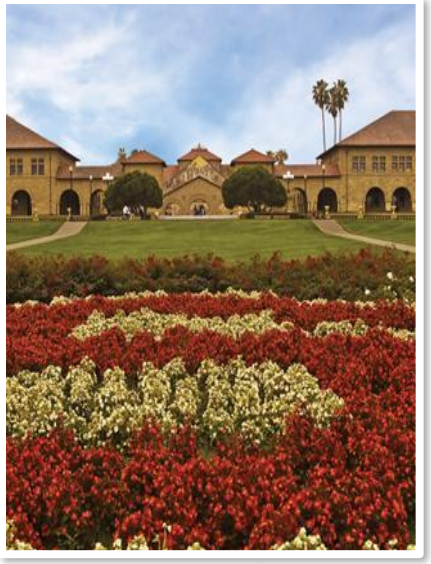
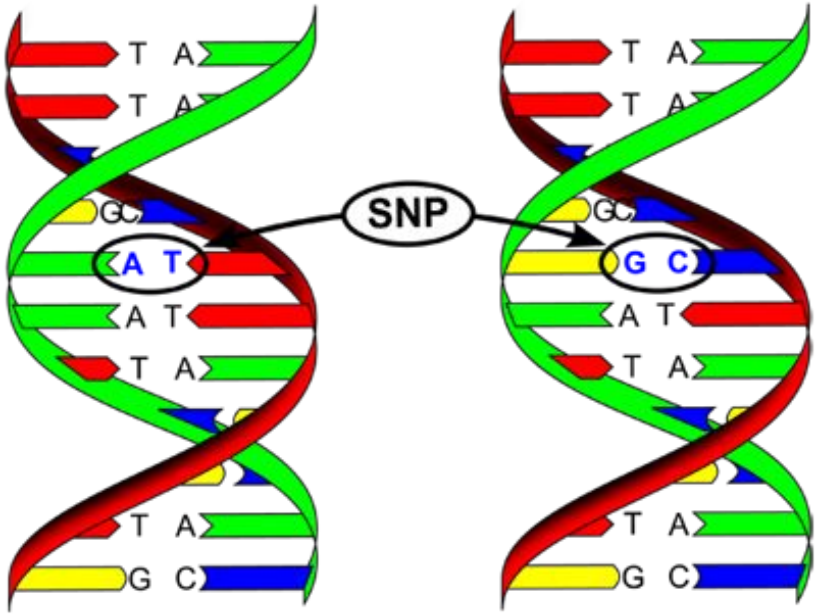
An introduction to polygenic scores

Oxford NCRM Summer School
Introduction to Using Molecular Genetic Data in the Social Sciences

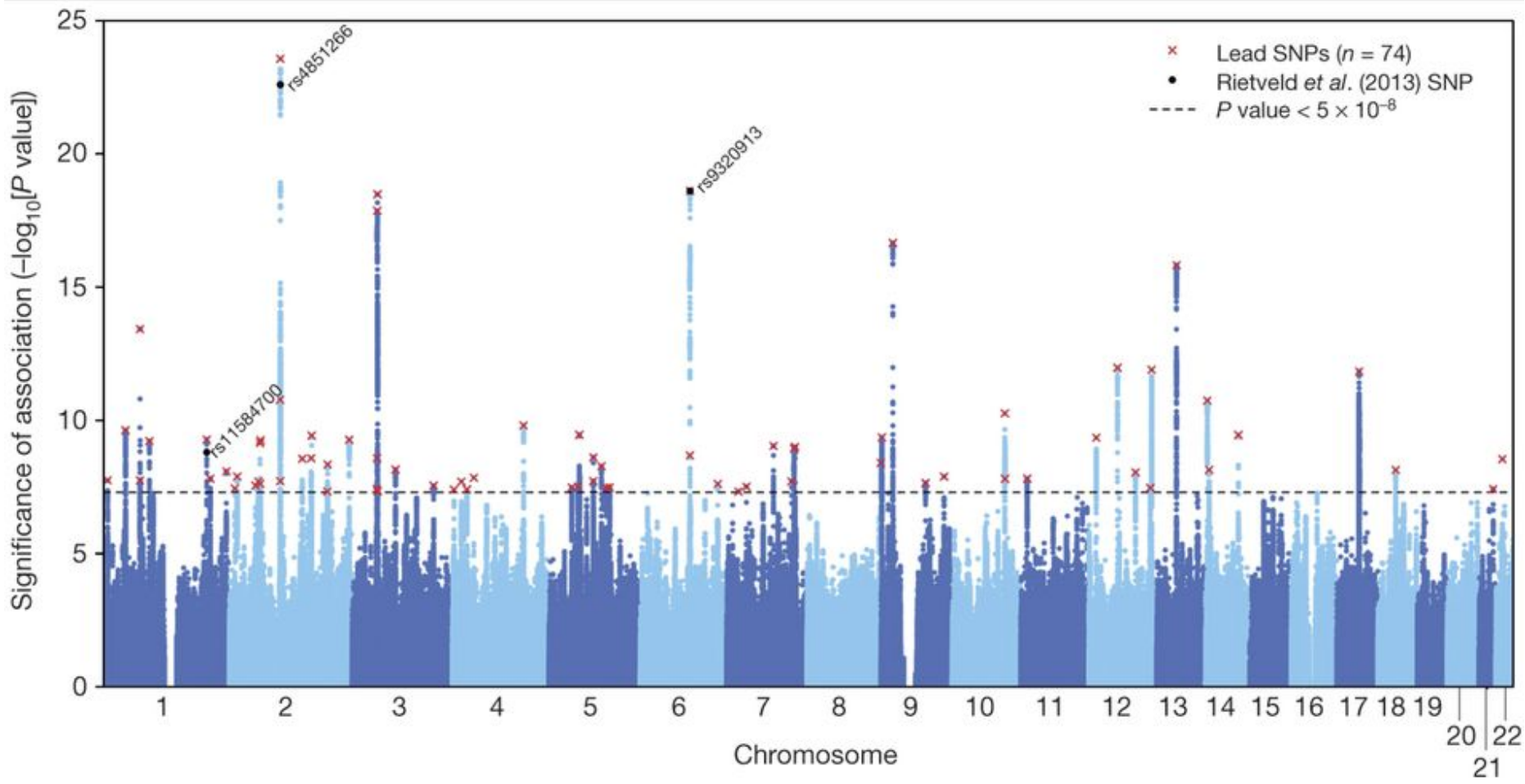


BEN DOMINGUE
bdomingue@stanford.edu

Single Nucleotide Polymorphisms (SNPs)



SNP Effects



Okbay et al. 2016

Individual SNPs are weak predictors

Weak prediction of individual genetic variants.

- Some exceptions (APOE)

Abstract

A genome-wide association study (GWAS) of educational attainment was conducted in a discovery sample of 101,069 individuals and a replication sample of 25,490. Three independent single-nucleotide polymorphisms (SNPs) are genome-wide significant (rs9320913, rs11584700, rs4851266), and all three replicate. Estimated effects sizes are small (coefficient of determination $R^2 \approx 0.02\%$), approximately 1 month of schooling per allele. A linear polygenic score from all measured SNPs accounts for ≈2% of the variance in both educational attainment and cognitive function. Genes in the region of the loci have previously been associated with health, cognitive, and central nervous system phenotypes, and bioinformatics analyses suggest the involvement of the anterior caudate nucleus. These findings provide promising candidate SNPs for follow-up work, and our effect size estimates can anchor power analyses in social-science genetics.

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

4th law of BG

1. All human behavioral traits are heritable. [That is, they are affected to some degree by genetic variation.]
2. The effect of being raised in the same family is smaller than the effect of genes.
3. A substantial portion of the variation in complex human behavioral traits is not accounted for by the effects of genes or families.

