Gene Discovery II: GWAS

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 • 13 June 2019

Outline

1. ***Genome-Wide Association Studies (GWAS)***
2. Example: Educational Attainment

GWAS

- Atheoretical testing of all $K \ll J$ SNPs measured on the chip ($K \approx$ 0.5-2.5 million) or imputed ($K \approx$ 9 million).
- Genome-wide significance: $\alpha = 5 \times 10^{-8}$.
  - Frequentist justification: Bonferroni correction for ~1 million effectively indep. loci in Europeans.
  - Bayesian justification: Need stringent significance threshold given low prior for any specific locus.
- Causal model:

$$\tilde{y}_i = \sum_{k=1}^{K} \beta_k x_{ik} + \mathbf{z}_i' \boldsymbol{\gamma} + \eta_i.$$

The Dimensionality Problem

Cannot estimate this regression:

$$\tilde{y}_i = \sum_{k=1}^{K} \beta_k x_{ik} + \boldsymbol{z}_i' \boldsymbol{\gamma} + \eta_i$$

unless $N > K + |\boldsymbol{\gamma}| \approx 9$ million!

Standard sol'n: estimate univariate regressions:

$$\tilde{y}_i = \beta_k^{GWAS} x_{ik} + \boldsymbol{z}_i' \boldsymbol{\gamma}_k + \epsilon_{ik}.$$

Why Not Machine Learning?

To date, two main reasons:

1. Well-powered GWAS result from meta-analyses of many samples.
   - Not individual-level data, due to IRB/privacy.
2. Computational constraints.

But these barriers are lifting:

1. Increasingly large samples directly accessible (e.g., UK Biobank).
2. Secure cloud computing.

Interpreting GWAS Coefficients

Causal model: $\tilde{y}_i = \sum_{k=1}^{K} \beta_k x_{ik} + \boldsymbol{z}_i' \boldsymbol{\gamma} + \eta_i$

Estimation: $\tilde{y}_i = \phi_k + \beta_k^{GWAS} x_{ik} + \boldsymbol{z}_i' \boldsymbol{\gamma}_k + \epsilon_{ik}.$

Key point: $\beta_k^{GWAS} \neq \beta_k.$

β_k^{GWAS} is β_k plus effects of SNPs in LD with k.

GWAS Hits

Because of LD, each genome-wide significant SNP is typically correlated with others.

