Gene Discovery I: Power, Candidate Genes

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

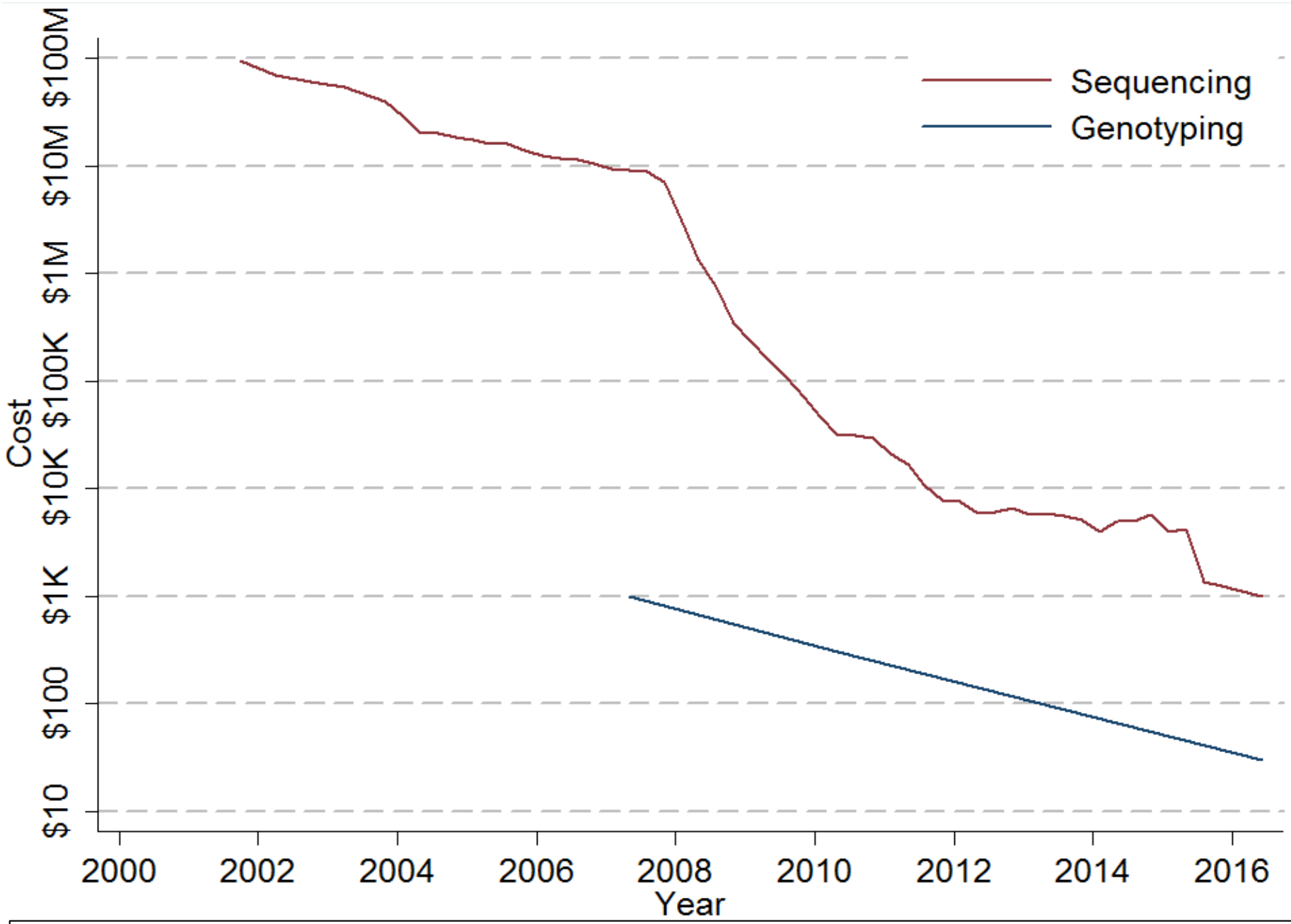
RSF Summer Institute in Social-Science Genomics • 12 June 2019

Outline

1. ***Genetic Data and Datasets***
2. Gene Discovery
3. Traditional Candidate-Gene Studies
4. The Power Problem

Obtaining Genetic Data

1. Obtain tissue sample and send to lab.
   - Blood: requires medical professional, higher DNA concentration (less relevant today).
   - Saliva (more common): simple instructions to spit into tube, can be shipped at room temperature.
2. Extract DNA and make many copies to analyze.
3. Machine reads genotype data from the DNA.
   - Genotyping (more common): Measure a set of genetic variants specified on an "array." Typically 0.5-2.5 million SNPs.
   - Sequencing (will eventually dominate): Measure every base pair, or a random sample of base pairs.



