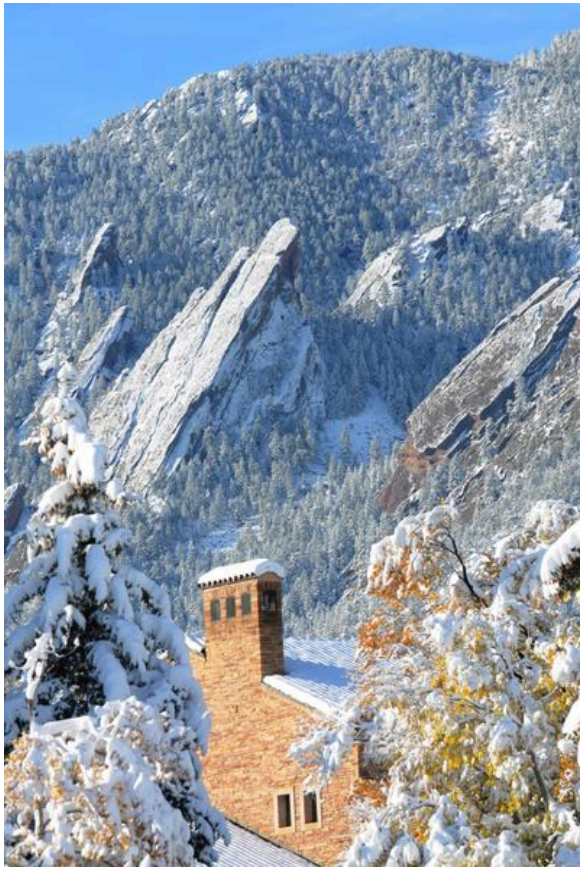
University of Colorado Boulder

Lecture 4: Heritability and Genetic Risk Scores

Matt McQueen | Associate Professor

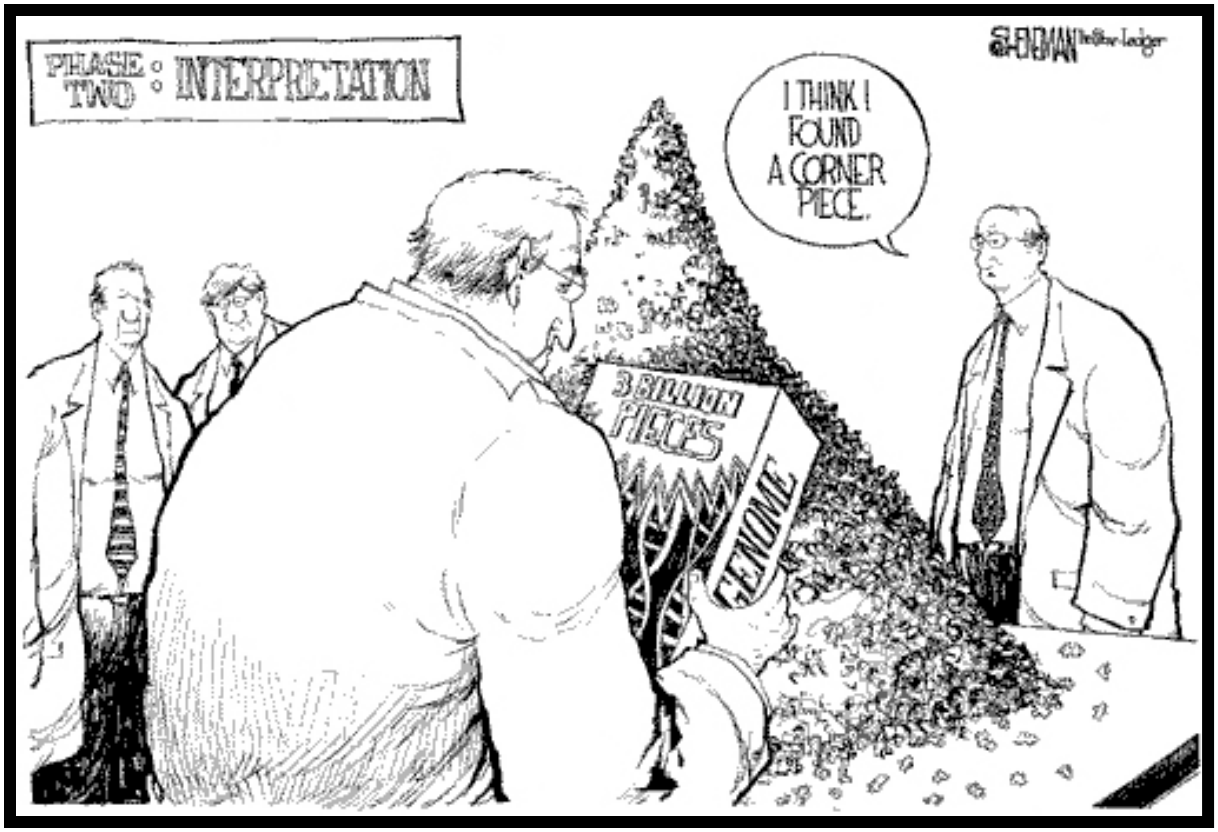
Department of Integrative Physiology
Institute for Behavioral Genetics
Institute of Behavioral Science
University of Colorado Boulder

Department of Epidemiology (secondary)
Colorado School of Public Health
University of Colorado



Let’s begin by taking a look back…

Interpretation Issues…



Interpretation Issues

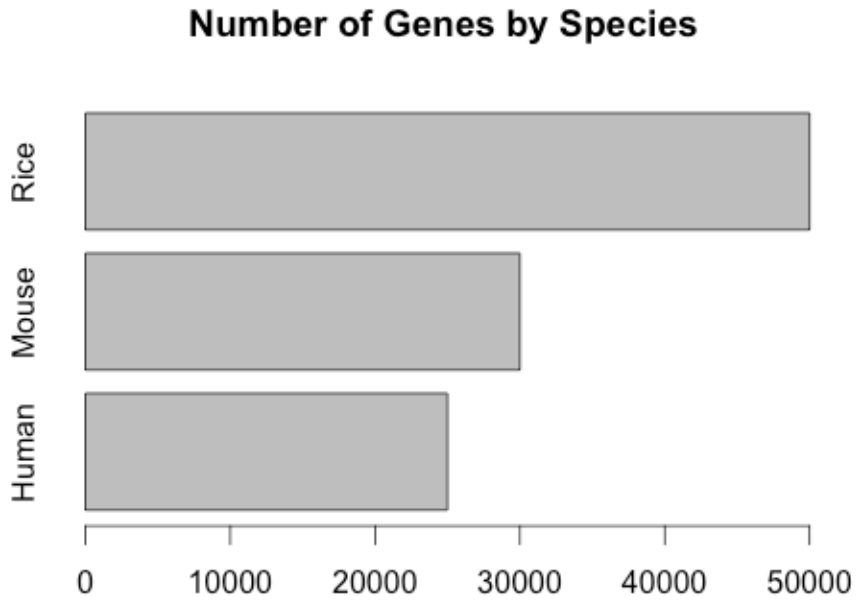
- Do we know what we are looking for?

Let (recent) history be our guide…

- Before the Human Genome Project (2003)
  - Human genome ~ 100,000 – 120,000 genes
- Before the ENCODE Project (2007)
  - Vast majority of genome is 'Junk' ("Junk DNA")
  - Genes are sufficient to understanding biology

Number of genes

Post-HGP



Gene-Centric View

Post ENCODE Project

"I once wrote a book called *The Language of the Genes*, but now biologists are beginning to face up to the uncomfortable truth that they have only been looking at the nouns in life's lexicon — the crudest and most basic elements of any tongue. Now we are reading the spaces in between — verbs, adverbs, adjectives, pronouns and all the rest, and they are complicated indeed."

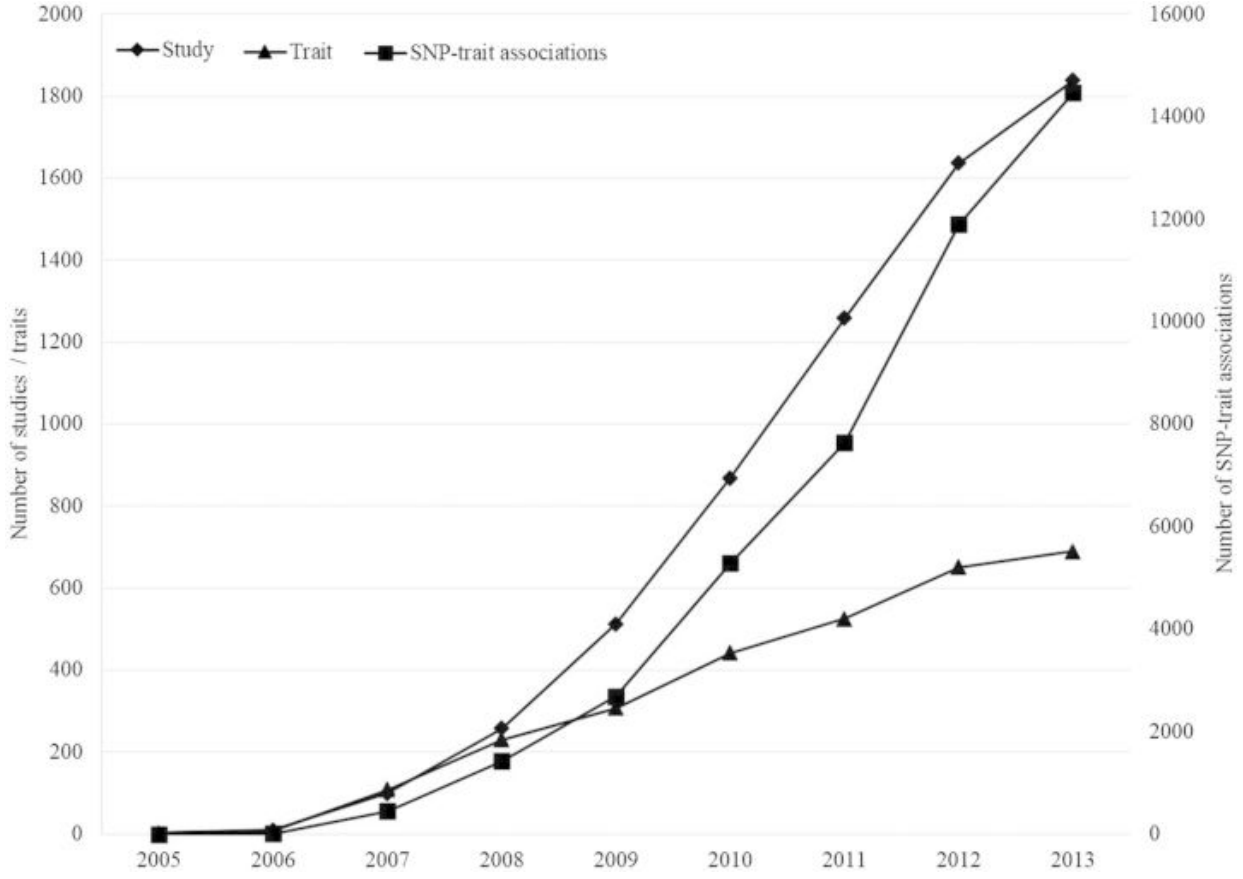
Dr. Steve Jones, University College London (2007)

Genetic Nostradamus?

"HGP scientists thought, and still do, that they could find a small number of genes that were key to these diseases. However, this strategy is flawed, because for most multifactorial diseases affected by many genes those genes have small, not large effects. **And genes with small effects are very hard to find**."

Dr. Richard Strohman, UC Berkeley (**2001**)

Where are we?


