Polygenic risk scores

Alicia Martin, PhD
Stanley Center Global Plenary 2018
September 12, 2018

armartin@broadinstitute.org  @genetisaur
Outline

• What are polygenic risk scores?
• How to compute them
• Methods, interpretations, and uses
• Ancestry, health disparities, and ongoing/future directions
How Genetics Is Changing Our Understanding of ‘Race’

By David Reich

March 23, 2018

Race, Genetics and a Controversy

April 2, 2018
An Enormous Study of the Genes Related to Staying in School

Researchers have found 1,271 gene variants associated with years of formal education. That’s important, but not for the obvious reasons.

---

Opinion

Why Progressives Should Embrace the Genetics of Education

By Kathryn Paige Harden
Dr. Harden is a psychologist who studies how genetic factors shape adolescent development.

July 24, 2018

---

Why We Shouldn't Embrace the Genetics of Education

It's a trap!

By John Warner // July 26, 2018
Forecasts of genetic fate just got a lot more accurate

Clues to Your Health Are Hidden at 6.6 Million Spots in Your DNA

By Gina Kolata

Aug. 13, 2018

With a sophisticated new algorithm, scientists have found a way to forecast an individual’s risks for five deadly diseases.
How scientists are learning to predict your future with your genes

But what are the limits?

By Brian Resnick | @B_resnick | brian@vox.com | Updated Aug 25, 2018, 9:35am EDT

Insight & Intelligence

August 22, 2018

Why Do Polygenic Risk Scores Get So Much Hype?

GWAS for Common Disease Variants Gains Prominence

Julianna LeMieux, Ph.D.
The rise of the polygenic risk score

Previous criticisms: limited sample size
Cheap test for insights into many diseases
Integrate with other clinical factors for therapeutic decision-making

“We propose the time has come to incorporate genetic risk scores into clinical practice”

A long shared history between PRS and breeding values

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

- The dark days of low-powered GWAS
- PRS show value of GWAS even in the absence of genome-wide significant loci
What is a polygenic risk score?

Genetic prediction of an individual’s phenotype

\[ Y = \sum_{j=1}^{m} g_j \beta_j \]

Sum the products of genotypes × effect size estimates from a GWAS across the genome.
What is a polygenic risk score?

**Fundamental choices:**
- Which SNPs to include
- What weights to apply

**Considerations:**
- LD
- P-value thresholds

**Genetic prediction of an individual’s phenotype**

\[
Y = \sum_{j=1}^{m} g_j \beta_j
\]

Sum the products of genotypes × effect size estimates from a GWAS across the genome.
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible discovery samples
2. Obtain independent target samples with genome-wide data
3. Identify SNPs in common between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with \( p < \text{various thresholds} \) (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. Calculate PRS as sum of risk alleles weighted by \( \beta \) from GWAS
7. Evaluate PRS accuracy by regressing trait in target sample onto PRS (e.g. \( R^2 \))

Content shamelessly borrowed and modified from Matthew Keller
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible discovery samples
2. Obtain independent target samples with genome-wide data
3. Identify SNPs in common between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with \( p < \) various thresholds (e.g., \( 5 \times 10^{-8} \), \( 1 \times 10^{-6} \), \( 1 \times 10^{-4} \), 0.05, 1)
6. Calculate PRS as sum of risk alleles weighted by \( \beta \) from GWAS
7. Evaluate PRS accuracy by regressing trait in target sample onto PRS (e.g. \( R^2 \))
1. Obtain large GWAS

Trait info: http://www.ukbiobank.ac.uk/data-showcase/
All things UK Biobank GWAS: http://www.nealelab.is/uk-biobank/
More powerful GWAS = more accurate predictor

What do GWAS summary statistics contain?

