Estimating Effect Sizes and Polygenic Risk Scores

Biostat 666

3/29/21

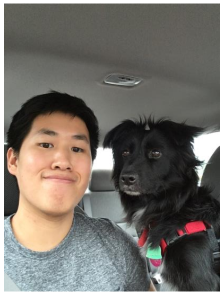
Kevin Liao

About Me

- 4th year Biostatistics student working with Sebastian
  - Former Genome Science Training Program Trainee

- From Chapel Hill, NC
- Hobbies: Tennis, golf, painting

- Current Research
  1) Polygenic risk scores for admixed individuals
  2) Genetic architecture of complex traits across diverse human populations



Lecture Outline

- Review: GWAS

- Estimating Effect Sizes
  - Measures of association: Risk Ratio vs Odds Ratio
  - LD confounding, Winner's Curse, Replication Studies

- Polygenic Risk Scores
  - Popular Methods of Construction
  - Strengths and Pitfalls

Review: GWAS

Review: Complex Traits

- Early genetic studies focused on Mendelian diseases
  - Single gene diseases that follow mendelian inheritance patterns
- “One gene, one mutation, out outcome” Model
- Well known monogenic diseases:

Disease	Type of Inheritance	Gene Responsible
Huntington’s Disease	Autosomal Dominant	Huntingtin (HTT)
Cystic Fibrosis	Autosomal Recessive	CFTR
Sickle Cell Anemia	Autosomal Recessive	Beta Hemoglobin (HBB)

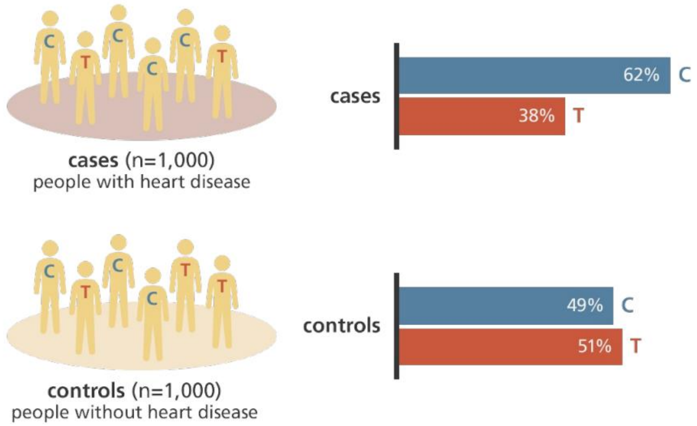
Review: Complex Traits

- Complex traits are traits influenced by many genes across the genome
  - Exp. Height, Type 2 Diabetes, Coronary Artery Disease, etc
- Studies of complex traits facilitated by sequencing technology
- Most commonly studied genetic variation are single nucleotide polymorphisms (SNPs)

Review: GWAS

- Genome wide association study (GWAS) used to study genetics of complex traits
- Basic idea of GWAS
  - 1) Collect sample of cases and controls for a trait
  - 2) Many loci across genome are genotyped/sequenced
  - 3) Associations tested by comparing frequency of alleles in cases and controls for each loci

Review: GWAS



Estimating Effect Sizes

Motivation

- GWAS allows framework to test SNPs for association with a phenotype
- Estimated effect sizes for each SNP provide insight into genetic architecture of disease
  - Which variants truly affect the disease?
  - Protective or Damaging?
  - How much of the phenotypic variance does genetics explain?

Study Designs

Prospective Study

- Cohorts followed over time to see who develops outcome
- Forward in time

Retrospective Study

- Outcome is established at start of study
- **GWAS are almost always retrospective case control studies**

Measure of association for GWAS

	Cases	Controls	Total
aA or AA	a	b	Unknown 1
aa	c	d	Unknown 2

Row totals unknown b/c of case ctrl sampling

- Would like to know the relative risk:
  - Risks easily interpretable: P(Disease)

$$RR = \frac{\Pr(Disease \mid genotype\ aA\ or\ AA)}{\Pr(Disease \mid genotype\ aa)} = \frac{a/Unknown_1}{c/Unknown_2}$$

- **Can’t get RR from retrospective case control study because you don’t known denominator!**

Measure of association for GWAS

- Odds ratios used for GWAS instead
  - Odds: Probability of event / Probability of no event