Study
Trait
SNP-trait associations
Number of studies / traits
Number of SNP-trait associations
2000
1800
1600
1400
1200
1000
800
600
400
200
0
16000
14000
12000
10000
8000
6000
4000
2000
0
2005
2006
2007
2008
2009
2010
2011
2012
2013

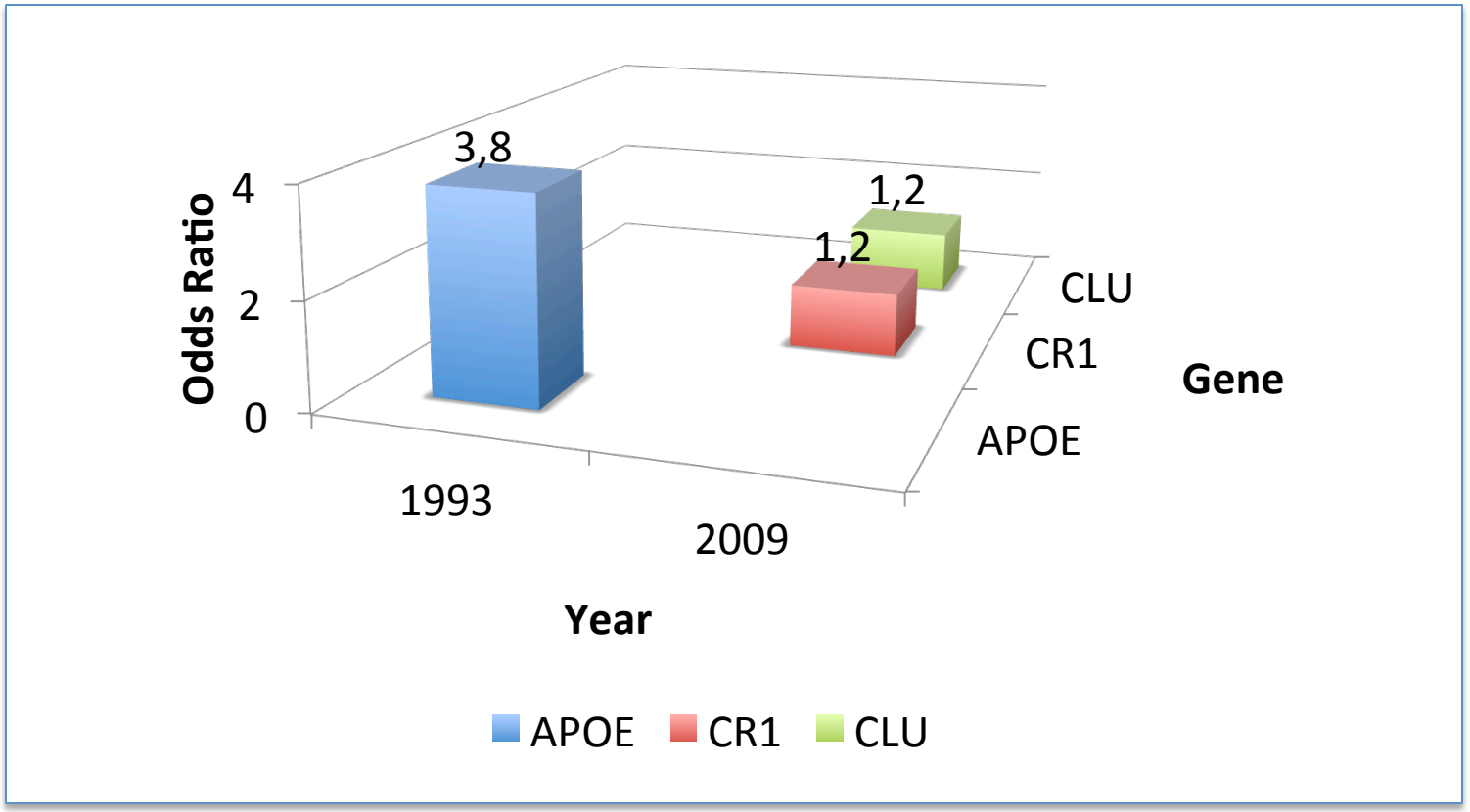
What did we hope to uncover with GWAS?

- Common variants underlying common disease
- We wanted more “APOEs”
  - e4 allele ~ 10-15% frequency
  - Effect size ~ 3-4 (odds ratio) for AD

What did we, in fact, uncover?

What did we find?

Small effects



Has GWAS Been Successful for Complex Disease?

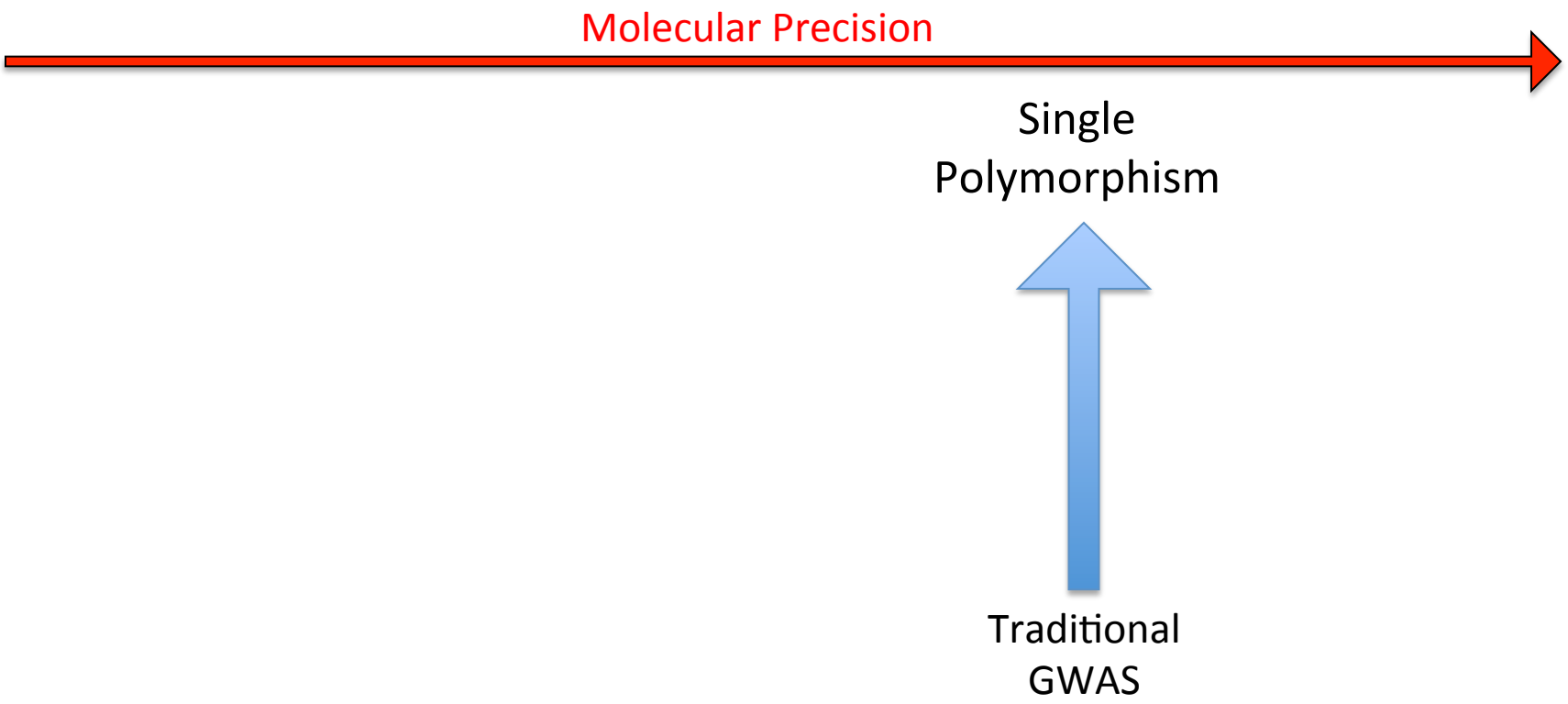
- It Depends…

- But more importantly, it depends on what you define as success

Defining Success for GWAS

- Gain insights into biology, population genetics, gene-flow, evolution
  - Generally, a success but…
    - *It was a really* expensive experiment
  - Good for advancing basic biological knowledge
  - Minimal public health impact on its own

Single Polymorphism



Single Polymorphism

- The original GWAS approach
- Some (*limited*) success
  - Novel targets have been identified
  - However…
    - Minimal public health impact
    - Can't explain more than 5-10% of heritability
    - *Missing Heritability*

2009

Vol 461|8 October 2009|doi:10.1038/nature08494 nature

REVIEWS

Finding the missing heritability of complex diseases

Teri A. Manolio[1], Francis S. Collins[2], Nancy J. Cox[3], David B. Goldstein[4], Lucia A. Hindorff[5], David J. Hunter[6], Mark I. McCarthy[7], Erin M. Ramos[5], Lon R. Cardon[8], Aravinda Chakravarti[9], Judy H. Cho[10], Alan E. Guttmacher[1], Augustine Kong[11], Leonid Kruglyak[12], Elaine Mardis[13], Charles N. Rotimi[14], Montgomery Slatkin[15], David Valle[9], Alice S. Whittemore[16], Michael Boehnke[17], Andrew G. Clark[18], Evan E. Eichler[19], Greg Gibson[20], Jonathan L. Haines[21], Trudy F. C. Mackay[22], Steven A. McCarroll[23] & Peter M. Visscher[24]

Missing Heritability

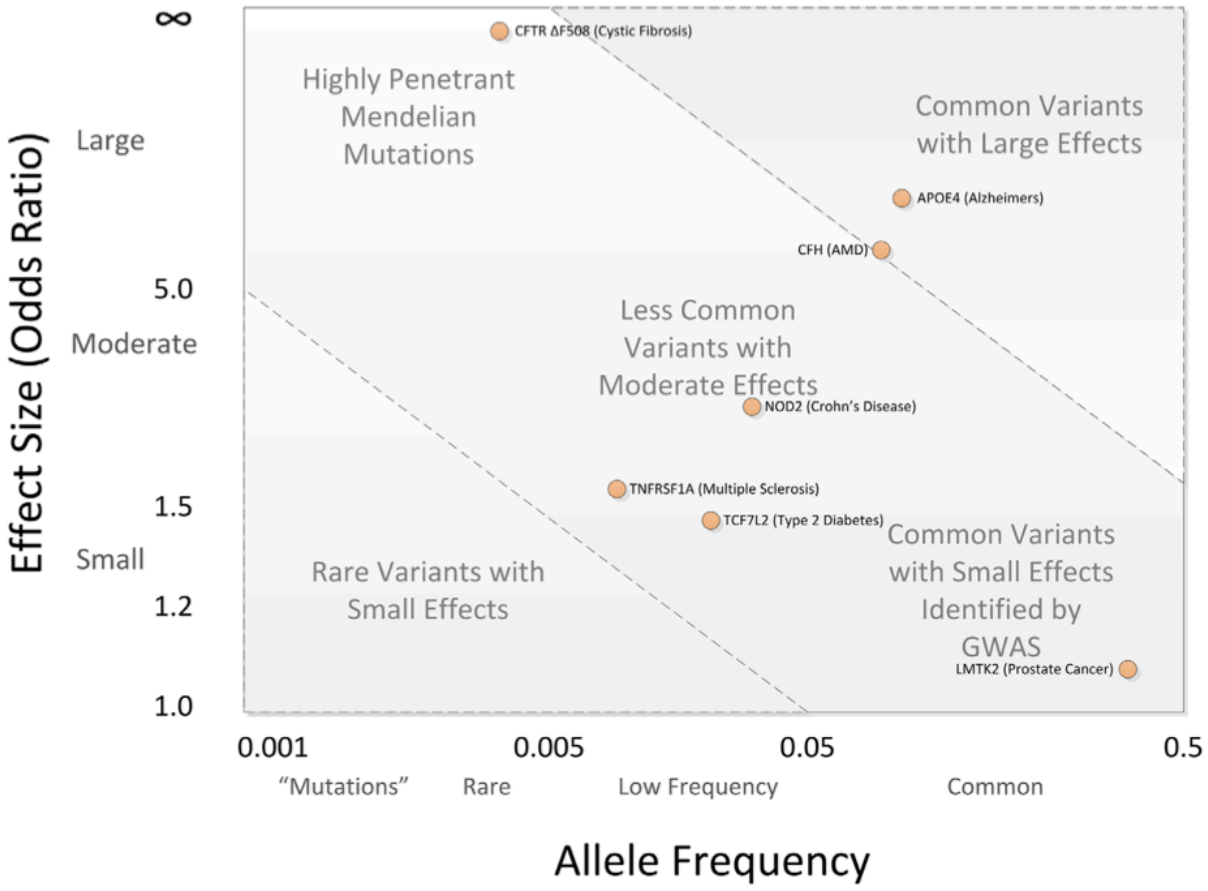
	Variance Explained (h^2)	
Trait/Disease	**Family Studies**	**SNP by SNP**
Height	0.80	0.10
Body Mass Index	0.40-0.60	0.05-0.10
Type 2 diabetes	0.30-0.60	0.05-0.10
HDL cholesterol	0.50	0.10
Breast cancer	0.30	0.08
Multiple sclerosis	0.30-0.80	0.10
Schizophrenia	0.70-0.80	0.01
Bipolar disorder	0.60-0.70	0.02

Adapted from: Visscher et al. (2012), AJHG.