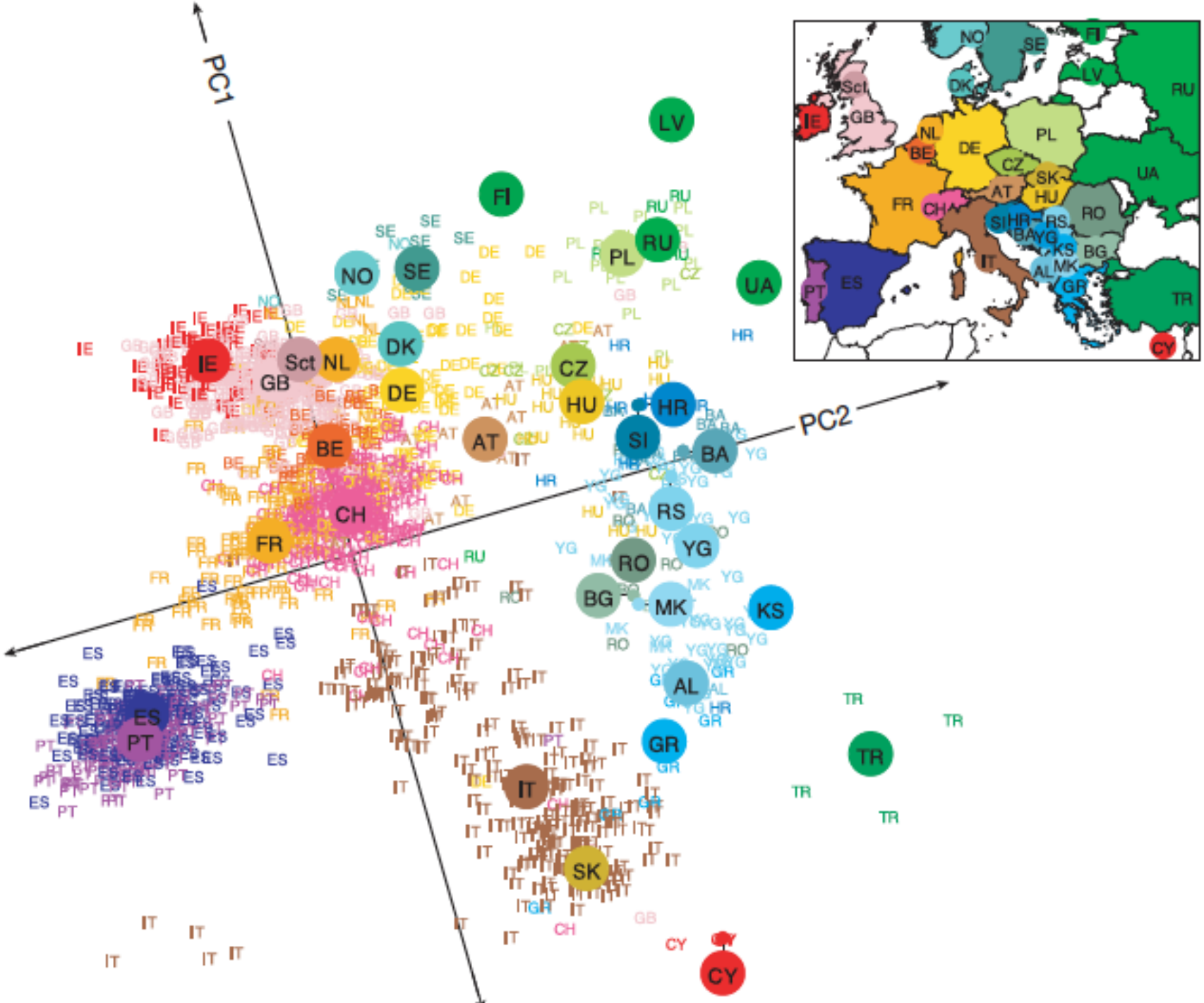
Each block of correlated SNPs is called a genome-wide significant *locus*.

Lead SNP: The SNP in a genome-wide significant locus with the smallest *p*-value.

By construction, the set of lead SNPs are therefore approximately uncorrelated with each other.

Addressing the Problems

1. Multiple hypothesis testing (MHT)
   - *All* measured SNPs are tested.
   - Genome-wide significance threshold addresses.
   - (But with more large-scale GWAS data available, MHT may occur with phenotypes.)
2. Population stratification
   - Can use genome-wide data to estimate PCs.
   - Then control for PCs. (Price et al., 2006)
   - (But bigger problem for tiny effects.)
3. Low power
   - Problem addressed by GWAS consortia.



Source: Novembre et al. (2008).

Examples of Large-Scale GWAS

- Height (Wood et al., 2014)
- BMI (Locke et al., 2015)
- Schizophrenia (Ripke et al., 2014)
- Smoking (Liu et al., 2019)
- SWB / Depr. / Neuroticism (Okbay et al., 2016)
- Educational attainment (Lee et al., 2018)
- Fertility (Barban et al., 2016)
- Risk tolerance (Linner et al., 2019)

Some Uses of GWAS Results

- Follow-up work to identify causal SNP and causal gene.
- Bioinformatics to examine differential expression by tissue, cell type, etc.
- Use GWAS results as weights to create a polygenic score in other datasets.
- Combine with other GWAS results to:
  - Calculate genetic correlation.
  - Discover novel associations with a genetically correlated phenotype.

Candidate-Gene Studies, Revisited

Candidate-gene studies have been problematic in practice but not in principle:

1. Multiple hypothesis testing: Could pre-register analyses and use stringent significance threshold.
2. Population stratification: In datasets with genome-wide data, could control for PCs.
3. Low power: Could use large samples.

Could choose "empirical candidates" from well-powered GWAS (not indirect hypotheses).

- "Proxy-phenotype approach" (Rietveld et al., 2014).

Outline

1. Genome-Wide Association Studies (GWAS)
2. ***Example: Educational Attainment***

Welcome to the Social Science Genetic Association Consortium (SSGAC).

The SSGAC is a cooperative enterprise among medical researchers and social scientists that coordinates genetic association studies for social science outcomes and provides a platform for interdisciplinary collaboration and cross-fertilization of ideas. The SSGAC also tries to promote the collection of harmonized and well-measured phenotypes.

