Genetics and Economics: Models, Methods and Findings

Victor Ronda
Center for the Economics of Human Development

MAY 4, 2022

Main objectives:

- → Have an overall understanding of the literature on genetics and socio-economic outcomes.
- → Understand key findings and methodological approaches.
- → Understand the issues and limitations of current methods.

Why Genetics?

→ DNA differences makes us biologically distinct.

→ Genes are pre-determined at conception and randomly assigned (conditional on parents).

→ Genes are finite and fully measurable and have distinctive and well understood features.

Why should economists care?

- → Within family variation.
  - → Do parents allocate resources across siblings based on genetic differences?
  - → Why siblings respond differently to parental and schooling inputs?
- → Uncovering (previously unobserved) Heterogeneity.
  - → Can genetics help us measure heterogeneity in response to interventions or socio-economic conditions.
  - → Gene-by-environment interactions.
- → Biological mechanisms - Epigenetics.
  - → Can we identify underlying biological mechanisms and design policy interventions for the individuals 'at risk' ?

Different approaches

- → Kinship studies:
  - → genome is latent
  - → exploits genetic overlaps between twins, siblings, cousins, etc.
  - → second moments
- → Genomic studies:
  - → genome is observed
  - → exploits DNA data and in particular 10 million single nucleotide polymorphism (SNPs) variables → big data methods
  - → first moments
- → Most excitement developments are on the latter approach.

- Epigenetic studies
  - Organization and regulation of the DNA
  - Integration of genome-wide mapping of DNA methylation and histone modifications with RNA expression
  - Methods and findings are still in infancy.
  - No evidence of "transgenerational epigenetic inheritance", but strong evidence of environmental impact on epigenetics *within a lifetime*.
  - Epigenetics suggest a dynamic interaction between genes and environment in the spirit of the human capital formation.

- Genetics Background.
- Genetics Model.
- Gene Discovery (GWAS).
- Polygenic Scores.
- Epigenetics.

- → **Genetics Background.**
- → Genetics Model.
- → Gene Discovery (GWAS).
- → Polygenic Scores.
- → Epigenetics.

- **Genetics Background**
  - Human DNA.
  - Inheritance.
  - Molecular genetic data.

Human genome:

- 3 billion genetic addresses.
- In each address we observe a base nucleotide-pair:
  - Adenine-thymine pair (A or T).
  - Guanine-cytosine pair (G or C).
- The nucleotide-pair is fixed in 99% of such addresses.
- The remaining addresses are mostly biallelic.
- The DNA sequence is the sequence of these nucleotide-pairs.

▸ Details on Transcription and Translation ▸ Details on Genetic Variation

SNPs:

- → Most of the variation in the human genome comes from variation in a single base pair in a DNA sequence.
- → These base-pairs are called single-nucleotide polymorphisms (SNPs).
- → In humans there are 3 million to 10 million SNPs.
- → In humans, the vast majority of SNPs are biallelic.

Other Genetic variation consists of:

- → Indels: small insertions of deletions of base-pairs (1-10,000 base pairs).
- → Structural variants: insertion and deletions of large sections of the genome.
  - → $\sim$ 2000-2500 structural variants
  - → On average, affecting $\sim$ 20 million bases of sequence.

- → Almost all genetics papers being written in Economics and other Social Sciences focus on SNP variation.
- → SNPs are easy and cheap to measure,
- → They explain a large fraction of genetic influences,
- → And simplifies the technical notation and models.

Focusing on SNPs, the genetic endowment of individual i; hence, $\mathbf{g_i}$ is a vector of nucleotide base pairs:

$$\mathbf{g_i} = \{g_{i1}, ..., g_{iS}\} \tag{1}$$

where g_{is} is the base pair variant for individual i at position s, and S is the total number of SNPs, and

$$g_{is} \in \{0, 1, 2\} \tag{2}$$

Example: $g_{is}(\{AA\}) = 0$, $g_{is}(\{AT, TA\}) = 1$, and $g_{is}(\{TT\}) = 2$

- SNPs and other genetic variation are inherited from parents to children.
- Only a small fraction of genetic variation is determined after conception (de novo mutation)
- This has two consequences:
  - First, in expectation, the child's will have the same number of minor allele as their parents.
  - Second, genetic variation is randomly determined at the SNP level, conditional on the parents' DNA.

▸ Details on Cell Division and Inheritance

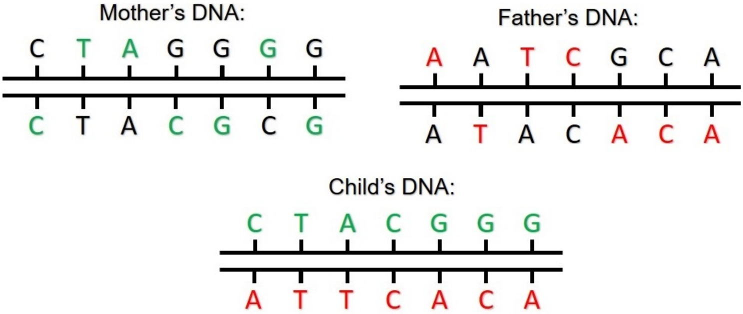
Formally, for each base pair s, we have that:

$$E[g_{is}] = 0.5g_{is}^{f} + 0.5g_{is}^{m} \tag{3}$$

where g_{is}^{f} is the minor allele frequency for the child's father at position s and g_{is}^{m} for the child's mother.

Importantly, $g_{is} - E[g_{is}]$ is random by nature and exogenous to any environmental effects before conception.

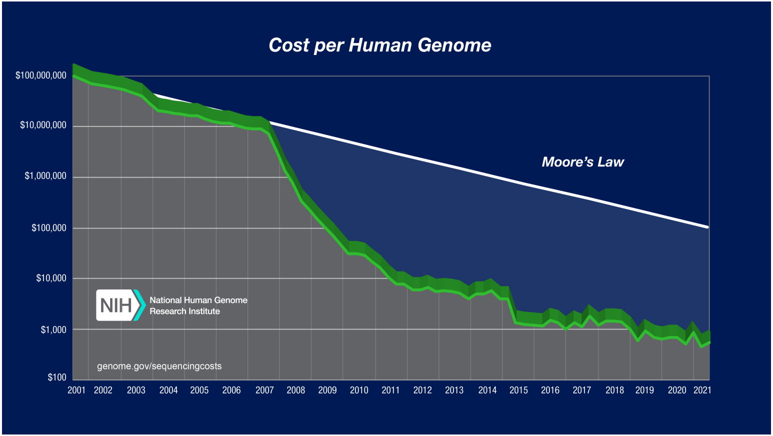
FIGURE: FAMILY GENETIC DATA - EXAMPLE



Genetic endowments:

- → Multidimensional, comprising of millions of individual variants.
- → Determined from the parental genetic pool and thus correlated with parental endowments.
- → Variation in child endowments conditional on parental endowments is random.

Rapid technological advances in measures genetic variation and falling costs of genotyping have made genomic data increasingly available in socioeconomic dataset



- Researchers collect tissue samples from blood (costly) or saliva (cheaper and more common).
- DNA is extracted from tissue and copies are made for analysis.
- Genotyped data is read from the DNA, two methods:
  - Genotyping (cheaper and more common): measure a set of SNPs specified on an “array”. Around 500k to 2.5M SNPs.
    - Current cost per individual $\sim \$50$
  - Sequencing (will become common in the future): Measure whole sequence of base pairs. Complete picture of genome.
    - High Accuracy: Current cost per individual $\sim \$1000$
    - Low Accuracy: Current cost per individual $\sim \$100$

- → Genotyping arrays usually only measure variation in SNPs.
- → Most of the molecular genetics research focus on SNP variation.
- → However, it is important to understand that genetic variation in humans goes beyond variation in SNPs.
- → Once sequencing becomes more common researchers will pay more attention to other rarer variation.

Datasets with molecular genetic data:

- → Medical studies with smaller sample sizes (N$<$5k).
    - → Formed to study a particular medical condition (e.g. cancer, depression, cardiovascular disease).
    - → e.g. Framingham Heart Study.
- → Social-science datasets (N = 5k-20k)
    - → Ideal for social-science research given extensive information on individuals.
    - → e.g. HRS, Add-Health, NCDS, Millennium Cohort Study.
    - → Some genetic data (e.g., polygenic scores) are publicly available on some of these datasets!

Datasets with molecular genetic data:

- Biobanks with very large sample sizes (N: 100k-1M)
  - Limited information on participants.
  - Formed to study medical conditions or to be included in discovery studies.
  - Eg. UK Biobank (500k), iPsych (Denmark, 160k), All of Us (US, 1-2M, in progress).
- Personal genomics datasets (N $>$ 500k)
  - Self-reported data where customers must consent to participate in research.
  - Ex: 23andMe, DeCode Genetics (all of Iceland)

- → Genetics Background.
- → **Genetics Model.**
- → Gene Discovery (GWAS).
- → Polygenic Scores.
- → Epigenetics.

A statistical framework

From a statistical perspective the role of genes in socioeconomic outcomes can be described by

$$\omega_i = m\left(\mathbf{g}_i, \mathbf{E}_i\right) + u_i \tag{4}$$

ω_i denote a characteristic of a trait of an individual i

$\mathbf{g}_i = \{g_{i1}, ..., g_{iS}\}$ the genotype of individual i

g_{is} is the base-pair variant s for individual i, and S is the number of base-pairs.

$\mathbf{E}_i$ denotes environmental factors such as family

$m(\cdot, \cdot)$ is an unknown function that captures the systematic information of the underlying phenomenon and

u_i is a regression error $E(u_i|\mathbf{g}_i, \mathbf{E}_i) = 0$.

The standard model imposes the following three non-innocuous assumptions

Assumption A.1. No gene-environment interaction

$$m\left(\mathbf{g}_i, \mathbf{E}_i\right) = m_G\left(\mathbf{g}_i\right) + m_E\left(\mathbf{E}_i\right). \tag{5}$$

Assumption A.2. No epistasis (no genetic interactions).

$$m_G\left(\mathbf{g}_i\right) = \sum_{s=1}^{S} m_{G_s}\left(g_{is}\right) \tag{6}$$

Assumption A.3. Additive effects (no dominance effects)

$$m_{G_s}\left(g_{is}\right) = \beta_s g_{is} \tag{7}$$

if g_{is} are individual SNPs, then g_{is} measures the number of minor alleles ($g_{is} \in \{0, 1, 2\}$) and β_s captures the predictive ability of increasing SNP s by one additional minor allele.

Standard model in molecular genetics

Then under Assumptions A.1-A.3 the standard model in molecular genetics is given by

$$\omega_i = \sum_{s=1}^{S} \beta_s g_{is} + m_E\left(\mathbf{E}_i\right) + u_i \tag{8}$$

- → Note: This framework is used on a variety of applications, from GWAS to twin studies and the construction of polygenic scores.
- → Note 2: Assumption 2 and 3 have empirical backing but Assumption 1 does not!

Additive Variance (Assumptions 2 and 3): Hill, Goddard and Visscher, 2008 ▸ Link

- → Empirical evidence and theory describe that most genetic variance can be explained by the additive component.
- → Additive variance typically accounts for over half, and often close to 100% of the total genetic variance.
- → Intuition:
  - → For most variants, minor allele frequency is very small (< 0.01).
  - → i.e. the distribution of minor allele frequencies is L-shaped.
  - → As a result, Epistasis (genetic interaction) and dominance (gene x gene) components have very low frequencies.
  - → See http://cnsgenomics.com/shiny/Falconer/

Additive Model in Practice:

- → Heritability: $h^2 = \frac{var(\sum_s \beta_s g_{is})}{var(\omega_i)}$
- → GWAS: Estimating $\hat{B}_s$
- → Polygenic Scores: $PGS = \sum_s \hat{\beta}_s g_{is}$

How can we recover these genetic effects?

- → Study a population where $\mathbf{g_i}$ is randomly assigned.
  - → Lab experiments in animals (e.g. mice or bees).
  - → This is done via gene-editing and or animal breeding.
  - → Not ethical in humans.
- → Find set of control variables $\mathbf{E}_i$ such that $\mathbf{g_i}$ is as good as random conditional on $\mathbf{E}_i$.
  - → Conditional on parental genes, variation in child's genes are random!
  - → As a result, we can exploit variation in inheritance from parents to children or across siblings to recover these effects!

- → Genetics Background.
- → Genetics Model.
- → **Gene Discovery (GWAS)**.
- → Polygenic Scores.
- → Epigenetics.

Conceptually, this area of research is interested in the average effect of changing a genetic variant at conception while keeping everything else constant.

$$\beta_s = \frac{\partial \omega_i}{\partial g_{is}} \tag{9}$$

where g_{is} is the sth variant of individual i, ω_i is the phenotype of interest

Assumptions A.1-A.3 imply that β_s is the same for all individuals, does not depend on the individual's genotype and on environmental influences.

Genome-wide association studies (GWAS) scan the entire genome for associations between genetic variants and socio-economic outcomes.

- → height (Allen et al., 2010, Wood et al., 2014)
- → BMI (Locke et al., 2015, Yengo et al., 2018)
- → depression (Wray et al., 2018)
- → intelligence (Sniekers et al., 2017)
- → educational attainment (Rietveld et al., 2013, Okbay et al., 2016, Lee et al., 2018)
- → risky behaviors (Karlsson Linner et al., 2018).
- → hundreds of traits from UKBiobank ▸ Link

The standard GWAS model is given by (5) using single SNPs (one at-a-time) and linear environmental controls

$$\omega_i = \beta_s^{GWAS} g_{is} + \gamma' \mathbf{E}_i + u_i \quad \forall s \in S \tag{10}$$

Each GWAS estimates a vector of s association coefficients $\beta^{GWAS} = \{\beta_s^{GWAS}\}_s^S$.

Key Challenges: GWAS studies face the following challenges

→ Ultra high-dimensionality of genetic data ($S > 1$ million)

→ Low explanatory power of single variants ($\beta_s \sim 0$)

→ Unobserved environmental effects correlated with the genotype (poor $\mathbf{E}_i$ controls)

→ Linkage disequilibrium (SNP correlation)

→ Unobserved or poorly measured phenotype ($\omega_i = \omega_i^T + \epsilon_i$)

Multiple Testing

- → The number of SNPs is much larger than the sample size.
- → The standard approach to deal with this problem is to consider the predictive ability of single SNPs (one at-a-time) and then correct for multiple testing.
- → There are about 10^6 tests in each GWAS study. Controlling the overall size of the test implies that one needs to lower the size of the test (α) of each individual test at very low levels.

- → Bonferroni correction is the most popular method: $\alpha' = \alpha/\#tests$;
- → It assumes that the GWAS association tests are independent.
- → Bonferroni correction slightly conservative but it becomes highly conservative with with vast loss of power when SNPs that are not truly independent. This can happen when many SNPs lie within regions of strong linkage disequilibrium (LD).

Side note: Why not LASSO or Machine Learning (in the past)?

- → GWAS requires large sample sizes.
- → Genetic data available in many small datasets.
- → These datasets cannot be combine due to IRB and privacy concerns.
- → No group has direct access to many of these datasets.
- → Consortia rely on data administrators with limited statistical knowledge.
- → Solution: GWAS parameters are estimated by meta-analysis using GWAS summary statistics (β_s^{GWAS}).

Side note:

- UK BioBank has recently released information on 500k individuals.
- Alternative methods now possible.
- I am unaware of any study that tried doing that.

The literature should pay more attention to alternative methods:

- Multiple testing
  - Permutation and bootstrapping methods when the entire genome is sequenced can account for the true dependence structure of the individual test statistic in order to obtain powerful tests without size distortions. However they are computationally intensive.
  - Chudnik, Kapetanios, Pesaran (ECTA, 2018)
- Multivariate methods
  - Lasso (Tibshirani, 1996)
  - Bayesian model averaging ((Raftery, 1995; Hoeting et al., 1999; Flutre et al. (2013))

Low Explanatory Power

$\rightharpoondown$ Each genetic variant explains only a very small percentage of the variation in the phenotype.