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

Datasets With Genome-Wide Data

1. Medical-genetic "cohorts" or twin registries ($N \approx 500$ to 5,000)
   - Medical cohorts typically formed to study particular conditions, e.g., cardiovascular disease, cancer.
   - To access data, generally must collaborate with cohort.
   - Ex.: Framingham Heart Study, Swedish Twin Registry.
2. Social-science surveys ($N \approx 5{,}000$ to 20,000)
   - Several recently genotyped participants.
   - Data publicly available subject to IRB/privacy procedures to access the genetic data, usually via dbGaP.
   - Ex.: Health and Retirement Study, Add Health.

3. National biobanks (*N* ≈ 50,000 to 500,000)
   - Initiatives underway in many countries.
   - Data usually available to qualified researchers via application process.
   - Ex.: UK Biobank, Estonian Biobank.
4. Personal genomics co.'s (*N* ≈ 50,000 to 500,000)
   - Customers consent to participate in research.
   - To access data, generally must collaborate with company, though application process for qualified researchers may be available.
   - Ex.: 23andMe, DeCode Genetics.

Outline

1. Genetic Data and Datasets
2. ***Gene Discovery***
3. Traditional Candidate-Gene Studies
4. The Power Problem

Gene Discovery

Heritability estimates do not tell us which genetic variants matter, how much, or why.

Nor do they allow us to construct predictive variables from genetic data.

"Gene discovery": identifying associations between genetic variants and a phenotype.

Necessary step before knowledge can be used for doing social-science research.

Gene-Discovery Causal Model

Focus on estimating additive model to reduce number of parameters:

$$\tilde{y}_i = A(\boldsymbol{x}_i) + \boldsymbol{z}_i\boldsymbol{\gamma} + \epsilon_i = \sum_{j=1}^{J} \beta_j x_{ij} + \boldsymbol{z}_i\boldsymbol{\gamma} + \epsilon_i.$$

Goal: Estimate the average causal effects β_j's.

Potential concern: Might miss variants that matter via interactions if the interactions average to zero.

- Common view: Better to reduce power/MHT concerns first by focusing on additive effects, then find interactions among those variants at a later stage.

The Three Big Problems of Gene Discovery

1. Multiple hypothesis testing (MHT)
   - Typically many measured loci in the dataset.
   - Typically several phenotypes.
   - Many possible specifications/controls.
2. Population stratification
   - Is $\boldsymbol{z}_i$ sufficient to eliminate potential confounds?
3. Low Power
   - Is N large enough, given true effect size β_j and significance threshold (and prior)?

