

Economics & Genetics

Lecture 3

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

OFFICE MANDEVILLE T18-29

Tutorial 2

PLINK: Make sure to have downloaded the version compatible with *your* operating system

Make sure you are working in the directory in which you placed your PLINK executable

- In MacOS, you can check your current path with the "pwd" command
- The PLINK executable is in some downloads in a subfolder (e.g., plink_mac_20201019) → Copy to folder with your data
- Data should also be in this folder

Same holds for GCTA (Tutorial 3)

```

Opdrachtprompt
Microsoft Windows [Version 10.0.19041.572]
(c) 2020 Microsoft Corporation. Alle rechten voorbehouden.

C:\Users\Niels>cd Documents
C:\Users\Niels\Documents>cd EUR
C:\Users\Niels\Documents\EUR>cd "E&G"
C:\Users\Niels\Documents\EUR\E&G>cd Tutorials
C:\Users\Niels\Documents\EUR\E&G\Tutorials>plink
PLINK v1.90b6.18 64-bit (16 Jun 2020)      www.cog-genomics.org/plink/1.9/
(C) 2005-2020 Shaun Purcell, Christopher Chang  GNU General Public License v3

plink <input flag(s)...> [command flag(s)...] [other flag(s)...]
plink --help [flag name(s)...]

Commands include --make-bed, --recode, --flip-scan, --merge-list,
--write-snp-list, --list-duplicate-vars, --freqx, --missing, --test-mishap,
--hardy, --mendel, --ibc, --impute-sex, --indep-pairphase, --r2, --show-tags,
--blocks, --distance, --genome, --homozyg, --make-rel, --make-grm-gz,
--rel-cutoff, --cluster, --pca, --neighbour, --ibs-test, --regress-distance,
--model, --bd, --gxe, --logistic, --dosage, --lasso, --test-missing,
--make-perm-pheno, --tdt, --qfam, --annotate, --clump, --gene-report,
--meta-analysis, --epistasis, --fast-epistasis, and --score.

plink --help | more" describes all functions (warning: long).


C:\Users\Niels\Documents\EUR\E&G\Tutorials>
  
```

Today's agenda

Main questions:

- Are unrelated but genetically similar individuals more similar in their behavior?
- What does it mean that some behaviors are genetically related?
- Are genetically similar individuals attracted to each other? (*Presentation 1*)

Literature:

- Rietveld et al. (2013), "Molecular genetics and subjective well-being"
 - Boardman et al. (2015), "What can genes tell us about the relationship between education and health?"
- 

Economics & Genetics

Lecture 3

MAIN QUESTION 1





All human behavioral traits are heritable (first law of behavioral genetics)

But where are the underlying responsible genes? - Are the twin study estimates incorrect?

Leading approach

Lecture 1: To estimate heritability, you need to separate the effect of genes from the effect of the family environment

- The classical twin study uses a within family approach (i.e., common environment is constant)

Leading idea this lecture: Among unrelated individuals, there is no common environment (=family) effect

- Methodological approach developed to solve "the case of the missing heritability" (2008)
- We measure genetic variation across unrelated individuals using SNPs (Lecture 2 & Tutorial 2)

Finding SNP-outcome associations ("working horse" framework, Benjamin et al., 2012)

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

$i \in [1, \dots, N]$

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

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

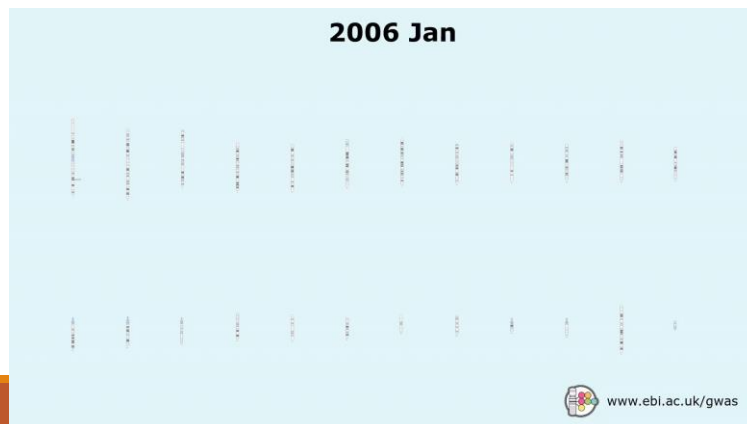
If $\beta_j \neq 0$, there is an association (not yet clear if it is a causal effect) between SNP j and outcome y

7

Finding SNP-outcome associations: Genome-wide association studies (GWASs)

Next week you will hear more about how to deal with the overidentification problem

- This video simply shows the number of revealed significant associations and their location in the DNA: <https://www.ebi.ac.uk/gwas/docs/diagram-downloads>