Social Science Genetic Association Consortium

Click here to learn about our upcoming training sessions in Genome-Wide Data Analysis!

Click here to learn about the 2017 Polygenic Prediction and its Application in Social Science Conference

Current Initiatives

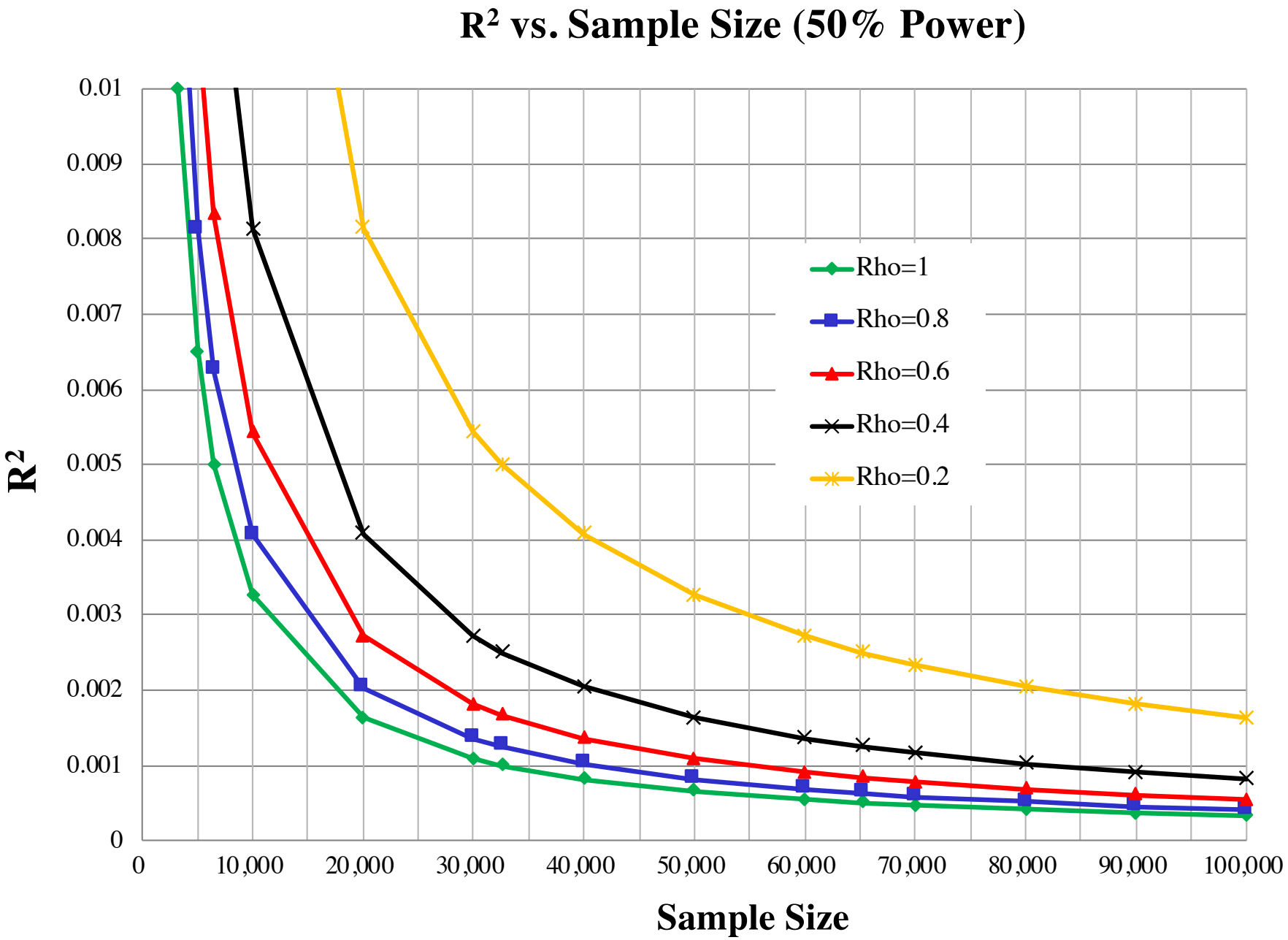
Data

SSGAC in the News

The SSGAC

- In 2011, David Cesarini, Philipp Koellinger, and I founded the Social Science Genetic Association Consortium (SSGAC).
- What outcomes to study?
  - In tradeoff between larger sample and higher-quality measure, given plausible effect sizes, larger sample gives more power. (Chabris et al., 2013)