Searching for the Missing Heritability

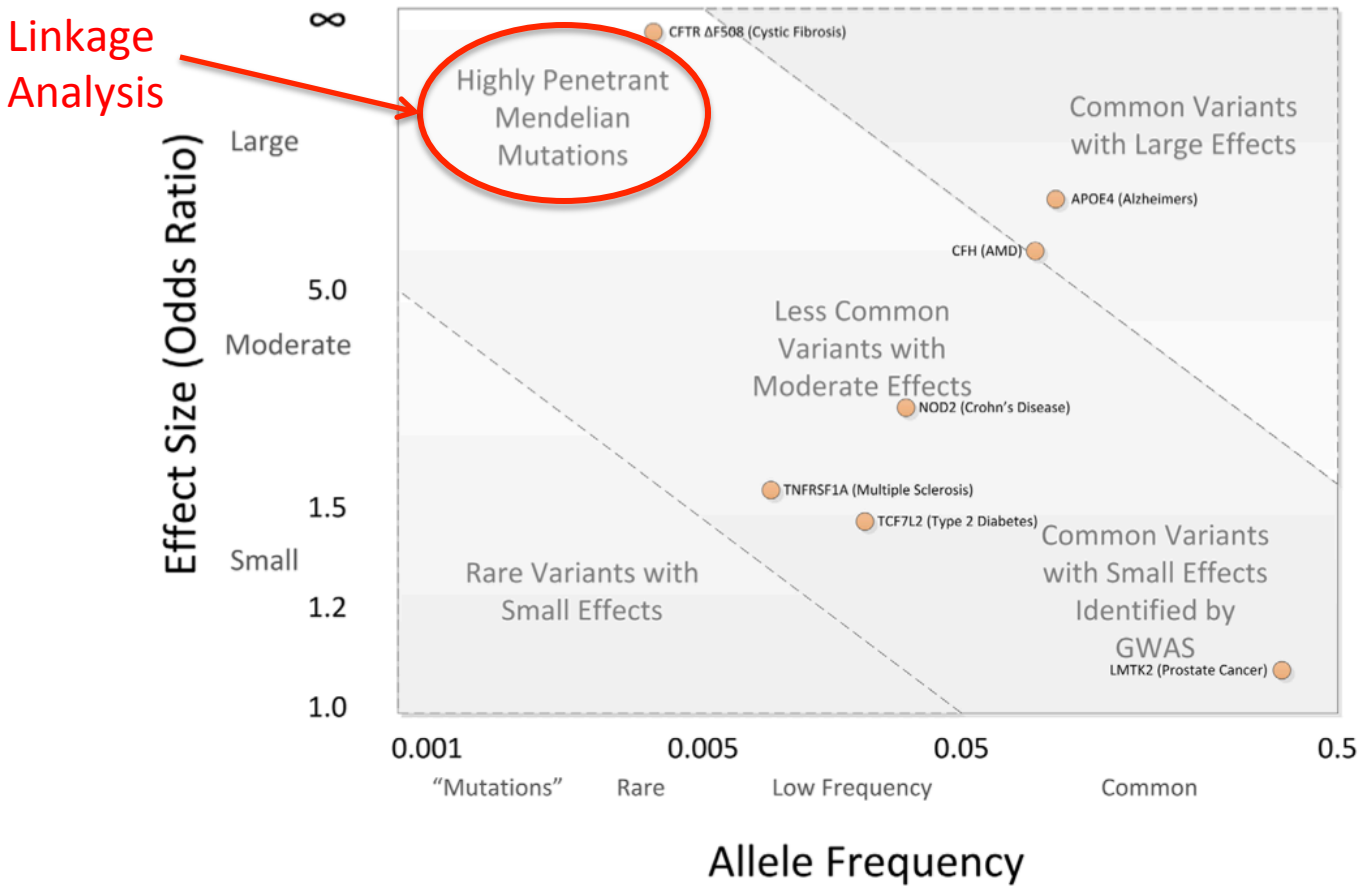
Why are GWAS signals so small?

Allelic Spectrum of Disease



Bush & Moore (2012) in *PLOS Comp Biol*.

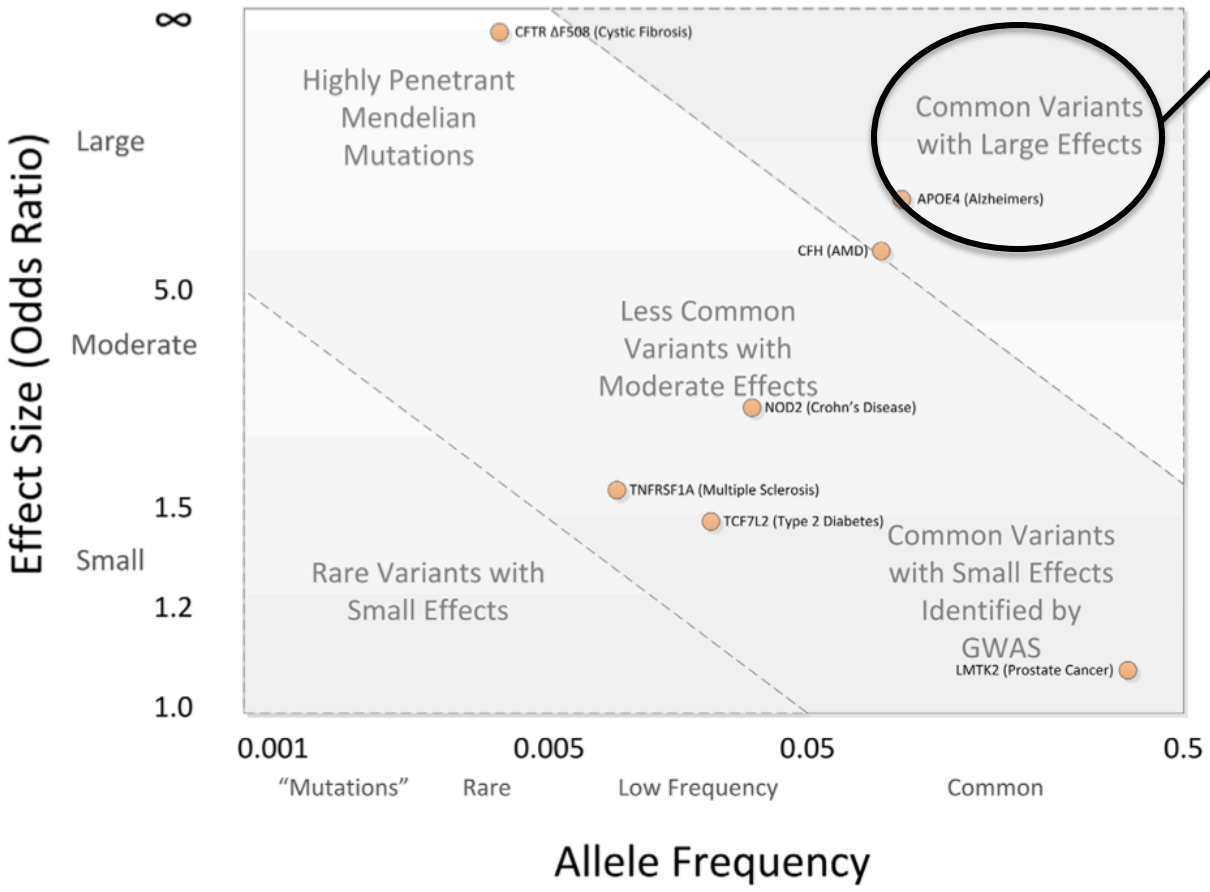
Allelic Spectrum of Disease



Bush & Moore (2012) in *PLOS Comp Biol*.

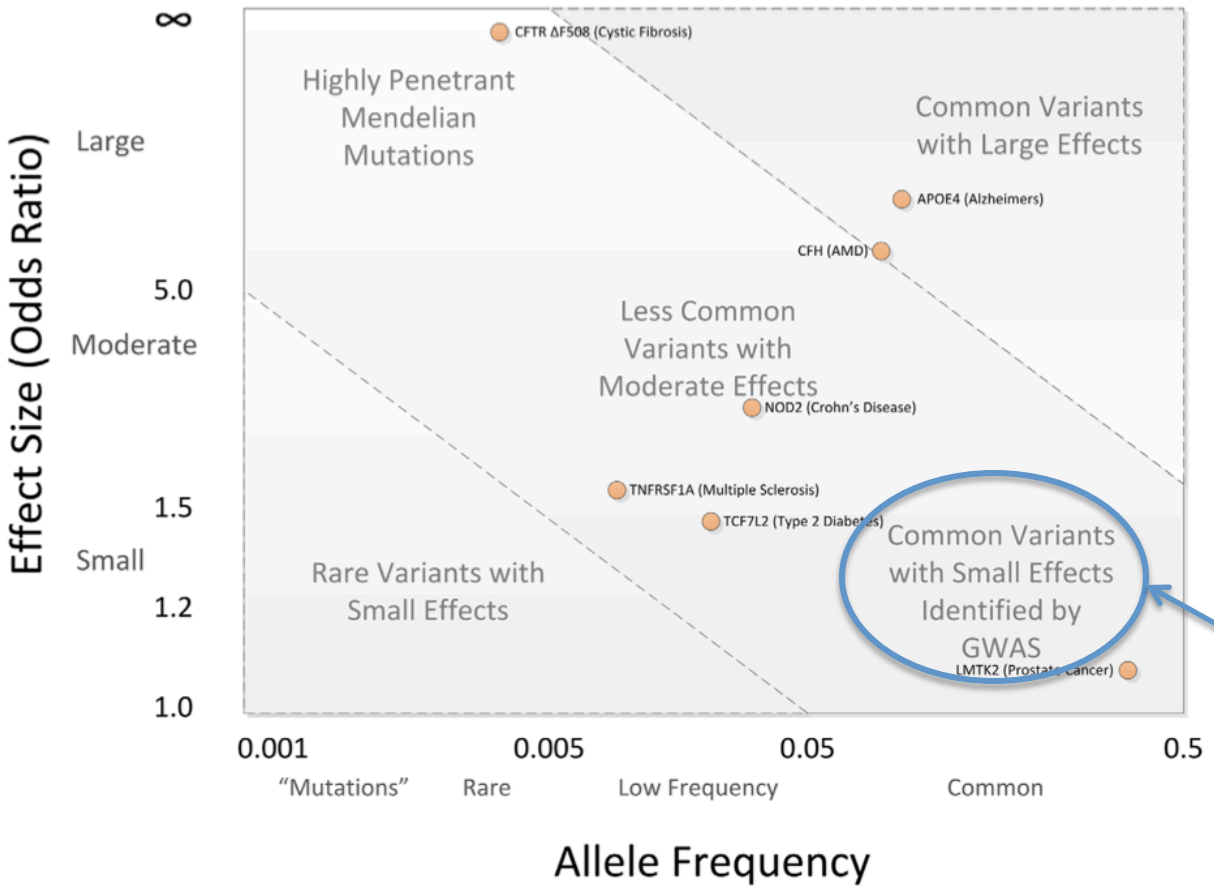
Allelic Spectrum of Disease

APOE-4 and AD



Bush & Moore (2012) in *PLOS Comp Biol*.