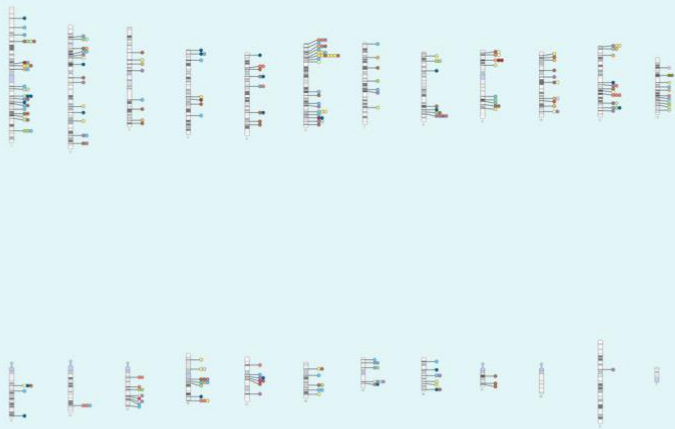
2008 Oct

Publication time “The case of the missing heritability”

Associations: 951

Studies: 240

Papers: 194



www.ebi.ac.uk/gwas

SNP-outcome associations, per chromosome



Statistically significant associations, *but tiny effects*

So few associations, with only tiny effects...

How to reconcile this with the twin study heritability estimates?

- Biased twin estimates? Look at other types of genetic variation than SNPs?

nature
genetics

ANALYSIS

Common SNPs explain a large proportion of the heritability for human height

Jian Yang¹, Beben Benyamin¹, Brian P McEvoy¹, Scott Gordon¹, Anjali K Henders¹, Dale R Nyholt¹, Pamela A Madden², Andrew C Heath², Nicholas G Martin¹, Grant W Montgomery¹, Michael E Goddard³ & Peter M Visscher¹

GREML: Intuition

- Genome-based Restricted Maximum Likelihood (GREML) estimation (implemented in GCTA software, see Tutorial 3)

- Starting point is the simple SNP-outcome association framework

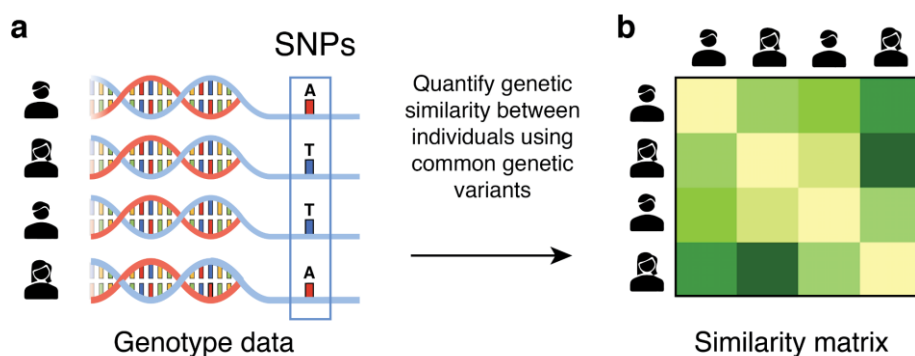
$$y_i = \mu + \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i$$

- We want to estimate heritability, and hence we want to know $\sum_{j=1}^J \beta_j x_{ij}$ rather than all the β_j 's
- Two tricks:
 - We assume that all SNP effects β originate from a Normal distribution with mean 0
 - Thereafter (like twin studies), we analyze the variances rather than the levels (we take the variance of both sides of the equation)

GREML: Intuition

- In a set of unrelated individuals, the observed variance of $\mu + \sum_{j=1}^J \beta_j x_{ij} + \varepsilon_i$ comes from the middle term only
- We already made an assumption about the distribution of β_j , so we can focus on the SNPs (x_{ij})
 - For a variance-analysis, we have to construct the “Genetic Relationship Matrix” (GRM), which captures the pairwise genetic relationship between each pair of individuals in the sample based on SNPs (computational details follow later)
- For comparison: The classical twin study works with expected (family-based) genetic relationships (MZ = 1, DZ = 0.5)
 - GREML uses observed genetic relationships (based on SNPs) in a sample of unrelated individuals
 - Genetic relationships ~ 0
- Method not completely new: Inspired by animal breeding literature
 - However: The use of observed genetic variation (i.e., SNPs) was really innovative

GREML: Intuition



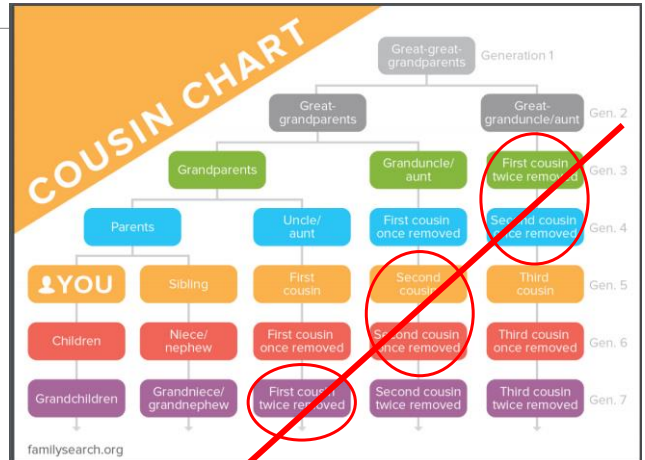
GREML: Intuition

Key assumption: (Unique) environmental influences are independent of genetic relatedness among unrelated individuals