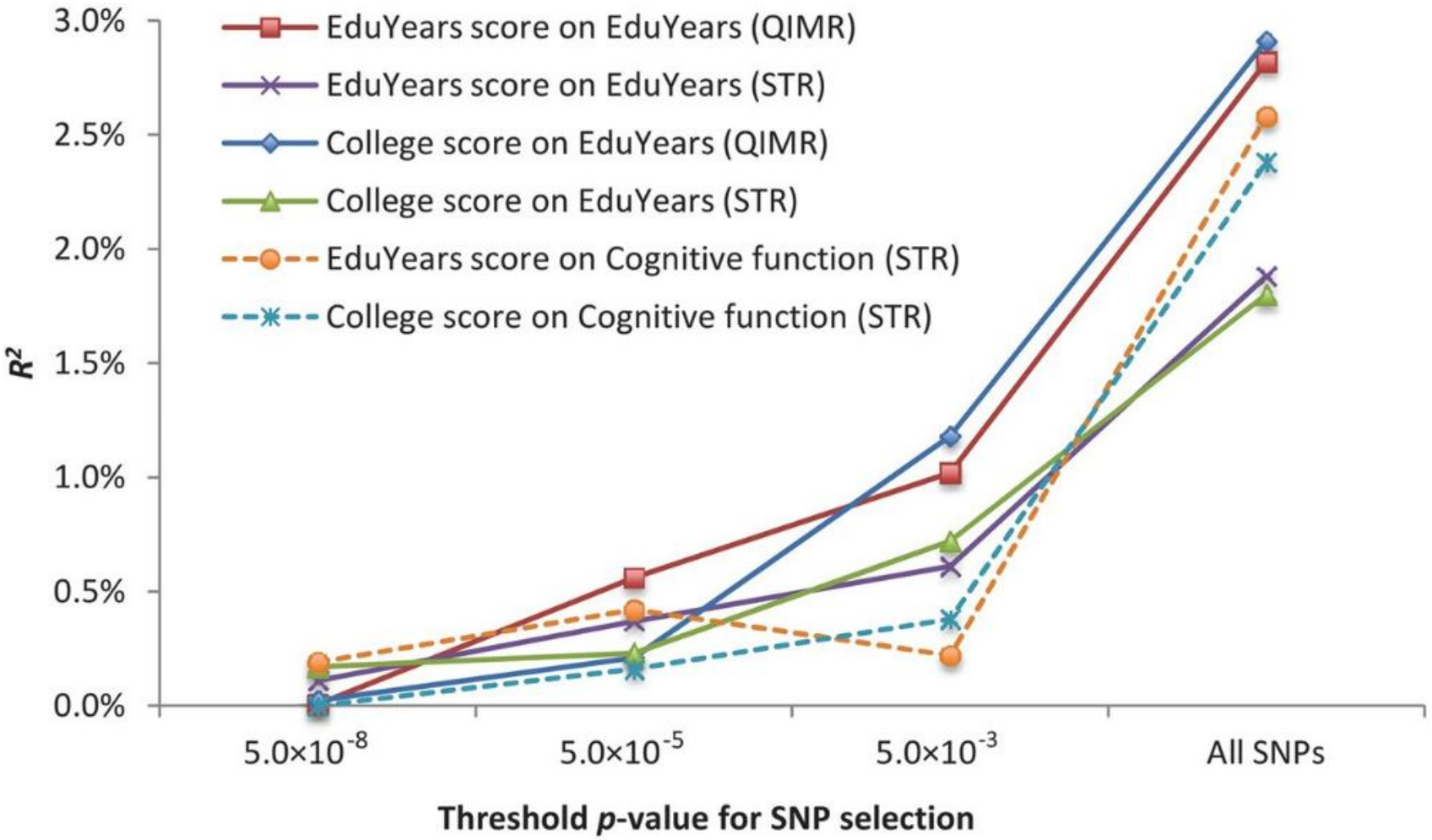
4th law of BG

1. All human behavioral traits are heritable. [That is, they are affected to some degree by genetic variation.]
2. The effect of being raised in the same family is smaller than the effect of genes.
3. A substantial portion of the variation in complex human behavioral traits is not accounted for by the effects of genes or families.
4. A typical human behavioral trait is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral variability.

The Fourth Law of Behavior Genetics

Christopher F. Chabris, James J. Lee, David Cesarini
Daniel J. Benjamin, David I. Laibson

One solution: aggregate information across SNPs



Rietveld et al., 2013, Science

Perhaps an analogy: Test scores

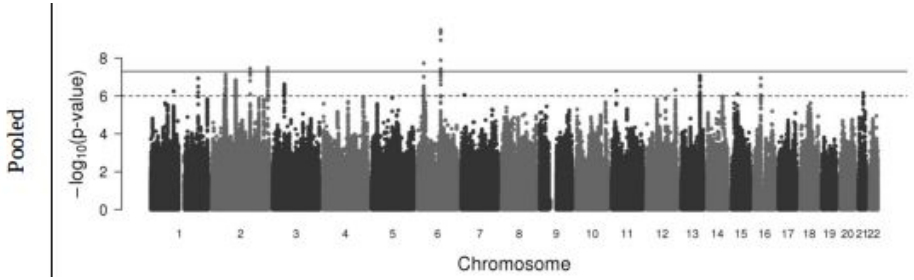
- Individual item responses are noisy indicators of student ability.
- To get a better measure of student ability, people typically aggregate over many item responses to derive a test score.
- Note: conventional wisdom in psychometrics is that all the individual item responses should be correlated

Perhaps an analogy: Test scores

- Individual item responses are noisy indicators of student ability.
- To get a better measure of student ability, people typically aggregate over many item responses to derive a test score.
- Note: conventional wisdom in psychometrics is that all the individual item responses should be correlated

- This analogy breaks down in one key sense: the SNPs won't necessarily be correlated.
- They might be (e.g., linkage), which in fact may cause problems.
  - Pruning/Clumping
- When they aren't, we are aggregating information across multiple biological systems.

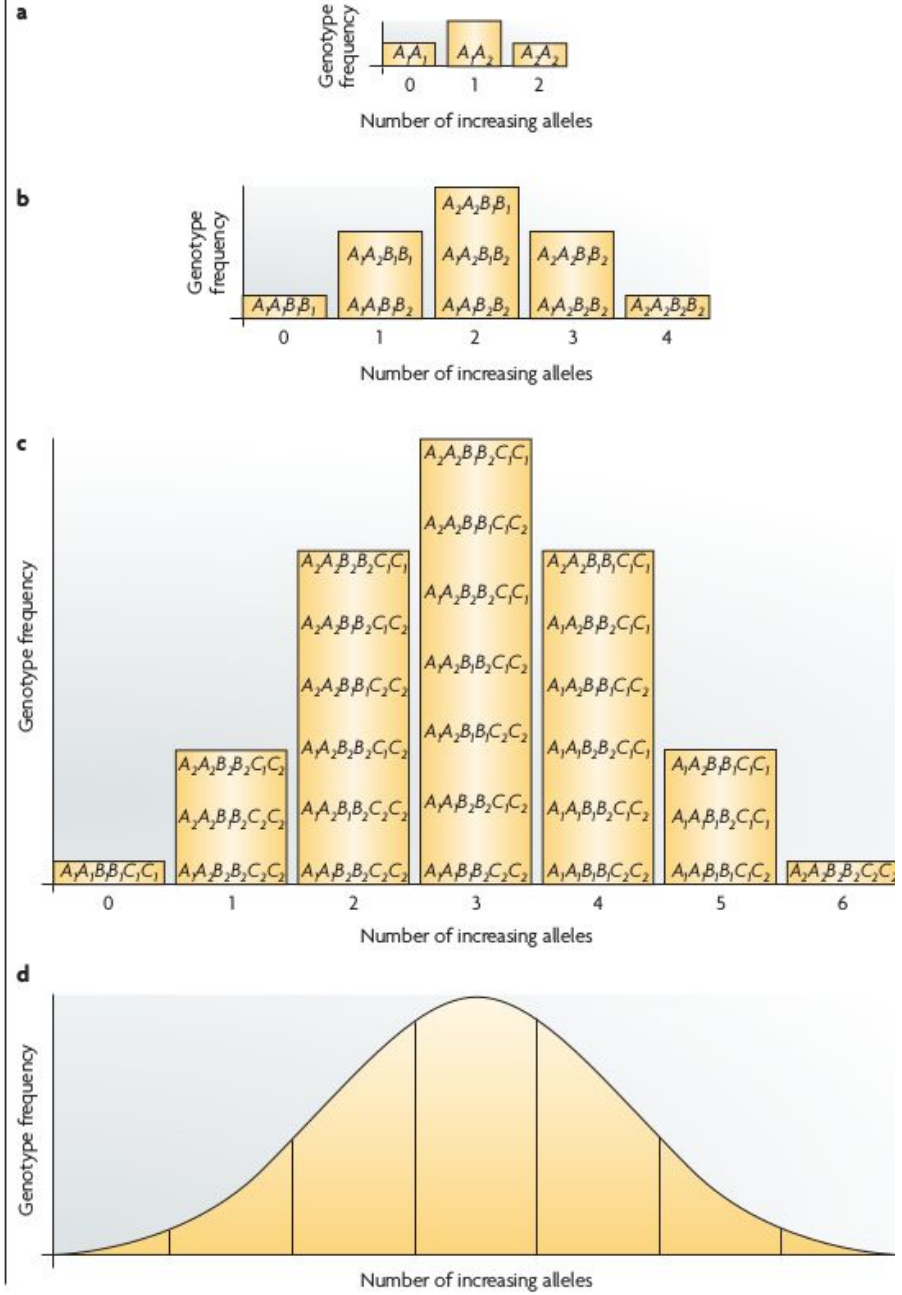
Score Construction



+

	SNP 1	SNP 2	⋯	SNP 1,000,000
P1	0	1	⋯	2
P2	1	0	⋯	0
P3	1	2	⋯	1
⋮	⋮	⋮	⋱	⋮
P1000	2	1	⋯	2

$1,000 \times 1,000,000$ matrix; each cell $\in \{0, 1, 2\}$.