$\rightharpoondown$ This, combine with the fact that most SNPs have small allele frequencies, can lead to an under-estimation of the number of true associations in many studies.

$\rightharpoondown$ To see this, one can easily show that the power power to detect an association depends on three main things: the study sample size, the allele frequency of variant s and the true association of variant s with the phenotype of interest.

$$E[\chi_j^2] = Np_j(1 - p_j)\beta_s \tag{11}$$

$\rightharpoondown$ The most common solution seems to be to increase the sample size N via meta-analysis.

Unobserved Environmental Effects Correlated with the Genotype

- Typically, this problem refers to systematic differences in allele frequencies due to:
  - Population Stratification
    - Individuals are more related than other individuals either because of common ancestry among individuals or because some individuals are closely related.
    - The literature developed many methods to correct for population stratification; see Price (2010)
  - Genetic Nurture
    - Parental (or sibling) genes can directly influence the individual phenotype
    - E.g. parental genetic propensity for higher cognition influence quality and quantity of parental investments; see Houmark, Rosholm, and Ronda (2021)
    - Solution: directly controlling for parental genes in the GWAS analysis (within-family GWAS).

Unobserved Environmental Effects Correlated with the Genotype

- → Solution:
  - → Within-family GWAS (controlling for parental genotype on the GWAS)
    - → Conditional on parental SNPs, variation in child SNPs are random
    - → Main issue is sample size. Parental genes are rarely collected.
    - → Big focus on developments on this area.
    - → First paper on educational attainment EA4 (https://www.nature.com/articles/s41588-022-01016-z),
    - → About half of the associations are due to non-causal, environmental, effects.

Linkage Disequilibrium

- → It emphasizes genetic architecture.
- → The correlation between genotypes across two loci is called linkage disequilibrium (LD).
- → Reasons: Recent origin of a mutation, selection for certain alleles (assortative mating) or haplotypes, migration, genetic drift.
- → LD results in omitted variable bias since the standard GWAS approach tests one variant at a time.

Linkage Disequilibrium

→ Assuming no environmental confounds, such as population stratification, we have that

$$\beta_s^{GWAS} = \beta_s + \sum_{k=1}^{K} \beta_k r_{jk}^2 \tag{12}$$

where β_s is the effect of variant s on the phenotype of interest and r_{jk}^2 is a measure of linkage disequilibrium that captures the correlation between variants s and k. r_{jk}^2 can take values in the [0,1] range, where a value of 1 is generally described as perfect linkage disequilibrium

- Linkage disequilibrium is a major challenge for the discovery of functional links.

- A result of linkage disequilibrium, is that it is not possible to distinguish a functional link between variant s and phenotype ω from the link between variant s and variant S' that influences the phenotype.

- Thus, GWAS alone cannot inform about the biological function of different genetic variants but they can give a direction.

- Understanding of biological function requires further functional studies, possibly in animal studies.

Unobserved or poorly measured phenotype

- → Solution: Genomic SEM (Grotzinger et al., 2019)
- → Idea: GWAS of related traits provide information about genetic influences on the latent trait.
- → Genomic SEM is very flexible and useful beyond dealing with unobserved or poorly measured phenotypes.

Formally, let ω be the latent trait of interest, and m^k be a related observed phenotype, we can extend the GWAS model as follows:

$$
\begin{aligned}
m_i^k &= \lambda_k \omega_i + \epsilon_{ik} \forall k \in K \\
\omega_i &= \beta_s^{GWAS} g_{is} + \gamma' \mathbf{E}_i + u_i \forall s \in S
\end{aligned}
$$

β_s^{GWAS} can be obtained indirectly from the β_{sk}^{GWAS}, estimated for each related observed phenotype m^k, and the genetic covariance matrix across the observed phenotypes .

General Finding of GWAS

- → The key insight from GWAS studies is that human traits and behaviors are highly polygenic.

- → This insight lead to the so called “forth law of behavioral genetics” that “a typical human behavioral trait is associated with very many genetic variants, each of which accounts for a very small percentage of the behavioral variability.” (Chabris et al., 2015).

- → As an example, the most recent GWAS for educational attainment discovered 1,271 independent genetic markers associated with educational attainment each with very small effects (Lee et al., 2018).

EA1 - Rietveld et al., 2013 ▸ Link

- 41 datasets and N = 101,069
- Findings:
  - 1 genome-wide association with years of education.
  - 2 genome-wide associations with college attainment.
  - All 3 hit replicated in 12 independent samples with N=25,490.

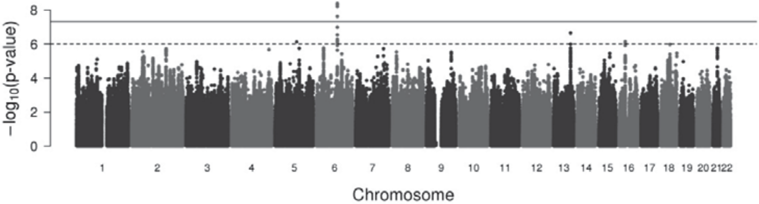


FIGURE: MANHATTAN PLOT FOR EDUCATIONAL ATTAINMENT: (Rietveld et al., 2013 ▸ Link)

EA2 - Okbay et al., 2016 ▸ Link

- 63 datasets and N = 293,723
- Findings:
  - 74 genome-wide associations with years of education.
  - Replication in UK Biobank (N = 110,000).
  - 72 out of 74 lead SNPs with consistent sign.
  - 52 significant (5% level) and 7 at genome-wide significant level.

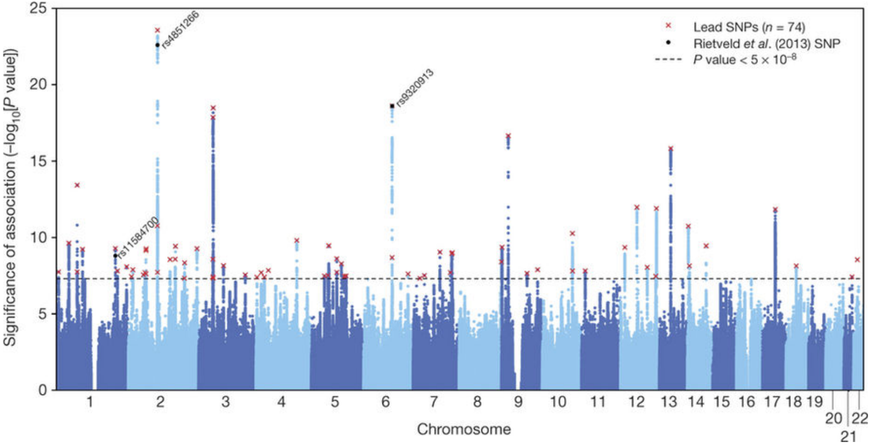


FIGURE: MANHATTAN PLOT FOR EDUCATIONAL ATTAINMENT:
(Okbay et.al. 2016)

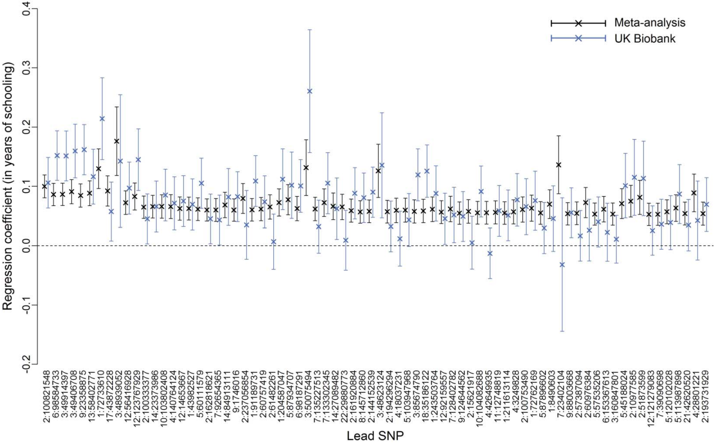


FIGURE: GENETIC EFFECTS: EA2 Replication (Okbay et.al. 2016

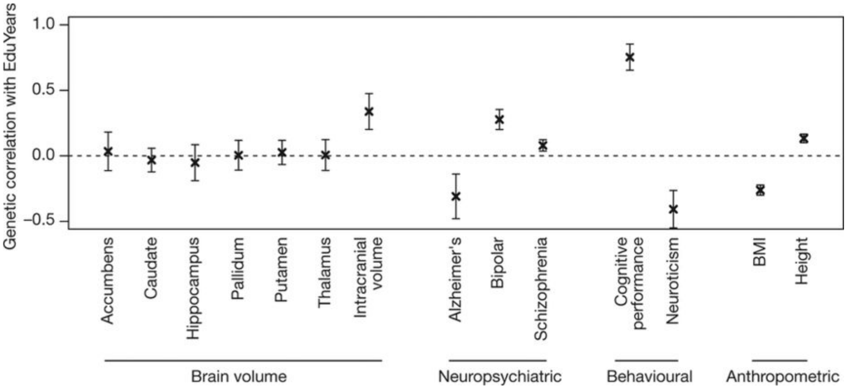


FIGURE: GENETIC CORRELATION: Educational Attainment and Other Phenotypes (Okbay et.al. 2016)

EA3 - Lee et al., 2018 ▸ Link

- → N = 1.1 million individuals
- → 1,271 genome-wide associations with years of education.

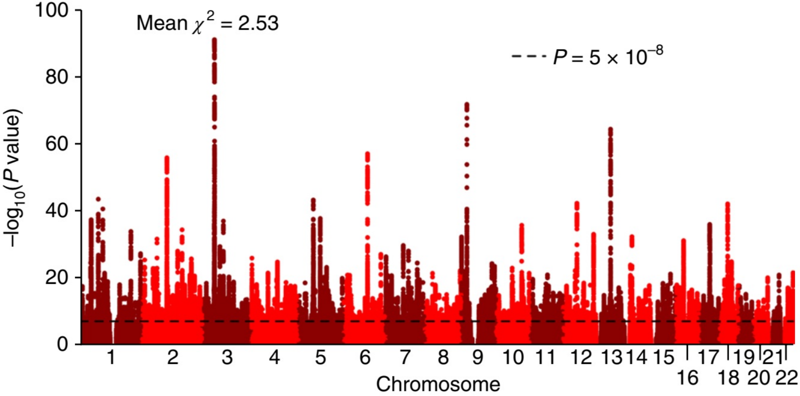


FIGURE: MANHATTAN PLOT FOR EDUCATIONAL ATTAINMENT: (Lee et.al. 2018)

EA4 - Okbay et al., 2022 ▸ Link

- N = 3 million individuals
- 3,952 genome-wide associations with years of education.

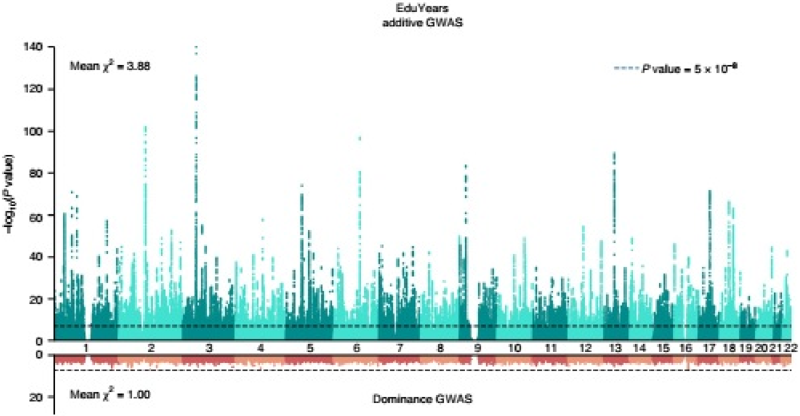


FIGURE: MANHATTAN PLOT FOR EDUCATIONAL ATTAINMENT: (Okbay et al., 2022)

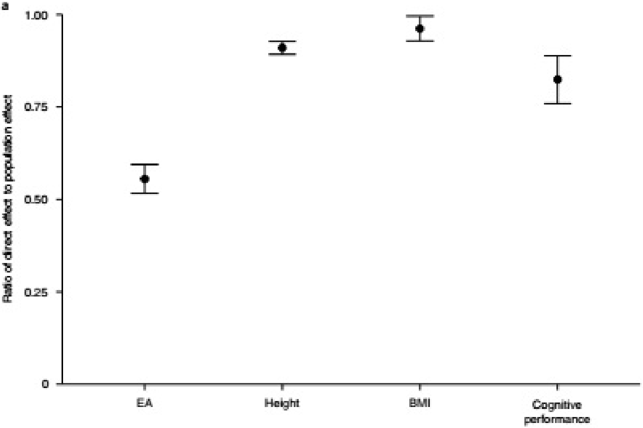


Figure: Share of direct vs non-direct effects: (Okbay et al., 2022)

- Genetics Background.
- Genetics Model.
- Gene Discovery (GWAS).
- **Polygenic Scores.**
- Epigenetics.

Polygenic scores (PGS) are constructed using transformed coefficients ($\tilde{B}_s$) that account for correlation across SNPs:

$$PGS_i^y = \sum_s \tilde{B}_s g_{is} \tag{13}$$

where the $\tilde{B}_s$ usually come from a GWAS study. (e.g. the EA PGS uses the betas from the EA GWAS.)

- Problem with PGS construction:
  - $\beta_s^{Gwas} \neq \beta_s$
  - GWAS parameter measures both causal effect of SNP s and effects of SNPs in linkage disequilibrium with SNP s.
- Solutions:
  - Prunning: Select only one SNP per causal loci. Selected SNPs should be uncorrelated.
  - LDpred: Bayesian method that accounts for linkage disequilibrium between SNPs.
- In both cases PGS will be measured with error - which decreases explanatory power.

Advantages of using PGSs:

- High explanatory power.
  - Educational Attainment PGS explains $\sim 10\%$ of the variation in Years of Education
- Out of sample reliability.
  - The same score has a similar effect across different samples/populations.

Disadvantages of using PGSs:

- Mechanisms not easily identifiable (what is being captured?).
- Same issues as in GWAS:
  - Parental genes as confounders (PGS capture environmental effects) - this is being solved with within-family GWAS analyses.
  - Relies on Assumptions A.1-A.3.

Educational attainment PGS

- → The vast majority of papers using polygenic scores in economics rely on the EA PGS.
- → It has a very high explanatory power for education, and thus explains a variety of socio-economic outcomes,
- → It is reliable and have a similar predictive power across countries and settings.

