Genetic Architecture of Economic Preferences

David Cesarini

Conference on the Biological Basis of Economics, May 2012

Collaborators

Dan Benjamin, Cornell
Jonathan Beauchamp, Harvard University
Christopher Chabris, Union College
Magnus Johannesson, Stockholm School of Economics
Philipp Koellinger, Erasmus School of Economics
David Laibson, Harvard University
Matthijs van der Loos, Erasmus School of Economics

Some Motivating Facts

- In the last few years, there have been rapid, continual advances in understanding of human genetics.
- The cost of genotyping is falling faster than Moore's Law.
- Some large-scale social surveys are now collecting genetic data on respondents.
- If data are there, economists will analyze it.
- How, if at all, can genetic data contribute to the social sciences, and how quickly should we expect that these goals will be realized?

Overview

- Behavior and Molecular Genetics
- Promises of molecular genetic data
- Challenges
- Some Productive Ways Forward

Molecular Genetics Basics

- Human DNA is a sequence of ~3 billion nucleotides (spread across 23 chromosomes).
- This sequence 20,000-25,000 subsequences called genes.
- Genes provide instructions for building proteins that in turn affect body function.
- At the vast majority of locations, there is no variation in nucleotides across individuals.

Molecular Genetics Basics (cont'd)

- Single-nucleotide polymorphisms (SNPs): The $<1\%$ of nucleotides (~20 million) where individuals differ. (There are also other types of variation.)
- A vast majority of SNPs are biallelic: there are only 2 possible nucleotides.
- From each parent, may inherit either allele; SNP unaffected by which received from whom.
- Genotype for each SNP: #minor alleles (0,1,2).

Genetic Effects

Let i index individuals; j index the causal SNPs. Let y_i denote some outcome of interest. The simplest model of genetic effects:

$$y_i = \mu + \sum \beta_j x_{ij} + \epsilon_i$$

- x_{ij} : genotype$\in \{0, 1, 2\}$ of person i for SNP j.
- β_j : causal effect of SNP j.
- ϵ_i : causal effect of residual factors.

Genetic Effects

$$y_i = \mu + \sum \beta_j x_{ij} + \epsilon_i$$

- β_j is the treatment effect from changing an individual's SNP at conception. Can be done in animals; hypothetical in humans.
- Now established that there is an effect of at least one SNP in the gene FTO on body weight. In a sample of ~40,000, Frayling et al. (2007) found that people with 2 major alleles weigh 3 kg more than people with 2 minor alleles.
- One proposed mechanism is preference for energy-rich foods (Cecil et al., 2008).

Interpreting Genetic Effects

$$y_i = \mu + \sum \beta_j x_{ij} + \epsilon_i$$

- ϵ_i is often called the "environmental" effect, but this is imprecise and potentially misleading.
- E.g., the component of caloric intake induced by variation in FTO is not part of ϵ_i.
- It captures environmental factors that are not endogenous to genotype (Jencks, 1980).
    - Consider the thought experiment of separating a pair of MZ twins at birth and randomly assigning them to families. Assume similarity in uterine environments can be ignored.
    - Any measured similarity in outcome can ultimately be traced to their shared genes.

Extensions of the Simple Model

$$y_i = \mu + \sum \beta_j x_{ij} + \epsilon_i$$

- "Dominance effects": the effect of x_{ij} on the outcome is non-linear.
- "Gene-gene interaction": x_{ij} interacts with $x_{ij'}$ in affecting the outcome.
- "Gene-environment interaction": x_{ij} interacts with ϵ_i in affecting the outcome.
- E.g., the effect of FTO on body weight is strongly affected by birth cohort (Rosenquist et al., 2012).

Behavior Genetics and Molecular Genetics

$$y_i = \mu + \underbrace{\sum \beta_j x_{ij}}_{\equiv g_i} + \epsilon_i$$

- g_i is individual i's genetic endowment, the effect of genes taken as a whole.
- Behavior genetics pre-dates availability of data on genotypes.
- Treats g_i as a latent variable and draws inferences about it by contrasting the similarity in outcomes of different relatives.

Extending the Simple Model (cont'd)

$$y_i = \mu + \underbrace{\sum \beta_j x_{ij}}_{\equiv g_i} + \epsilon_i$$

- Much of behavior genetics is about estimating heritability $Var(g_i)/Var(y_i)$.
- If g_i is independent of ϵ_i, then heritability is the population R^2 of the regression from y on all J SNPs.
- Can estimate $Var(g_i)/Var(y_i)$ by contrasting the resemblance of different types of relatives.

Heritability

- Economic Outcomes
  - Educational attainment ∼40% (Behrman et al., 1975; Miller et al., 2001; Scarr and Weinberg, 1994; Lichtenstein et al., 1992)
  - Income ∼30% (Björklund, Jäntti and Solon, 2005; Sacerdote, 2007; Taubman, 1976)
- Economic Preferences
  - Risk preferences ∼20%(Cesarini et al., 2009; Zhong et al. 2009; Zyphur et al. 2009)
  - Bargaining behavior, altruism and trust ∼20% (Wallace et al., 2007; Cesarini et al., 2008)
- Economic Behaviors
  - Financial decision-making ∼30% (Barnea et al., 2010; Cesarini et al, 2010)
  - Susceptibility to decision-making anomalies ∼30%(Cesarini et al., 2011)

Heritability (cont'd)

- Compared to other traits (e.g., height, personality), the heritabilities of economic phenotypes are lower, often ~30-40%.
- These differences are diminished when measurement error/transitory variance is accounted for.
  - MZ correlation in income rises to 0.55 when smoothing out transitory fluctuations by taking a 20 year average (Benjamin et al., forthcoming).
  - MZ correlation in a measure of risk aversion rises to 0.70 when adjusting for low reliability (Beauchamp et al., 2011).

