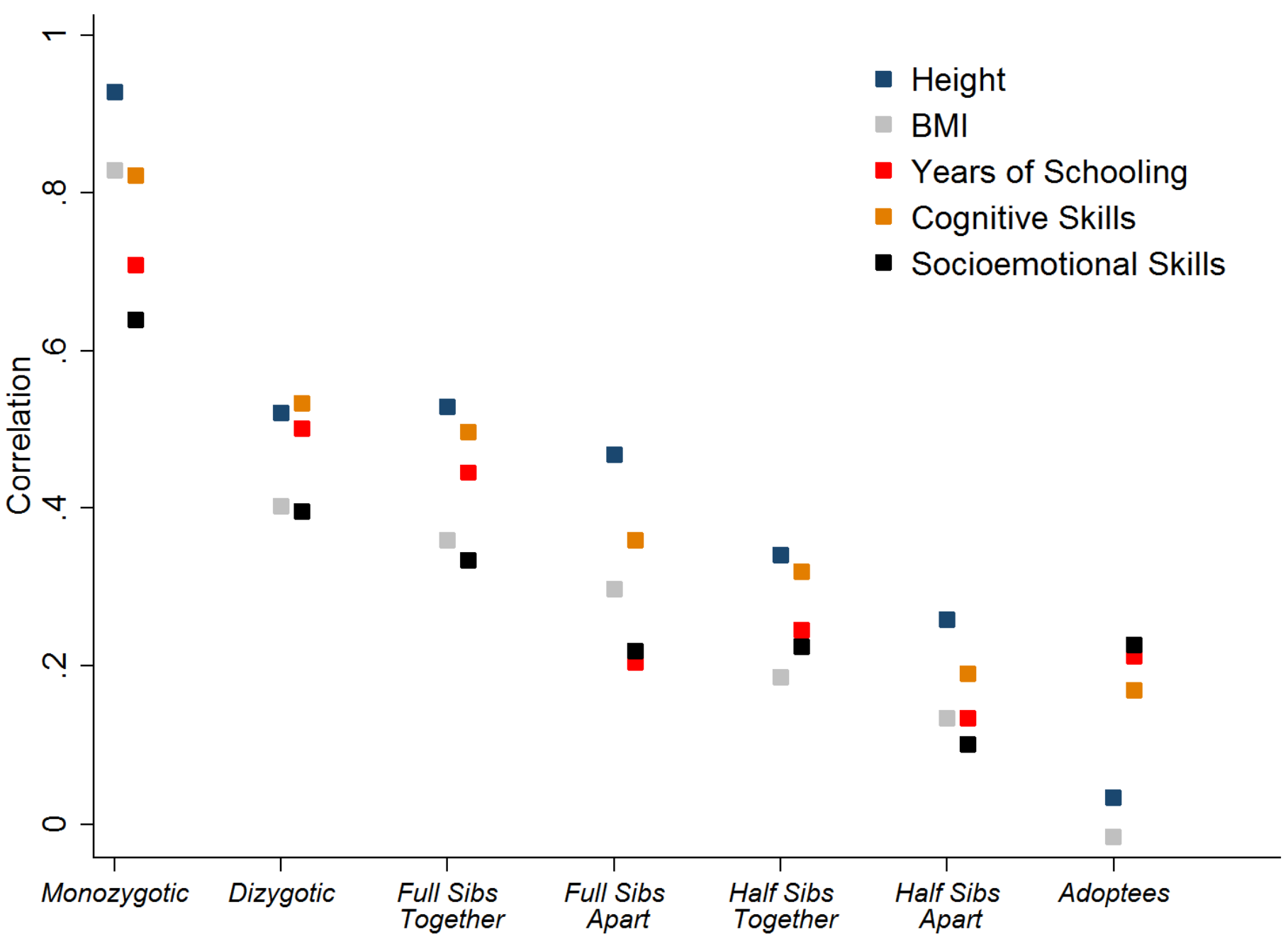
Predicting Behavioral Traits from Genomic Data

