

Economics & Genetics

Lecture 4

DR. NIELS RIETVELD, NRIETVELD@ESE.EUR.NL

OFFICE MANDEVILLE T18-29

GCTA output

What is most important to report?

- h_{SNPs}^2 (with its standard error)
- p -value
- Sample size

1	Source	Variance	SE
2	V(G)	3.467299	0.644886
3	V(e)	2.395653	0.607410
4	Vp	5.862952	0.188332
5	V(G)/Vp	0.591391	0.105080
6	logL	-2755.555	
7	logL0	-2770.189	
8	LRT	29.268	
9	df	1	
10	Pval	3.1507e-08	
11	n	2000	
12			

Today's agenda

Main questions:

- How can we identify the genes that influence behavior?
- Why is the “gene for X” story flawed?

Literature:

- Nicolaou et al. (2011), “A polymorphism associated with entrepreneurship: Evidence from dopamine receptor candidate genes”
- Rietveld et al. (2013), “GWAS of 126,559 individuals identifies genetic variants associated with educational attainment”

HEGAB MULDER | SCIENCE | 07-31-2018 07:00 AM

Are Diplomas in Your DNA?

Last week researchers announced more than 1,000 genetic variants associated with how far a person gets through school—along with warnings for how *not* to use that data. But earlier results from the same group are already available in a consumer product.



DESIGNS | FEATURED

QUARTZ

EMAILS | EDITORIALS | DEC


ANCESTRYHEALTH

Ancestry's genetic tests can now tell you about your health

By Katherine Ellen Foley • October 15, 2019

The difficulty to find SNP-outcome associations
The tiny effects of SNPs

Two methodologies and their pitfalls

1. Candidate gene studies and why they don't work
 - Pitfalls: Illusive theoretical rigor & insufficient “statistical power”
 2. Genome-wide association studies (GWASs) and why they do work
 - Pitfalls: Multiple testing & “population stratification”
- 

Economics & Genetics Lecture 4

CANDIDATE GENE STUDIES



“Working horse” model from last week

y_i is the value of the outcome variable for individual i

μ is the intercept (constant)

β_j is the effect of SNP j (assumption of additivity, every allele has the same effect)

x_{ij} reflects the number of reference alleles for a SNP (0, 1, 2)

ε_i is the effect of exogenous residual factors

$i \in [1, \dots, N]$, N usually 5K-10K in a dataset

$j \in [1, \dots, J]$, $J > 1,000,000$ [Overidentification!]

$$y_i = \mu + \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i$$

How to deal with overidentification, i.e. $J > N$? (regular regression is not going to work)

- Candidate gene study: Do not consider all J SNPs
- *Genome-wide Association study: Consider all J SNPs, but do not consider them all in one model (one model per SNP)*

Candidate gene studies

Main idea: Do not consider all SNPs (“genes”), but make a clever selection

- *In the past, selection also result of using small (relatively cheap) genotyping arrays*

Choice of SNPs (x_{ij}) based on *theory* or *biological insights*

- The classic way of coming to a testable hypothesis
- If you find a significant relationship, you are pretty sure about *why* you find it

Example

- Nicolaou et al. (2011) attempt to find SNP-entrepreneurship associations (*cf. Lecture 1*)
- Risk-taking central in almost all entrepreneurship theories:
 - “Since dopamine receptor genes have been associated with novelty seeking/sensation seeking, and attention deficit hyperactivity disorder (ADHD) has been reported to occur at greater rates among entrepreneurs, **we examined the association between five dopamine receptor genes and four ADHD-associated genes**”

Example candidate gene study

- $N = 1,335$ (mostly females) from the UK Twin Registry, “In your working life, have you started a new business?” (Yes/No)
- *Not clear how many SNPs were tested*, but they find 18 SNPs with p -value < 0.05
- Authors consider the 1 SNP with p -value $< 6 \times 10^{-4}$ as truly significant
 - Significance threshold (α) corrected for testing several correlated SNPs (“Linkage disequilibrium”: SNPs physically close together in the genome are likely to be inherited together)

Table 2 Association between being an entrepreneur and all single nucleotide polymorphisms for all associations significant at $p < 0.05$

Marker	Position	Case	Control	Allele	Frequency	z	p value	Chromosome	Gene
rs1486011 ^a	115,374,300	282	975	C	0.054	-3.716	0.0002	3	DRD3
rs393795	1,481,514	301	1,028	T	0.198	-3.102	0.0019	5	SLC6A3
rs409588	1,483,834	302	1,033	T	0.197	-3.108	0.0019	5	SLC6A3
rs456082	1,483,515	302	1,033	G	0.197	-3.108	0.0019	5	SLC6A3
rs458860	1,483,933	302	1,033	A	0.197	-3.108	0.0019	5	SLC6A3
rs460000	1,485,825	302	1,033	T	0.197	-3.108	0.0019	5	SLC6A3
rs460700	1,482,969	302	1,033	C	0.197	-3.108	0.0019	5	SLC6A3
rs462370	1,484,164	302	1,033	C	0.197	-3.108	0.0019	5	SLC6A3

Candidate gene studies

- “This result is the first evidence of the association of a specific gene with entrepreneurship”
- They know why they found this association: “We postulate that subjects carrying allele G at the rs1486011 polymorphism of DRD3 are more likely to be sensation-seeking due to the higher arousal threshold needed to achieve a given level of dopamine in the brain. This higher arousal threshold may lead to greater odds of engaging in entrepreneurship, among other sensation-seeking activities.”
- “The small amount of variance explained by rs1486011 (**0.5%**) suggests that this gene alone does not predict the tendency to be an entrepreneur but, rather, it is one of many genes involved”
 - Effect small if we consider the h^2 estimate in Nicolaou et al. (2008) ($h^2 = 0.48$, cf. Lecture 1)
 - *But: Remember this explained variance of 0.5%*
- If this is a real effect, then we should be able to see it in other samples as well...

Non-replication

- Van der Loos et al. (2011) fail to replicate this finding using a much larger, independent dataset
 - $N = 9,365$ individuals from the Rotterdam Study a representative population sample (three subsamples) of elderly from Rotterdam
 - *Sweet memories: The first scientific paper I contributed to*

Table 2 Association results using two logit models of at least once self-employment for RS-II

SNP	Allele	Chromosome	Frequency	Model 1		Model 2	
				Beta	p value	Beta	p value
rs1486011	C	3	0.057	0.020	0.9330	0.017	0.9420
rs393795	T	5	0.203	-0.038	0.7811	-0.037	0.7860
rs409588	T	5	0.200	-0.038	0.7792	-0.037	0.7852
rs456082	G	5	0.200	-0.038	0.7789	-0.037	0.7848
rs458860	A	5	0.200	-0.038	0.7793	-0.037	0.7854
rs460000	T	5	0.199	-0.038	0.7792	-0.037	0.7855
rs460700	C	5	0.203	-0.038	0.7810	-0.037	0.7859
rs463379	C	5	0.200	-0.038	0.7791	-0.037	0.7853
rs464528	T	5	0.200	-0.038	0.7792	-0.037	0.7853
rs250682	C	5	0.203	-0.037	0.7814	-0.037	0.7861
rs456774	C	5	0.214	-0.011	0.9314	-0.013	0.9241
rs1486000	T	3	0.050	0.001	0.9960	0.000	0.9711

Explaining non-replications

Should we trust the “discovery” findings or the “replication” findings?