	Cases	Controls	Total
aA or AA	a	b	Unknown
aa	c	d	Unknown

Discussion: Why do the unknown row totals not matter?

$$OR = \frac{\Pr(Disease \mid genotype\ aA\ or\ AA) / \Pr(No\ disease \mid genotype\ aA\ or\ AA)}{\Pr(Disease \mid genotype\ aa) / \Pr(No\ disease \mid genotype\ aa)} = \frac{a/b}{c/d} = \frac{a * d}{b * c}$$

OR can approximate RR

- OR approximates RR when disease/health outcome is rare (i.e affecting < 10% in population)

Assume had data on all subjects

	Cases	Controls	Total
Exposed	a	b	a+b
Unexposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$OR = \frac{a/b}{c/d} \approx \frac{\frac{a}{a+b}}{\frac{c}{c+d}} = RR$$

Approximation holds when a & c small

How to estimate effect sizes

- Logistic regression often used to estimate effect sizes instead
  - Chi square test can't adjust for covariates
- Model Setup:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 G + \beta_2 X$$

where

π is the probability of being affected, $\Pr(Y = 1)$

$\log[\pi/(1-\pi)]$ - log odds of disease (logit)

G - genotype coded according to assumed model

X - other covariate (e.g., ancestry, age, gender, etc.)

How to estimate effect sizes

- Model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 G + \beta_2 X$$

- Genotype Coding:

Model	*aa*	*aA*	*AA*
Dominant	0	1	1
Recessive	0	0	1
Additive/multiplicative	0	1	2
Co-dominant*	0	1	0
(genotypic)	0	0	1

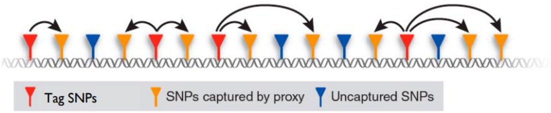
- Under additive model (most common):
  - β_1: change in log odds of disease for each additional minor allele
  - OR = e^{β_1}: odds of disease are increase by factor of X per each additional minor allele

Additional Factors when Estimating Effect Sizes

- Confounding of effect sizes due to LD
- Proportion of variance explained
- Winner’s Curse, Replication studies

Confounding of Effect Sizes due to LD

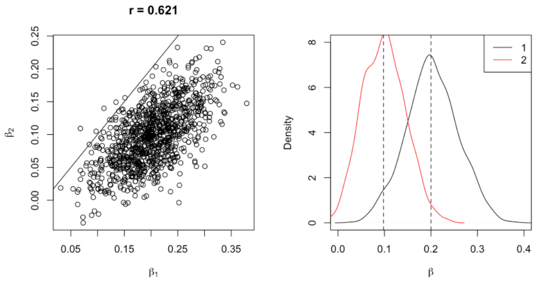
- Genotype arrays leverage LD to avoid genotyping all variants
  - Often tag variant genotyped rather than causal variant



- **Estimated marginal effect size for tag SNP j will depend on any causal SNPs in LD with**

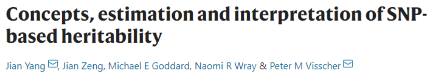
Simulation Experiment

- Run GWAS simulation experiment with two SNPs for 1000 times
  - SNP1 causal with effect size $\lambda_1 = 0.2$ and MAF = 0.2
  - SNP2 not causal with effect size $\lambda_2 = 0$ and MAF = 0.4
  - LD between SNPs: $r_{12}^2 = 0.60$



Discussion: What do you see from simulation results?

Proportion of Variance Explained



Concepts, estimation and interpretation of SNP-based heritability

Jian Yang, Jian Zeng, Michael E Goddard, Naomi R Wray & Peter M Visscher

- Decompose variance of phenotype

$$Y = \sum_{SNPs} x_j \beta_j + \epsilon$$

  - Var(Y): Total phenotypic variance

- SNP-based heritability h^2 is proportion of variance explained (PVE) due to set of SNPs

$$h^2 = \frac{var\left(\sum_{SNPs} x_j \beta_j\right)}{var(y)}$$

Proportion of Variance Explained

- Phenotypic variance explained for single SNP j:

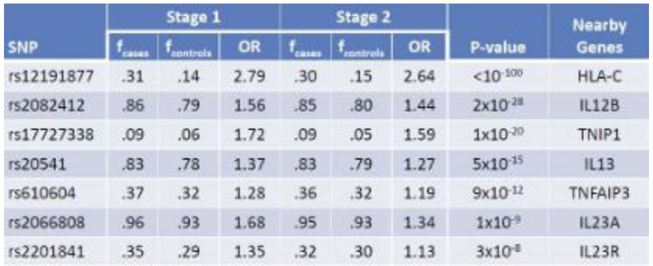
$$\mathrm{Var}(x_j\beta_j) = 2f_j(1-f_j){\beta_j}_{true}^2$$

Estimated using $\hat{\beta}_j$

- Impact of SNP j on PVE depends on:
  - Marginal effect size: β_j
  - **Allele frequency: $\boldsymbol{f_j}$**

Winner's Curse

- Significant associations likely stronger in GWAS sample than general population



SNP	Stage 1			Stage 2			P-value	Nearby Genes
	f_{cases}	$f_{controls}$	OR	f_{cases}	$f_{controls}$	OR		
rs12191877	.31	.14	2.79	.30	.15	2.64	$<10^{-100}$	HLA-C
rs2082412	.86	.79	1.56	.85	.80	1.44	2×10^{-28}	IL12B
rs17727338	.09	.06	1.72	.09	.05	1.59	1×10^{-20}	TNIP1
rs20541	.83	.78	1.37	.83	.79	1.27	5×10^{-15}	IL13
rs610604	.37	.32	1.28	.36	.32	1.19	9×10^{-12}	TNFAIP3
rs2066808	.96	.93	1.68	.95	.93	1.34	1×10^{-9}	IL23A
rs2201841	.35	.29	1.35	.32	.30	1.13	3×10^{-8}	IL23R

Winner's Curse

- Caused by thresholding on statistical significance.
  - Significant associations may have effects overestimated in a particular sample due to chance
- Winner's curse effect "stronger" when power of discovery GWAS low
- Solution: Larger sample sizes or Meta Analysis

Replication Studies

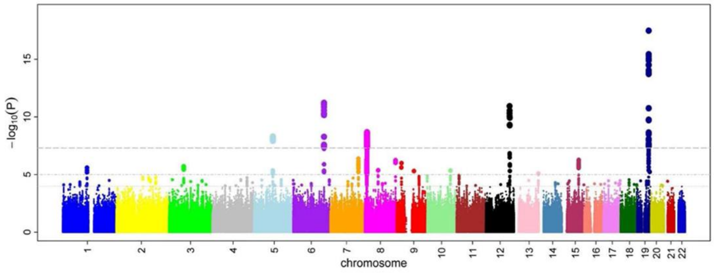
- Gold standard to validate genetic association is replication in another sample
- Replication sample should be independent and drawn from same population as original GWAS

Discussion: Will replication sample sizes ideally be smaller or larger than discovery GWAS sample size?

Break Time!

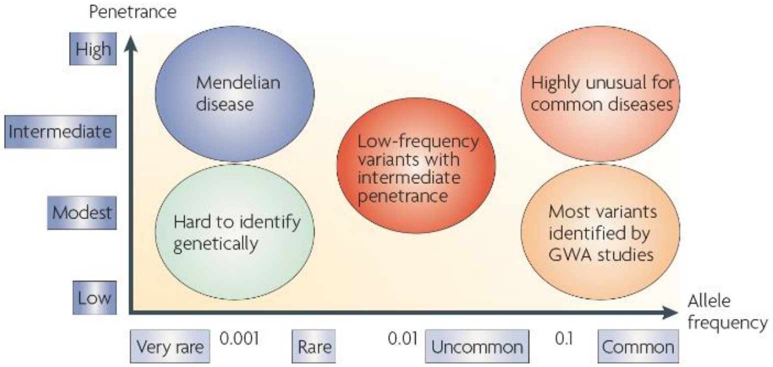
Polygenic Risk Scores

What to do after GWAS?



- GWAS has estimated effect sizes and identified risk variants
- Can we predict phenotypes using genetic information?