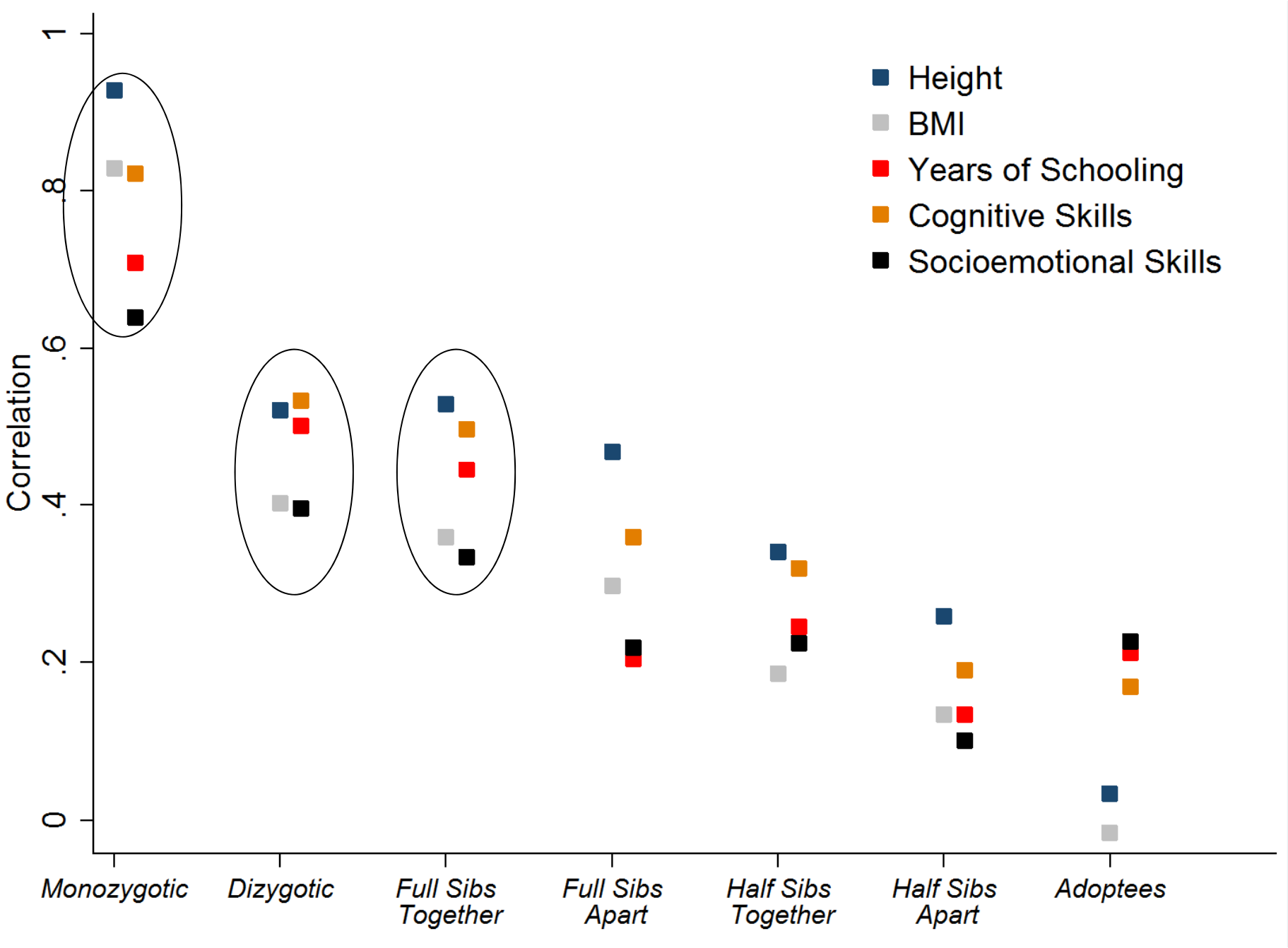
David Cesarini
NYU

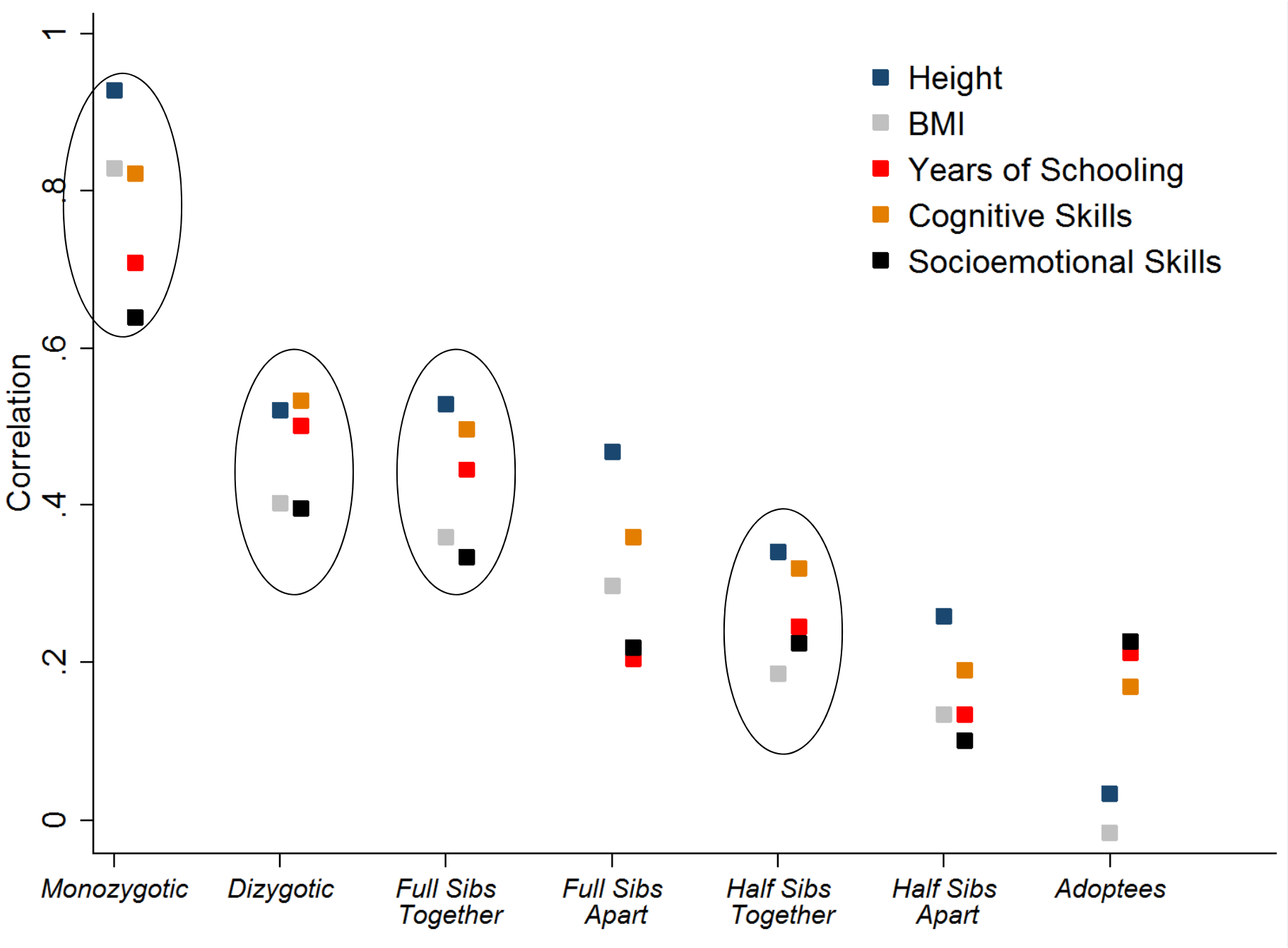
Social and Behavioral Sciences for National Security

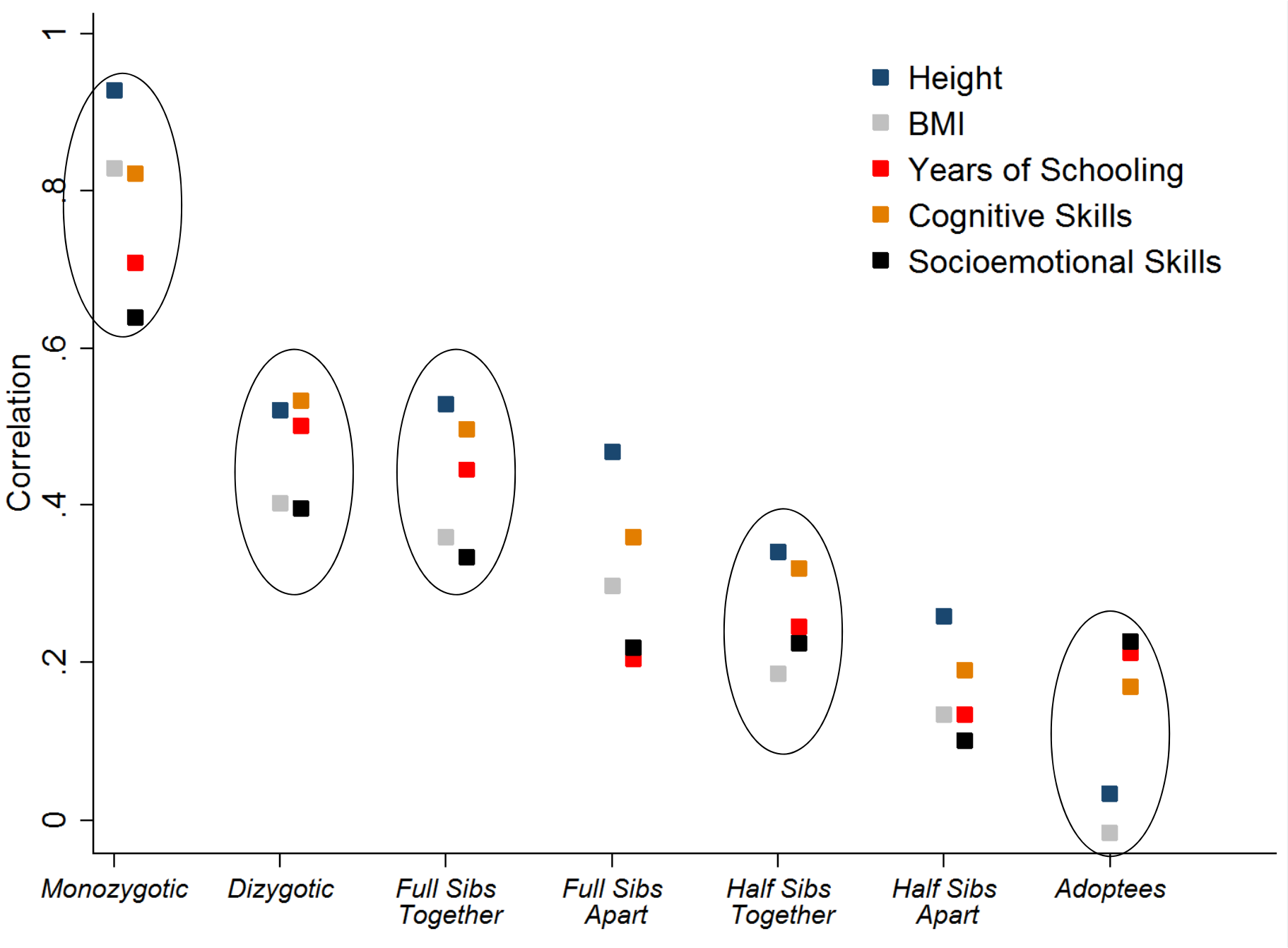
Outline

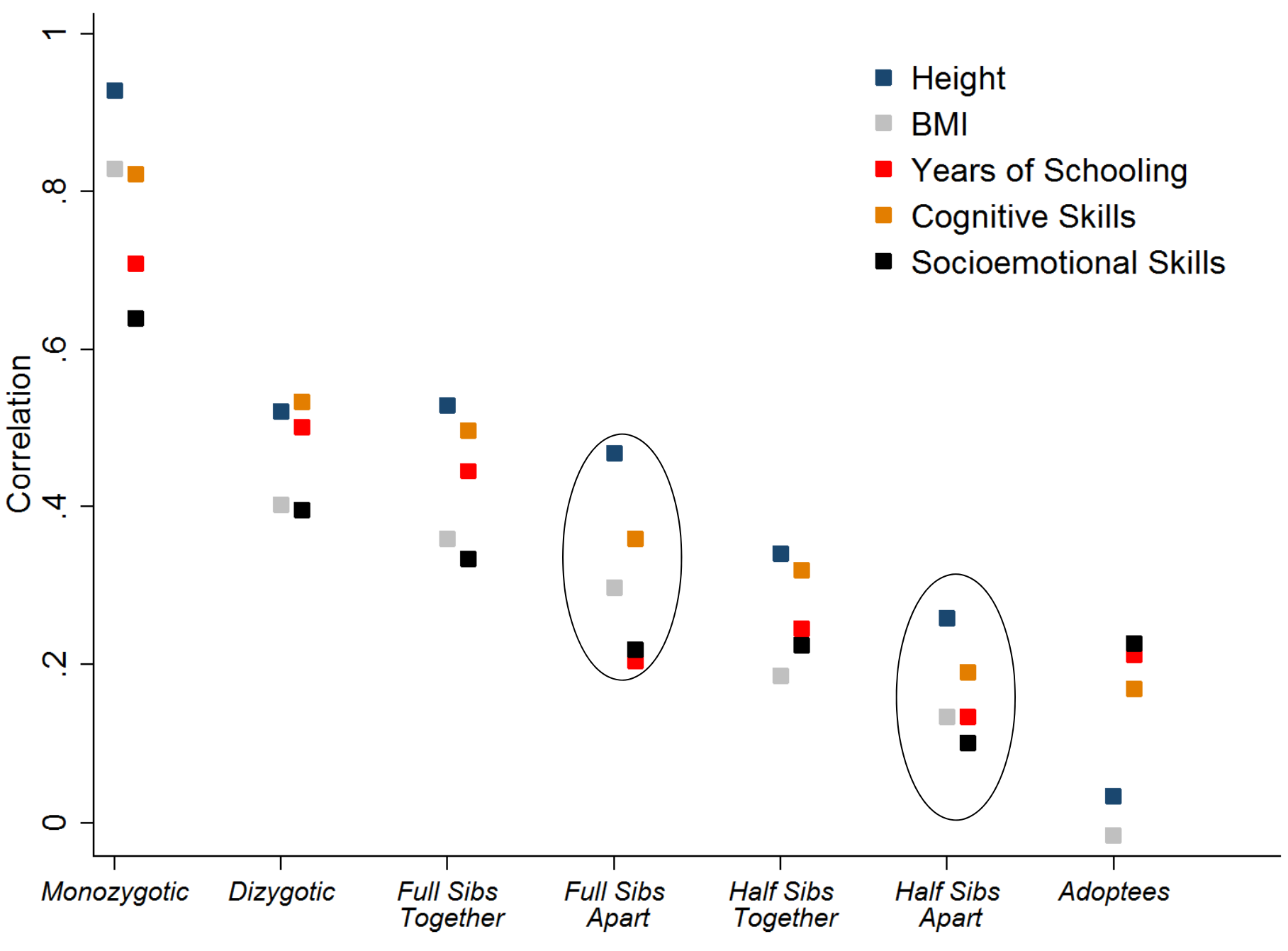
1. **Preliminaries**
    1. **Twin- and Family Studies**
    2. Sequencing Costs
2. Molecular Genetics Research
    a. Strategies for Gene Discovery
    b. Canonical Findings
    c. Predicting Behavioral Traits from DNA
    d. Organizing the Evidence
3. Conclusion


