Estimation of additive genetic variance and covariance using individual-level data

Loic Yengo, PhD

Institute for Molecular Bioscience

The University of Queensland

l.yengo@imb.uq.edu.au

What quantitative genetics tells us about heritability?

Part 1

Outline

- Definition of heritability (and genetic correlation)
- What is heritability used and useful for?

Definitions

Heritability (*h^2*) quantifies the degree to which inter-individual differences and resemblance in the population are due to genetic factors.

Chial, H. (2008) Polygenic inheritance and gene mapping. Nature Education 1(1):17

Definitions

If the value, Y, of trait (=phenotype) can be modelled as

$$Y = \underbrace{G}_{\text{Genetic factors}} + \underbrace{E}_{\text{Non-genetic factors}}$$

then h^2 = var(G) / var(Y), i.e. proportion of trait variance explained by genetic factors.

Chial, H. (2008) Polygenic inheritance and gene mapping. Nature Education 1(1):17

Nice definition but not very useful unless we can observe G!

Definitions

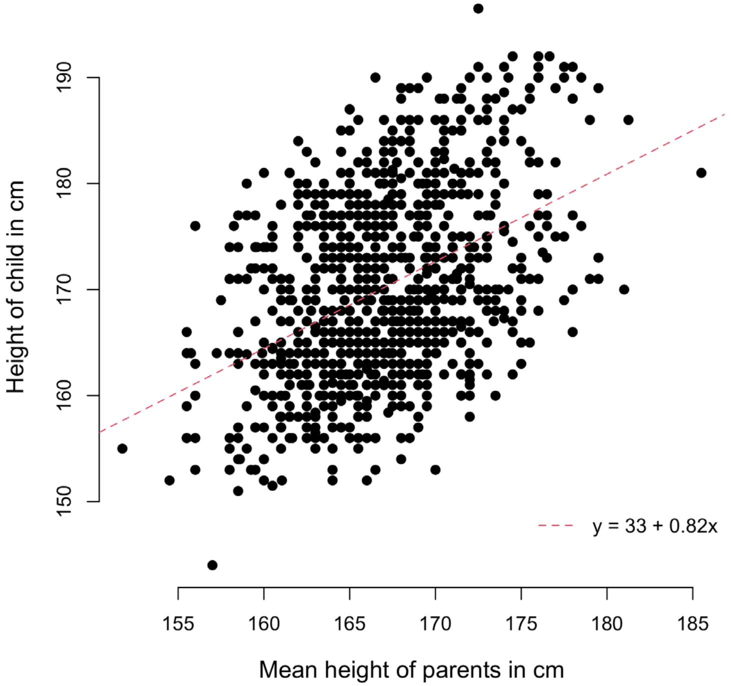
Another view of h^2 is the given by the phenotypic correlation between relatives.

E.g.,

- How much is the height of siblings correlated?
- How much having a family history of schizophrenia predisposes you to also develop it?

Quantitative Genetics theory answers those questions using the following Equation

$$\mathrm{corr}(Y_i, Y_j) = h^2 R_{ij} + \mathrm{Residual}$$



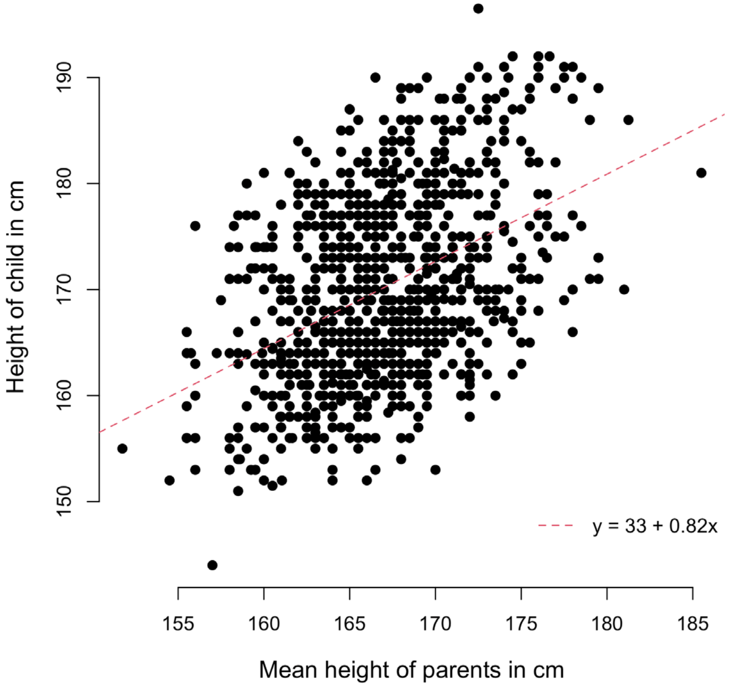
Data from UK Biobank participants
(Application number 12505)