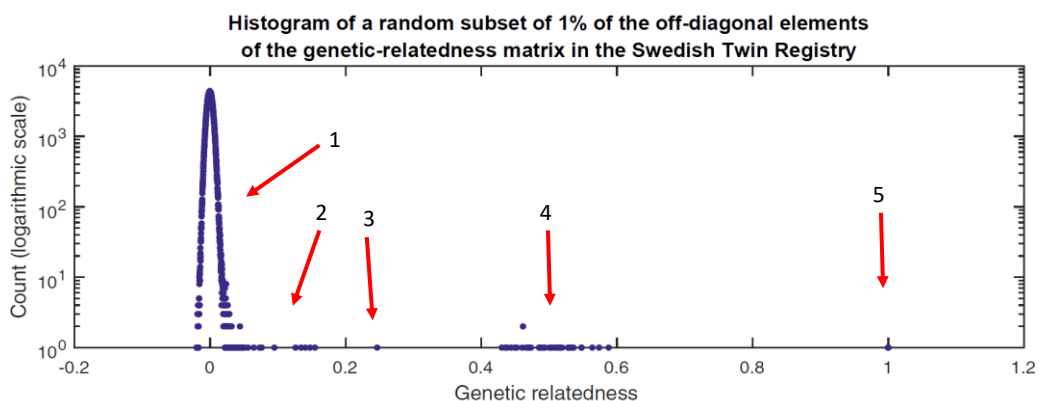
Therefore: Exclude cousins two/three times removed and closer related

- What would happen if you do not do this?

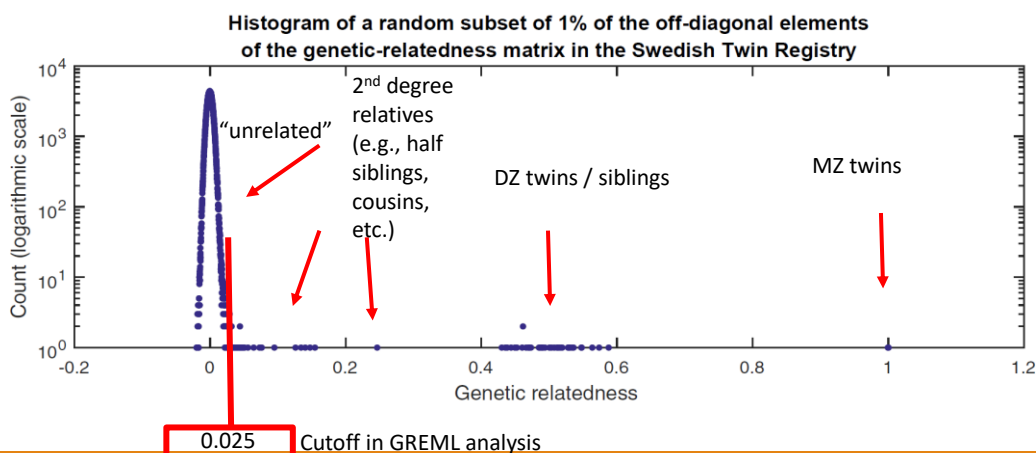
Do you think you “share the environment” with your cousins two/three times removed?



Exercise: What kind of family relations do we see in this GRM? (cf. Lecture 1)



Exercise: What kind of family relations do we see in this GRM? (cf. Lecture 1)