Correlation
0
.2
.4
.6
.8
1
Height
BMI
Years of Schooling
Cognitive Skills
Socioemotional Skills
Monozygotic
Dizygotic
Full Sibs Together
Full Sibs Apart
Half Sibs Together
Half Sibs Apart
Adoptees


Correlation
0
.2
.4
.6
.8
1
Height
BMI
Years of Schooling
Cognitive Skills
Socioemotional Skills
Monozygotic
Dizygotic
Full Sibs Together
Full Sibs Apart
Half Sibs Together
Half Sibs Apart
Adoptees

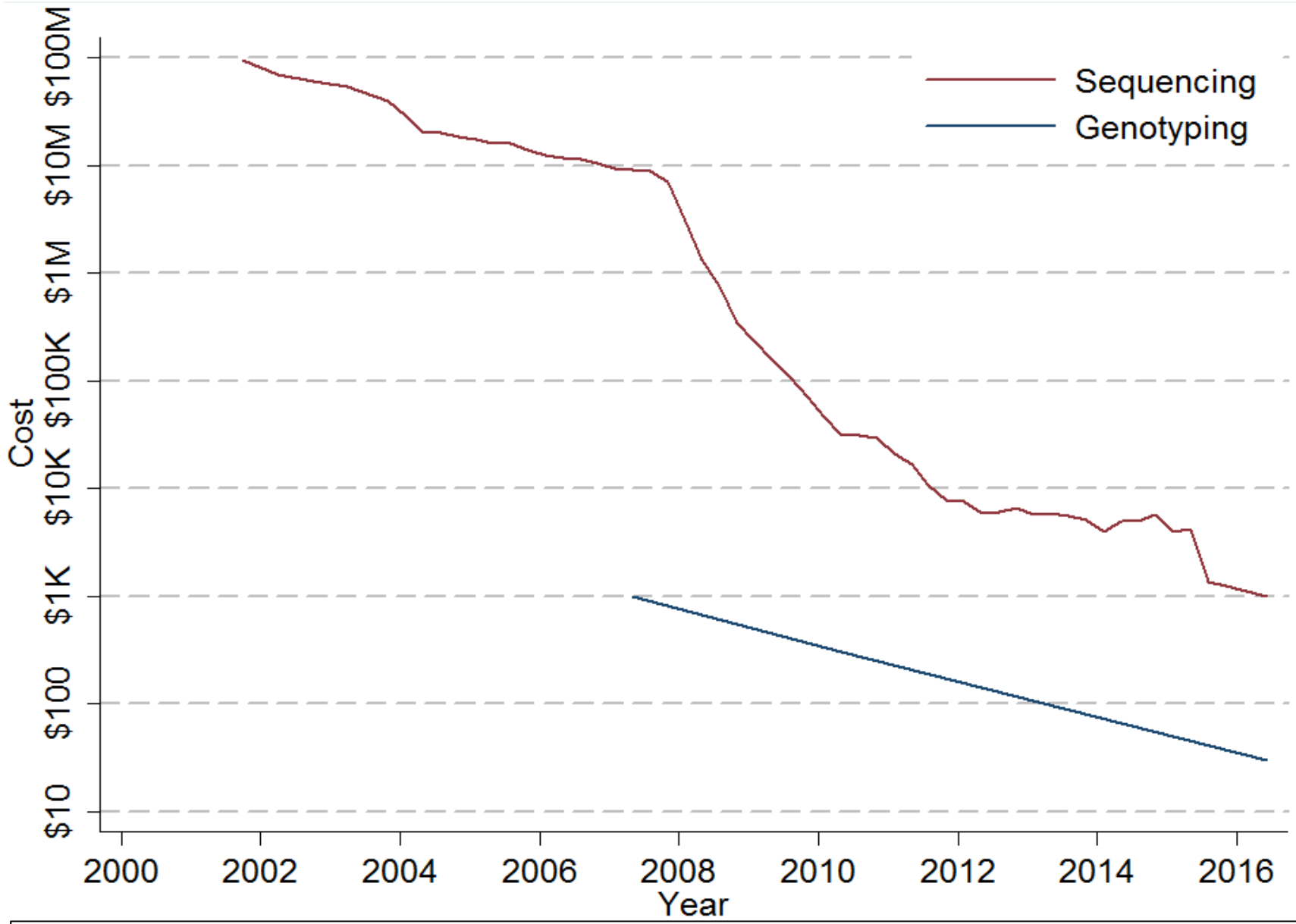

Correlation
0
.2
.4
.6
.8
1
Height
BMI
Years of Schooling
Cognitive Skills
Socioemotional Skills
Monozygotic
Dizygotic
Full Sibs Together
Full Sibs Apart
Half Sibs Together
Half Sibs Apart
Adoptees


Correlation
0
.2
.4
.6
.8
1
Height
BMI
Years of Schooling
Cognitive Skills
Socioemotional Skills
Monozygotic
Dizygotic
Full Sibs Together
Full Sibs Apart
Half Sibs Together
Half Sibs Apart
Adoptees


Correlation
0
.2
.4
.6
.8
1
Height
BMI
Years of Schooling
Cognitive Skills
Socioemotional Skills
Monozygotic
Dizygotic
Full Sibs Together
Full Sibs Apart
Half Sibs Together
Half Sibs Apart
Adoptees

Outline

1. **Preliminaries**
    1. Twin- and Family Studies
    2. **Sequencing Costs**
2. Molecular Genetics Research
    a. Strategies for Gene Discovery
    b. Canonical Findings
    c. Predicting Behavioral Traits from DNA
    d. Organizing the Evidence
3. Conclusion

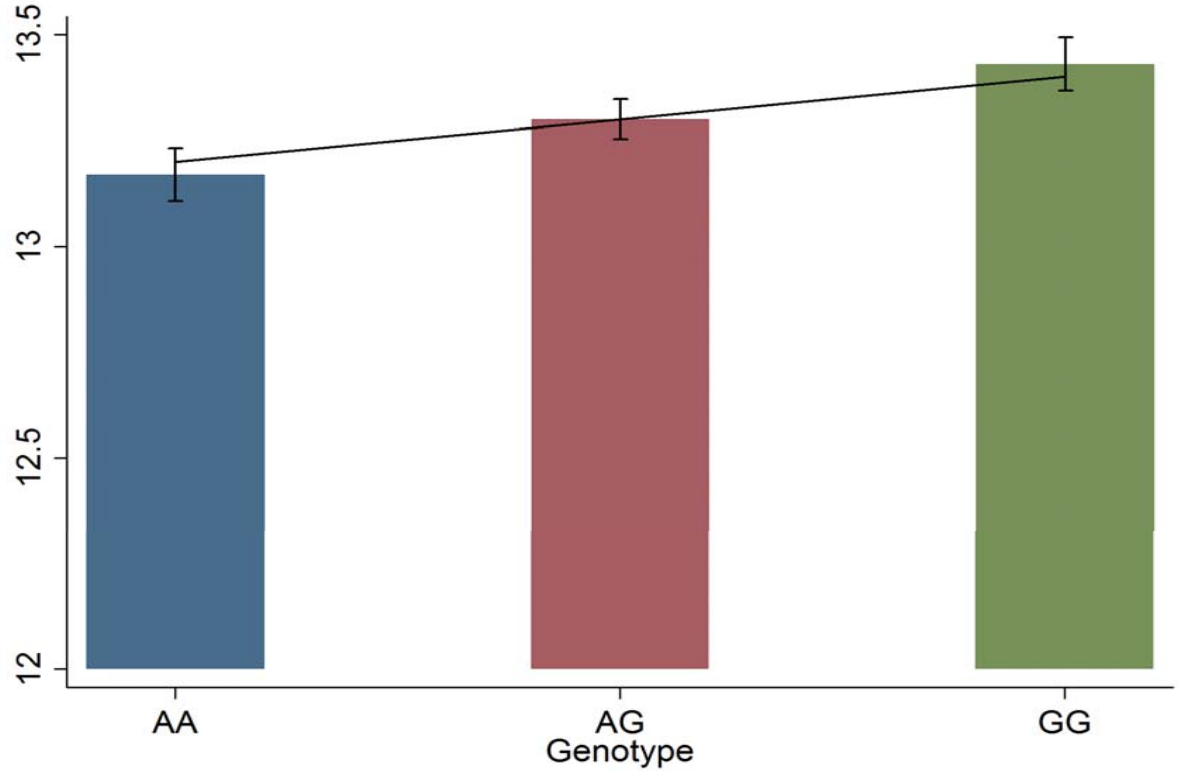


Note. Genotyping costs from multiple sources. Sequencing costs from *NIH* (genome.gov/sequencingcosts).