Reminder: Individual effect sizes small


Penetrance
High
Intermediate
Modest
Low
Mendelian disease
Hard to identify genetically
Low-frequency variants with intermediate penetrance
Highly unusual for common diseases
Most variants identified by GWA studies
Allele frequency
Very rare
0.001
Rare
0.01
Uncommon
0.1
Common

Polygenic Risk Score

- Polygenic risk scores (PRS) aggregate information from multiple small effect variants genome wide into a single score

- Each individual has a unique genetic portfolio of risk variants

$$PRS = \sum_{i=1}^{n} \beta_i G_i$$

Typically use GWAS estimated effect size $\hat{\beta}$

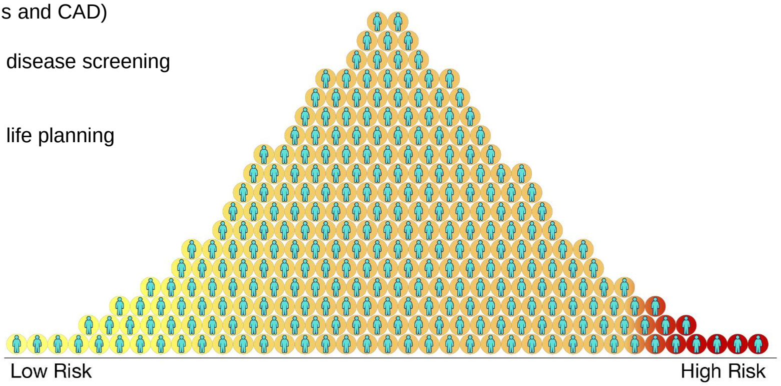
Strengths of PRS

The personal and clinical utility of polygenic risk scores

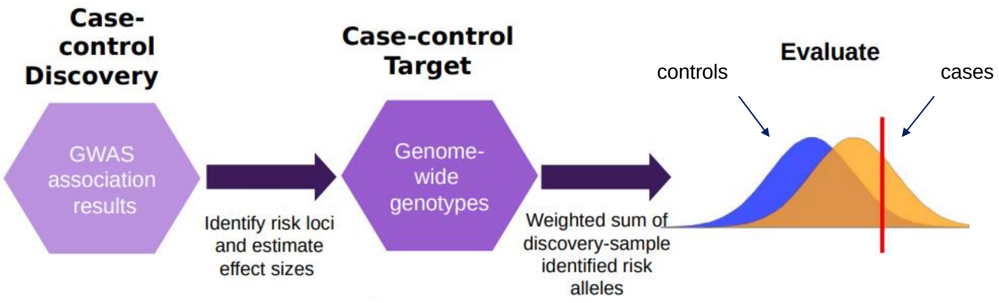
Ali Torkamani, Nathan E. Wineinger & Eric J. Topol

Nature Reviews Genetics **19**, 581–590(2018) | Cite this article

- Inform treatment use (Statins and CAD)
- Inform disease screening
- Inform life planning



Construction of Polygenic Risk Score



$PRS = \sum \beta_i G_i$

Discussion

Grad student Kevin has genetic data (~500,000 SNPs) for n=10,000 subjects and wants to make a PRS for disease X. He performs a GWAS for disease X to estimate effect sizes and makes a PRS using all 500,000 SNPs:

$$PRS = \sum \beta_i G_i$$

What’s the problem?

How did Kevin mess up

1) Overfitting!

- Kevin estimated effect sizes and made PRS in same data
- Overfitting falsely improves PRS

2) Including non-risk variants!

- Only a handful of variants are true risk variants.
- Adding noise hurts PRS

Solutions:

1) Overfitting!

- Use external set of summary statistics for PRS
- Ensure no sample overlap

2) Including non-risk variants!

- Prune out variants in high LD
- Variable selection/Shrinkage

Two Main Computational Frameworks

1) Shrinkage of β's

- Clumping and Thresholding
- Lassosum

2) Adjusting β's for LD

- LDpred

1) Clumping and Thresholding

Step 1: Clumping

- Remove correlated SNPs
- Clumping – Looks at most significant variants and removes nearby variants above some specified r^2

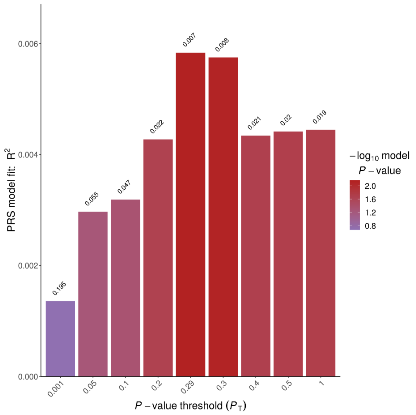
Step 2: Thresholding

- Try multiple p-value thresholds with SNPs under retained
- For each p-value threshold construct PRS and assess model fit
- **Note: Thresholding effectively shrinks β's to 0 for SNPs failing threshold**

1) Clumping and Thresholding

- PRSice is popular software for Clumping and Thresholding
- Here, P_T: 0.29 gives best PRS

Discussion: What is a problem of clumping and thresholding?