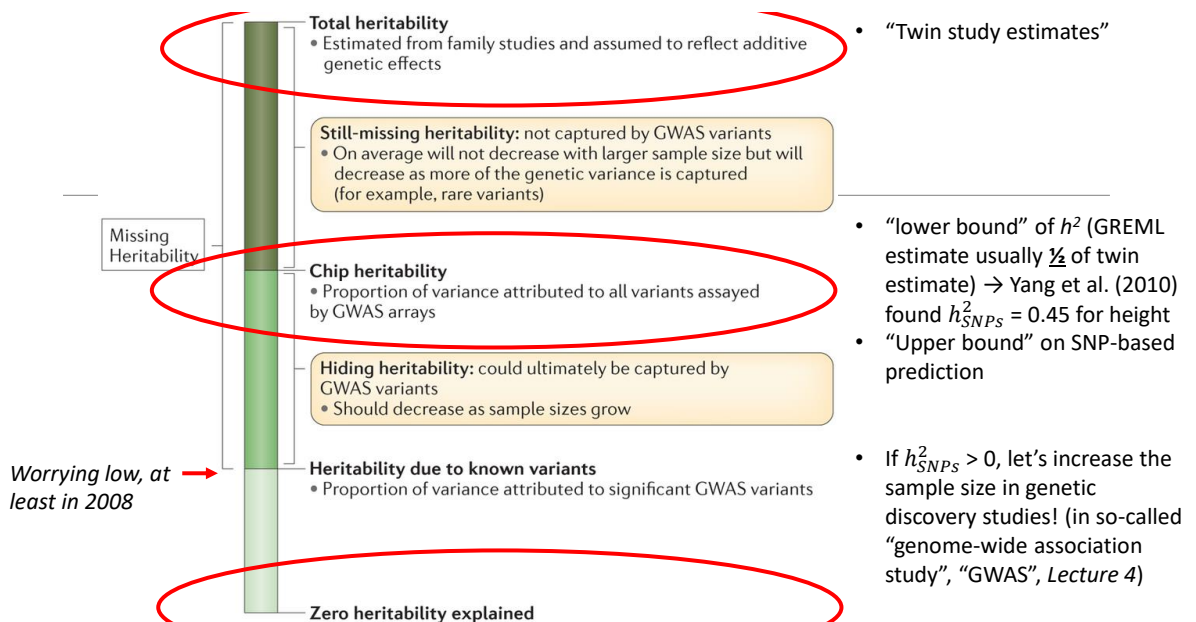


GREML: Intuition

- Construct the GRM in your sample
- Exclude individuals who are "too related" (based on the GRM, cutoff 0.025)
- Examine whether individuals who are genetically more similar are also trait-wise more similar
- Resulting estimate can be interpreted as proportion of variance in a trait accounted for by the genotyped SNPs

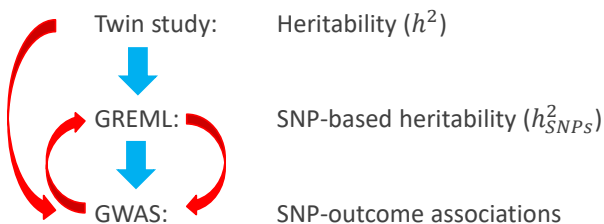
$$h_{\text{SNPs}}^2 = \frac{\sigma_A^2}{\sigma_y^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_E^2}$$

- σ_y^2 is the variance of trait (phenotype) y
- A is the genetic component (as in the classical twin study reflecting additive genetic effects)
- No family component (C) in GREML
- E reflects unique environmental variance (E)
- GRML only estimates σ_A^2



Source: Witte et al., 2014. The contribution of genetic variants to disease depends on the ruler. *Nature Reviews Genetics* 15, 765-776.

Schematic overview methods



The SNP heritability of subjective well-being

Rietveld et al. (2013), “Molecular genetics and subjective well-being”

Subjective well-being (SWB) is defined as 'a person's cognitive and affective evaluations of his or her life' (Diener, Lucas, & Oshi, 2002, p. 63) → Increasingly popular “broad prosperity” measure

- Affective state present situation (e.g., during past week, did you enjoy life)
- Cognitive evaluation (e.g., satisfied with life/work)

Twin study heritability estimates range between 30%-40%

- Should we launch a gene discovery study on SWB? Let’s use GREML first!

Data from the Rotterdam Study (RS) & the **Swedish Twin Registry (STR)**

Phenotype (trait) distribution

2 SWB traits: How happy were you last week? / Did you enjoy life last week?

Table S3. Descriptive statistics: Single-question SWB measures

SWB measure	Sample	n	% Rarely or none of the time (less than 1 d)	% Some or a little of the time (1–2 d)	% Occasionally or a moderate amount of time (3–4 d)	% Most or all of the time (5–7 d)
Happy	RS-I	3,842	7.0	7.7	15.9	69.5
	RS-II	2,075	4.5	11.0	22.3	62.2
	RS-III	2,992	2.9	9.8	20.3	67.1
	STR	6,675	5.1	11.0	44.9	39.0
	STR+RS	15,584	5.1	10.0	30.0	55.0
Enjoy	RS-I	3,866	6.1	7.0	13.3	73.6
	RS-II	2,080	4.2	10.3	17.9	67.6
	RS-III	2,990	2.6	8.7	15.7	72.9
	STR	6,751	2.4	3.7	27.0	66.9
	STR+RS	15,687	3.6	6.3	20.3	69.8

This table provides summary statistics for the SWB measures used in the GREML analysis.

Categorical measure, but GREML can only handle continuous and binary traits...

- Recoding to binary variables: “Most or all of the time” (1) versus other categories (0)

Results



It does make sense to launch a genetic discovery study for SWB!

In 2016, we were able to publish about the first genetic variants associated with subjective well-being (Okbay et al., 2016, *Nature Genetics*)

Fig. 1. This figure shows the GREML for *Happy*, *Enjoy*, and *Combined*. The error bars represent the point estimate ± 1 SD. The sample pools the three Rotterdam cohorts (RS) and the Swedish Twin Registry TwinGene sample (TG).