- *Can different environments explain the non-replication, or is there something more fundamentally wrong?*

Theory about biological mechanisms for behavior is often weak, and therefore theoretical rigor of candidate gene studies is illusive

- For example, >14,000 (70%) genes (and hence the SNPs in them) influence brain expression and we don’t know yet what all these genes are doing precisely

Thought example:

- With $\alpha = 0.05$, you get a “false positive” 1 out of 20 times. Hence, even under the null hypothesis of no association, for every 20 regressions you run you expect 1 “false positive”
- If you run 1.000.000 regressions (the approximate number of “independent” SNPs in the genome), you expect 50.000 “false positives”
- If you analyze only a subset of the 1.000.000 SNPs, how do you know you did not end up with a relatively large share of the 50.000 “false positives”?

“Ban” on candidate gene studies

Editors of “Behavior Genetics” (2012)

“The literature on candidate gene associations is full of reports that have not stood up to rigorous replication...”

As a result, the ... 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.”

Solution? “Adequately powered” studies

- Idea has implications beyond behavioral genetics

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

Received: 6 September 2011 / Accepted: 8 September 2011 / Published online: 18 September 2011
© Springer Science+Business Media, LLC 2011

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).

studies of complex traits, especially when reporting complex interaction effects based on novel phenotypes and groupings. Of course, we understand that this has not been done routinely—sometimes it is not practical—and so authors have preferred to publish the initial paper without such replication. We also recognize that there are historical examples where early failures to replicate were themselves misleading because of heterogeneity or poor methodology.

However, for a candidate gene or candidate gene-by-environment interaction study of a complex trait to be considered for publication in *Behavior Genetics* it should usually have one or more of the following characteristics:

- It is a rigorously conducted, adequately powered, direct replication study of a previously reported result; for well-conducted replication studies, there is no editorial

The “credibility crisis”

There is so much empirical data available, that researchers run the risk of post-rationalizing empirical results

- Finding a theory (explanation) after seeing the empirical results

However, Nicolaou et al. (2011)....

- Presented more SNP-entrepreneurship associations on a conference, why ended they up publishing only this subset?
- Used only one empirical measure for entrepreneurship, whereas the heritability paper by Nicolaou et al. (2008) using the same dataset used several other ones...

If statistical power would have been high, it would have been likely that findings were replicable

Statistical power

Strong focus on statistical significance (α) in empirical research, little attention for power (π)

- α : The probability of rejecting the null hypothesis (i.e. $H_0: \beta = 0$) when it is true (**lower is better**)
 - Example: You'll find an association between a SNP and entrepreneurship, although $\beta = 0$
- π : The probability of rejecting the null hypothesis (i.e. $H_0: \beta = 0$) when it is not true (**higher is better**)
 - Example: You'll find there is no association (high p -value) between a SNP and entrepreneurship, although $\beta \neq 0$

What are the consequences from low statistical power?

- Low chance of finding true effects ("false negatives")
 - A pity, but not so problematic usually...
- If something is found (i.e., a significant p -value for a SNP), the effect sizes will be (dramatically) over-estimated due to chance and they will have the wrong sign in many cases
 - Very problematic!

A widespread problem....

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P.A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a true relationships among th

PLoS Medicine (2005): The power problem explained for gene discovery

THE ECONOMIC JOURNAL

The Economic Journal, 117 (October), F296-F295. Doi: 10.1111/eoj.12461 © 2017 Royal Economic Society. Published by John Wiley & Sons, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

THE POWER OF BIAS IN ECONOMICS RESEARCH*

John P. A. Ioannidis, T. D. Stanley and Hristos Doucouliagos

We investigate two critical dimensions of the credibility of empirical economics research: statistical power and bias. We survey 159 empirical economics literatures that draw upon 64,076 estimates of economic parameters reported in more than 6,700 empirical studies. Half of the research areas have nearly 90% of their results under-powered. The median statistical power is 18%, or less. A simple weighted average of those reported results that are adequately powered (power $\geq 80\%$) reveals that nearly 80% of the reported effects in these empirical economics literatures are exaggerated; typically, by a factor of two and with one-third inflated by a factor of four or more.

Statisticians routinely advise examining the power function, but economists do not follow the advice.

McCloskey (1985, p. 204)

In Economics, median statistical power is 18% (80% is adequate)

Nearly 80% of the reported effects are exaggerated...

Power

Candidate gene study of Nicolaou et al. (2011) was also statistically **underpowered**

- Power (given a particular significance level α) is a function of the sample size and effect size of the SNP

Recall our conceptual model: $y_i = \mu + \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i$

Because of overidentification ($J > i$), we are going to estimate for a single SNP: $y_i = \mu + \beta x_i + \varepsilon_i$

- We drop subscript j to simplify notation
- Simplifying assumption: No covariates

Power (illustration)

H_0 (null hypothesis): $\beta = 0$

H_1 (alternative hypothesis): $\beta \neq 0$

Run the regression in your sample of size N , to obtain an estimate for β and its standard error

Compute test statistic $t = \frac{\hat{\beta}}{s.e.(\hat{\beta})}$

Compare test statistic with critical values of the Normal distribution, to obtain the p -value

- E.g. $t = -1.96 / 1.96$ gives p -value = 5% (2-sided test)
- Statistical significance threshold: α
- Coefficient β is declared “significant” if $p \leq \alpha$

Power (illustration)

What if H_0 is true, i.e., $\beta = 0$

- If β is truly 0 in the population, you should get $\hat{\beta} = 0$ in the regression in the full population

However, we usually only have a (hopefully representative) *subsample* of the full population

- Your $\hat{\beta}$ in this subsample depends on the true value of β , but also on random characteristics of the individuals you drew

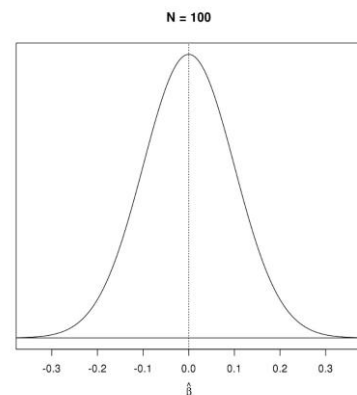
Let's run some simulations

- You randomly draw N individuals from the total population and estimate β in this subsample
- Repeat the first step 1,000 times, and every time store $\hat{\beta}$
 - Every time you have a different subsample; Therefore, every time your $\hat{\beta}$ is influenced by other random characteristics in your subsample
- Let's look at the distribution of these 1,000 stored $\hat{\beta}$'s

