Genetics and Economic Outcomes

Kevin Thom
Dept. of Economics, UW-Milwaukee

February 2, 2023

Genetic Factors and Economic Outcomes

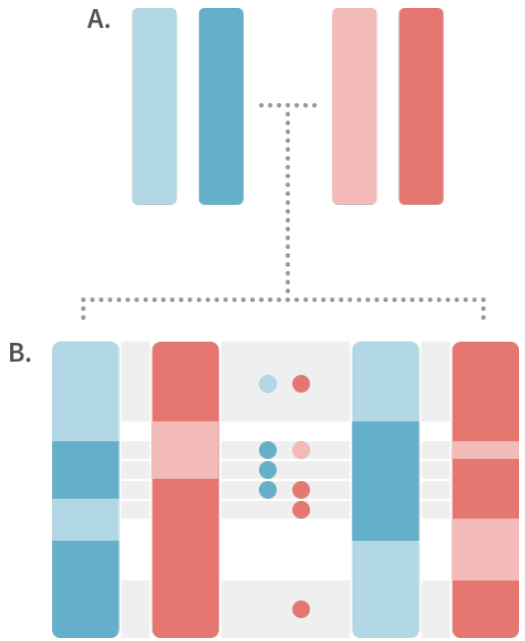
- Very old question in social sciences
- "Nature v.s. Nurture" (quite the cliche)
- Why do we see variability across individuals in important economic outcomes?
- Intuition that traits (e.g. eye color) "run in families" - does this extend to economic outcomes like education, earnings, risk preferences, etc.
- Huge problem: parents are passing along both genetic material and rearing environments. How can we identify these things separately?

Outline

1. Twin
2. Molecular Genetics and GWAS
3. GWAS results for Educational Attainment
4. Polygenic Scores - Construction, Interpretation
5. Applications for Molecular Genetics:
   - Mendelian Randomization
   - Gene-by-Environment Interactions
   - Learning about Mechanisms

- Human DNA is a sequence of approximately 3 billion nucleotide molecules spread across 23 chromosomes.
- Each human has two copies of each chromosome: one from each parent.

- Panel A: parental genetic material at a particular chromosome.
- Panel B: genetic material of two siblings (not identical twins).



- Process of genetic inheritance creates the following regularities:
  - Full siblings will share 50 percent of the same genetic material.
  - Identical (monozygotic, or MZ) twins share 100 percent of their genetic material.
  - Non-identical (dizygotic, or DZ) twins share 50 percent of their genetic material (just like regular siblings)
- The basic rationale for the twin study approach to identifying the contribution of genetic factors to an outcome:
  - Compare pairs of MZ twins to DZ twins
  - Both share same parents, neighborhood, in utero environment, etc.
  - Big difference - MZ twins have same genetic material, DZ twins only share half.

Twin Studies - A Basic Model

- Assume that outcome y_i is determined by a genetic component (g_i) and an environmental component, (ϵ_i):

$$y_i = g_i + \epsilon_i \tag{1}$$

- One goal would be to estimate the heritability of the outcome y_i:

$$h^2 = \frac{Var(g_i)}{Var(y_i)} \tag{2}$$

- This is the fraction of the variance of y_i accounted for by variation in genetic factors.
- Notice, making strong assumptions here about independence of distribution of g_i and ϵ_i

Twins Studies

- Assume that y_i is determined by a genetic component (g_i) and an environmental component, (ϵ_i):

$$y_i = g_i + \epsilon_i \tag{3}$$

- Suppose we have pairs of observations for twins, $< y_i, y_{i'} >$:
- Importantly, there are two varieties of twins:
  - **Monozygotic Twins:** Share all genetic material, so $g_i = g_{i'}$
  - **Dizygotic Twins:** Share approximately 50 percent of their genetic material.

Twin Studies .

- We can estimate heritability using the covariance in y for both monozygotic twins and dizygotic twins:
- For monozygotic pairs:

$$Cov(y_i^m, y_{i'}^m) = Var(g_i^m) + Cov(\epsilon_i^m, \epsilon_{i'}^m) \tag{4}$$

- For dizygotic pairs we have :

$$Cov(y_i^d, y_{i'}^d) = \frac{1}{2} Var(g_i^d) + Cov(\epsilon_i^d, \epsilon_{i'}^d) \tag{5}$$

- Note: There are some assumptions that go into these covariance formulae - especially the lack of assortative mating (which would cause the genetic covariance for dizygotic pairs to be higher).

Twin Studies

- Make some assumptions:
  1. Common Environments Assumption: $Cov(\epsilon_i^m, \epsilon_{i'}^m) = Cov(\epsilon_i^d, \epsilon_{i'}^d)$
  2. $Var(g_i^m) = Var(g_i^d)$
- Then, we have the following system:

$$Cov(y_i^m, y_{i'}^m) = Var(g_i) + Cov(\epsilon_i, \epsilon_{i'}) \quad (6)$$

$$Cov(y_i^d, y_{i'}^d) = \frac{1}{2}Var(g_i) + Cov(\epsilon_i, \epsilon_{i'}) \quad (7)$$

- Which permits the following estimator for heritability:

$$\frac{\widehat{Var(g_i)}}{Var(y_i)} = 2\left(\frac{\widehat{Cov(y_i^m, y_{i'}^m)}}{Var(y_i^m)} - \frac{\widehat{Cov(y_i^d, y_{i'}^d)}}{Var(y_i^d)}\right) \quad (8)$$

Twin Studies

- Some terminology: Often the canonical model is refereed to as the ACE model:

$$y_i = a_i + c_i + e_i \tag{9}$$

- a_i is the **additive genetic component**
- c_i is the **common environmental component** (shared by all siblings within a house)
- e_i is the **idiosyncratic environmental component** (specific to each individual).
- So $\epsilon_i = c_i + e_i$ in our previous notation.

- From Branigan et al (2013) - a meta-analysis of many recent twins studies on educational attainment:

							h^2	c^2	e^2
Nationality ($k = 34$)	Sex	Cohort	r_{MZ}	N_{MZ}	r_{DZ}	N_{DZ}	$2(r_{MZ} - r_{DZ})$	$r_{MZ} - h^2$	$1 - r_{MZ}$
Australia (1)	Male	1	**0.70**	216	**0.53**	94	0.34	0.36	0.30
	Female	1	**0.77**	520	**0.55**	299	0.44	0.33	0.23
Australia (2)	Male	2	**0.74**	226	**0.47**	161	0.54	0.20	0.26
	Female	2	**0.75**	479	**0.49**	290	0.52	0.23	0.25
Australia (3)	Male	2	**0.674**	282	**0.532**	164	0.284	0.39	0.326
	Female	2	**0.705**	320	**0.319**	158	0.772	–0.067	0.295
Denmark	Male	2	**0.62**	4370	**0.444**	7068	0.352	0.268	0.38
Finland	Male	1	**0.83**	1506	**0.58**	3504	0.50	0.33	0.17
	Female	1	**0.86**	2028	**0.62**	3870	0.48	0.38	0.14
Germany	Male	Mixed	**0.680**	133	**0.306**	47	0.748	–0.068	0.320
	Female	Mixed	**0.717**	421	**0.479**	172	0.476	0.241	0.283
Italy	Male	Mixed	**0.71**	752	**0.61**	406	0.20	0.51	0.29
	Female	Mixed	**0.79**	1342	**0.7**	712	0.18	0.61	0.21
Norway (1)	Male	1	**0.86**	259	**0.77**	313	0.18	0.68	0.14
	Female	1	**0.89**	405	**0.75**	425	0.28	0.61	0.11
Norway (2)	Male	1	**0.82**	253	**0.48**	284	0.68	0.14	0.18
	Female	1	**0.85**	342	**0.68**	400	0.34	0.51	0.15
Norway (3)	Male	2	**0.85**	370	**0.47**	463	0.76	0.09	0.15
	Female	2	**0.89**	518	**0.66**	576	0.46	0.43	0.11

- Branigan et al (2013) table continued:

Spain	Male	Mixed	**0.758**	128	**0.519**	155	0.478	0.28	0.242
	Female	Mixed	**0.821**	228	**0.562**	231	0.518	0.303	0.179
Sweden	Mixed	1	**0.76**	2492	**0.55**	3368	0.42	0.34	0.24
United States									
AddHealth	Male	2	**0.611**	100	**0.477**	94	0.268	0.343	0.389
	Female	2	**0.623**	117	**0.650**	93	–0.054	0.677	0.377
Vietnam Veterans	Male	1	**0.76**	1019	**0.54**	907	0.44	0.32	0.24
Minnesota	Male	Mixed	**0.65**	512	**0.42**	772	0.46	0.19	0.35
	Female	Mixed	**0.72**	758	**0.57**	1154	0.30	0.42	0.28
WW2 Veterans (NAS-NRC)	Male	1	**0.764**	1234	**0.545**	1167	0.438	0.326	0.236
MIDUS	Male	Mixed	**0.668**	164	**0.538**	124	0.293	0.375	0.332
	Female	Mixed	**0.707**	186	**0.561**	198	0.260	0.447	0.293
SRI	Male	Mixed	**0.65**	170	**0.48**	28	0.34	0.31	0.35
	Female	Mixed	**0.68**	390	**0.50**	123	0.36	0.32	0.32
United Kingdom (1)	Mixed	1	**0.717**	457	**0.521**	393	0.391	0.326	0.283
United Kingdom (2)	Mixed	2	**0.593**	388	**0.474**	247	0.238	0.355	0.407
			Total	23,085		28,460			

Critiques of Twin Methodologies

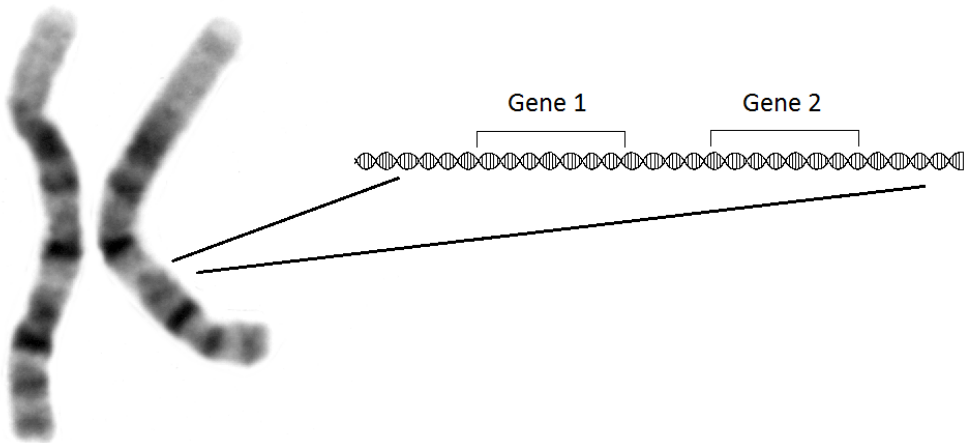
- Equal Environments assumption is strong - parents may treat identical twins more similarly than non-identical twins.
- Model assumes lack of interactions between genes and environments - evidence that this may not be true.
- Assume additivity and lack of interactions between genes, or non-linearities in effects of genetic variants.
- What do you do with these estimates? If heritability is high (or low) - doesn't this only indicate that genes matter more (or less) in the particular environments studied?

From Twins Studies to Molecular Genetics

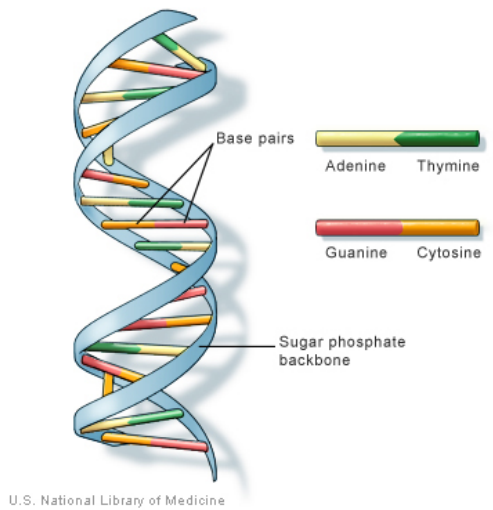
- Twins studies are useful, but face some limitations:
  - Cannot tell us **which genes** matter.
  - Difficult to explore interactions between genes and environments, and mechanisms.
  - Need strong assumptions about how twins are reared, how genes and environments interact.
- Recent advances in molecular genetics allow us to start studying individual genetic markers.

- Human DNA is a sequence of approximately 3 billion nucleotide molecules spread across 23 chromosomes.
- Each human has two copies of each chromosome: one from each parent.

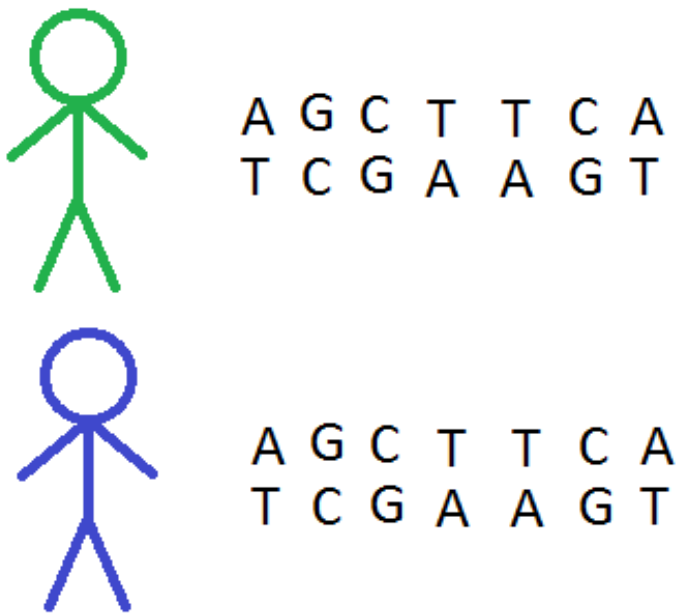
- If we zoom in further, we see that each chromosome contains subsequences of genetic material that are referred to as **genes**.
- There are between 20,000-25,000 genes in the human genome.
- Genes provide instructions for synthesizing proteins that affect body function.