Liability threshold model (cf. Lecture 2)

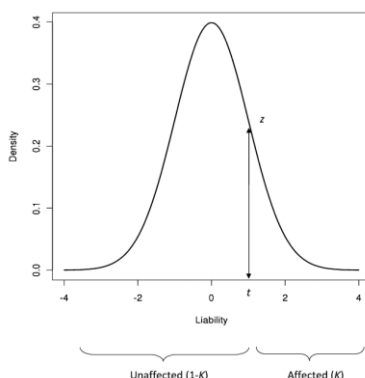


Figure 1. The Liability Threshold Model for a Disease Prevalence of K . An underlying continuous random variable determines disease status. If liability exceeds the threshold t , then individuals are affected.

For binary outcomes, we assume there is an underlying continuous liability

If your liability is sufficiently high, you are “affected”

Transform the heritability on the observed (1/0) scale to the heritability on the continuous liability scale as follows:

$$h_l^2 = h_o^2 K(1 - K)/z^2.$$

K = “Prevalence” disease (sampling correction)

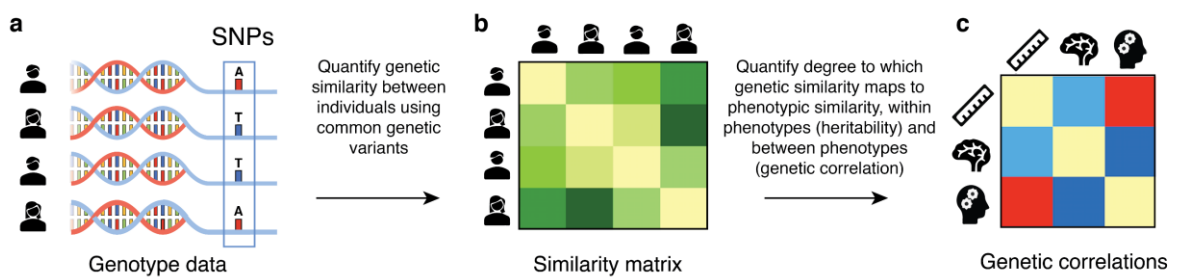
Transformation is implemented in GCTA, we’ll practice with it in *Tutorial 3*

Economics & Genetics

Lecture 3

MAIN QUESTION 2

Similarity *within* and *between* traits



GREML: A closer look at the GRM

Empirical examples in Tutorial 3

The Genetic Relationship Matrix (GRM) captures to what extent individuals are genetically similar

- Remember the bed/bim/fam structure of PLINK data, x takes value 0, 1, or 2
- Every individual i is related to his/herself, and all other $N-1$ individuals in the sample
- Individuals (i), and SNPs (j) with allele frequency AF_j (`--freq`)
- For each SNP, you calculate whether individuals are more related to each *than can be expected based on the allele frequency*

	$i = 1$	$i = 2$...
$i = 1$	$\frac{1}{J} \sum_j \frac{(x_{j,1} - 2AF_j)^2}{2AF_j(1 - AF_j)}$
$i = 2$	$\frac{1}{J} \sum_j \frac{(x_{j,2} - 2AF_j)(x_{j,1} - 2AF_j)}{2AF_j(1 - AF_j)}$
$i = 3$	$\frac{1}{J} \sum_j \frac{(x_{j,3} - 2AF_j)(x_{j,1} - 2AF_j)}{2AF_j(1 - AF_j)}$
.
.
.

GREML: A closer look at the GRM

By averaging over all (standardized by allele frequency) SNPs:

- You expect genetic similarity to be 0 for completely unrelated individuals
- However, due to “chance” (or very distant relationship), some individuals will be more similar to each other than others
- Or, a bit less similar than expected (genetic relationship can be negative)

	$i = 1$...
$i = 1$	$\frac{1}{J} \sum_j \frac{(x_{j,1} - 2AF_j)^2}{2AF_j(1 - AF_j)}$...
$i = 2$	$\frac{1}{J} \sum_j \frac{(x_{j,2} - 2AF_j)(x_{j,1} - 2AF_j)}{2AF_j(1 - AF_j)}$...
$i = 3$	$\frac{1}{J} \sum_j \frac{(x_{j,3} - 2AF_j)(x_{j,1} - 2AF_j)}{2AF_j(1 - AF_j)}$...
.
.
.

This is the genetic variation GREML exploits

GREML: A closer look at the GRM

What about genetic relatedness with yourself? (diagonal of the matrix)

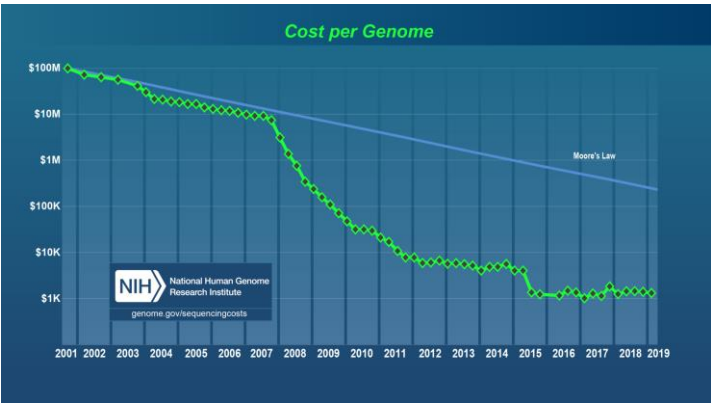
- You expect genetic similarity to be 1...
- However, you standardize SNPs by allele frequency in the sample
- If there is assortative mating of parents (“inbreeding”), individuals may be more or less homozygote (same SNP alleles, $x_{j,i} = 0$ or 2) than expected in a sample of unrelated individuals
- Therefore, values on diagonal of the GRM may be slightly different from 1