Heritability and Malleability

- "[These results] really tell the [Royal] Commission [on the Distribution of Income and Wealth] that they might as well pack up"

(Hans Eysenck, quoted in the Times of London)

(Goldberger, 1979)

Heritability and Malleability

- "[These results] really tell the [Royal] Commission [on the Distribution of Income and Wealth] that they might as well pack up"

(Hans Eysenck, quoted in the Times of London)

- "A powerful intellect was at work. In the same vein, if it were shown that a large proportion of the variance in eyesight were due to genetic causes, then the Royal Commission on the Distribution of Eyeglasses might as well pack up. And if it were shown that most of the variation in rainfall is due to natural causes, then the Royal Commission on the Distribution of Umbrellas could pack up too."

(Goldberger, 1979)

Why Care?

- Heritability quantifies how much i's outcome can be predicted from x_{ij}'s if β_j's were known.
  - Such prediction from genetic data will become increasingly practically relevant.
- All else equal, higher heritabilities imply greater potential for genetic factors to confound estimates of environmental effects.
  - E.g., parental income on children's outcomes.
- Provides guidance regarding which outcomes are more promising targets for gene discovery.
- Heritabilities of income, etc., are facts that may constrain the set of plausible theories regarding heterogeneity.
  - High heritabilities challenge blank-slate theories (Pinker, 2002).

Promise of Molecular Genetic Data

- Direct measures of latent parameters (preferences, abilities)
  - E.g., FTO genotype may be a measure of food preference. Some other gene may affect the production function of body weight from calorie consumption.
  - Could then use as variables of interest or controls.
- Biological mechanisms for social behavior.
  - Could help test existing hypotheses (oxytocin and trust).
  - Could suggest new hypotheses.
    - E.g., how to decompose crude concepts like "risk aversion" and "patience."
  - In medicine, genetic associations have led to discoveries about new pathways for age-related macular degeneration and for Crohn's disease.

Promise of Molecular Genetic Data

- Genes in Empirical Work
  - E.g., in epidemiology: effect of higher levels of alcohol consumption on blood pressure, using SNPs that cause variation in alcohol metabolism. (Chen, Davey Smith, Harbord, and Lewis, 2008)
  - Could, very mundanely, be used as control variables to lower the unexplained variation and reduce the standard errors on the coefficients of interest.
- Targeting social-science interventions
  - Much as envisioned for medical interventions.
  - E.g., children with dyslexia-susceptibility genotypes could be taught to read differently from an early age.
  - For adults, can often directly measure realized preferences and abilities, so targeting most likely to be done by parents.

Genetic Architecture

- The extent to which these promises of molecular genetic data will be fulfilled hinges crucially on the molecular genetic architecture of the traits in question (Benjamin, 2010; Beauchamp et al., 2011).
  - This architecture is the result of evolutionary forces, including mutation, selection and drift.
- Molecular genetic architecture: joint distribution of effect sizes and allele frequencies in a population.
  - Biological mechanisms typically requires an ability to identify individual variants or genes.

Genetic Architecture

- The extent to which these promises of molecular genetic data will be fulfilled hinges crucially on the molecular genetic architecture of the traits in question (Benjamin, 2010; Beauchamp et al., 2011).
  - This architecture is the result of evolutionary forces, including mutation, selection and drift.
- Molecular genetic architecture: joint distribution of effect sizes and allele frequencies in a population.
  - Biological mechanisms typically requires an ability to identify individual variants or genes.
  - For prediction, interventions and OVB, predictability may be enough.

Genetic Architecture

- The extent to which these promises of molecular genetic data will be fulfilled hinges crucially on the molecular genetic architecture of the traits in question (Benjamin, 2010; Beauchamp et al., 2011).
  - This architecture is the result of evolutionary forces, including mutation, selection and drift.
- Molecular genetic architecture: joint distribution of effect sizes and allele frequencies in a population.
  - Biological mechanisms typically requires an ability to identify individual variants or genes.
  - For prediction, interventions and OVB, predictability may be enough.
  - Genes as instrumental variables (Davey-Smith, 2002; Ding et al., 2007) requires detailed knowledge of the pathways through which the genetic variants affect the outcome of interest.

Roadmap

- Inferential Challenge.
- Describe our intellectual journey
  - Chabris et al. (2011), *Psychological Science*.
  - Beauchamp et al. (2011), *Journal of Economic Perspectives*.
  - Propose an interpretation of these patterns of results.
  - Benjamin et al. (2012), *PNAS*.

Two Approaches to Genetic Association

- Candidate gene studies - type a set of markers that have some believed or known biological function and test them for association with the trait of interest. This is by far the most common approach in economics and social science.
- GWAS studies - atheoretical mining of the genome, insisting on a very stringent significance threshold.