- Each gene consists of a sequence of **base pairs**.
- Pairs can either be adenine-thymine (AT) pairs,or guanine-cytosine (GC) pairs.
- So at each address in the human genome, we can either see (AT) or (GC).



- At the vast majority of locations in the human genome, there is no variation in the population.
- All individuals have the same nucleotide pair at such locations.



- A **single nucleotide polymorphism (SNP)** is a form of genetic variation in which individuals differ in which base pair (e.g. AT or GC) resides at a particular genetic address.
- We will refer to specific positions in the genetic code by names, such as rs1051730
- Alleles:
  - The **major allele** at a position is the more common allele in pop.
  - The **minor allele** at a position is the less common allele.

- Suppose that there is variation at rs1051730, and that the major allele is AT.
- Then individuals can differ in terms of how many copies of the minor or major allele (AT) they possess (0, 1, or 2 since there are two copies of each chromosome).
- An individual's genotype at a particular SNP is the number of copies of the reference allele that they possess:

$$rs1051730_i \in \{0, 1, 2\}$$

- Genome Wide Association Study (GWAS)
- Basic Procedure:
  - Regress the outcome against individual SNPs, one at a time:

$$y_i = \mu + \beta_j x_{ij} + Z'\gamma + \epsilon_i$$

  - Z includes controls - especially some number of principal components of the genetic data.
  - Collect the GWAS coefficients $\widehat{\beta}_j$ and the associated p-values.
  - Associations with sufficiently small p-values are considered **genome-wide significant**.
- Key to addressing multiple hypothesis testing: apply stringent p-value thresholds (typically 5×10^{-8}).

PGS Construction

- Using the coefficients from a GWAS one can form a **polygenic score** as follows:

$$PGS_i = \sum_j \widetilde{\beta}_j SNP_{ij}$$

- One issue: SNPs may be correlated. Two SNPs that are correlated are said to be in **linkage-disequilibrium**
- If the SNPs are correlated, then unadjusted coefficients $\widehat{\beta}_j$ may over or underestimate the influence of an individual SNP.
- The sum $\sum_j \widehat{\beta}_j SNP_{ij}$ could double-count certain SNPs.

PGS Construction

- Using the coefficients from a GWAS one can form a **polygenic score** as follows:

$$PGS_i = \sum_j \widetilde{\beta}_j SNP_{ij}$$

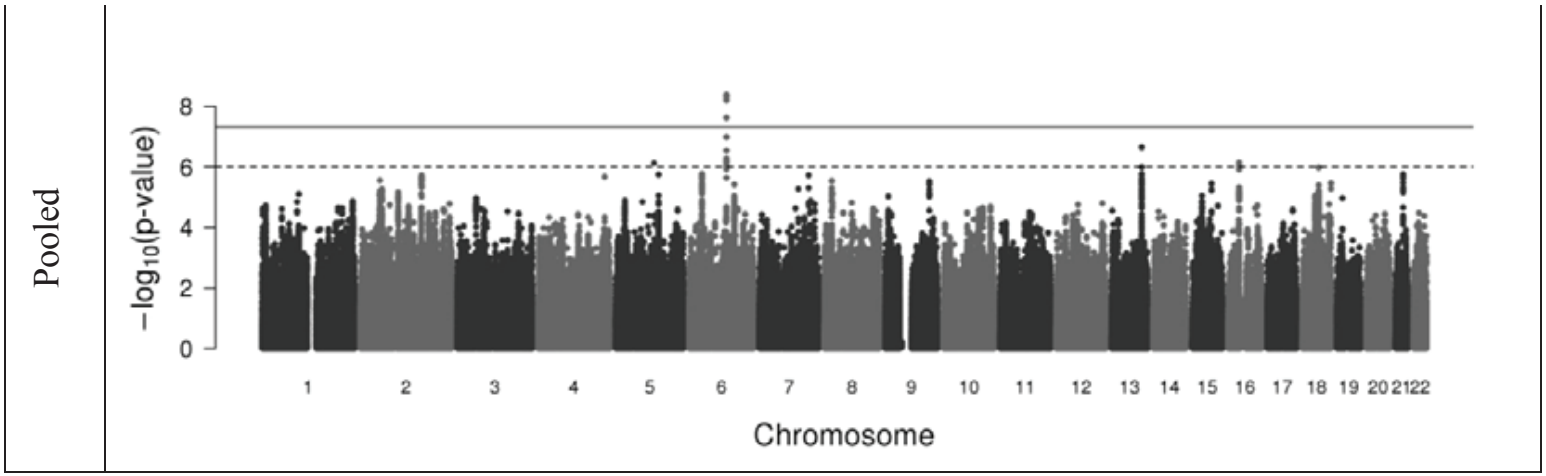
- Various algorithms to adjust for correlated SNPs: incorporate information about SNP covariances to adjust for correlation (LDPred)

- Other choices here - how many SNPs? P-value thresholds?

Educational Attainment

- A series of GWAS have studied educational attainment (EA)
- First GWAS of educational attainment Rietveld et al (2013):
  - Overall discovery sample of size N=126,559
  - Identified three SNPs with association sizes reaching genome-wide significance: rs9320913, rs11584700, and rs4851266
- Subsequent EA GWAS:
  - Okbay et al (2016): Discovery sample of N 300,000
  - Lee et al (2018): Discovery sample of N 1.1 million
  - Okbay et al (2022): Discovery sample of N 3 million

- GWAS results are often depicted graphically using a **Manhattan plot**.
- Each position on the X-axis represents a loci or position on genome (arranged by chromosomes). The associated p-values are plotted.



- Rietveld et al (2013)
- Sample size of N=126,559
- 3 genome-wide significant associations

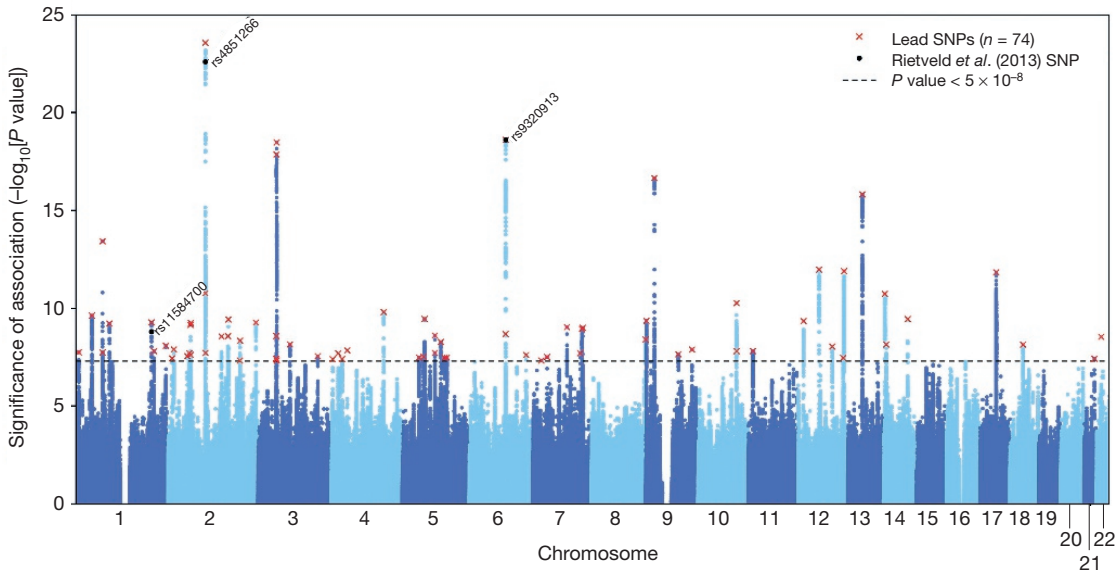


Figure 1 | Manhattan plot for EduYears associations ($n = 293,723$). The x axis is chromosomal position, and the y axis is the significance on a $-\log_{10}$ scale (two-tailed test). The black dashed line shows the genome-wide significance level (5×10^{-8}). The red crosses are the 74 approximately independent genome-wide significant associations (lead SNPs). The black dots labelled with rs numbers are the three SNPs identified in ref. 1.

- Okbay et al (2016)
- Sample size of N 300,000
- 74 genome-wide significant associations

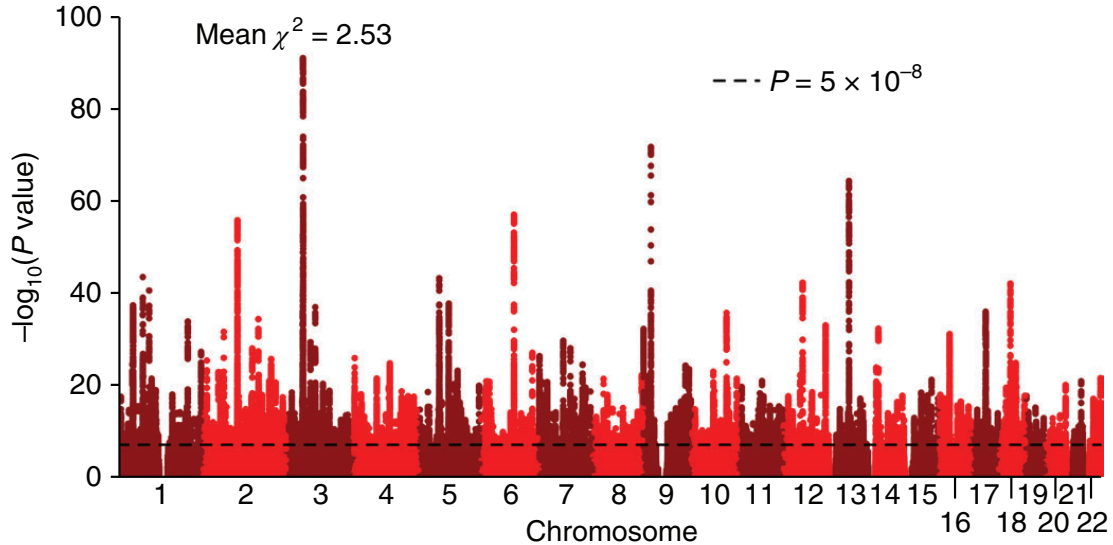
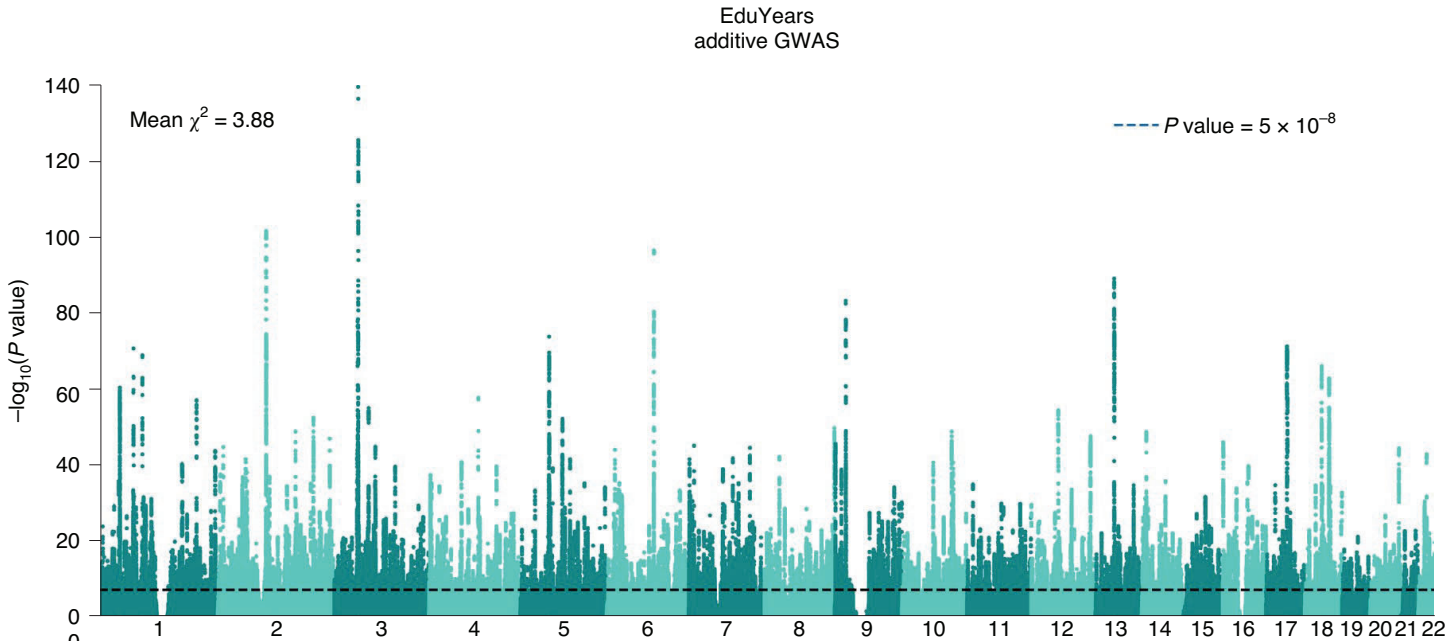