	$i = 1$...
$i = 1$	$\frac{1}{J} \sum_j \frac{(x_{j,1} - 2AF_j)^2}{2AF_j(1 - AF_j)}$...
$i = 2$	$\frac{1}{J} \sum_j \frac{(x_{j,2} - 2AF_j)(x_{j,1} - 2AF_j)}{2AF_j(1 - AF_j)}$...
$i = 3$	$\frac{1}{J} \sum_j \frac{(x_{j,3} - 2AF_j)(x_{j,1} - 2AF_j)}{2AF_j(1 - AF_j)}$...
.
.
.

Why GREML is feasible

GREML exploits (very) small genetic differences
You'll need a few thousand genotyped individuals

Costs of genotyping have decreased dramatically...
Moore's law is the observation that the number of transistors in a dense integrated circuit doubles about every two years



Genetic data: Cheap but valuable

Nowadays, it is possible to collect high-accuracy measures of an individual's genome at reasonable cost

- Sequencing (all ~3 billion genetic variants) ~ 1,000€ p.p.
- Common genetic variants (~1 million SNPs) ~ 30€ p.p.
- Customized sub-sets of common genetic variants (e.g., for specific diseases) ~ 20€ p.p.

Drastic decreases in genotyping costs made

- Large-scale surveys starting to collect data
- Private companies stepping in...

Some interesting datasets

National Longitudinal Study of Adolescent to Adult Health (Add Health)

- Used in Paper 1 Group assignment, <http://www.cpc.unc.edu/projects/addhealth> (thesis option)

Avon Longitudinal Study of Parents and Children (ALSPAC)


- Used in Paper 3 Group assignment, <http://www.bristol.ac.uk/alspac/>

UK Biobank (SNP data ~1 TB!)

- $N \approx 500,000$, <http://www.ukbiobank.ac.uk/>


Health and Retirement Study (HRS)

- Thesis option, <http://hrsonline.isr.umich.edu/>



OUR SERVICES ▾HOW IT WORKS ▾REPORTSSTORESSHOP ▾

SIGN INREGISTER KITHELP ▾




Live in the know™


Discover what 90+ personalized reports have to say about your ancestry, health, wellness and more.

[shop](#)


SLEEP MOVEMENT



MUSCLE COMPOSITION



CAFFEINE CONSUMPTION




Health + Ancestry service

Don't trust "health predictions"!

[More about prediction in upcoming lectures]

Ancestry Service




Experience your ancestry in a new way! Get a breakdown of your global ancestry by percentages, connect with DNA relatives and more. [learn more](#)

€99

[add to cart](#)

Health + Ancestry Service



Get an even more comprehensive understanding of your genetics. Receive 90+ online reports on your ancestry, traits and health - and more. **New** BRCA1/BRCA2 (Selected Variants)* report just added! [learn more](#)

€169

[add to cart](#)



[Important Test Info](#)

Would you be interested in this kind of information?

Sort by

Percent Related ▾

 Showing 1018 out of 1018 relatives

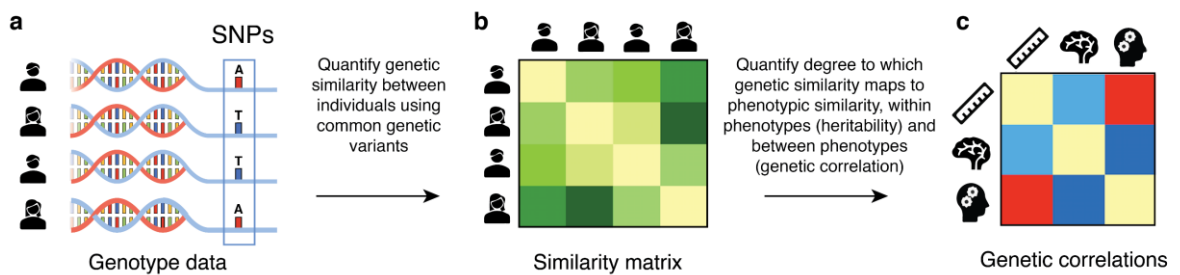
★	Name	Relationship	Sharing
☆	<div>NK</div> Nico Kleyn Male	Second to Third Cousin 1.08% DNA shared, 6 segments	●
☆	<div>AS</div> Anton Schep Male	Third to Fourth Cousin 0.65% DNA shared, 5 segments	●
☆	<div>AS</div> AS Female	Third to Fourth Cousin 0.54% DNA shared, 4 segments	●
☆	 Kees Been Male	Third to Fifth Cousin 0.47% DNA shared, 3 segments	●
☆	<div>RK</div> Rosemary Kievit Female	Third to Fifth Cousin 0.43% DNA shared, 3 segments	●
☆	<div>AB</div> AB Male	Third to Fifth Cousin 0.41% DNA shared, 2 segments	●
☆	 Jackie Van Drunen Female	Third to Sixth Cousin 0.40% DNA shared, 1 segment	●