$N = 100$

$\hat{\beta}$ is Normally distributed around true $\beta = 0$

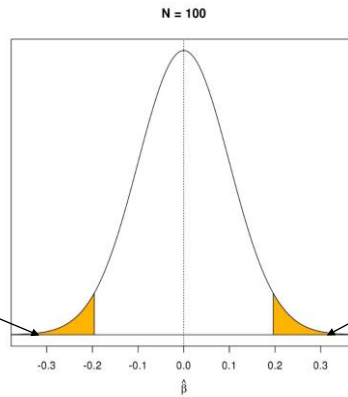
The variance of the estimator is equal to $\frac{\sigma^2}{N}$
 σ^2 = population variance of the error term
 N = sample size



$N = 100$

Even if the true $\beta = 0$, it is possible to obtain a regression coefficient of considerable size in a subsample of the population

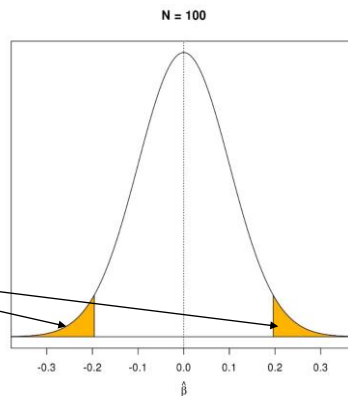
Left 2.5% of distribution



Right 2.5% of distribution

$N = 100$

Statistically significant at $\alpha = 0.05$



If you get a significant β at $p < 0.05$ in a sample of $N = 100$ (happens 1 in 20 times)...

It is a “false positive”, because the true $\beta = 0$

Nevertheless, your estimated effects will be $> |0.2|$

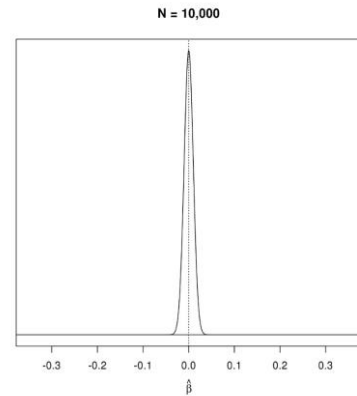
You will think you’ve found a moderate correlation, although there was nothing real at all...

$$N = 10,000$$

Things get better when you increase the sample size

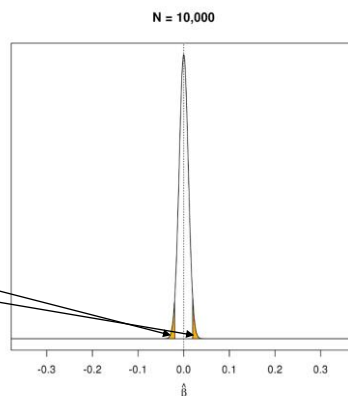
$\hat{\beta}$ is Normally distributed around true $\beta = 0$

The variance of the estimator is equal to $\frac{\sigma^2}{N}$
 σ^2 = population variance of the error term
 N = sample size



$$N = 10,000$$

Statistically
significant at
 $\alpha = 0.05$

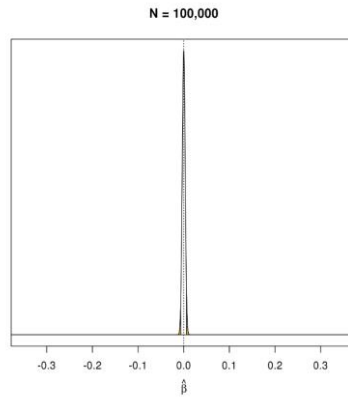


With $N = 10,000$, you'll get a much tighter distribution around the true value

A false positive at $\alpha = 0.05$ in $N = 10,000$ still occurs with a 5% chance

However, we won't be thinking anymore we found a large effect
 The left and the right 2.5% of the distribution are much closer to the true value $\beta = 0$

$$N = 100,000$$



If N increases it becomes better and better

We still get a significant estimate with 5% chance, but if so the effect is estimated to be fairly small

“The higher N , the better”

Power (illustration)

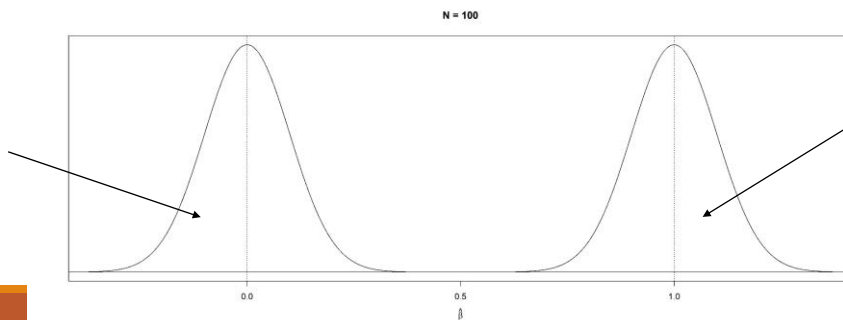
What if H_1 is true, i.e., $\beta \neq 0$

- H_1 (alternative hypothesis): $\beta \neq 0$

$\hat{\beta}$ is Normally distributed around true $\beta \neq 0$, with a variance equal to $\frac{\sigma^2}{N}$

Example: true $\beta = 1$

Distribution of $\hat{\beta}$
under null
hypothesis



Distribution of $\hat{\beta}$
under alternative
hypothesis

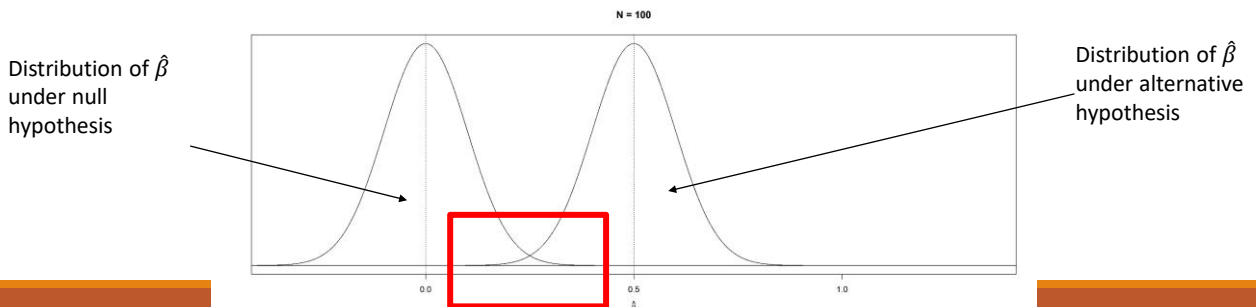
Power (illustration)

What if H_1 is true, i.e., $\beta \neq 0$

- H_1 (alternative hypothesis): $\beta \neq 0$

$\hat{\beta}$ is Normally distributed around true $\beta \neq 0$, with a variance equal to $\frac{\sigma^2}{N}$

