Polygenic Scores in Add Health

David B. Braudt
braudt@unc.edu
Polygenic Scores (PGSs): An Overview

- PGSs capture the combined additive genetic influence of SNPs across the entire genome on a specific trait/behavior in a single measure

$$\text{PGS}_i = \mathbf{X}$$

<table>
<thead>
<tr>
<th>ID</th>
<th>SNP 1</th>
<th>SNP 2</th>
<th>SNP 3</th>
<th>...</th>
<th>SNP m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>...</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>n</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Molecular genetic data

GWAS summary statistics

Add Health

The National Longitudinal Study of Adolescent to Adult Health
Polygenic Scores (PGSs): An Overview

- PGSs capture the combined additive genetic influence of SNPs across the entire genome on a specific trait/behavior in a single measure.

\[ PGS_i = \sum_{j=1}^{k} \beta_j SNP_{ij} \]  

standardized within ancestry groups  

\[ \mu_{PGS} = 0 \quad \text{and} \quad sd = 1 \]
Polygenic Scores (PGSs): An Overview

• PGSs capture the combined additive genetic influence of SNPs across the entire genome on a specific trait/behavior in a single measure

\[ PGS_i = \sum_{j=1}^{k} \beta_j SNP_{ij} \]

\[ \mu_{PGS} = 0 \quad \text{and} \quad sd = 1 \]

• Whole genome PGSs
  • Include genetic association from across the entire genome, but eliminate the possibility to testing hypotheses related to specific biological pathways (Belsky and Israel, 2014)
Why/How to use PGSs

• Four possible uses
  • Nature net of Nurture
  • Nurture net of Nature
  • How Nurture modifies the effect(s) of Nature
  • How Nature modifies the effect(s) of Nurture
Why/How to use PGSs

• Four possible uses
  • Nature net of Nurture
  • Nurture net of Nature
  • How Nurture modifies the effect(s) of Nature
  • How Nature modifies the effect(s) of Nurture
Why/How to use PGSs

• Four possible uses
  • Nature net of Nurture
  • Nurture net of Nature (Control for genetic influences… Endogeneity problem)
  • How Nurture modifies the effect(s) of Nature
  • How Nature modifies the effect(s) of Nurture
Why/How to use PGSs

• Four possible uses
  • Nature net of Nurture
  • Nurture net of Nature
  • How Nurture modifies the effect(s) of Nature
  • How Nature modifies the effect(s) of Nurture
Why/How to use PGSs

• Four possible uses
  • Nature net of Nurture
  • Nurture net of Nature
  • How Nurture modifies the effect(s) of Nature
  • How Nature modifies the effect(s) of Nurture

Gene-Environment Correlations
Why/How to use PGSs

The Genetics of Success: How Single-Nucleotide Polymorphisms Associated With Educational Attainment Relate to Life-Course Development

Daniel W. Belsky¹,², Terrie E. Moffitt³,⁴,⁵,⁶, David L. Corcoran⁵, Benjamin Domingue⁷, HonaLee Harrington³, Sean Hogan⁶, Renate Houts², Sandhya Ramrakha⁸, Karen Sugden⁵, Benjamin S. Williams⁵, Richie Poulton⁸, and Avshalom Caspi³,⁴,⁵,⁶
The Genetics of Success: How Single-Nucleotide Polymorphisms Associated With Educational Attainment Relate to Life-Course Development

Abstract

A previous genome-wide association study (GWAS) of more than 100,000 individuals identified molecular-genetic predictors of educational attainment. We undertook in-depth life-course investigation of the polygenic score derived from this GWAS using the four-decade Dunedin Study (N = 918). There were five main findings. First, polygenic scores predicted adult economic outcomes even after accounting for educational attainments. Second, genes and environments were correlated: Children with higher polygenic scores were born into better-off homes. Third, children's polygenic scores predicted their adult outcomes even when analyses accounted for their social-class origins; social-mobility analysis showed that children with higher polygenic scores were more upwardly mobile than children with lower scores. Fourth, polygenic scores predicted behavior across the life course, from early acquisition of speech and reading skills through geographic mobility and mate choice and on to financial planning for retirement. Fifth, polygenic-score associations were mediated by psychological characteristics, including intelligence, self-control, and interpersonal skill. Effect sizes were small. Factors connecting DNA sequence with life outcomes may provide targets for interventions to promote population-wide positive development.
The Genetics of Success: How Single-Nucleotide Polymorphisms Associated With Educational Attainment Relate to Life-Course Development