Outline

1. Genetic Data and Datasets
2. Gene Discovery
3. ***Traditional Candidate-Gene Studies***
4. The Power Problem

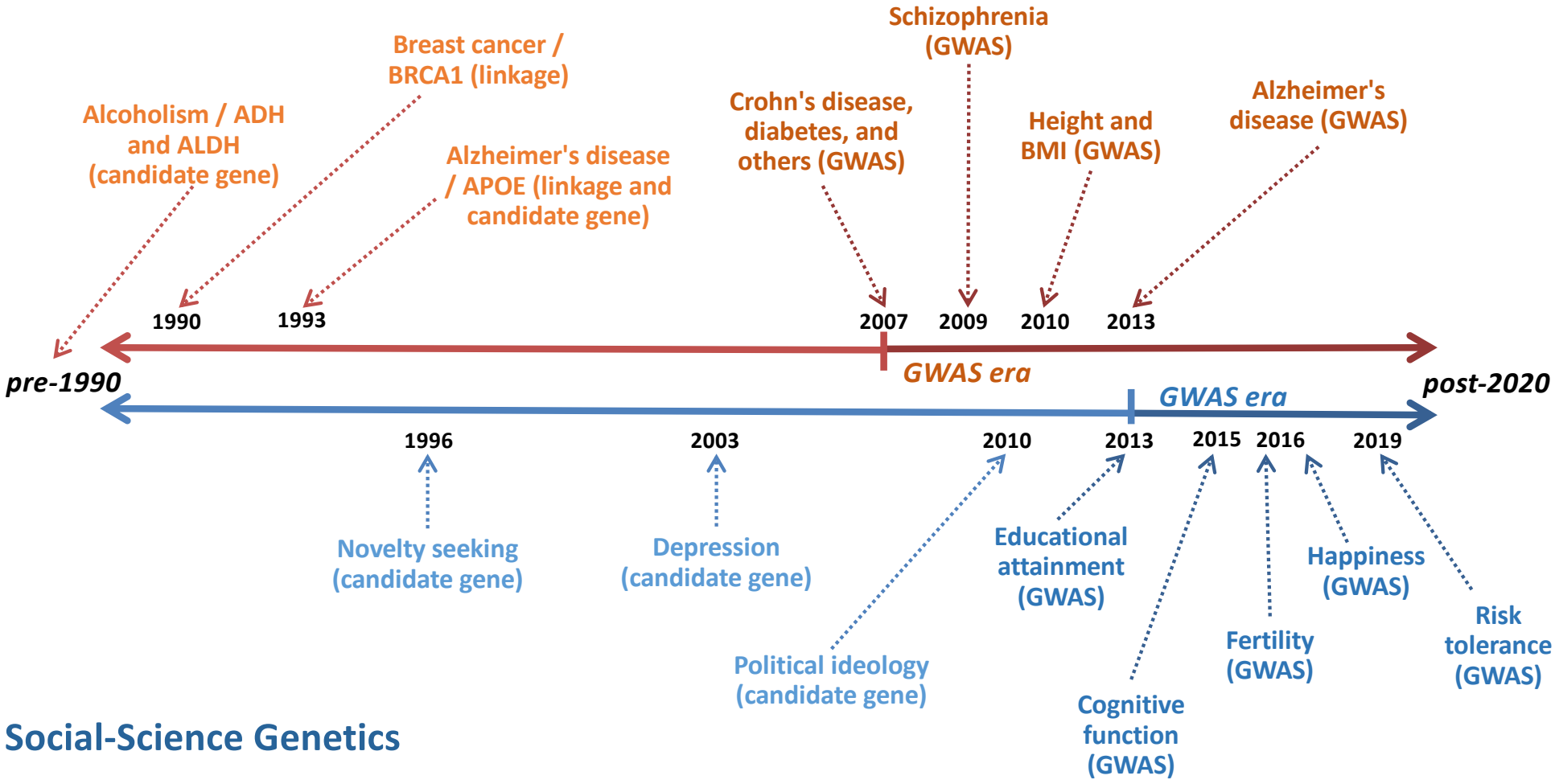
Traditional Candidate-Gene Study

- Specify *ex ante* hypotheses about small set of $K \ll J$ SNPs (often $K = 1$) based on believed biological function.
- Estimate

$$\tilde{y}_i = \beta_k x_{ik} + \mathbf{z}_i' \boldsymbol{\gamma}_k + \eta_{ik}.$$

- Set significance threshold α = .05 / *K*.
- Virtually all existing work in social-science genetics. (Reviews: Ebstein, Israel, Chew, Zhong, and Knafo, 2010; Beauchamp et al., 2011; Benjamin et al., 2012)

Medical Genetics



Social-Science Genetics

- Eminently reasonable, and has worked when hypotheses are direct. (e.g., *APOE* and Alzheimer's disease)
- But in social-science genetics, replication record has been inconsistent.
  - My own Icelandic saga.
  - Example: candidate genes for cognitive function.