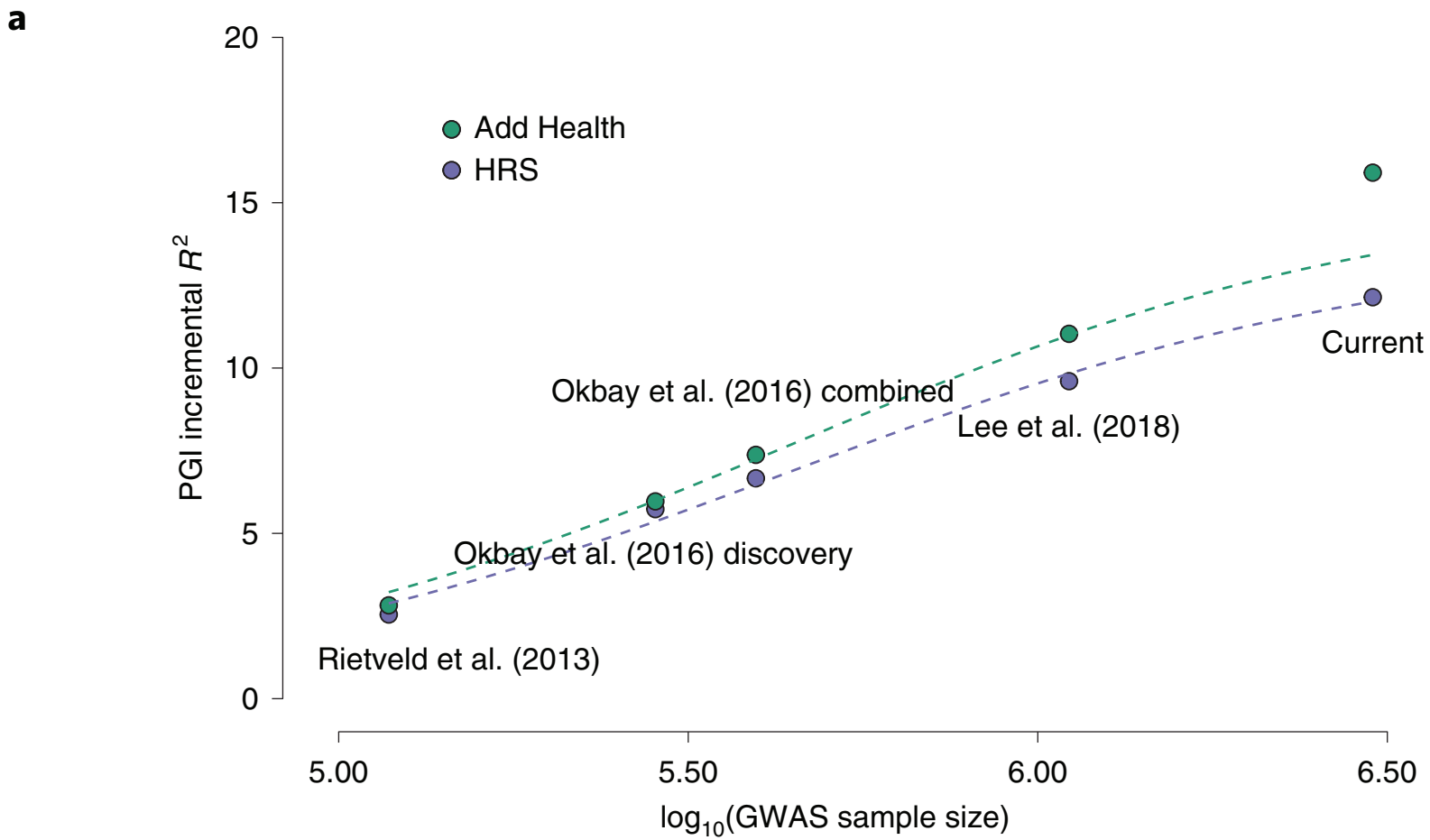
Fig. 1 | Manhattan Plot for GWAS of EduYears. The *P* value and mean χ^2 value are based on inflation-adjusted test statistics. The *x* axis is chromosomal position and the *y* axis is the significance on a $-\log_{10}$ scale. The dashed line marks the threshold for genome-wide significance ($P = 5 \times 10^{-8}$) ($n = 1{,}131{,}881$).

- Lee et al (2018)
- Sample size of N 1.1 million
- 1,271 genome-wide significant associations



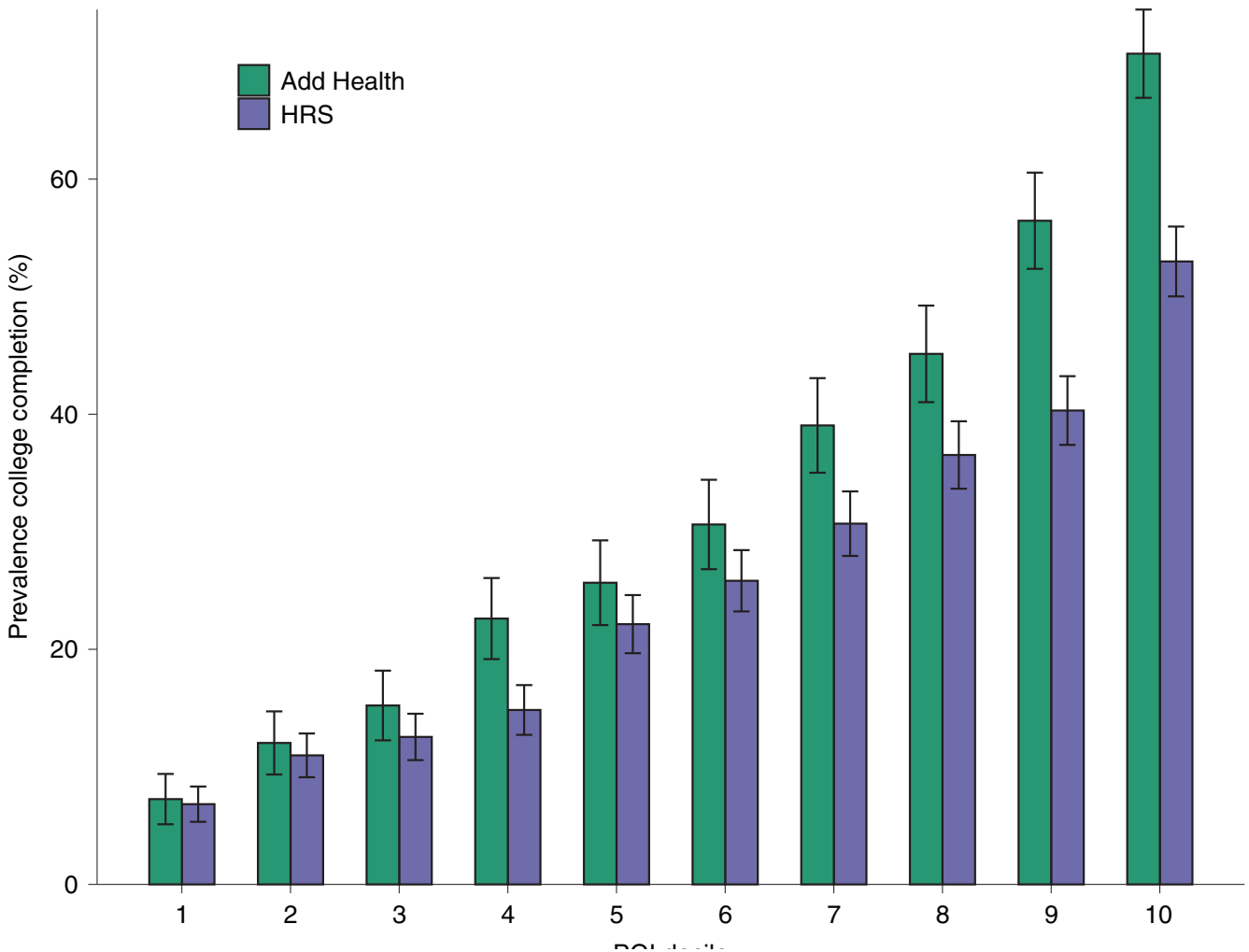
- Okbay et al (2022)
- Sample size of N 3 million
- 3,952 genome-wide significant associations

- From Okbay et al (2022)



- Increasing incremental R^2 of PGS in predicting educational attainment.

- From Okbay et al (2022)



- College Completion by PGS Decile

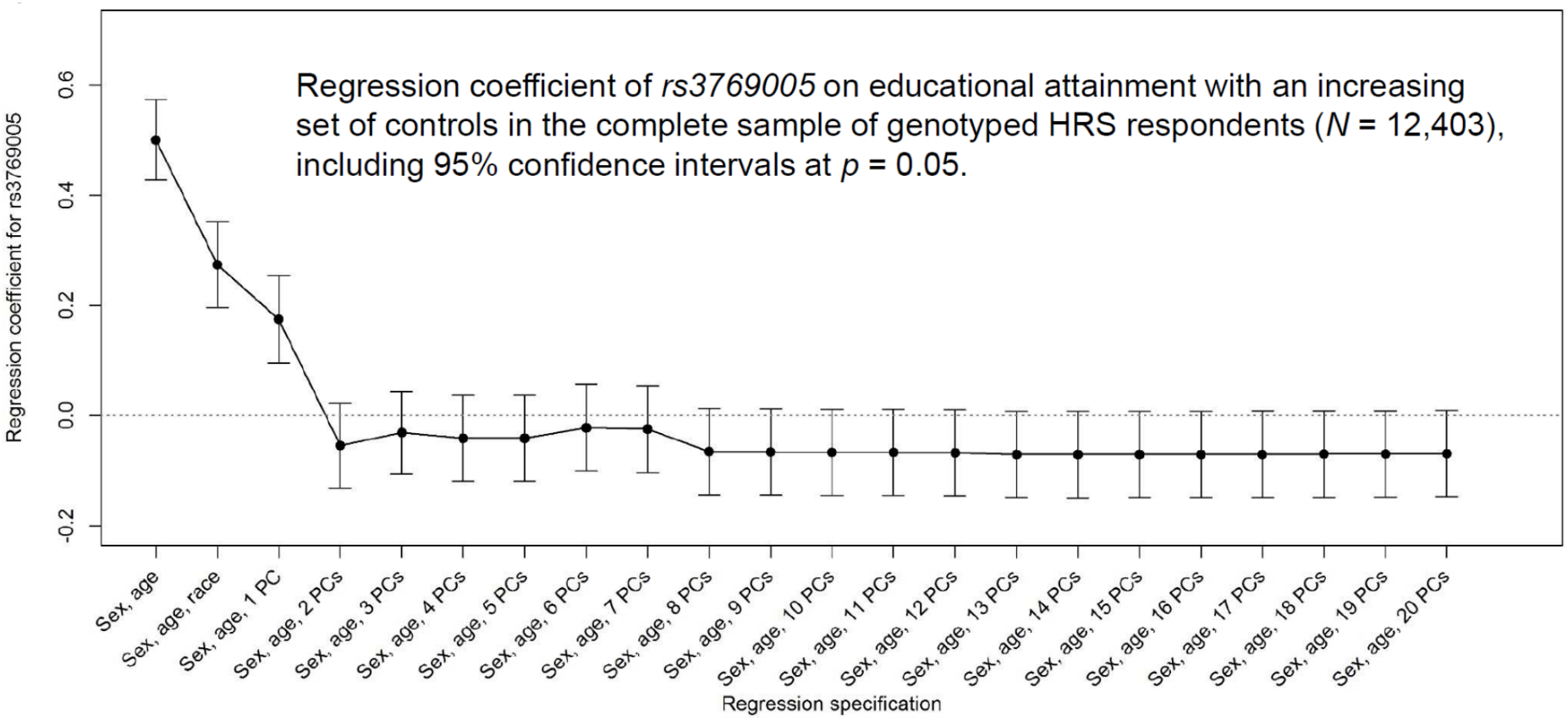
Interpreting GWAS Associations

- What do these associations reflect? Are these causal? And what do we even mean by causal here?
- An enormous set of questions.
- Before thinking about more satisfactory methods (e.g. within-family analyses), let's look at some features of GWAS that might alleviate concerns.
  - Controlling for Principal Components
  - Examining biological annotation

Controlling for Principal Components

- Recall - principal components of the SNP-level data are added as essential controls in GWAS
- May be concerned that variation in markers reflect **population stratification**: different markers could be associated with an outcome because of correlation with an ancestral history.
- Genetic ancestry groups share big blocks of genetic material - principal components of the SNP data do a good job of capturing variation due to ancestral clustering.