Outline

1. Preliminaries
    1. Twin- and Family Studies
    2. Sequencing Costs
2. **Molecular Genetics Research**
    a. **Strategies for Gene Discovery**
    b. Canonical Findings
    c. Predicting Behavioral Traits from DNA
    d. Organizing the Evidence
3. Conclusion

Gene Discovery



Test of null that the (regression-adjusted) means of individuals with different genotypes are the same.

Important considerations:

- Determining which *J* variants to test for association.
- Minimizing problems caused by stratification biases.
- Multiple-hypotheses adjustment.

Outline

1. Preliminaries
    1. Twin- and Family Studies
    2. Sequencing Costs
2. **Molecular Genetics Research**
    a. Strategies for Gene Discovery
    b. Canonical Findings
    c. Predicting Behavioral Traits from DNA
    d. Organizing the Evidence
3. Conclusion

Candidate-Gene Study (*J* small)

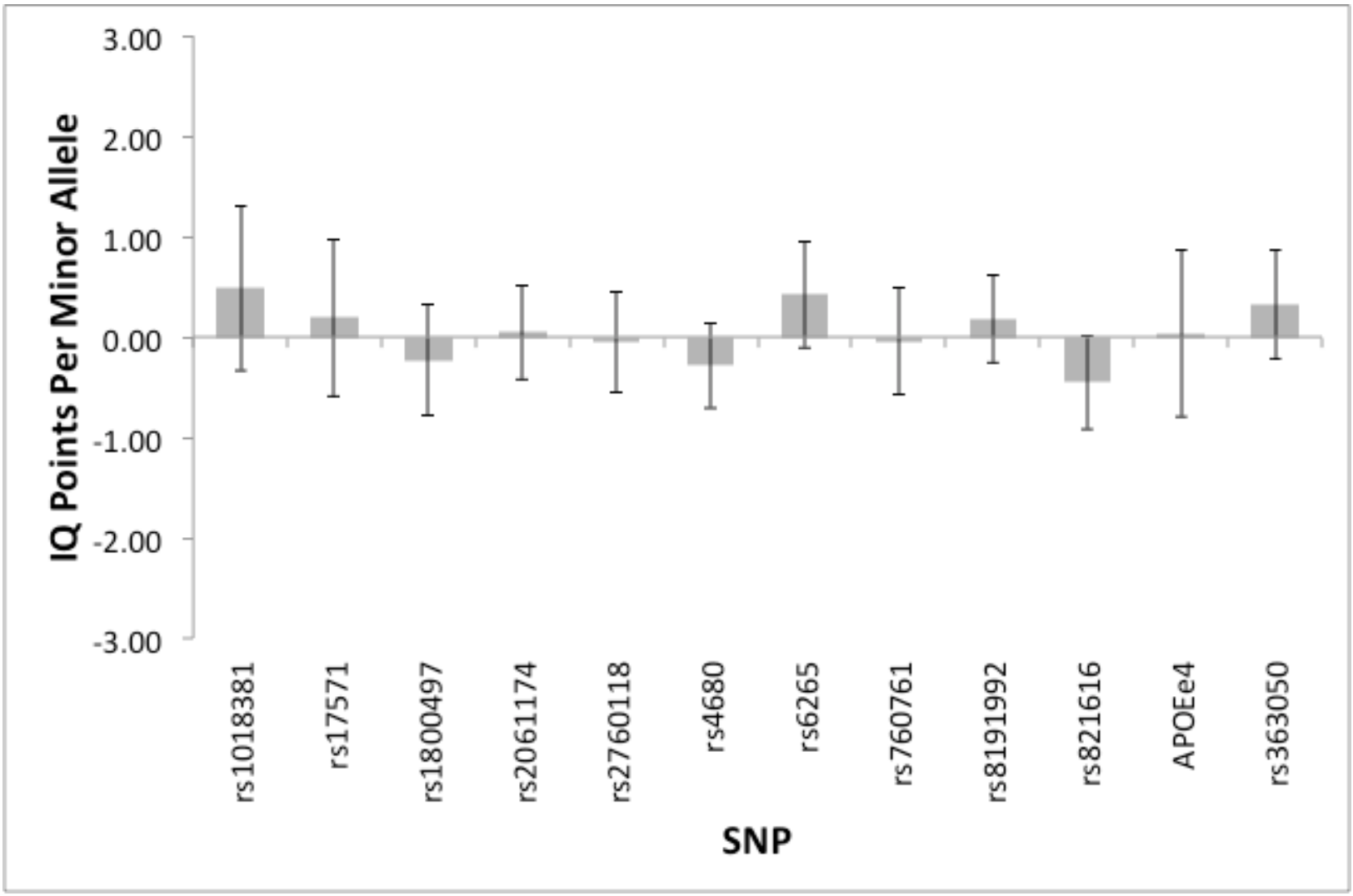
- Specify *ex ante* hypotheses about small set of SNPs based on believed biological function.
- Typical significance threshold: 0.05/ *J*.
- Eminently reasonable, and has worked when hypotheses are direct. (e.g., *APOE* and Alzheimer's)
- But most reported associations with behavioral traits have failed to replicate.
  - Weak hypotheses (except for highly proximal behaviors).
  - Low power (in the small samples typically used).
  - Population stratification.
  - Uncorrected multiple hypothesis testing / publication bias.

Most Reported Genetic Associations With General Intelligence Are Probably False Positives

Psychological Science
23(11) 1314–1323
© The Author(s) 2012
Reprints and permissi
sagepub.com/journalsP
DOI: 10.1177/095679
http://pss.sagepub.com
SAGE

Christopher F. Chabris[1], Benjamin M. Hebert[2], Daniel J. Benjamin[3], Jonathan Beauchamp[2], David Cesarini[4], Matthijs van der Loos[5], Magnus Johannesson[6], Patrik K. E. Magnusson[7], Paul Lichtenstein[7], Craig S. Atwood[8], Jeremy Freese[9], Taissa S. Hauser[10], Robert M. Hauser[10], Nicholas Christakis[11,12], and David Laibson[2]