Example: true $\beta = 0.5$



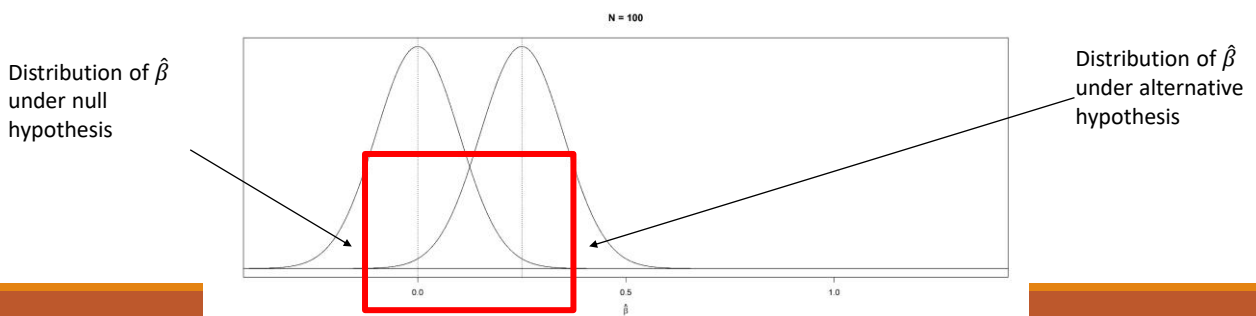
Power (illustration)

What if H_1 is true, i.e., $\beta \neq 0$

- H_1 (alternative hypothesis): $\beta \neq 0$

$\hat{\beta}$ is Normally distributed around true $\beta \neq 0$, with a variance equal to $\frac{\sigma^2}{N}$

Example: true $\beta = 0.25$



Distribution of $\hat{\beta}$ under H_1

What if H_1 is true, i.e., $\beta \neq 0$

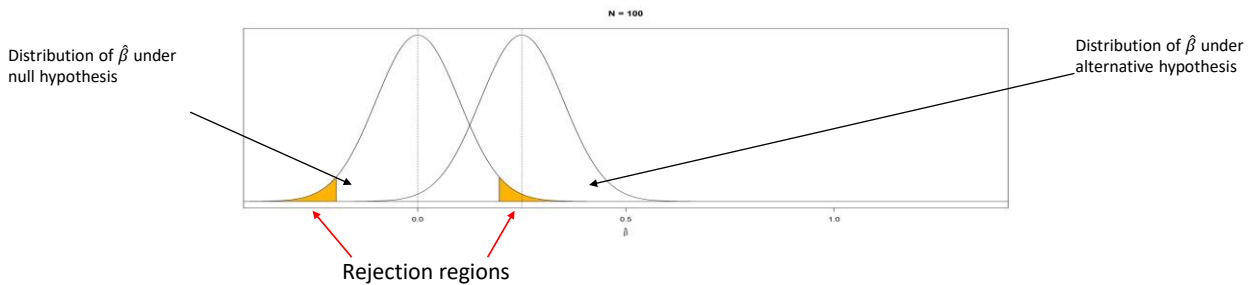
- H_1 (alternative hypothesis): $\beta \neq 0$

$\hat{\beta}$ is Normally distributed around true $\beta \neq 0$, with variance $\frac{\sigma^2}{N}$

- Example: true $\beta = 0.25$

π : The probability of rejecting the null hypothesis (i.e. $H_0: \beta = 0$) when it is not true

= Area under alternative hypothesis curve which is in the rejection region of the null hypothesis curve



Distribution of $\hat{\beta}$ under H_1

What if H_1 is true, i.e., $\beta \neq 0$

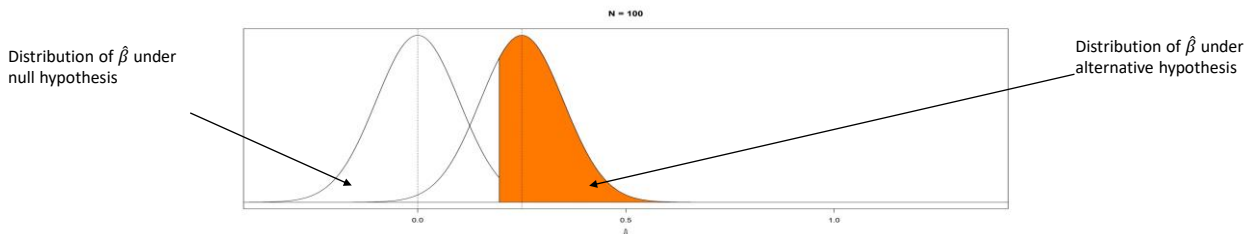
- H_1 (alternative hypothesis): $\beta \neq 0$

$\hat{\beta}$ is Normally distributed around true $\beta \neq 0$, with variance $\frac{\sigma^2}{N}$

- Example: true $\beta = 0.25$

π : The probability of rejecting the null hypothesis (i.e. $H_0: \beta = 0$) when it is not true

= Area under alternative hypothesis curve which is in the rejection region of the null hypothesis curve



Distribution of $\hat{\beta}$ under H_1

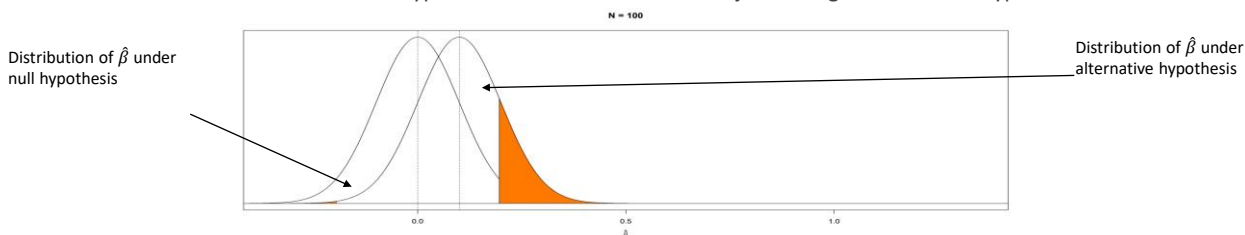
What if H_1 is true, i.e., $\beta \neq 0$

- H_1 (alternative hypothesis): $\beta \neq 0$

$\hat{\beta}$ is Normally distributed around true $\beta \neq 0$, with variance $\frac{\sigma^2}{N}$

Example: true $\beta = 0.10$

Power: Area under alternative hypothesis curve which is in the rejection region of the null hypothesis curve