Quality of Measure vs. Sample Size

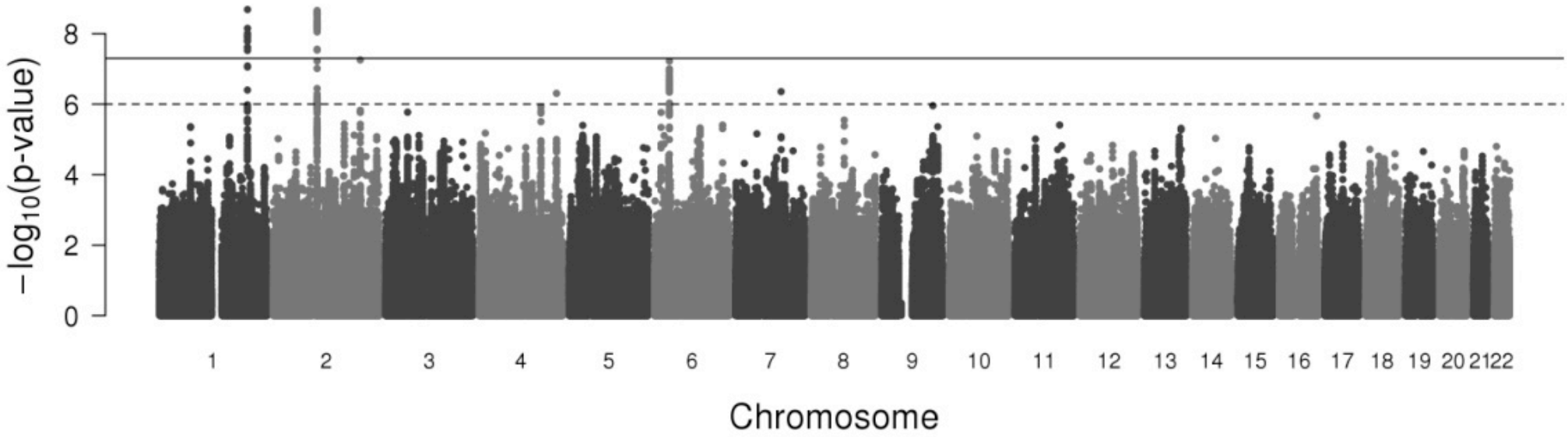


The SSGAC

- In 2011, David Cesarini, Philipp Koellinger, and I founded the Social Science Genetic Association Consortium (SSGAC).
- What outcomes to study?
  - In tradeoff between larger sample and higher-quality measure, given plausible effect sizes, larger sample gives more power. (Chabris et al., 2013)
- Proof-of-concept: education attainment, available in many medical datasets.