- From Rietveld et al (2014)

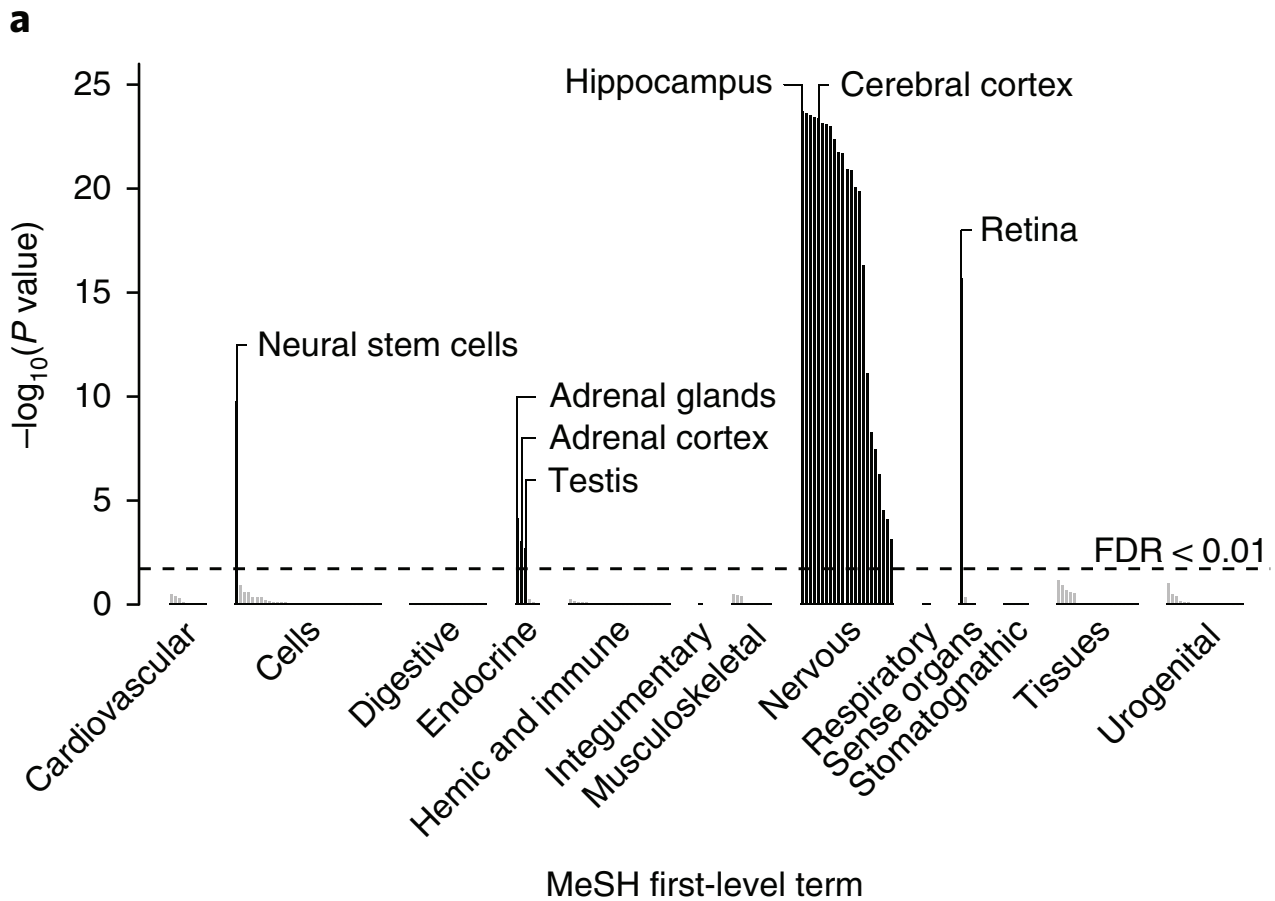


- Here $rs3769005$, which affects lactose metabolism, has a significant association with educational attainment, but not after controlling for first two PCs.

Biological Annotation

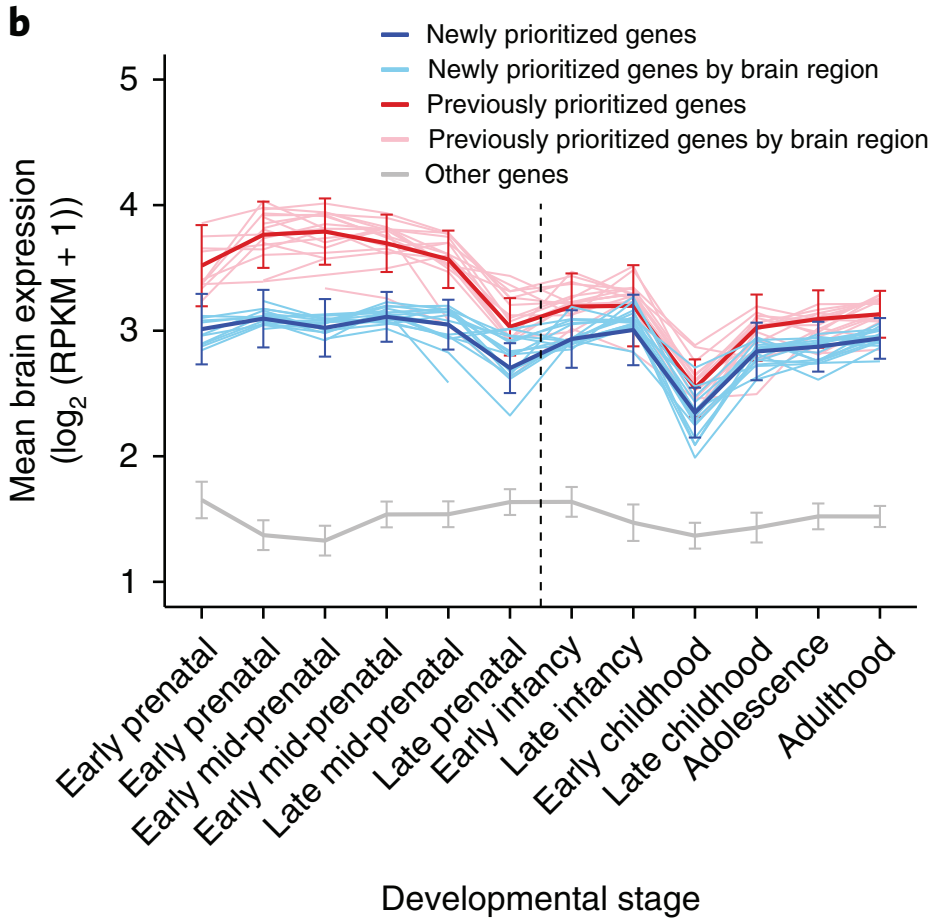
- One exercise that can be performed after a GWAS is **biological annotation**
- Biologists have some information on pathways of genes:
  - In what kinds of tissues these genes are expressed - that is, where they are being used to code for proteins or perform regulatory functions.
  - Also know when these genes are expressed.
  - Can ask - are the SNPs that are more heavily weighted in the GWAS found in genes that are expressed more in a particular tissue or at a particular time?

- From Lee et al (2018)



- Genes associated with genome wide significant SNPs largely linked to expression in central nervous system.

- From Lee et al (2018)

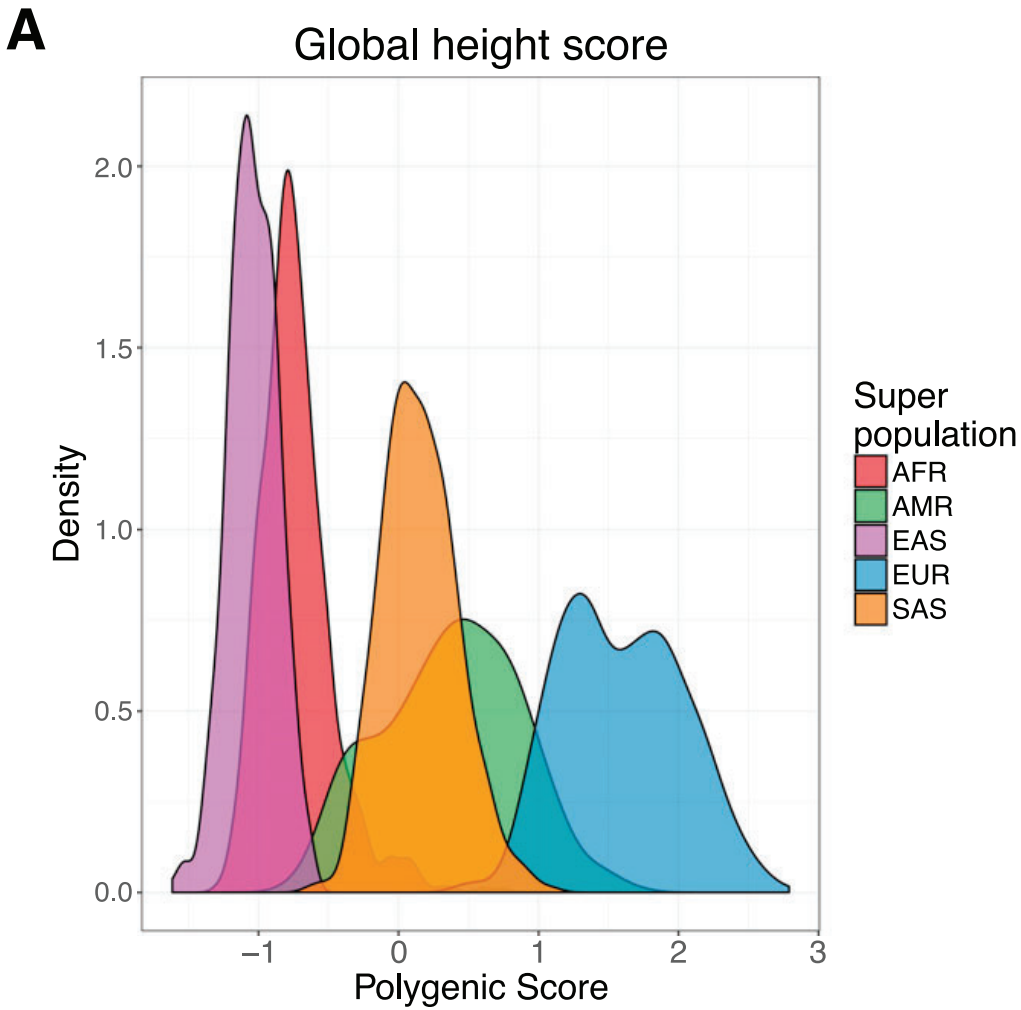


- Newly identified SNPs show expression both prenatally and during adulthood.
- Glial cells not implicated (but just as numerous as neurons)

Polygenic Variation Across Ancestry Groups

- An important caveat - current GWAS are overwhelmingly performed on samples of European ancestry.
- A limitation here is that these results (and PGS constructed from them) cannot be used to learn about differences between population ancestry groups (or different associations between PGS across ancestry groups).
- Martin et al (2017) discusses the issue in detail
- Allele frequencies differ across ancestry groups, and patterns of linkage-disequilibrium may be different as well.
- For example - using a PGS for height constructed from a European ancestry GWAS predicts that individuals of African ancestry should have average heights that are several standard deviations lower than the European average - this is clearly erroneous.

- From Martin et al (2017)



Within-Family Variation

- Perhaps the most convincing approach to addressing causality in genetic associations - within-family variation.
- Conditional on having the same parents, variation in the genotypes of two siblings is purely random.
- Family fixed-effects designs can identify causal effects of variation in genetic measures, under some assumptions
- Bottom Line: Within family estimates of polygenic associations tend to shrink cross-sectional estimates by about 50%. Controlling for parental education / background can account for much of this.

- From Belskey et al. (2018), cross-sectional v.s. within-family estimtes of correlation between EA PGS and various outcomes