Minimal useful info: variant ID, chromosome, position, risk and protective allele, sample size, p-value, effect size, standard error

Example: standing height

<table>
<thead>
<tr>
<th>variant</th>
<th>minor_allele</th>
<th>minor_AF</th>
<th>low_confidence_variant</th>
<th>n_complete_samples</th>
<th>AC</th>
<th>ytx</th>
<th>beta</th>
<th>se</th>
<th>tstat</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:15791:C:T</td>
<td>T</td>
<td>5.44E-09</td>
<td>TRUE</td>
<td>360388</td>
<td>3.92E-03</td>
<td>3.47E-03</td>
<td>1.80E+01</td>
<td>1.78E+02</td>
<td>1.01E-01</td>
<td>9.19E-01</td>
</tr>
<tr>
<td>1:69487:G:A</td>
<td>A</td>
<td>5.76E-06</td>
<td>TRUE</td>
<td>360388</td>
<td>4.15E+00</td>
<td>-8.75E-02</td>
<td>-4.13E-02</td>
<td>3.5E-01</td>
<td>-1.18E-01</td>
<td>9.06E-01</td>
</tr>
<tr>
<td>1:69569:T:C</td>
<td>C</td>
<td>1.88E-04</td>
<td>TRUE</td>
<td>360388</td>
<td>1.36E+02</td>
<td>-2.08E+00</td>
<td>-4.70E-02</td>
<td>6.27E-02</td>
<td>-7.5E-01</td>
<td>4.54E-01</td>
</tr>
<tr>
<td>1:139853:C:T</td>
<td>T</td>
<td>5.67E-06</td>
<td>TRUE</td>
<td>360388</td>
<td>4.09E+00</td>
<td>-1.06E-01</td>
<td>-4.21E-02</td>
<td>3.5E-01</td>
<td>-1.2E-01</td>
<td>9.04E-01</td>
</tr>
<tr>
<td>1:692794:CA:C</td>
<td>C</td>
<td>1.11E-01</td>
<td>FALSE</td>
<td>360388</td>
<td>7.97E+04</td>
<td>1.02E+02</td>
<td>7.97E-04</td>
<td>2.90E-03</td>
<td>2.75E-01</td>
<td>7.83E-01</td>
</tr>
<tr>
<td>1:693731:A:G</td>
<td>G</td>
<td>1.16E-01</td>
<td>FALSE</td>
<td>360388</td>
<td>8.35E+04</td>
<td>-6.93E+01</td>
<td>-1.44E-03</td>
<td>2.74E-03</td>
<td>-5.24E-01</td>
<td>6.00E-01</td>
</tr>
<tr>
<td>1:707522:G:C</td>
<td>C</td>
<td>9.73E-02</td>
<td>FALSE</td>
<td>360388</td>
<td>7.01E+04</td>
<td>-7.86E+00</td>
<td>2.47E-04</td>
<td>3.08E-03</td>
<td>8.02E-02</td>
<td>9.36E-01</td>
</tr>
<tr>
<td>1:717587:G:A</td>
<td>A</td>
<td>1.57E-02</td>
<td>FALSE</td>
<td>360388</td>
<td>1.13E+04</td>
<td>5.47E+00</td>
<td>1.13E-03</td>
<td>7.35E-03</td>
<td>1.54E-01</td>
<td>8.77E-01</td>
</tr>
<tr>
<td>1:723329:A:T</td>
<td>T</td>
<td>1.73E-03</td>
<td>FALSE</td>
<td>360388</td>
<td>1.25E+03</td>
<td>3.87E+01</td>
<td>2.22E-02</td>
<td>2.17E-02</td>
<td>1.02E+00</td>
<td>3.06E-01</td>
</tr>
</tbody>
</table>
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible discovery samples
2. Obtain independent target samples with genome-wide data
3. Identify SNPs in common between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with $p < \text{various thresholds}$ (e.g., $5\times10^{-8}$, $1\times10^{-6}$, $1\times10^{-4}$, 0.05, 1)
6. Calculate PRS as sum of risk alleles weighted by $\beta$ from GWAS
7. Evaluate PRS accuracy by regressing trait in target sample onto PRS (e.g. $R^2$)

Content shamelessly borrowed and modified from Matthew Keller
2. Independent target cohort must be independent

Prediction “accuracy” measures will be overestimated if discovery and target are not independent. This can arise if:

- The same people are in both cohorts
- There are close relatives between the two
- SNPs are selected from meta-analysis of discovery + target
Choose your favorite dataset

Most people like phenotypes, but...

Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible discovery samples
2. Obtain independent target samples with genome-wide data
3. Identify SNPs in common between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with $p <$ various thresholds (e.g., $5 \times 10^{-8}$, $1 \times 10^{-6}$, $1 \times 10^{-4}$, 0.05, 1)
6. Calculate PRS as sum of risk alleles weighted by $\beta$ from GWAS
7. Evaluate PRS accuracy by regressing trait in target sample onto PRS (e.g. $R^2$)

Content shamelessly borrowed and modified from Matthew Keller
3. Use SNPs in common

Array data

Illumina array

Axiom array

Imputed data

Phase and impute data to help maximize overlap

Content shamelessly borrowed and modified from Matthew Keller
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible discovery samples
2. Obtain independent target samples with genome-wide data
3. Identify SNPs in common between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with $p < \text{various thresholds}$ (e.g., $5\times10^{-8}$, $1\times10^{-6}$, $1\times10^{-4}$, 0.05, 1)
6. Calculate PRS as sum of risk alleles weighted by $\beta$ from GWAS
7. Evaluate PRS accuracy by regressing trait in target sample onto PRS (e.g. $R^2$)

Content shamelessly borrowed and modified from Matthew Keller
4. Account for LD

Two primary approaches:

- LD clumping (heuristic, less good)
- In PLINK, --clump
- Model LD! LDPred (better, but harder to run)
Clumping with PLINK

Example:

plink --bfile [reference LD panel] \ 
   --clump [summary statistics] \ 
   --clump-field [summary statistics p-value column name] \ 
   --clump-snp-field [summary statistics snp column name] \ 
   --clump-p1 1 \ 
   --clump-p2 1 \ 
   --clump-r2 0.5 \ 
   --clump-kb 250 \ 
   --out [output filename]
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible **discovery samples**
2. Obtain independent **target samples** with genome-wide data
3. Identify **SNPs in common** between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with $p < \text{various thresholds}$ (e.g., $5\times10^{-8}$, $1\times10^{-6}$, $1\times10^{-4}$, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by $\beta$ from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. $R^2$)

Content shamelessly borrowed and modified from Matthew Keller
5. Use various $p$ thresholds

Use $p$-thresholds from $5e^{-8}$, $1e^{-7}$, $0.05$,...1
Report results from all thresholds
For PLINK

Create a file with multiple thresholds, for example:

<table>
<thead>
<tr>
<th>Threshold name</th>
<th>lower bound</th>
<th>upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>0</td>
<td>0.000000005</td>
</tr>
<tr>
<td>s2</td>
<td>0</td>
<td>0.0000001</td>
</tr>
<tr>
<td>s3</td>
<td>0</td>
<td>0.0001</td>
</tr>
<tr>
<td>s4</td>
<td>0</td>
<td>0.001</td>
</tr>
<tr>
<td>s5</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>s6</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>s7</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>s8</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>s9</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>s10</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible discovery samples
2. Obtain independent target samples with genome-wide data
3. Identify SNPs in common between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with $p < \text{various thresholds}$ (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. **Calculate PRS** as sum of risk alleles weighted by $\beta$ from GWAS
7. Evaluate **PRS accuracy** by regressing trait in target sample onto PRS (e.g. $R^2$)

Content shamelessly borrowed and modified from Matthew Keller
6. Calculate PRS

• \( \text{PRS}_j = \sum [\beta_{i, \text{discovery}} \times \text{SNP}_{ij}] \)
  
  • \( \beta_{i, \text{discovery}} = \) effect size in discovery sample from
    
    • linear regression (continuous trait)
    
    • logistic regression (binary trait; \( \beta = \log(\text{OR}) \))

• \( \text{SNP}_{ij} = \) # alleles (0,1,2) for SNP \( i \) of person \( j \) in target sample

• In PLINK, --score.
In PLINK

Example:

plink --bfile [best guess genotypes] \ 
--extract [clumped snps] \ 
--q-score-range [range file] [summary stats] [variant ID column #] [p-value column #] [header] \ 
--score [summary stats] [variant ID column #] [allele column #] [effect size column #] \ 
--out [output file]
Most common steps to calculate PRS

1. Obtain GWAS summary statistics from the largest possible discovery samples
2. Obtain independent target samples with genome-wide data
3. Identify SNPs in common between both datasets
4. Deal with association redundancy due to LD
5. Restrict to SNPs with $p <$ various thresholds (e.g., 5e-8, 1e-6, 1e-4, 0.05, 1)
6. Calculate PRS as sum of risk alleles weighted by $\beta$ from GWAS
7. Evaluate PRS accuracy by regressing trait in target sample onto PRS (e.g. $R^2$)

Content shamelessly borrowed and modified from Matthew Keller
7. Evaluate PRS accuracy

- For continuous traits, this is simply the $R^2$ from regressing trait $\sim$ PRS in target + covariates

- Trickier for binary (e.g., case-control) data due to ascertainment

- Often Nagelkerke’s $R^2$ is reported. Unfortunate, because this depends on prevalence and case:control ratio.

