The Future of Social-Science Genomics

Daniel J. Benjamin
Center for Economic and Social Research,
Behavioral and Health Genomics Center, and Economics Department
University of Southern California

David Cesarini
Economics Department and Center for Experimental Social Science
New York University

Patrick Turley
Massachusetts General Hospital and Harvard-MIT Broad Institute

RSF Summer Institute in Social-Science Genomics • 20 June 2019

Some Logistics

- No more access to Add Health data.
- Breakfast on Friday starting 7:30am; bus for LAX departs at 8:30am sharp.
- You have received reimbursement instructions from us.
- Please fill out the anonymous feedback forms you will receive from RSF.
- Please let others know about the camp!
- Activity tonight! Meet in lobby at 6:40pm.

Outline

1. ***Some Key Ideas***
2. The New Social-Science Genomics
3. The Next Five Years
4. What Research Makes Sense?
5. The Job Market
6. Wrap Up

Genetic Effects

- Definition: The treatment effect from changing the genotype at conception.
  - The additive genetic effect is the average treatment effect (averaged over any dominance, GxG, and GxE).
- Therefore, genetic effects can operate through the environment.
  - Any such endogenous environmental effects are not part of the G-E correlation that contributes additional phenotypic variance.

Heritability

- High heritability does not imply that (endogenous) environmental factors are unimportant. (Jencks critique)
- Heritability is *not* an index of policy effectiveness. (Goldberger critique)
- Without data on other family members, twin studies estimate some mix of additive and other genetic variance components.
- Heritability is an upper bound on the predictive power of polygenic scores.

Environmental Effects

- Definition: The treatment effect from modifying the environment.
- Unless the environment is exogenous, it is difficult to interpret apparent GxE effects. E.g.:
  - If E is caused by G, could be just a G effect.
  - If E is correlated with G, could be a GxG effect.

Additive Genetic Effects

- The additive genetic component is the best linear predictor of the phenotype.
  - It captures the sum of all average genetic effects (averaged over dominance, GxG, and GxE).
  - SNP-based heritability estimators (e.g., GREML, LD Score Regression, RDR) estimate the contribution of the additive genetic component to the phenotype for the measured SNPs.

The Three Big Problems

- Population stratification
  - The biggest single confound.
  - Controlling for self-reported race is not sufficient.
- Multiple hypothesis testing
  - Many variants.
  - Many potential sample splits.
  - Many potential GxG effects.
  - Many potential GxE effects.
  - Many possible specifications.

- Low Power
  - The "4th Law": individual variants generally have small effects on behavioral phenotypes.
  - With low power, results are uninformative—whether or not they are significant.
  - With low power, significant results run a high risk of wrong signs and exaggerated effect sizes (winner's curse).

Outline

1. Quick Review
2. ***The New Social-Science Genomics***
3. The Next Five Years
4. What Research Makes Sense?
5. The Job Market
6. Wrap Up

The New Social-Science Genomics

- Dramatic decline in cost of genotyping → rapidly growing availability of genome-wide data.
- Large-scale GWAS:
  - Explosion in the number of identified loci.
  - Polygenic scores.
- Beginning to see applications in the social sciences.

Some Emerging Applications

- GxE.
- Polygenic scores as controls, in tests of balance, or to study mediation.
- Heritability and genetic correlation from genome-wide data.
- Estimating genetic nurture or other G-E corr.
- Sibling interactions.
- Mendelian randomization.
- Assortative mating.
- Understanding evolutionary history.

Better Research Practices

- Many results in social-science genetics have not been robust.
  - Has led to much debate and not enough cumulative knowledge.
- Rising standards for empirical work (along with many other sciences).
  - Recognition of realistic effect sizes.
  - Ex ante power calculations.
  - Awareness about multiple hypothesis testing and publication bias.
  - Assessing Bayesian credibility of findings.
  - Often, pre-registration of analysis plans.

Taking Ethical Responsibility

- Given bad history, need to re-build trust in social-science genomics research.
- Balancing risks of data breach and re-identifiability with value of data sharing.
- Informed consent clear about purposes of the research.
- Avoid hype, and emphasize limitations.
- Actively communicate what findings do and do not mean.

Outline

1. Quick Review
2. The New Social-Science Genomics
3. ***The Next Five Years***
4. What Research Makes Sense?
5. The Job Market
6. Wrap Up