Allelic Spectrum of Disease

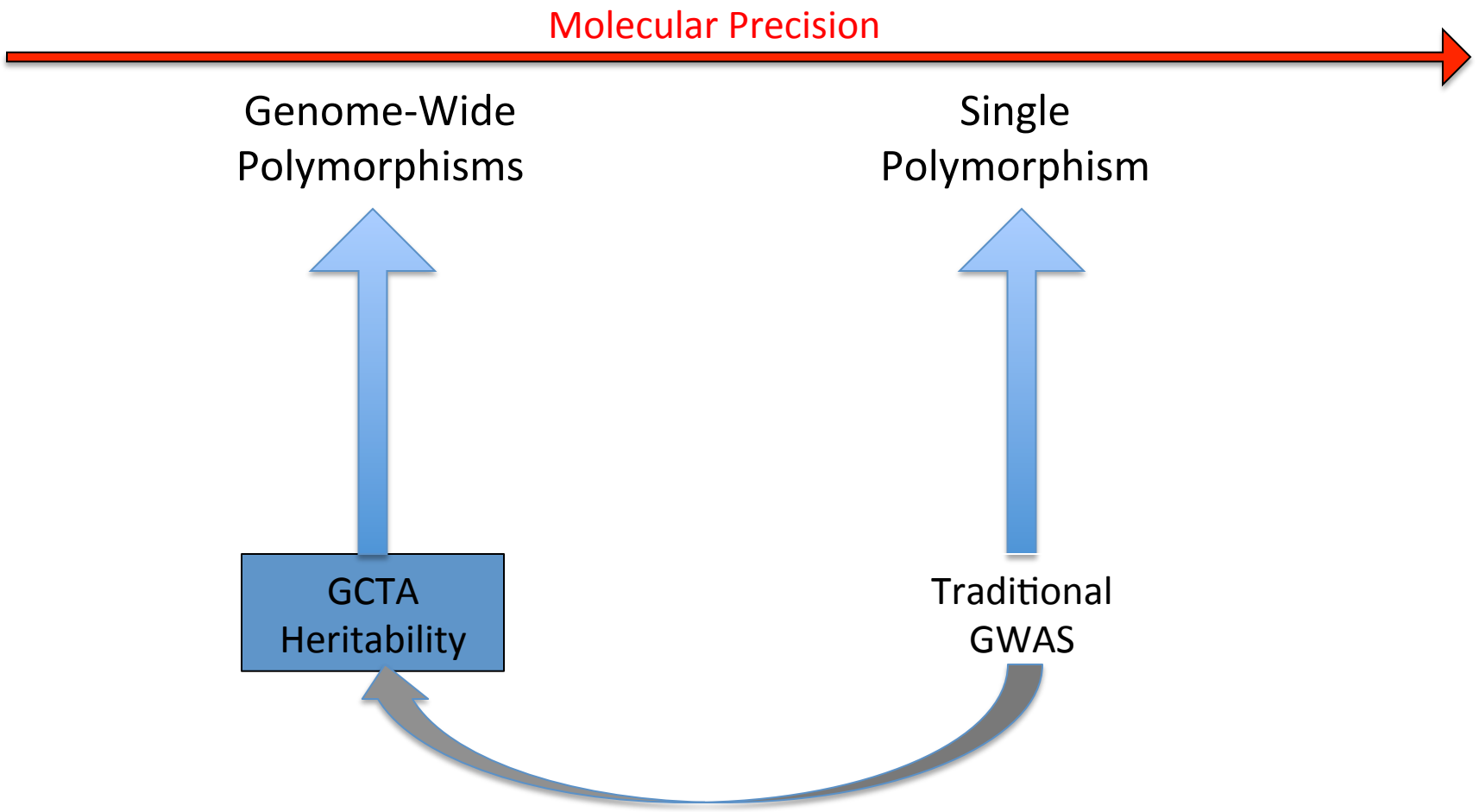


Genome-Wide Association

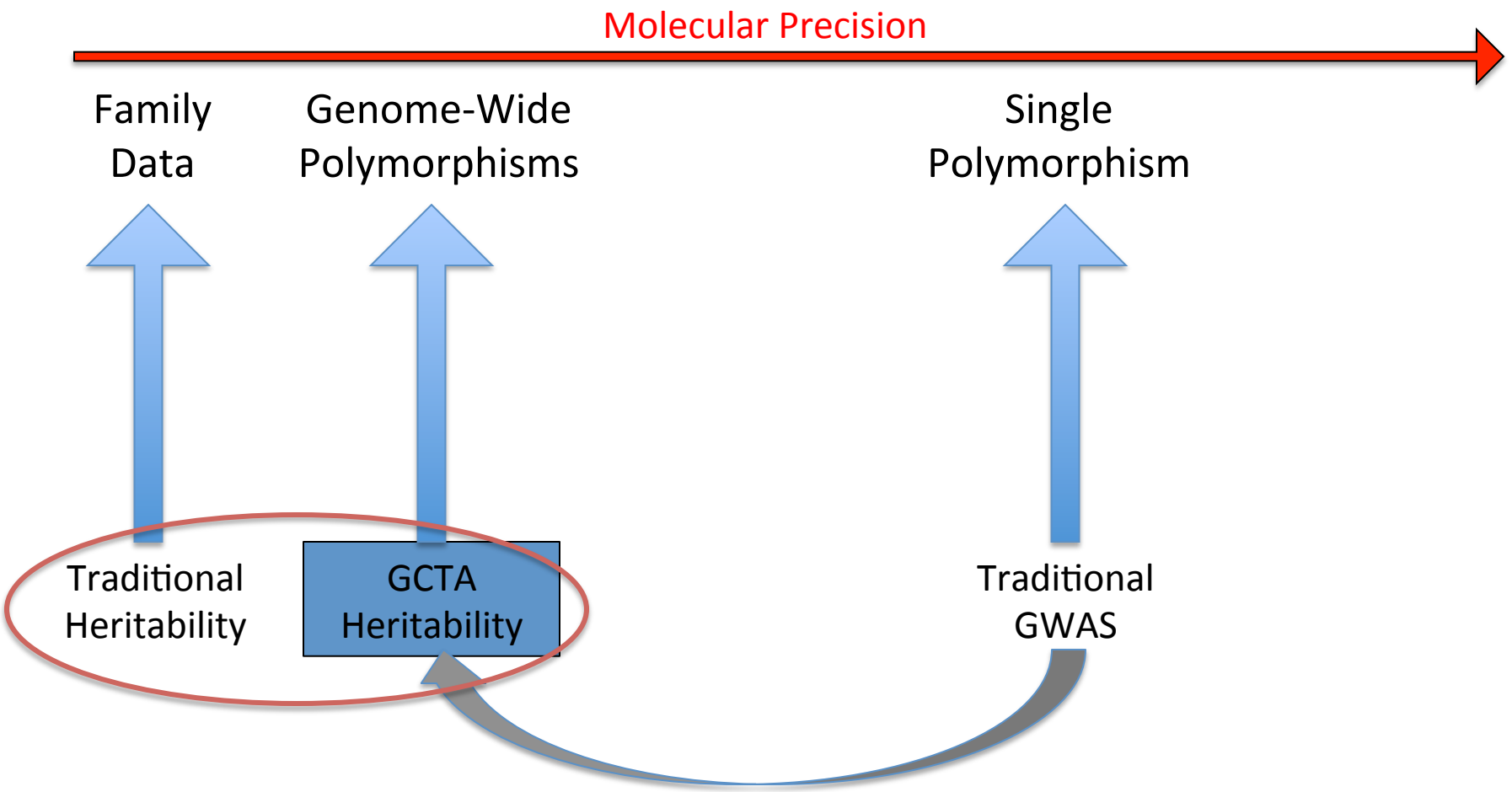
Bush & Moore (2012) in *PLOS Comp Biol*.

Where is this heritability?

Is the heritability really there?



Is the heritability really there?



2010

Common SNPs explain a large proportion of the heritability for human height

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

Genome-wide Complex Trait Analysis (GCTA)

- What does it do?
  - Provides an estimate of heritability using genome-wide data from unrelated samples
- How is this different than a family/twin study?
  - Disentangles issues related to the family environment inherent in twin studies
- Limitation
  - Need fairly large sample sizes (> 5000) to adequately power the approach

Finding Heritability

	Variance Explained (h^2)		
Trait/Disease	**Family Studies**	**SNP by SNP**	**All SNPs**
Height	0.80	0.10	0.50
Body Mass Index	0.40-0.60	0.05-0.10	0.20
Type 2 diabetes	0.30-0.60	0.05-0.10	TBD
HDL cholesterol	0.50	0.10	TBD
Breast cancer	0.30	0.08	TBD
Multiple sclerosis	0.30-0.80	0.10	TBD
Schizophrenia	0.70-0.80	0.01	0.30
Bipolar disorder	0.60-0.70	0.02	0.40

Adapted from: Visscher et al. (2012), AJHG.

Finding Heritability

- It appears to be there, just very difficult to find
- The question then becomes…
  - *How do we find it?*

Do we need to find it at all?

- Research groups focused exclusively on finding the *missing heritability*
- Research groups abandoning genetic studies
- Research groups approaching genetic factors as one of many components underlying a phenotype/disease

Limitations of GCTA as a solution to the missing heritability problem

Siddharth Krishna Kumar[a,1], Marcus W. Feldman[a], David H. Rehkopf[b], and Shripad Tuljapurkar[a]

[a]Department of Biology, Stanford University, Stanford, CA 94305-5020; and [b]School of Medicine, Stanford University, Stanford, CA 94305-5020

Edited by Mary-Claire King, University of Washington, Seattle, WA, and approved November 20, 2015 (received for review October 9, 2015)

Bottom line: GCTA performs quite poorly in stratified samples!

Also in 2009…

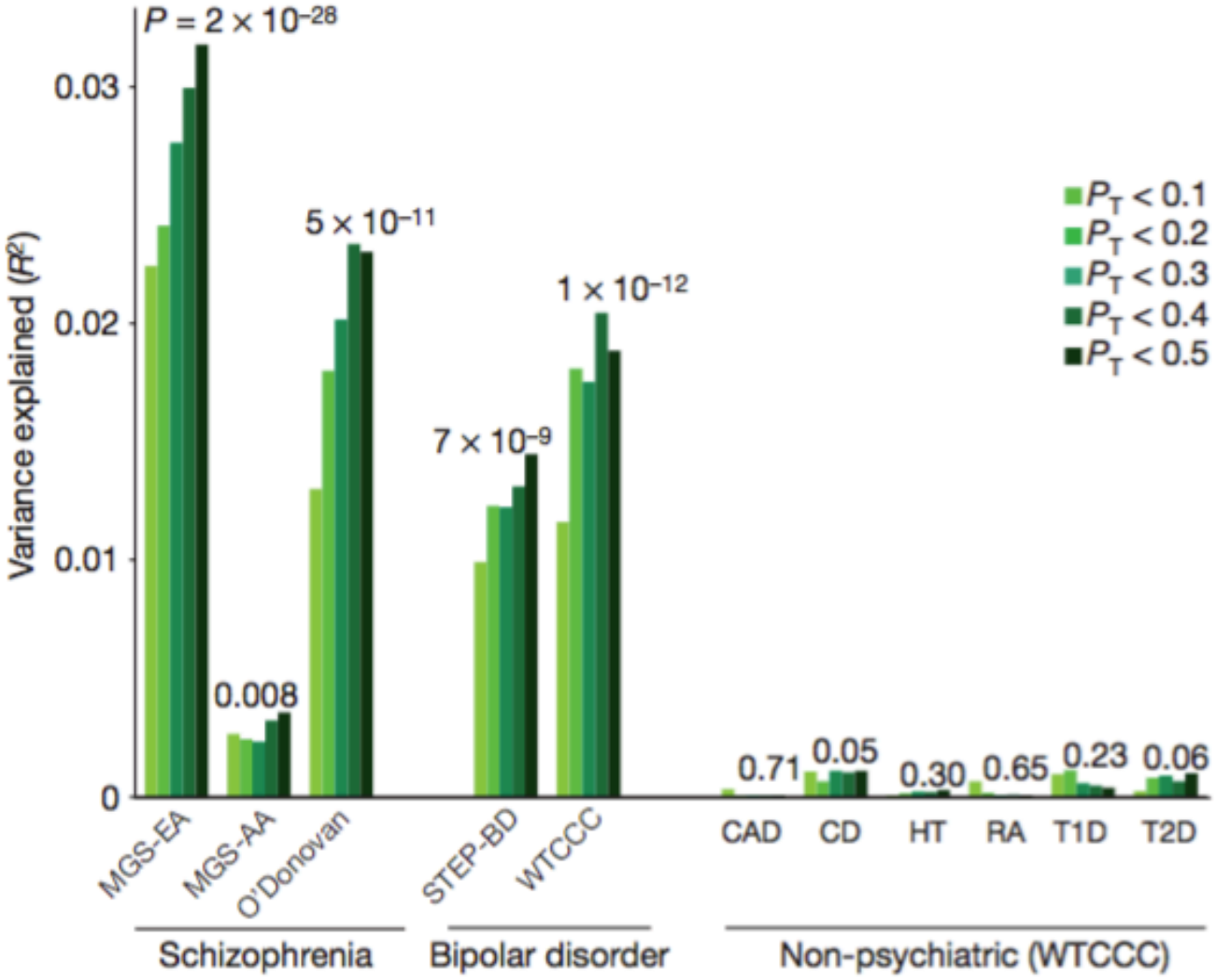
nature Vol 460|6 August 2009|doi:10.1038/nature08185

