Genetics, economics, and education

Philipp Koellinger
University of Amsterdam
Road map

• Why social scientists should be excited about genetics
• The example of educational attainment
• The road ahead
## Heritability estimates for various traits

<table>
<thead>
<tr>
<th>Trait</th>
<th>( h^2_{\text{twin}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monogenetic disorders (e.g. Huntington disease)</td>
<td>≈ 100%</td>
</tr>
<tr>
<td>Height</td>
<td>≈ 90%</td>
</tr>
<tr>
<td>BMI</td>
<td>≈ 70%</td>
</tr>
<tr>
<td>Personality</td>
<td>≈ 50%</td>
</tr>
<tr>
<td>Educational attainment</td>
<td>≈ 40%</td>
</tr>
<tr>
<td>Self-employment</td>
<td>≈ 40%</td>
</tr>
<tr>
<td>Happiness</td>
<td>≈ 35%</td>
</tr>
<tr>
<td>Overconfidence</td>
<td>≈ 20%</td>
</tr>
<tr>
<td>Risk preferences</td>
<td>≈ 15%</td>
</tr>
</tbody>
</table>

**Notes:** Heritability is defined as \( h^2 = \frac{VG}{VP} \) where \( VP \) is the variance of a trait (“phenotype”) and \( VG \) is phenotypic variance due to genotypic variance. \( h^2_{\text{twin}} \) refers to estimates obtained from twin studies.
Why study the genetics of social-scientific variables?

1. Identifying causal pathways

2. Informing theory

3. Better understanding of environmental effects
   - Gene-environment interactions
   - Using polygenic scores as control variables

4. Links between behavior and health
Challenges in gene discovery

• Three well-known challenges to discover genetic associations:
  1. Theory is not a good guide *yet* for studying the effects of genes
     • >14,000 genes expressed in the brain
     • Most published findings from candidate gene studies could not be replicated
  2. Multiple testing: ~1 mio independent statistical tests in a GWAS
     • Genome-wide significance: $p < 5 \times 10^{-8}$
  3. Very small effect sizes of common variants
     • Need very large samples for GWAS and prediction

• Solutions:
  1. International collaboration of data providers and meta-analysis (e.g., the Social Science Genetic Association Consortium - SSGAC)
  2. Replication
GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

Rietveld et al., Science, 340, 1467-1471, 2013
GWAS on educational attainment

- 42 + 12 = 54 cohorts participated ($N \sim 125,000$)
- Two measures for educational attainment
  - Years of schooling ($EduYears$)
  - College degree yes/no ($College$)
- Strict quality controls
  - Data
    - MAF, call rate, imputation quality etc.
  - Population stratification
    - First 4 principle components from cohort-specific genetic relatedness matrix
    - Genomic control
- Many follow-up analyses
Three replicated SNPs

- Max $R^2 = 0.022\%$ for *EduYears*, ~2 months difference between the two homozygotes
- Max *Odds* = 0.904 for *College*, ~5.4% difference in chance to complete college between the two homozygotes
### Replicated loci from discovery stage

<table>
<thead>
<tr>
<th>SNP</th>
<th>College rs11584700</th>
<th>Beta</th>
<th>p-value</th>
<th>Replication rs4851266</th>
<th>Beta</th>
<th>p-value</th>
<th>Combined rs9320913</th>
<th>Beta</th>
<th>p-value</th>
</tr>
</thead>
</table>
|         |                   | -0.101| 2.07×10^{-9} | -0.101                | 0.001 |           | -0.101              | 8.24×10^{-12}
| EduYears|                   | 0.064 | 2.20×10^{-9} | 0.075                | 0.007 |           | 0.066               | 5.33×10^{-11}
|         |                   | 0.076 | 4.19×10^{-9} | 0.062                | 0.024 |           | 0.076               | 3.50×10^{-10}

- Largest effect sizes of replicated SNPs:
  - Max $R^2 = 0.022\%$ for EduYears, ~2 months difference between the two homozygotes
  - Max Odds = 0.904 for College, ~5.4\% difference in chance to complete college between the two homozygotes
Genetic prediction analyses on education

![Graph showing the relationship between R² and threshold p-value for SNP selection. The graph illustrates the predictive power for different educational outcomes such as EduYears score and College score, comparing QIMR and STR methods.](image-url)

- Results replicate in additional sample that used very stringent controls for population stratification
  - Sample: 23andMe
- Results hold in mixed linear models that control for entire genetic relationship matrix
  - Sample: STR
- Results are robust in within-family tests
  - Samples: STR, QIMR, FHS

→ Extremely unlikely that association results are driven by population stratification
The road ahead

1. Exponential growth in genetic data
   – Genetic data is getting cheap!
   – Decades of genetic discovery ahead
   – $R^2$ of polygenic scores will increase

2. Identifying causal pathways / informing theory
   – New methods / data / models

3. Better understanding of environmental effects
   – Gene-environment interactions $\rightarrow$ EA2.0
   – Using polygenic scores as control variables $\rightarrow$ EA2.0

4. Links between behavior and health
   – EA & dementia
   – EA & Schizophrenia