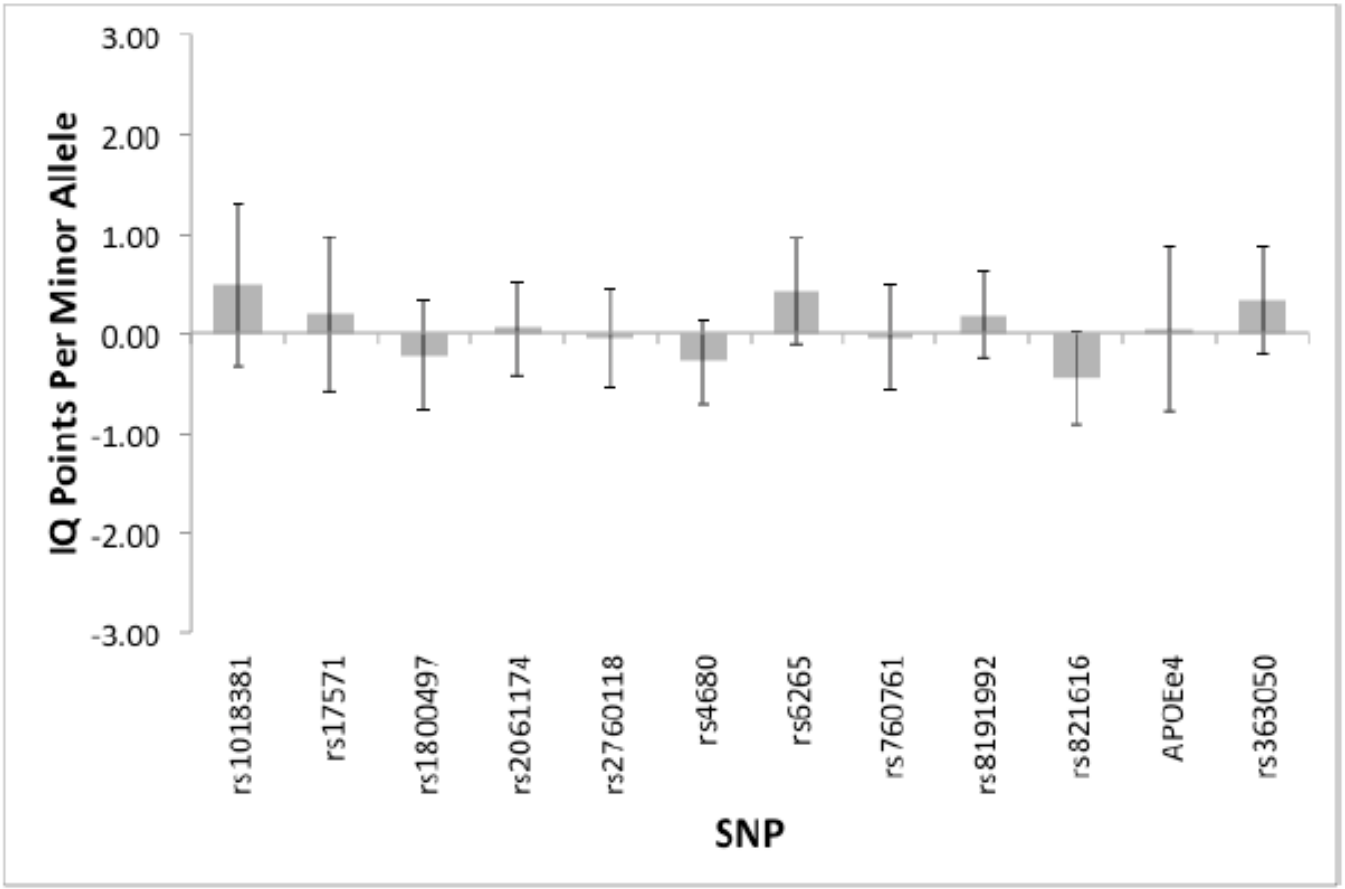
Two Tales

- Genetics of Cognitive Ability (Chabris et al., 2011).
- Genome Wide Association Study of Educational Attainment (Beauchamp et al., 2011).

Study 1 "Most Published Associations with General Intelligence Are False Positive" (Chabris et al., 2011)

- Use three different datasets, the WLS, the Framingham Heart Study and a Swedish sample of twins, to try to replicate the associations between 13 SNPs with published g associations.
- Selected the candidate SNPs from an authoritative review (Payton, 2008).
- Our total sample is just shy of 10,000 individuals (so power is excellent).
- In none of the samples were we able to replicate any of the associations reported in the literature.
- Even worse, we cannot reject the hypothesis that the SNPs jointly have any explanatory power for g.

Meta-Analyses



Study 2 Molecular Genetics and Economics (Beauchamp et al., 2011)

- Can we find genetic markers that predict educational attainment?
- Data comes from Framingham Heart Study
- Of the 14,428 participants, 9,237 have been genotyped. Original Cohort: 29%, Offspring Cohort: 73%, Third Generation Cohort: 95%
- "Years of education" constructed using survey respones.
- Final sample with genetic, educational & demographic data: N=8,496

Methods

In a GWAS, tens of thousands of regressions are run, one for each SNP in the array that passes quality control filters,

$$y = \mu + \beta_j x_j + PC \cdot \beta_2 + X \cdot \beta_3 + \varepsilon,$$

where *Edu* is years of education, x_j is the number of copies of the minor allele (0, 1, or 2) and the vector X includes a cubic of age, gender, their interactions and the first ten princincipal components of the variance covariance matrix of the genotypic data.

Complications

- Genotyping Errors
- Population Stratification.
- Multiple Hypothesis Testing.
- Family Based.