Abstract
A previous genome-wide association study (GWAS) of more than 100,000 individuals identified molecular-genetic predictors of educational attainment. We undertook an in-depth life-course investigation of the polygenic score derived from this GWAS using the four-decade Dunedin Study (N = 918). There were five main findings. First, polygenic scores predicted adult economic outcomes even after accounting for educational attainments. Second, genes and environments were correlated: Children with higher polygenic scores were born into better-off homes. Third, children’s polygenic scores predicted their adult outcomes even when analyses accounted for their social-class origins. Social-mobility analysis showed that children with higher polygenic scores were more upwardly mobile than children with lower scores. Fourth, polygenic scores predicted behavior across the life course, from early acquisition of speech and reading skills through geographic mobility and mate choice and on to financial planning for retirement. Fifth, polygenic-score associations were mediated by psychological characteristics, including intelligence, self-control, and interpersonal skill. Effect sizes were small. Factors connecting DNA sequence with life outcomes may provide targets for interventions to promote population-wide positive development.
Why/How to use PGSs

Gene-Environment Correlations

Abstract
A previous genome-wide association study (GWAS) of more than 100,000 individuals identified molecular-genetic predictors of educational attainment. We undertook in-depth life-course investigation of the polygenic score derived from this GWAS using the four-decade Dunedin Study (N = 918). There were five main findings. First, polygenic scores predicted adult economic outcomes even after accounting for educational attainments. Second, polygenic scores were correlated: Children with higher polygenic scores were born into better-off homes. Third, children’s polygenic scores predicted their adult outcomes even when analyses accounted for their social-class origins; social-mobility analysis showed that children with higher polygenic scores were more upwardly mobile than children with lower scores. Fourth, polygenic scores predicted behavior across the life course, from early acquisition of speech and reading skills through geographic mobility and mate choice and on to financial planning for retirement. Fifth, polygenic-score associations were mediated by psychological characteristics, including intelligence, self-control, and interpersonal skill. Effect sizes were small. Factors connecting DNA sequence with life outcomes may provide targets for interventions to promote population-wide positive development.
Cardiovascular Health
- Coronary Artery Disease
- Myocardial Infarction
- Plasma Cortisol
- Low-density Lipoprotein Cholesterol
- High-density Lipoprotein Cholesterol
- Total Cholesterol
- Triglycerides

Anthropomorphic Traits
- Body-Mass Index
- Waist Circumference
- Waist-to-Hip Ratio
- Height

Reproductive Traits
- Age at Menarche
- Age at Menopause
- Number of Children Ever Born
- Age at First Birth
Add Health Polygenic Scores – Release 1

Tobacco use
- Current/Ever Smoker
- Number of Cigarettes per day

Mental Health / Personality
- Bipolar Disorder
- ADHD
- Major Depressive Disorder
- Schizophrenia
- Mental Health Cross Disorder
- Extraversion

Cognitive Health / Education
- Alzheimer’s Disease
- Educational Attainment
### Phenotype GWAS Ancestry Group(s)