Content shamelessly borrowed and modified from Matthew Keller
7. Evaluate PRS accuracy

- For continuous traits, this is simply the $R^2$ from regressing trait $\sim$ PRS in target + covariates
- Trickier for binary (e.g., case-control) data due to ascertainment
  - Often Nagelkerke’s $R^2$ is reported. Unfortunate, because this depends on prevalence + case:control ratio.
  - Better: liability-scale $R^2$

Content shamelessly borrowed and modified from Matthew Keller
7. Please report comparable $R^2$!

(trilling stuff, I know)

TABLE I. Brief description of $R^2$ measures used in this study and their theoretical expectation

<table>
<thead>
<tr>
<th>Brief description</th>
<th>Notation and formula</th>
<th>Expectation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ on the observed scale</td>
<td>$R^2 = 1 - \frac{\sum_i (y_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$</td>
<td>$h^2$</td>
</tr>
<tr>
<td>Cox and Snell’s $R^2$ on the observed scale</td>
<td>$R^2_{C&amp;S} = 1 - \left( \frac{\text{Likelihood}<em>{null}}{\text{Likelihood}</em>{full}} \right)^{2/N}$</td>
<td>$h^2$</td>
</tr>
<tr>
<td>Nagelkerke’s $R^2$ on the observed scale</td>
<td>$R^2_N = \frac{R^2_{C&amp;S}}{1-(\text{Likelihood}_{null})^{2/N}}$</td>
<td>$R^2_{\text{Nag}}$</td>
</tr>
<tr>
<td>$R^2$ on the liability scale</td>
<td>$R^2_l = R^2_0 \frac{K(1-K)}{P^2}$</td>
<td>$h^2_l$</td>
</tr>
<tr>
<td>$R^2$ on the probit liability scale</td>
<td>$R^2_{\text{probit}} = \frac{\text{var}(\theta_{\text{probit}})}{\text{var}(\theta_{\text{probit}}) + 1}$</td>
<td>$h^2_l$</td>
</tr>
<tr>
<td>$R^2$ on the logit liability scale</td>
<td>$R^2_{\text{logit}} = \frac{\text{var}(\theta_{\text{logit}})}{\text{var}(\theta_{\text{logit}}) + 3.29}$</td>
<td>$h^2_l$</td>
</tr>
<tr>
<td>$R^2$ on the liability scale using AUC</td>
<td>$R^2_{\text{AUC}} = \frac{Q^2}{(m_2 + m_1)^2 - Q^2 m_1} * \frac{m_1}{m_2}$</td>
<td>$h^2_l$</td>
</tr>
<tr>
<td>$R^2$ on the liability scale when using ascertained case-control studies</td>
<td>$R^2_{\text{CC}} = \frac{R^2_{\text{CC}} C}{1 + R^2_{\text{CC}}}$</td>
<td>$h^2_l$</td>
</tr>
</tbody>
</table>

$y_i$ observations that are 0 or 1 for unaffected and affected individuals; $h^2_l$, heritability on the liability scale, in this context the proportion of variance on the liability scale explained by the genetic profile; $K$, population prevalence; $z$, the height of a normal density curve at the point according to $K$; $g$, the sum of all additive genetic factors in the estimated genetic predictor; $\beta$, regression coefficient from generalized linear model; $m$, the mean liability for cases; $m_2$, the mean liability for controls; $t$, the threshold on the normal distribution that truncates the proportion of disease prevalence $K$; $Q$, the inverse of the cumulative density function of the normal distribution up to values of AUC; C and $\theta$, correcting factors for ascertainment.

So now you have a PRS...

- What are polygenic risk scores?
- How to compute them
- Methods, interpretations, and uses
- Ancestry, health disparities, and ongoing/future directions
The rise of the polygenic risk score

No discussion of ancestry!