Framingham Results

SNP (Chromosome)	$\hat{\beta}$	p-value	Bonf.	Sample	M.A.
rs11758688 (6)	-0.253	$2.97 \cdot 10^{-7}$	0.107	7572	T
rs12527415 (6)	-0.253	$3.03 \cdot 10^{-7}$	0.109	7570	T
rs17365411 (2)	0.260	$3.73 \cdot 10^{-7}$	0.134	7559	C
rs7655595 (4)	-0.266	$3.99 \cdot 10^{-7}$	0.144	7486	G
rs17350845 (1)	-0.291	$6.22 \cdot 10^{-7}$	0.224	7415	C
rs12691894 (2)	-0.246	$6.67 \cdot 10^{-7}$	0.240	7572	G
rs9646799 (2)	0.271	$7.41 \cdot 10^{-7}$	0.267	7478	T
rs11722767 (4)	-0.257	$7.77 \cdot 10^{-7}$	0.280	7574	C
rs10947091 (6)	-0.245	$9.03 \cdot 10^{-7}$	0.325	7574	T
rs6536456 (4)	0.230	$1.32 \cdot 10^{-6}$	0.474	7513	C

Discussion

- 4 of the 7 most "significant" SNPs are in or near known genes
- 12 of 20 reported SNPs are in or near known genes
- Our top 2 hits are close to *IER3* (Immediate Early Response 3) gene, involved in apoptosis (regulation of cell death); apoptosis is believed to have an important impact on cognitive development (Arora et al, 2009)
- 3 of these are in the *MAPKAP2* gene, which encodes a protein involved in stress and inflammatory responses, among others; hypothesized link to neuronal death and regeneration (Harper et al., 2001).

Replication in the Rotterdam Study

- The Rotterdam Study also comprises three cohort.
- The initial cohort started in 1990 with 7,983 men and women aged 55 years and over.
- Two more cohorts have subsequently been recruited.
  - None of the top twenty hits replicated and 11 had the "wrong" sign.

Possible Interpretations

- False positive due to multiple hypothesis testing.
- Population stratification.
- True treatment effect local to environmental circumstances in Framingham.
- True treatment effect local to Framingham's gene-pool.

Are We Alone?

Editorial Policy on Candidate Gene Association and Candidate Gene-by-Environment Interaction Studies of Complex Traits

John K. Hewitt

The literature on candidate gene associations is full of reports that have not stood up to rigorous replication. This is the case both for straightforward main effects and for candidate gene-by-environment interactions (Duncan and Keller 2011). As a result, the psychiatric and behavior genetics literature has become confusing and it now seems likely that many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge. The reasons for this are complex, but include the likelihood that effect sizes of individual polymorphisms are small, that studies have therefore been underpowered, and that multiple hypotheses and methods of analysis have been explored; these conditions will result in an unacceptably high proportion of false findings (Ioannidis 2005).

Bayesian Calculation (Benjamin et al., forthcoming)

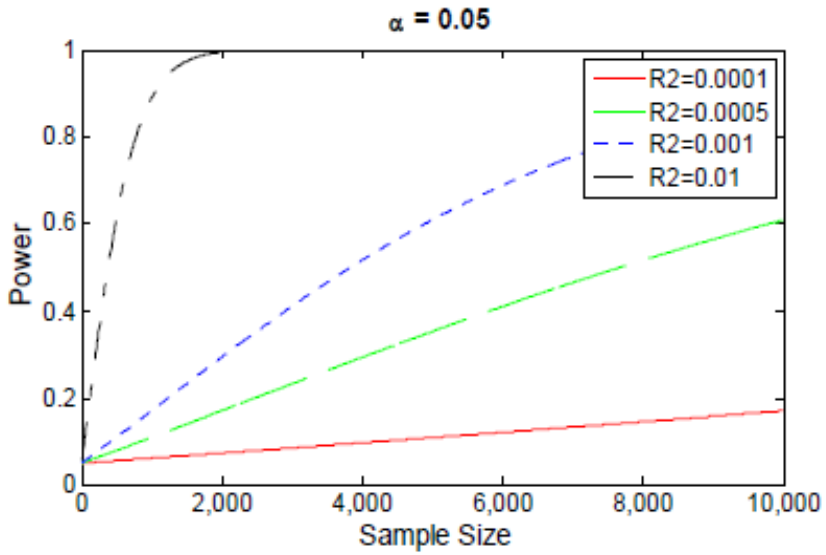
- Two alleles: High and Low.
- Equal frequency of High and Low.
- Phenotype distributed normally.
- Two states of the world: true association or not.
- If associated, $R^2 = .1\%$ (large for behavior).
- Sample size for 80% power:
- Now suppose significant association at $\alpha = 0.05$.

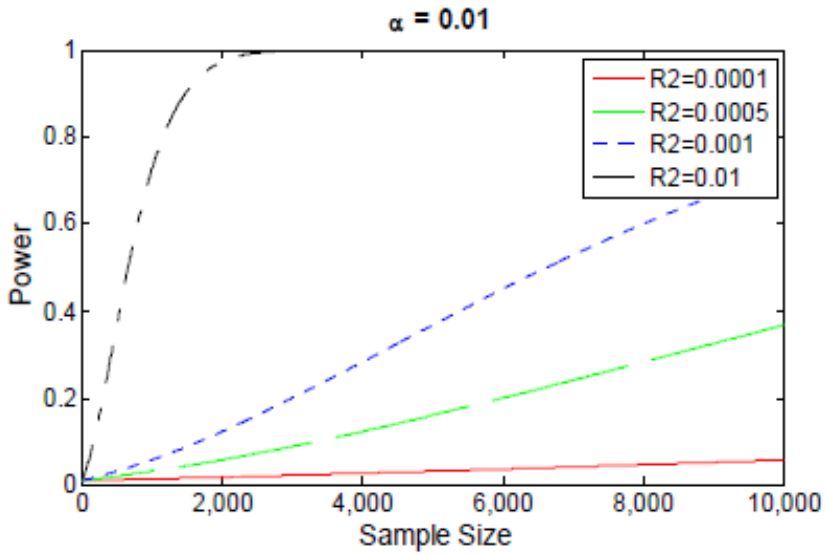
Posteriors as a Function of Sample Size and Priors