Polygenic scores are weighted sums

(person-i, SNP-j)

$$\mathrm{PGS}_i = \sum_j \beta_j \mathrm{Genotype}_{ij}$$

Key Questions

$$\mathrm{PGS}_i = \sum_j \beta_j \mathrm{Genotype}_{ij}$$

1. Which j's? (Which SNPs to include?)
   a. Strand Ambiguity
   b. P-value thresholds
   c. Independence of SNPs
   d. Genotyped or imputed?
2. What are beta-weights?
3. Which i's? (For whom should they apply?)

Mechanical Problem: Strand ambiguity

Identifying the allele that predisposes towards an outcome can be trickier than it would seem at first blush.

GWAS Results

MarkerName	Effect_Allele	Other_Allele	EAF	Beta	SE	Pvalue
rs4075116	T	C	0.78	0.001	0.005	0.8374
rs3934834	T	C	0.13	-0.007	0.005	0.1763
rs3766193	C	G	0.4	-0.006	0.004	0.1371
rs3766192	T	C	0.61	0.007	0.004	0.0978
rs3766191	T	C	0.11	-0.008	0.005	0.133
rs9442371	T	C	0.62	0.006	0.004	0.1567
rs9442372	A	G	0.37	-0.005	0.004	0.2273
rs10907177	A	G	0.87	0.008	0.005	0.1198
rs3737728	A	G	0.26	-0.003	0.004	0.4356

GWAS Results

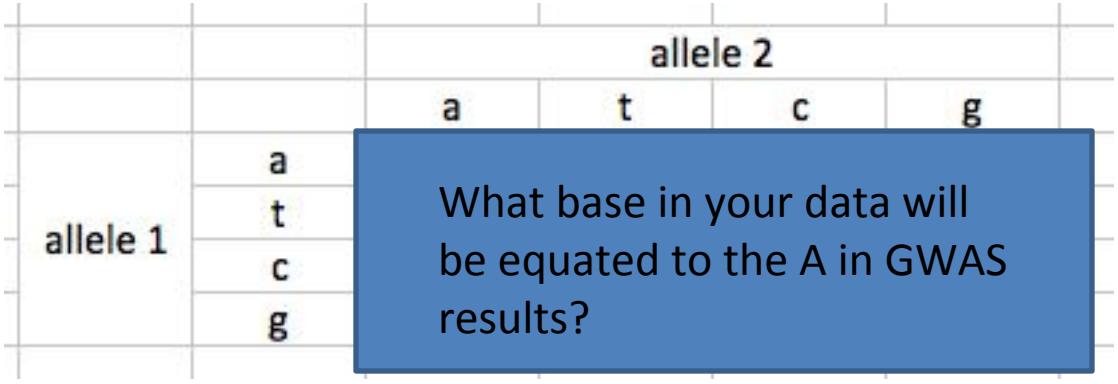
MarkerName	Effect_Allele	Other_Allele	EAF	Beta	SE	Pvalue
rs4075116	T	C	0.78	0.001	0.005	0.8374
rs3934834	T	C	0.13	-0.007	0.005	0.1763
rs3766193	C	G	0.4	-0.006	0.004	0.1371
rs3766192	T	C	0.61	0.007	0.004	0.0978
rs3766191	T	C	0.11	-0.008	0.005	0.133
rs9442371	T	C	0.62	0.006	0.004	0.1567
rs9442372	A	G	0.37	-0.005	0.004	0.2273
rs10907177	A	G	0.87	0.008	0.005	0.1198
rs3737728	A	G	0.26	-0.003	0.004	0.4356

Empirical Data

1 rs3934834 0 1005806 A G

Strand issues

- Paired bases: A-T, C-G
- Fwd strand A/C implies reverse strand T/G
- Suppose GWAS identifies A as risk allele. [note: they may not give you other allele.]
- Your data:



Strand issues

- Paired bases: A-T, C-G
- Fwd strand A/C implies reverse strand T/G
- Suppose GWAS identifies A as risk allele. [note: they may not give you other allele.]
- Your data:

		allele 2			
		a	t	c	g
allele 1	a		?	a	a
	t	?		t	t
	c	a	t		?*
	g	a	t	?*	

Choice of p-value

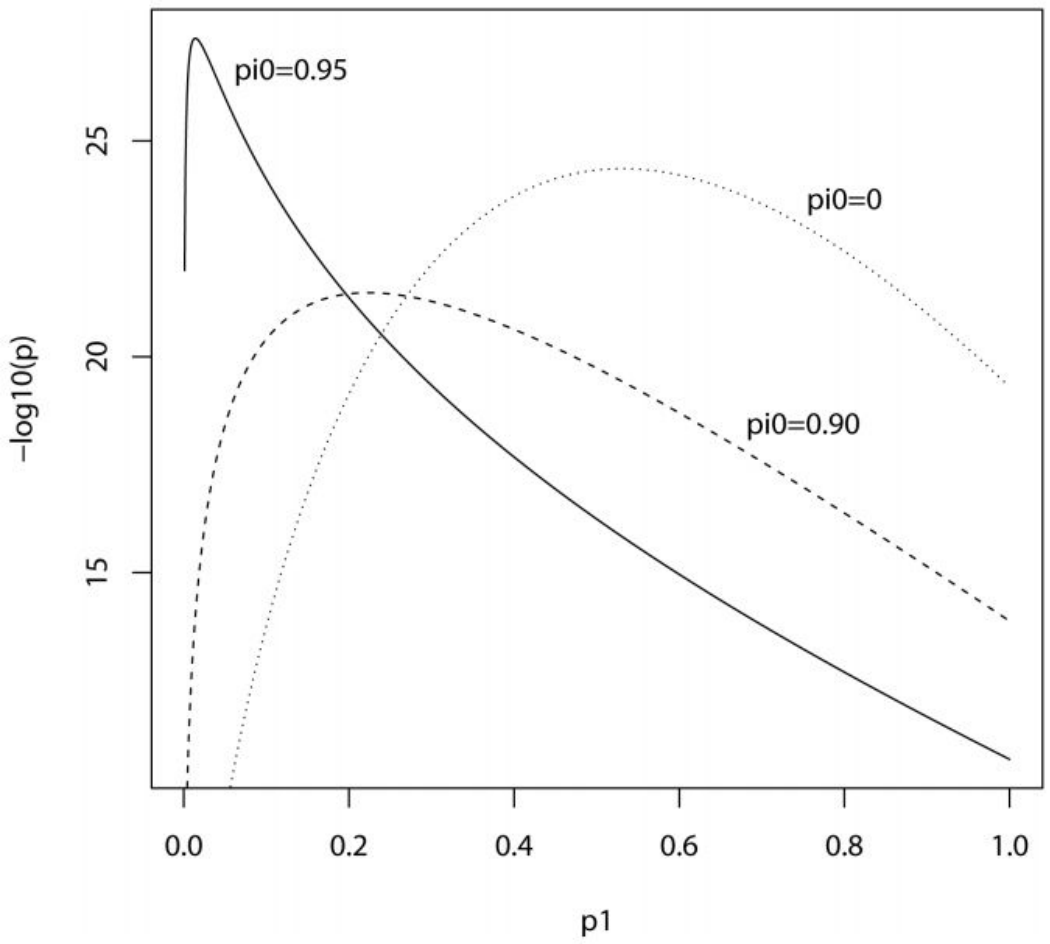


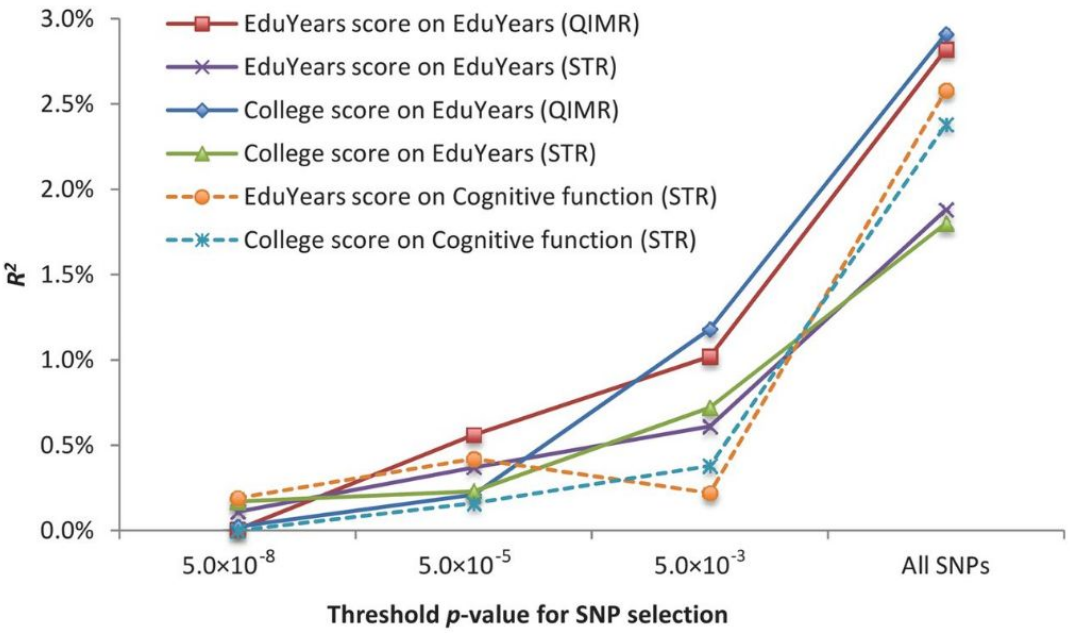
Figure 2. Expected $-\log_{10}(P)$ of allele score estimate as a function of *P*-value threshold for selecting markers into the polygenic score. Training sample, 3322 cases and 3587 controls; replication sample, 2687 cases and 2656 controls. Marker panel of 74062 independent SNPs. Variance explained by markers, 28.7%. pi0, proportion of markers with no effect on disease.
doi:10.1371/journal.pgen.1003348.g002

- There is a theoretical optimum
- It depends on knowledge of trait's genetic architecture we don't generally have.

P-value threshold

General rule:
More genetic information yields better prediction.

- Most research use lenient p-value threshold.
- Low weights for low p-values



Two points re p-values

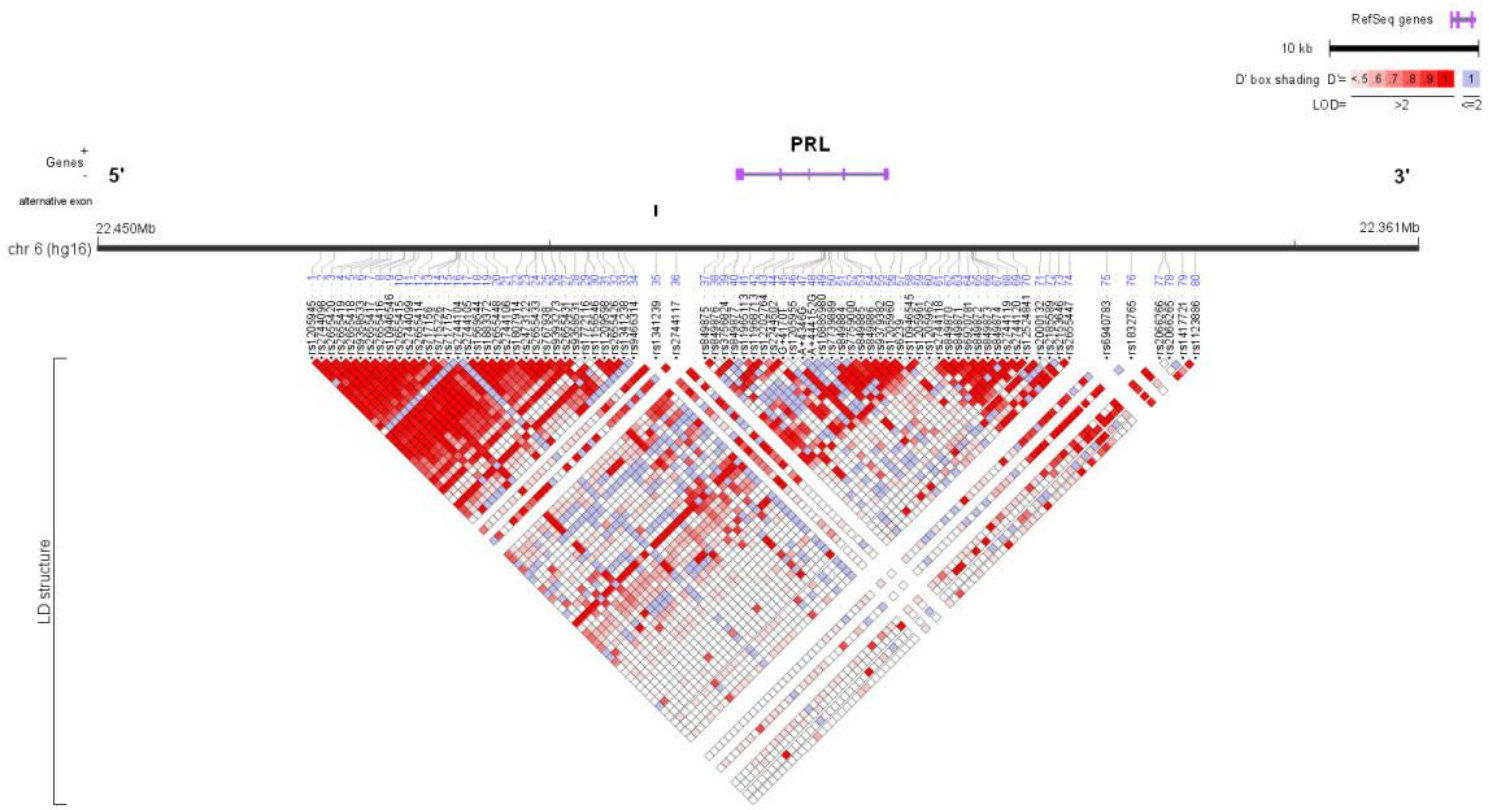
Top hits scores

- Scores constructed using only genome-wide significant hits.
- Useful when you might be interested in understanding the relevant biology.
  - Mendelian randomization?

Beware overfitting!

- Optimization of p-value threshold should not take place in your analytic sample.

Linkage


RefSeq genes
10 kb
D' box shading
LOD=
PRL
Genes
5'
3'
alternative exon
22.450Mb
22.361Mb
chr 6 (hg16)
LD structure

Why is linkage relevant here?

- What do we want out of PGS?
  - Create a proxy that captures biological predisposition or risk.
- If too many SNPs in linkage with a causal locus are included, we might overemphasize that locus in score.
  - Chip design
- Tricky problem
  - Removing non-causal (but associated) SNPs necessitates knowledge of causal SNPs which we don't have.

Many options

- Use all SNPs [aka do nothing]
- Pruning
  - Traditional approach to identifying independent markers involved random selection.
- Clumping
  - Select variants in LD blocks that are most highly associated with outcome.

LDpred

"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."

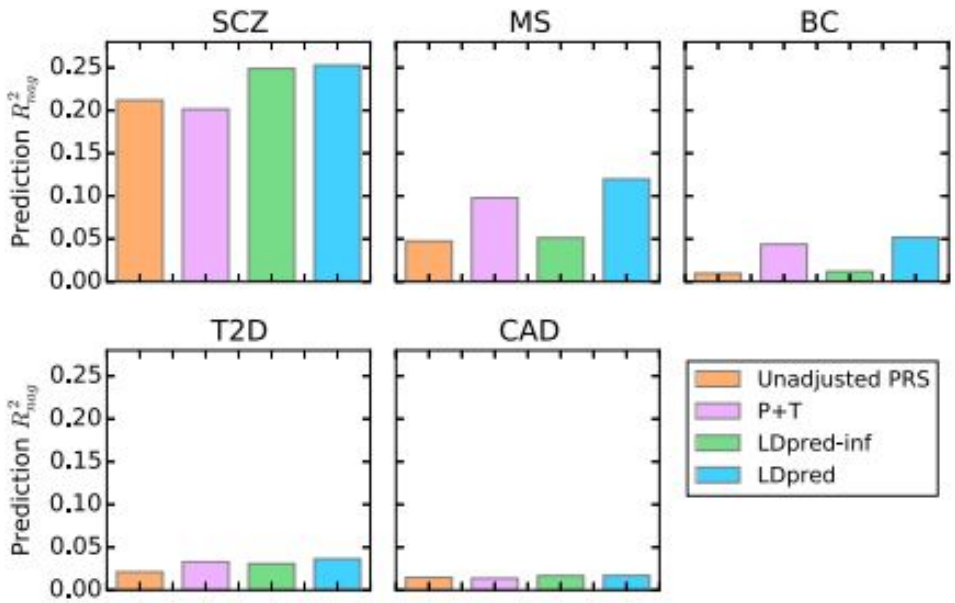


Figure 4. Comparison of Methods Training on Large GWAS Summary Statistics for Five Different Diseases

The prediction accuracy is shown for five different diseases: schizophrenia (SCZ), multiple sclerosis (MS), breast cancer (BC), type 2 diabetes (T2D), and coronary artery disease (CAD). The risk scores were trained with large GWAS summary-statistics datasets and used for predicting disease risk in independent validation datasets. The Nagelkerke prediction R^2 is shown on the y axis (see Table S5 for other metrics). Compared to LD pruning + thresholding (P+T), LDpred improved the prediction R^2 by 11%–25%. SCZ results are shown for the SCZ-MGS validation cohort used in recent studies,[9,12,14] but LDpred also produced a large improvement for the independent SCZ-ISC validation cohort (Table S5).