But what about these stories?



https://www.youtube.com/watch?v=1BgWM5LE_i8

Similarity *within* and *between* traits

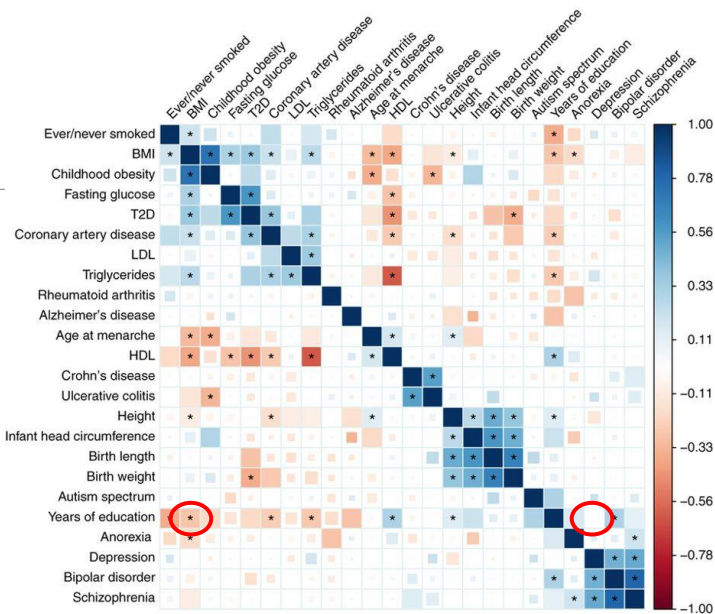


Bivariate GREML

Yang et al. (2010) originally developed GREML for estimating SNP-heritability (univariate analysis)

Lee et al. (2012) made it possible to estimate the SNP heritability of two traits simultaneously

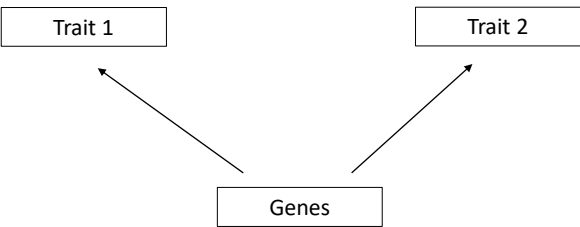
- This enabled the estimation of the genetic correlation between two traits
- To which extent are traits related at the genetic level?



Source: Bulik-Sullivan et al. (2014), “An atlas of genetic correlations across human diseases and traits.”

Genetic correlations

Enthusiastic reactions...



Explaining genetic correlations

...but interpretation of genetic correlations not straightforward!

- Different mechanisms can be behind it (figure from Boardman et al. (2015))

Genetic correlation ≠ Pleiotropy: A gene influencing more than one trait at the same time

- No pleiotropy (A) [two traits may influence each other, but different genes influence the two traits]
- Mediation pleiotropy (B and C)
- Biological pleiotropy (D) [same genes influencing two traits]

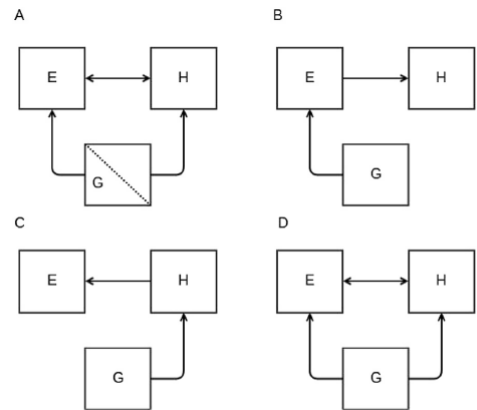


Fig. 3. Examples of four hypothetical causal relationships between Education (E), Health (H), and Genetics (G).

Education and health at the genetic level

Boardman et al. (2015), “What can genes tell us about the relationship between education and health?”

Table 3
Bivariate genome wide covariance estimates for education and three health outcomes.

	Body mass index	Depression	Self-rated health
Genetic variance			
Health	10.668	0.007	0.128
Education	2.139	2.173	2.142
Cov (health,education)	-0.159	-0.089	-0.477
Environmental variance			
Health	14.677	0.028	0.576
Education	4.059	4.025	4.055
Cov (health,education)	-0.788	-0.003	-0.178
Phenotypic variance			
Health	25.345	0.034	0.704
Education	6.197	6.198	6.198
Heritability			
Health	0.421	0.193	0.181
Education	0.345	0.351	0.346
rG	-0.033	-0.746	-0.912
95% CI (rg)	(-0.297, .331)	(-1.0, -0.201)	(-1.0, -0.374)
logL	-14860.413	-837.535	-7089.678
logL0 (rG = 0)	-14860.429	-841.035	-7094.394
LRT	0.032	6.999	9.432
df	1	1	1
pr. <	0.4	0.004	0.001

Note: Data come from the Health and Retirement Study; n = 4233.