Definitions

Quantitative Genetics theory answers those questions using the following Equation

$$\mathrm{corr}(Y_i,Y_j) = h^2 R_{ij} + \mathrm{Residual}$$

where R_{ij} is the **coefficient of genetic relationship** between individual i and individual j (e.g., R_{ij}=0.5 for full siblings).



Data from UK Biobank participants
(Application number 12505)

this definition is a bit more useful
As we can observe both corr(Y_i,Y_j)
and R_{ij} **(Part 2).**

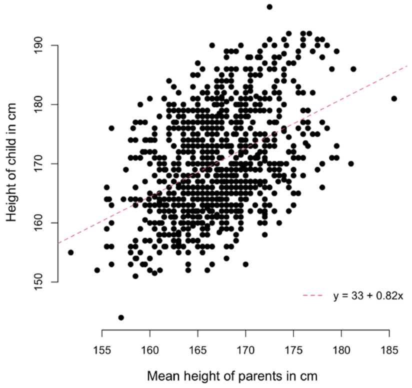
Definitions

Heritability (h^2) quantifies the degree to which inter-individual differences and resemblance in the populations are due to genetic factors.

Heritability can be approached in terms of

- Differences between people in the population: h^2 = var(G) / var(Y),

- Resemblance between relatives (in families): $\mathrm{corr}(Y_i, Y_j) = h^2 R_{ij} + \mathrm{Residual}$



Is heritability a universal constant?

No!

Heritability is a property of a trait, in a population, at a given time.

Heritability can change over time (e.g., Rimfeld et al. NHB 2018, heritability of educational attainment in Estonia before/after the cold war).

Larger is not necessarily better.

Genetic correlation

The genetic correlation (r_g) between two traits Y_1 and Y_2 can be defined similarly as the heritability

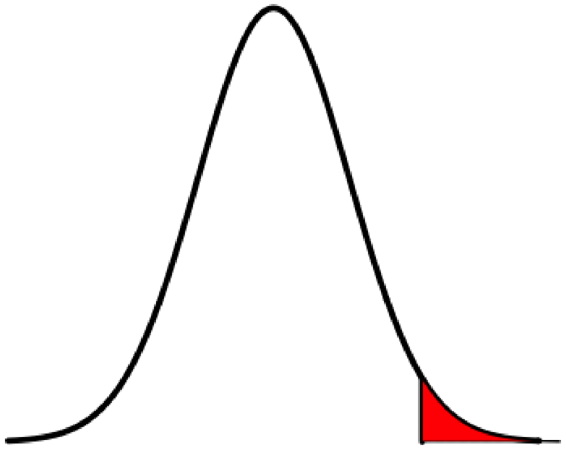
Population-based definition

$$\begin{aligned} Y_1 &= G_1 + E_1 \\ Y_2 &= G_2 + E_2 \end{aligned} \Rightarrow r_g = \mathrm{corr}(G_1, G_2)$$

(Co)variance-based definition

$$\mathrm{corr}\left(Y_{i1,} Y_{j2}\right) = r_g \sqrt{h_1^2 h_2^2 R_{ij}} + \mathrm{Residual}$$

Heritability of binary traits



Phenotypic liability

A binary trait (e.g., a disease) can be modelled as an extreme form of a certain liability (liability threshold model).