LETTERS

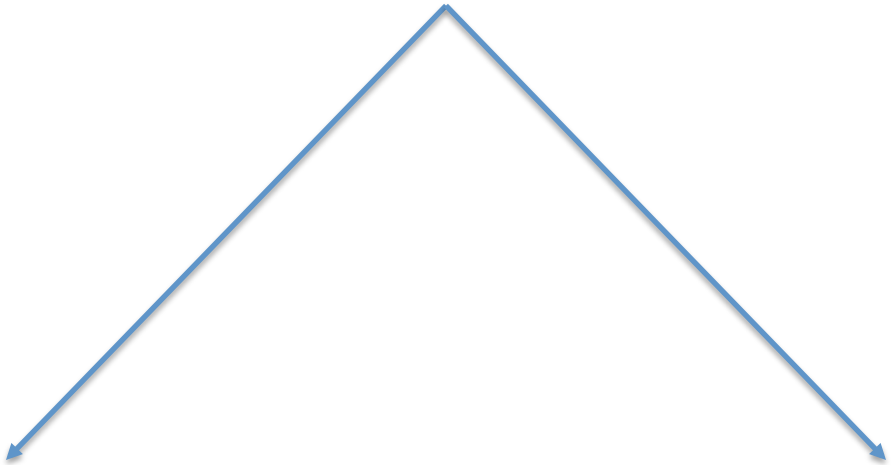
Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

Schizophrenia Consortium



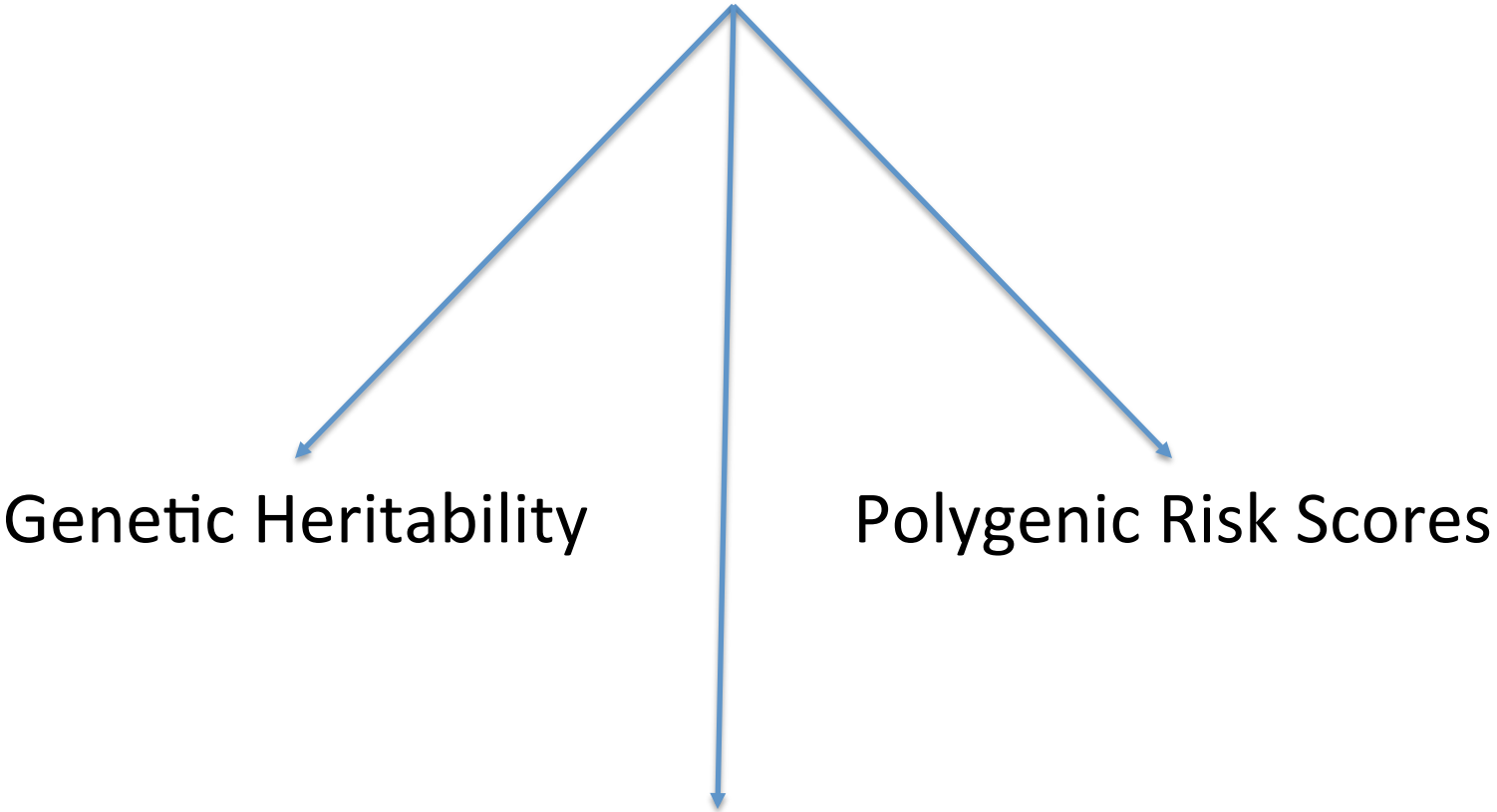
”Post-GWAS Era”



Genetic Heritability

Polygenic Risk Scores

”Post-GWAS Era”

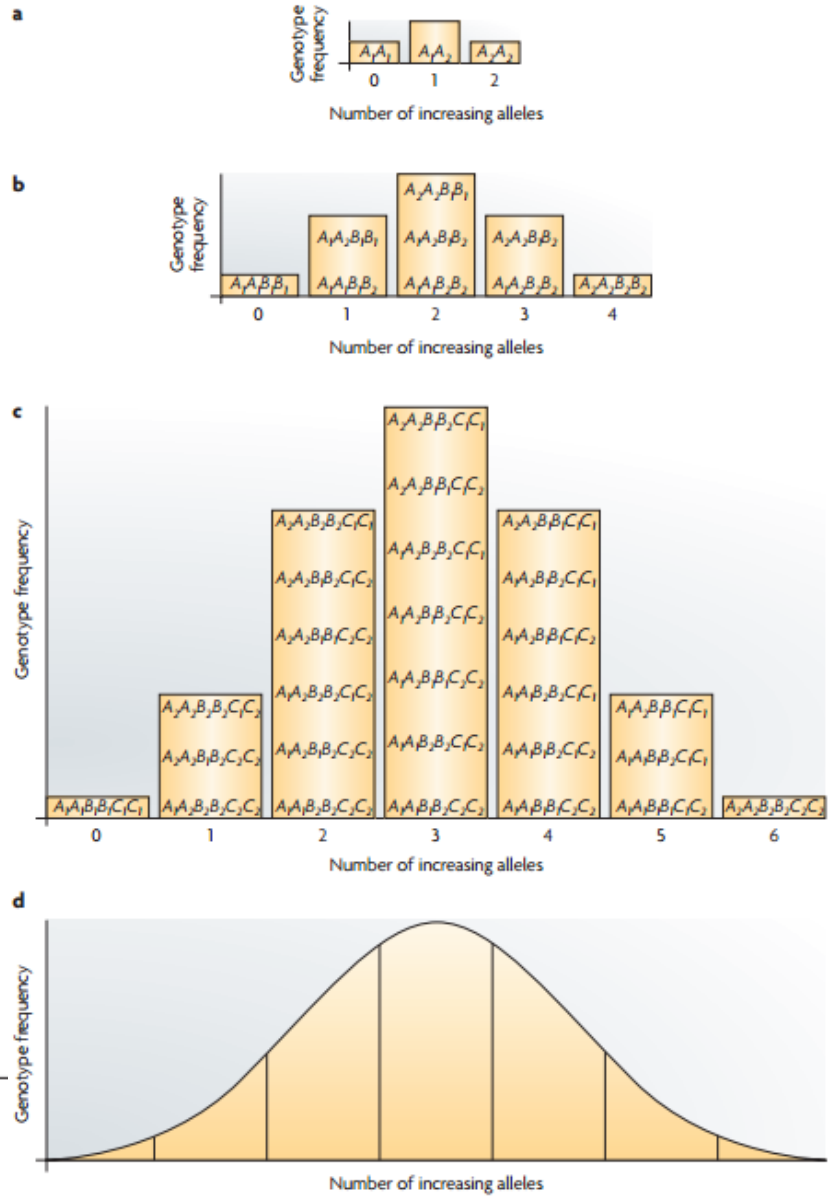


Polygenic Risk Scores

Polygenic Risk Scores

- Overview
- How are they used?
- How are they constructed?

Polygenic Scores (PGS)



GENOME-WIDE ASSOCIATION STUDIES — OPINION

Common disorders are quantitative traits

What is a PGS?

- Sum of the copies of an allele (usually the "risk" allele) within an individual
- Often a *weighted* sum of alleles weighted by parameter estimates (regression) from a previous study

Different forms of the PGS

- Top hits
  - Only SNPs that reach some (stringent) level of statistical significance (1e-8)
- Genome-wide
  - All SNPs or all below some nominal level of significance (0.5)
  - All SNPs available

Where does the data come from?

- Generally, large consortia will generate the association results that ultimately become a polygenic score

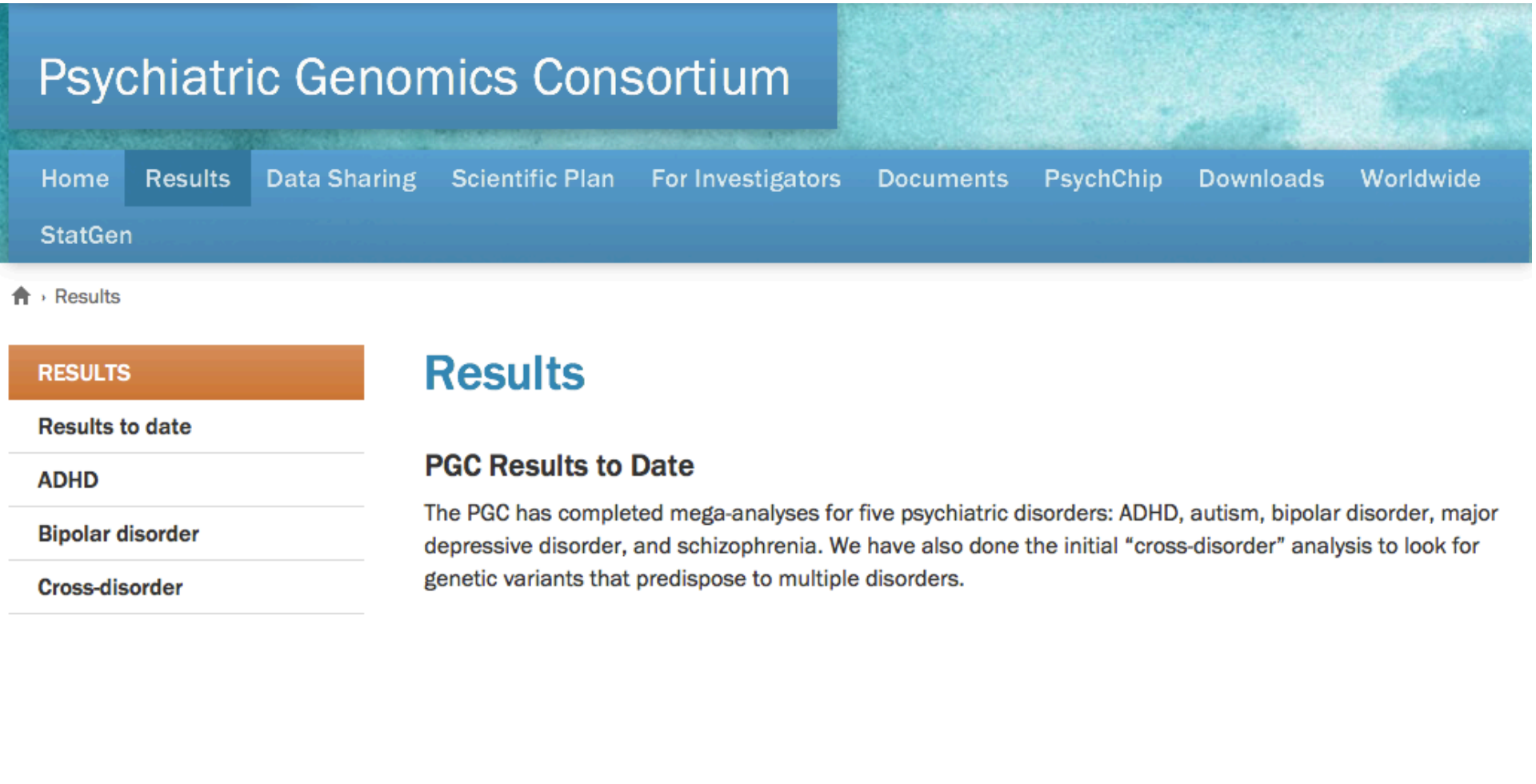
GIANT Consortium

nature genetics

Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index

Obesity is globally prevalent and highly heritable, but its underlying genetic factors remain largely elusive. To identify genetic loci for obesity susceptibility, we examined associations between body mass index and ~2.8 million SNPs in up to 123,865 individuals with targeted follow up of 42 SNPs in up to 125,931 additional individuals. We confirmed 14 known obesity susceptibility loci and identified 18 new loci associated with body mass index ($P < 5 \times 10^{-8}$), one of which includes a copy number variant near *GPRC5B*. Some loci (at *MC4R*, *POMC*, *SH2B1* and *BDNF*) map near key hypothalamic regulators of energy balance, and one of these loci is near *GIPR*, an incretin receptor. Furthermore, genes in other newly associated loci may provide new insights into human body weight regulation.