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>GWAS Ancestry Group(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coronary Artery Disease</td>
<td>European</td>
</tr>
<tr>
<td>Myocardial Infarction</td>
<td>European, South Asian, East Asian</td>
</tr>
<tr>
<td>Plasma Cortisol</td>
<td>European</td>
</tr>
<tr>
<td>Low-density Lipoprotein Cholesterol</td>
<td>European</td>
</tr>
<tr>
<td>High-density Lipoprotein Cholesterol</td>
<td>European</td>
</tr>
<tr>
<td>Total Cholesterol</td>
<td>European</td>
</tr>
<tr>
<td>Triglycerides</td>
<td>European</td>
</tr>
<tr>
<td>Type II Diabetes (2012)</td>
<td>European</td>
</tr>
<tr>
<td>Type II Diabetes (2018)</td>
<td>European, East Asian, South Asian,</td>
</tr>
<tr>
<td></td>
<td>Mexican/Mexican-American</td>
</tr>
<tr>
<td>BMI</td>
<td>European</td>
</tr>
<tr>
<td>Waist Circumference</td>
<td>European</td>
</tr>
<tr>
<td>Waist-to-Hip Ratio</td>
<td>European</td>
</tr>
<tr>
<td>Height</td>
<td>European</td>
</tr>
<tr>
<td>Age at Menarche</td>
<td>European</td>
</tr>
<tr>
<td>Age at Menopause</td>
<td>European</td>
</tr>
<tr>
<td>Number of Children Ever Born</td>
<td>European</td>
</tr>
<tr>
<td>Age at First Birth</td>
<td>European</td>
</tr>
<tr>
<td>Ever/Current Smoker</td>
<td>European</td>
</tr>
<tr>
<td>Number of Cigarettes per day</td>
<td>European</td>
</tr>
<tr>
<td>Extraversion</td>
<td>European</td>
</tr>
<tr>
<td>Attention-deficit/hyperactivity Disorder (2010)</td>
<td>European</td>
</tr>
<tr>
<td>Attention-deficit/hyperactivity Disorder (2017)</td>
<td>European, Chinese</td>
</tr>
<tr>
<td>Bipolar Disorder</td>
<td>European</td>
</tr>
<tr>
<td>Major Depressive Disorder (2013)</td>
<td>European</td>
</tr>
<tr>
<td>Major Depressive Disorder (2018)</td>
<td>European</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>European, East Asian</td>
</tr>
<tr>
<td>Mental Health Cross Disorder</td>
<td>European</td>
</tr>
<tr>
<td>Alzheimer's Disease</td>
<td>European</td>
</tr>
<tr>
<td>Educational Attainment (2016)</td>
<td>European</td>
</tr>
<tr>
<td>Educational Attainment (2018)</td>
<td>European</td>
</tr>
</tbody>
</table>

PGSs for individuals not of the same ancestry group(s) as the GWAS from which summary statistics are retrieved may be less predictive (Martin et al. 2017; Ware et al. 2017).
Race/Ethnicity VS Genetic Ancestry

• While genetic ancestry and self-identified race/ethnicity are strongly correlated (0.89 in Add Health), they are two separate constructs.
  • Race and ethnicity are social constructs based on a multitude of factors, of which ancestry may be included depending on historical and societal differences in racialization (Omi and Winant 1994).
  • Genetic ancestry is defined using the first two principal components of the genotyped data.
Population Structure and Genetics
Add Health Genetic Ancestry Groups
Add Health Genetic Ancestry Groups
Add Health Genetic Ancestry Groups

Add Health Self-identified Race/Ethnicity

- Western Europe & Italy
- African ancestry (South West USA), Kenya, Nigeria
- Mexican ancestry (Los Angeles, CA)
- Gujarati Indians in Houston, TX
- Han Chinese (Beijing), Chinese (Den, CO), Japanese (Tokyo)
### Add Health Genetic Ancestry Groups

<table>
<thead>
<tr>
<th>Self-Identified Race/Ethnicity</th>
<th>European</th>
<th>African</th>
<th>East Asian</th>
<th>Hispanic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Hispanic White</td>
<td>5,644</td>
<td>5</td>
<td>0</td>
<td>105</td>
<td>5,754</td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>0</td>
<td>1,939</td>
<td>0</td>
<td>1</td>
<td>1,940</td>
</tr>
<tr>
<td>Native American</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>Asian</td>
<td>0</td>
<td>1</td>
<td>422</td>
<td>26</td>
<td>449</td>
</tr>
<tr>
<td>Hispanic</td>
<td>70</td>
<td>27</td>
<td>14</td>
<td>850</td>
<td>961</td>
</tr>
<tr>
<td>Total Sample Size</td>
<td>5,728</td>
<td>1,974</td>
<td>436</td>
<td>988</td>
<td>9,126</td>
</tr>
</tbody>
</table>