Obesity = BMI > 30 kg/m^2

Type 2 diabetes = Fasting glucose > 7mmol/L.

Therefore, heritability is well defined as the **heritability of the continuous liability**.

Falconer (1965): The inheritance of liability to certain diseases, estimated form incidence among relatives.

A few applications

1) The heritability of a trait gives an upper bound for the accuracy of genetic predictors of that trait.

2) The heritability predicts the response to (natural) selection.

3) The heritability predicts an individual’s risk to develop a certain disease knowing they have affected relatives.

4) The heritability influences the statistical power of genome-wide association studies (GWAS)

Part 2 will address

Definition of genetic relatedness

How to calculate it from SNP data?

How to use it to estimate heritability and genetic correlation.

Concepts and tools for estimating heritability using individual-level data?

Part 2

International Statistical Genetics Workshop

IBG

• Boulder, Colorado •

Outline

- Concepts underlying estimation methods
- Genetic relationship matrices (GRM)

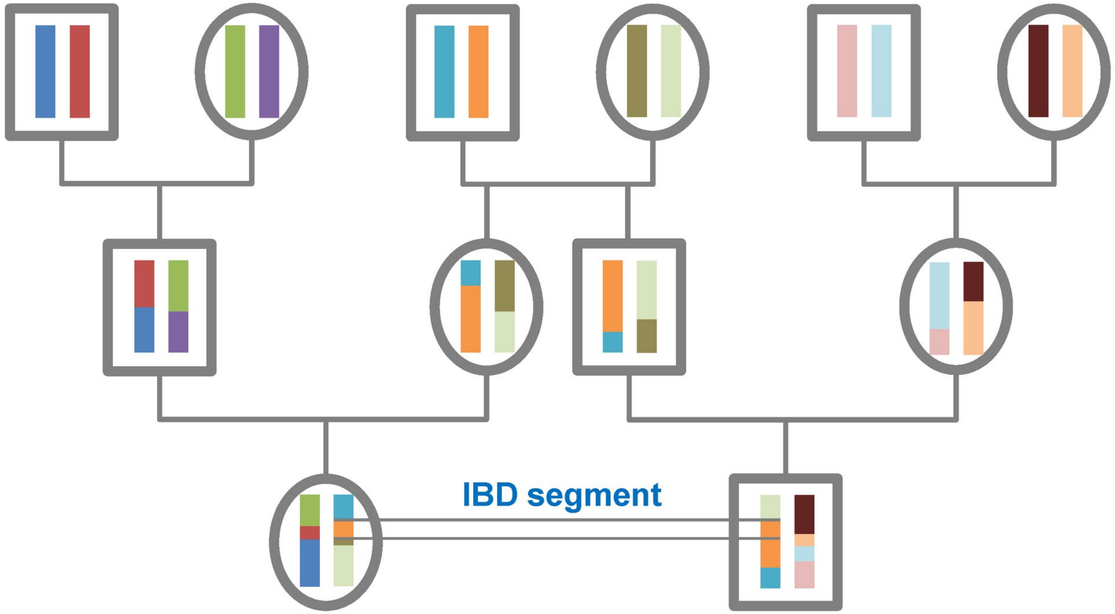
Coefficient of genetic relationship?

In Part 1, we introduced the following Equation:

$$\mathrm{corr}(Y_i, Y_j) = h^2 R_{ij} + \mathrm{Residual}$$

But what is R_{ij}?

R_{ij} is defined as 2x the probability that two alleles picked at random in individual i and individual j are Identical-by-descent (IBD).



Source: Wikipedia

Genetic relatedness is the key

Most methods for estimating heritability (using individual-level data) are based upon this fundamental theorem of QG.

$$\mathrm{corr}(Y_i, Y_j) = h^2 R_{ij} + \mathrm{Residual}$$

Methods differs in the ways they utilize they combine information from R_{ij} and that from $\mathrm{corr}(Y_i, Y_j)$.