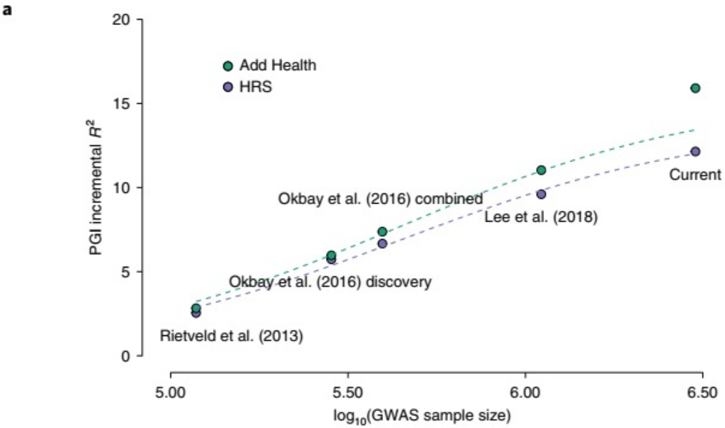


FIGURE: PREDICTIVE POWER EA OVERTIME: (Okbay et al., 2022)

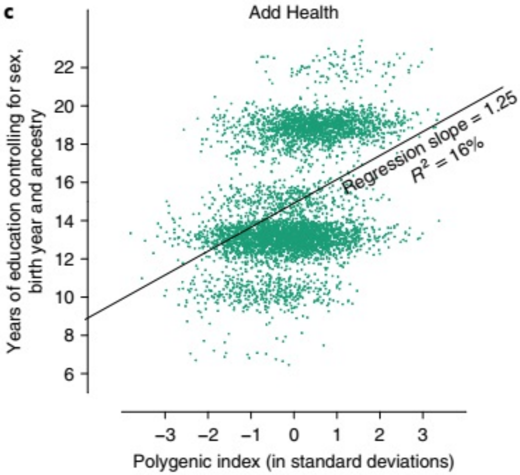


FIGURE: PREDICTIVE POWER OF EA4 ON YEARS OF EDUCATION: (Okbay et al., 2022)

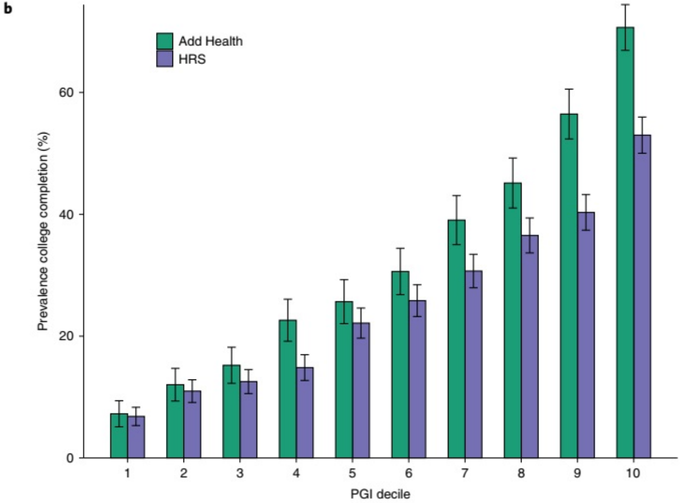


Figure: Predictive Power of EA4 on College Attainment: (Okbay et al., 2022)

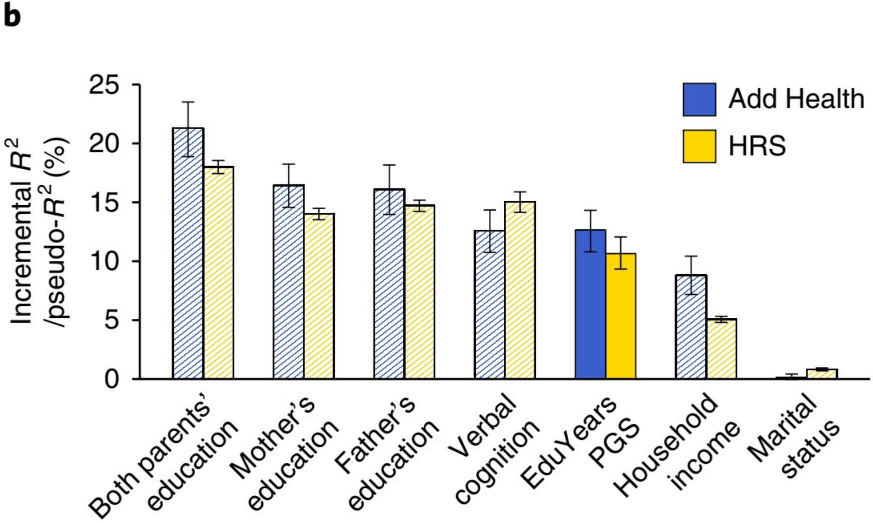
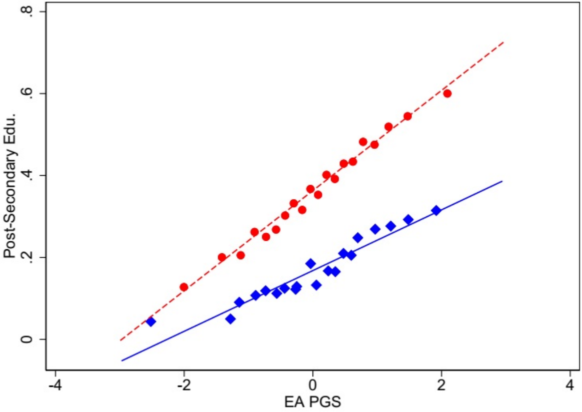


FIGURE: FIGURE FROM LEE ET AL., 2018 ▸ LINK



(a) Full Sample

FIGURE: EA3 AND COLLEGE FOR LOW/HIGH SES IN DENMARK - FIGURE FROM RONDA ET AL., 2019.

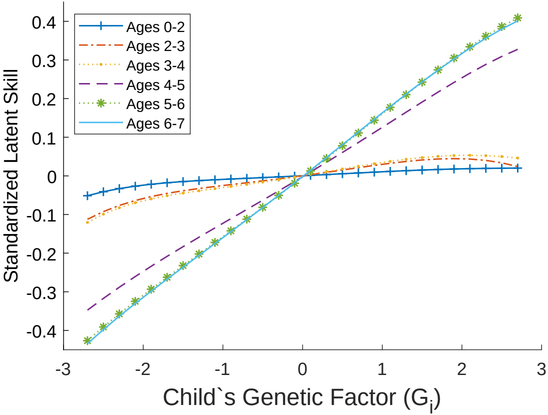


FIGURE: EA3 AND SKILL FORMATION IN THE UK - FIGURE FROM HOUMARK, ROSHOLM, AND RONDA (2021)

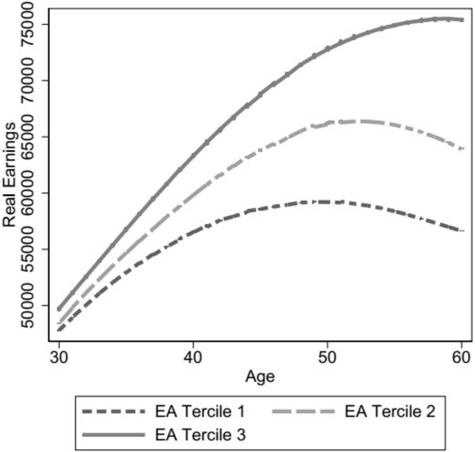


FIGURE: EA3 AND EARNINGS IN THE HRS - FIGURE FROM PAPAGEORGE AND THOM, 2017 ▸ LINK

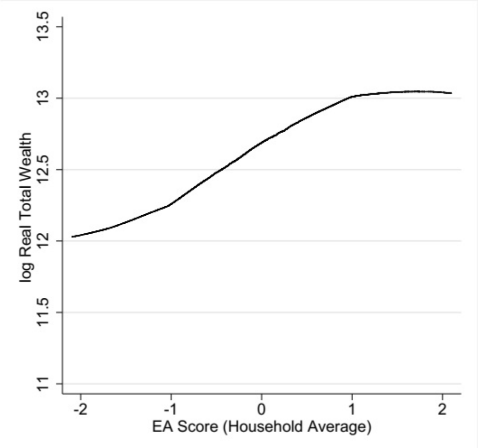


FIGURE: EA3 AND WEALTH IN THE HRS - FIGURE FROM BARTH, PAPAGEORGE AND THOM, 2019 ▸ LINK

→ PGSs for other traits also have high explanatory power.

→ Results also replicate across regions and settings.

→ E.g. polygenic score for body mass index (BMI).

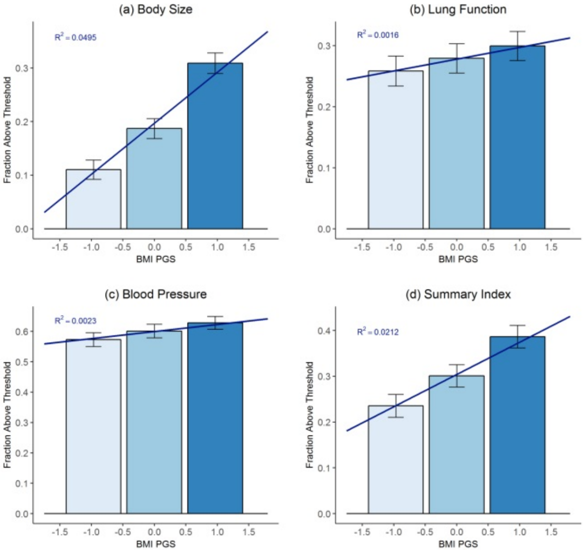


Figure: Figure from Barcellos et al., 2018 ▸ Link

GWAS weights and PGSs

1. Do not translate across ethnic groups.
2. Carry signal of population stratification.
3. Carry information about parental genes and influences.

GWAS weights and PGSs

1. Do not translate across ethnic groups.
   - → GWAS identifies SNPs that correlate with the outcome.
   - → SNP could be ‘causal’ - a change a conception would translate into a different phenotype.
   - → or, SNP could correlate with genetic variation that is ‘causal’.
   - → The main issue is that genetic correlation (linkage disequilibrium) is different across population groups.

Figure 1. Comparison of semi-partial R^2 for polygenic scores created from directly genotyped and imputed SNPs, by ethnicity and trait

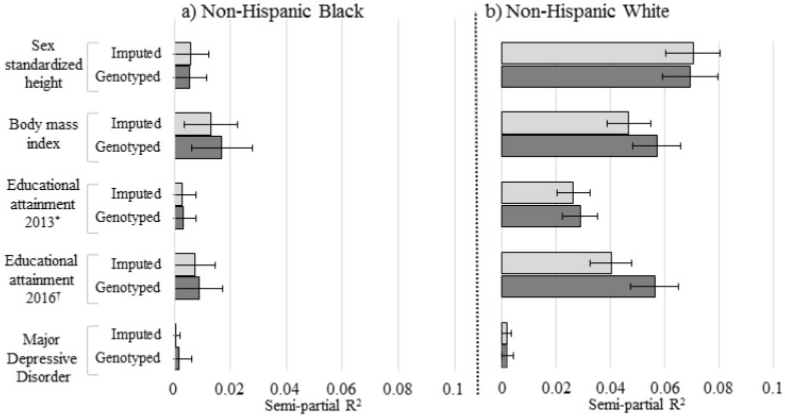


FIGURE: FIGURE FROM WARE ET AL., 2017 ▸ LINK

GWAS weights and PGSs

2 Carry signal of population stratification.

- → Phenotype and genotype vary across population groups.
- → If a genotype has higher frequency at one group that also has higher phenotype, GWAS will identify the genotype as having a positive effect.
- → This is specially concerning in consortium studies.
- → If population stratification is not properly corrected, PGSs will carry information about regional variation.

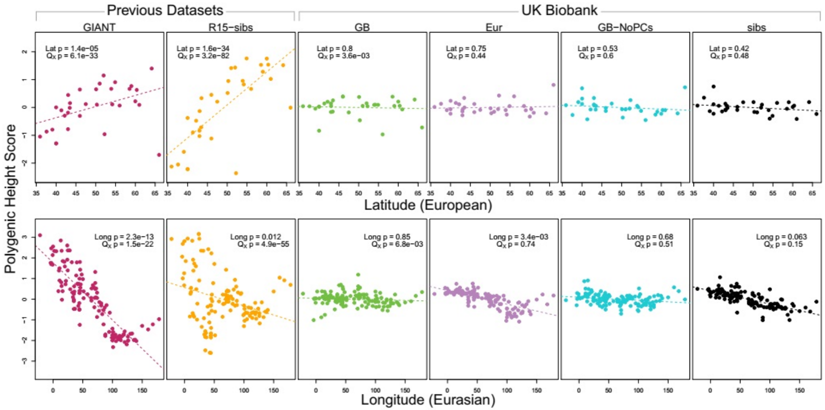


FIGURE: FIGURE FROM BERG ET AL., 2019 ▸ LINK

GWAS weights and PGSs

3 Carry information about parental genes and influences.
  - → Genes are determined from parental genetic pool.
  - → Parental genes influence the environment - “Genetic Nurture”.
  - → As a result, G could be capturing environmental effects.

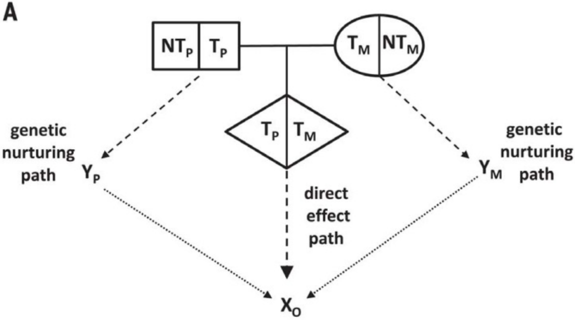


FIGURE: FIGURE FROM KONG ET AL., 2018 ▸ LINK

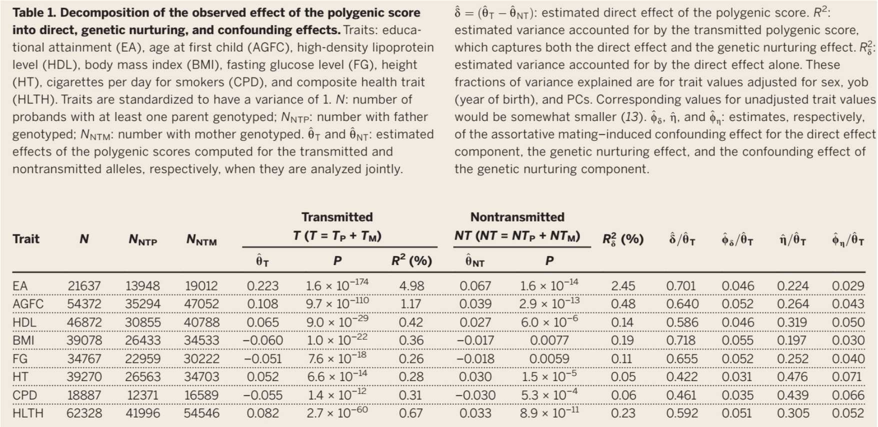


Table 1. Decomposition of the observed effect of the polygenic score into direct, genetic nurturing, and confounding effects. Traits: educational attainment (EA), age at first child (AGFC), high-density lipoprotein level (HDL), body mass index (BMI), fasting glucose level (FG), height (HT), cigarettes per day for smokers (CPD), and composite health trait (HLTH). Traits are standardized to have a variance of 1. N: number of probands with at least one parent genotyped; N_{NTP}: number with father genotyped; N_{NTM}: number with mother genotyped. $\hat{\theta}_{\mathrm{T}}$ and $\hat{\theta}_{\mathrm{NT}}$: estimated effects of the polygenic scores computed for the transmitted and nontransmitted alleles, respectively, when they are analyzed jointly. $\hat{\delta} = (\hat{\theta}_{\mathrm{T}} - \hat{\theta}_{\mathrm{NT}})$: estimated direct effect of the polygenic score. R^2: estimated variance accounted for by the transmitted polygenic score, which captures both the direct effect and the genetic nurturing effect. R^2_{δ}: estimated variance accounted for by the direct effect alone. These fractions of variance explained are for trait values adjusted for sex, yob (year of birth), and PCs. Corresponding values for unadjusted trait values would be somewhat smaller (*13*). $\hat{\phi}_{\delta}$, $\hat{\eta}$, and $\hat{\phi}_{\eta}$: estimates, respectively, of the assortative mating–induced confounding effect for the direct effect component, the genetic nurturing effect, and the confounding effect of the genetic nurturing component.