Psychiatric outcomes

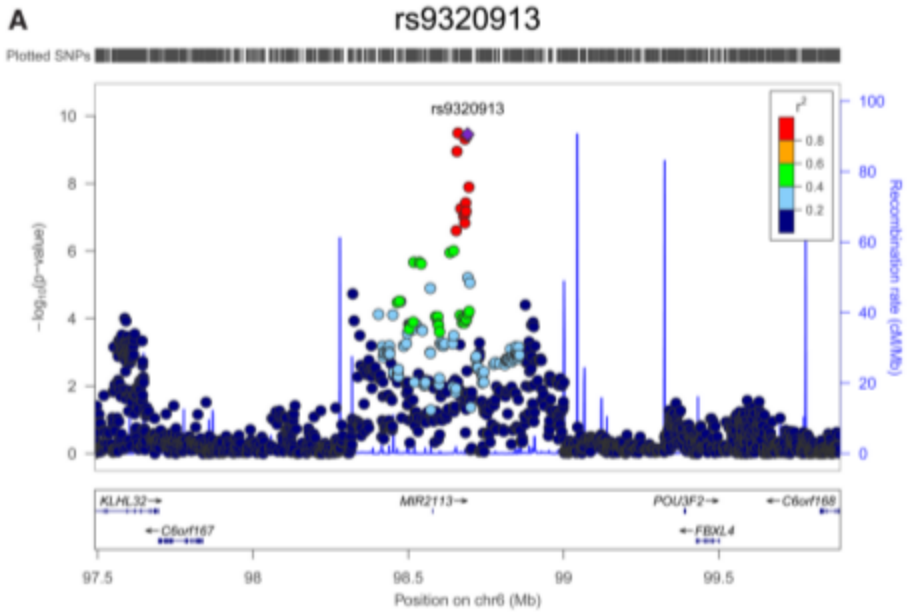


Educational Attainment

Sciencexpress

GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

All authors with their affiliations appear at the end of this paper.



Accessing the data

- Quite accessible
  - Most results are available online to download!

- GIANT Consortium

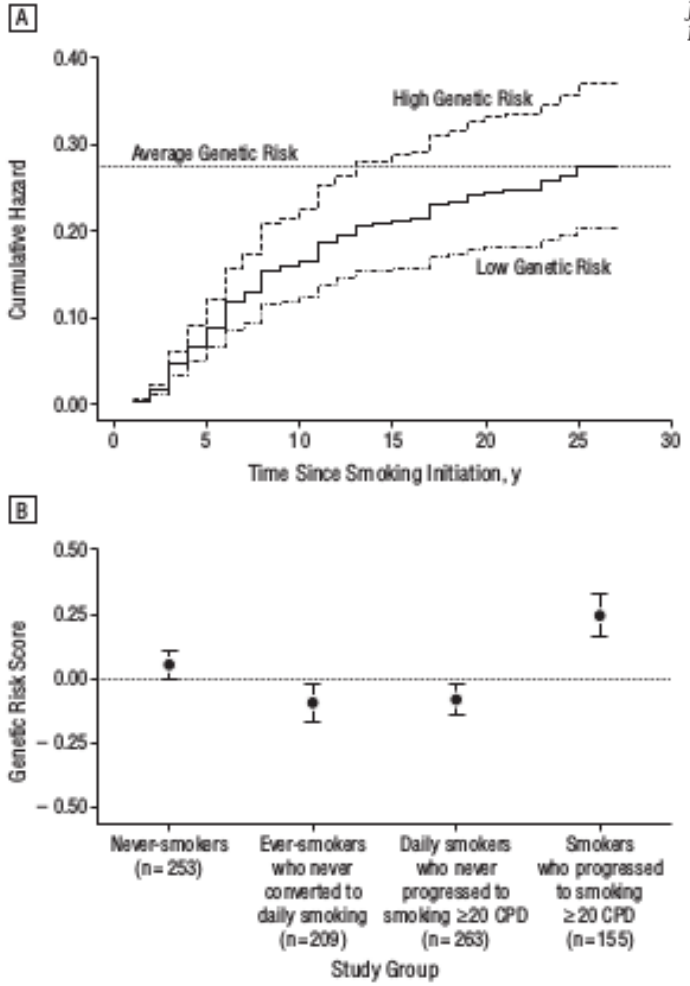
How are PGS being used?

Characterizing an outcome

Polygenic Risk and the Developmental Progression to Heavy, Persistent Smoking and Nicotine Dependence

Evidence From a 4-Decade Longitudinal Study

Daniel W. Belsky, PhD; Terrie E. Moffitt, PhD; Timothy B. Baker, PhD; Andrea K. Biddle, PhD; James P. Evans, MD, PhD; HonaLee Harrington, BA; Renate Houts, PhD; Madeline Meier, PhD; Karen Sugden, PhD; Benjamin Williams, BS; Richie Poulton, PhD; Avshalom Caspi, PhD

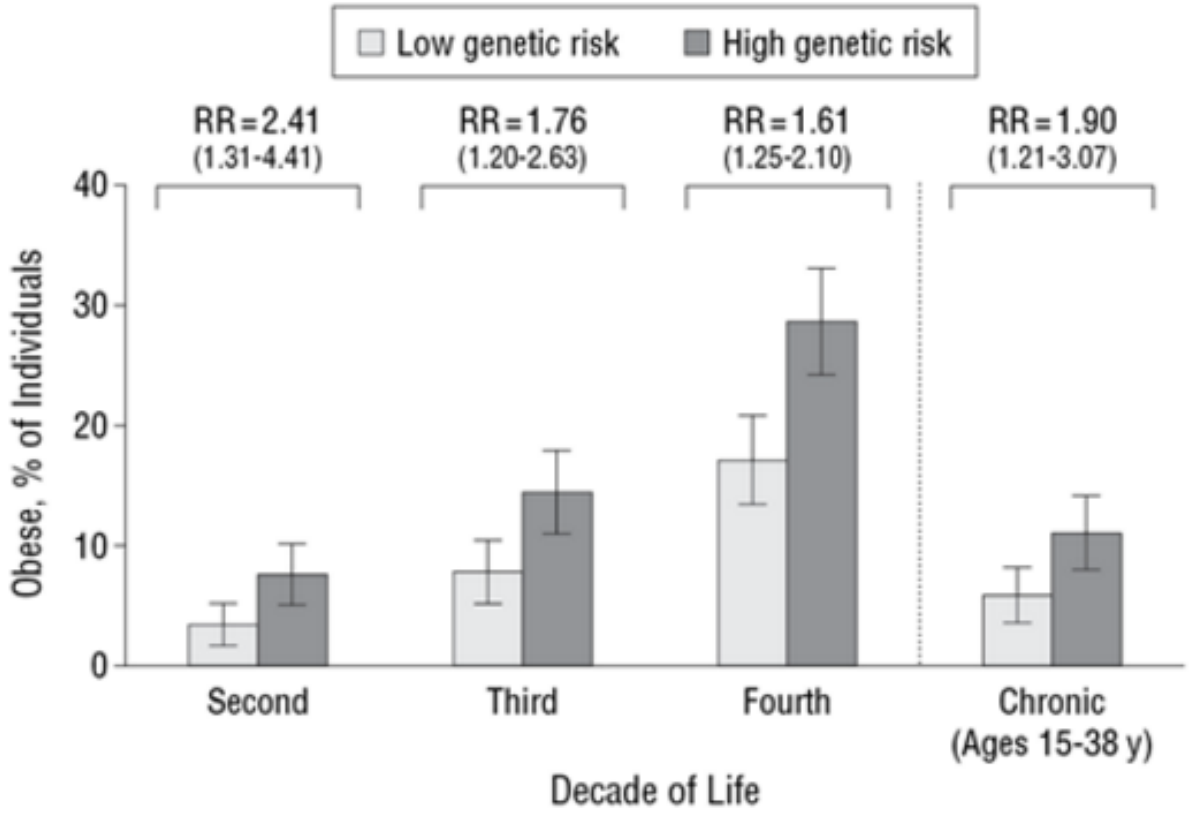


Polygenic Risk, Rapid Childhood Growth, and the Development of Obesity

Evidence From a 4-Decade Longitudinal Study FREE

Daniel W. Belsky, PhD; Terrie E. Moffitt, PhD; Renate Houts, PhD; Gary G. Bennett, PhD; Andrea K. Biddle, PhD; James A. Blumenthal, PhD; James P. Evans, MD, PhD; HonaLee Harrington, BA; Karen Sugden, PhD; Benjamin Williams, BS; Richie Poulton, PhD; Avshalom Caspi, PhD

[+] Author Affiliations



Sugar-Sweetened Beverages and Genetic Risk of Obesity

Qibin Qi, Ph.D., Audrey Y. Chu, Ph.D., Jae H. Kang, Sc.D., Majken K. Jensen, Ph.D., Gary C. Curhan, M.D., Sc.D., Louis R. Pasquale, M.D., Paul M. Ridker, M.D., M.P.H., David J. Hunter, M.B., B.S., Sc.D., Walter C. Willett, M.D., Dr.P.H., Eric B. Rimm, Sc.D., Daniel I. Chasman, Ph.D., Frank B. Hu, M.D., Ph.D., and **Lu Qi, M.D., Ph.D.**