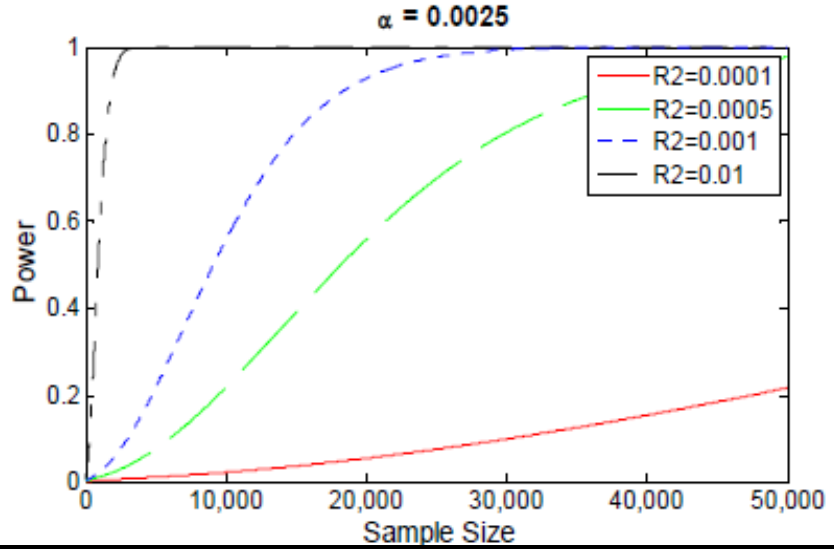
		Sample Size		
		n=100	n=5000	n=30,000
Prior	.01%	.01%	.12%	.20%
	1%	1%	11%	17%
	10%	12%	58%	69%

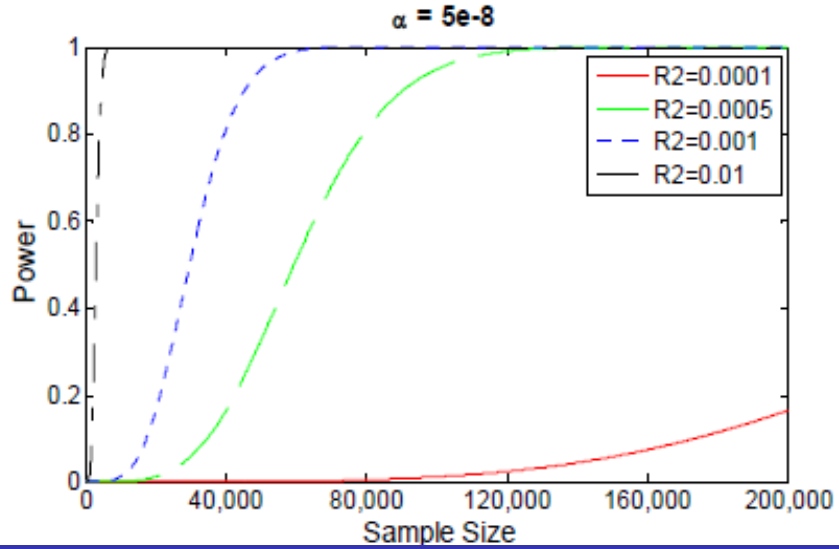
Note: Posteriors computed using Bayes' law.

Power Graphs









More on the Power Problem

- Low power is due to small effect sizes and the problem is likely exacerbated by
  - Multiple hypothesis testing
  - Publication bias.
- Evidence for low power:
  - Many published associations not reproducible (Ioannidis, 2007), especially so in the social sciences (Beauchamp et al., 2011; Benjamin et al., 2011).
  - Associations are especially likely to fail to replicate when the original sample was small.

Constructive Response 1

- Currently forming a consortium (with Dan Benjamin and Phil Koellinger), pooling data from several large samples, hopefully with a final sample exceeding 100,000 individuals.
- Modern studies use huge samples and impose extremely strict confidence levels.
- As an empirical matter, the results that emerge out of such efforts tend to be much more reliable.
- Our effort is embedded in the CHARGE consortium and 41 different cohorts are enrolled.
- Preliminary results - a handful of SNPs with $p < 5 \cdot 10^{-8}$.