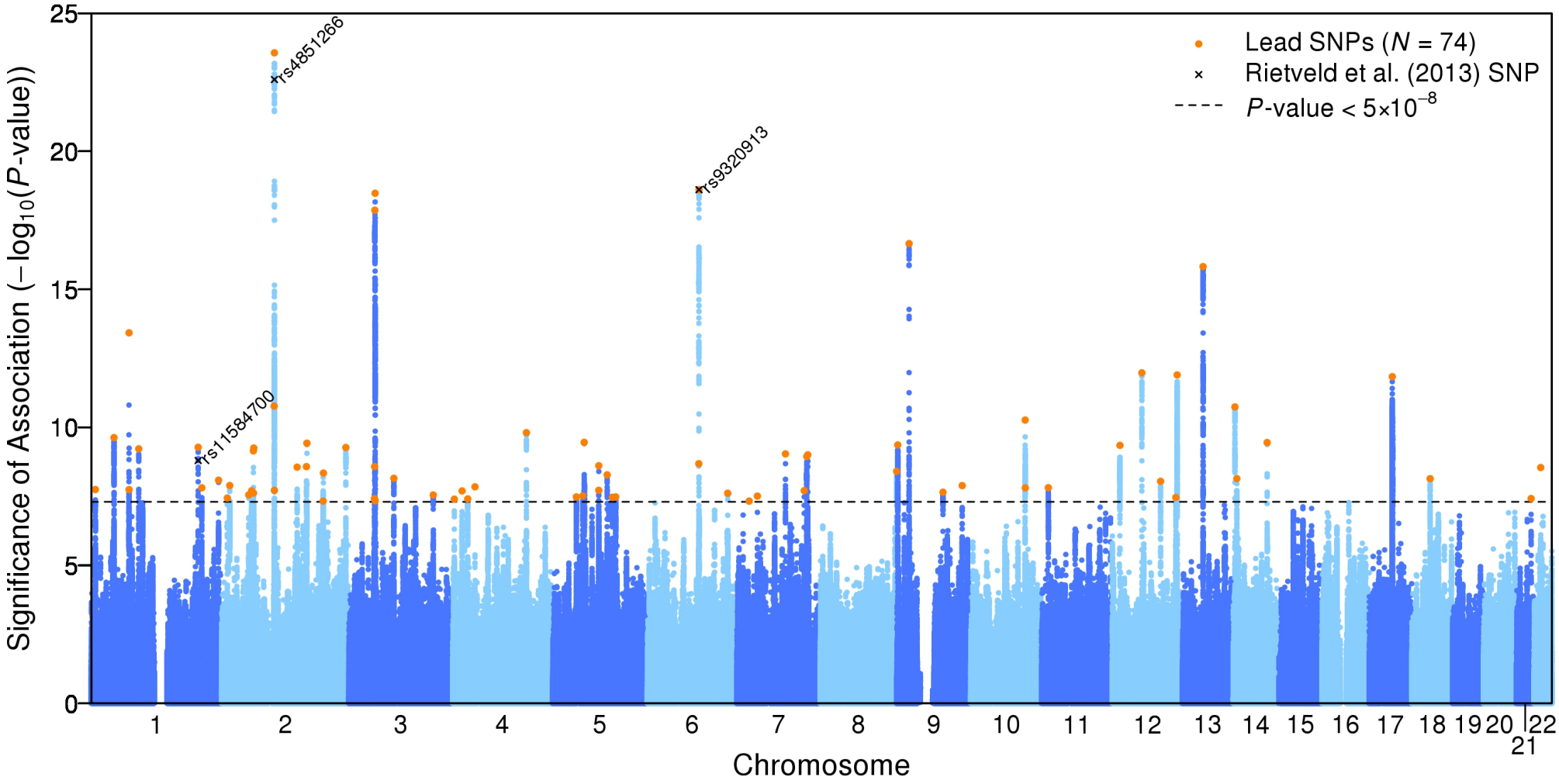
Discovery: N = 101,069 individuals (41 datasets).
Replication: N = 25,490 individuals (12 datasets).

Controls: age, sex, genome-wide PCs.