Genetic Relation Matrix (GRM)

A GRM is a matrix (of dimension say n,m), which entries are coefficients of genetic relationship.

GRMs can be quantified based expected IBD sharing (e.g., 0.5 for fullsibs) or using actual SNP data.

There are many ways to calculate a GRM using SNP data (nearly an infinite number) but we will focus on a standard estimator implemented in the software **GCTA**.

Standard GRM estimator

$$\hat{\pi}_{jk} = \frac{1}{m}\sum_i \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}$$

where, x_{ij} and x_{ik} are the minor allele count (x_{ij}, x_{ik} = 0,1 or 2) at SNP i for individuals j and k respectively, p_i the minor allele frequency (MAF) of SNP I and m the number of SNPs used to calculate the GRM.

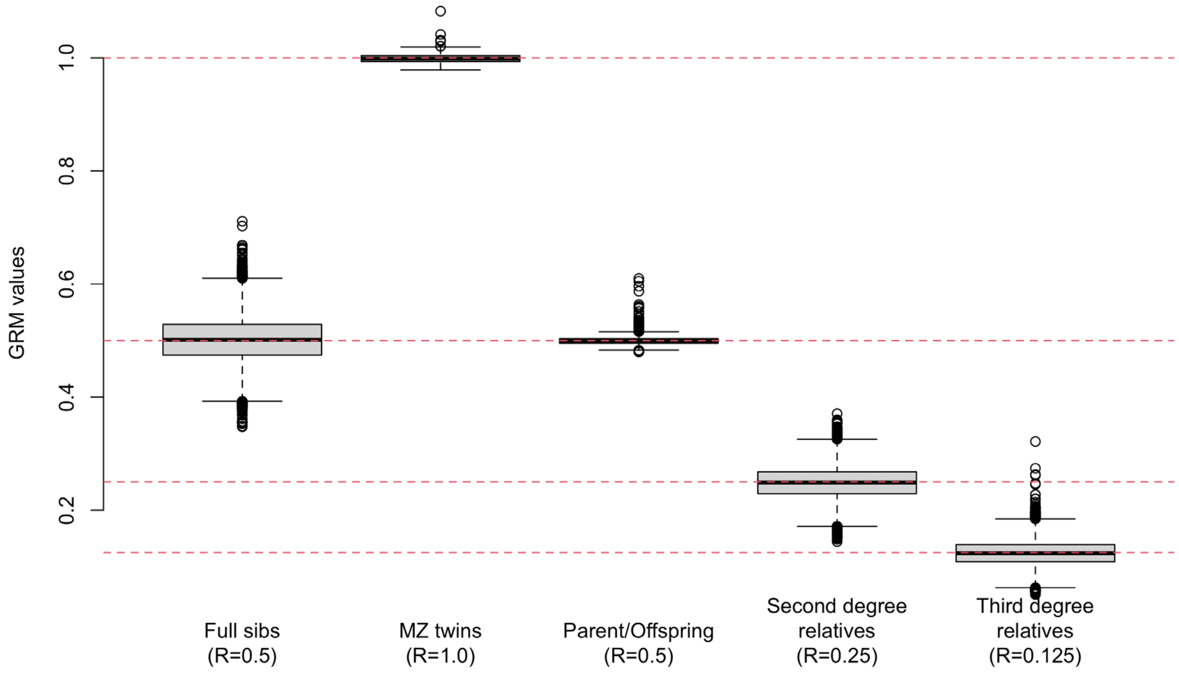
Example of GRM between *N*=3 individuals (over m=1000 SNPs)

```
[$bash] zless myGRM.grm.gz
1 1 1000  0.99
1 2 1000 -0.01
1 3 1000  0.01
2 2 1000  1.03
2 3 1000  0.03
3 3 1000  1.01
```

Distribution of GRM values

The expectation (over a large sample of relatives) of the $\hat{\pi}_{jk}$ is exactly R_{jk}.

Observed relatedness may be still vary within a type of pedigree relationship.



