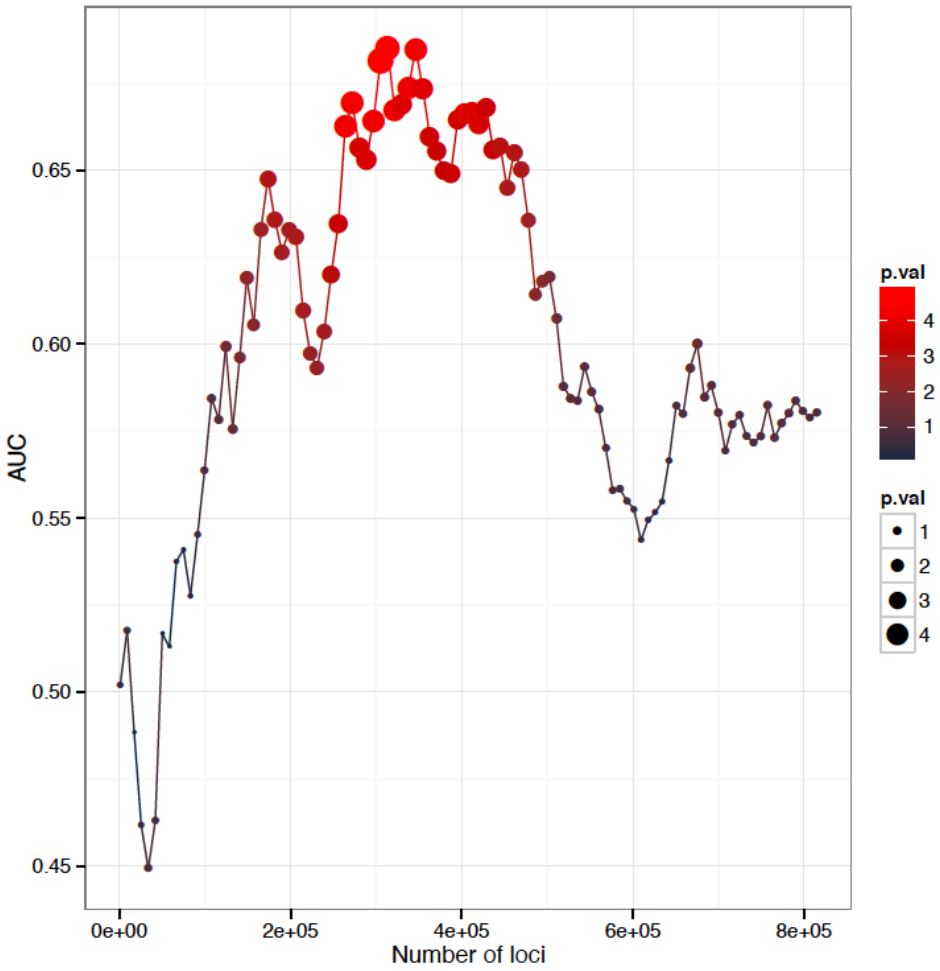
Poly-Omic Prediction of Complex Traits: OmicKriging

Heather E. Wheeler, Keston Aquino-Michaels, Eric R. Gamazon, Vassily V. Trubetskoy, M. Eileen Dolan, R. Stephanie Huang, Nancy J. Cox, Hae Kyung Im

(Submitted on 7 Mar 2013 (v1), last revised 12 Sep 2013 (this version, v2))


Physical variable
(ex. rainfall)
?
Prediction by kriging
?
= w1
+ w2
+ w3
Closer locations get
larger weights
Locations
Physical proximity
Physical similarity


Physical variable
(ex. rainfall)
Complex trait
(ex. height)
?
?
Prediction by kriging
?
= w1
+ w2
+ w3
Closer locations get
larger weights
More related individuals
get larger weights
Locations
Individuals
Physical proximity
Omic proximity
Physical similarity
Phenotypic similarity



- **Build large-scale predictors for hypertension using GWAS meta-analysis on 20,000+ subjects**
- **Test quality of prediction for bevacizumab-induced hypertension in clinical trials data (80303)**
- **AUC ~ 0.68 for polygenic prediction**

Large-scale -Omic Predictors

- Can be used in much the same way as biomarkers for risk prediction
- Can be built using data on 10's to 100's of thousands of individuals
- Can be tested and validated in high-throughput using information in CRDWs and existing biobanks
- Can be combined with other –omic, biomarker, and EMR usage-based predictors

We Have Been Picking the Cherries

Cox Lab

Eric Gamazon

Lea Davis
(Bridget)

Jason Torres

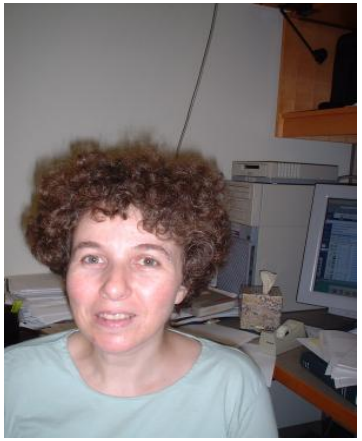
Anna
Tikhomirov

Anuar Konkashbaev

Keston Aquino-
Michaels

Carolyn Jumper

Vasily Trubetskoy



Anna Pluzhnikov

Colleagues & Collaborators

Dan Nicolae

M. Eileen Dolan

Haky Im

Bob Grossman

Chun-yu Liu

Andrey Rzhetsky

The GTEx Consortium Investigators (GTEx Pilot phase)

cancer Human Biobank (caHUB)