“We propose the time has come to incorporate genetic risk scores into clinical practice”

- Previous criticisms: limited sample size
- Cheap test for insights into many diseases
- Integrate with other clinical factors for therapeutic decision-making

Genomics has a diversity problem

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)
Causal effects are mostly shared across populations... but what about other effects?

European
East Asian
European
East Asian

...but what about other effects?
Predictable basis of PRS disparities

Prediction accuracy decays with $F_{ST}$. Why?

- GWAS best-powered to discover common variants
- LD differences across populations
- More sparse association
- Environmental, selection, and more complicated differences
Polygenic height scores are substantially different across populations.

These differences are not meaningful.
Coalescent model for simulation framework


Model parameters
- $N_e$: population size
- $m$: migration rates
- $T$: time
- $r$: growth

Model parameters:
- $N_A$: population size
- $N_{AF}$: population size
- $N_B$: population size
- $N_{EU0}$: population size
- $N_{AS0}$: population size
- $T_{AF}$: time
- $T_B$: time
- $T_{EuAs}$: time

Africans

Europeans

East Asians

$N_A$: 7300
$N_{AF}$: 14474
$N_B$: 1861
$N_{EU0}$: 1032
$N_{AS0}$: 550
$T_{AF}$: 148kya
$T_B$: 51kya
$T_{EuAs}$: 23kya

$m_{AfB}$
$m_{AfAs}$
$m_{AfEu}$
$m_{EuAs}$
Simulation overview

1. Simulate genotypes (AFR, EUR, EAS)

2. Assign evenly spaced causal variants

3. Compute $\text{PRS}_{\text{TRUE}}$

$$X = \sum_{i=1}^{m} g_i \beta_i$$

4. Define EUR cases, controls (10k each)

5. Run a EUR GWAS

6. Compute $\text{PRS}_{\text{INFER}}$ across populations

$$X = \sum_{i=1}^{m} g_i \beta_i$$
\textbf{PRS}_{\text{TRUE}} \text{ is not significantly different across populations}

\textit{True causal variants}
**PRSI_{INFER}** is highly stratified across populations

**True causal variants**

**GWAS inferred variants**
Unpredictable PRS biases across populations
Unpredictable PRS biases across populations

Analogous to different traits:

Height

Schizophrenia

T2D

For a given trait, impossible to predict *a priori* which population will have highest inferred risk!
Staggering PRS disparities across populations

<table>
<thead>
<tr>
<th>Trait/study</th>
<th>Target population</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI (Belsky)</td>
<td>African American</td>
</tr>
<tr>
<td>BMI (Ware)</td>
<td>East Asian</td>
</tr>
<tr>
<td>EA (Domingue)</td>
<td>European</td>
</tr>
<tr>
<td>EA (Lee)</td>
<td>SCZ (Ripke)</td>
</tr>
<tr>
<td>EA (Ware)</td>
<td>SCZ (Vassos)</td>
</tr>
<tr>
<td>Height (Ware)</td>
<td>SCZ (Vilhjalmsson)</td>
</tr>
<tr>
<td>SCZ (Ripke)</td>
<td>Mean (African American)</td>
</tr>
<tr>
<td>SCZ (Vassos)</td>
<td>EA (Ware)</td>
</tr>
<tr>
<td>SCZ (Vilhjalmsson)</td>
<td>SCZ (Vilhjalmsson)</td>
</tr>
</tbody>
</table>

Variance explained

<table>
<thead>
<tr>
<th>Proportion variance explained</th>
<th>relative to Europeans</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.00</td>
</tr>
<tr>
<td>B</td>
<td>0.25</td>
</tr>
<tr>
<td>C</td>
<td>0.50</td>
</tr>
<tr>
<td>D</td>
<td>0.75</td>
</tr>
<tr>
<td>E</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities *(in prep)*
Staggering PRS disparities across populations

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)
Consistent promise from diversifying efforts

Despite 3X larger sample sizes in Europeans, prediction in East Asians performs best with matched training data.

Other examples:
- BMI (Akiyama et al, 2018 Nat Gen)
- SCZ (Li et al, 2017 Nat Gen)
Consistent promise from diversifying efforts

**Goal**: Compare PRS accuracy for 17 traits in UKBB and BBJ

- Randomly set aside 5,000 individuals from each biobank
  - Match BBJ proportion with disease ascertainment
- Run GWAS on all other BBJ individuals. Match numbers in UKBB.