Trait	N	N_{NTP}	N_{NTM}	**Transmitted** T $(T = T_{\mathrm{P}} + T_{\mathrm{M}})$			**Nontransmitted** NT $(NT = NT_{\mathrm{P}} + NT_{\mathrm{M}})$		R^2_{δ} **(%)**	$\hat{\delta}/\hat{\theta}_{\mathrm{T}}$	$\hat{\phi}_{\delta}/\hat{\theta}_{\mathrm{T}}$	$\hat{\eta}/\hat{\theta}_{\mathrm{T}}$	$\hat{\phi}_{\eta}/\hat{\theta}_{\mathrm{T}}$
				$\hat{\theta}_{\mathrm{T}}$	P	R^2 **(%)**	$\hat{\theta}_{\mathrm{NT}}$	P					
EA	21637	13948	19012	0.223	1.6×10^{-174}	4.98	0.067	1.6×10^{-14}	2.45	0.701	0.046	0.224	0.029
AGFC	54372	35294	47052	0.108	9.7×10^{-110}	1.17	0.039	2.9×10^{-13}	0.48	0.640	0.052	0.264	0.043
HDL	46872	30855	40788	0.065	9.0×10^{-29}	0.42	0.027	6.0×10^{-6}	0.14	0.586	0.046	0.319	0.050
BMI	39078	26433	34533	−0.060	1.0×10^{-22}	0.36	−0.017	0.0077	0.19	0.718	0.055	0.197	0.030
FG	34767	22959	30222	−0.051	7.6×10^{-18}	0.26	−0.018	0.0059	0.11	0.655	0.052	0.252	0.040
HT	39270	26563	34703	0.052	6.6×10^{-14}	0.28	0.030	1.5×10^{-5}	0.05	0.422	0.031	0.476	0.071
CPD	18887	12371	16589	−0.055	1.4×10^{-12}	0.31	−0.030	5.3×10^{-4}	0.06	0.461	0.035	0.439	0.066
HLTH	62328	41996	54546	0.082	2.7×10^{-60}	0.67	0.033	8.9×10^{-11}	0.23	0.592	0.051	0.305	0.052

Figure: Table from Kong et al., 2018 ▸ Link

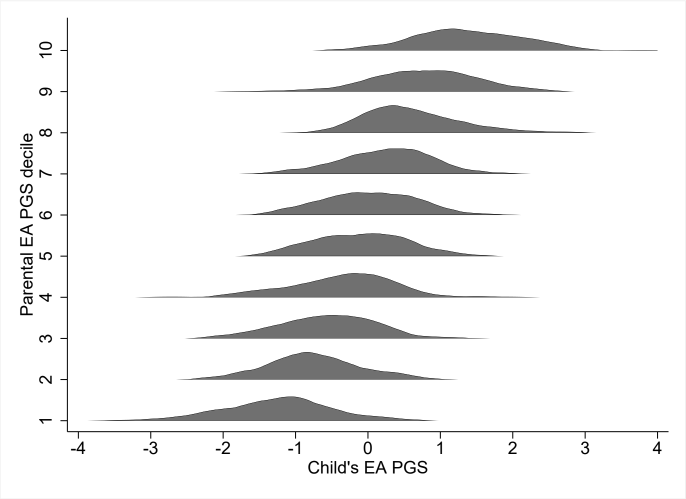


Figure: Figure from Houmark, Rosholm, and Ronda (2021)

In practice, two solutions:

1. Directly control for parental genes (PGS)
   - → e.g. Houmark, Rosholm, and Ronda (2021)
2. Exploit variation in sibling genes
   - → Within family analysis allow us to control for omitted parental genes.
   - → Siblings face the same genetic pool.
   - → Genetic differences between siblings are exogenous.
   - → e.g. Ronda et al., 2018.

Table: EA PGS and Skills by Age

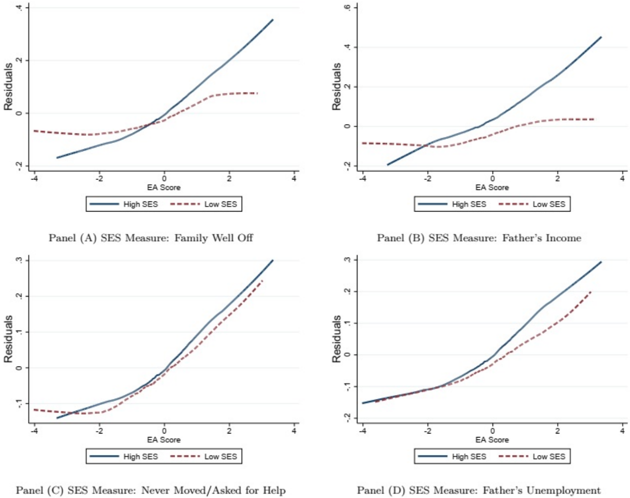
Ages:	[0-2]	[2-3]	[3-4]	[4-5]	[5-6]	[6-7]	[Pooled]
Panel A:							
Child's PGS	0.029	0.041	0.067	0.121	0.132	0.091	0.080
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.010)
N	4510	4510	4510	4510	4510	4510	27060
Panel B:							
Child's PGS	0.015	0.007	-0.013	0.071	0.076	0.050	0.034
	(0.028)	(0.027)	(0.027)	(0.027)	(0.027)	(0.028)	(0.018)
Mother's PGS	0.032	0.037	0.069	0.055	0.057	0.041	0.049
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.013)
Father's PGS	-0.006	0.019	0.061	0.028	0.034	0.027	0.027
	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.023)	(0.015)
N	4510	4510	4510	4510	4510	4510	27060

EA PGS and Child Skills - from Houmark, Rosholm, and Ronda (2021)

	(1)	(2)	(3)	(4)
Dep. Var.	Y.Edu.	P.S.E.	Danish	Math
EA PGS	0.561	0.114	6.248	6.722
	(0.053)	(0.010)	(0.469)	(0.558)
Family F.E.	(N)	(N)	(N)	(N)
EA PGS	0.296	0.069	2.774	3.616
	(0.094)	(0.020)	(0.842)	(0.982)
Family F.E.	(Y)	(Y)	(Y)	(Y)
N	1,487	1,487	1,838	1,793

Within Family: EA PGS and Human Capital Outcomes - from Ronda et al., 2019

- → Polygenic scores can also be extended to estimate gene-environment interactions.
- → One use is to estimate heterogeneous returns to genes across environments.
- → Another is to estimate heterogeneous responses to policies and interventions.



Panel (A) SES Measure: Family Well Off

Panel (B) SES Measure: Father's Income

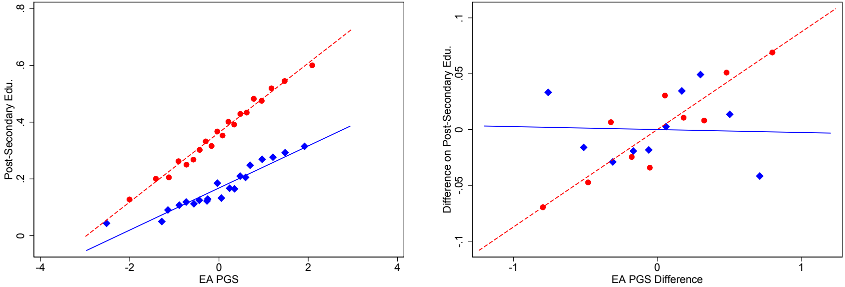
Panel (C) SES Measure: Never Moved/Asked for Help

Panel (D) SES Measure: Father's Unemployment

Figure 6: Non-parametric (Lowess) estimation relating the probability of completing a college degree

Figure: Figure from Papageorge and Thom, 2017 ▸ Link

FIGURE: Post-Secondary Education by Disadvantage



(a) Between-Family

(b) Sibling Differences

FIGURE: *Notes*: Disadvantaged group in blue and non-disadvantage in red. Dots represent mean outcome values across PGS percentile. - from Ronda et al., 2019

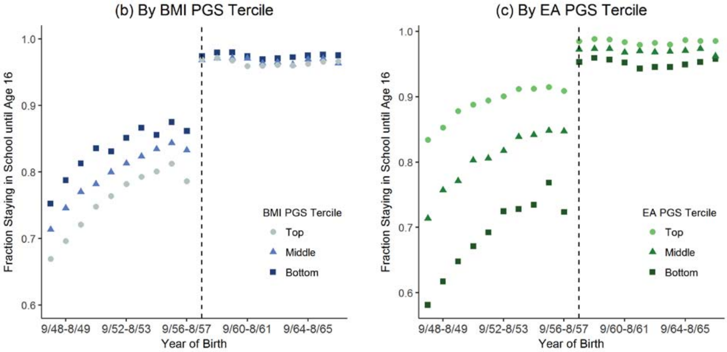


Fig. 2. Fraction staying in school until age 16 by year of birth for **(a)** full sample, **(b)** bottom, middle, and top terciles of the BMI PGS distribution, and **(c)** bottom, middle, and top terciles of the EA PGS distribution. Dashed vertical lines mark the first birth cohort affected by the raising of the school-leaving age from 15 to 16.

FIGURE: FIGURE FROM BARCELLOS ET AL., 2018 ▸ LINK

- → Polygenic Scores are a powerful tool to understand genetic influences.
- → The educational attainment polygenic score is the most widely used due to its high reliability and predictive power.
- → Research on this area is still at is infancy. There is a lot of untapped questions and low-hanging fruits.
- → Many of the issues with GWAS and polygenic scores are being addressed.
- → I am hopeful for the near future!

- → Genetics Background.
- → Genetics Model.
- → Gene Discovery (GWAS).
- → Polygenic Scores.
- → **Epigenetics.**

Epigenetics

- The study of cellular and physiological trait variations **not** caused by changes in DNA.
- Changes induced by environmental factors.
- Epigenetic changes were thought to be not heritable - this view is currently being challenged.
- Examples:
  - Methylation (most common)
  - Acetylation
  - Small non-coding RNAs.

- Epigenetics is a new and exciting field of research as it allows for dynamic interactions between genes and the environment.
- Current research has mainly focused on establishing a link between environmental conditions (stress, smoking and BMI) and methylation.

Methylation

- A process by which methyl groups are added to the DNA molecule.
- Change the activity of a DNA segment without changing the sequence.
- Methylation acts to repress gene transcription.

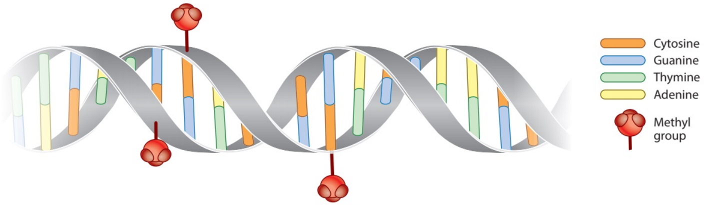


Figure 1

Depiction of a DNA molecule showing the methylation of some, but not all, cytosine bases, as often occurs in the cell.

Figure: Figure from Muligan et al., 2016 (Annual Review of Anthropology) ▸ Link

→ Many examples of epigenetic changes, measured via methylation, across SES.

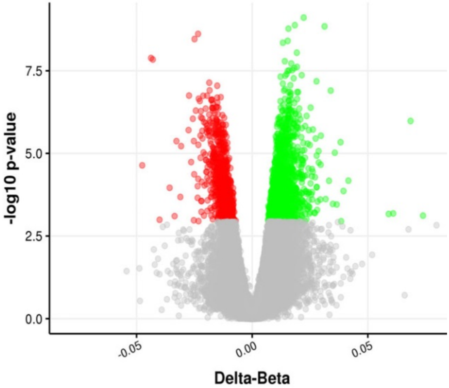


FIGURE 1 Volcano plot comparing low/low to high/high SES score (high/high as reference). Each point represents the difference in methylation between groups, with colored points representing significant down-methylation (red) and up-methylation (green) after accounting for false discovery (FDR $q < 0.05$)

FIGURE: FIGURE FROM MCDADE ET AL., 2019 (AJ OF PHYSICAL ANTHROPOLOGY) ▸ LINK

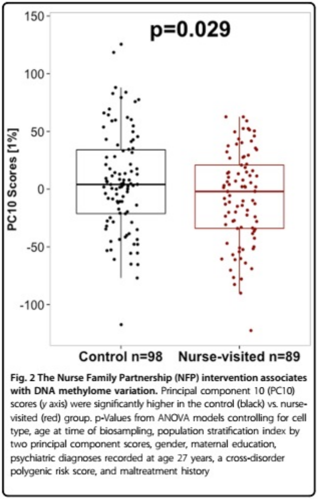


Fig. 2 The Nurse Family Partnership (NFP) intervention associates with DNA methylome variation. Principal component 10 (PC10) scores (*y* axis) were significantly higher in the control (black) vs. nurse-visited (red) group. p-Values from ANOVA models controlling for cell type, age at time of biosampling, population stratification index by two principal component scores, gender, maternal education, psychiatric diagnoses recorded at age 27 years, a cross-disorder polygenic risk score, and maltreatment history

FIGURE: FIGURE FROM O'DONNELL ET AL., 2018 (TRANSLATIONAL PSYCHIATRY) ▸ LINK

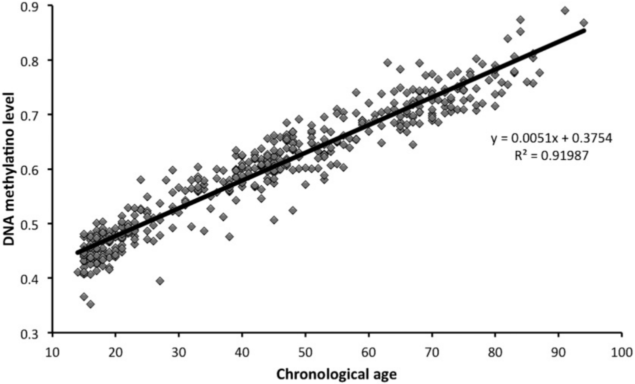


FIGURE: FIGURE FROM JOHHANSON ET AL., 2013 (PLOS ONE) ▸ LINK

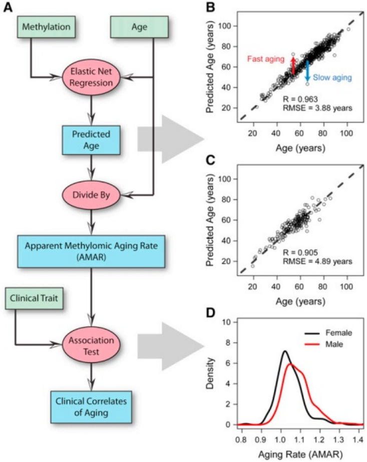


FIGURE: FIGURE FROM HANNUN ET AL., 2013 (MOLECULAR CELL) ▸ LINK

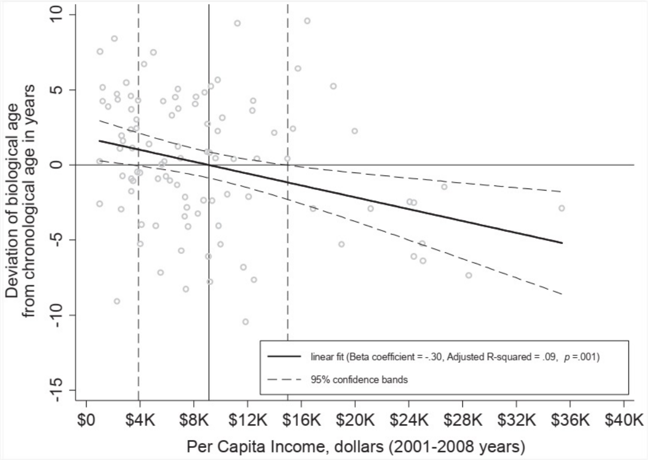


FIGURE: FIGURE FROM SIMONS ET AL., 2016 (SOCIAL SCIENCE & MEDICINE) ▸ LINK

→ Trans-generational epigenetics is very controversial.

→ Evidence on humans is week.

→ See the discussion ▸ Here and ▸ Here.

→ Key reason for skepticism of results is that epigenetic patters are “erased” at the embryonic stage.