Article

Switch to Standard View

Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores

Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong-Hee Won,
⊞ Schizophrenia Working Group of the Psychiatric Genomics Consortium,
⊞ Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study, Sekar Kathiresan, Michele Pato, Carlos Pato, Rulla Tamimi, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Gillian Belbin, Eimear E. Kenny, Mikkel H. Schierup, Philip De Jager, Nikolaos A. Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Benjamin Neale, Michael Goddard, Peter M. Visscher, Peter Kraft, Nick Patterson, Alkes L. Price

Imputed or Genotyped SNPs

- Imputation quality may vary as a function of ancestry
- Probably worth knowing whether key results are sensitive to this choice re score construction.

Weights

- Each SNP will be weighted by the magnitude of its effect.
  - For continuous traits, the estimated coefficient.
  - For dichotomous traits, log of the OR.
- If you're doing a top-hits score, then may not need to weight.

Weights

- Each SNP will be weighted by the magnitude of its effect.
  - For continuous traits, the estimated coefficient.
  - For dichotomous traits, log of the OR.
- If you're doing a top-hits score, then may not need to weight.

But where do we get estimated effect?

Very Important

- You probably need big samples to get reasonable weights
- Use consortia-led GWAS rather than doing yourself.

Table 6. Numbers of subjects (in 1000s, rounded up) required to attain a specified correlation with a normal trait using a panel of 1,000,000 markers that explains the full heritability.

	Correlation	$\pi_0 = 0.999$	$\pi_0 = 0.99$	$\pi_0 = 0.90$	$\pi_0 = 0.75$	$\pi_0 = 0$
$h^2 = 0.8$ (Max = 0.894)	0.8046 = 0.9*Max	31 (0.00007)	227 (0.0004)	1601 (0.007)	3231 (0.03)	5329 (1)
	0.8493 = 0.95*Max	55 (0.00007)	411 (0.0003)	3004 (0.005)	6250 (0.02)	11571 (1)
	0.88506 = 0.99*Max	213 (0.00007)	1546 (0.0002)	12171 (0.003)	26724 (0.01)	61565 (1)
$h^2 = 0.4$ (Max = 0.632)	0.5688 = 0.9*Max	61 (0.00007)	453 (0.0004)	3201 (0.007)	6461 (0.03)	10658 (1)
	0.6004 = 0.95*Max	109 (0.00007)	821 (0.0003)	6007 (0.005)	12500 (0.02)	23141 (1)
	0.62568 = 0.99*Max	426 (0.00007)	3092 (0.0002)	24341 (0.003)	53448 (0.01)	123128 (1)

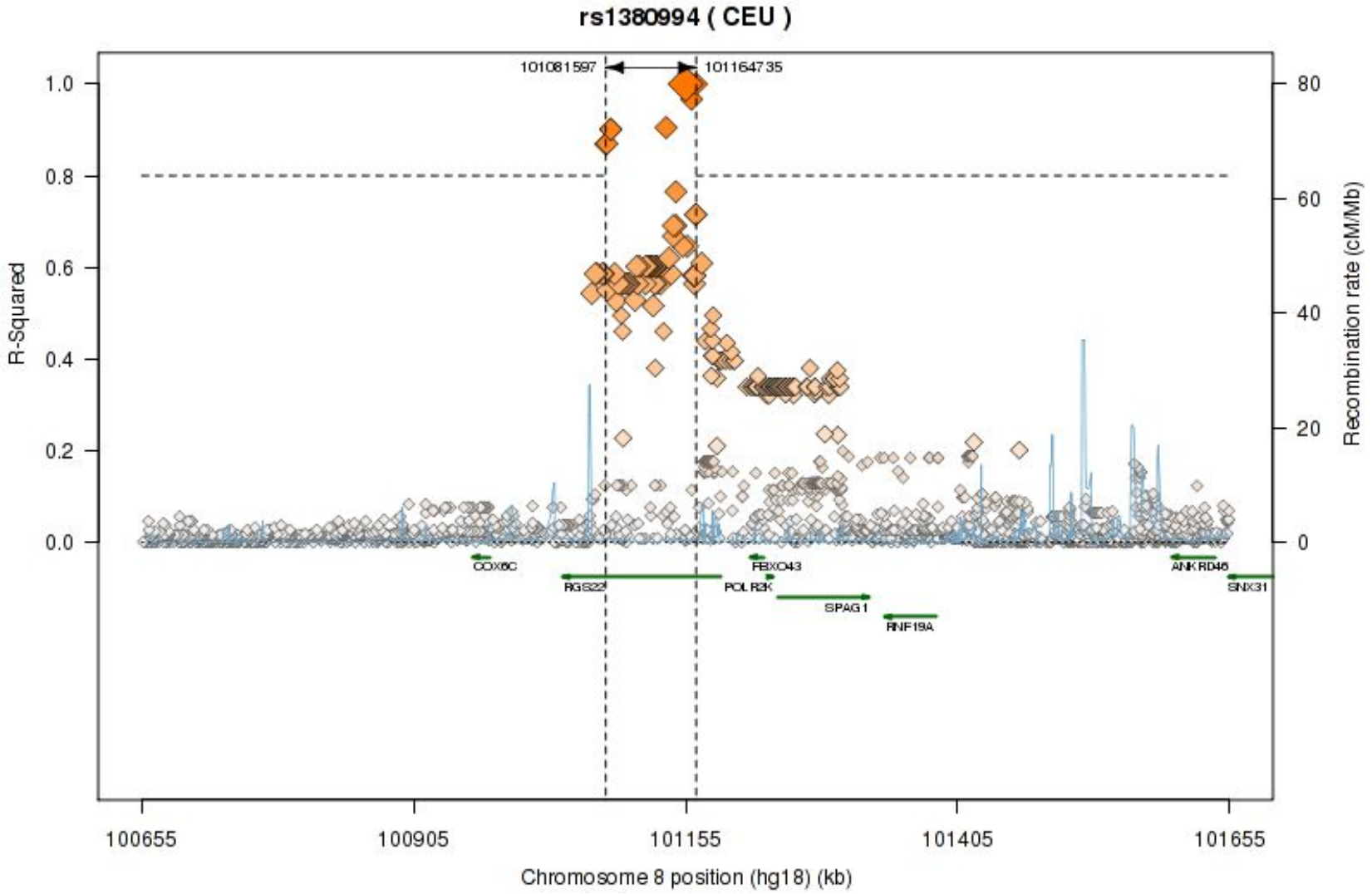
π_0, proportion of SNPs having no effect on the trait. Max, maximum correlation achievable given the genetic variance of the marker panel. In parentheses, *P*-value threshold that maximises the correlation.