Replication Results



Behav Genet (2012) 42:1–2
DOI 10.1007/s10519-011-9504-z

BRIEF COMMUNICATION

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

John K. Hewitt

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

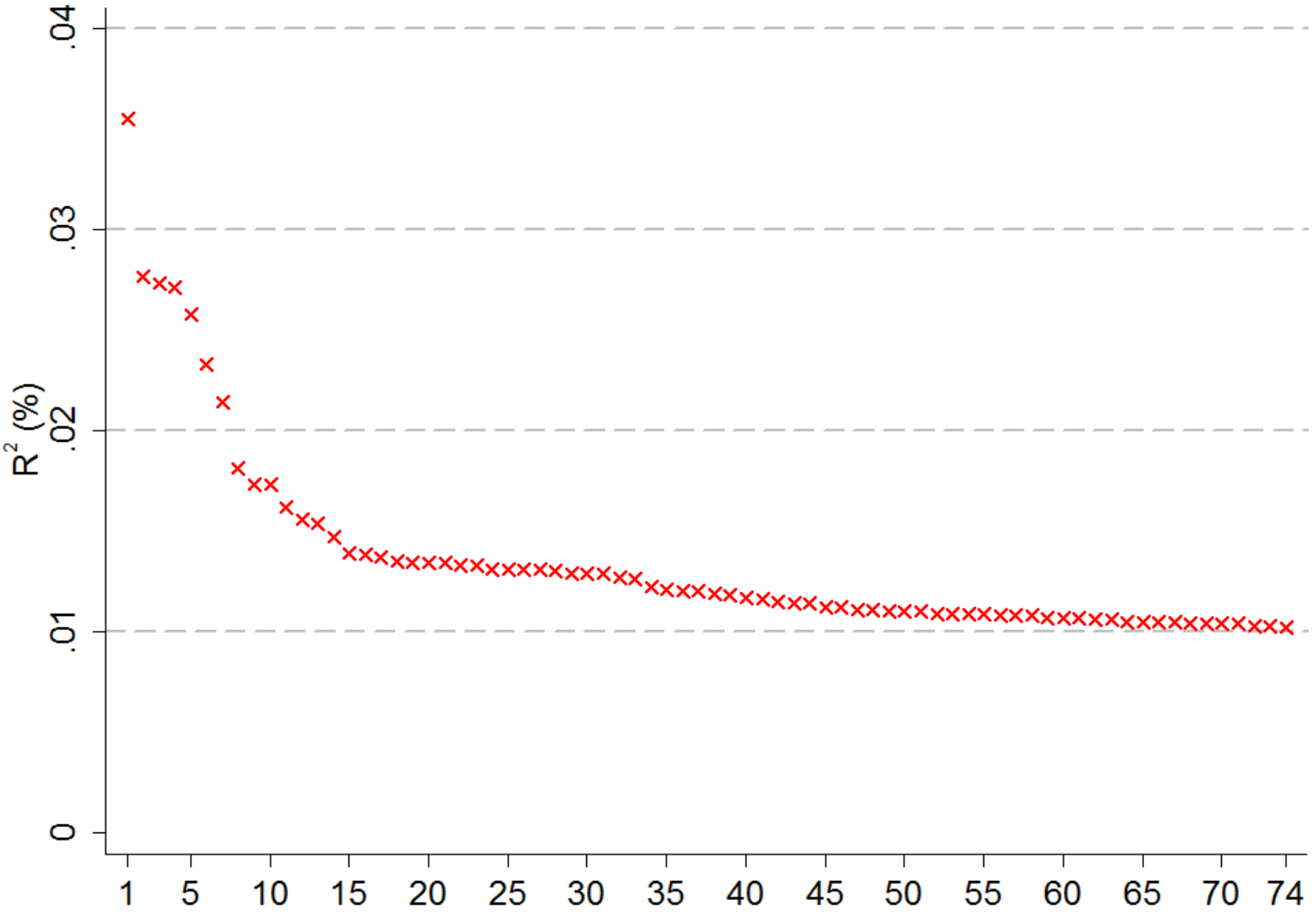
Genome-Wide Association Study (GWAS) (*J* large)

- Atheoretical testing of all SNPs measured using modern technologies (~0.5-2.5M).
- Set significance threshold $\alpha = 5 \times 10^{-8}$ (since ≈1M independent SNPs in genome).
- Some advantages of GWAS:
  - Hypothesis-free design makes the need to correct for multiple hypothesis testing transparent.
  - Genome-wide data makes it easier to minimize stratification biases.
  - Conditional on genome-wide significance, almost certain to be true.

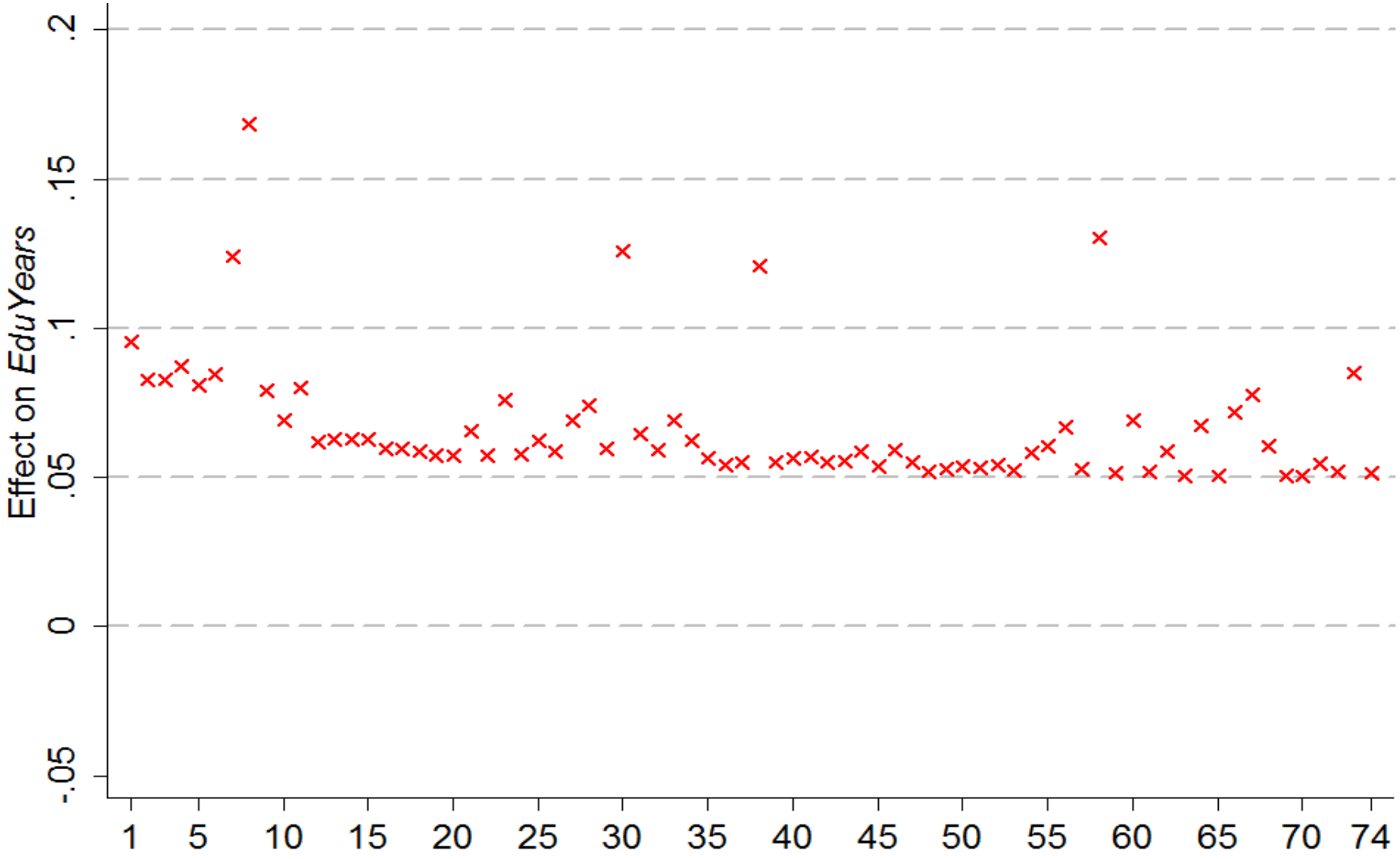
Table 1. Sample Size and Number of Genome-Wide Significant Associations

Years of Education			Height			Body-mass index		
Ref.	***N***	**#Hits**	**Ref.**	***N***	**#Hits**	**Ref.**	***N***	**#Hits**
[1]	9,538	0	[6]	11,536	1	[12]	11,536	0
[2]	7,500	0	[7]	15,821	12	[13]	123,865	19
[3]	101,069	1	[8]	16,482	20	[14]	339,224	97
[4]	126,069	4	[9]	30,968	27			
[5]	293,723	74	[10]	183,727	180			
[6]	405,072	162	[11]	253,288	697			