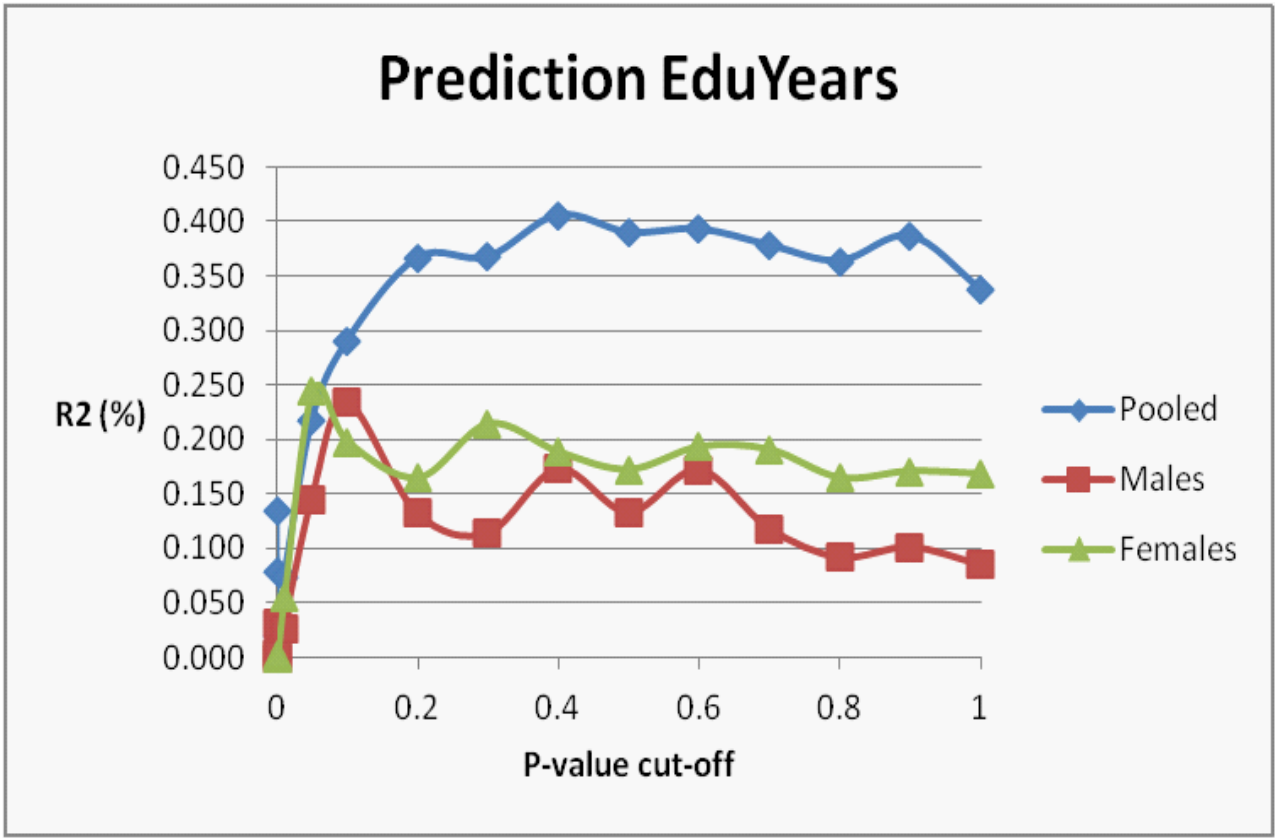
Risk Prediction

- The basic insight behind polygenic risk prediction (e.g., Purcell et al., 2009) is even when effect sizes are small, it may still be possible to make statistically efficient use of the joint predictive power of a large number of SNPs.

Risk Prediction

- Construct a genetic risk score by forming a discovery sample (80%) and a validation sample (20%).
- Take the estimated regression coefficients from the discovery sample and form a genetic score $X\hat{\beta}$.
- Do this for a pruned sample of ~100,000 SNPs which are in approximately uncorrelated.
- Correlate $X\hat{\beta}$ with the phenotype in the validation sample.
- Can set some regression coefficients to 0 if they are estimated with too much imprecision.

Prediction



Discovery sample N = 94,775 (61% female)

Prediction sample N = 6,774 (52% female)

Constructive Response 2

- Recognize that with presently attainable sample size it may not be feasible to detect individual marker associations with most complex traits.
- Use other techniques more suitable for complex traits in samples with comprehensively genotyped subjects
    - GREML analyses (Yang et al., 2010)

Constructive Response 2

- Recognize that with presently attainable sample size it may not be feasible to detect individual marker associations with most complex traits.
- Use other techniques more suitable for complex traits in samples with comprehensively genotyped subjects
    - GREML analyses (Yang et al., 2010)

Benjamin et al., 2012 "Genetic Architecture of Economic and Political Preferences"

- Use the method of Yang et al. (2001) - GREML - for estimating the proportion of variance explained jointly by all the SNPs measured in a GWAS.
- Carry out prediction.
- Conduct GWAS analysis.

GREML: Key Identifying Assumption

- Idea is to see how the correlation in phenotype between pairs of individuals relates to the genetic distance between those individuals.
- Among individuals who are unrelated—i.e., distantly related, since all humans are related to some extent—environmental factors are uncorrelated with differences in the degree of genetic relatedness.
- We should expect the estimated relationship between phenotype and genetic relatedness will be attenuated because relatedness is measured imperfectly; the common SNPs typed on the genotyping chip capture may not be perfectly representative of the causal variants (Yang et al., 2010; Visscher, Yang and Goodard, 2010).

GREML: Interpretation

- GREML estimates are a lower bound of narrow heritability.
- Output can be interpreted as the ultimate predictive value that can be obtained from dense SNP data.
- Can test for diffuse effects by checking whether longer chromosomes explain more variation.
- Applying the method, Yang et al. (2010) found that the measured SNPs could account for 45% of the variance in human height (missing heritability).
- Davies et al. (2011) apply the method to cognitive ability and obtained estimates in the 40-50% range.
- Used the GCTA software (Yang et al., 2011) to estimate "heritability" of some economic and political phenotypes.