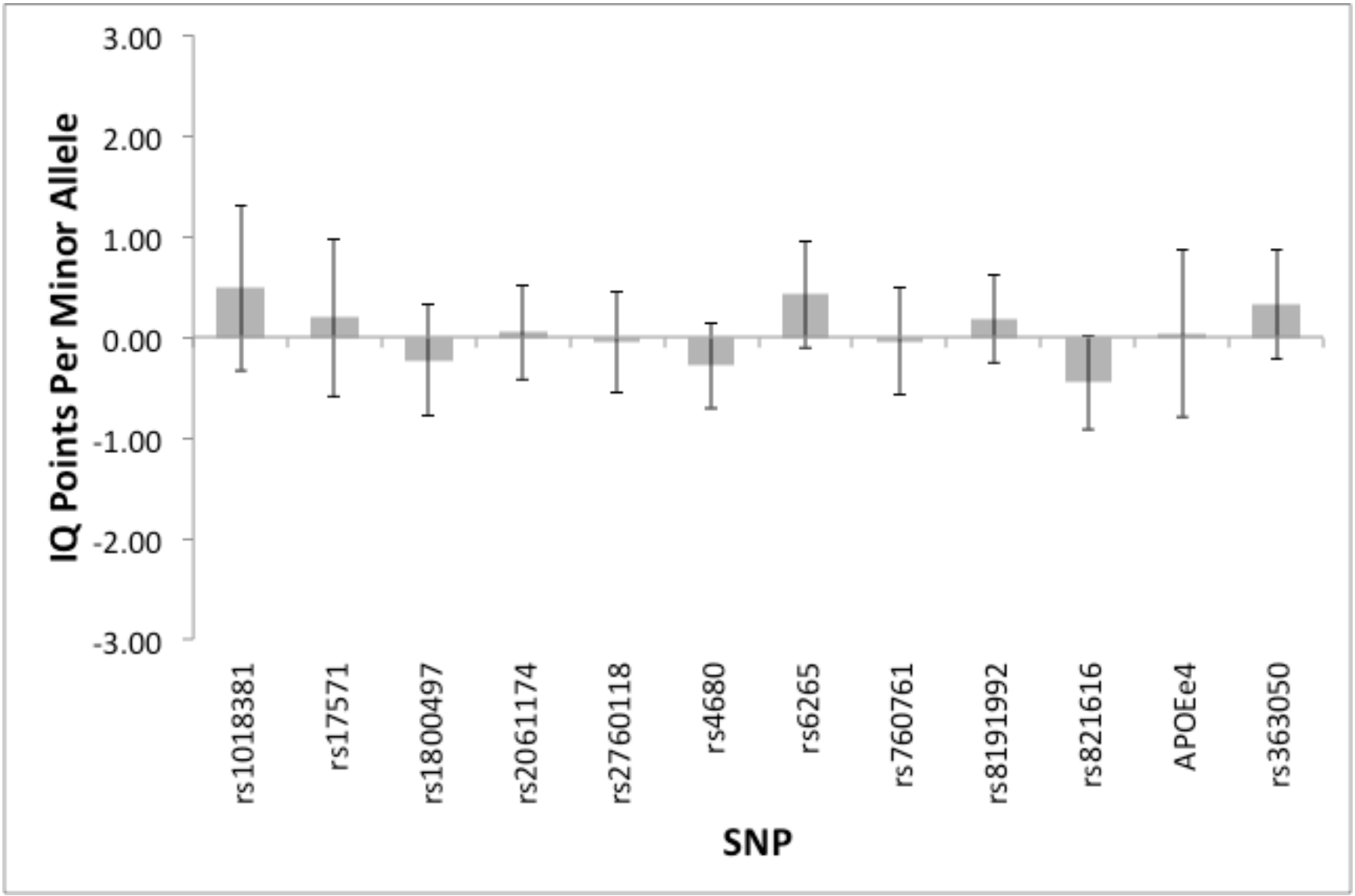
Most Reported Genetic Associations With General Intelligence Are Probably False Positives

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]

Psychological Science
23(11) 1314–1323

DOI: 10.1177/095679
http://pss.sagepub.com

Pooled estimates (11 SNPs + *APOE*)



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).

- Eminently reasonable, and has worked when hypotheses are direct. (e.g., *APOE*/Alzheimer's)
- But in social-science genetics, replication record has been inconsistent.
  - My own Icelandic saga.
  - Example: candidate genes for cognitive function.
- Problem with premise: assumed a few genes have large effects (detectable with $N \approx 100$ to 3,000).
  - But we now know that effects of *any* particular genetic variant likely to be very small.
  - Having plausible hypotheses lent spurious credibility to statistically suspicious findings.

(Not) Addressing the Problems

1. Multiple hypothesis testing (MHT)
   - Seemingly common and usually uncorrected-for.
   - (A problem more generally in social science.)
2. Population stratification
   - Typically no genome-wide (or family) data.
   - Sometimes restrict to Whites or control for self-reported ethnicity, but may not be sufficient.
3. Low power

Calibration: Power Analysis

Suppose:

- $\alpha = 0.05$
- $R^2 \approx 0.02\%$
- $N = 3{,}000$

What is power? 12%.

Sample size for 80% power? 39,150.