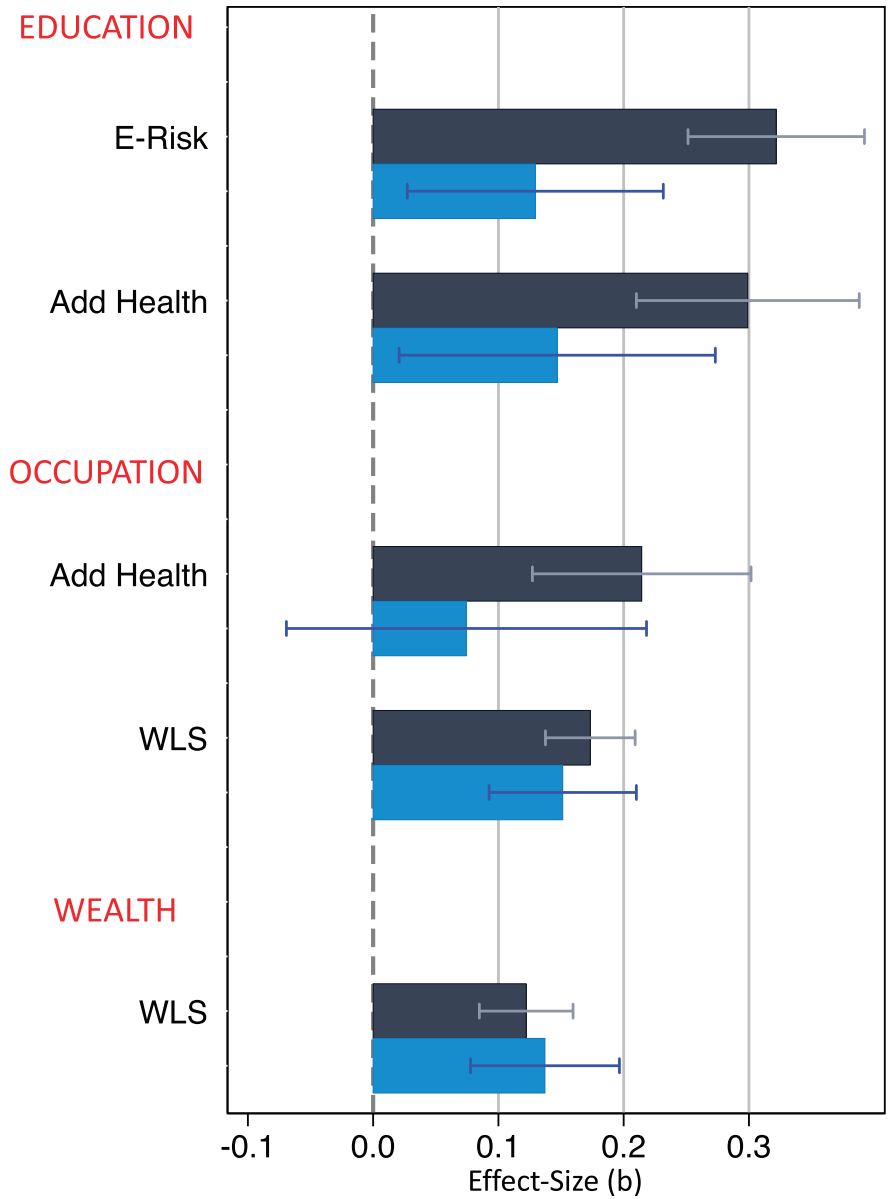


Fig. 3. Sibling-difference effect-size estimates for education, polygenic

- From Ronda et al (2020) - Sample of Danish Siblings

Table 3: Sibling Sample: EA PGS and Human Capital Formation

Dep. Var.	(1) Y. Edu.	(2) Any P.S.E.	(3) Danish	(4) Math.
Panel A:				
EA PGS	0.561	0.114	6.248	6.722
	(0.053)	(0.010)	(0.469)	(0.558)
Family Controls	(N)	(N)	(N)	(N)
Family F.E.	(N)	(N)	(N)	(N)
R^2	0.123	0.123	0.179	0.103
Incr. R^2 EA PGS	0.070	0.070	0.077	0.072
Panel B:				
EA PGS	0.352	0.073	4.542	4.780
	(0.055)	(0.011)	(0.491)	(0.570)
Family Controls	(Y)	(Y)	(Y)	(Y)
Family F.E.	(N)	(N)	(N)	(N)
Panel C:				
EA PGS	0.296	0.069	2.774	3.616
	(0.094)	(0.020)	(0.842)	(0.982)
Family Controls	(N)	(N)	(N)	(N)
Family F.E.	(Y)	(Y)	(Y)	(Y)
N	1,487	1,487	1,838	1,793

Applications

- Suppose you are convinced that molecular genetic measures are picking up something real. What do you do with this?
- Suggest a few possible applications:
  - Mendelian Randomization (genes as IVs)
  - Gene-by-Environment Interactions
  - Understanding Structure of Heterogeneity (mechanisms in a structural model)

Mendelian Randomization Studies

- If genetic variation within families is truly random, then could use within-family variation as an instrumental variable for various outcomes.
- Some major challenges here:
  - Very unlikely to satisfy exclusion restructions.
  - Pleiotropy - gene can affect multiple outcomes through one or multiple mechanisms.

Educational Attainment, Polygenic Scores, and Labor Market Outcomes

- Summarize four papers:
  - Houmark, Ronda, Rosholm (2020) "The Nurture of Nature and Nature of Nurture"
  - Papageorge and Thom (2020) "Genes, Education, and Labor Market Outcomes"
  - Barth, Papageorge, and Thom (2020), "Genetic Endowments and Wealth Inequality"
  - Barth, Papageorge, Thom, and Velasquez-Giraldo (2020), "Genetic Endowments, Income Dynamics, and Wealth Accumulation Over the Lifecycle"
- Broad goals:
  - Understand mechanisms through which genes seem to operate.
  - Understand how environments (which policy can affect) might interact with endowments.
  - Important for building better structural models, especially models with overlapping generations.

Incorporating Genes into Technology of Skill Formation (Houmark, Ronda, Rosholm 2020).

- Genetic measures can potentially have a large impact on study of skill formation and child development.
- Some basic questions:
  - Where do genetic factors show up in the skill production technology?
  - How do parental genes and child genes interact in the production process?
  - Given role of genes and the dynamics of skill formation, how can policy affect (genetic) inequality.

Incorporating Genes into Technology of Skill Formation (Houmark, Ronda, Rosholm 2020).

- Use data from the ALSPAC (Avon Longitudinal Study of Parents and Children)
- Features large number of **genetic trios** - family observations with genetic data for a child and both parents.
- Basic idea - incorporate genetic variation at both the child and parent level into a model of skill formation (in spirit of Cunha and Heckman (2007), Cunha, Heckman, and Schennach (2010).

- Model evolution of skills, θ_{it} across six periods: ages 0-2 $(t = 0)$, 2-3 $(t = 1)$, 3-4 $(t = 2)$, 4-5 $(t = 3)$, 5-6 $(t = 4)$ and 6-7 $(t = 5)$
- Cobb-Douglas Production Technology

$$\ln \theta_{it+1} = \ln A + \delta_1 \ln \theta_{it} + \delta_2 \ln I_{it} + \delta_3 \text{pgs}_i + \delta_4 \text{pgs}_i^p + \epsilon_{it} \tag{10}$$

- Parental Investments are modelled as:

$$\ln I_{it} = \gamma_1 \ln \theta_{it} + \gamma_2 \text{pgs}_i + \gamma_3 \text{pgs}_i^p + \gamma_x X_{it}^I + \eta_{it} \tag{11}$$

- Initial Skills:

$$\ln \theta_{i0} = \alpha_1 \text{pgs}_i + \alpha_2 \text{pgs}_i^p + \alpha_x X_i^{\theta_0} + \epsilon_{i0} \tag{12}$$

Role for genetic endowments:

- *Direct effects*: child's own PGS can affect skills directly.
- *Nature of Nurture*: Parental PGS may show up directly in investment function
- *Nurture of Nature*: Parents may be responding to higher PGS children by investing more.

Table 1: EA PGS and Skills by Age

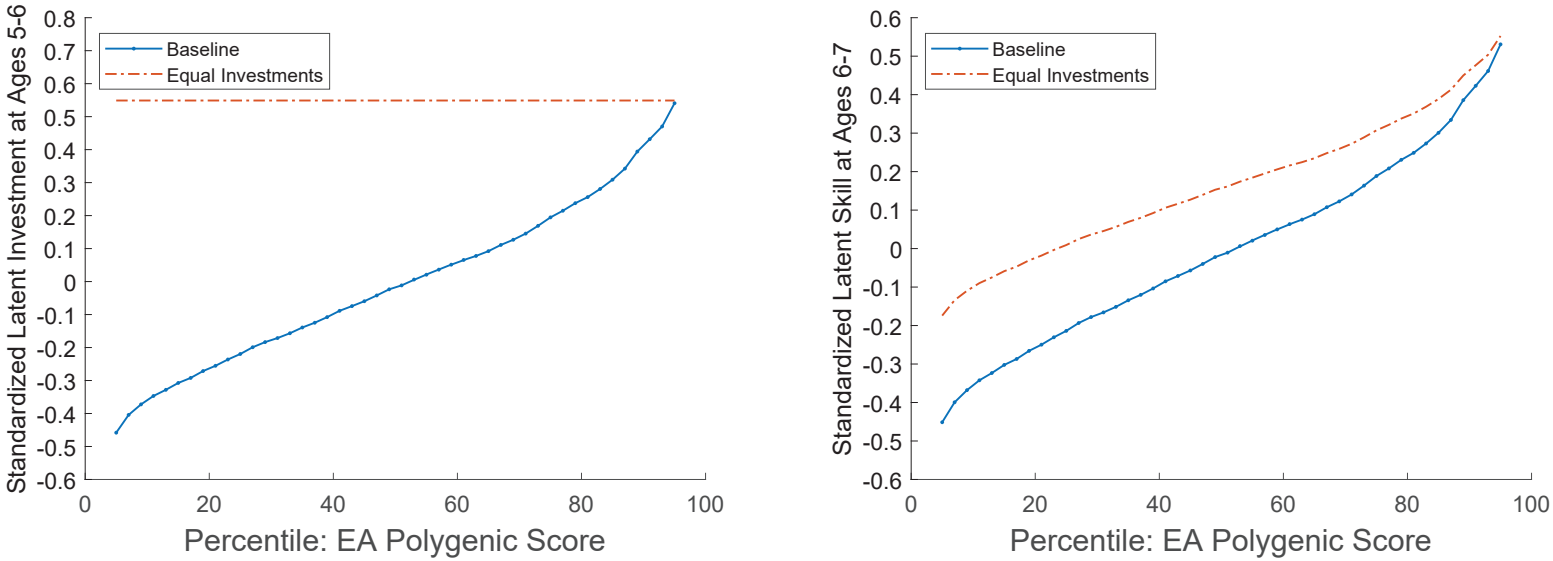
Ages:	[0-2[	[2-3[	[3-4[	[4-5[	[5-6[	[6-7[	[Pooled]
Panel A:							
Child's PGS	0.047*	0.047*	0.097***	0.158***	0.169***	0.101***	0.103***
	(0.029)	(0.027)	(0.028)	(0.028)	(0.028)	(0.028)	(0.021)
R_2	0.002	0.003	0.009	0.027	0.028	0.010	0.012
N	1267	1267	1267	1267	1267	1267	7602
Panel B:							
Child's PGS	0.045	0.009	0.024	0.076*	0.099**	0.039	0.049
	(0.044)	(0.042)	(0.042)	(0.043)	(0.043)	(0.043)	(0.032)
Parental PGS	0.003	0.051	0.097**	0.108**	0.092**	0.082*	0.072**
	(0.043)	(0.042)	(0.042)	(0.042)	(0.043)	(0.043)	(0.031)
N	1267	1267	1267	1267	1267	1267	7602

Notes: This table reports parameter estimates from regressions used to link the polygenic score for educational attainment to children's skills across childhood. To test the effect of the EA PGS, we regress at each age the skill measure on the polygenic score, controlling for gender and the first 15 principal components of the genetic matrix. In Panel B, we add the parental polygenic score to the regressions. Skills have been standardized as described in the data section, with missing values set equal to the median for that measure, allowing for a maximum of ten such imputations per summary score. Standard errors are reported in parenthesis. In the pooled specification, standard errors are clustered at the individual level.

Table 3: Main Parameter Estimates

	$\ln \theta_{i0}$	$\ln \theta_{it+1}$	$\ln I_{it}$
pgs_i	0.022	0.016	0.013
	[0.002 , 0.039]	[0.005 , 0.032]	[-0.001 , 0.027]
pgs_i^p	-0.001	0.020	0.041
	[-0.018 , 0.018]	[0.010 , 0.037]	[0.023 , 0.056]
$\ln \theta_{it}$	.	0.469	0.265
	.	[0.419 , 0.538]	[0.180 , 0.303]
$\ln I_{it}$	.	0.205	.
	.	[0.120 , 0.293]	.
Constant	1.463	1.151	3.076
	[1.434 , 1.494]	[0.672 , 1.567]	[2.985 , 3.295]

Notes: The parameter estimates for the initial skill equation (Equation 15) are reported in the first column, for the technology of skill formation (Equation 13) in the second column, and for the investment policy function (Equation 14) in the third column. 90% bootstrap confidence intervals in brackets.



(a) Latent Investment

(b) Latent Skills

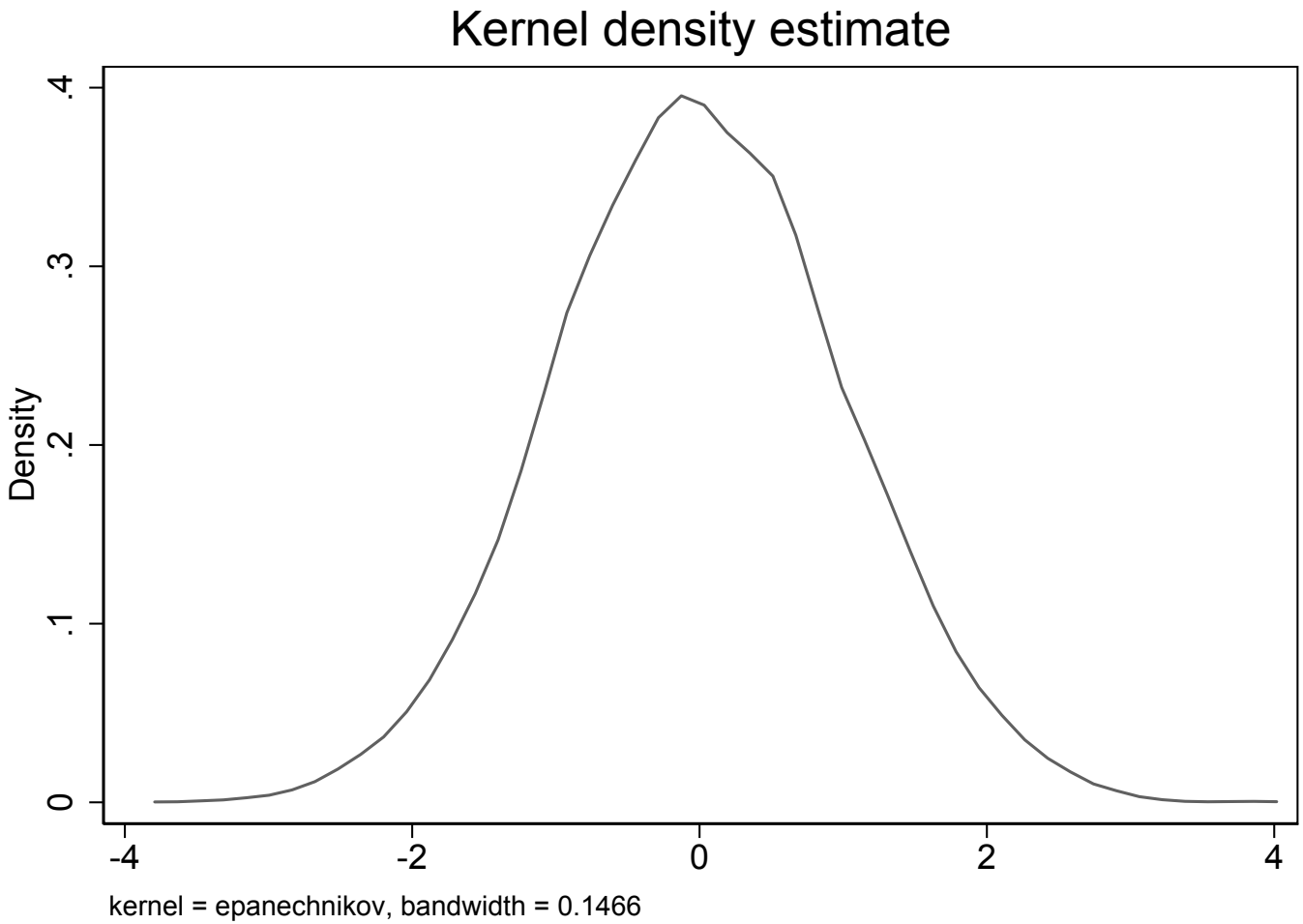
Figure 5: Equalising Investments: These figures compare baseline and simulated skills and investments when investments are equalized at the 95th percentile. We demonstrate graphically how a decrease in social inequality, via equalising parental investments, leads to a decrease in genetic inequality.