SALTY

- The SALTY survey, administered in 2010, contains an entire section dedicated to measuring economic behaviors, attitudes and outcomes.
- Respondents can be matched to other administrative data.
- Subjects born between 1943 and 1958.
- The survey generated a total of 11,743 responses, a response rate of ~50%. Finally, 800 people were asked to complete the survey twice.
- Approximately 4,000 SALTY respondents have been comprehensively genotyped as part of the TwinGene sample.

Economic Preferences

- Risk - Questions from Barsky et al. (1997) and Dohmen et al. (2006).
- Trust - Questions from World Value Survey.
- Fairness - Questions from World Kahneman Knatsch and Thaler (1986)
- Discounting - Three Questions Comparing Immediate to Delayed Payoffs.

Political Attitudes

- Derived from a factor analysis of a 34 item battery of policy proposals.
- Results suggest five distinct factors: attitudes toward immigration, economic policy, environmentalism, feminism and international affair.

Economic Outcomes

- Years of educational attainment from SALT survey.

GREML Analyses

	Economics					Political				
	Edu	Risk	Patient	Fair	Trust	Crime	Econ Pol	Environ	Femin	Foreign
v(g)	0.158	0.137	0.085	0.000	0.242	0.203	0.344	0.000	0.000	0.354
p	0.004	0.186	0.285	0.150	0.146	0.079	0.012	0.500	0.500	0.009
N	5,727	2,327	2,399	2,376	2,410	2,368	2,368	2,368	2,368	2,368
Chrom	0.442	0.118	-0.195	-0.111	0.460	0.118	0.496	-0.311	0.247	0.462
p	0.039	0.601	0.623	0.031	0.031	0.601	0.019	0.159	0.268	0.030

Constructive Response 3

- Focus on large and replicated associations with more biologically proximate traits which have survived the challenges of replication and analyze these associations through the prism of economic theory.
- Example: addictive goods.
- Cigarettes (The Tobacco and Genetics Consortium, 2010, *CHRNA3*)
- Coffee (Cornelis et al., 2011, *CYP1A2*)
- Alcohol (Li et al., 2011, *ADH1B*)
- BMI (Frayling et al., *FTO*)

Conclusion

- These results consistent with these traits having a complex architecture, with highly diffuse and small genetic effects scattered across the genome.
- These results are relevant for evaluating the extent to which the promises of "genoeconomics" are likely to be realized any time soon.
- As we pursue these questions, it is important that we stop recapitulating the mistakes of medical genetics and set high standards.

GREML

We let j index individuals and i indexes SNPs. Let m be the number of causal SNPs and J the number of individuals in our sample. We assume that $e_j \sim N(0, \sigma^2_\epsilon)$. We define,

$$y_j = \mu + g_j + e_j,$$

where $g_j = \sum_{i=1}^{m} z_{ij} u_i$. Let f_i be the frequency of reference allele i and,

$$z_{ij} = \frac{x_{ij} - 2f_i}{\sqrt{2f_i\left(1 - f_i\right)}}$$

where $x_{ij} \in \{0, 1, 2\}$ is the number of references alleles individual j is endowed with at locus i. The standardization ensures that $var\left(z_{ij}\right) = 1$ and $\mathbf{E}(z_{ij}) = 0$.

GREML (continued)

We now write the model in matrix notation,

$$\mathbf{y} = \boldsymbol{\mu} \cdot \mathbf{1} + \mathbf{g} + \mathbf{e},$$

where $\mathbf{g} = \mathbf{Zu}$ and,

$$u \sim N\left(0, I\sigma_u^{\mathbf{2}}\right).$$

This implies that g_j is normal with mean 0 and variance $\sigma_u^2 \equiv \sigma_g^2$.

$$VCOV\left(yy'\right) = E\left(yy'\right) = ZZ'\sigma_u^2 + I\sigma_e^{\mathbf{2}}$$

$$\equiv G\sigma_g^2 + I\sigma_e^{\mathbf{2}}.$$

GREML (continued)

$$y \sim N\left(0, G\sigma_g^2 + I\sigma_e^{\mathbf{2}}\right)$$

G here is the genetic relatedness matrix estimated from the causal *SNP*s. We do not know what the causal SNPs are. An estimator for G is,

$$A = \frac{W'W}{N},$$

where W is the N by j matrix of genotypic data.

Genotyping Errors

- Following usual practices (Manolio et al, 2008; Sullivan and Purcell, 2008), we first applied a number of quality control measures.
  - First, 499 individuals were dropped because they had a "missingness" larger than 0.05.
- Next, we excluded individual SNPs which failed one of three additional quality controls.
  - SNPs with a missing data frequency greater than 2.5% were deleted.
  - We eliminated SNPs with a "minor allele frequency" less than 1%
  - We excluded SNPs which failed a test of Hardy-Weinberg equilibrium at the 10^{-6} level.

Genotyping Errors (cont'd)

From the original 500,568 SNPs on the array:

- 76,764 did not satisfy the missingness criteria
- 61,293 did not satisfy the minor allele frequency criteria
- 16,991 did not pass the Hardy-Weinberg test.

Applying all quality controls leaves a total of 363,776 SNPs for analysis.

Population Stratification

Population stratification: differences in allele frequencies across subpopulations.