Do we see symmetric, comparable PRS accuracy?

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (*in prep*)
Consistent promise from diversifying efforts

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)
Consistent promise from diversifying efforts

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)
Consistent promise from diversifying efforts

Note: differing axes... why?

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)
Trans-ethnic genetic correlation is quite high
Cohort definition matters!

- UKBB has a “healthy volunteer” bias (healthier than average population)

- BBJ cohort is ascertained for 47 diseases (sicker than average population)

  - Manually transcribe patients’ data from medical records in each hospital, read through and re-enter into BBJ’s electronic database
Cohort definition matters!

<table>
<thead>
<tr>
<th>Trait</th>
<th>Observed $h^2$ (BBJ)</th>
<th>Observed $h^2$ (UKBB)</th>
<th>SE (BBJ)</th>
<th>SE (UKBB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basophil</td>
<td>0.0441</td>
<td>0.0213</td>
<td>0.0121</td>
<td>0.0050</td>
</tr>
<tr>
<td>BMI</td>
<td>0.1361</td>
<td>0.1955</td>
<td>0.0087</td>
<td>0.0090</td>
</tr>
<tr>
<td>DBP</td>
<td>0.0430</td>
<td>0.0984</td>
<td>0.0051</td>
<td>0.0068</td>
</tr>
<tr>
<td>Eosinophil</td>
<td>0.0586</td>
<td>0.1354</td>
<td>0.0093</td>
<td>0.0167</td>
</tr>
<tr>
<td>Hb</td>
<td>0.0452</td>
<td>0.1054</td>
<td>0.0053</td>
<td>0.0107</td>
</tr>
<tr>
<td>Height</td>
<td>0.3059</td>
<td>0.3675</td>
<td>0.0187</td>
<td>0.0208</td>
</tr>
<tr>
<td>Ht</td>
<td>0.0457</td>
<td>0.0942</td>
<td>0.0056</td>
<td>0.0093</td>
</tr>
<tr>
<td>Lymphocyte</td>
<td>0.0516</td>
<td>0.1318</td>
<td>0.0073</td>
<td>0.0118</td>
</tr>
<tr>
<td>MCH</td>
<td>0.1309</td>
<td>0.1942</td>
<td>0.0184</td>
<td>0.0210</td>
</tr>
<tr>
<td>MCHC</td>
<td>0.0481</td>
<td>0.0402</td>
<td>0.0080</td>
<td>0.0052</td>
</tr>
<tr>
<td>MCV</td>
<td>0.1447</td>
<td>0.1994</td>
<td>0.0178</td>
<td>0.0201</td>
</tr>
<tr>
<td>Monocyte</td>
<td>0.0448</td>
<td>0.1331</td>
<td>0.0090</td>
<td>0.0177</td>
</tr>
<tr>
<td>Neutrophil</td>
<td>0.0758</td>
<td>0.1153</td>
<td>0.0097</td>
<td>0.0131</td>
</tr>
<tr>
<td>Platelet</td>
<td>0.1260</td>
<td>0.2012</td>
<td>0.0148</td>
<td>0.0179</td>
</tr>
<tr>
<td>RBC</td>
<td>0.0818</td>
<td>0.1586</td>
<td>0.0093</td>
<td>0.0141</td>
</tr>
<tr>
<td>SBP</td>
<td>0.0574</td>
<td>0.1041</td>
<td>0.0063</td>
<td>0.0070</td>
</tr>
<tr>
<td>WBC</td>
<td>0.0778</td>
<td>0.1286</td>
<td>0.0074</td>
<td>0.0114</td>
</tr>
</tbody>
</table>
...but a lot of room for growth

Note: differing axes

Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)
New statistical approaches for genetic prediction

**Study**
- Multi-population
- Single population mismatch
- Recently admixed population

**Target**
- AND
- OR

**Approach**
- Multi-ancestry meta-analysis (MAMA)
- Kalman filter
- Personalized LD Panel

**Under construction**
Genetic prediction with GWAS from multiple populations

- **Approach**: Consider cross-population LD to recalibrate effect sizes in each population
- **Related methods**: LD score regression, MTAG
- **Status**: Implementing across global biobanks

**Approach**

- Multi-population
  - AND
  - OR

**Target**

- Multi-ancestry meta-analysis (MAMA)

- Patrick Turley
- Hui Li
- Raymond Walters
GWAS stats differ across populations due to LD

\[
\hat{\beta}_j = \sum_{k=1}^{M} r_{j,k} b + \epsilon
\]

• Causal effect sizes tend to be the same...