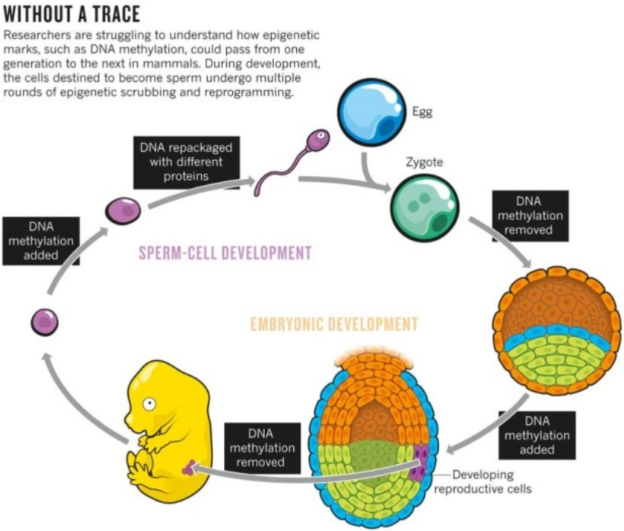

WITHOUT A TRACE
Researchers are struggling to understand how epigenetic marks, such as DNA methylation, could pass from one generation to the next in mammals. During development, the cells destined to become sperm undergo multiple rounds of epigenetic scrubbing and reprogramming.
Egg
Zygote
DNA repackaged with different proteins
DNA methylation removed
DNA methylation added
SPERM-CELL DEVELOPMENT
EMBRYONIC DEVELOPMENT
DNA methylation removed
DNA methylation added
Developing reproductive cells

FIGURE: FIGURE FROM HUGHES ET AL., 2014 (NATURE REVIEWS) ▸ LINK

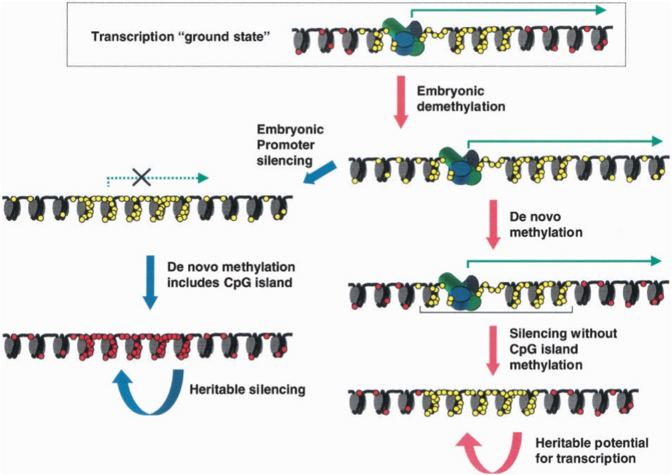

Transcription "ground state"
Embryonic demethylation
Embryonic Promoter silencing
De novo methylation
De novo methylation includes CpG island
Silencing without CpG island methylation
Heritable silencing
Heritable potential for transcription

FIGURE: FIGURE FROM BIRD 2002 (GENES AND DEVELOPMENT) ▸ LINK

APPENDIX

Transcription and Translation:

- → DNA contains exons (protein coding regions) and introns (regions not translated into protein).
- → Most of the DNA is regulatory/structural , only ~3-4% of DNA is translated into proteins.

▸ Back

Transcription and Translation:

- Transcription:
  - DNA is transcribed into an RNA molecule.
  - Most of the bases are associated with at least one primary transcript.

▸ Back

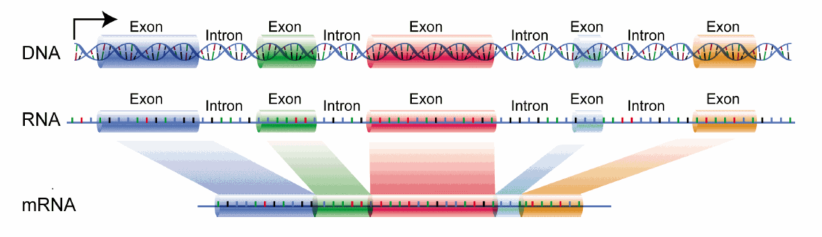


FIGURE: TRANSCRIPTION AND TRANSLATION ▸ SOURCE

▸ Back

Transcription and Translation:

- Translation (gene expression):
  - mRNA (messenger RNA), which is composed of exons, is decoded into amino acids.
  - 3 base pairs (codonds) make an amino acid - 20 possibilities out of 64 possible combinations.
  - Proteins are long chains of bonded amino acids.
  - Majority of Mendelian phenotypes associated with protein coding changes.

▸ Back

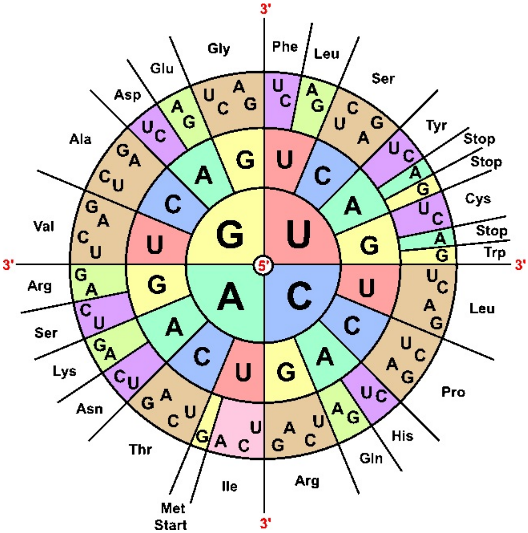


Figure: Codonds and Amino Acids ▸ Source

▸ Back

Chromosomes:

- → DNA is organized into chromosomes
- → Human cells have 23 chromosomes
- → Chromosomes have different sizes
- → Chromosomes condense to carry out cell division

▸ Back

Genes:

- → The DNA sequence is usually separated into genes.
- → A gene is part of the DNA that is a template to make RNA.
- → How many genes?
  - → Protein coding genes ~20500 (transcription and translation).
  - → IncRNA genes ~9500 (transcription only).
  - → Take Away: ~30k genes and only ~21k lead to translation into a protein (aminoacids).
  - → Note: Gene annotation is an ongoing process (still being updated).

▸ Back

Typical genome vs. reference (The 1000 Genomes Project Consortium, 2015 ▸ Link).

- → Each individual genome combines inherited alleles and new variation (de novo mutation).
- → Individual genome differs from reference human genome at 4-5 million sites:
  - → ∼ 3.5-4.5 million SNPs
  - → ∼ 500-600 thousand Indels
  - → ∼ 1 thousand large deletions
  - → Other rare-variants.

▸ Back

Genetic Variation In Humans

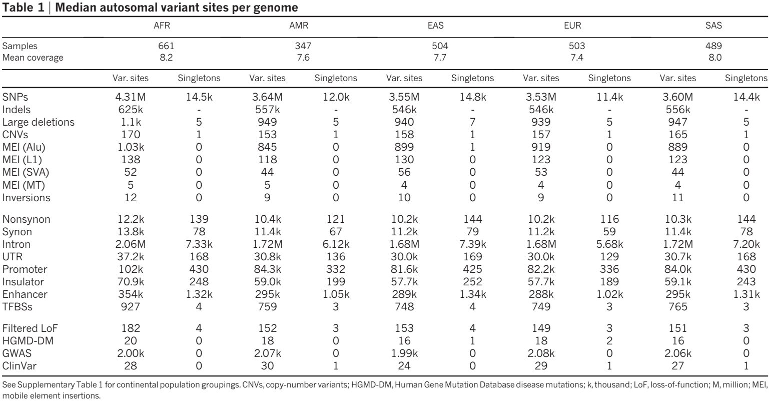


Table 1 | Median autosomal variant sites per genome

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

Figure: Table from The 1000 Genomes Project Consortium, 2015 ▸ Link

▸ Back

Genetic variation:

- 99% of variation are **SNPs** and indels.
- Single nucleotide polymorphism (SNP) is a variation in a single base pair in a DNA sequence.
- **Most of the new empirical papers in economics focus on SNP variation.**
- Indels consist of small insertions of deletions of base-pairs (1-10,000 base pairs).
- However, structural variants affect more base-pairs:
  - ∼2000-2500 structural variants
  - ∼1000 large deletions, ∼160 copy-number variants, ∼915 Alu insertions, ∼128 insertions, ∼51 SVA insertions , ∼4 NUMTs and ∼10 inversion)
  - All affecting ∼ 20 million bases of sequence.

▸ Back

Example of structural variants:

- Short tandem repeat (STR) expansions.
- 2 to 6 nucleotides repeated hundreds of times.
- Small number of repeats (dozen) create no problem.
- However, larger repeats are prone to “replication slippage”, which results in high variability in the number of repeat elements.
- Mutation rate of STRs exceeds that of any other type of genetic variation (Ballantyne et al., 2010 ▸ Link).
- Expansion repeats are considered causal of over 20 neurological disorders.

▸ Back

Cell Division:

- Main function of cell division is growth and repair.
- Mitosis:
  - Asexual cell division.
  - Results in an equal number of chromosomes in a cell.
- Meiosis:
  - Sexual cell reproduction.
  - Results in an half the number of chromosomes.
  - Novel combination of genes → changes across generations.

▸ Back

Crossing over during meiosis:

- A segment of DNA from on chromosome switches places with a segment in the homologous chromosome
- Quite common in humans. Usual one or two crossovers along each pair of homologous chromosomes.
- Multiple segments of of homologous chromosomes ‘swap positions’.
- Since exchange happen between homologous chromosomes the same genes are present but with a different combination of alleles.

▸ Back

Crossing over during meiosis:

- → Genes that are further apart are more likely to become separated during crossover (linkage equilibrium).
- → Alleles that are close to each other are said to be linked (linkage disequilibrium).
- → Also, group of alleles that are close-linked and tend to be inherited together are called ‘Haplotypes’.

▸ Back

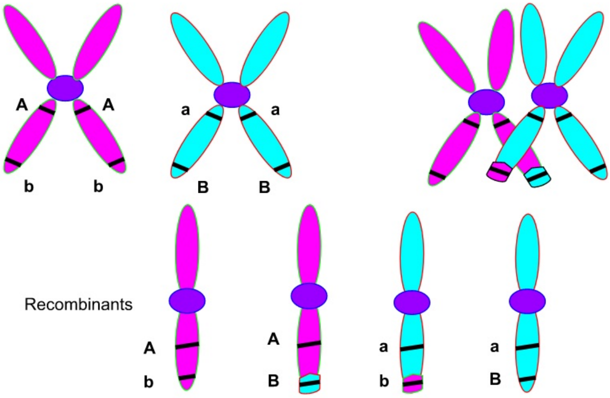


FIGURE: CROSSING OVER - MEIOSIS ▸ SOURCE

▸ Back

- → Genetics Background.
- → Statistical Model of Population Genetics.
- → **Heritability - Kinship Studies.**
- → Candidate Genes.
- → Gene Discovery.
- → Polygenic Scores.
- → Epigenetics.

Broad Heritability:

$$h_G^2 = \frac{Var\left(m_{G_j}\left(g_{is}\right)\right)}{Var(\omega_i)} \tag{14}$$

- → Broad heritability is the fraction of the variance in the phenotype explained by the genetic factor.
- → Assumption: $G(\mathbf{g_i})$ independent of $\mathbf{E_i}$.

Narrow Heritability:

$$h_A^2 = \frac{Var(\mathbf{g_i}\beta)}{Var(\omega_i)} \tag{15}$$

- → Narrow heritability is the fraction of the variance in the phenotype explained by the additive component.
- → By definition: $h_A^2 \leqslant h_G^2$.
- → When we talk about heritability estimates we are often talking about narrow heritability.
- → h_A^2 is the R^2 from the population regression of ω_i on $\mathbf{g_i}$ controlling for $\mathbf{E_i}$.

Broad vs Narrow Heritability:

- For most phenotypes h_A^2 is very close to h_G^2 (E.g. educational attainment and height).
- This is due to small minor allele frequencies on most causal variants.
- See Hill, Goddard and Visscher, 2008 ▸ Link.

Estimating Heritability:

- → Adoption Studies.
- → Twin studies.
- → Kinship/Family studies .
- → Molecular-genetic data (GCTA, LD Score regression).

Intuition Behind Behavioral Genetic Models:

- Correlation in traits between relatives due to:
  - Shared family environment.
  - Shared genetic pool (genetic correlation).
- The estimation approach is to:
  - Estimate observed correlation in phenotype between different relatives.
  - Impose a theoretical genetic correlation between relatives.
  - Decompose trait variation due to environment and genetics.

Formally, start by reorganizing and standardizing the causal model:

$$\omega_i = G(\mathbf{g_i}) + \mathbf{E_i}\gamma + \epsilon_i \tag{16}$$

$$\omega_i = G_i + U_i \tag{17}$$

$$\tag{18}$$

where

- $\omega_i = \frac{\omega_i - E[\omega_i]}{std(\omega_i)}$
- $G_i = \frac{G(\mathbf{g_i})}{std(G(\mathbf{g_i}))}$
- $U_i = \frac{\mathbf{E_i}\gamma + \epsilon_i}{std(\mathbf{E_i}\gamma + \epsilon_i)}$
- G_i captures genetic effects
- U_i captures environmental effects

Further decompose environment into a common (shared) environment between relatives and an unshared component, so that:

$$\omega_i = G_i + C_i + E_i \tag{19}$$

where

- E_i is assumed to be independent of C_i and G_i.
- C_i is assumed to be independent of G_i
- Question: When are these assumptions violated and when they are not?

This assumptions imply that:

$$
\begin{aligned}
var(\omega_i) =& var(G_i) + var(C_i) + var(E_i) \qquad (20)\\
1 =& h_g^2 + c^2 + e^2 \qquad (21)
\end{aligned}
$$

and the observed correlation in phenotypes between relatives can be described as:

$$Cov(\omega_i, \omega_i') = Cov(G_i, G_i') + Cov(C_i, C_i') \qquad (22)$$

→ h_G^2 is identified by assuming $Cov(G_i, G_i')$ for a different set of relatives.

Additive (ACE) model:

$$var(\omega_i) = var(A_i) + var(C_i) + var(E_i) \tag{23}$$

$$1 = h_a^2 + c^2 + e^2 \tag{24}$$

$\rightharpoondown$ Ignores non-additive effects in order to simplify calculations.

$\rightharpoondown$ h_a^2 is identified by assuming $Cov(A_i, A_i')$ for a different set of relatives.

The “ACE Model”:

$$\omega_i = A_i + C_i + E_i \tag{25}$$

Key assumptions:

- → All genetic variance is additive.
- → No gene-environment correlation:
  - → $Cov(A_i, C_i) = 0$ and $Cov(A_i, E_i) = 0$
  - → and $Cov(A_i, C_i') = 0$ and $Cov(A_i, E_i') = 0$
- → Equal environments assumption:
  - → Adoption studies: $Cov(C_i, C_i')_{na,na} = Cov(C_i, C_i')_{ad,na} = c^2$
  - → Twin studies: $Cov(C_i, C_i')_{mz} = Cov(C_i, C_i')_{dz} = c^2$

Adoption studies:

- Correlation between adopted and non-adopted siblings:
    - $Cov(G_i, G_i')_{ad,na} = 0$
    - so, $Cov(\omega_i, \omega_i')_{ad,na} = Cov(C_i, C_i') = c^2$
- While, correlation between non-adopted siblings:
    - $Cov(\omega_i, \omega_i')_{na,na} = Cov(G_i, G_i')_{na,na} + Cov(C_i, C_i')$

- On average, full-siblings (non-twins) share 50% of their genes.
- This implies, that under some assumptions (e.g. random mating):
    - Siblings share 50% of genetic additive components,
    - 25% of dominance components,
    - and 12.5'% of epistatic components.

As a result:

- $Cov(G_i, G_i')_{na,na} = 0.5h_A^2 + 0.25h_D^2 + 0.125h_{AA}^2$
- $0.5h_A^2 \leqslant Cov(G_i, G_i')_{na,na} \leqslant 0.5h_G^2$

Adoption studies:

- Obtain data on additional pairs of relatives to disentangle additive from non-additive components.
- Most common: assume genetic factor is purely additive (ACE Model):
  - $Cov(\omega_i, \omega_i')_{na,na} = 0.5h_A^2 + c^2$
  - $Cov(\omega_i, \omega_i')_{ad,na} = c^2$
  - So, $h_a^2 = 2\,[Cov(\omega_i, \omega_i')_{na,na} - Cov(\omega_i, \omega_i')_{ad,na}]$

Sacerdote (2007) ▸ Link

- → Follow a sample of Korean-born American adoptees.
- → Information on adoptees, on adoptive parents, and on parents' biological children.
- → Adoptee-parent assignment is plausibly random.
- → Large samples: 1650 adoptees and 1196 biological children.