Exercise: For each trait, which mechanism do you consider most plausible given the r_G with education?

Table 3
Bivariate genome wide covariance estimates for education and three health outcomes.

	Body mass index	Depression	Self-rated health
Genetic variance			
Health	10.668	0.007	0.128
Education	2.139	2.173	2.142
Cov (health,education)	-0.159	-0.089	-0.477
Environmental variance			
Health	14.677	0.028	0.576
Education	4.059	4.025	4.055
Cov (health,education)	-0.788	-0.003	-0.178
Phenotypic variance			
Health	25.345	0.034	0.704
Education	6.197	6.198	6.198
Heritability			
Health	0.421	0.193	0.181
Education	0.345	0.351	0.346
r_G	-0.033	-0.746	-0.912
95% CI (r_G)	(-0.297, .331)	(-1.0, -0.201)	(-1.0, -0.374)
logL	-14860.413	-837.535	-7089.678
logL0 ($r_G = 0$)	-14860.429	-841.035	-7094.394
LRT	0.032	6.999	9.432
df	1	1	1
pr. <	0.4	0.004	0.001

Note: Data come from the Health and Retirement Study; $n = 4233$.

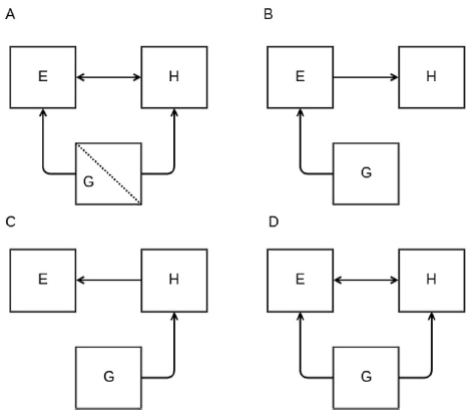


Fig. 3. Examples of four hypothetical causal relationships between Education (E), Health (H), and Genetics (G).

Question: For each trait, which mechanism do you consider most plausible given the r_G with education?

Education & BMI \rightarrow A? (If we assume r_G to be truly 0...)

Education & Depression \rightarrow B, C, D? (But even A is possible...)

Education & Self-rated health \rightarrow B, C, D? (But even A is possible...)

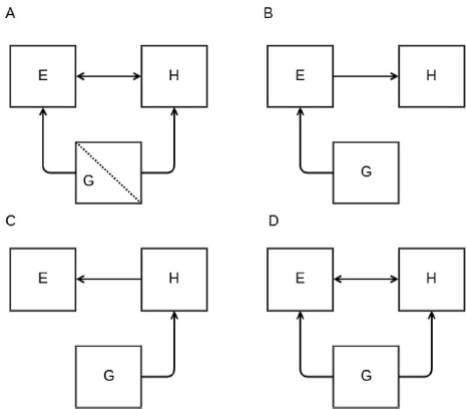


Fig. 3. Examples of four hypothetical causal relationships between Education (E), Health (H), and Genetics (G).

Implications

Genetic correlations are interesting, but are not informative about the underlying causal mechanism

Conclusion Boardman et al. (2015): “In the future, causal analyses of the relationship between education and health should consider whether more accurate estimates might be obtained if **shared** genes, even amongst unrelated individuals, are accounted for in the estimation of the causal relationship”

- However, doing so properly requires additional knowledge of the nature of the pleiotropic effects in question...
- Biological pleiotropy (D) extremely complicated to solve (you can't “block” a pathway in your model)
- Mediation pleiotropy (B and C) easier to fix

Summary of Lecture 3

1. Are unrelated but genetically similar individuals more similar in their behavior?
 - Third law of behavioral genetics: All human behavioral traits are heritable
 - GREML heritability estimate (h_{SNPs}^2) usually $\frac{1}{2}$ of twin study estimate: Grounded hope for SNP discovery!
2. Are behaviors related at the genetic level?
 - Bivariate GREML analysis to estimate genetic correlations
 - Interpretation of genetic correlation not straightforward (genetic correlation \neq “biological pleiotropy”)
3. Are genetically similar individuals attracted to each other?
 - SNP data makes it possible to study such questions (“10 Shocking Results from DNA Ancestry Tests”)
 - Presentation 1: Domingue et al. (2018), “The social genome of friends and schoolmates in the National Longitudinal Study of Adolescent to Adult Health”

What comes next?

- Tutorial 3
 - Practice with GREML estimation using GCTA (and work further on Individual assignment)
- Lecture 4
 - How can we identify the genes that influence behavior?
 - Why is the “gene for X” story flawed?



Presentation 1

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 3

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

OFFICE MANDEVILLE T18-29