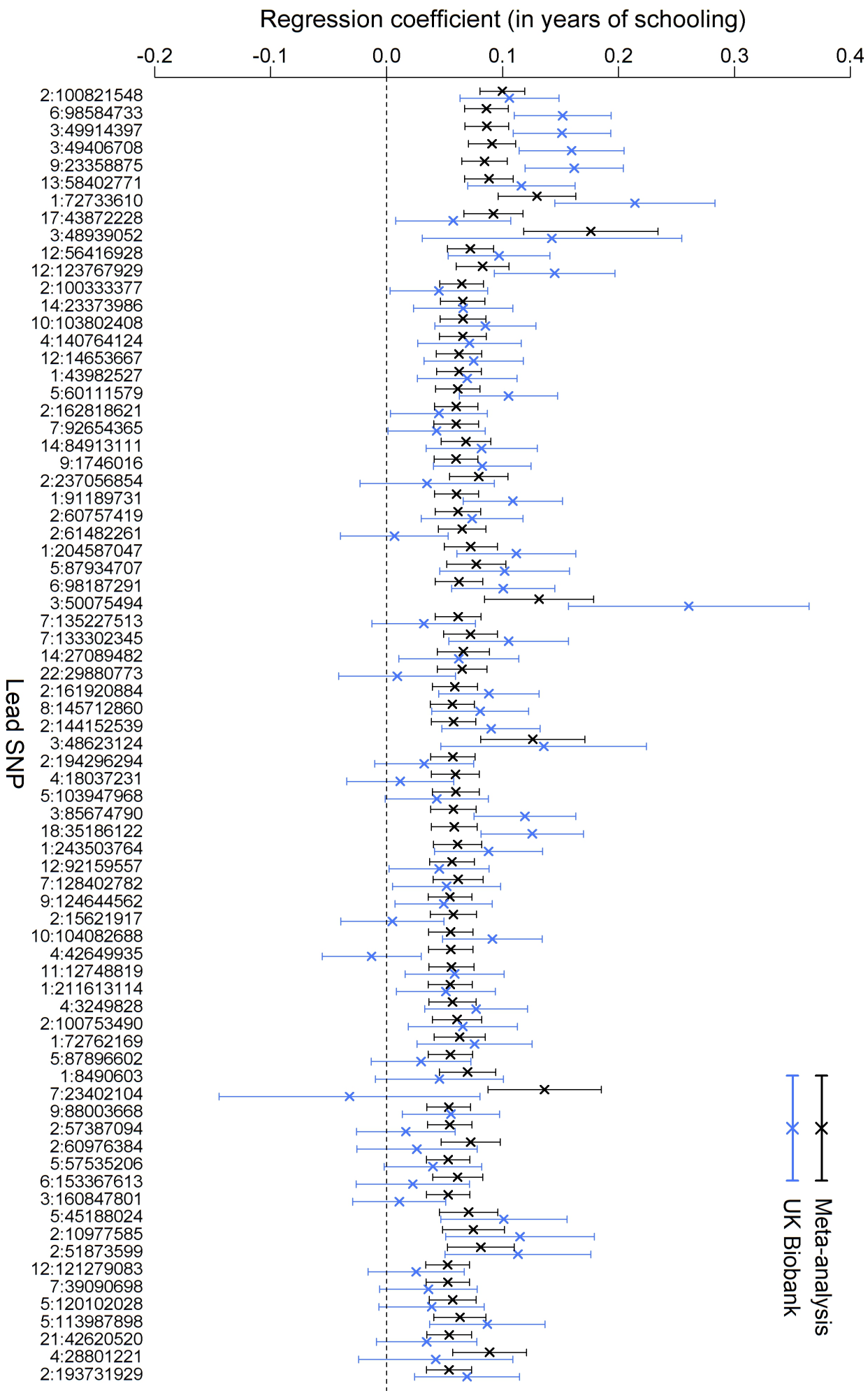


Source: Rietveld et al. (2013, *Science*).