- Can be important source of false positives in GWAS.
- The classic "chopstick example" (Hamer and Sirota, 2000)
- GWAS's thus require ethnically homogenous sample (Campbell et al., 2005).

Population Stratification (cont'd)

- Used EIGENSTRAT method to control for remaining stratification (Price et al. 2006).
- Idea: use principal components analysis to explicitly model ancestry differences.
- The correction is specific to a candidate marker's variation in frequency across ancestral populations.

Non-Independence of Errors

- In what follows, the subscripts i or j refer to individuals, $f \in \{1, ..., F\}$ indexes families, and $g \in \{1, 2, 3\}$ refers to the three generations in the data.
- Our sample is family based, so we cannot assume that $\mathrm{E}(\epsilon_{if}\epsilon_{-if})=0$ for two individuals in the same family. Let $E[\varepsilon\varepsilon'] = \Omega$. Assume that the error terms of individuals from different families are independent. Then we can write,

$$\Omega = \mathrm{diag}(\Omega_1, \Omega_2, ..., \Omega_F),$$

- Our strategy is to model the correlation structure of Ω_f.

Modeling Family Correlation

To model the correlation structure of Ω_f, we follow the basic ACE model from the behavioral genetics literature (Falconer and Mackay, 1996; Neale and Cardon, 1992)

$$\varepsilon = \sigma_\varepsilon(aA_{-SNP_S} + cC + eE),$$

where $\sigma_\varepsilon = \sqrt{\sigma^2_\varepsilon}$, $\sigma^2_\varepsilon = var(\varepsilon)$, and A_{-SNP_S}, C, and E are, respectively, the latent additive genetic (with SNP_S partialled out), common environmental, and individual environmental factors underlying educational attainment.

Genetic Relatedness

Biometrical genetic theory implies that, if mating is random,

$$E[A_{-SNP_S,i}, A_{-SNP_S,j}] = r_{ij},$$

where r_{ij} is Sewall Wright's coefficient of relationship. Wright's coefficient of relationship for two individuals is the probability that the alleles of the two individuals at a random locus are identical copies of the same ancestral allele.

Environmental Resemblance

Modelling the transmission of common environment from parent to child is more complicated and no generally agreed upon model exists (See Feldman et al., 2000, for an accessible introduction). We assume that,

$$E[C_{i\ g}, C_{j\ g+1}] = \gamma,$$

Predicted Correlation Structure

	$E[A_i A_j]$	$E[C_i C_j]$	$E[\varepsilon_i \varepsilon_j]$
Relatedness			
Full siblings	$\frac{1}{2}$	1	$\sigma^2_\varepsilon(\frac{1}{2}a^2 + c^2)$
Half siblings	$\frac{1}{4}$	$\frac{1}{2}$	$\sigma^2_\varepsilon(\frac{1}{4}a^2 + \frac{1}{2}c^2)$
Parent-child	$\frac{1}{2}$	γ	$\sigma^2_\varepsilon(\frac{1}{2}a^2 + \gamma c^2)$
Grandparent-grandchild	$\frac{1}{4}$	γ^2	$\sigma^2_\varepsilon(\frac{1}{4}a^2 + \gamma^2 c^2)$
Full cousins	$\frac{1}{8}$	γ^2	$\sigma^2_\varepsilon(\frac{1}{8}a^2 + \gamma^2 c^2)$
Half cousins	$\frac{1}{16}$	$\frac{1}{2}\gamma^2$	$\sigma^2_\varepsilon(\frac{1}{16}a^2 + \frac{1}{2}\gamma^2 c^2)$
Aunt/uncle-nephew	$\frac{1}{4}$	γ	$\sigma^2_\varepsilon(\frac{1}{4}a^2 + \gamma c^2)$
Half aunt/uncle-nephew	$\frac{1}{8}$	$\frac{1}{2}\gamma$	$\sigma^2_\varepsilon(\frac{1}{8}a^2 + \frac{1}{2}\gamma c^2)$

Obtaining Estimates of the Elements of Omega

Solve the system of equations,

$$\hat{\rho}_{FS}(\varepsilon_i, \varepsilon_j | i, j \text{ are full siblings}) = \tfrac{1}{2}\hat{a}^2 + \hat{c}^2,$$
$$\hat{\rho}_{PC}(\varepsilon_i, \varepsilon_j | i, j \text{ are parent-child}) = \tfrac{1}{2}\hat{a}^2 + \hat{\gamma}\hat{c}^2,$$
$$\hat{\rho}_{AUC}(\varepsilon_i, \varepsilon_j | i, j \text{ are Aunt/uncle-nephew/niece}) = \tfrac{1}{4}\hat{a}^2 + \hat{\gamma}\hat{c}^2,$$

and from these estimates obtain $\hat{\Omega}$, so a consistent estimator of the variance covariance matrix of the regression coefficients is:

$$\hat{var}(\hat{\beta}) = (\Sigma_{f=1}^{F} X_f^T X_f)^{-1}(\Sigma_{f=1}^{F} X_f^T \hat{\Omega}_f X_f)(\Sigma_{f=1}^{F} X_f^T X_f)^{-1}.$$

Multiple Hypothesis Testing

- Total of 363,776 regressions => expect to find 5% of them significant" even if just noise
- Bonferroni correction: divide all p-values by number of regressions run. Probably overly conservative because of linkage disequilibrium blocks.
- Duggal et al. (2008) propose the following taxonomy: p-values less than 1.49×10^{-5} are "suggestive" and 7.47×10^{-7} "significant".