doi:10.1371/journal.pgen.1003348.t006

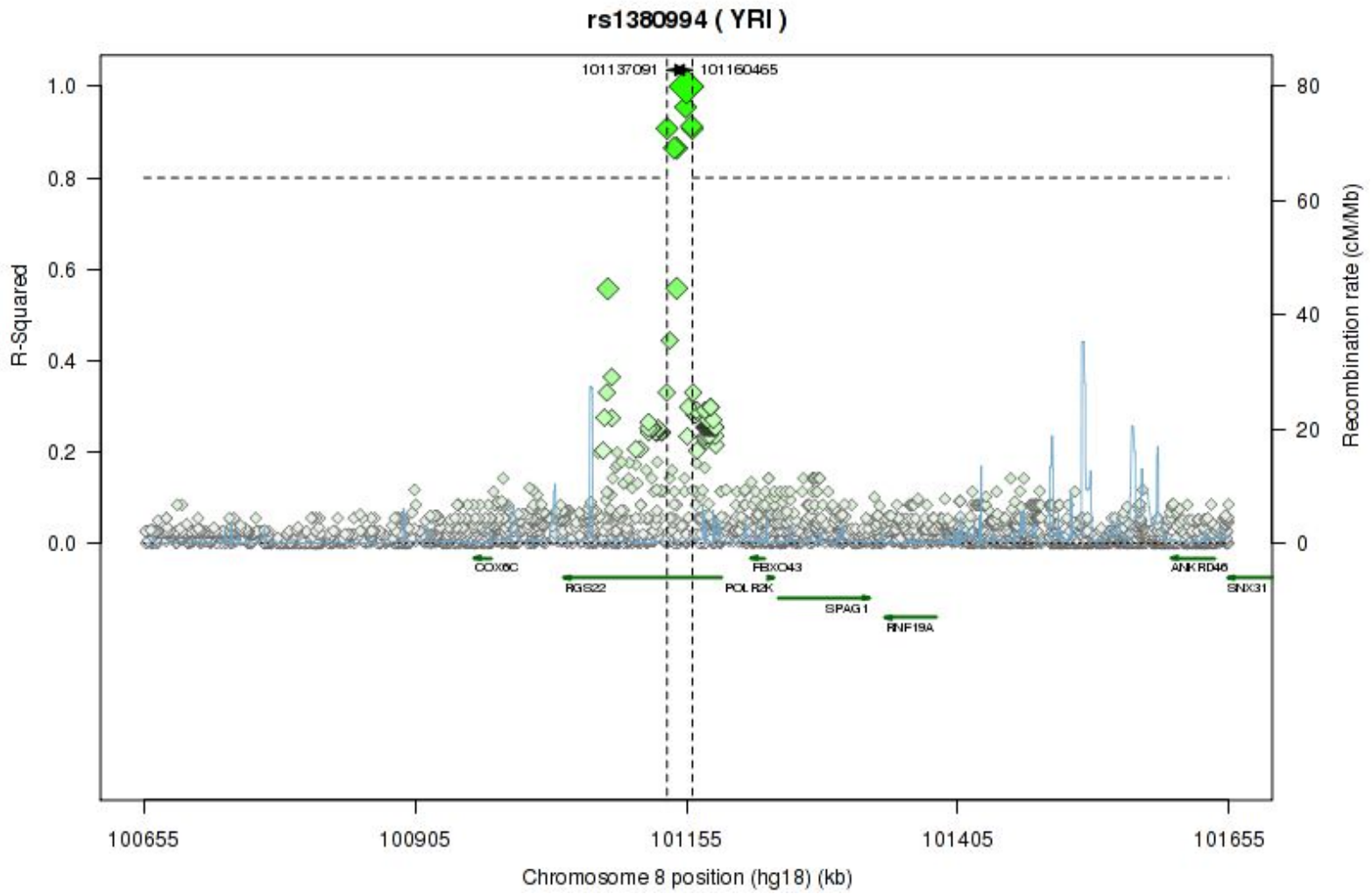
Sensitivity to ancestral background

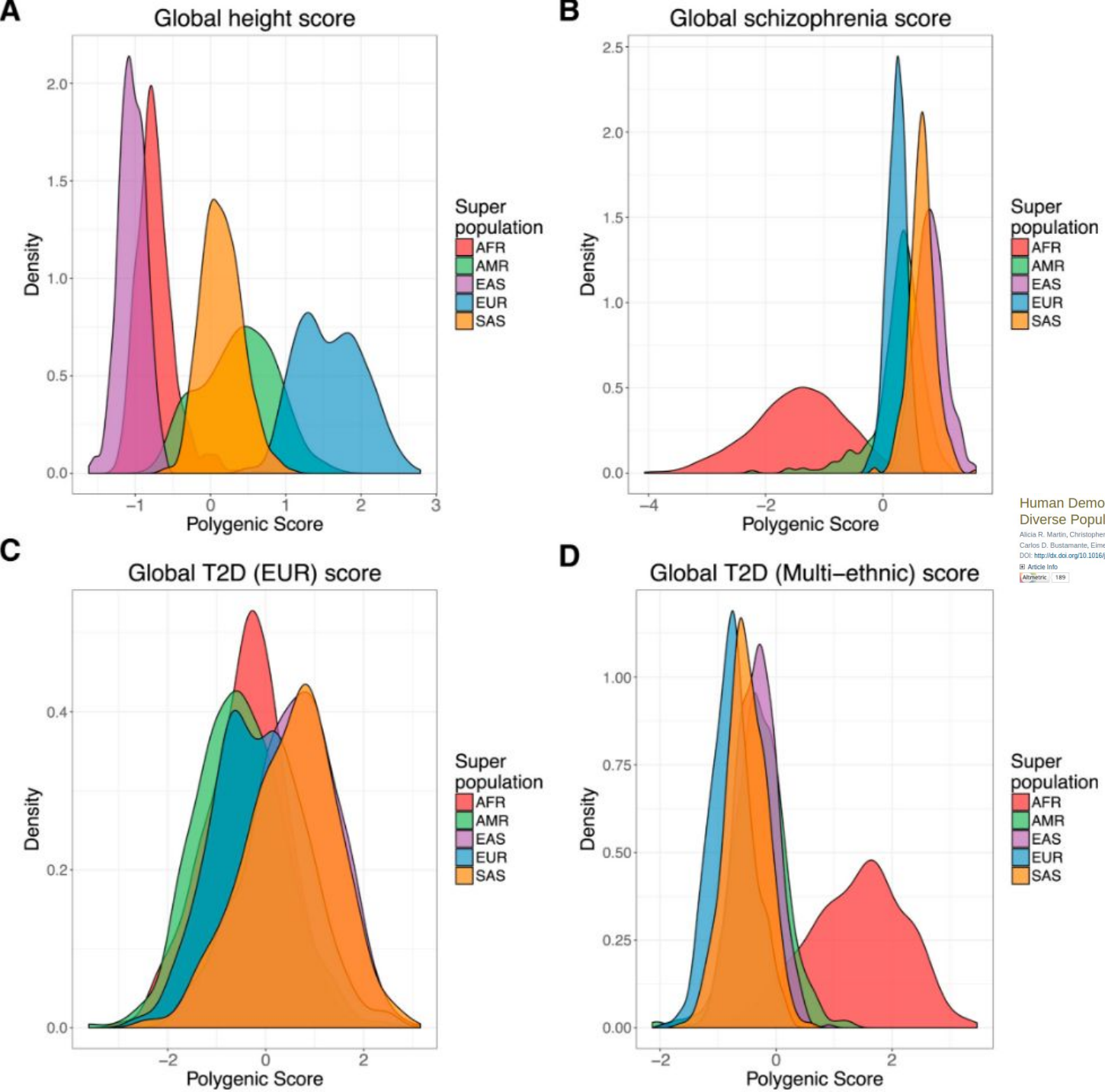
Vast majority of GWAS results are based on those of euro ancestry

Tag SNPs


rs1380994 (CEU)
101081597
101164735
1.0
0.8
0.6
0.4
0.2
0.0
R-Squared
80
60
40
20
0
Recombination rate (cM/Mb)
COX6C
RGS22
FBXO43
POLR2K
SPAG1
RNF19A
ANKRD45
SNX31
100655
100905
101155
101405
101655
Chromosome 8 position (hg18) (kb)

Tag SNPs


rs1380994 (YRI)
101137091
101160465
1.0
0.8
0.6
0.4
0.2
0.0
R-Squared
80
60
40
20
0
Recombination rate (cM/Mb)
COX6C
RGS22
FBXO43
POLR2K
SPAG1
RNF19A
ANKRD46
SNX31
100655
100905
101155
101405
101655
Chromosome 8 position (hg18) (kb)



Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations

Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, Elmear E. Kenny

DOI: http://dx.doi.org/10.1016/j.ajhg.2017.03.004 | CrossMark

Article Info

Altmetric 189

Simulation

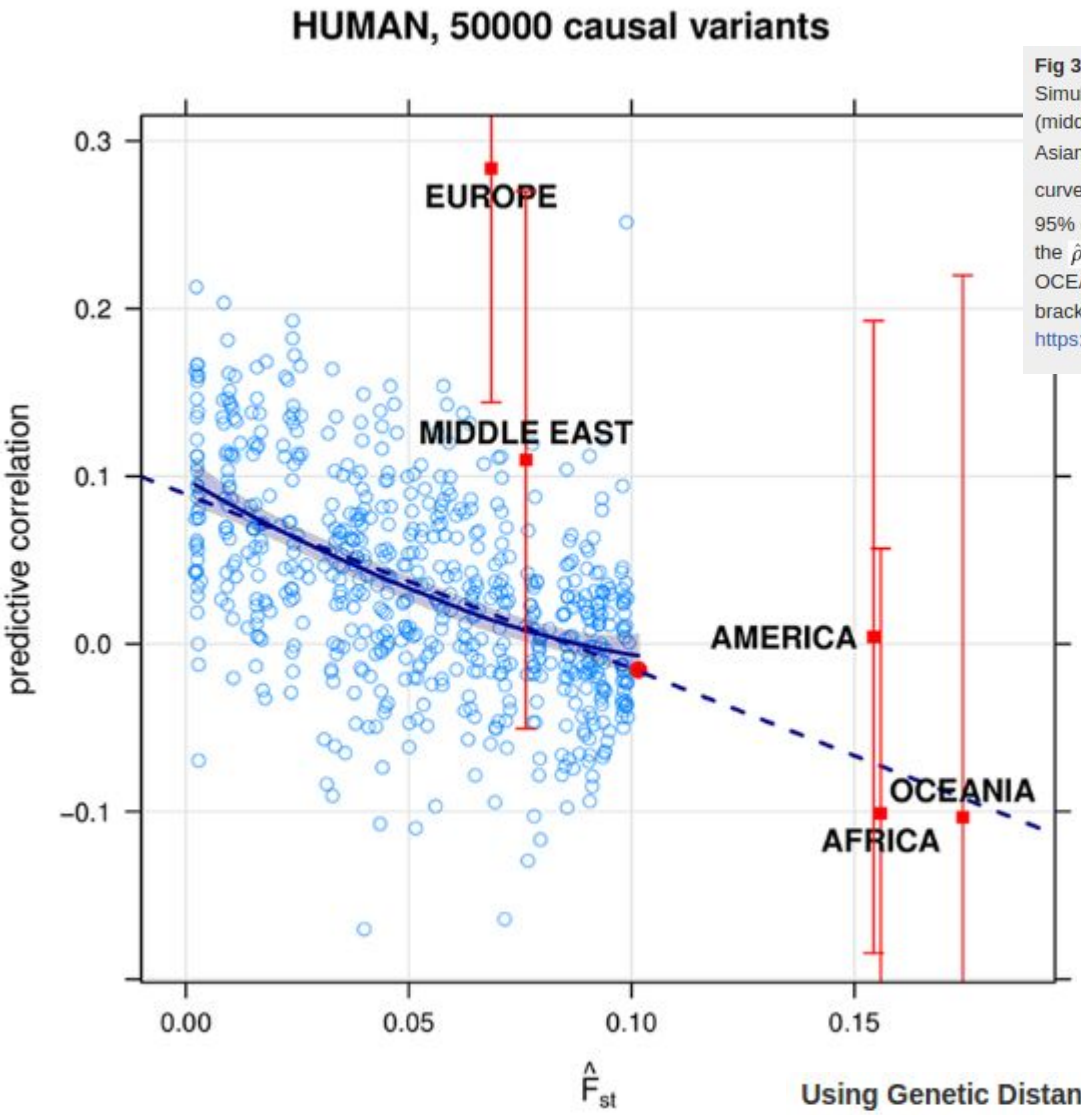


Fig 3. Simulation of quantitative traits from the HUMAN data.
Simulation of quantitative traits with 5 (top left), 20 (top right), 100 (middle left), 2000 (middle right), 10000 (bottom left) and 50000 (bottom right) causal variants from the Asian individuals in the HUMAN data. The blue circles are the $\hat{\rho}_{\mathrm{D}}^{(m)}$ used to build the curve, and the red point is $\hat{\rho}_{\mathrm{D}}^{(0)}$. The blue line is the mean decay trend, with a shaded 95% confidence interval, and the dashed blue line is the linear interpolation provided by the $\hat{\rho}_{\mathrm{L}}$. The red squares labelled EUROPE, MIDDLE EAST, AMERICA, AFRICA and OCEANIA correspond to the $\hat{\rho}_{\mathrm{P}}$ for the individuals from those continents, and the red brackets are the respective 95% confidence intervals.
https://doi.org/10.1371/journal.pgen.1006288.g003

Practical Implications of PGS construction

Table 1. Estimation approaches for polygenic score creation for each trait

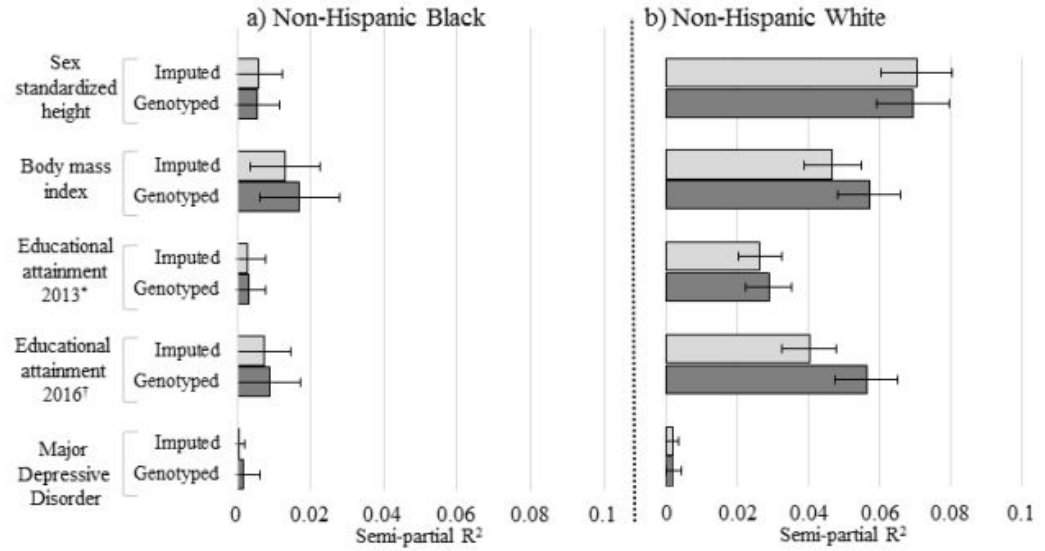
Accounting for linkage disequilibrium	Imputed vs genotyped		Race/ Ethnicity*		P-value threshold for inclusion†		Linkage disequilibrium R^2		Genomic region size		Sliding window size		Total
None	2	x	2	x	6							=	24
Clumping‡§	2	x	2	x	6	x	3	x	3			=	216
Pruning¶#‖	2	x	2	x	6	x	2	x	2	x	2	=	192
Sub Total													432**
Top SNPs			2										2
Total													434

*non-Hispanic Black and non-Hispanic White
†GWAS meta-analysis p-value thresholds for SNP inclusion (pT)
‡LD R^2 threshold for clumping: 0.1, 0.5, 0.8
§Genomic distance for clumping: 100, 250, 500 kilobases (kb)
¶Pairwise LD R^2 threshold for pruning: 0.5, 0.8
#Genomic window size for pruning: 25, 50kb
‖Increment for sliding window for pruning: 2, 5 SNPs
**Since Major Depressive Disorder (MDD) did not have any published top SNPs, the total for MDD is 432

Heterogeneity in polygenic scores for common human traits

Erin B Ware MPH PhD[*1], Lauren L Schmitz PhD[1], Jessica Faul MPH PhD[1], Arianna Gard MA[2], Colter Mitchell PhD[1], Jennifer A Smith MPH PhD[1,3], Wei Zhao PhD[3], David Weir PhD[1], Sharon LR Kardia PhD[3]

Figure 1. Comparison of semi-partial R^2 for polygenic scores created from directly genotyped and imputed SNPs, by ethnicity and trait



No LD trimming, pT = 1.0. BMI: body mass index, EA: educational attainment, MDD: Major Depressive Disorder. Semi-partial R^2 and 95% confidence intervals from bootstrapped models with 1,000 repetitions. Models include covariates reflective of the meta-analysis GWAS from which the SNP weights came.