Outline

1. Genetic Data and Datasets
2. Gene Discovery
3. Traditional Candidate-Gene Studies
4. ***The Power Problem***

Mistakes About Power

- Recall that power = $\Pr(\text{sig}|H_1)$, calculated using the true effect size β_j.
  - When estimate power, need to use *ex ante* plausible effect size.
  - Do not use *ex post* $\hat{\beta}_j$: if turns out to be stat. sig. by chance, power will appear to have been large.
- Misconceptions:
  - Power matters for whether I'll find an effect when it's true, but isn't relevant any more once I've found an effect.
  - If I find an effect despite having had low power, then my result is even more convincing.

The Problem With Low Power

- Low power essentially equivalent to: true $\mathrm{SE}(\hat{\beta}_j)$ is large relative to $\beta_j \neq 0$.
  - If $\mathrm{SE}(\hat{\beta}_j)$ were small, then $\hat{\beta}_j$ would likely turn out to be close to β_j.
  - And the confidence interval around $\hat{\beta}_j$ would likely exclude zero.
  - But that means the estimate would likely be significant—i.e., power is high.
- $\mathrm{SE}(\hat{\beta}_j)$ large relative to $\beta_j \leftrightarrow \hat{\beta}_j$ uninformative.
  - Can't conclude much from failing to reject H_0.
  - *Also* can't conclude much from rejecting H_0.

Manifestation #1: Sign and Magnitude Errors

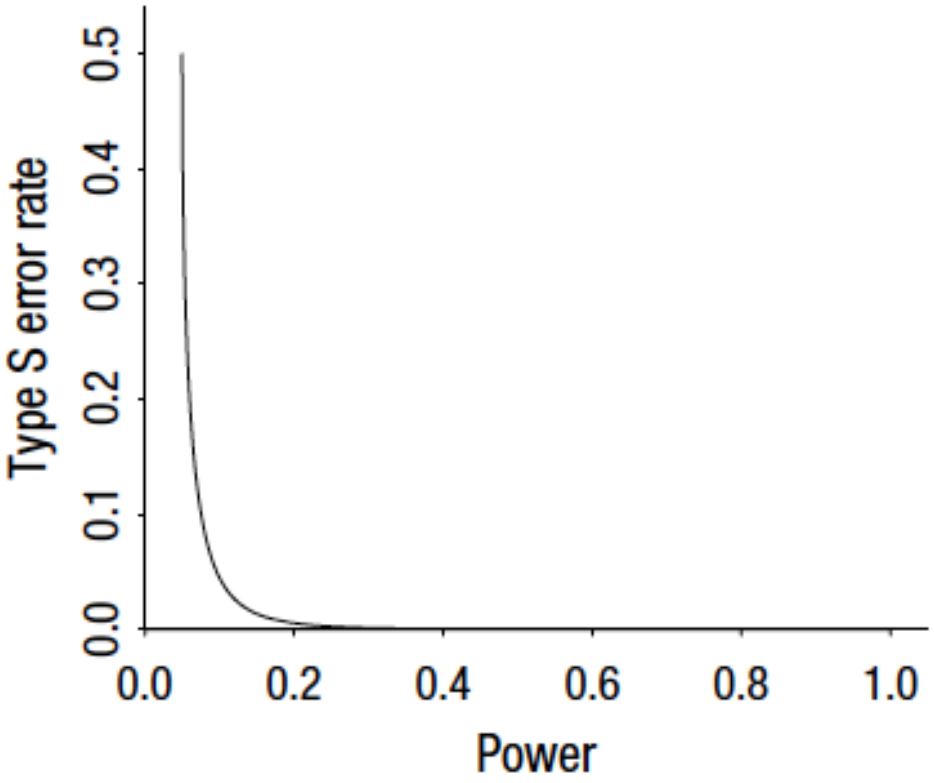
(Gelman and Carlin, 2014)