Data from UK Biobank participants
(Application number 12505)

Summary and next part

Observing simultaneously corr(Y_i,Y_j) and R_{ij} is key to estimating h^2

GRMs can be quantified using actual SNP data (show more variation than expected genetic relatedness)

Software **GCTA** can calculate GRM and use them for estimating h^2 (Part 3)

Methods for estimating heritability using individual-level data?

Part 3

International Statistical Genetics Workshop

IBG

• Boulder, Colorado •

Outline

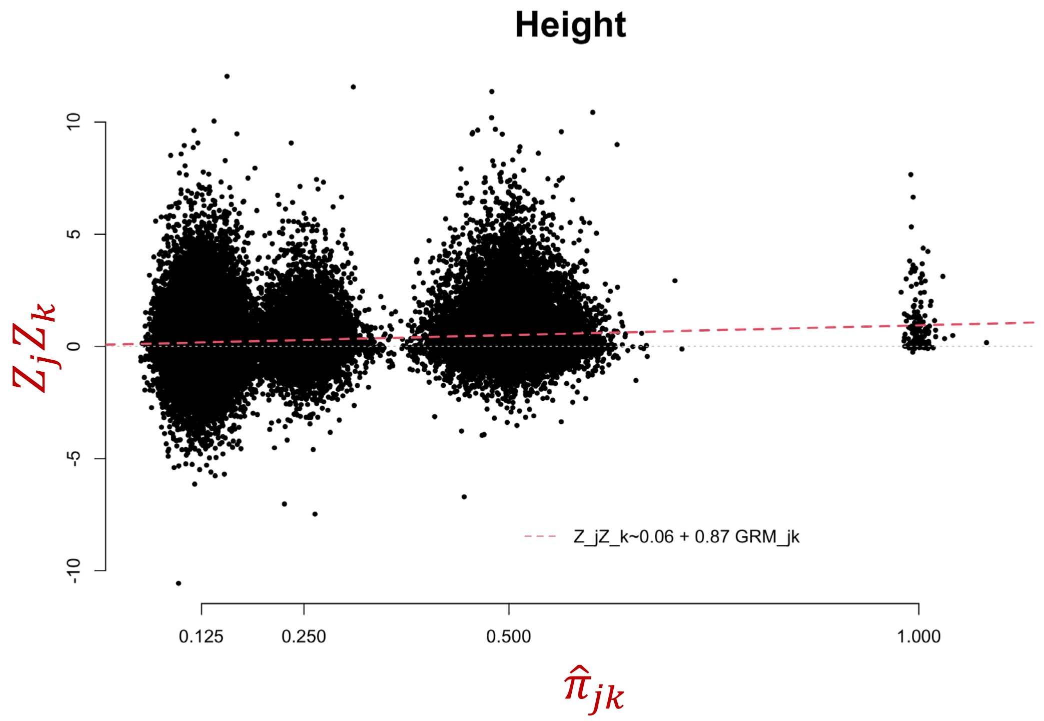
- Estimation using Haseman-Elston (HE) regression
- Estimation using Genome-based REstricted Maximum Likelihood (GREML)

Haseman-Elston (HE) regression

HE regression estimates h^2 by regressing $Z_j Z_k$ onto $\hat{\pi}_{jk}$,

where $Z_j = (Y_j - \text{mean}(Y))/\text{sd}(Y)$ and $Z_k = (Y_k - \text{mean}(Y))/\text{sd}(Y)$, i.e.