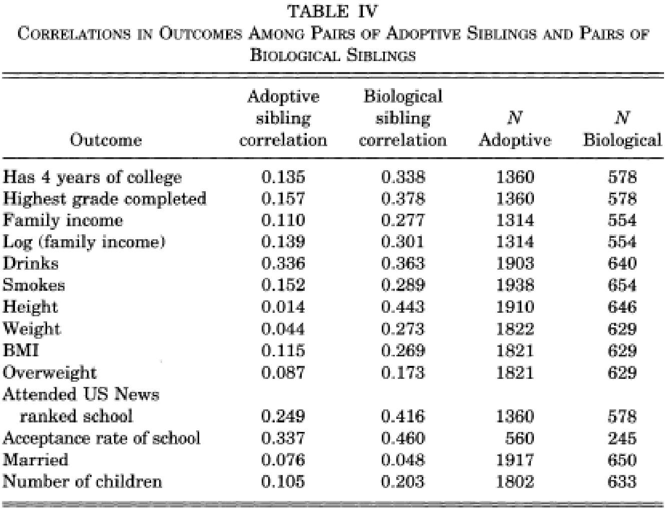


TABLE IV
CORRELATIONS IN OUTCOMES AMONG PAIRS OF ADOPTIVE SIBLINGS AND PAIRS OF BIOLOGICAL SIBLINGS

Outcome	Adoptive sibling correlation	Biological sibling correlation	N Adoptive	N Biological
Has 4 years of college	0.135	0.338	1360	578
Highest grade completed	0.157	0.378	1360	578
Family income	0.110	0.277	1314	554
Log (family income)	0.139	0.301	1314	554
Drinks	0.336	0.363	1903	640
Smokes	0.152	0.289	1938	654
Height	0.014	0.443	1910	646
Weight	0.044	0.273	1822	629
BMI	0.115	0.269	1821	629
Overweight	0.087	0.173	1821	629
Attended US News ranked school	0.249	0.416	1360	578
Acceptance rate of school	0.337	0.460	560	245
Married	0.076	0.048	1917	650
Number of children	0.105	0.203	1802	633

FIGURE: TABLE FROM SACERDOTE (2007) ▸ LINK

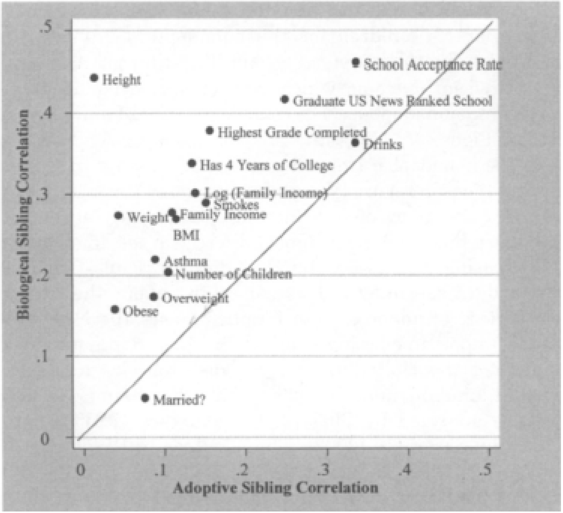


FIGURE: FIGURE FROM SACERDOTE (2007) ▸ LINK

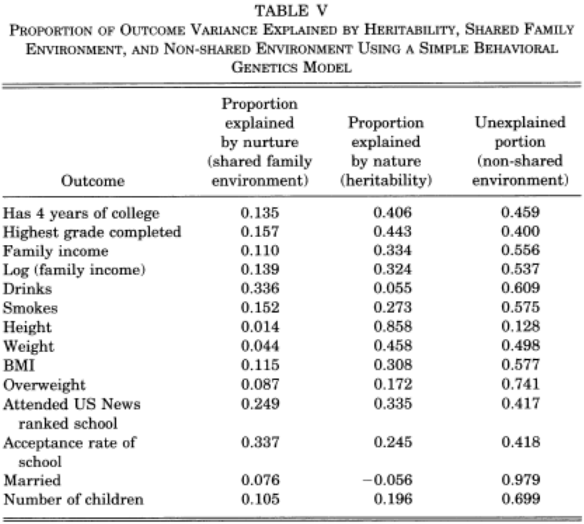


TABLE V

PROPORTION OF OUTCOME VARIANCE EXPLAINED BY HERITABILITY, SHARED FAMILY ENVIRONMENT, AND NON-SHARED ENVIRONMENT USING A SIMPLE BEHAVIORAL GENETICS MODEL

Outcome	Proportion explained by nurture (shared family environment)	Proportion explained by nature (heritability)	Unexplained portion (non-shared environment)
Has 4 years of college	0.135	0.406	0.459
Highest grade completed	0.157	0.443	0.400
Family income	0.110	0.334	0.556
Log (family income)	0.139	0.324	0.537
Drinks	0.336	0.055	0.609
Smokes	0.152	0.273	0.575
Height	0.014	0.858	0.128
Weight	0.044	0.458	0.498
BMI	0.115	0.308	0.577
Overweight	0.087	0.172	0.741
Attended US News ranked school	0.249	0.335	0.417
Acceptance rate of school	0.337	0.245	0.418
Married	0.076	−0.056	0.979
Number of children	0.105	0.196	0.699

FIGURE: TABLE FROM SACERDOTE (2007) ▸ LINK

Criticism of adoption studies:

- → Adoptive families are non-representative.
- → Non-random assignment of adoptees to families.
- → Common environmental effects different for adopted and non-adopted children.
  - → i.e. $Cov(C_i, C_i')_{na,na} \neq Cov(C_i, C_i')_{ad,na}$
- → Assumes away gene-environment interactions.

Twin Studies:

- Compare observed correlation in phenotype from monozygotic ('identical') and dizygotic ('fraternal') twins.
- Idea (assuming all genetic variance is additive):
  - Monozygotic: $Cov(\omega_i, \omega_i')_{mz} = 1h_A^2 + c^2$
  - Dizygotic: $Cov(\omega_i, \omega_i')_{dz} = 0.5h_A^2 + c^2$

  And the statistics of interest are:
  - heritability: $\hat{h_A^2} = 2\,[Cov(\omega_i, \omega_i')_{mz} - Cov(\omega_i, \omega_i')_{dz}]$
  - $\hat{c^2} = 2Cov(\omega_i, \omega_i')_{dz} - Cov(\omega_i, \omega_i')_{mz}$
  - $\hat{e^2} = 1 - Cov(\omega_i, \omega_i')_{mz}$

Polderman et al., (2015) ▸ Link

- “Meta-analysis of the heritability of human traits based on fifty years of twin studies.”
- Authors report a meta-analysis of twin correlation for
  - 17,804 traits
  - 2,748 publications
  - 14,558,903 partly dependent twin pairs
- Findings:
  - Average heritability of 49% across traits.
  - Common family environment variance less than 20%.
  - Average heritability for cognitive traits around 50%.

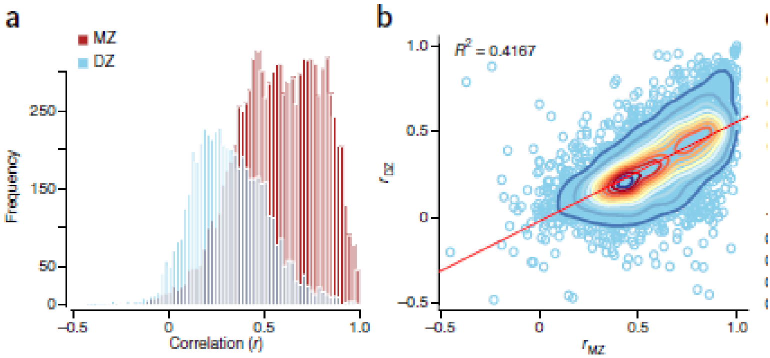


Figure: Figure from Polderman et al., (2015) ▸ Link

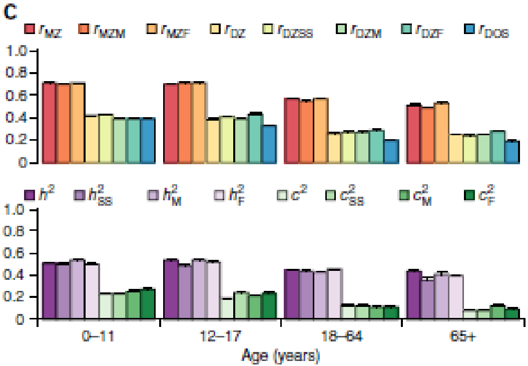


Figure: Figure from Polderman et al., (2015) ▸ Link

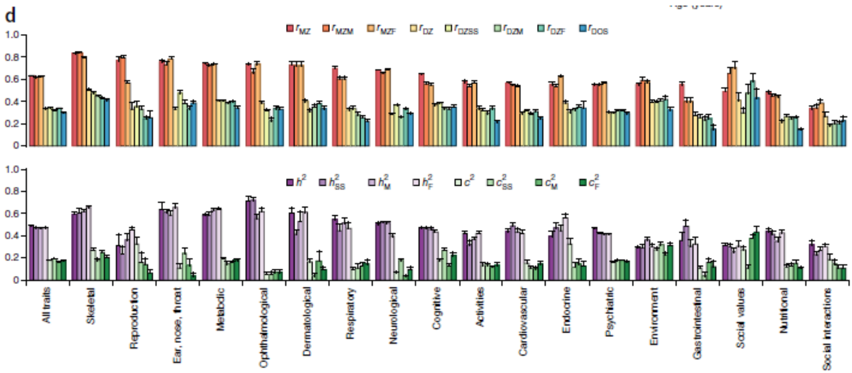


Figure: Figure from Polderman et al., (2015) ▸ Link

Three “Laws” of Behavioral Genetics - Turkheimer 2000

- → All human behavioral traits are heritable.
- → The effect of being raised in the same family is smaller than the effect of the genes.
- → A substantial portion of the variation in complex human behavioural traits is not accounted for by the effects of genes or families.

Criticism of twin studies - Goldberger Critique.

- Common environmental correlation different for twin and non-twin siblings.
  - i.e. $Cov(C_i, C_i')_{mz} \neq Cov(C_i, C_i')_{dz}$
  - OK: MZ twins self-select into the same environment (both like playing soccer).
  - Not OK: Parents enrol twin 2 in soccer practice because twin 1 wanted to play soccer.
- Twin families might not be representative of overall population.
- Assumes away gene-environment interactions.

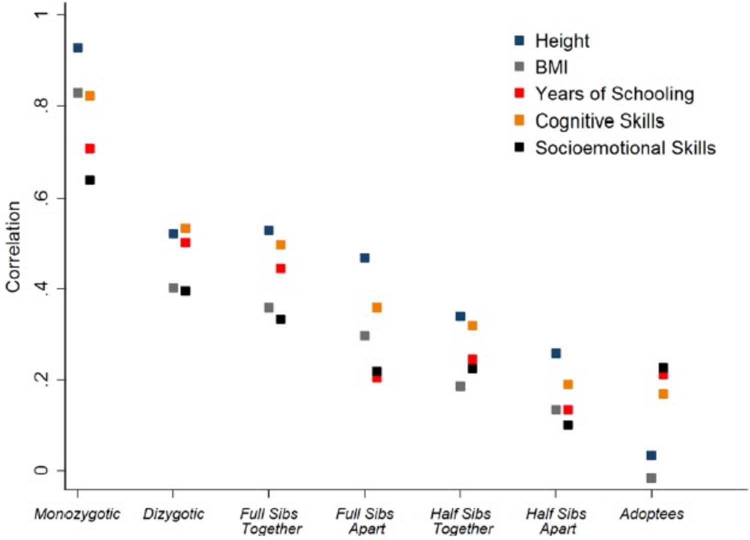


FIGURE: FIGURE FROM CESARINI AND VISSCHER (2017) ▸ LINK

Heritability estimates are stable across kinship models:

- → Educational attainment is estimated to be ∼40% heritable.
- → Cognitive skills are estimated to be ∼50% heritable.
- → Socioemotional skills are estimated to be ∼50% heritable.

Criticism of assumptions:

⇁ Taubman (1976)

⇁ Behrman and Taubman (1989)

⇁ Bjorklund, Jantti, and Solon (2005)

Bjorklund, Jantti, and Solon (2005)

- Explore a variety of sibling types to relax assumptions in traditional ACE model.
- The objective is to compare heritability estimates for earnings under different model assumptions.
- They estimate four models:
  - Standard model
  - Relax assumptions that A and C are uncorrelated.
  - Allow for assortative matting.
  - Allow for differences in shared environment across sibling pairs.

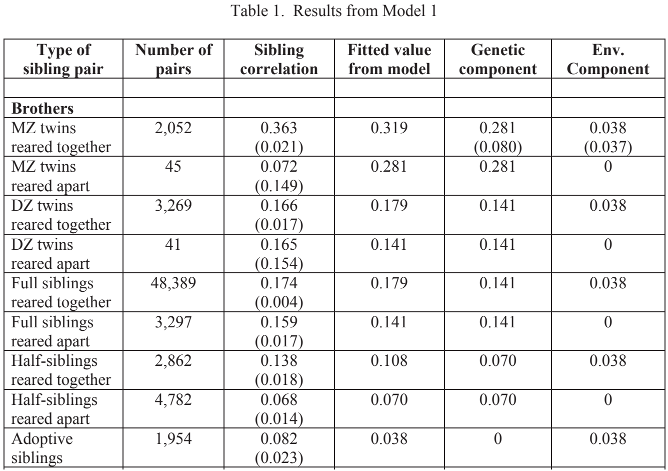


Table 1. Results from Model 1

Type of sibling pair	Number of pairs	Sibling correlation	Fitted value from model	Genetic component	Env. Component
Brothers					
MZ twins reared together	2,052	0.363 (0.021)	0.319	0.281 (0.080)	0.038 (0.037)
MZ twins reared apart	45	0.072 (0.149)	0.281	0.281	0
DZ twins reared together	3,269	0.166 (0.017)	0.179	0.141	0.038
DZ twins reared apart	41	0.165 (0.154)	0.141	0.141	0
Full siblings reared together	48,389	0.174 (0.004)	0.179	0.141	0.038
Full siblings reared apart	3,297	0.159 (0.017)	0.141	0.141	0
Half-siblings reared together	2,862	0.138 (0.018)	0.108	0.070	0.038
Half-siblings reared apart	4,782	0.068 (0.014)	0.070	0.070	0
Adoptive siblings	1,954	0.082 (0.023)	0.038	0	0.038

Figure: Table from Bjorklund, Jantti, and Solon (2005)

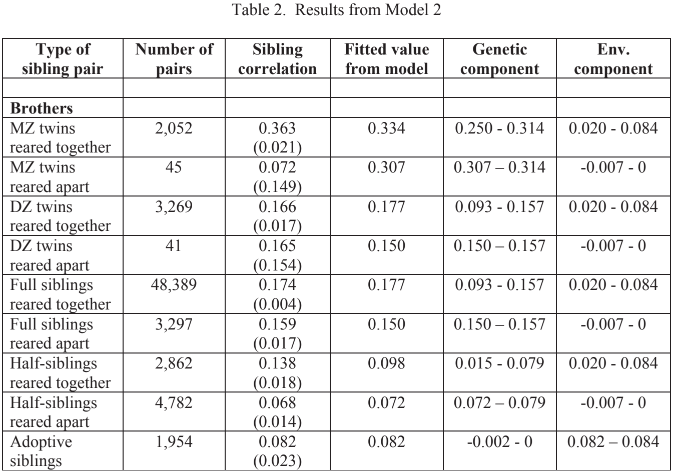


Table 2. Results from Model 2

Type of sibling pair	Number of pairs	Sibling correlation	Fitted value from model	Genetic component	Env. component
Brothers					
MZ twins reared together	2,052	0.363 (0.021)	0.334	0.250 - 0.314	0.020 - 0.084
MZ twins reared apart	45	0.072 (0.149)	0.307	0.307 – 0.314	-0.007 - 0
DZ twins reared together	3,269	0.166 (0.017)	0.177	0.093 - 0.157	0.020 - 0.084
DZ twins reared apart	41	0.165 (0.154)	0.150	0.150 – 0.157	-0.007 - 0
Full siblings reared together	48,389	0.174 (0.004)	0.177	0.093 - 0.157	0.020 - 0.084
Full siblings reared apart	3,297	0.159 (0.017)	0.150	0.150 – 0.157	-0.007 - 0
Half-siblings reared together	2,862	0.138 (0.018)	0.098	0.015 - 0.079	0.020 - 0.084
Half-siblings reared apart	4,782	0.068 (0.014)	0.072	0.072 – 0.079	-0.007 - 0
Adoptive siblings	1,954	0.082 (0.023)	0.082	-0.002 - 0	0.082 – 0.084