Genetic Data and the HRS.

- Shift attention to Papageorge and Thom (2020), Barth, Papageorge and Thom (2020), and Barth et al (2022).
- Longitudinal sample of U.S. over age 50.
- Surveys begin 1992; occur every two years.
- Individuals genotyped in four waves (2006, 2008, 2010, 2012).
- Score we use constructed for the first two waves.
- Individuals had to survive until at least 2006 to be included.

Analytic Samples.

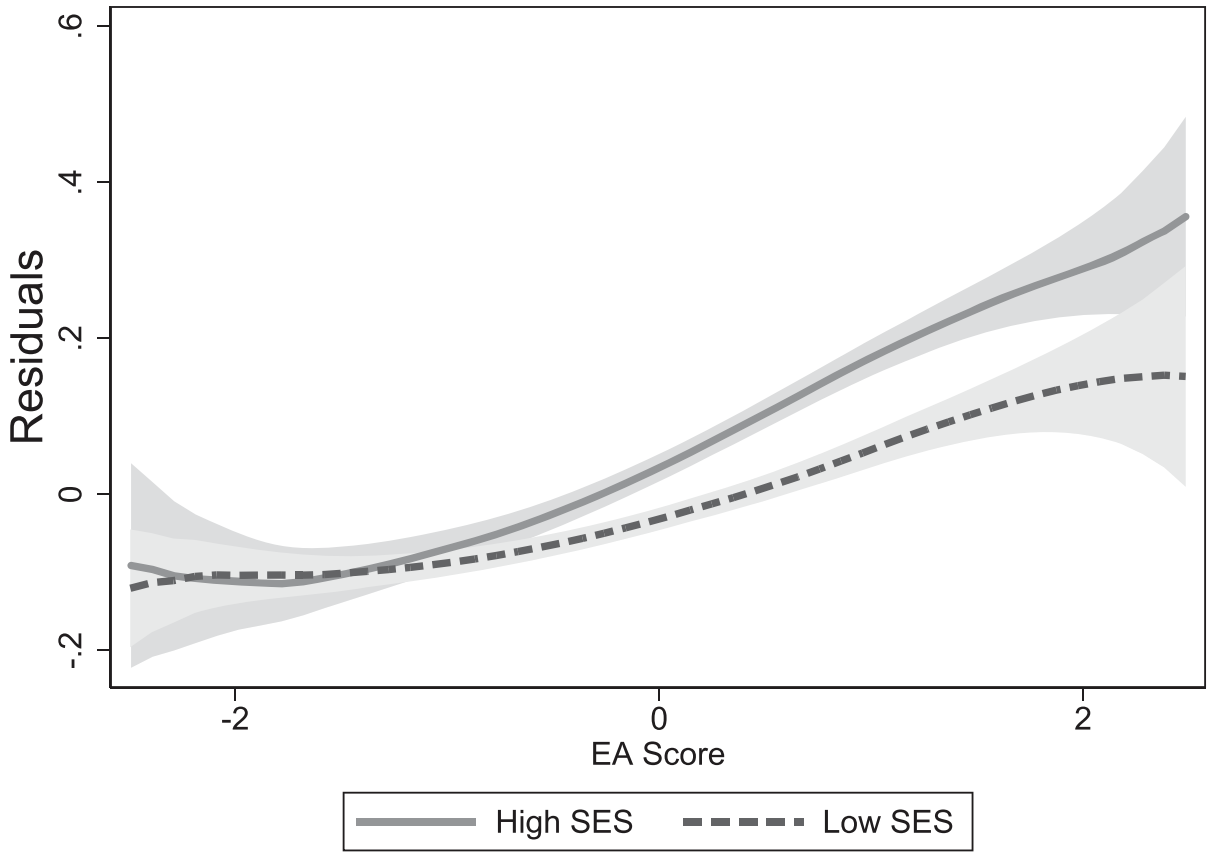
- We restrict attention to:
  - Genetic Europeans.
  - Born before 1965.
  - For Income Sample: Men earnings at least $10,000 (2010 dollars) in a person-year, ages 25-64
  - For Wealth Sample: Retired in 1996, 1998, 2002-2012, ages 65-75
- Resulting sample sizes:
  - 8,537 individuals (men and women) in cross-sectional sample.
  - 3,140 men in the SSA Earnings Sample.
  - 2,590 households and 5,701 household-year observations.



Notes: EA Score Distribution among HRS Individuals.

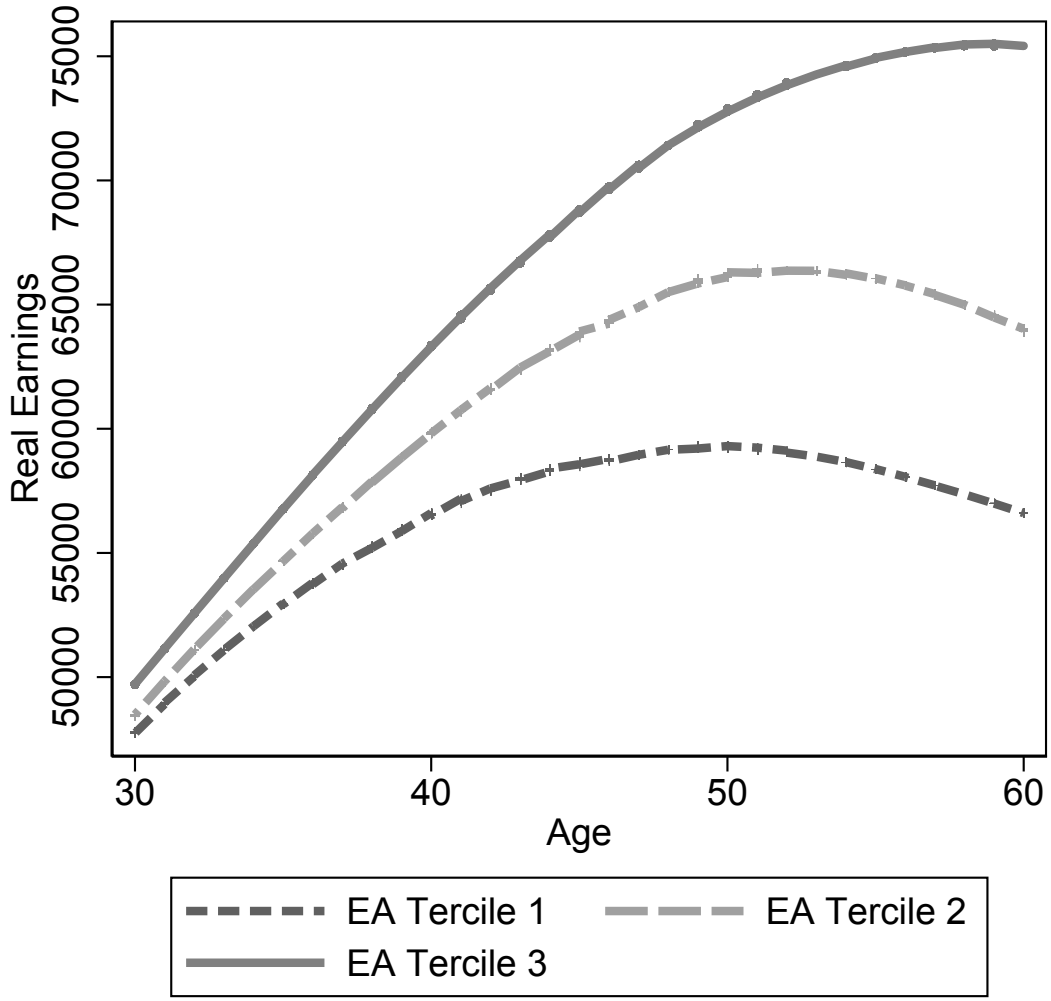
Polygenic Score and Educational Attainment

	(1)	(2)	(3)	(4)	(5)
EA Score	0.844***	0.614***	0.610***	0.589***	0.587***
	(0.046)	(0.043)	(0.043)	(0.045)	(0.032)
Father Educ		0.147***	0.144***	0.107***	0.109***
		(0.013)	(0.013)	(0.016)	(0.013)
Mother Educ		0.172***	0.170***	0.149***	0.150***
		(0.016)	(0.016)	(0.016)	(0.015)
Child Health: Very Good			-0.141	-0.100	-0.128*
			(0.126)	(0.116)	(0.070)
Child Health: Good			-0.259**	-0.190	-0.422***
			(0.127)	(0.123)	(0.090)
Child Health: Fair			-0.197	-0.114	-0.407***
			(0.168)	(0.175)	(0.145)
Child Health: Poor			-0.651	-0.549	-0.853
			(0.579)	(0.572)	(0.573)
Child Health: Missing			1.561***	1.054	1.995
			(0.415)	(1.159)	(1.243)
Obs.	8537	8537	8537	8537	8537
R^2	0.253	0.361	0.363	0.380	0.515
Child SES Measures	N	N	N	Y	Y
Child Region	N	N	N	N	Y
Religion	N	N	N	N	Y
Incr. R^2, EA score	0.075	0.038	0.037	0.034	0.034

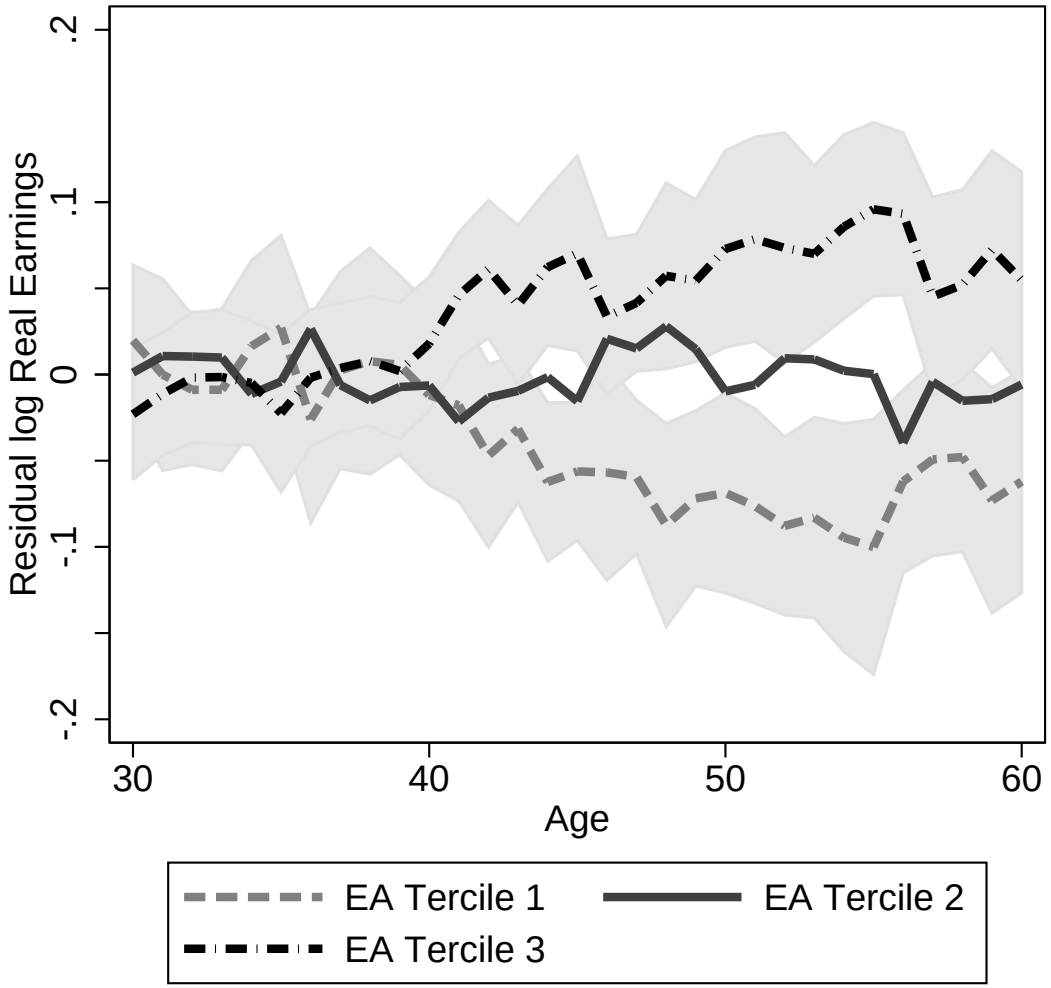


Panel (A) SES Measure: Father's Income

- Relationship between EA PGS and College Completion.
- Offers an example of a Gene-by-Environment Interaction



Panel (A) Earnings Over the Life-Cycle by EA Score Terciles.