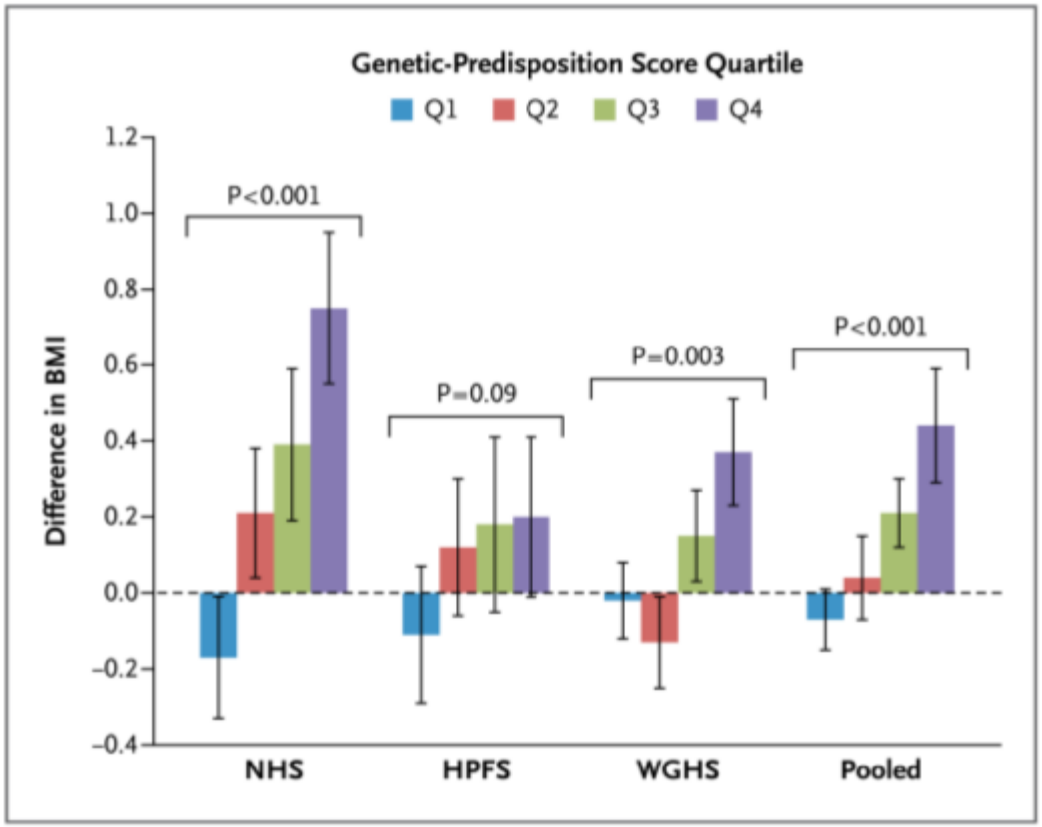
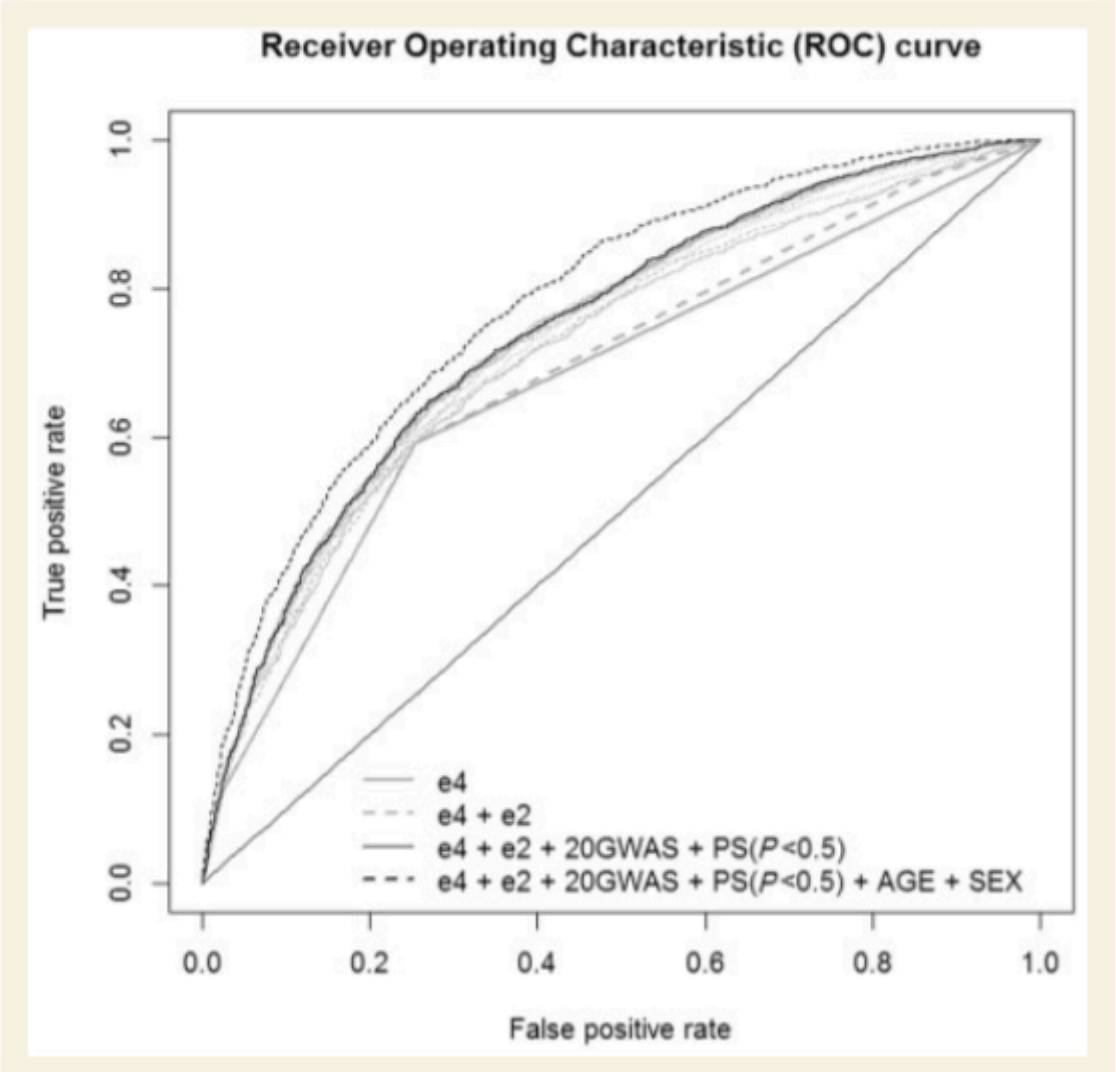


Figure 2. Difference in BMI Associated with One Serving of a Sugar-Sweetened Beverage per Day, According to the Quartile of the Genetic-Predisposition Score

Improving prediction

Alzheimer's Disease



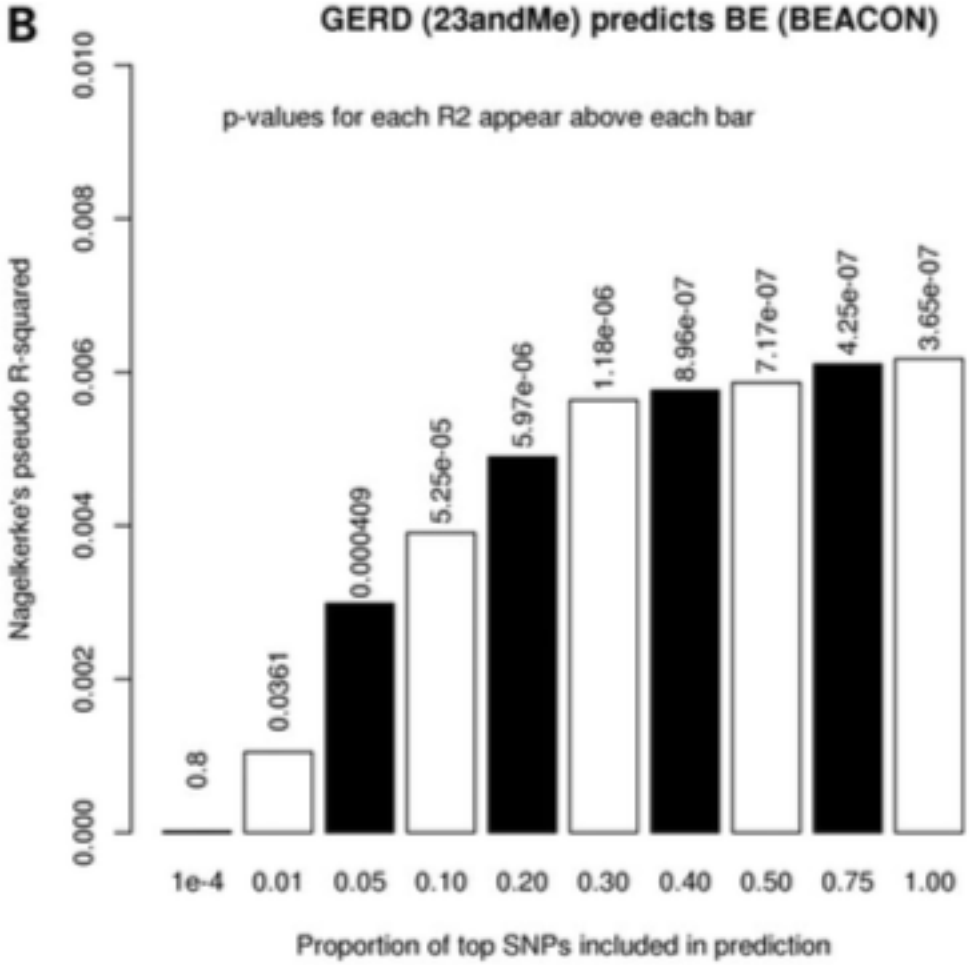
Identifying shared vulnerability

ASSOCIATION STUDIES ARTICLE

Chronic gastroesophageal reflux disease shares genetic background with esophageal adenocarcinoma and Barrett's esophagus

Puya Gharahkhani[1,*], Joyce Tung[3], David Hinds[3], Aniket Mishra[1,4], Barrett's and Esophageal Adenocarcinoma Consortium (BEACON), Thomas L. Vaughan[5], David C. Whiteman[2] and Stuart MacGregor[1], on behalf of the BEACON study investigators

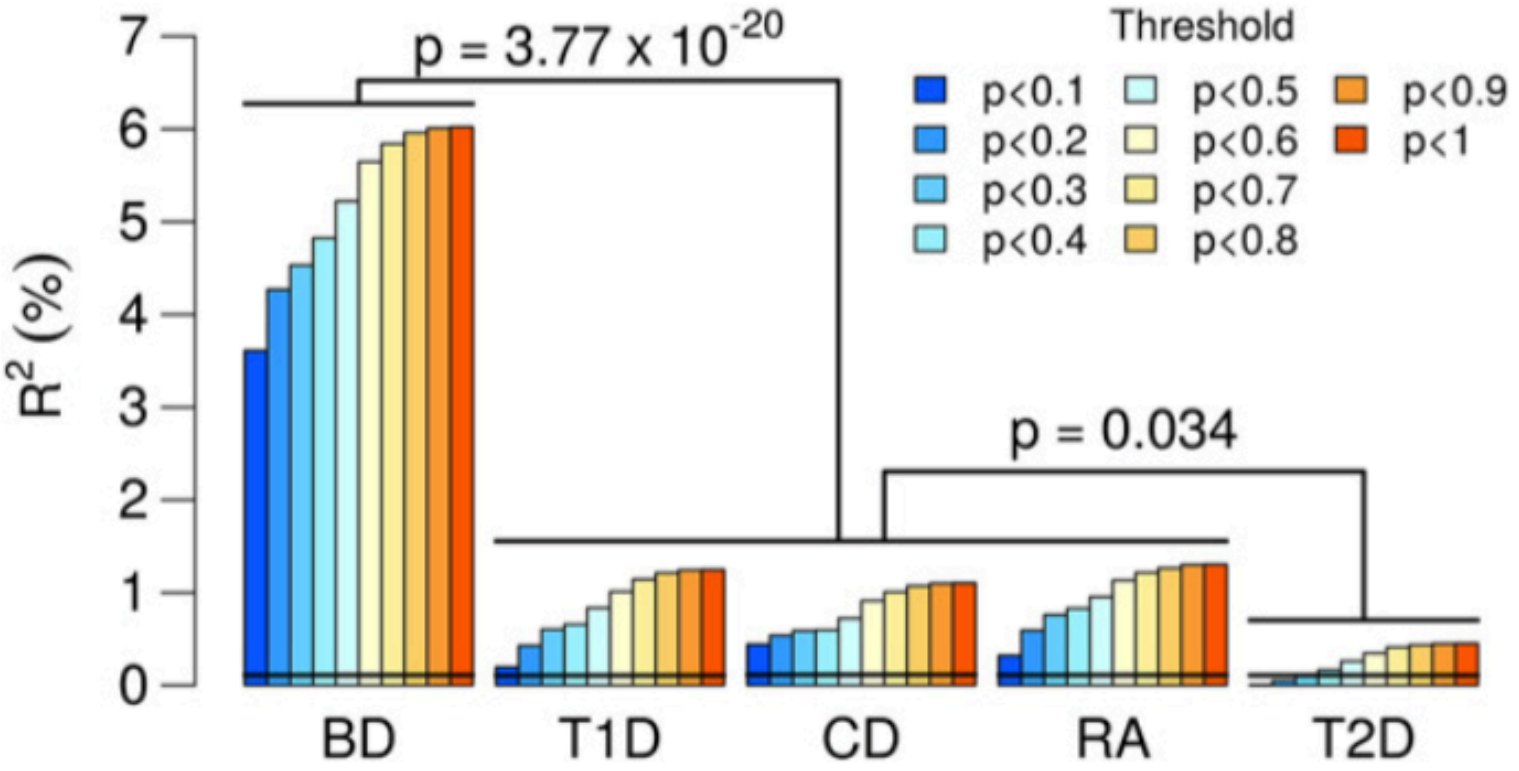
Human Molecular Genetics, 2016, Vol. 25, No. 4 828–835



Genetic liability for schizophrenia predicts risk of immune disorders

Sven Stringer [a,b,*], René S. Kahn [b], Lot D. de Witte [b], Roel A. Ophoff [b,c], Eske M. Derks [a]

[a] Department of Psychiatry, Amsterdam Medical Center, Amsterdam, The Netherlands
[b] Department of Psychiatry, Rudolf Magnus Institute of Neuroscience, University Medical Center, Utrecht, The Netherlands
[c] University California Los Angeles, Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, Los Angeles, CA, USA



Constructing a PGS

Some issues to consider

- Additive Assumptions
- Strand Ambiguity
- Weights
- LD Considerations

Additive Assumptions

- Polygenic scores are constructed additively
  - Sum of the count of alleles over many SNPs
- However, entirely reasonable to consider other models if they are established
  - Dominant coding, recessive coding on a SNP-by-SNP basis

Strand Ambiguity

Strand Orientation

- Strand unambiguous
  - A/G alleles -> T/C alleles
- Strand ambiguous
  - A/T and C/G

Strand Reporting

- Probe/Target
  - Generically as A/B
  - Illumina/Affymetrix
- Plus (+) / Minus (-)
  - 5’ end at the tip of the short arm = plus
  - 1000 genomes, HapMap
- FWD / REV
  - Based upon submitted flanking DNA sequence
  - dbSNP (NCBI)
- TOP/BOT
  - Based upon flanking sequences
  - Illumina

Strand issues

- In a “top hits” PGS, you can examine SNP-by-SNP to determine the correct strand
  - Authoritative info hard to come by
  - Allele frequencies have to align
- In a full genome-wide PGS, much more difficult
  - People tend to discard a lot of ambiguous SNPs to simplify the issue

Weights

Weights

- Each SNP weighted by the *magnitude* of its estimated effect.
  - For continuous traits : the estimated regression coefficient (’beta’)
  - For dichotomous traits : log(OR)
- Pay attention to the direction of the effect!
  - Negative effect estimate
    - Alternative allele = “risk” allele"

GIANT Consortium Data

LD Considerations

LD Considerations

- Too many variants in LD could lead to overemphasis of that region in the PGS
- Instead of LD pruning, LD **clumping**:
  - Select variants in LD blocks that are most highly associated with outcome

Hot off the press…

Am J Hum Genet. 2015 Oct 1;97(4):576-92. doi: 10.1016/j.ajhg.2015.09.001.

Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.

Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, Genovese G, Loh PR, Bhatia G, Do R, Hayeck T, Won HH; Schizophrenia Working Group of the Psychiatric Genomics Consortium; Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Kathiresan S, Pato M, Pato C, Tamimi R, Stahl E, Zaitlen N, Pasaniuc B, Belbin G, Kenny EE, Schierup MH, De Jager P, Patsopoulos NA, McCarroll S, Daly M, Purcell S, Chasman D, Neale B, Goddard M, Visscher PM, Kraft P, Patterson N, Price AL; Discovery Biology and Risk of Inherited Variants in Breast Cancer DRIVE study.

⊕ **Collaborators (368)**

Abstract

Polygenic risk scores have shown great promise in predicting complex disease risk and will become more accurate as training sample sizes increase. The standard approach for calculating risk scores involves linkage disequilibrium (LD)-based marker pruning and applying a p value threshold to association statistics, but this discards information and can reduce predictive accuracy. We introduce LDpred, a method that infers the posterior mean effect size of each marker by using a prior on effect sizes and LD information from an external reference panel. Theory and simulations show that LDpred outperforms the approach of pruning followed by thresholding, particularly at large sample sizes. Accordingly, predicted R(2) increased from 20.1% to 25.3% in a large schizophrenia dataset and from 9.8% to 12.0% in a large multiple sclerosis dataset. A similar relative improvement in accuracy was observed for three additional large disease datasets and for non-European schizophrenia samples. The advantage of LDpred over existing methods will grow as sample sizes increase.

Simple Example

SNP	Risk Allele	Weight
1	A	0.06
2	A	0.07
3	A	0.13
4	C	0.22

Simple Example

SNP	Risk Allele	Weight
1	A	0.06
2	A	0.07
3	A	0.13
4	C	0.22

ID	SNP1	SNP2	SNP3	SNP4	PGS(u)	PGS(w)
1	AG (1)	AA (2)	GG (0)	CT (1)	4	0.55
2	GG (0)	AG (1)	AG (1)	TT (0)	2	0.20
3	AA (2)	AA (2)	AA (2)	CC (2)	6	0.96

Overall limitations

- Polygenic Scores only capture GWAS variants
  - Common variants
  - Relatively constant prediction across environments
- Conservative method for testing GxE
  - Need to establish PGS within environments
- Ethnic homogeneity
  - Most (if not all) scores derive from predominantly European samples

Polygenic Risk Predicts Obesity in Both White and Black Young Adults

Benjamin W. Domingue[1], Daniel W. Belsky[2], Kathleen Mullan Harris[3], Andrew Smolen[4], Matthew B. McQueen[5], Jason D. Boardman[1]*

1 Institute of Behavioral Science, University of Colorado Boulder, Boulder, CO, United States of America, **2** Center for the Study of Aging and Human Development, Duke University Medical Center, Durham, NC, United States of America, **3** Sociology Department and the Carolina Population Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States of America, **4** Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, CO, United States of America, **5** Department of Integrative Physiology, University of Colorado Boulder, Boulder, CO, United States of America

Polygenic Scores

- European PGS
  - Based upon 32 SNPs in Speliotes et al
- African-American PGS
  - Based upon 8 SNPs in Monda et al

Table 3. Characteristics of white and black young adults in the Add Health Sibling Pairs sample.

	Whites (N = 918)		Blacks (N = 677)		p-value for difference
	Mean	SD	Mean	SD	
% Male	0.48	0.50	0.46	0.50	0.44
BMI-Wave 3	25.78	5.80	26.39	6.32	0.07
BMI-Wave 4	27.86	6.60	29.34	7.44	0.00
BMI Change	2.10	3.93	2.69	3.99	0.01
Waist/Height-Wave 4	0.57	0.10	0.58	0.11	0.15
% Obese-Wave 3	0.22	0.42	0.26	0.44	0.13
% Obese-Wave 4	0.33	0.47	0.40	0.49	0.01

Note: Data are for the Sibling Pairs of the National Longitudinal Study of Adolescent Health [17].

doi:10.1371/journal.pone.0101596.t003

Table 4. Genetic associations with body-mass index and obesity in white and black young adults in the Add Health Sibling Pairs sample estimated using the genetic risk score for Europeans (GRS-E).

	Obesity Phenotype	White Sample		Black Sample		*p-value for difference*
Unweighted		B [95% CI]				
	BMI-Wave 3	0.16***	[0.09, 0.23]	0.14**	[0.06, 0.23]	*0.96*
	BMI-Wave 4	0.17***	[0.10, 0.24]	0.13**	[0.04, 0.21]	*0.76*
	Change	0.06*	[0.01, 0.10]	0.01	[−0.04, 0.05]	*0.23*
		OR [95% CI]				
	Obesity-Wave 3	1.42**	[1.14, 1.78]	1.19	[0.96, 1.48]	*0.38*
	Obesity-Wave 4	1.54***	[1.30, 1.83]	1.19	[0.98, 1.45]	*0.06*
	Change	1.43**	[1.14, 1.79]	1.09	[0.83, 1.45]	*0.12*
Weighted		B [95% CI]				
	BMI-Wave 3	0.16***	[0.09, 0.23]	0.16***	[0.07, 0.24]	*0.83*
	BMI-Wave 4	0.18***	[0.10, 0.25]	0.14***	[0.06, 0.22]	*0.85*
	Change	0.06**	[0.02, 0.11]	0.01	[−0.03, 0.06]	*0.21*
		OR [95% CI]				
	Obesity-Wave 3	1.37**	[1.10, 1.71]	1.25*	[1.01, 1.56]	*0.68*
	Obesity-Wave 4	1.56***	[1.31, 1.85]	1.22*	[1.00, 1.48]	*0.05*
	Change	1.48***	[1.18, 1.86]	1.10	[0.83, 1.46]	*0.07*

* p<.05; ** p<.01; *** p<.001.

Note: All data come from the National Longitudinal Study of Adolescent Health Sibling Pairs [17]. Genetic risk was measured using the genetic risk score for Europeans (GRS-E). Regressions were estimated using multi-level models [20] to account for the clustering of observations within families and adjusted for age and sex. Change models were estimated by including Wave 3 outcomes as covariates in regression models predicting Wave 4 outcomes.

doi:10.1371/journal.pone.0101506.t004

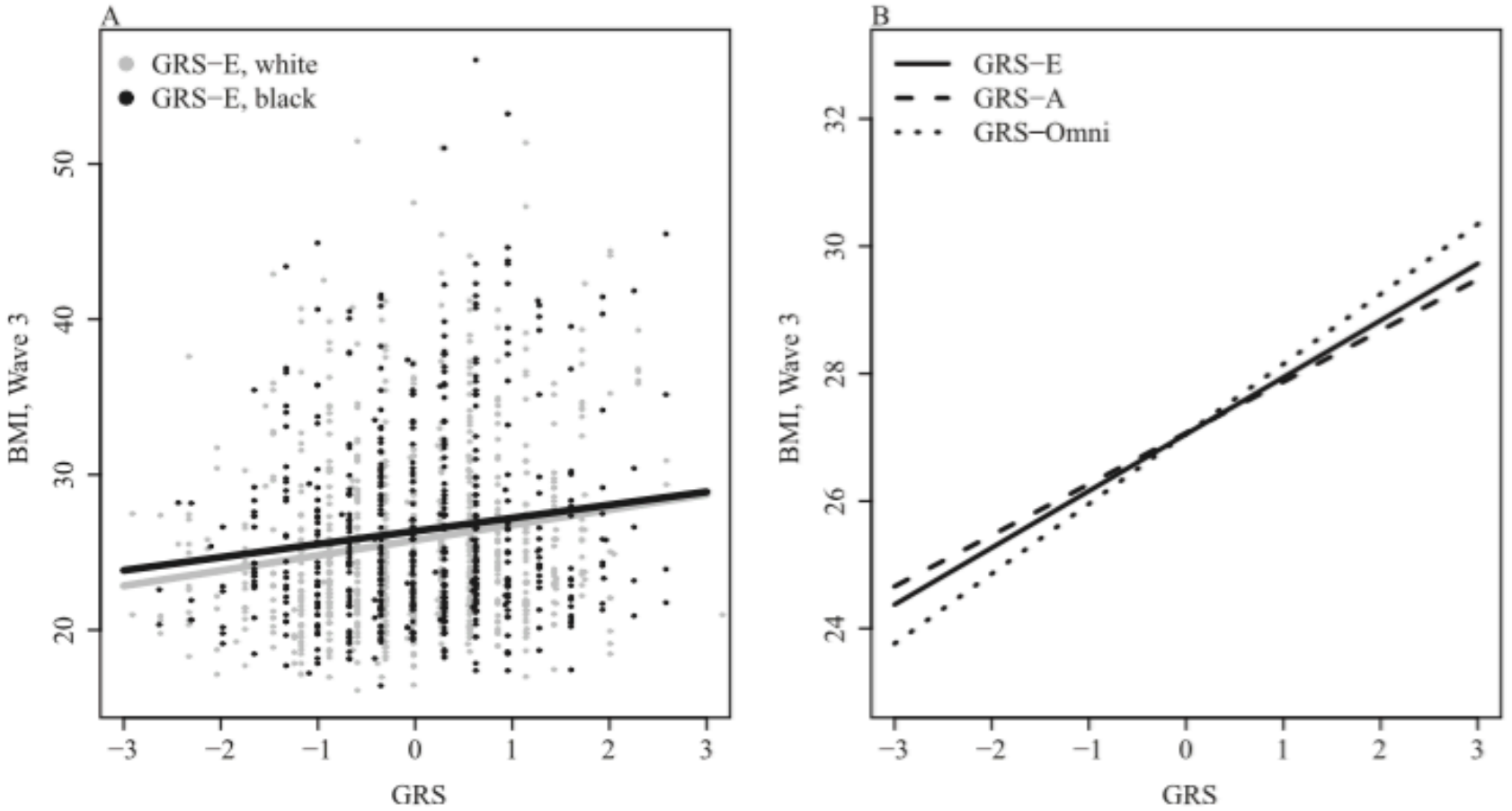


Figure 1. Comparison of GRS predictions. Panel A compares the predictive performance of GRS-E in both white and black samples of Add Health respondents based on a model where Wave 3 BMI is predicted by only GRS (separately in each racial group). Panel B focuses on predictions based on the three risk scores for only the black sample of respondents. The fitted lines are based on linear models controlling for age, sex, and one of the risk scores. The predictions assume an age of 21 and female.

doi:10.1371/journal.pone.0101596.g001

PGS in Whites/Blacks

- Performed very similarly
- PGS (European) was slightly weaker in blacks
- More inter-continental and global samples are necessary to fully appreciate how PGS impact health