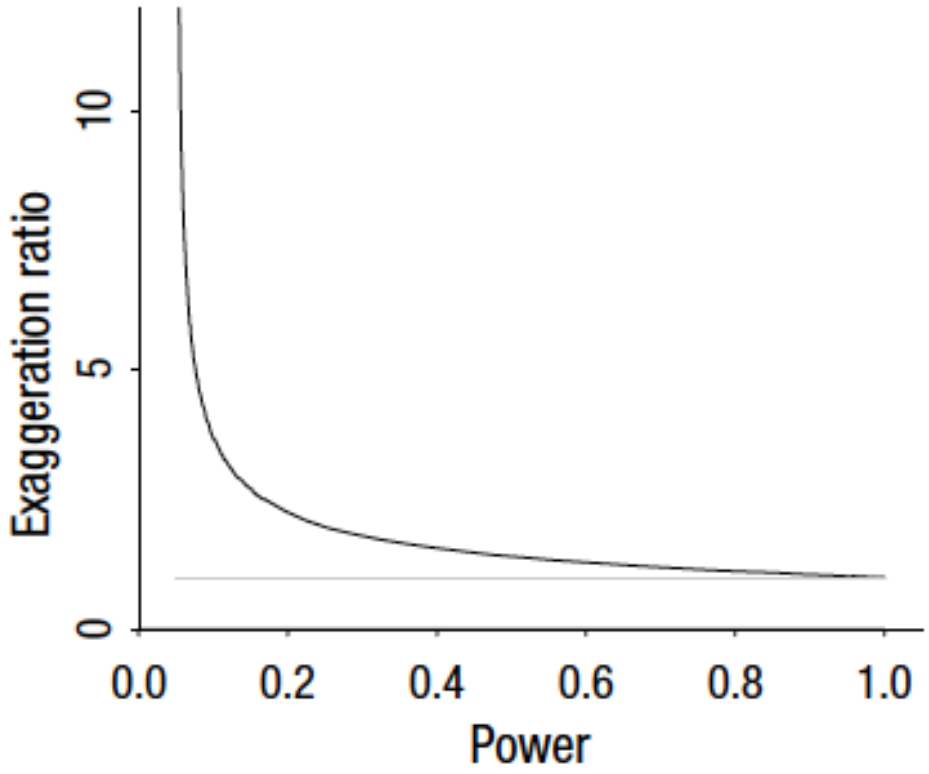
When $\hat{\beta}_j$ is statistically significant:

- "Type S error": probability that $\hat{\beta}_j$ has the incorrect sign.
- "exp. Type M error" (exaggeration ratio): $\frac{\mathrm{E}|\widehat{\beta}_j|}{\mathrm{E}|\beta_j|}$.

When power is lower, both are larger.

(Exaggeration ratio related to winner's curse.)





Source: Gelman and Carlin (2014). Assumes unbiased estimate that is normally distributed.

Manifestation #2: High Risk of "False Positive"

(based on Wacholder et al., 2004; Ioannidis, 2005; WTCCC, 2006; Benjamin et al., 2012; Bayarri et al., 2016)

Power is about $\Pr(\text{sig}|H_1)$.

But what we really want to know is $\Pr(H_1|\text{sig})$. That is, is the hit a true finding?

The formula for how to switch the order of conditioning is Bayes' Rule.

Reminder: Bayes' Rule

For any two events A and B,

$$\Pr(A|B) \equiv \frac{\Pr(A \cap B)}{\Pr(B)}$$

$$= \frac{\Pr(A \cap B)}{\Pr(B \cap A) + \Pr(B \cap \sim A)}$$

$$= \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\sim A)\Pr(\sim A)}.$$

Assumptions

- There are two states of the world:
  - H_1: A true effect of size $\beta_j > 0$.
  - H_0: Zero effect $\beta_j = 0$.
  - Useful for illustrative purposes, but realistically should specify anticipated effect size *distribution*.

- The information we have available is whether $\hat{\beta}_j$ is significant or not.
  - If already run the study, better to condition on the data (e.g., value of $\hat{\beta}_j$, or *p*-value).
  - Can think of the analysis we will do as "pre-experimental." (Bayarri et al., 2016)

$$\Pr(H_1|\text{sig})$$

$$= \frac{\Pr(\text{sig}|H_1)\Pr(H_1)}{\Pr(\text{sig}|H_1)\Pr(H_1) + \Pr(\text{sig}|H_0)\Pr(H_0)}$$

$$= \frac{\frac{\Pr(\text{sig}|H_1)}{\Pr(\text{sig}|H_0)}\frac{\Pr(H_1)}{\Pr(H_0)}}{\frac{\Pr(\text{sig}|H_1)}{\Pr(\text{sig}|H_0)}\frac{\Pr(H_1)}{\Pr(H_0)} + 1}.$$