FIGURE: TABLE FROM BJORKLUND, JANTTI, AND SOLON (2005)

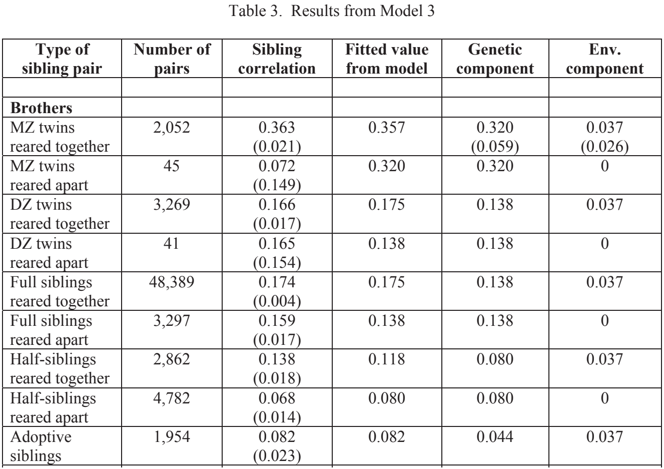


Table 3. Results from Model 3

Type of sibling pair	Number of pairs	Sibling correlation	Fitted value from model	Genetic component	Env. component
Brothers					
MZ twins reared together	2,052	0.363 (0.021)	0.357	0.320 (0.059)	0.037 (0.026)
MZ twins reared apart	45	0.072 (0.149)	0.320	0.320	0
DZ twins reared together	3,269	0.166 (0.017)	0.175	0.138	0.037
DZ twins reared apart	41	0.165 (0.154)	0.138	0.138	0
Full siblings reared together	48,389	0.174 (0.004)	0.175	0.138	0.037
Full siblings reared apart	3,297	0.159 (0.017)	0.138	0.138	0
Half-siblings reared together	2,862	0.138 (0.018)	0.118	0.080	0.037
Half-siblings reared apart	4,782	0.068 (0.014)	0.080	0.080	0
Adoptive siblings	1,954	0.082 (0.023)	0.082	0.044	0.037

FIGURE: TABLE FROM BJORKLUND, JANTTI, AND SOLON (2005)

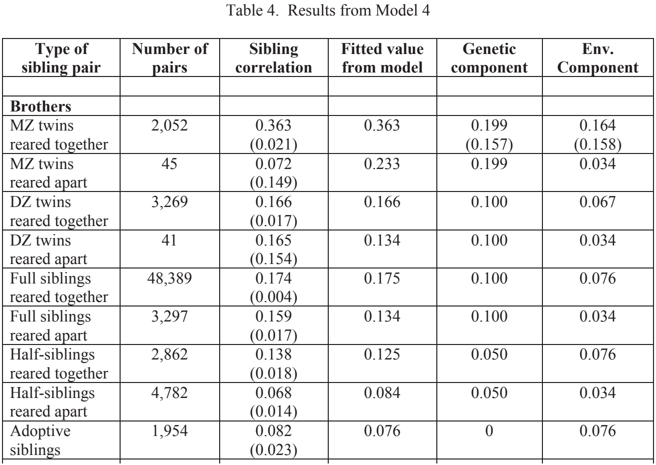


Table 4. Results from Model 4

Type of sibling pair	Number of pairs	Sibling correlation	Fitted value from model	Genetic component	Env. Component
Brothers					
MZ twins reared together	2,052	0.363 (0.021)	0.363	0.199 (0.157)	0.164 (0.158)
MZ twins reared apart	45	0.072 (0.149)	0.233	0.199	0.034
DZ twins reared together	3,269	0.166 (0.017)	0.166	0.100	0.067
DZ twins reared apart	41	0.165 (0.154)	0.134	0.100	0.034
Full siblings reared together	48,389	0.174 (0.004)	0.175	0.100	0.076
Full siblings reared apart	3,297	0.159 (0.017)	0.134	0.100	0.034
Half-siblings reared together	2,862	0.138 (0.018)	0.125	0.050	0.076
Half-siblings reared apart	4,782	0.068 (0.014)	0.084	0.050	0.034
Adoptive siblings	1,954	0.082 (0.023)	0.076	0	0.076

FIGURE: TABLE FROM BJORKLUND, JANTTI, AND SOLON (2005)

Bjorklund, Jantti, and Solon (2005) - Takeaways:

- → Estimated results on heritability are very sensitive to assumptions on environmental similarity.
- → The traditional ACE model seems to exaggerate the importance of nature.
- → Nonetheless, all models point to a significant role of genetic variance ($>$ 10%).
- → Non-shared environment seems to explain a large percentage of the variation in earnings ($\sim$ 65%).

→ Goldberger (1979)

→ Manski (JEP, 2011): Genes, Eyeglasses, and Social Policy

Consider Goldberger's use of distribution of eyeglasses as the intervention. For simplicity, suppose that nearsightedness derives entirely from the presence of a particular allele of a specific gene. Suppose that this gene is observable, taking the value $g = 0$ if a person has the allele for nearsightedness and $g = 1$ if he has the one that yields normal sight.

Let the outcome of interest be effective quality of sight, where "effective" means sight when augmented by eyeglasses, should they be available. A person has effective normal sight either if he has the allele for normal sight or if eyeglasses are available. A person is effectively nearsighted if that person has the allele for nearsightedness and eyeglasses are unavailable.

Now suppose that the entire population lacks eyeglasses. Then the heritability of effective quality of sight is one. What does this imply about the usefulness of distributing eyeglasses as a treatment for nearsightedness? Nothing, of course. The policy question of interest concerns effective quality of sight in a conjectured environment where eyeglasses are available. However, the available data only reveal what happens when eyeglasses are unavailable.

Molecular Genetics based Heritability

- → Does not rely on family study assumptions.
- → Rely on unrelated individuals instead of relatives.
- → Heritability estimates ‘potentially’ uncontaminated by shared environment.
- → Can help us understand genetic mechanisms of heritability.

Idea:

- → Unrelated individuals vary in their genetically similarity and phenotypic similarity.
- → If a trait is genetically influenced, then individuals who are more genetically similar should be more phenotypically similar.

Key Assumption:

- → It excludes individuals that have high genetic similarity (siblings, cousins and second cousins) and focus on genetically unrelated individuals.
- → The underlying assumption is that, for this group, genetically similarity is unrelated to environmental similarity.

- Two approaches:
  - GCTA or genomic-relatedness based restricted maximum-likelihood (GREML).
  - LD Score regression (LDSC).
- The two methods are very similar and rely on similar assumptions.
- This talk will only focus on GCTA-GREML.

GCTA-GREML Model:

$$\tilde{\omega}_i = \sum_{s=1}^{S} \beta_s g_{is} + \varepsilon_i \tag{26}$$

where

- $g_{is} = \frac{g_{is} - 2p_j}{\sqrt{2p_j(1-p_j)}}$ is the standardized SNP s for individual i
- where p_j is the minor allele frequency of SNP s,
- and $\tilde{\omega}_i = std(\omega_i - \mathbf{E_i}\gamma)$ is the standardized residual phenotype.
- Note: I use the residual phenotype here for exposition only, in practice γ is estimated jointly.
- Note 2: Also, in practice $\mathbf{E_i}$ usually includes the first few principal components of the genetic matrix.

GCTA-GREML Approach (Yang et al., 2015) ▸ Link :

- $\rightharpoondown$ Instead of estimating the values for each β_s (not possible since $S > n$)
- $\rightharpoondown$ Assume $\beta \sim N(0, \sigma^2_\beta)$
- $\rightharpoondown$ Estimate σ^2_β instead.
- $\rightharpoondown$ Underlying assumption: rare (small minor allele frequency) SNPs have larger effects than common SNPS - since g_{is} has been standardized.
- $\rightharpoondown$ The heritability is given by:

$$h^2_{SNP} = \frac{Var(\sum_{s=1}^{S} \beta_s g_{is})}{Var(\tilde{\omega}_i)} = \sum_{s=1}^{S} \sigma^2_\beta = J\sigma^2_\beta \qquad (27)$$

GCTA-GREML:

$$V(\mathbf{Y}) = \mathbf{X}\mathbf{X}'\sigma_{\beta}^{2} + \mathbf{I}\sigma_{\epsilon}^{2} \tag{28}$$

where:

- $\rightharpoondown$ $V(\mathbf{Y})$ is the n-by-n phenotypic variance-covariance matrix.
- $\rightharpoondown$ $\mathbf{X}\mathbf{X}'$ is the n-by-n genetic relatedness matrix.
- $\rightharpoondown$ $\mathbf{I}$ is the identity matrix.
- $\rightharpoondown$ σ_{ϵ}^{2} is the environmental variance.

Genetic Relatedness: Each element of $\mathbf{XX}'$ is given by:

$$\hat{\pi}_{ik} = \frac{1}{J} \sum_j \frac{(g_{is} - 2p_j)(G_{kj} - 2p_j)}{2p_j(1 - p_j)} \tag{29}$$

where:

- g_{is} is the reference allele for SNP s of individual i.
- p_j is the frequency of the reference allele.
- Note: $\hat{\pi}_{ii} \neq 1$. Distance to 1 is a measure of inbreeding (cousin matting: 1.05 +).

In twin studies:

- → $\hat{\pi}_{ik} = 1$ for monozygotic twin pairs.
- → $\hat{\pi}_{ik} = 0.5$ for dizygotic twin pairs.

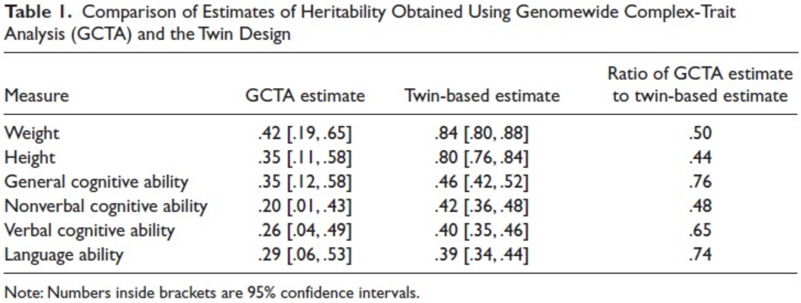


Table 1. Comparison of Estimates of Heritability Obtained Using Genomewide Complex-Trait Analysis (GCTA) and the Twin Design

Measure	GCTA estimate	Twin-based estimate	Ratio of GCTA estimate to twin-based estimate
Weight	.42 [.19, .65]	.84 [.80, .88]	.50
Height	.35 [.11, .58]	.80 [.76, .84]	.44
General cognitive ability	.35 [.12, .58]	.46 [.42, .52]	.76
Nonverbal cognitive ability	.20 [.01, .43]	.42 [.36, .48]	.48
Verbal cognitive ability	.26 [.04, .49]	.40 [.35, .46]	.65
Language ability	.29 [.06, .53]	.39 [.34, .44]	.74

Note: Numbers inside brackets are 95% confidence intervals.

FIGURE: FIGURE FROM PLOMIN ET AL., (2013) ▸ LINK

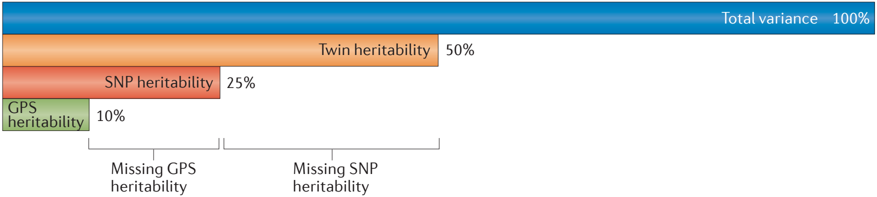


Figure: Heritability of intelligence.

Figure: Figure from Plomin and Stumm (2018) ▸ Link

Interpreting h^2_{SNP} estimates

- In theory $h^2_{SNP} \leqslant h^2_A$
    - h^2_{SNP} captures the proportion of variation in a phenotype due to genetic influences captured by the SNPs included in the model or the SNPs in LD with the ones included in the model.
    - It does not capture the genetic effects from rare variants not capture in SNP arrays.
- As imputation methods improve and/or sequencing data becomes available, h^2_{SNP} should approach true h^2_A.
- If close relatives are included (e.g. twins or full siblings), then h^2_{SNP} equal h^2_A estimated from family-based method.
    - Intuition: $\hat{\pi}_{ik}$ from siblings around 0.5 - outliers.
- In practice, unrelated individuals are considered those with $\hat{\pi}_{ik} < 0.05$

Good:

- ‘Unrelateds’: non-additive genetic effects are very small (not -confounders).
- Different assumptions than in family models.

Bad:

- h^2_{SNP} depends on the correlation (LD) between *causal* SNPs with *tagged* SNPs.
- Different LD assumptions lead to under-overestimation or over-estimation of true h^2_A.

Other relevant points:

- → h^2_{SNP} is the upper bound of r^2 GWAS can detect.
- → Improvements in imputation and sample size: $h^2_{SNP} \to h^2_A$
- → Different bias from family studies → triangulation of true h^2_A.
- → Flexible approach means it can be modified in many different ways:
  - → Estimate heritability for different SNPs groups (say heritability due to each chromosome).
  - → Allow for gene-environment correlation.
  - → Estimation of genetic effects in structural models.

→ GREML extension allows to estimate the genetic variation explained by different set of SNPs.

→ e.g. variation explained by different chromosomes or SNPs associated with different organ tissues.

→ Idea: Jointly estimate different σ^2_βs for different set of SNPs.

Partitioning of Genetic Variation (Yang et al., (2011) ▸ Link)

$$
\begin{aligned}
\tilde{\omega}_i &= \sum_{s=1}^{S} \beta_s g_{is} + \varepsilon_i && (30) \\
&= \sum_{c=1}^{C} \sum_{j_c=1}^{J_c} \beta_{j_c} G_{ij_c} + \varepsilon_i && (31)
\end{aligned}
$$

where

- $c \in \{1, ..., C\}$ are genetic groups
- and $\beta_{j_c} \sim N(0, \sigma^2_{\beta,c})$.
- The model jointly estimates the heritability for each genomic partitioning group ($\sigma^2_{\beta,c}$)
- Example: each group corresponds to SNPs in each chromosome.

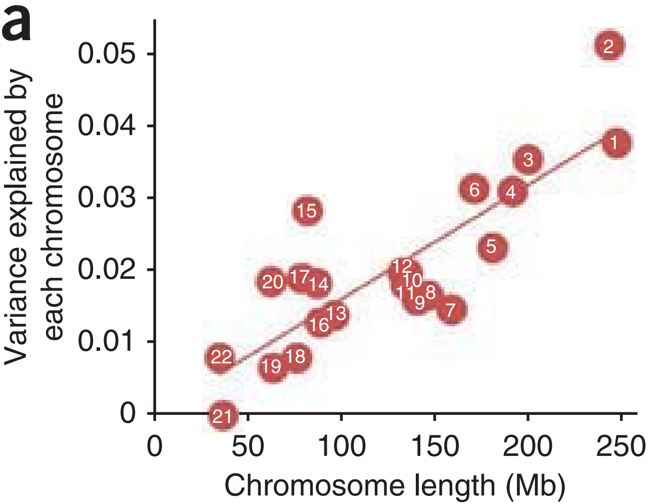


Figure: Figure from Yang et al., (2011) ▸ Link

- → GREML extension allows to estimate the genetic correlation across different phenotypes.
- → That is, whether SNPs associated with phenotype # 1 are similar to SNPs associated with phenotype# 2.
- → Idea: Jointly estimate σ^2_βs for different traits in addition to a correlation parameter.