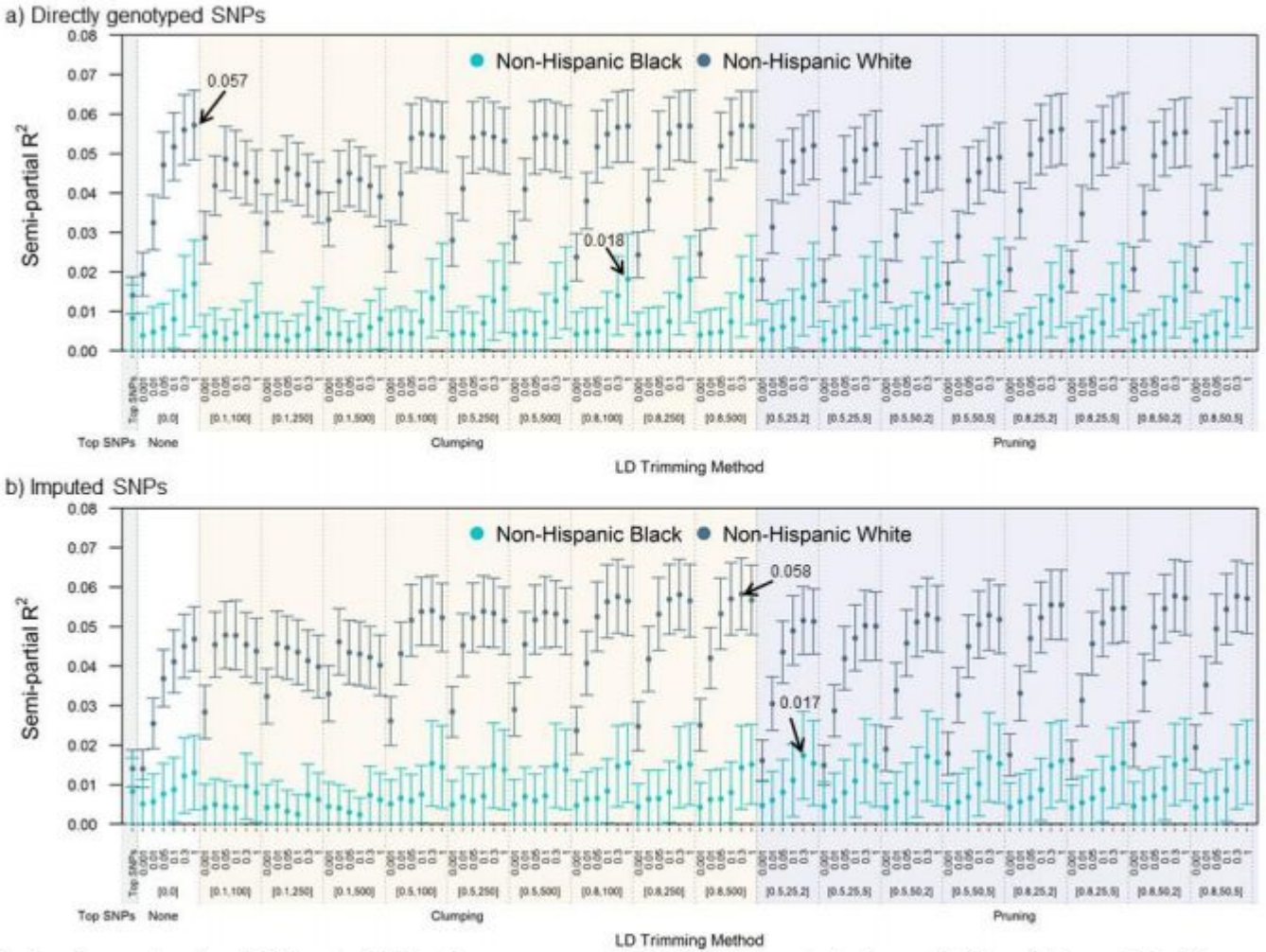
- HRS is reporting scores based on:
  - No p-value
  - No LD
  - genotyped
- Good for replication
- How sensitive are scores to this set of choices?

Table 2. Comparing polygenic scores from directly genotyped SNPs to imputed SNPs across estimation approaches

		Linkage disequilibrium (LD) trimming method		
		None (m*=6)	Clumping (m*=54)	Pruning (m*=48)
		†$R^2_{genotyped} > R^2_{imputed}$	†$R^2_{genotyped} > R^2_{imputed}$	†$R^2_{genotyped} > R^2_{imputed}$
Race/ethnicity	**Trait**	n‡ (%)	n‡ (%)	n‡ (%)
Non-Hispanic Black	Body mass index	2 (33.3)	15 (27.8)	8 (16.7)
	Educational attainment 2013†	5 (83.3)	18 (33.3)	11 (22.9)
	Educational attainment 2016§	4 (66.7)	47 (87)	46 (95.8)
	Sex standardized and adjusted height	3 (50)	28 (51.9)	9 (18.8)
	Major Depressive Disorder	5 (83.3)	45 (83.3)	30 (62.5)
Non-Hispanic White	Body mass index	6 (100)	27 (50)	24 (50)
	Educational attainment 2013§	5 (83.3)	24 (44.4)	11 (22.9)
	Educational attainment 2016¶	6 (100)#	48 (88.9)	48 (100)
	Sex standardized and adjusted height	6 (100)	35 (64.8)	23 (47.9)
	Major Depressive Disorder	4 (66.7)	30 (55.6)	16 (33.3)

- PGS based on genotyped SNPs generally more predictive than those based on imputed SNPs
  - Irrespective of how LD is handled

Figure 3. Semi-partial R^2 with 95% standard error bars for regression of body mass index onto the polygenic score



a) directly genotyped and **b)** imputed SNPs. Arrows represent the highest amount of observed trait-variation explained for each race/ethnicity. Semi-partial R^2 and 95% standard errors were created from 1,000 bootstrapped repetitions, adjusting for covariates that were used in the GWAS meta-analysis from which the SNP weights came.
BMI: body mass index, LD: Linkage Disequilibrium. None: no LD trimming; LD Pruning [LD R^2, window size (kb), sliding increment (number of SNPs)], LD Clumping [LD R^2, region size (kb)]

- Clumping > Pruning
- Patterns related to p-values consistent across genotyped & imputed.
- Some sensitivity to mechanics of clumping

Table 3. Recommendations for researchers and GWAS consortia regarding polygenic score use

FOR RESEARCHERS ESTIMATING POLYGENIC SCORES
• Use effect estimates from large, replicated genome-wide association meta-analyses that do not include the study of interest
• Provide complete documentation for the source of SNP weights, p-value threshold(s), LD trimming strategy, and a list of SNPs that are in the PGS
• Consider using whole-genome scores instead of "top SNPs" scores
• Calculate multiple scores and include sensitivity analysis, and examine the correlation between scores
• If distributing PGSs to the wider scientific community, use no LD trimming so that colleagues can verify or replicate the PGS (i.e. both pruning and clumping are stochastic)
FOR GWAS CONSORTIA:
• Make available whole genome results for discovery meta-analyses as opposed to selected or replicated SNPs only

• Provide more detailed reporting of the meta-analysis results (both discovery and replication) including chromosome, position, both alleles, beta/OR, standard error, and full sample size per SNP to:
  • Allow comparison between builds
  • Allow for the development and use of multiple software packages

• Offer participating studies meta-analysis results with their study removed to enable them to estimate independent PGSs for their studies
• Clearly report imputation panel and build in the documentation for downloadable results
• Call for large scale GWAS efforts in individuals of non-European ancestry

Table 3. Recommendations for researchers and GWAS consortia regarding polygenic score use

FOR RESEARCHERS ESTIMATING POLYGENIC SCORES
• Use effect estimates from large, replicated genome-wide association meta-analyses that do not include the study of interest
• Provide complete documentation for the source of SNP weights, p-value threshold(s), LD trimming strategy, and a list of SNPs that are in the PGS
• Consider using whole-genome scores instead of "top SNPs" scores
• Calculate multiple scores and include sensitivity analysis, and examine the correlation between scores
• If distributing PGSs to the wider scientific community, use no LD trimming so that colleagues can verify or replicate the PGS (i.e. both pruning and clumping are stochastic)
FOR GWAS CONSORTIA:
• Make available whole genome results for discovery meta-analyses as opposed to selected or replicated SNPs only
• Provide more detailed reporting of the meta-analysis results (both discovery and replication) including chromosome, position, both alleles, beta/OR, standard error, and full sample size per SNP to: • Allow comparison between builds • Allow for the development and use of multiple software packages
• Offer participating studies meta-analysis results with their study removed to enable them to estimate independent PGSs for their studies
• Clearly report imputation panel and build in the documentation for downloadable results
• Call for large scale GWAS efforts in individuals of non-European ancestry

- One caveat: no LDpred scores.
- My advice:
  - Perform sensitivity analyses
  - Don't fret overly maximizing predictive power. Unlikely that is really the goal of your work.

How to get PGS in practice

- Software Options
  - PRSice, http://prsice.info/
  - plink, https://www.cog-genomics.org/plink2
  - LDpred, https://github.com/bvilhjal/ldpred
- Some are now available.
  - HRS, https://hrs.isr.umich.edu/data-products/genetic-data/products/pgs-list

THANKS!

bdomingue@stanford.edu