Distribution of $\hat{\beta}$ under H_1

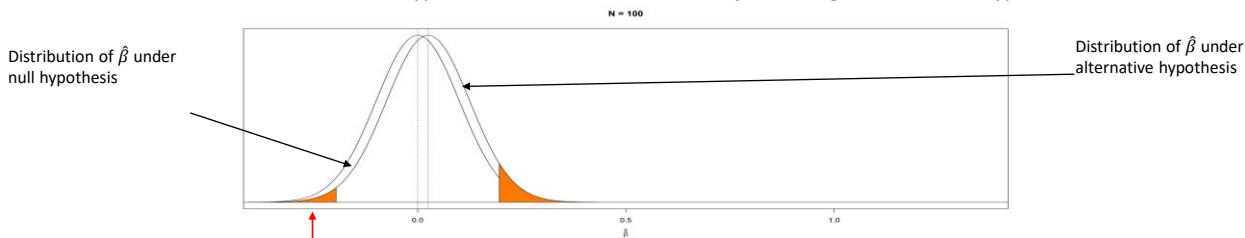
What if H_1 is true, i.e., $\beta \neq 0$

- H_1 (alternative hypothesis): $\beta \neq 0$

$\hat{\beta}$ is Normally distributed around true $\beta \neq 0$, with variance $\frac{\sigma^2}{N}$

Example: true $\beta = 0.05$

Power: Area under alternative hypothesis curve which is in the rejection region of the null hypothesis curve



With low power, there is even a chance the sign of your relationship is not correctly estimated

Power (illustration)

Nicolaou et al. (2011) had 6% power to detect and effect of $R^2=0.5$ at $\alpha=0.05$

GWAS on educational attainment (Rietveld et al., 2013)

- The genome-wide significant SNP with largest explanatory power has $R^2 = 0.02\%$, which implies a standardized effect $\beta = 0.014$ (i.e., $\sqrt{0.0002}$)

Chabris et al. (2015) argue this is a realistic maximum R^2 (*upper bound*) for any behavioral trait

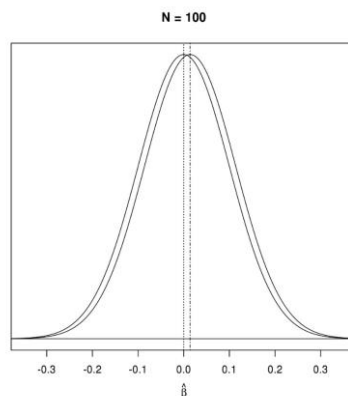
- Note that the effect is 25x smaller than the 0.5% reported by Nicolaou et al. (2011)!

$N = 100$

$$\tilde{\beta}_1 = 0.014$$

Left curve: Distribution of $\hat{\beta}$
under H_0

Right curve: Distribution of $\hat{\beta}$
under H_1



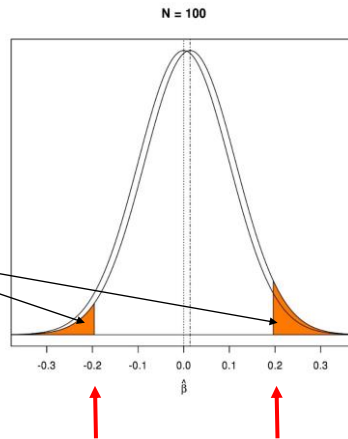
Overlaying the distributions of $\beta_1=0.014$ and $\beta_0=0$ shows that the distributions of estimated effects under H_0 and H_1 are very, very similar (with $N=100$)

$N = 100$

$$\tilde{\beta}_1 = 0.014$$

$$\tilde{\beta}_0 = 0$$

Power
(two-tailed)
at $\alpha = 0.05$



Statistical power: The probability of rejecting H_0 , given it is false (and the alternative, H_1 , is true)

So, your "draw" from the H_1 distribution should be in the "rejection region" of the H_0 curve

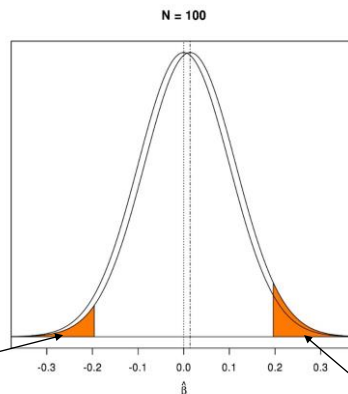
Power: Area under the H_1 curve in the rejection region of H_0 curve
In Tutorial 4 we use a function in R to calculate this area

A little more than 5% in this case

$N = 100$ (~6% power)

$$\tilde{\beta}_1 = 0.014$$

$$\tilde{\beta}_0 = 0$$



Wrong sign!

With $N = 100$, we have very limited power (a little less than 6%).

Hence, in most cases (~94%), we do not reject H_0 although it is false [a pity, but not too problematic, because we know "no evidence of effect" is not "evidence of no effect"]

However, if we DO reject H_0 , we do so based on a dramatically overestimated effect [very problematic!]

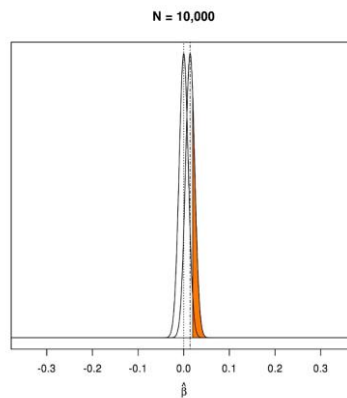
And in many cases, we'll even get the sign of the effect wrong! [extremely problematic!]

Overestimated effect size (>14x)!

$N = 10,000$ ($\sim 30\%$ power)

$$\tilde{\beta}_1 = 0.014$$

$$\tilde{\beta}_0 = 0$$



If N increases, the overlap of the distributions under H_0 and H_1 decreases

Higher power

If we do find something significant, the effects aren't as much inflated as with $N=100$

Also, it is much, much less likely that we'll find an estimate with a wrong sign

Summary

Candidate gene studies suffer from low statistical power

- Power is (given significance level α) a function of sample size and effect sizes of the SNP
 - Even if we would take the $R^2=0.5\%$ of Nicolaou et al. (2011) seriously, they only had 6% power to detect the effect at $\alpha=0.05$
- Low power results in i) "false negatives" and ii) overestimation of effects

If you set up a genetic discovery study (*Individual assignment!*), you want to be adequately powered (at least 80%) but collecting data is costly

- Therefore, you want to determine the minimum sample size you need in a genetic discovery study given:
 - Effect size of the SNP (not in your hands, make reasonable assumption based on earlier empirical findings)
 - Statistical significance level (the lower we set α , the lower π)
 - Required power (usually 80%)

We practice with "pre-study power analysis" in Tutorial 4 (in R)

Economics & Genetics

Lecture 4

GENOME-WIDE ASSOCIATION STUDIES




Genome-wide association studies (GWAS)

Alternative for candidate gene study: Hypothesis-free scan of all J SNPs (No selection on *weak* theory)

- Crude approach: 1 SNP per regression (millions of regressions, because of overidentification)
- Transparent correction for “multiple testing”
 - Europeans have ≈ 1 million independent common SNPs; Bonferroni-adjusted level of statistical significance is $\alpha = 0.05 / 1,000,000 = 5 \times 10^{-8} = \text{“Genome-wide significance”}$
- Power reduction because of low α , therefore very large N needed
- Large N via meta-analysis of GWAS results from multiple samples (GWAS conducted based on standardized protocols)