• ... but effect size estimates vary with LD differences across populations

\[
\hat{\beta}_{A,j} \neq \hat{\beta}_{B,j}
\]
Key elements of MAMA

$\Omega_j$

- Variance-covariance of genetic component
- More information shared when LD patterns and conditional effects are similar

$\Sigma_j$

- Variance-covariance of error and biases
- Less information shared when estimates are noisy or biased
Applications in real data

- Psychiatric disorders

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Population/ Location</th>
<th>N Cases</th>
<th>N Controls</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia</td>
<td>Europe</td>
<td>34,989</td>
<td>113,075</td>
<td>PGC</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>East Asia</td>
<td>13,305</td>
<td>16,244</td>
<td>PGC</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>African Americans</td>
<td>6,981</td>
<td>2,564</td>
<td>PGC</td>
</tr>
<tr>
<td>Bipolar/Schizophrenia</td>
<td>Hispanic/Latinos</td>
<td>3,982</td>
<td>4,553</td>
<td>PGC</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>Africa</td>
<td>~18,000</td>
<td>~18,000</td>
<td>NeuroGAP</td>
</tr>
<tr>
<td>PTSD</td>
<td>U.S. minorities</td>
<td>21,845</td>
<td>58,769</td>
<td>PGC/CVB</td>
</tr>
</tbody>
</table>

- Anthropometric traits (height, BMI, blood panels, etc)

<table>
<thead>
<tr>
<th>Biobank</th>
<th>Code</th>
<th>Sample sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK Biobank</td>
<td>UKBB</td>
<td>~500k</td>
</tr>
<tr>
<td>Finnish biobank</td>
<td>Finrisk</td>
<td>~50k</td>
</tr>
<tr>
<td>BioBank Japan Project</td>
<td>BBJ</td>
<td>~162k</td>
</tr>
<tr>
<td>China Kadoorie Biobank</td>
<td>CKB</td>
<td>~100k</td>
</tr>
<tr>
<td>PAGE (US minorities)</td>
<td>PAGE</td>
<td>~50k</td>
</tr>
</tbody>
</table>
Lots of nice resources!

Some nice reviews:


Coming soon:

• Martin, Kanai, Daly. Clinical use of genetic risk scores will exacerbate existing health disparities (in prep)

• Martin, Daly, Robinson, Hyman, & Neale. Predicting polygenic risk of psychiatric disorders (in revision)
Conclusions

- Polygenic risk scores have the potential to improve clinical models, but are currently likely to increase health disparities due to Eurocentric GWAS biases.

- We need more diverse GWAS studies and new methods to address these major issues.

- We are developing new methods that can use biobank-scale data from diverse populations to improve the generalizability of genetic prediction across populations.
Future directions

• How will we use PRS in the future?

• Tricky issues to resolve:
  • Pleiotropy
  • Healthcare economics: $ and life disparities by ethnicity?
Acknowledgments

MGH/Broad
• Mark Daly
• Ben Neale
• Patrick Turley
• Raymond Walters
• Duncan Palmer
• Hailiang Huang
• Chia-Yen Chen
• Masahiro Kanai
• Elizabeth Atkinson
• Caitlin Carey
• Alex Bloemendal
• Karestan Koenen
• Giulio Genovese
• Elise Robinson
• Steve Hyman

Mt. Sinai
• Eimear Kenny

UC Denver
• Chris Gignoux

McGill
• Simon Gravel

FIMM
• Samuli Ripatti
• Juuka Koskela
• Juulia Partanen
• Sini Kerminen
• Matti Pirinen
• Aarno Palotie

RIKEN
• Y. Okada
• Yoichiro Kamatani

Consortia
1000 Genomes Project
NeuroGAP-Psychosis
PGC-SCZ
• Max Lam
SSGAC
• Hui Li
• Dan Benjamin
• David Cesarini
• Meghan Zacher

Funding
• K99MH117229

Questions/comments?
armartin@broadinstitute.org
@genetisaur