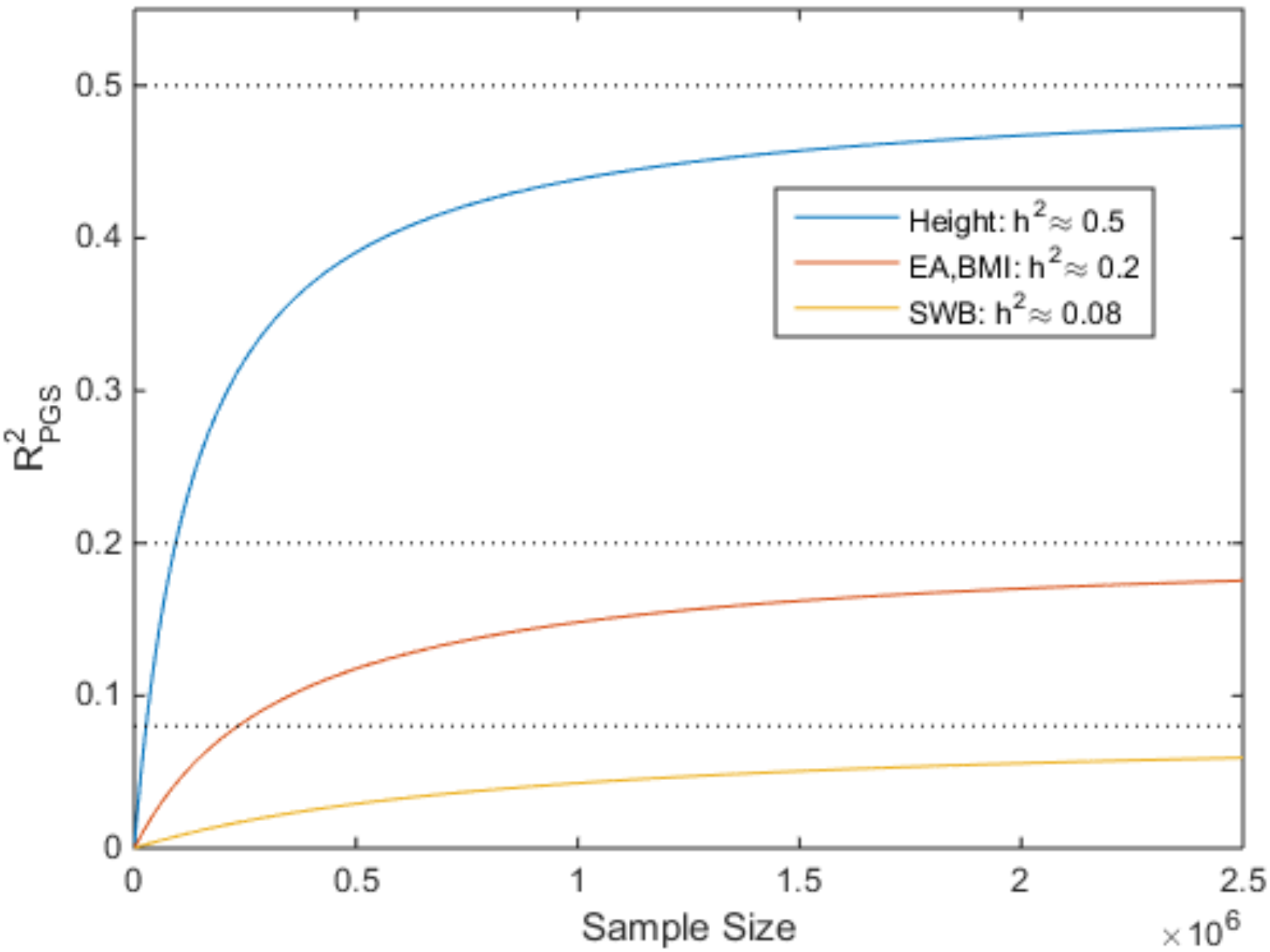
Ever More Datasets With Genome-Wide Data

- Add Health ($N \approx 6{,}500$)
- Health and Retirement Study ($N \approx 20{,}000$)
- Wisconsin Longitudinal Study ($N \approx 8{,}000$)
- UK Biobank ($N \approx 500{,}000$)
- Understanding Society ($N \approx 10{,}000$)
- Avon Longitudinal Study of Parents and Children (ALSPAC) ($N \approx 17{,}000$)
- http://www.thessgac.org/#!participating-cohorts/u23gj

More Discoveries

- With growing sample sizes, explosion in number SNPs identified.
- Improved understanding of biological and behavioral mechanisms.
- More application of new methods (such as LD Score regression, MR-Egger regr.).
- Development of new methods.
  - Esp. take advantage of newly available large, *individual-level* datasets (not meta-analysis).

Theoretical Projection for R^2_{PGS}



More Phenotypes

- GWAS on behavioral phenotypes has been limited by data availability
  - EA, SWB, neuroticism, fertility, risk tolerance.
- Many additional interesting phenotypes!
- Medical geneticists unlikely to measure many of these without social scientists.

Outline

1. Quick Review
2. The New Social-Science Genomics
3. The Next Five Years
4. ***What Research Makes Sense?***
5. The Job Market
6. Wrap Up

Questions You Should Always Ask

- Is the analysis well-powered?
- How have you dealt with multiple hypothesis testing?
- What about population stratification?
- How much winner's curse is there?
- How should I interpret "genetic effects"?
- (GxE) Is E exogenous?
- (MR) Does the exclusion restriction (or InSIDE assumption) hold?

PGS vs. Specific Variants

- Given 4th Law, often makes sense to focus on collections of variants, or on PGS.
- Probably impossible to pin down one specific biological mechanism.
- But nonetheless useful in many settings.