GWAS works

- GWAS results are replicable and have therefore replaced candidate gene studies for gene discovery
 - High power and hence low chance that findings are due to chance
 - *However, follow-up work is needed to understand the (biological) working of significant SNPs*
- 

Challenges in GWAS

Combining samples through meta-analysis may introduce bias

Population stratification = Difference in allele frequencies between (sub)populations (due to different ancestry)

- Problematic if these differences are correlated with environmental factors

Extreme example: “Beware the chopsticks gene” by Hamer & Sirota (2000)

- Allele frequencies of a SNP differ between Chinese and American individuals
- Way of eating (i.e., using chopsticks) culturally determined



Chopstick example

Sample 1 Americans: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

In both samples, no relation
between the SNP alleles
and use of chopsticks

Chopstick example

Sample 1 Americans: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

There is a clear allele frequency difference between Americans and Chinese

Chopstick example

Sample 1 Americans: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

There is a clear difference between Americans and Chinese in proportion of "cases" and "controls"

Chopstick example

Sample 1 Americans: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	320	640
Allele 2	80	80	160
Total	400	400	800

Sample 2 Chinese: $\chi^2=0, p=1$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	320	20	340
Allele 2	320	20	340
Total	640	40	680

Sample 1 + 2 = Americans + Chinese: $\chi^2=34.2, p=4.9 \times 10^{-9}$			
	Use of chopsticks		
	Yes	No	Total
Allele 1	640	340	980
Allele 2	400	100	500
Total	1040	440	1480

In the combined sample, we observe a significant relationship between the SNP and the use of chopsticks!

Implications

GWAS needs to be conducted in a genetically homogeneous sample

- In practice, this means that most GWAS use individuals from recent European ancestry because of data availability, but GWAS in e.g., Asians on the rise

Even within these samples, there can be subtle population stratification

- See, e.g. Abdellaoui et al. (2013), "Population structure, migration, and diversifying selection in the Netherlands"
- Inclusion of so-called "principal components" of the GRM can fix this (cf. Tutorial 3)
- Principal component analysis (mathematical details beyond the scope of this course) extracts latent factors from the GRM that explain genetic relatedness among individuals (example on next sheet)
- *We'll practice with the construction and use of "PCs" in Tutorial 4*

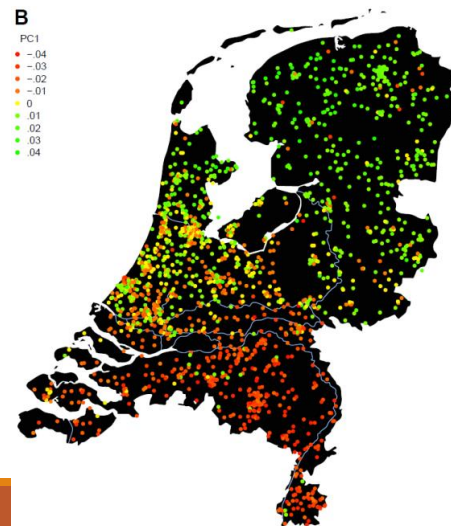
Population structure NL

Abdellaoui et al. (2013): Principal components of individuals Netherlands Twin Registry

Color of the dots represent the mean "PC" value per postal code (based on current living address of the individuals in the sample)

Although the Netherlands is relatively small, subtle genetic differences are present between the north and south

- Reason?
- Control for PCs in GWAS (4 PCs often works fine)



GWAS educational attainment (EA1)



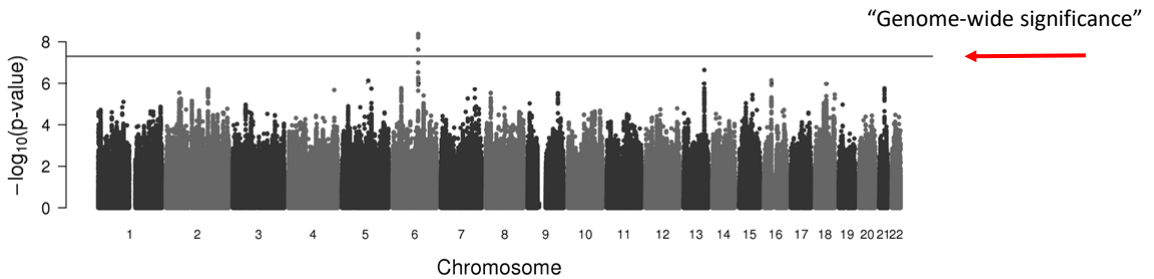
GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

Project kick-off in 2010, paper published in 2013

Traits: Years of education *and* Having a college degree or not

"The field's most high-profile article" (Bliss, 2018, "Social by nature", p. 141)

Manhattan plot (Years of education)



Every dot in the plot is a SNP

Genome-wide significance level: $-\log_{10}(5 \times 10^{-8}) = 7.3$

Manhattan plot (Correlation in the genome: SNPs physically close are inherited together)

Replicated SNPs from discovery stage

		Discovery ($N \sim 100K$)		Replication ($N \sim 25K$)		Combined ($N \sim 125K$)	
	SNP	<i>Beta</i>	<i>p</i> -value	<i>Beta</i>	<i>p</i> -value	<i>Beta</i>	<i>p</i> -value
Years of education	rs9320913	0.076	4.19×10^{-9}	0.062	0.024	0.076	3.50×10^{-10}

1 replicated SNP for years of education (two other SNPs for college yes/no):

- Effects very similar in discovery and replication stage
- Max $R^2 = 0.02\%$ for *Years of education*, ~ 2 months difference between the two homozygotes
- *Bioinformatics analyses suggest the involvement of central nervous system (anterior caudate nucleus)*

QQ-plot (Years of education)

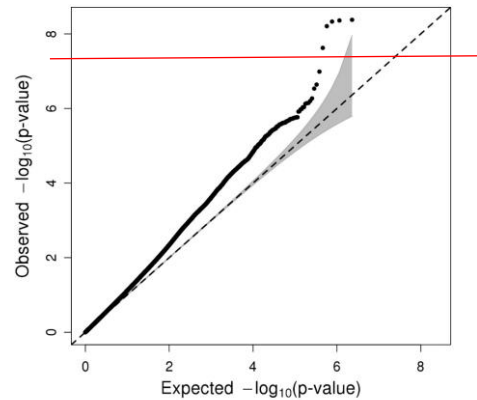
Manhattan plot already indicates *specific* SNP associations

In a QQ-plot you plot *all* observed p -values against the expected p -values

- Are GWAS results “enriched” for association? (i.e., are all p -values together lower than expected under H_0 of no association?)
- Useful check, particularly in small samples

Under H_0 , you expect the p -values to be uniformly distributed on the interval $[0, 1] \rightarrow$ 45 degree line

- Above diagonal \rightarrow More association than expected under H_0 of no association