2) Lassosum

Genetic Epidemiology

INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY

RESEARCH ARTICLE | Full Access |

Polygenic scores via penalized regression on summary statistics

- Lassosum computes PRS using penalized regression (LASSO) on all summary statistics
- LASSO Overview:
  - Normal linear regression: $\boldsymbol{y} = \boldsymbol{XB} + \boldsymbol{\epsilon}$
    - $f(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta})$
  - LASSO minimizes objective function:
    - $f(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{2\lambda}\big|\big|\boldsymbol{\beta}\big|\big|_{\boldsymbol{1}}^{\boldsymbol{1}}$
  - **LASSO penalty provides shrinkage of $\boldsymbol{\beta}$'s (even to 0)**

2) Lassosum

- Lassosum objective function:

Note: lassosum doesn't use genotypes of your data set

$$f(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X\beta}) + \boldsymbol{2\lambda}\big|\big|\boldsymbol{\beta}\big|\big|_{\boldsymbol{1}}^{\boldsymbol{1}}$$

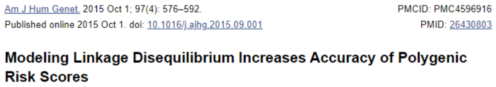
$$= \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X\beta} - \boldsymbol{2\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{2\lambda}\big|\big|\boldsymbol{\beta}\big|\big|_{\boldsymbol{1}}^{\boldsymbol{1}}$$

X^TX is LD matrix from external reference

X^Ty is correlations between SNP and phenotype from external data

- β estimates from minimizing function used to compute PRS for target sample: $PRS = \sum \beta_{i,lasso} G_i$

3) LDpred



Am J Hum Genet. 2015 Oct 1; 97(4): 576–592. PMCID: PMC4596916

Published online 2015 Oct 1. doi: 10.1016/j.ajhg.2015.09.001 PMID: 26430803

Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores

- LDpred is a Bayesian method that estimates posterior mean causal effect sizes given:
  - LD from an external reference panel
  - Prior on genetic architecture of trait
- Adjusts each variant's marginal effect β for nearby variants in LD with

3) LDpred

Step 1: Compute LD Matrix using external reference panel

Step 2: Define prior on genetic architecture

- Infinitesimal model:

$$\beta_i \sim_{iid} N(0, \frac{h_g^2}{M})$$

h_g^2 is SNP-based heritability estimated from effect sizes

- Non-infinitesimal model:

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) \text{ with probability } p \\ 0 \text{ with probability } (1-p), \end{cases}$$

3) LDpred

Step 3: Estimate posterior effect sizes

- Infinitesimal model:

$$E\left(\beta^l \middle| \tilde{\beta}^l, D\right) \approx \left(\frac{M}{Nh_g^2} I + D_l\right)^{-1} \tilde{\beta}^l.$$

LD matrix

- Non-infinitesimal model:
  - Analytical expression for posterior mean hard. Uses MCMC Gibbs sampler instead

Step 4: Use posterior effect sizes to construct PRS

- $PRS = \sum \beta_{i,post} G_i$

Evaluating PRS performance

Regression Model:

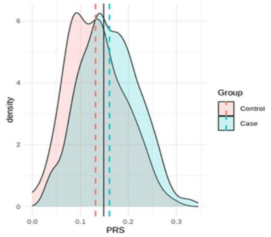
$$Phenotype = \beta_0 + \beta_1 PRS + \beta Covariates$$

1) P-value for β_1 corresponding to null of no association

- Sensitive to sample size

2) Case control Separation

- T-test for difference in means



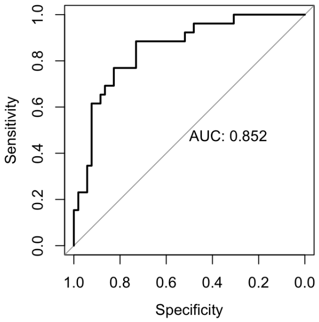
Evaluating PRS performance

3) R^2 metrics

- Quantitative: R^2 is proportion of variance explained
- Binary: Nagelkerke R^2
  - Sensitive to proportion of cases in testing data