Panel (B) Residual log Earnings Over the Life-Cycle by EA Score Terciles

Polygenic Score and Earnings

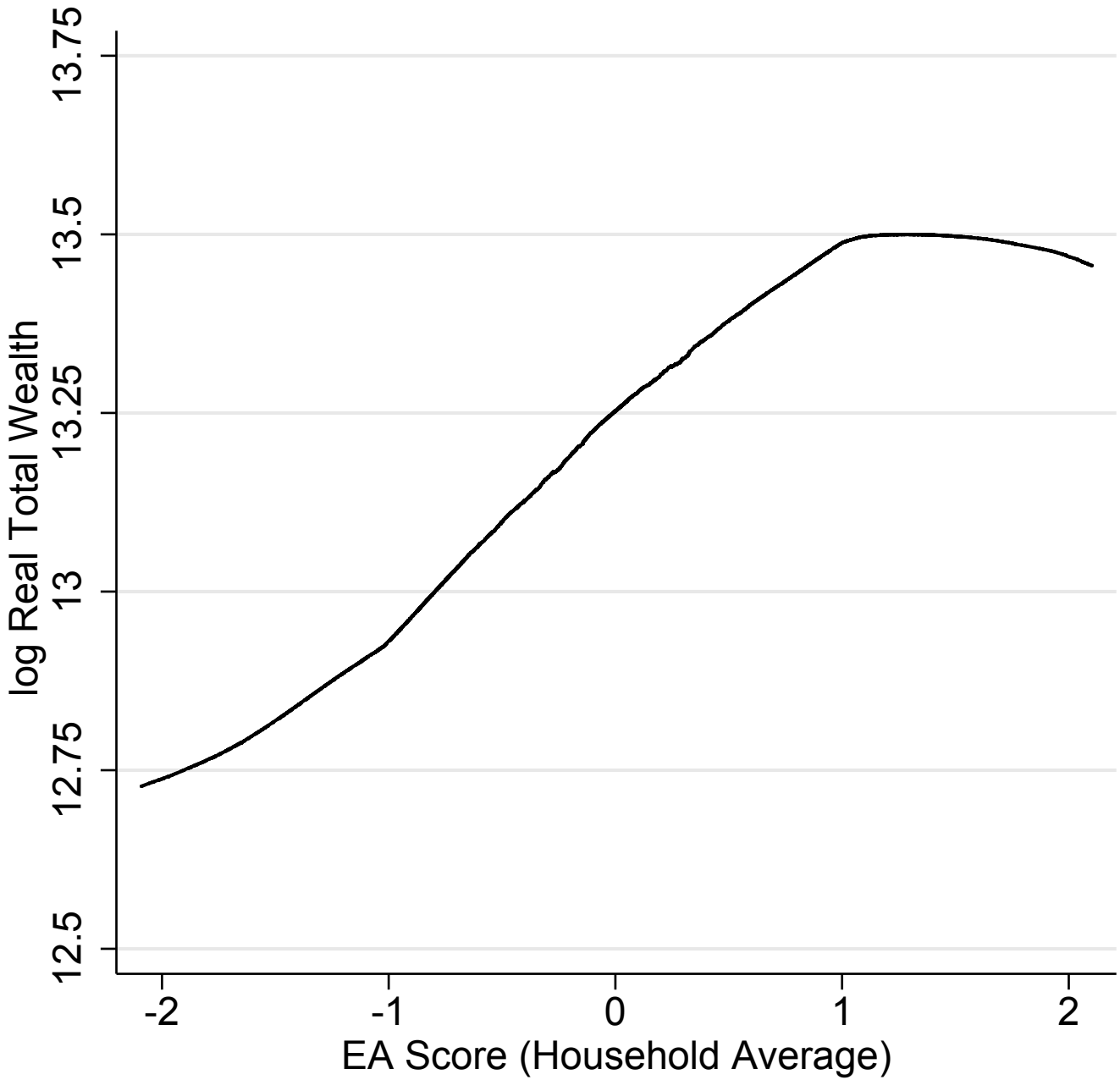
Panel A: Log Earnings				
Basic Specifications	(1)	(2)	(3)	(4)
EA Score	0.079***	0.032***	0.025**	0.041***
	(0.009)	(0.009)	(0.010)	(0.011)
EA Score x College			0.016	
			(0.020)	
Obs.	96721	96721	96510	57469
R^2	0.143	0.189	0.192	0.150
Age Group	25-64	25-64	25-64	40-64
Period	All Years	All Years	All Years	All Years
Educ. Controls	N	Y	Y	Y
Parent Controls	N	Y	Y	Y

Polygenic Score and Earnings

Panel B: Log Earnings By Time and Cohorts	(1)	(2)	(3)	(4)	(5)
EA Score	-0.010	0.009	0.018**	0.026***	0.011
	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
EA Score x Post 1980	0.077***	0.039***			0.043***
	(0.013)	(0.013)			(0.010)
EA Score x BY $>$ 1942			0.031*	0.009	-0.010
			(0.019)	(0.019)	(0.019)
College x Post 1980		0.276***			0.256***
		(0.031)			(0.024)
College x BY $>$ 1942				0.152***	0.041
				(0.045)	(0.044)
Obs.	96721	96510	96721	96510	96510
R^2	0.194	0.204	0.192	0.196	0.206
Ed. Groups	All	All	All	All	All
Period	All Years	All Years	All Years	All Years	All Years
Educ. Controls	Y	Y	Y	Y	Y
Parent Controls	Y	Y	Y	Y	Y

Results on Wealth (Barth, Papageorge and Thom 2020).

- Given the results with earnings, we expect a relationship between the EA Score and household wealth
- Question - are there other channels besides earnings that might link the two?
- Complication: wealth is a household-level outcome. We consider household average of the EA score (renormalized to have mean 0, variance 1)
- We construct a measure of total household financial wealth, including the present discounted value of annuity and defined benefit pension flows.


log Real Total Wealth
12.5
12.75
13
13.25
13.5
13.75
-2
-1
0
1
2
EA Score (Household Average)

Average Household EA Score and Household Wealth

Dep. Var: **Log Wealth**	[1]	[2]	[3]	[4]	[5]	[6]	[7]
EA Score	0.246***	0.221***	0.218***	0.085***	0.070***	0.179***	0.047**
	(0.022)	(0.020)	(0.020)	(0.021)	(0.023)	(0.020)	(0.022)
Male Educ				0.061***			
				(0.009)			
Female Educ				0.122***			
				(0.010)			
Log Income						0.316***	0.263***
						(0.039)	(0.038)
Obs.	5621	5621	5621	5621	5621	5308	5308
R^2	0.054	0.251	0.279	0.368	0.435	0.349	0.479
Standard Controls		X	X	X	X	X	X
Principal Comp.			X	X	X	X	X
Years of Educ.				X			
Full Educ. Controls					X		X

Average Household EA Score and Portfolio Decisions

Panel A	**Owns**	**Owns**	**Owns**	**Owns**	**Owns**	**Owns**
Dep. Var:	**House**	**Business**	**Stocks**	**House**	**Business**	**Stocks**
	[1]	[2]	[3]	[4]	[5]	[6]
EA Score	0.003	0.005	0.052***	-0.008	-0.001	0.040***
	(0.008)	(0.006)	(0.011)	(0.008)	(0.006)	(0.011)
Log Income	0.033***	-0.004	0.062***	0.002	-0.021**	0.021
	(0.008)	(0.006)	(0.011)	(0.008)	(0.008)	(0.013)
Lagged Log Wealth				0.122***	0.047***	0.151***
				(0.009)	(0.007)	(0.016)
Obs.	6460	6460	5450	4649	4649	4196
R^2	0.304	0.160	0.348	0.399	0.217	0.435
Mean outcome	0.84	0.08	0.46	0.83	0.08	0.47
Standard Controls	X	X	X	X	X	X
Principal Comp.	X	X	X	X	X	X
Full Educ. Controls	X	X	X	X	X	X

Average Household EA Score and Portfolio Decisions

Panel B Dep. Var: Log Wealth	[1]	[2]	[3]	[4]	[5]
EA Score	0.049**	0.046**	0.046**	0.016	0.018
	(0.023)	(0.021)	(0.022)	(0.021)	(0.019)
Owns Stocks				0.624***	0.507***
				(0.034)	(0.029)
Has Business			0.594***		0.530***
			(0.049)		(0.044)
Owns Home		0.887***			0.741***
		(0.054)			(0.052)
Obs.	4912	4912	4912	4912	4912
R^2	0.487	0.551	0.504	0.540	0.599
Standard Controls	X	X	X	X	X
Principal Comp.	X	X	X	X	X
Full Educ. Controls	X	X	X	X	X
Log Income	X	X	X	X	X

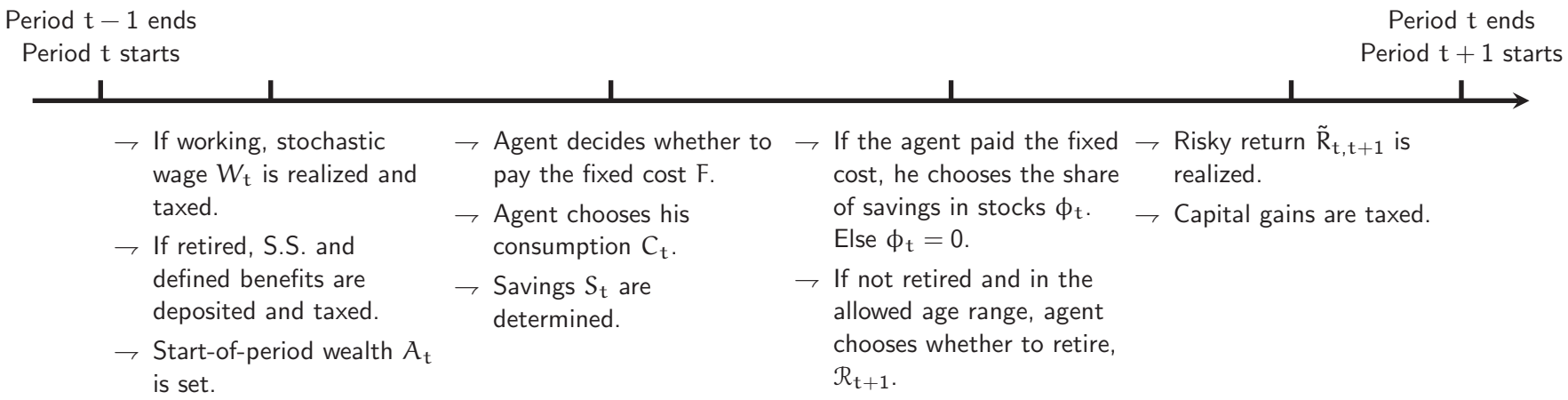
Pensions and Household Wealth

Dep. Var:	**Has Pension** [1]	**Pension Wealth** [2]	**Log Wealth** [3]	**Log Wealth** [4]
EA Score	0.003	0.030	0.069***	0.125***
	(0.011)	(0.035)	(0.022)	(0.035)
DB Pension			0.385***	0.181***
			(0.035)	(0.051)
EA Score x DB Pension				-0.096***
				(0.036)
Obs.	5621	3226	5621	5621
R^2	0.215	0.400	0.460	0.474
Mean outcome	0.57	$234,021		
Standard Controls	X	X	X	X
Principal Comp.	X	X	X	X
Full Educ. Controls	X	X	X	X

EA PGS in a Life-Cycle Model

- Barth et al (2022): Incorporates genetic variation (from PGS for Education) into a life-cycle model of income dynamics, savings, portfolio choice, retirement.
- Better understand mechanisms through which basic associations arise.
- Perform theoreticallyinformed gene-by-environment analysis:
  - Ex-ante GxE
  - Get at GxE in *welfare*

Summary and timing of the household's problem.



Where the EA score shows up.