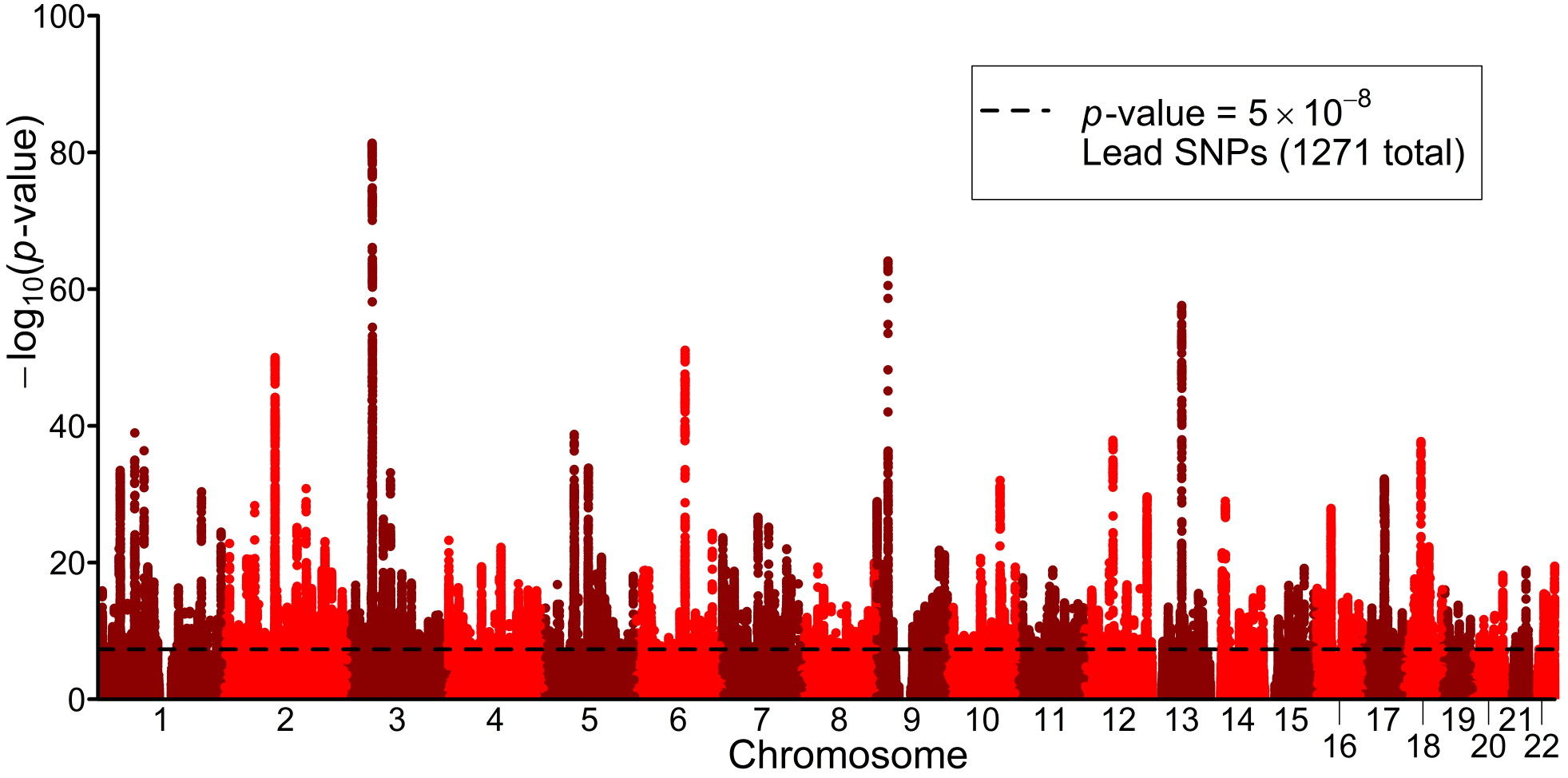
Discovery: N = 293,723 individuals (63 datasets).
Replication: N = 111,349 individuals (UK Biobank 1st release).



Source: Okbay et al. (2016, *Nature*).


Regression coefficient (in years of schooling)
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
Lead SNP
2:100821548
6:98584733
3:49914397
3:49406708
9:23358875
13:58402771
1:72733610
17:43872228
3:48939052
12:56416928
12:123767929
2:100333377
14:23373986
10:103802408
4:140764124
12:14653667
1:43982527
5:60111579
2:162818621
7:92654365
14:84913111
9:1746016
2:237056854
1:91189731
2:60757419
2:61482261
1:204587047
5:87934707
6:98187291
3:50075494
7:135227513
7:133302345
14:27089482
22:29880773
2:161920884
8:145712860
2:144152539
3:48623124
2:194296294
4:18037231
5:103947968
3:85674790
18:35186122
1:243503764
12:92159557
7:128402782
9:124644562
2:15621917
10:104082688
4:42649935
11:12748819
1:211613114
4:3249828
2:100753490
1:72762169
5:87896602
1:8490603
7:23402104
9:88003668
2:57387094
2:60976384
5:57535206
6:153367613
3:160847801
5:45188024
2:10977585
2:51873599
12:121279083
7:39090698
5:120102028
5:113987898
21:42620520
4:28801221
2:193731929
Meta-analysis
UK Biobank

Discovery: N = 1,131,881 individuals (70 datasets).
(Replication of Okbay et al. (N = 405,073) in new data (N = 726,808), and vice-versa.)



Source: Lee et al. (2018, *Nature Genetics*).

Takeaways for Your Research

- Beware which gene-discovery research you rely on.
- In general, individual loci have small effects on behavioral phenotypes.
  - "4th Law of Behavior Genetics." (Chabris et al., 2015)
- Best chance of adequate power to focus on:
  - Well-established genetic variants with relatively large effects (e.g., FTO on BMI, Mr. Big on smoking, APOE on Alzheimer's).
  - Polygenic scores (will discuss on Friday).
- Do power calculations!