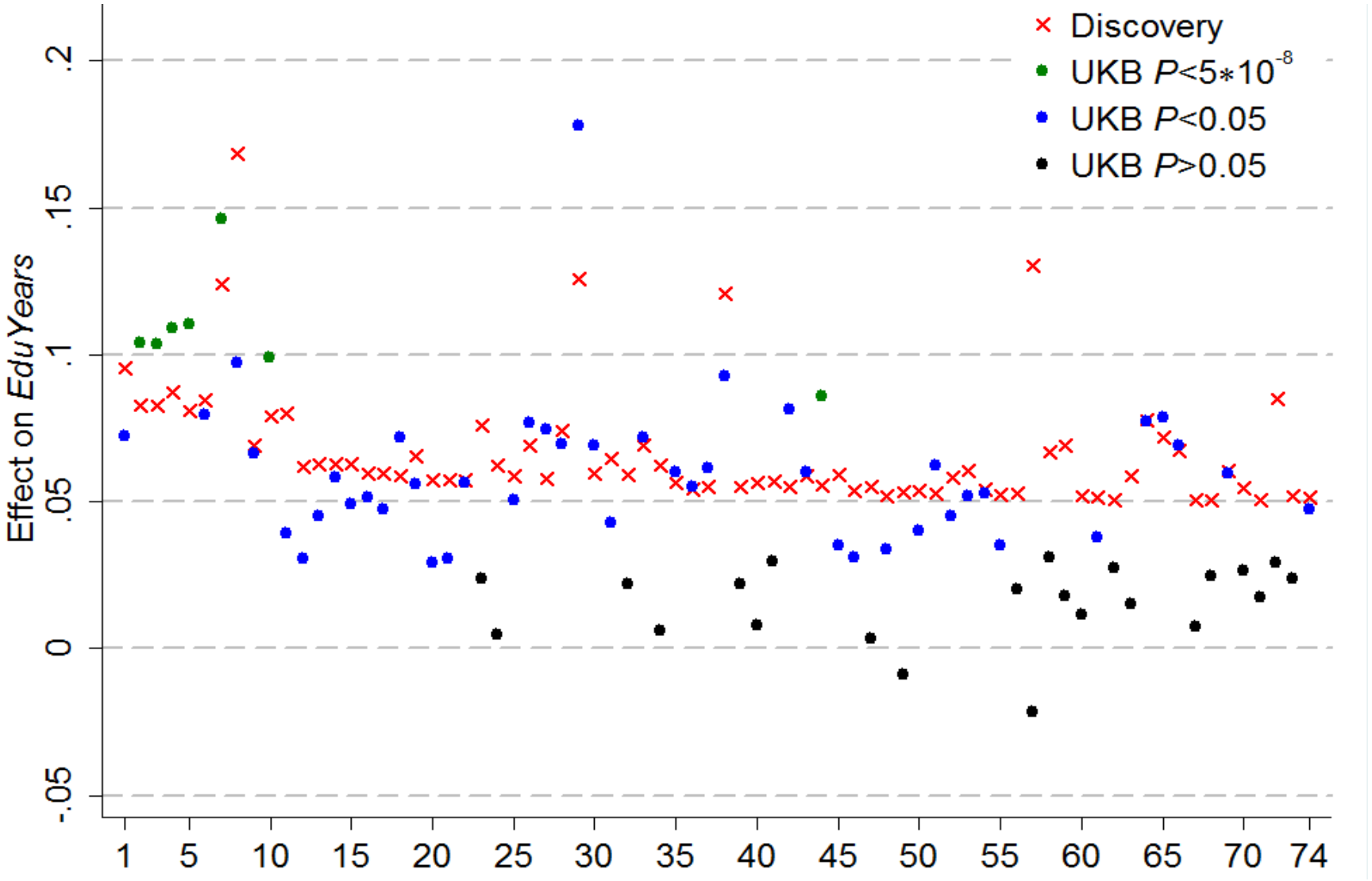
Note. Relationship between discovery sample size and the number of independent loci (“hits”) identified at genome-wide significance.



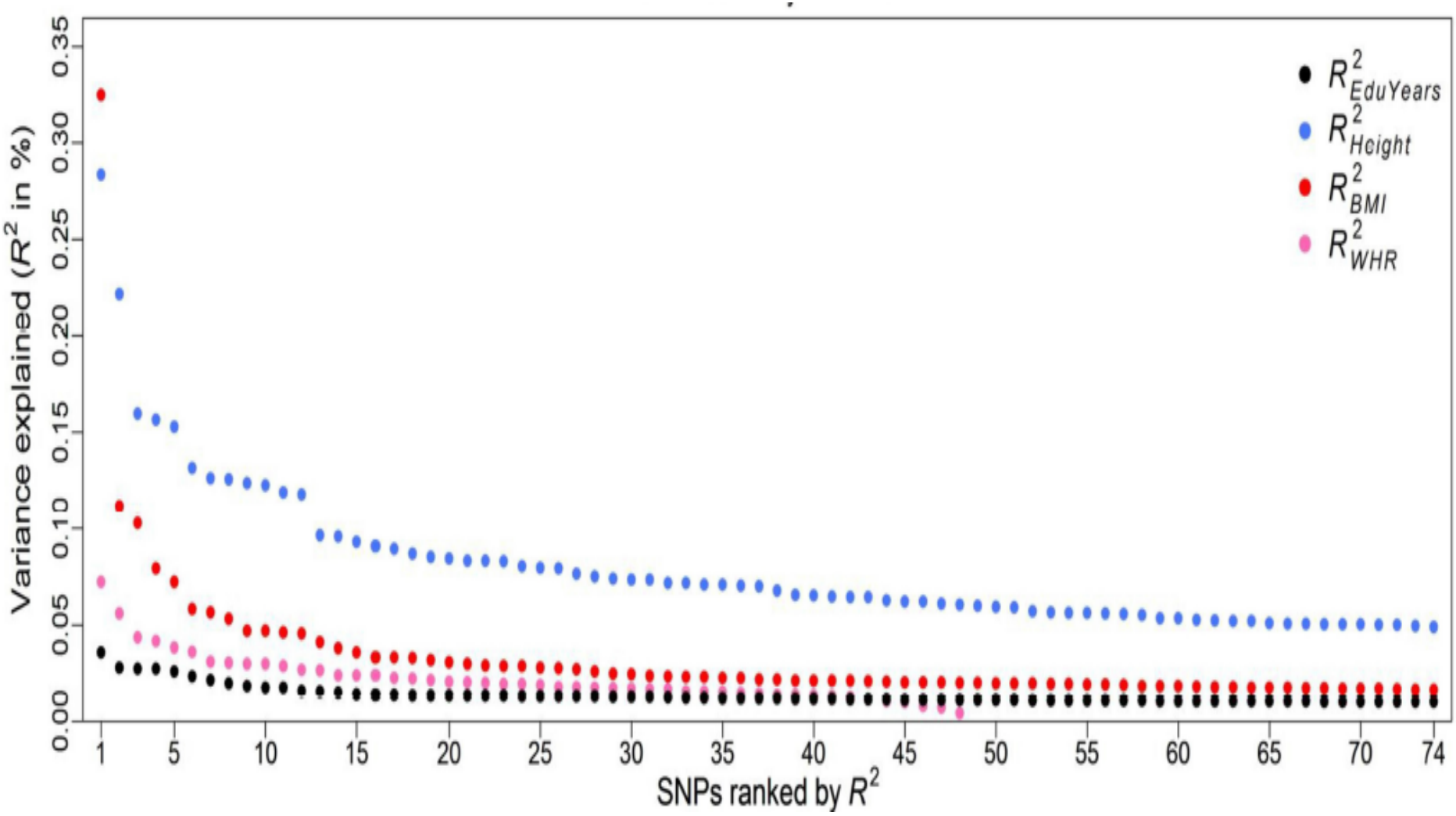
Note. The distribution of effect sizes (R^2) for the 74 hits reported by Okbay *et al.* (2016) for educational attainment.



Note. The distribution of effect sizes (years per allele) for the 74 hits reported by Okbay *et al.* (2016) for educational attainment.



Note. In UKB sample (N = 111,349), 72/74 SNPs have predicted sign, 52 replicate at $P < 0.05$ and 7 at $P < 5 \times 10^{-8}$.



Note. Effects (R^2) benchmarked against the top 74 genome-wide significant hits reported for height and body mass index.

Outline

1. Preliminaries
    1. Twin- and Family Studies
    2. Sequencing Costs
2. Molecular Genetics Research
    a. Strategies for Gene Discovery
    b. Canonical Findings
    c. Predicting Behavioral Traits from DNA
    d. Organizing the Evidence
3. Conclusion

Polygenic Scores

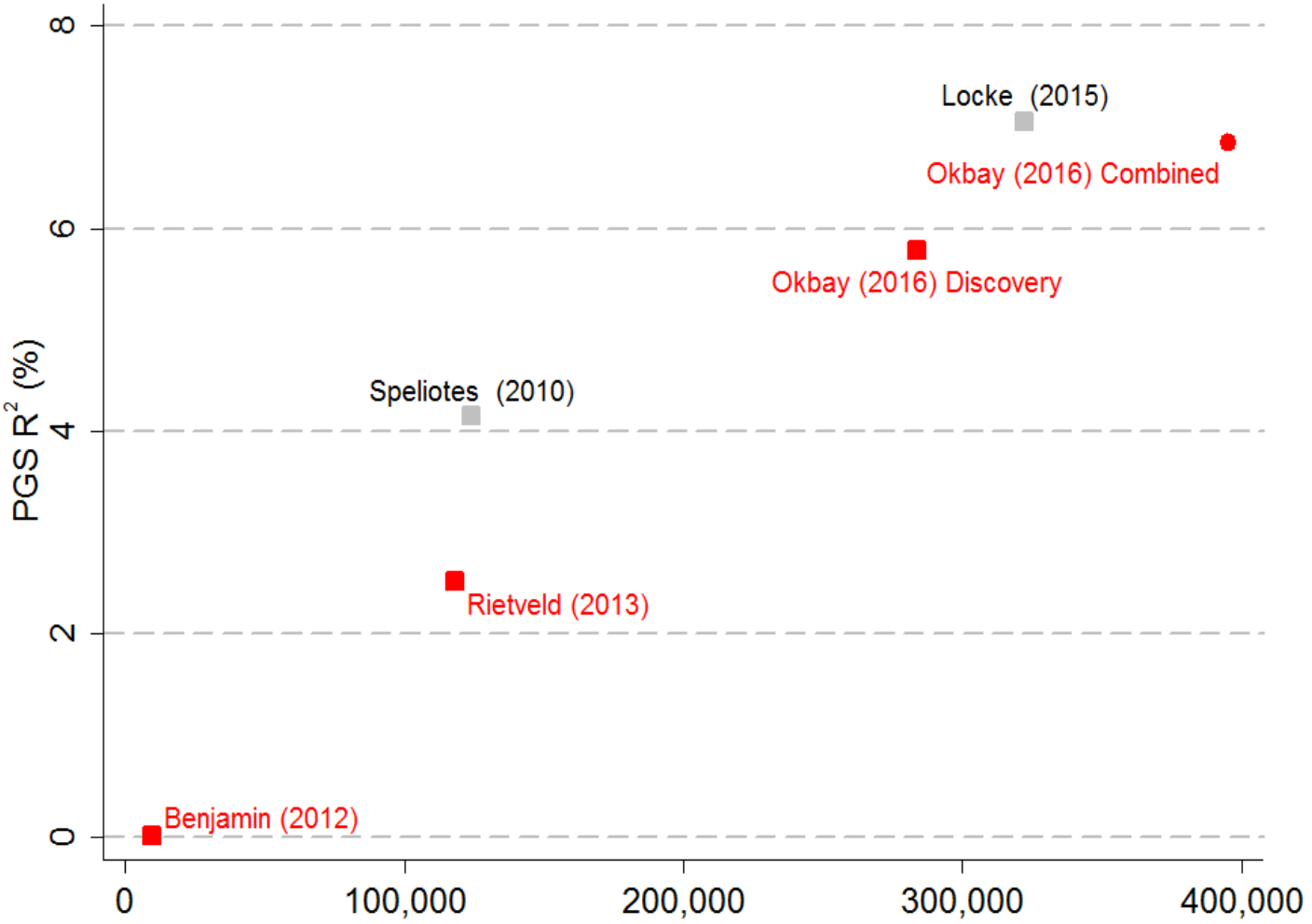
- Can use GWAS estimates to predict *i*'s outcome from *J* measured genetic variants:

$$\hat{g}_i = \sum_{j=1}^{J} x_{ij}\hat{\beta}_j$$

- x_{ij} is individual *i*'s genotype (0,1,2) at variant *j* and $\hat{\beta}_j$ is our preferred estimate of variant *j*'s effect.
- Predictive power: $\hat{r}^2(\hat{g}_i, y_i)$.
- As $N \to \infty$, better estimates of β_j, and

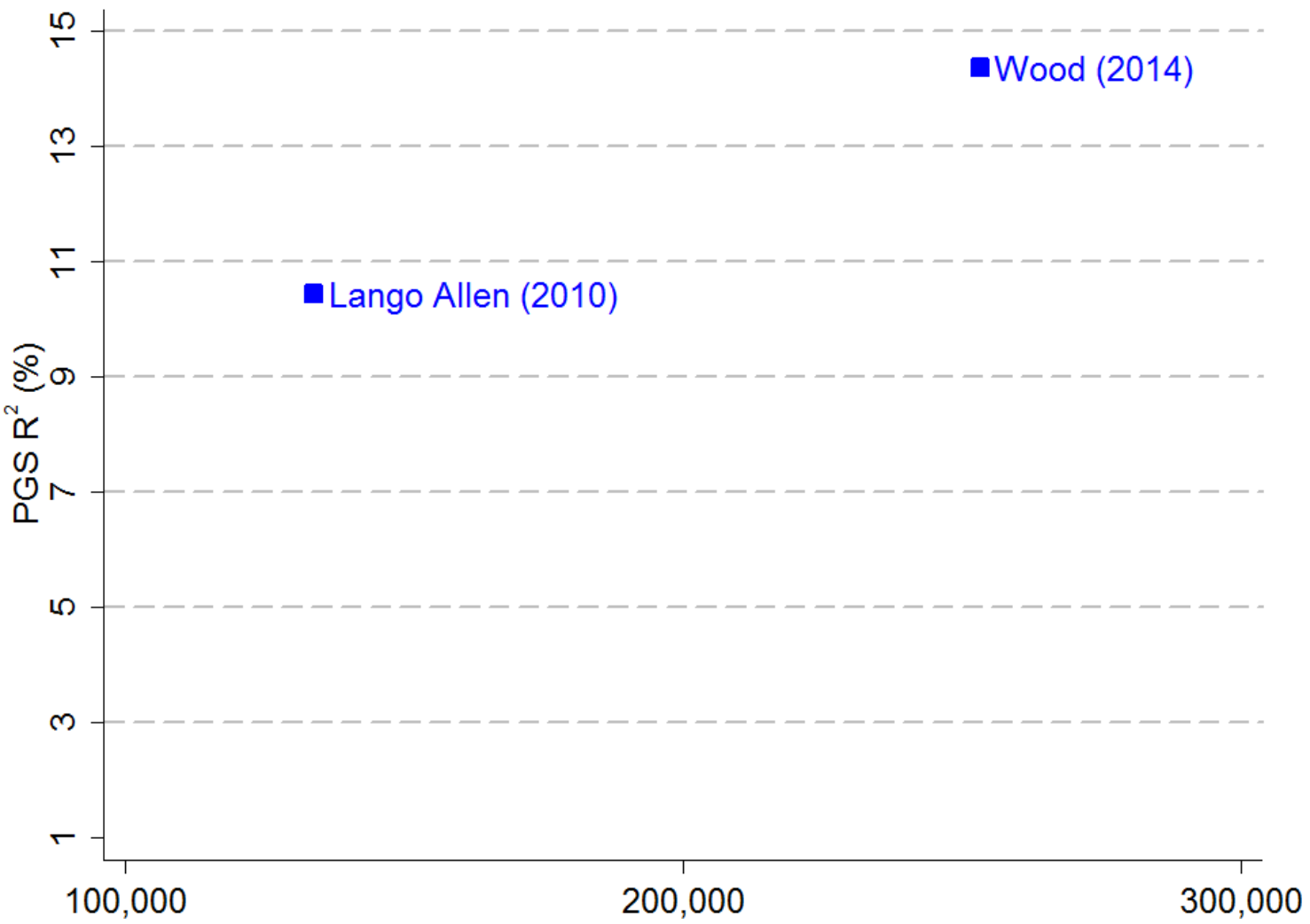
$$\hat{r}^2(\hat{g}_i, y_i) \to r^2(g_i, y_i).$$

Predicting BMI and Education



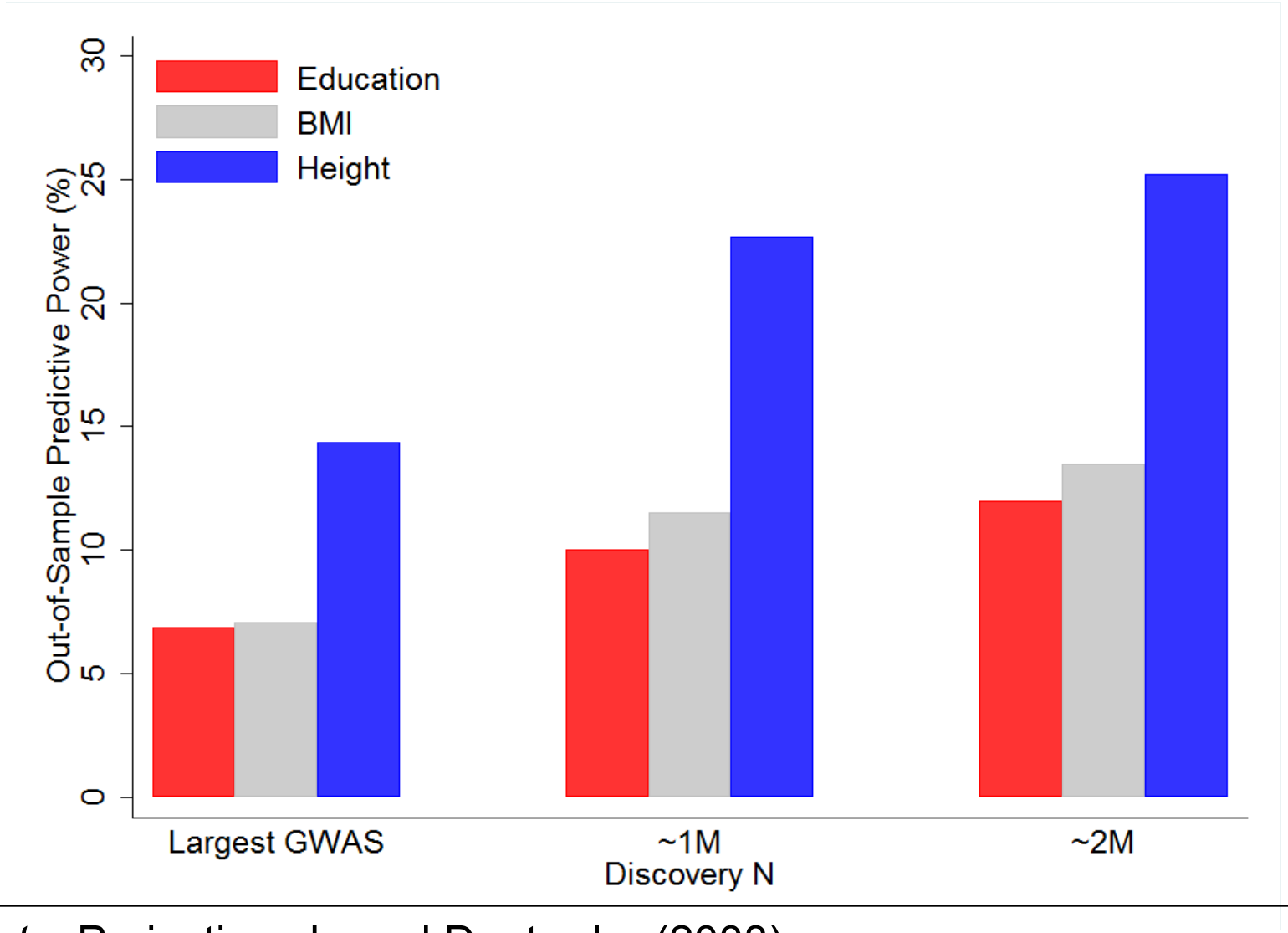
Note. Polygenic scores estimated using LD Pred (Vilhjalmsson *et al.* 2015). All analyses in European-ancestry subjects in HRS.