Implications

Successful gene discovery

- As a result of the GWAS of educational attainment, launch of many other GWAS studies on behavioral outcomes (such as subjective well-being, risk preferences, and *income*)
- Large samples needed for GWAS with sufficient power, but associations then replicable

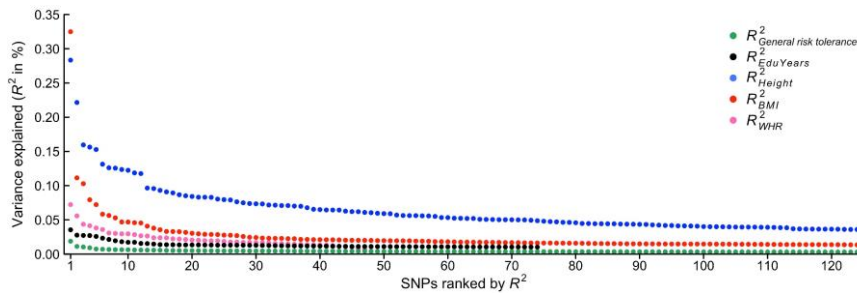
Individual genetic variants have tiny effects, there is no “gene for X”

- Small effects: “The fourth law of behavioral genetics”

The fourth law of behavior genetics

Most traits are influenced by hundreds / thousands of SNPs with small effect sizes (“polygenic trait”)

- Chabris et al. (2015). The **fourth** law of behavior genetics. *Current Direction in Psychological Science*, 24(4), 304-312.



Notes: Effect sizes of genome-wide significant SNPs in R^2 from various traits. The SNP with the lowest p -value for each approximately independent locus is displayed. Source: Linnér et al. 2018,

<https://www.biorxiv.org/content/early/2018/02/08/261081>

Educational Attainment 2.0

Higher power through higher N

- Fewer “false negatives”: More significant SNPs!
- Paper reports 1 SNP per genomic LD region, i.e. per “locus”

Okbay et al., 2016 (Nature)

LETTER

doi:10.1038/nature17671

Genome-wide association study identifies 74 loci associated with educational attainment

A list of authors and their affiliations appears in the online version of the paper.

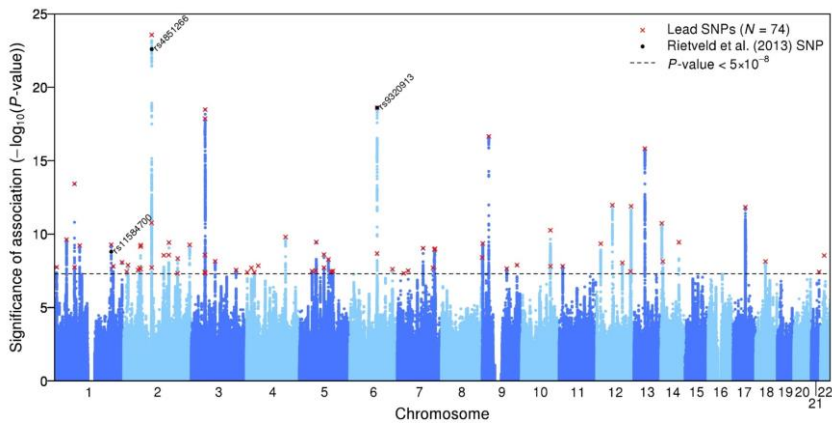
Educational attainment is strongly influenced by social and other environmental factors, but genetic factors are estimated to account for at least 20% of the variation across individuals¹. Here we report the results of a genome-wide association study (GWAS) for educational attainment that extends our earlier discovery sample² of 101,069 individuals to 293,723 individuals, and a replication study in an independent sample of 111,349 individuals from the UK Biobank. We identify 74 genome-wide significant loci associated with the number of years of schooling completed. Single-nucleotide polymorphisms associated with educational attainment are disproportionately found in genomic regions regulating gene expression in the fetal brain. Candidate genes are preferentially expressed in neural tissue, especially during the prenatal period, and enriched for biological pathways involved in neural development. Our findings demonstrate that, even for a behavioural phenotype that is mostly environmentally determined, a well-powered GWAS identifies replicable associated genetic variants that suggest biologically relevant pathways. Because educational attainment is measured in large numbers of individuals, it will continue to be useful as a proxy phenotype in efforts to characterize the genetic influences of related phenotypes, including cognition and neuropsychiatric diseases.

Our meta-analysis identified 74 approximately independent genome-wide significant loci. For each locus, we define the ‘lead SNP’ as the SNP in the genomic region that has the smallest P value (Supplementary Information section 1.6.1). Figure 1 shows a Manhattan plot with the lead SNPs highlighted. This includes the three SNPs that reached genome-wide significance in the discovery stage of our previous GWAS meta-analysis of educational attainment². The quantile–quantile (Q–Q) plot of the meta-analysis (Extended Data Fig. 1) exhibits inflation ($\lambda_{GC} = 1.28$), as expected under polygenicity³.

Extended Data Fig. 2 shows the estimated effect sizes of the lead SNPs. The estimates range from 0.014 to 0.048 standard deviations per allele (2.7 to 9.0 weeks of schooling), with incremental R^2 in the range 0.01% to 0.03%. To quantify the amount of population stratification in the GWAS estimates that remains even after the stringent controls used by the cohorts (Supplementary Information section 1.4), we used linkage-disequilibrium (LD) score regression⁴. The regression results indicate that ~8% of the observed inflation in the mean χ^2 is due to bias rather than polygenic signal (Extended Data Fig. 3a), suggesting that stratification effects are small in magnitude. We also found evidence for polygenic association signal in several within-family analyses, although these are not powered for individual SNP association testing.

Manhattan plot EA2

EA2 as further replication of EA1



Educational Attainment 3.0

Lee et al., 2018 (Nature Genetics)



N > 1,000,000 (mainly from UK Biobank and 23andMe)