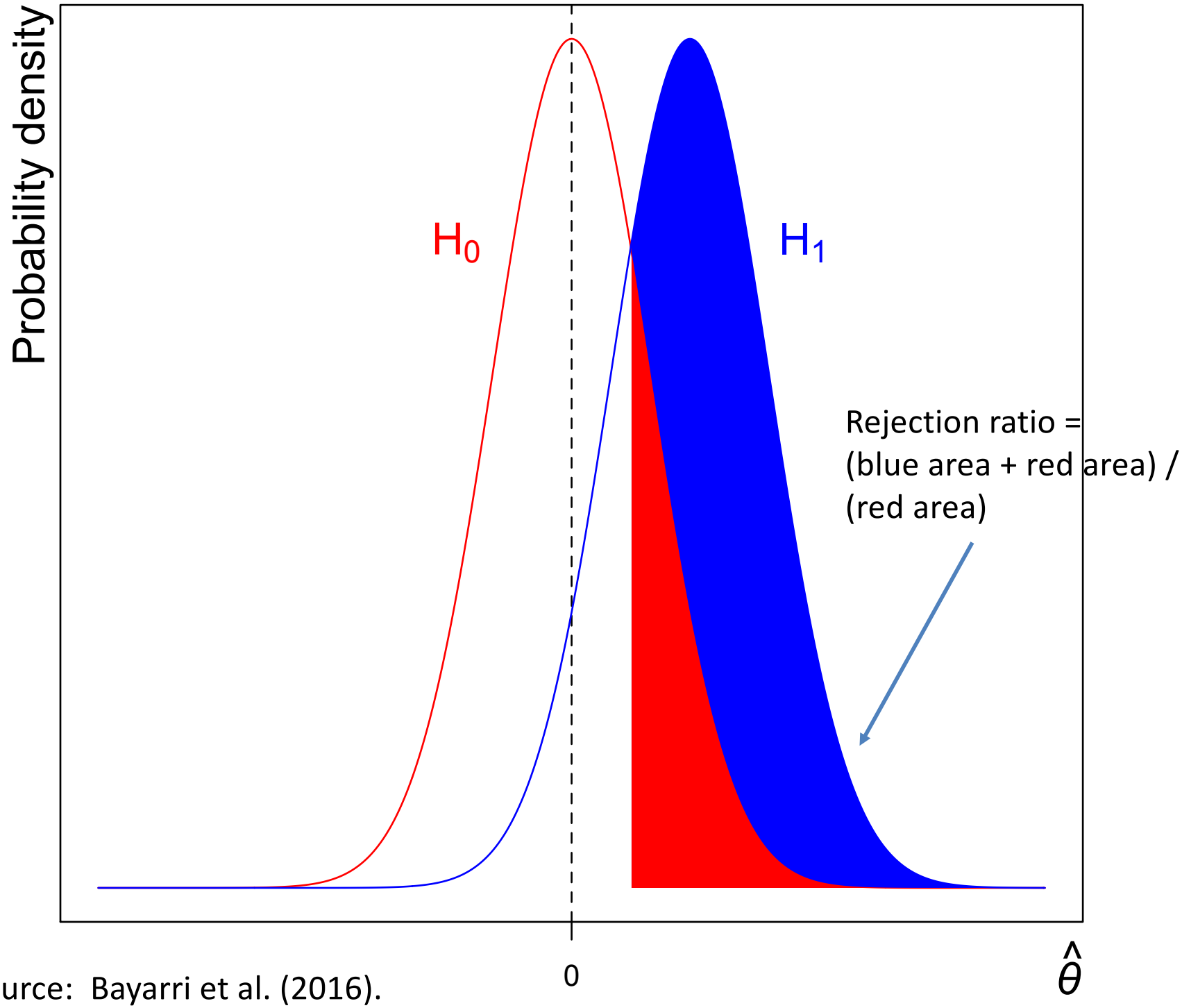
We can similarly derive $\Pr(H_0|\text{sig})$.

Dividing $\Pr(H_1|\text{sig})$ by $\Pr(H_0|\text{sig})$ gives a formula that is easier to remember:

$$\underbrace{\frac{\Pr(H_1|\text{sig})}{\Pr(H_0|\text{sig})}}_{\text{Posterior ratio}} = \underbrace{\frac{\Pr(\text{sig}|H_1)}{\Pr(\text{sig}|H_0)}}_{\text{“Rejection ratio”}} \underbrace{\frac{\Pr(H_1)}{\Pr(H_0)}}_{\text{Prior ratio}}.$$

Therefore:

$$(\text{Posterior ratio}) = \underbrace{\left(\frac{\text{power}}{\alpha}\right)}_{\text{Rejection ratio is strength of evidence from rejecting } H_0} (\text{prior ratio}).$$



Source: Bayarri et al. (2016).