- Use well-understood specific variants if need to know mechanism:
  - Incorporating genetic variants into structural models.
  - Mendelian randomization.
  - GxE.
- Examples:
  - Smoking and Mr. Big: $R^2 \approx 0.5\%$.
  - BMI and FTO: $R^2 \approx 0.3\%$.
  - Huntington's disease, eye color, etc.
- Tend to be variants with largest, most biologically direct effects.

Should I Collect My Own Data?

- Probably—can always freeze and genotype/sequence in the future.
- Many great research questions can be answered with large, existing datasets.
- But polygenic scores and ancestry estimates can be studied in small samples:

Power at $\alpha = 0.05$

	$N = 50$	$N = 75$	$N = 100$	$N = 200$	$N = 400$	$N = 900$
$R^2 = 2\%$	17%	23%	29%	52%	81%	99%
$R^2 = 11\%$	68%	84%	93%	100%	100%	100%

What Kind of Genotyping?

- Candidate gene
  - Many tagged by SNPs in genome-wide data.
  - These days, only necessary for complex structural genetic variants.

- SNP chip data (roughly $40/person)
  - Captures most common variation.
  - Allows calculating PCs and polygenic scores.
  - Doesn't tag very rare variants.

- Low-density sequencing (roughly $40/person)
  - Measures many more SNPs than SNP chip.
  - But won't tag well all the SNPs on a SNP chip.
  - Better coverage for non-European ancestries.

- Sequencing data (roughly $500/person)
  - Captures virtually all genetic variation.

Outline

1. Quick Review
2. The New Social-Science Genomics
3. The Next Five Years
4. What Research Makes Sense?
5. ***The Job Market***
6. Wrap Up

Economics

- No ideological opposition to genetics, but want to see relevance for economics.
- Econ job market candidates in genoeconomics (that we know of):
  - David Cesarini (2010, MIT) → NYU.
  - Jonathan Beauchamp (2011, Harvard) → postdoc → Toronto.
  - Justin Cook (2012, LSU) → postdoc → UC Merced.
  - Pietro Biroli (2015, Chicago) → Zurich.
  - Lauren Schmitz (2015, New School) → postdoc → UW-M.
- A thinner market is always riskier.

Sociology

Dalton Conley: “Of all the social sciences, my sense is that both the interest in and resistance to genetic integration is the highest in sociology for ideological reasons. Sociologists are wary of inequality-rationalizing consequences of genetic analysis. Indeed, sociology was founded against a backdrop of 19th Century biological theories of society such as Herbert Spencer's organicism. That said, I do not conclude that anyone doing G or GxE work in sociology is a dead duck on the job market.”

“I think that a number of sociologists and sociology departments are embracing this torrent of genotypic data as they become increasingly common in survey studies that we often analyze. The very inclusion of genetic data into WLS, Add Health, and HRS, just to name a few, suggest that sociologists are not all hostile to this sort of analysis. While the N is still small, my observations detect a paradox: The top departments are the most receptive to this sort of analysis, because they are the most positivist and least ideological.”

“But the students from these departments (with a couple exceptions like UNC and Colorado) are less likely to do genetic work in their graduate work because they are risk averse to damaging their job market prospects. But those interested in the topic from lower ranked departments are more willing to roll the dice and take a chance. I have seen this pay off. I should also say that there are a number of post-docs -- at Princeton, Michigan and UNC, for instance -- that specifically look for people doing biomarker work.”

“That said, many people do the socio-genomics as side work and do a more straight sociology dissertation. But I always believe that you cannot be strategic about what you study. You need to study what you are passionate about because you will be the most innovative about what you are obsessed about. You will also work harder. So if this is what you are passionate about, go for it!”

Some Common Themes

- The career risks of social-science genomics are like those of any new field.
  - Thin market.
  - Unanticipated negative shocks (to the field, to the job market) could have a big effect.
- Also appears to earn a return.
- Our (biased) view: Obvious that growth in social-science genomics research will continue to be exponential.
- You are in a unique position to be an early leader.

Social-Science Genomics Listserv

- We maintain a listserv for former RSF SI participants and other members of the social-science genomics community.
- Emails primarily contain open job postings.
  - Also notifications for conferences and workshops.
  - Anyone who wants to be added should email Chelsea Watson <cwatson.usc@gmail.com>.
- If you have open positions at your institution that you'd like us to post, let us know!

Outline

1. Quick Review
2. The New Social-Science Genomics
3. The Next Five Years
4. What Research Makes Sense?
5. The Job Market
6. ***Wrap Up***

Thank-You's

- Paige Harden
- Dalton Conley
- Andrea Byrnes
- Michelle Meyer
- Aysu Okbay
- Raymond Walters
- Alicia Martin
- James Lee
- Loïc Yengo
- Matt Keller
- Alex Young
- Jonathan Beauchamp
- Lauren Schmitz
- Tõnu Esko
- Silvia Barcellos
- Phil Koellinger
- Hui Li
- Rosie Li

Thank-You's, continued

- Samantha Cherney
- Best Western Plus Pepper Tree Inn
- Maribel Estrada
- RSF
- YOU!