1,271 approximately independent SNPs

EA 4.0 underway

Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals

James J. Lee^{1,58}, Robbee Wedow^{2,3,4,58}, Aysu Okbay^{6,58*}, Edward Kong⁷, Omeed Maghzi⁸, Meghan Zacher⁹, Tuan Anh Nguyen-Viet⁹, Peter Bowers⁷, Julia Sidorenko^{10,11}, Richard Karlsson Linnér^{6,4,2}, Mark Alan Fontana¹², Tushar Kundu⁶, Chanwook Lee⁷, Hui Li⁷, Ruoxi Li⁹, Rebecca Royer⁹, Pascal N. Timshel^{14,15}, Raymond K. Walters^{16,17}, Emily A. Willoughby¹⁸, Loïc Yengo¹⁹, 23andMe Research Team¹⁸, COGENT (Cognitive Genomics Consortium)¹⁹, Social Science Genetic Association Consortium¹⁸, Maris Alver¹¹, Yanchun Bao²⁰, David W. Clark²¹, Felix R. Day²², Nicholas A. Furlotte²³, Peter K. Joshi^{21,24}, Kathryn E. Kemper²⁵, Aaron Kleinman²³, Claudia Langenberg²⁶, Reedik Mägi¹¹, Joey W. Trampush^{25,26}, Shefali Setia Verma²⁷, Yang Wu²⁸, Max Lam^{28,29}, Jing Hua Zhao³⁰, Zhili Zheng^{30,30}, Jason D. Boardman^{3,14}, Harry Campbell³¹, Jeremy Freese³¹, Kathleen Mullan Harris^{32,33}, Caroline Hayward³⁴, Pamela Herd^{35,35}, Meena Kumari³⁶, Todd Lencz^{36,37,38}, Jian'an Luan³⁹, Anil K. Malhotra^{40,33,39}, Andres Metspalu^{41,39}, Lili Milani⁴², Ken K. Ong⁴³, John R. B. Perry⁴⁴, David J. Porteous⁴⁵, Marylyn R. Ritchie⁴⁶, Melissa C. Smart⁴⁷, Blair H. Smith^{48,42}, Joyce Y. Tung⁴⁹, Nicholas J. Wareham²², James F. Wilson^{50,51}, Jonathan P. Beauchamp⁴⁴, Dalton C. Conley⁴⁴, Tõnu Esko⁵², Steven F. Lehar^{44,44,47}, Patrik K. E. Magnusson⁴⁸, Sven Oskarsson⁴⁹, Tune H. Pers^{54,16}, Matthew R. Robinson^{50,50}, Kevin Thom⁵, Chelsea Watson⁵, Christopher F. Chabris⁵², Michelle N. Meyer⁴³, David I. Laibson⁷, Jian Yang^{55,54}, Magnus Johannesson⁴⁵, Philipp D. Koellinger^{54,52}, Patrick Turley^{56,17,57}, Peter M. Visscher^{58,54,55*}, Daniel J. Benjamin^{58,47,54,59*} and David Cesarini^{47,51,57,59}

Here we conducted a large-scale genetic association analysis of educational attainment in a sample of approximately 1.1 million individuals and identify 1,271 independent genome-wide significant SNPs. For the SNPs taken together, we found evidence of heterogeneous effects across environments. The SNPs implicate genes involved in brain-development processes and neuron-to-neuron communication. In a separate analysis of the X chromosome, we identify 10 independent genome-wide significant SNPs and estimate a SNP heritability of around 0.3% in both men and women, consistent with partial dosage compensation. A joint (multi-phenotype) analysis of educational attainment and three related cognitive phenotypes generates polygenic scores that explain 11–13% of the variance in educational attainment and 7–10% of the variance in cognitive performance. This prediction accuracy substantially increases the utility of polygenic scores as tools in research.

How to deal with tiny effects?

Effects of individual SNPs are very small, but what if we aggregate SNPs?

General idea: We sum up all SNPs, but weight each SNP by effect size in the GWAS

- SNPs with largest effects get most importance in the composite measure
- SNPs with small (~ 0) effects do not contribute

Polygenic score (PGS): The individual-level genetic susceptibility for a trait

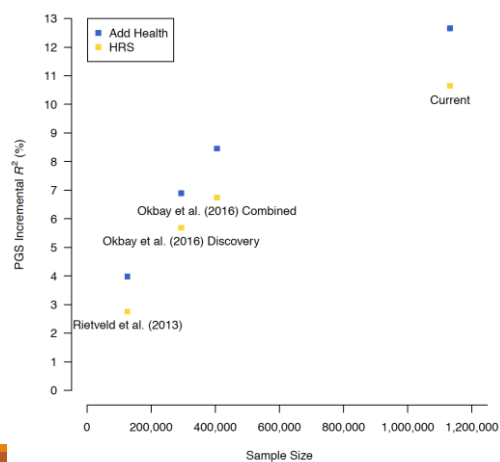
- For example, a high PGS value for EA indicates you have a high chance of attaining a high level of education

Polygenic prediction of educational attainment

Whereas a SNP explains only max. 0.02%, the PGS explains currently >10%!

Does the predictive power of the PGS have a limit? (cf. Lecture 3)

- Yes! Limit is the GREML heritability estimate ($\sim 25\%$ for EA)

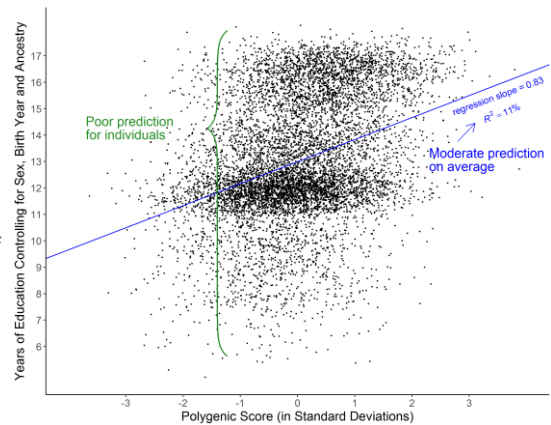


Polygenic prediction of educational attainment

PGSs with such large explanatory power are interesting for economist to include in empirical models (*cf. Lecture 2: The promises of geneoconomics*)

Two caveats

- PGS combines several (many different!) biological and environmental mechanisms: Interpretation of effect not straightforward
- R^2 is population level measure: Individual level prediction is poor!



MEGAN MOLTENT

SCIENCE 07.31.2019 07:00 AM

Are Diplomas in Your DNA?

Last week researchers announced more than 1,000 genetic variants associated with how far a person gets through school—along with warnings for how *not* to use that data. But earlier results from the same group are already available in a consumer product.



SESSIONS FEATURED

QUARTZ

EMAILS EDITIONS BE

ANCESTRYHEALTH

Ancestry's genetic tests can now tell you about your health

By Katherine Ellen Foley • October 15, 2019

Summary of Lecture 4

How can we identify the genes that influence behavior?

- Not through candidate gene studies: Their theoretical rigor is illusive and they are underpowered
- Through GWAS (but we need to deal with multiple testing and population stratification)

Why is the “gene for X” story flawed?

- Effect sizes of individual SNP are tiny (“the fourth law of behavior genetics”)
- Polygenic risk scores (based on GWAS results) explain a larger share of the variance, but
 - R^2 is population level measure: Individual level prediction is poor!




What comes next?

Tutorial 4

- R: Power analysis
- PLINK: Running a GWAS
- R: Visualizing GWAS results

Lecture 5

- How can we use GWAS results to identify causal effect in economic models?
 - How can polygenic scores be used to understand the interplay between nature and nurture?
- 

Presentation 2

The presenting group will be chosen randomly
(<https://www.random.org/integers/>)

- Format: 15 minutes + discussion
- The other teams that reviewed the same paper take the lead in the discussion



Economics & Genetics Lecture 4

DR. NIELS RIETVELD, NRIETVELD@ESE.EUR.NL

OFFICE MANDEVILLE T18-29