Bivariate GREML Model - Lee et al., 2012 ▸ Link:

$$\tilde{\mathbf{Y_1}} = \mathbf{G_1}\beta_1 + \varepsilon_1 \tag{32}$$

$$\tilde{\mathbf{Y_2}} = \mathbf{G_2}\beta_2 + \varepsilon_2 \tag{33}$$

where

- $\rightharpoondown$ $\tilde{\mathbf{Y_t}}$ is the 'residual' vector of observations for trait t.
- $\rightharpoondown$ β_t is the vector of random genetic effects for trait t.

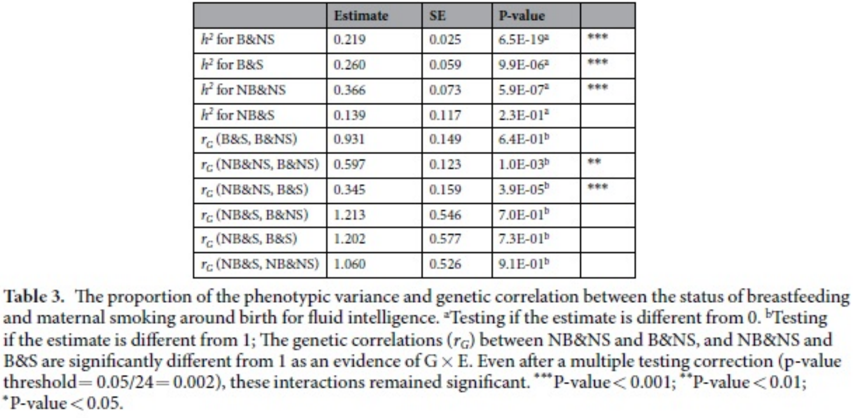
Then,

$$V([\mathbf{Y_1}, \mathbf{Y_2}]) = \begin{bmatrix} G_1AG_1'\sigma^2_{\beta_1} + I\sigma^2_{\epsilon_1} & G_1AG_2'\sigma^2_{\beta_1\beta_2} \\ G_1AG_2'\sigma^2_{\beta_1\beta_2} & G_2AG_2'\sigma^2_{\beta_2} + I\sigma^2_{\epsilon_2} \end{bmatrix} \tag{34}$$

where

- As before, A is the genomic similarity relationship matrix based on SNP information,
- $\beta_t \sim N(0, \sigma^2_{\beta_t})$
- and $\epsilon_t \sim N(0, \sigma^2_{\epsilon_t})$, is the environmental vector for trait t.
- The key parameter is: $\sigma^2_{\beta_1\beta_2}$ is the genetic correlation between the two phenotypes.

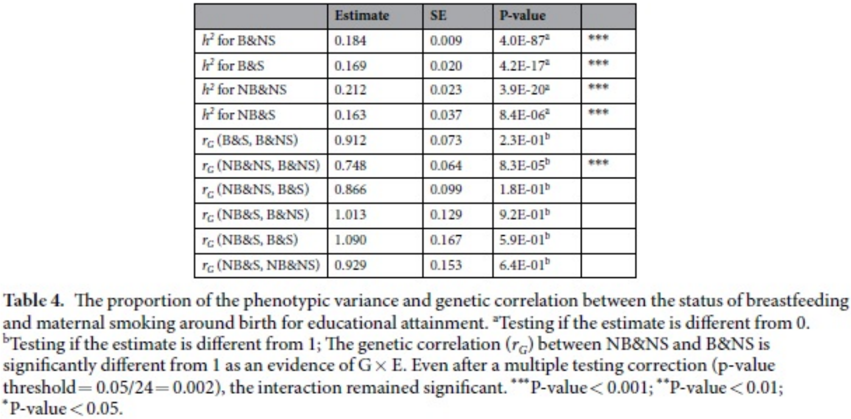
- The GREML model can also be extended to estimate gene-environment interactions.
- The idea is to allow σ^2_β to be different at different environments.
- Moreover, the bivariate GREML can be used to estimate genetic correlation for the same trait at different environments.
- e.g. Lee et al., (2017) ▸ Link estimates heritability for educational attainment and genetic correlation among different groups sorted by whether they experienced maternal smoking during pregnancy and breastfeeding.



	Estimate	SE	P-value	
h^2 for B&NS	0.219	0.025	6.5E-19[a]	***
h^2 for B&S	0.260	0.059	9.9E-06[a]	***
h^2 for NB&NS	0.366	0.073	5.9E-07[a]	***
h^2 for NB&S	0.139	0.117	2.3E-01[a]	
r_G (B&S, B&NS)	0.931	0.149	6.4E-01[b]	
r_G (NB&NS, B&NS)	0.597	0.123	1.0E-03[b]	**
r_G (NB&NS, B&S)	0.345	0.159	3.9E-05[b]	***
r_G (NB&S, B&NS)	1.213	0.546	7.0E-01[b]	
r_G (NB&S, B&S)	1.202	0.577	7.3E-01[b]	
r_G (NB&S, NB&NS)	1.060	0.526	9.1E-01[b]	

Table 3. The proportion of the phenotypic variance and genetic correlation between the status of breastfeeding and maternal smoking around birth for fluid intelligence. [a]Testing if the estimate is different from 0. [b]Testing if the estimate is different from 1; The genetic correlations (r_G) between NB&NS and B&NS, and NB&NS and B&S are significantly different from 1 as an evidence of G × E. Even after a multiple testing correction (p-value threshold = 0.05/24 = 0.002), these interactions remained significant. ***P-value < 0.001; **P-value < 0.01; *P-value < 0.05.

FIGURE: TABLE FROM LEE ET AL., (2017) ▸ LINK



	Estimate	SE	P-value	
h^2 for B&NS	0.184	0.009	4.0E-87[a]	***
h^2 for B&S	0.169	0.020	4.2E-17[a]	***
h^2 for NB&NS	0.212	0.023	3.9E-20[a]	***
h^2 for NB&S	0.163	0.037	8.4E-06[a]	***
r_G (B&S, B&NS)	0.912	0.073	2.3E-01[b]	
r_G (NB&NS, B&NS)	0.748	0.064	8.3E-05[b]	***
r_G (NB&NS, B&S)	0.866	0.099	1.8E-01[b]	
r_G (NB&S, B&NS)	1.013	0.129	9.2E-01[b]	
r_G (NB&S, B&S)	1.090	0.167	5.9E-01[b]	
r_G (NB&S, NB&NS)	0.929	0.153	6.4E-01[b]	

Table 4. The proportion of the phenotypic variance and genetic correlation between the status of breastfeeding and maternal smoking around birth for educational attainment. [a]Testing if the estimate is different from 0. [b]Testing if the estimate is different from 1; The genetic correlation (r_G) between NB&NS and B&NS is significantly different from 1 as an evidence of G × E. Even after a multiple testing correction (p-value threshold = 0.05/24 = 0.002), the interaction remained significant. ***P-value < 0.001; **P-value < 0.01; *P-value < 0.05.

FIGURE: TABLE FROM LEE ET AL., (2017) ▸ LINK

→ Another example, from Estonia before and after Soviet era.

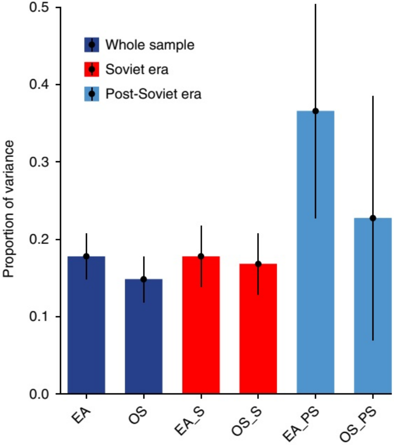


Fig. 2 | SNP heritabilities showing the proportion of variance explained by additive effects of common SNPs for the whole EGCUT sample and for the Soviet and post-Soviet groups using a cutoff of 15 years. SNP heritabilities

Figure: Table from Rimfeld et al. (2018) ▸ Link

Beyond Heritability:

- → Which genetic variants matter and why do they matter?
  - → Animal models.
  - → Gene discovery from association studies.
- → Can we construct predictive (individual-level) variables from molecular genetic data?
  - → Candidate genes.
  - → Genome-wide polygenic scores.

Animal Models

- 97% overlap of genes in humans in mice!
- Potential for:
  - 1- Identify causal variants in humans (GWAS)
  - 2- Identify the variants in mice
  - 3- Validate findings by inbreeding mice with same causal variants and test mechanisms
- Feasible way to establish causal mechanisms
- Also possible to validate using worms (~50% of known human disease match), fruit fly (~75%), etc...

└ CANDIDATE GENES

- Traditional approach was to examine *candidate genes*.
  - Selected from prior knowledge/research (e.g. Animal Models).
- Unfortunately, the candidate genes approach suffers from a severe replication crises.
  - Weak effects combined with small sample sizes.
  - Studies are underpowered and results prone to the ‘winner’s curse’.
  - Ignores linkage disequilibrium between variants.
- Editorial Statement at *Behavior Genetics*:
  - “Many of the published findings of the last decade are wrong or misleading and have not contributed to real advances in knowledge” (Hewitt 2012 ▸ Link).
- “Most reported genetic associations with general intelligence are probably false positives” (Chabris et al., (2012) ▸ Link)

└Candidate Genes

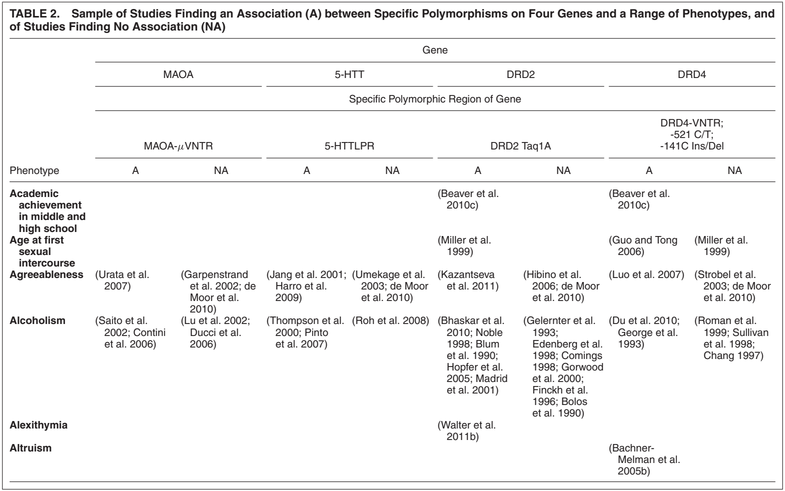
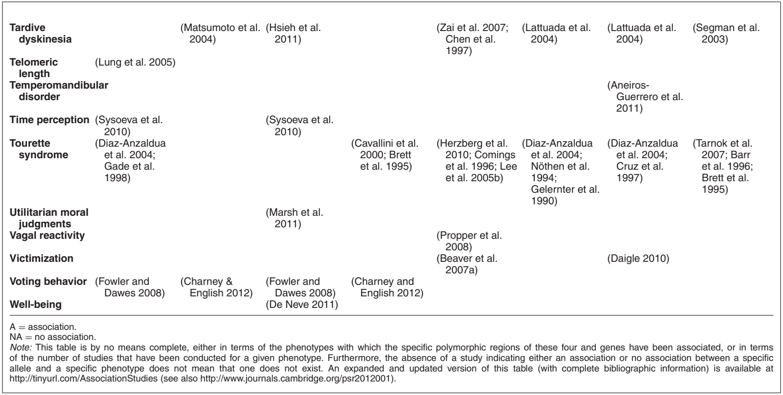


TABLE 2. Sample of Studies Finding an Association (A) between Specific Polymorphisms on Four Genes and a Range of Phenotypes, and of Studies Finding No Association (NA)

	Gene							
	MAOA		5-HTT		DRD2		DRD4	
	Specific Polymorphic Region of Gene							
	MAOA-μVNTR		5-HTTLPR		DRD2 Taq1A		DRD4-VNTR; -521 C/T; -141C Ins/Del	
Phenotype	A	NA	A	NA	A	NA	A	NA
Academic achievement in middle and high school					(Beaver et al. 2010c)		(Beaver et al. 2010c)	
Age at first sexual intercourse					(Miller et al. 1999)		(Guo and Tong 2006)	(Miller et al. 1999)
Agreeableness	(Urata et al. 2007)	(Garpenstrand et al. 2002; de Moor et al. 2010)	(Jang et al. 2001; Harro et al. 2009)	(Umekage et al. 2003; de Moor et al. 2010)	(Kazantseva et al. 2011)	(Hibino et al. 2006; de Moor et al. 2010)	(Luo et al. 2007)	(Strobel et al. 2003; de Moor et al. 2010)
Alcoholism	(Saito et al. 2002; Contini et al. 2006)	(Lu et al. 2002; Ducci et al. 2006)	(Thompson et al. 2000; Pinto et al. 2007)	(Roh et al. 2008)	(Bhaskar et al. 2010; Noble 1998; Blum et al. 1990; Hopfer et al. 2005; Madrid et al. 2001)	(Gelernter et al. 1993; Edenberg et al. 1998; Comings 1998; Gorwood et al. 2000; Finckh et al. 1996; Bolos et al. 1990)	(Du et al. 2010; George et al. 1993)	(Roman et al. 1999; Sullivan et al. 1998; Chang 1997)
Alexithymia					(Walter et al. 2011b)			
Altruism							(Bachner-Melman et al. 2005b)	

Figure: 4 Candidate Genes Predict everything from Academic achievement to well-being (Charney and English 2012 ▸ Link)

14 Pages latter...



Tardive dyskinesia		(Matsumoto et al. 2004)	(Hsieh et al. 2011)		(Zai et al. 2007; Chen et al. 1997)	(Lattuada et al. 2004)	(Lattuada et al. 2004)	(Segman et al. 2003)
Telomeric length	(Lung et al. 2005)							
Temperomandibular disorder							(Aneiros-Guerrero et al. 2011)	
Time perception	(Sysoeva et al. 2010)		(Sysoeva et al. 2010)					
Tourette syndrome	(Diaz-Anzaldua et al. 2004; Gade et al. 1998)			(Cavallini et al. 2000; Brett et al. 1995)	(Herzberg et al. 2010; Comings et al. 1996; Lee et al. 2005b)	(Diaz-Anzaldua et al. 2004; Nöthen et al. 1994; Gelernter et al. 1990)	(Diaz-Anzaldua et al. 2004; Cruz et al. 1997)	(Tarnok et al. 2007; Barr et al. 1996; Brett et al. 1995)
Utilitarian moral judgments			(Marsh et al. 2011)					
Vagal reactivity					(Propper et al. 2008)			
Victimization					(Beaver et al. 2007a)		(Daigle 2010)	
Voting behavior	(Fowler and Dawes 2008)	(Charney & English 2012)	(Fowler and Dawes 2008)	(Charney and English 2012)				
Well-being			(De Neve 2011)					

A = association.
NA = no association.
Note: This table is by no means complete, either in terms of the phenotypes with which the specific polymorphic regions of these four and genes have been associated, or in terms of the number of studies that have been conducted for a given phenotype. Furthermore, the absence of a study indicating either an association or no association between a specific allele and a specific phenotype does not mean that one does not exist. An expanded and updated version of this table (with complete bibliographic information) is available at http://tinyurl.com/AssociationStudies (see also http://www.journals.cambridge.org/psr2012001).

FIGURE: 4 CANDIDATE GENES PREDICT EVERYTHING FROM ACADEMIC ACHIEVEMENT TO WELL-BEING (CHARNEY AND ENGLISH 2012 ▸ LINK)

Candidate Genes done right:

- Larger samples.
- Focus on genes with strong genetic effects and or known function from animal models.
- Examples:
  - FTO gene (rs9930506) on BMI,
  - “Mr. Big” (rs16969968) on Smoking,
  - APOE genes (rs429358 and rs7412) on Alzheimer.
- Control for possible confounds due to linkage disequilibrium.

- → Two examples in Economics:
- → Biroli 2015 ▸ Link studies the effect of the FTO genes on obesity.
- → Benjamin et al., 2015 ▸ Link studies the effect of the “Mr. Big” genes on smoking.