Biospecimen Source Sites (BSS)

John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, *National Disease Research Interchange, Philadelphia, PA*
Richard Hasz, *Gift of Life Donor Program, Philadelphia, PA*
Gary Walters, *LifeNet Health, Virginia Beach, VA*
Nancy Young, *Albert Einstein Medical Center, Philadelphia, PA*

Laura Siminoff (ELSI Study), Heather Traino, Maghboeba Mosavel, Laura Barker, *Virginia Commonwealth University, Richmond, VA*

Barbara Foster, Mike Moser, Ellen Karasik, Bryan Gillard, Kimberley Ramsey, *Roswell Park Cancer Institute, Buffalo, NY*
Susan Sullivan, Jason Bridge, *Upstate New York Transplant Service, Buffalo, NY*

Comprehensive Biospecimen Resource (CBR)

Scott Jewell, Dan Rohr, Dan Maxim, Dana Filkins, Philip Harbach, Eddie Cortadillo, Bree Berghuis, Lisa Turner, Melissa Hanson, Anthony Watkins, Brian Smith, *Van Andel Institute, Grand Rapids, MI*

Pathology Resource Center (PRC)

Leslie Sobin, James Robb, *SAIC-Frederick, Inc., Frederick, MD*
Phillip Branton, *National Cancer Institute, Bethesda, MD*
John Madden, *Duke University, Durham, NC*
Jim Robb, Mary Kennedy, *College of American Pathologists, Northfield, IL*

Comprehensive Data Resource (CDR)

Greg Korzeniewski, Charles Shive, Liqun Qi, David Tabor, Sreenath Nampally, *SAIC-Frederick, Inc., Frederick, MD*

caHUB Operations Management

Steve Buia, Angela Britton, Anna Smith, Karna Robinson, Robin Burges, Karna Robinson, Kim Valentino, Deborah Bradbury, *SAIC-Frederick, Inc., Frederick, MD*
Kenyon Erickson, *Sapient Government Services, Arlington, VA*

Brain Bank

Deborah Mash, PI; Yvonne Marcus, Margaret Basile *University of Miami School of Medicine, Miami, FL*

Laboratory, Data Analysis, and Coordinating Center (LDACC)

Kristin Ardlie, Gad Getz, co-PIs; David DeLuca, Taylor Young, Ellen Gelfand, Tim Sullivan, Yan Meng, Ayellet Segre, Jules Maller, Pouya Kheradpour, Luke Ward, Daniel MacArthur, Manolis Kellis, *The Broad Institute of Harvard and MIT, Inc., Cambridge, MA*

Statistical Methods Development (R01)

Jun Liu, co-PI, *Harvard University, Boston, MA, USA*
Jun Zhu, co-PI; Zhidong Tu, Bin Zhang, *Mt Sinai School of Medicine, New York, NY*
Nancy Cox, Dan Nicolae, co-PIs; Eric Gamazon, Haky Im, Anuar Konkashbaev, *University of Chicago, Chicago, IL*
Jonathan Pritchard, PI; Matthew Stevens, Timothèe Flutre, Xiaoquan Wen, *University of Chicago, Chicago, IL*
Emmanouil T. Dermitzakis, co-PI; Tuuli Lappalainen, Pedro Ferreira, *University of Geneva, Geneva, Switzerland*
Roderic Guigo, co-PI; Jean Monlong, Michael Sammeth, *Center for Genomic Regulaton, Barcelona, Spain*
Daphne Koller, co-PI; Alexis Battle, Sara Mostafavi, *Stanford University, Palo Alto, CA*
Mark McCarthy, co-PI; Manuel Rivas, Andrew Morris, *Oxford University, Oxford, United Kingdom*
Ivan Rusyn, Andrew Nobel, Fred Wright, Co-PIs; Andrey Shabalin, *University of North Carolina - Chapel Hill, Chapel Hill, NC*

US National Institutes of Health

NCBI dbGaP

Mike Feolo, Steve Sherry, Jim Ostell, Nataliya Sharopova, Anne Sturcke, *National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD*

Program Management

Leslie Derr, *Office of Strategic Coordination (Common Fund), Office of the Director, National Institutes of Health, Bethesda, MD*
Eric Green, Jeffery P. Struewing, Simona Volpi, Joy Boyer, Deborah Colantuoni, *National Human Genome Research Institute, Bethesda, MD*
Thomas Insel, Susan Koester, A. Roger Little, Patrick Bender, Thomas Lehner, *National Institute of Mental Health, Bethesda, MD*
Jim Vaught, Sherry Sawyer, Nicole Lockhart, Chana Rabiner, Joanne Demchok, *National Cancer Institute, Bethesda, MD*