4) AUC – Area under the curve

- Prob that the PRS of a random case is larger than PRS of random control
- Nice property that independent of proportion of cases

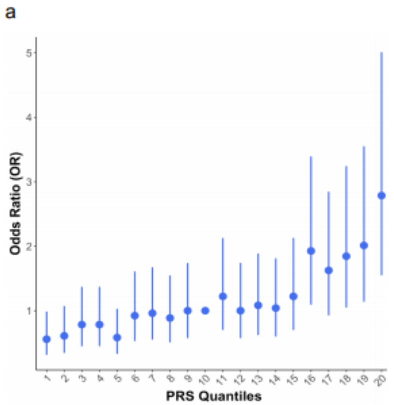


Evaluating PRS performance

5) Odds Ratio by PRS Quantiles

- Construct quantiles for PRS
- Fit logistic regression using quantiles as predictor

$$Phenotype = \beta_0 + \beta_1 PRS_{quant2} + \cdots + \beta_{19} PRS_{quant20}$$

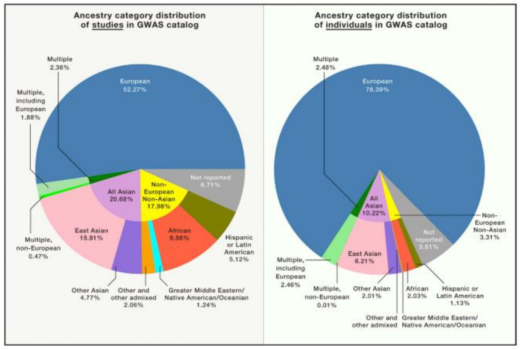


Pitfalls of PRS

The Missing Diversity in Human Genetic Studies

Giorgio Sirugo 6 • Scott M. Williams 6 • Sarah A. Tishkoff 6 • Show footnotes

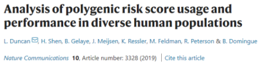
- Most genetic studies done in Europeans
- Genotype-phenotype associates can differ across populations
  - LD differences
  - Allele frequency differences
  - Unique environments



Pitfalls of PRS

The Missing Diversity in Human Genetic Studies

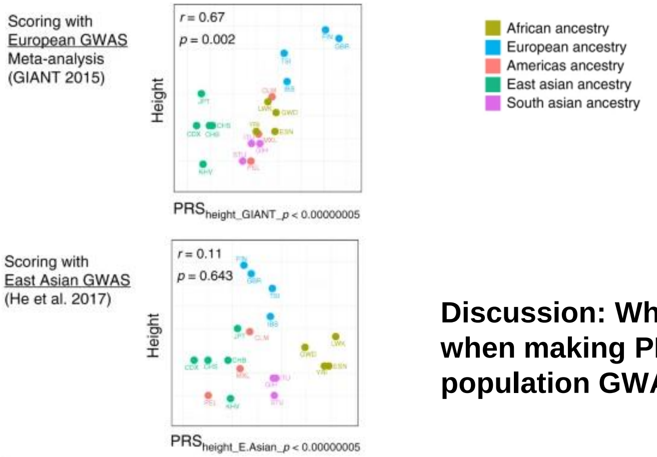
Giorgio Sirugo 6 • Scott M. Williams 6 • Sarah A. Tishkoff 6 • Show footnotes



Analysis of polygenic risk score usage and performance in diverse human populations

L. Duncan, H. Shen, B. Gelaye, J. Meijsen, K. Ressler, M. Feldman, R. Peterson & B. Domingue

Nature Communications **10**, Article number: 3328 (2019) | Cite this article



Discussion: What do you notice when making PRS with different population GWAS?

Future of PRS

- PRS methods development is active area of research
  - Construction of PRS
  - Transferring PRS across populations
- Increase clinical utility of PRS
  - Currently PRS only used for a handful of traits (CAD, prostate cancer, breast cancer, etc)
  - Informing physicians and public education regarding interpretation

Overview

- Measures of association for GWAS
- Factors to consider when estimating effect sizes
  - LD confounding
  - Going from β to proportion of variance explained
  - Winner’s curse, replication studies
- Polygenic risk scores

Thanks!