Application: Simple Experiment

- Treatment and control group, each with sample size *N*.
- Effect size $r = 0.21$, "typical" according to meta-analysis of studies in social psychology. (Richard, Bond, and Stoke-Zoota, 2003)

Per-condition *N*	10	20	30	40	50
Power	0.25	0.39	0.51	0.61	0.69
Rejection ratio	4.9	7.7	10.2	12.2	13.9

Per-condition *N*	65	75	100	150	200
Power	0.79	0.84	0.92	0.98	1.00
Rejection ratio	15.8	16.8	18.4	19.6	19.9

Source: Bayarri et al. (2016).

Application: Gene-Discovery for Educational Attainment

How to choose priors?

Often useful to consider a reasonable range.

Starting points:

- GREML estimate: variance explained by additive effects of all independent loci is roughly 20%.
- There are roughly 1 million independent loci (in European-descent populations).

One extreme: Each associated locus has effect size R^2 = 0.02%.

Then there are 20% / 0.02% = 1,000 associated loci in total.

Thus each locus in the genome has prior probability 1,000 / 1 million = 0.1%.

This is a lower bound on the prior: Most loci surely have smaller effects than R^2 = 0.02%, hence there are more than 1,000 loci.

The other extreme is trickier: the H_0 vs. H_1 setup assumes all loci have either of two particular effect sizes, R^2 = 0% or 0.02%.

Somewhat arbitrarily, let's assume effect size <0.002% is the moral equivalent of "null."

Then the other extreme: Each associated locus has effect size R^2 = 0.002%.

Hence there are 20% / 0.002% = 10,000 associated loci in total.

Thus each locus in the genome has prior probability 10,000 / 1 million = 1%.

Pre-Experimental Odds

(based on Wacholder et al., 2004; Ioannidis, 2005; WTCCC, 2006; Benjamin et al., 2012; Bayarri et al., 2016)

Given significant at α = .05, assuming effect size R^2 = 0.02%.

		Sample size		
		N = 100 (power = .05)	*N* = 10,000 (power = .29)	*N* = 100,000 (power = .99)
Prior prob-ability	0.1%	0.1%	0.6%	2%
	1%	1%	6%	17%
	5%	5.2%	24%	51%
	10%	10.4%	39%	69%

These calculations make clear that in a world of tiny R^2, we need large N.

But they also provide another perspective (besides correcting for multiple-hypothesis testing in GWAS) for why we need stringent significance thresholds—not only for GWAS, but also for candidate-gene studies.

$$\text{Posterior ratio} = \left(\frac{\text{power}}{\alpha}\right)(\text{Prior ratio}).$$

We've said prior ratio of $\frac{0.01}{0.99} \approx 0.01$ is optimistic.

Power is bounded above by 1 →

At $\alpha = 0.05$, $\left(\frac{\text{power}}{\alpha}\right)$ is bounded above by $\frac{1}{0.05} = 20$.

Pre-experimental odds ≤ $(20)(0.01) = 0.2$!

In a world of small priors, need stringent significance thresholds for findings to be credible. ("Extraordinary claims require extraordinary evidence." -Carl Sagan)

At $\alpha = 5\times10^{-8}$, $\left(\frac{\text{power}}{\alpha}\right)$ is bounded above by $\frac{1}{5\times10^{-8}} = 2\times10^{7}$.

Hence a hit at genome-wide significance from a well-powered study is credible even with pessimistic priors.

Now at Genome-Wide Significance…

(based on Wacholder et al., 2004; Ioannidis, 2005; WTCCC, 2006; Benjamin et al., 2012; Bayarri et al., 2016)

Given significant at $\boldsymbol{\alpha = 5 \times 10^{-8}}$, assuming effect size R^2 = 0.02%.

		Sample size		
		N = 100 (power = .00)	*N* = 10,000 (power = .00)	*N* = 100,000 (power = .16)
Prior prob-ability	0.1%	0.13%	36%	100%
	1%	1.3%	85%	100%
	5%	7%	97%	100%
	10%	13%	98%	100%

Other important insights: At $\alpha = 5\times10^{-8}$, need larger *N*… *but* 16% power is perfectly adequate!

$$\left(\frac{\text{power}}{\alpha}\right) = \frac{0.16}{5\times10^{-8}} = 3{,}200{,}000,$$

which is strong evidence!

Bottom line: Do power/credibility calculations!

Our usual intuitions are often off in a world of low priors, small effects, and genome-wide significance.