- Wages.

$$\ln \tilde{W}_{i,t} = f(\text{Age}_{i,t}, \text{EA}_i, \text{Coll}_i, \text{SES}_i, \text{DB}_i, \text{Year}_t, \text{Unemp}_t) + \mathcal{Z}_i^w + \epsilon_{i,t}^w$$

- Fixed costs of stock market participation.

$$F_i = \exp\{f_0 + f_c \times \text{Coll}_i + f_g \times \text{EA}_i + \zeta_i^f\}$$

- Stock market returns.

$$\ln \tilde{R}_{it} = \ln R_t^{\text{SP500}} - \mu^{\text{SP500}} \times \underbrace{\text{Logistic}(r_0 + r_c \times \text{Coll}_i + r_g \times \text{EA}_i + \zeta_i^r)}_{\text{Inefficiency}_i}$$

- Additive utility cost of labor.

$$d_{i,t} = d_0 + d_{\text{Coll}} \times \text{Coll}_i + d_{\text{EA}} \times \text{EA}_i + d_{\text{Age}} \times \max\{\text{Age}_{i,t} - 50, 0\}$$

Participation cost

$\ln F_i = f_0 + f_{Coll} \times Coll_i + f_{EA} \times EA_i + \zeta_i^F$

f_0	f_{Coll}	f_{EA}
−0.9867	0.0311	0.0066
(0.2092)	(0.0369)	(0.0143)

Risky asset returns

$\ln \tilde{R}_{i,t} = \ln R_t^{SP500} - \mu^{SP500} \times g(r_0 + r_{Coll} \times Coll_i + r_{EA} \times EA_i + \zeta_i^R)$

r_0	r_{Coll}	r_{EA}
−0.0366	−1.1055	−0.6610
(0.0608)	(0.2568)	(0.1326)

Disutility from work

$d_{i,t} = d_0 + d_{Coll} \times Coll_i + d_{EA} \times EA_i + d_{Age} \times \max\{Age_{i,t} - 50, 0\}$

d_0	d_{Coll}	d_{EA}	d_{Age}
0.3961	−0.0052	−0.0033	−0.0241
(0.0816)	(0.0015)	(0.0009)	(0.0062)

Unobserved heterogeneity

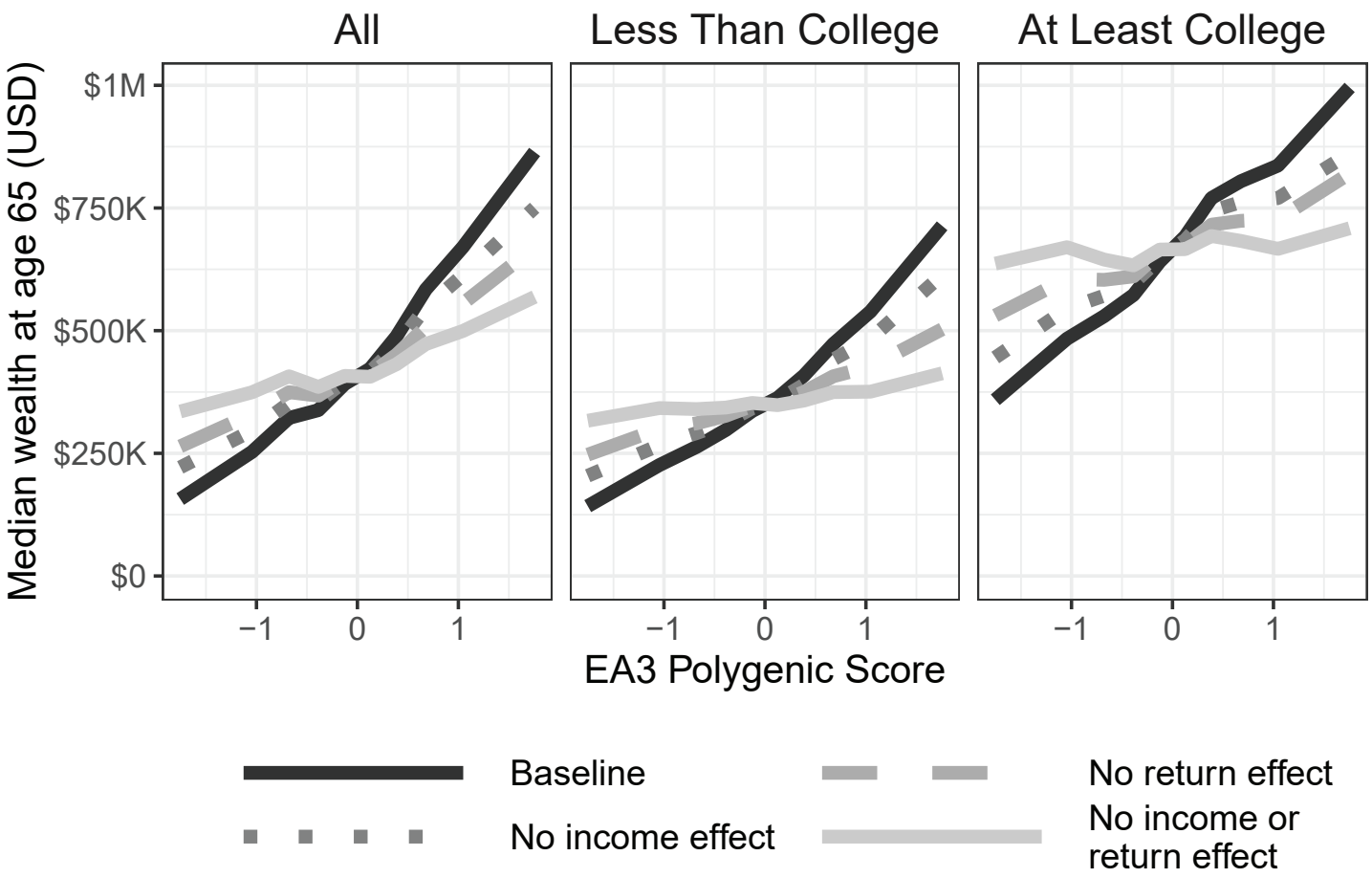
$\vec{\zeta}_i = \vec{z} \times SES_i + \vec{\mathcal{Z}}_i$

$\ln \sigma(\mathcal{Z}^F)$	$\ln \sigma(\mathcal{Z}^r)$	z_F	z_R
1.1838	−4.0326	−0.0434	−0.7250
(0.5698)	(1.3845)	(0.0475)	(0.1451)

Bequest motive

$\varphi(S_{i,t}) = \theta(S_{i,t} + \kappa)^{1-\omega}/(1-\omega)$

$\ln \kappa$	$\ln \theta$
7.0638	6.9423
(0.2474)	(0.5353)


All
Less Than College
At Least College
Median wealth at age 65 (USD)
$1M
$750K
$500K
$250K
$0
−1
0
1
EA3 Polygenic Score
Baseline
No income effect
No return effect
No income or return effect

Two counterfactual policy experiments aimed at lowering costs arising from an aging population.

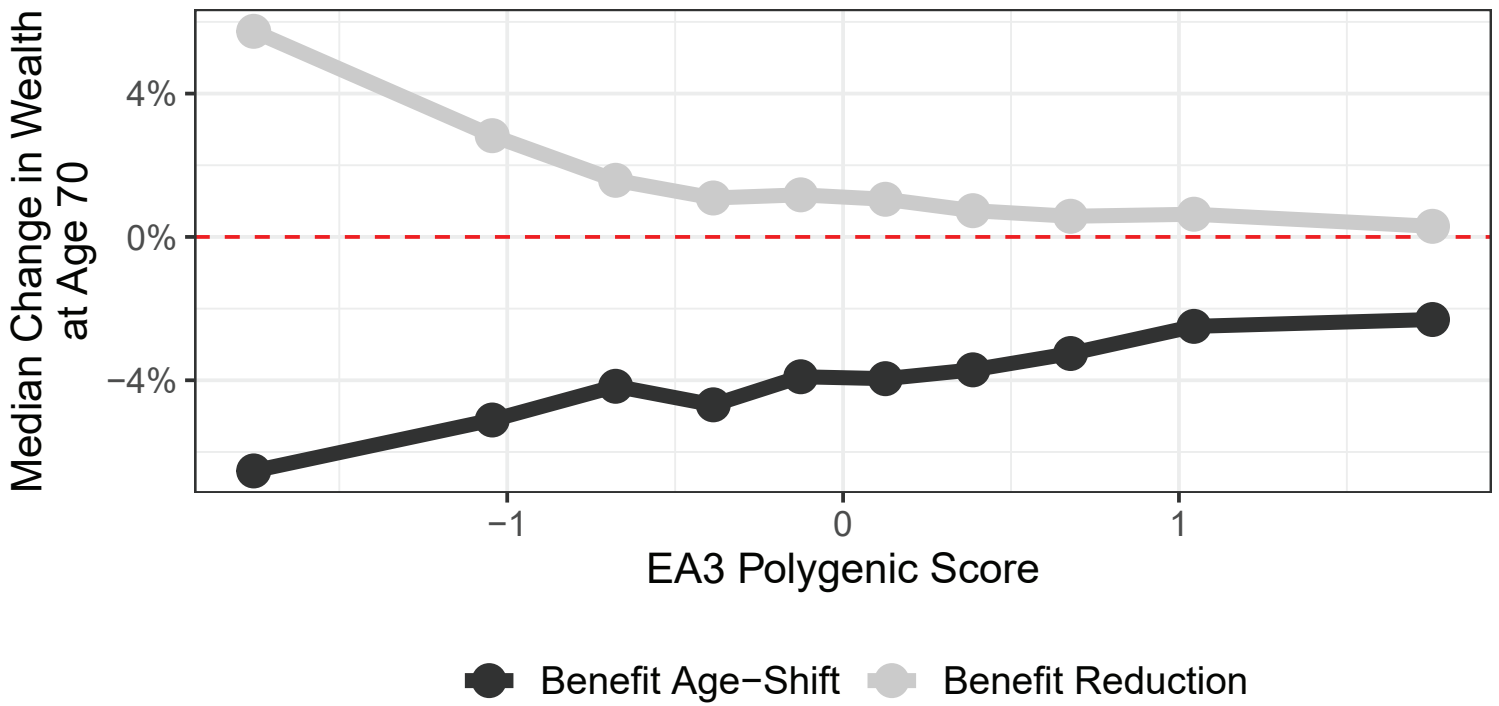
- Raise retirement age.
- Reduce social security.

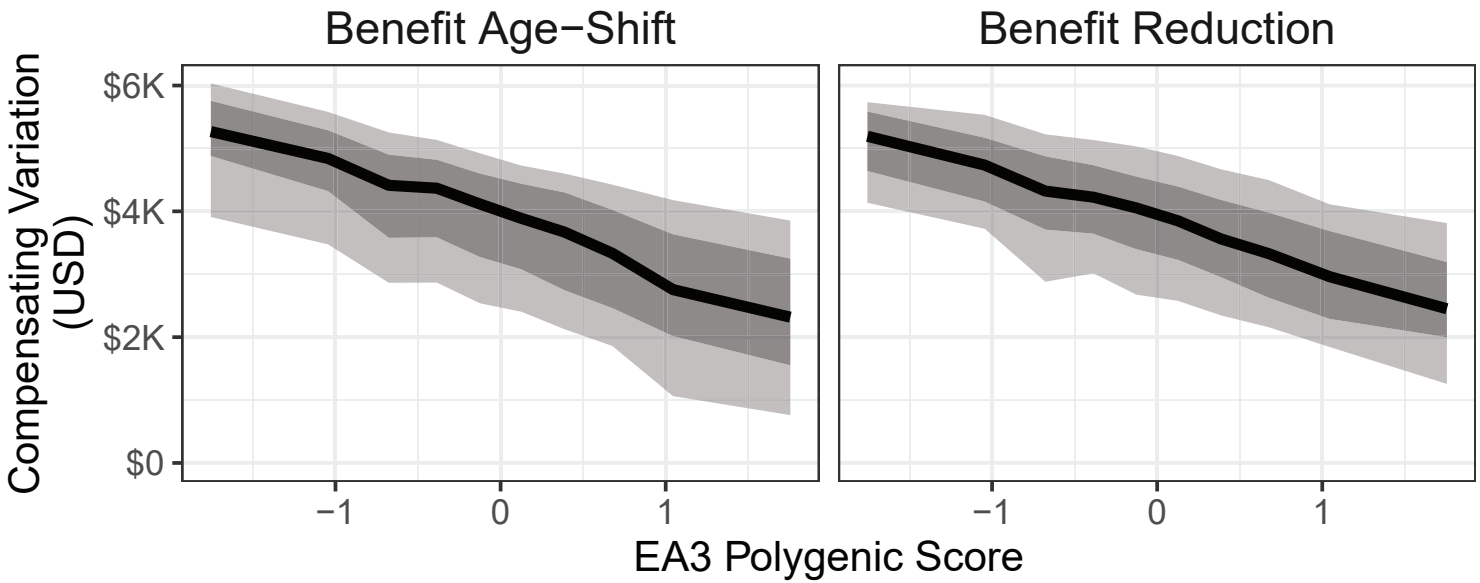
Counterfactual 1: Raise retirement age.

- Increase the earliest Social Security retirement age from 62 to 67, as has been proposed.
- Shift whole scheme for benefits 5 years forward.
- Full retirement age rises from from 67 to 72.

Counterfactual 2: Cut social security payments.

- Restore the retirement age and benefit schedule.
- Cut benefit amounts.
  - Find the reduction that makes revenue the same as in previous policy.
  - Current estimate: $\approx 29\%$ reduction.


Median Change in Wealth at Age 70
4%
0%
−4%
−1
0
1
EA3 Polygenic Score
Benefit Age−Shift
Benefit Reduction



Line: median. Inner shaded area: p25 - p75. Outer shaded area: p10 - p90.