The table above represents the distribution of ancestry among different self-identified race/ethnicity groups. The data includes European, African, East Asian, Hispanic, and total sample sizes for each group. The total sample size for the entire dataset is 9,126.
Using Add Health PGSs

• Accounting for ancestry
  • Separate analyses by ancestry
  • Ancestry-specific principal components

• Ancestry-specific principal components
  • Randomized in sets of five
    • 1-5, 6-10, 11-15, 16-20
SOME RESULTS
BMI (Waves I – IV)

BMI from Clinical Visit
All Ancestry Groups

BMI from Self-Report
All Ancestry Groups
BMI (Waves I – IV)

**European Ancestry Only**

- +1 sd PGS (BMI)
- -1 sd PGS (BMI)

**European Ancestry VS All Other Ancestry Groups**

- +1 sd PGS (BMI)
- -1 sd PGS (BMI)
BMI (Waves I – IV)

BMI from Clinical Visit
All Ancestry Groups

BMI from Self-Report
All Ancestry Groups
Depression (Waves I – IV)

All Ancestry Groups

![Graph showing depression index over years for different ancestry groups.]

-1 sd PGS (MDD 2018)
+1 sd PGS (MDD 2018)
Depression (Waves I – IV)

European Ancestry Only

European Ancestry VS All Other Ancestry Groups
Word Recall (Wave 4)

All Ancestry Groups

[Graph showing word recall over recall periods for different ancestry groups]
Word Recall (Wave 4)

European Ancestry Only

European Ancestry VS All Other Ancestry Groups

[Diagram showing recall periods and ancestry groups with statistical measures.]
New Data Releases

Wave IV: Ambient Air Pollutants Data
(Released: July 20, 2018)

AmbientAirPollutants_OrderForm

The Ambient Air Pollutants data include 365 daily exposure estimates of the following ambient air pollutants for each Add Health study participant in Wave IV:

- Particulate Matter: PM2.5, PM10, SO4, NO3, NH4, organic carbon (OC), elemental carbon (EC)
- Fuels: AI, S, TI, OI, Mg, K, Mn, Na, and Cl
- Gases: O3, NO2, HNO3, HONO, H2O2, CO, and SO2; 2007 only: acetaldehyde, benzene, butadiene, ethanol, formaldehyde, and naphthalene

Parents Phase 2 Data
(Released: July 20, 2018)

ParentStudy_OrderForm

The parent data files contain social, demographic, behavioral, and health data collected in 2015-2017 on a probability sample of Add Health parents who were originally interviewed in 1995. Data for 2,241 Wave I parents, representing 2,247 Add Health sample members, are available. Additionally, 939 current spouse/partner interviews are available.

Polygenic Scores (PGS)
(Released: July 20, 2018)

PGS_OrderForm

Thirty constructed polygenic risk scores (PGS) are available for Add Health respondents who provided archival saliva samples for genetic testing at Wave IV (N = 9,129). Scores are available for coronary artery disease, myocardial infarction, plasma cortisol, LDL cholesterol, HDL cholesterol, total cholesterol, triglycerides, type 2 diabetes (2), BMI, waist circumference, waist-to-hip ratio, height, age at menarche, age at menopause, number of children, age at first birth, evercurrent smoker, number of cigarettes per day, extramarital, attention deficit disorder (2), bipolar disorder, major depressive disorder (2), schizophrenia, mental health cross disorder, Alzheimer’s disease, and educational attainment (2).
PGS Contracts

http://www.cpc.unc.edu/projects/addhealth/contracts/data-releases
Thank you!

braudt@unc.edu

Funding Acknowledgements: This presentation was funded by grant R01 HD073342 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD). Data for this study come from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from NICHD, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. Additional support was received from the Population Research Training grant (T32 HD007168) and the Population Research Infrastructure Program (P2C HD050924) awarded to the Carolina Population Center at The University of North Carolina at Chapel Hill by NICHD.