Predicting Height



Note. Polygenic scores estimated using LD Pred (Vilhjalmsson *et al.* 2015). All analyses in European-ancestry subjects in HRS.

Future Polygenic Scores



Note. Projections based Daetwyler (2008).

Outline

1. Preliminaries
    1. Twin- and Family Studies
    2. Sequencing Costs
2. Molecular Genetics Research
    a. Strategies for Gene Discovery
    b. Canonical Findings
    c. Predicting Behavioral Traits from DNA
    d. Organizing the Evidence
3. Conclusion

Four Stylized Facts

1. Small-*N* candidate gene studies of behavioral traits have a weak replication track record.
2. Steady increase in genetic associations identified in GWA studies as $N \uparrow$. Strong replication record.
3. Despite the fact that GWAS has had some successes, much of the heritability is “missing”.
4. As larger *N* available to estimate weights for polygenic scores, their predictive power rises.

Fourth “Law” as a Unifying Principle

“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. Curr Dir in Psych Sci, 2014)

Calibration: Power Analysis

- Either there is a true association or not.
- If associated, $R^2 \approx 0.02\%$. Else, $R^2 \approx 0\%$.
- Sample size for 80% power: 39,240.
- Now suppose significant association at α = .05.
- What should we conclude?

Bayesian Analysis of a Candidate-Gene Study

(based on Wacholder et al., 2004; Ioannidis, 2005; Benjamin et al., 2012)

Given significant at α = .05, posterior probability of true association with effect size R^2 = 0.02%.

N		100	10K	100K
Power		0.052	0.294	0.993
Prior	0.1%	0.1%	0.6%	2%
	1%	1%	6%	17%
	5%	5.2%	24%	51%
	10%	10.4%	39%	69%

Bayes' Rule: $\mathrm{P}(True|Sig) = \frac{power \times prior}{power \times prior + 0.05 \times (1 - prior)}$.

Bayesian Analysis of a Candidate-Gene Study

(based on Wacholder et al., 2004; Ioannidis, 2005; Benjamin et al., 2012)

Given significant at α = .05, posterior probability of true association with effect size R^2 = 0.02%.

N		100	10K	100K
Power		0.052	0.294	0.993
Prior	0.1%	0.1%	0.6%	2%
	1%	1%	6%	17%
	5%	5.2%	24%	51%
	10%	10.4%	39%	69%

Bayes' Rule: $\mathrm{P}(True|Sig) = \frac{power \times prior}{power \times prior + 0.05 \times (1 - prior)}$.

Candidate-Gene Study: Design Calculations

(based on Gelman and Carlin 2014)

Suppose R^2 = 0.02% and significant at α = .05:

1. How often will estimate have the right sign?
2. How exaggerated is the magnitude of the estimate?

N	100	1K	10K	100K
Power	0.052	0.073	0.294	0.993
$\mathrm{P}(Right\ Sign \mid P < 0.05)$	66%	89%	99.9%	100%
$\mathrm{E}\left(abs\left(\hat{\beta}/\beta\right) \mid P < 0.05\right)$	16.7	5.3	1.8	1.0

Candidate-Gene Study: Design Calculations

(based on Gelman and Carlin 2014)

Suppose R^2 = 0.02% and significant at α = .05:

1. How often will estimate have the right sign?
2. How exaggerated is the magnitude of the estimate?

N	100	1K	10K	100K
Power	0.052	0.073	0.294	0.993
$\mathrm{P}(Right\ Sign \mid P < 0.05)$	66%	89%	99.9%	100%
$\mathrm{E}\left(abs\left(\hat{\beta}/\beta\right) \mid P < 0.05\right)$	16.7	5.3	1.8	1.0

Candidate-Gene Study: Design Calculations

(based on Gelman and Carlin 2014)

Suppose R^2 = 0.02% and significant at α = .05:

1. How often will estimate have the right sign?
2. How exaggerated is the magnitude of the estimate?

N	100	1K	10K	100K
Power	0.052	0.073	0.294	0.993
$\mathrm{P}(Right\ Sign \mid P < 0.05)$	66%	89%	99.9%	100%
$\mathrm{E}\left(abs\left(\hat{\beta}/\beta\right) \mid P < 0.05\right)$	16.7	5.3	1.8	1.0

Bayesian Analysis of a GWAS

(based on Wacholder et al., 2004; Ioannidis, 2005; Benjamin et al., 2012)

Given significant at $\alpha = 5 \times 10^{-8}$, posterior probability of true association with effect size $R^2 = 0.02\%$.

N		100	10K	100K
Power		6.6×10^{-8}	2.7×10^{-5}	0.157
Prior	0.1%	0.13%	36%	100%
	1%	1.3%	85%	100%
	5%	7%	97%	100%
	10%	13%	98%	100%

Bayes' Rule: $\mathrm{P}(True|Sig) = \frac{power \times prior}{power \times prior + 5 \times 10^{-8}\,(1 - prior)}$.

Bayesian Analysis of a GWAS

(based on Wacholder et al., 2004; Ioannidis, 2005; Benjamin et al., 2012)

Given significant at $\alpha = 5 \times 10^{-8}$, posterior probability of true association with effect size $R^2 = 0.02\%$.

N		100	10K	100K
Power		6.6×10^{-8}	2.7×10^{-5}	0.157
Prior	0.1%	0.13%	36%	100%
	1%	1.3%	85%	100%
	5%	7%	97%	100%
	10%	13%	98%	100%

Bayes' Rule: $\mathrm{P}(True|Sig) = \frac{power \times prior}{power \times prior + 5 \times 10^{-8}\,(1 - prior)}$.

Fourth “Law” as a Unifying Principle

- **Small-*N* candidate gene studies of behavioral traits have a weak replication track record.**
- Steady increase in genetic associations identified in GWA studies as $N \uparrow$. Strong replication record.
- Despite the fact that GWAS has had some successes, much of the heritability is “missing”.
- As larger *N* available to estimate weights for polygenic scores, their predictive power rises.

Fourth “Law” as a Unifying Principle

- Small-*N* candidate gene studies of behavioral traits have a weak replication track record.
- **Steady increase in genetic associations identified in GWA studies as $N \uparrow$. Strong replication record.**
- Despite the fact that GWAS has had some successes, much of the heritability is “missing”.
- As larger *N* available to estimate weights for polygenic scores, their predictive power rises.

Fourth “Law” as a Unifying Principle

- Small-*N* candidate gene studies of behavioral traits have a weak replication track record.
- Steady increase in genetic associations identified in GWA studies as $N \uparrow$. Strong replication record.
- **Despite the fact that GWAS has had some successes, much of the heritability is “missing”.**
- As larger *N* available to estimate weights for polygenic scores, their predictive power rises.

Fourth “Law as a Unifying Principle

- Small-*N* candidate gene studies of behavioral traits have a weak replication track record.
- Steady increase in genetic associations identified in GWA studies as $N \uparrow$. Strong replication record.
- Despite the fact that GWAS has had some successes, much of the heritability is “missing”.
- **As larger *N* available to estimate weights for polygenic scores, their predictive power rises.**

Concluding Remarks

1. Substantial progress in years ahead.
   - UKB, Precision Medicine and similar initiatives -> $N \approx 1$M samples available for hundreds of traits.
2. Could advance research in a number of ways:
   - Elucidating biological mechanisms
   - Non-genetic empirical research

     Control variables
     Instrumental variables
   - Better foundation for G × E and prediction

     E.g., older individuals with at-risk cognitive health.

Acknowledgments

Jonathan Beauchamp (Harvard University)
Dan Benjamin (USC)
Christopher Chabris (Union College)
Tõnu Esko (Broad Institute)
Magnus Johannesson (Stockholm School of Economics)
Philipp Koellinger (University of Amsterdam)
David Laibson (Harvard University)
Aysu Okbay (VU Amsterdam)
Niels Rietveld (Rotterdam University)
Patrick Turley (Broad Institute)
Peter Visscher (University of Queensland)

We gratefully acknowledge support from NIH's NIA and OBSSR, NSF, the Ragnar Söderberg Foundation, and the Swedish Research Council.