$$E[Z_j Z_k] = \text{corr}(Y_j, Y_k).$$



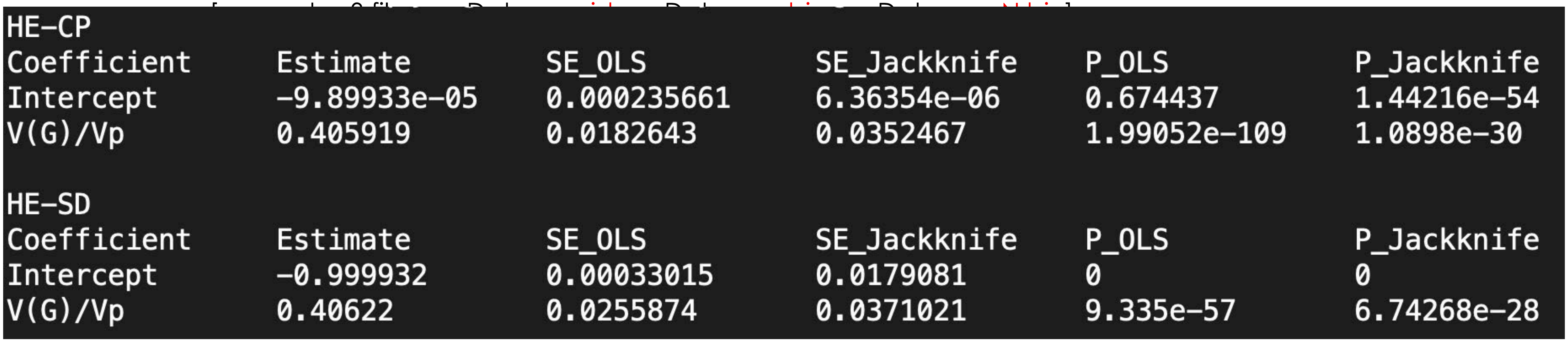
Data from UK Biobank participants
(Application number 12505)

$$E[Z_j Z_k | \hat{\pi}_{jk}] = 0.06 + 0.87\ \hat{\pi}_{jk} \Rightarrow \hat{h}^2_{HE} \sim 0.87.$$

HE regression with GCTA

Step 1: Calculate the GRM

gcta64 --bfile myDataInPLINKformat --make-grm-bin --out myData



```
HE-CP
Coefficient     Estimate        SE_OLS          SE_Jackknife    P_OLS           P_Jackknife
Intercept       -9.89933e-05    0.000235661     6.36354e-06     0.674437        1.44216e-54
V(G)/Vp         0.405919        0.0182643       0.0352467       1.99052e-109    1.0898e-30

HE-SD
Coefficient     Estimate        SE_OLS          SE_Jackknife    P_OLS           P_Jackknife
Intercept       -0.999932       0.00033015      0.0179081       0               0
V(G)/Vp         0.40622         0.0255874       0.0371021       9.335e-57       6.74268e-28
```

Step 3: Run GCTA to estimate heritability of trait 1 using HE regression

gcta64 --grm myData --pheno phenotype.txt --mpheno 1 --HEreg --out myHE_estimates

[generates 2 files: myHE_estimates.log, myHE_estimates.HEreg]

Genome-based Restricted Maximum Likelihood (GREML)

The Model: the “G” part in GREML

The Estimation: the “REML” in GREML.

The “G” part in GREML…

Recall from Part 1: $Y = g + e$ => $h^2 = \text{var}(g)/[\text{var}(g) + \text{var}(e)]$.

=> $h^2 = \sigma_g^2/[\sigma_g^2 + \sigma_e^2]$

GREML estimates h^2 by modelling g as a linear function of SNPs:

$$g_j = \sum_{i=1}^{m} \left(\underset{\underset{\text{Scaled minor allele count}}{\uparrow}}{\frac{(x_{ij} - 2p_i)}{\sqrt{2p_i(1-p_i)}}} \right) \times \underset{\underset{\text{Effect of SNP i}}{\uparrow}}{u_i} = \sum_{i=1}^{m} z_{ij} \times u_i$$

Under Hardy-Weinberg Equilibrium: $\text{var}(z_{ij}) = 1$.

The "G" part in GREML…

Challenge

with the "$g_j = \sum_{i=1}^{m} z_{ij} \times u_i$" model is that the number of SNPs (m) is often (if not always) much larger than the sample size (n).

Therefore: we cannot estimate each SNP effect u_i independently.

Solution

Assume that u_i's are independent, normally distributed random variables, with mean 0 and variance equal to $\mathrm{var}(u_i) = \sigma_g^2/m$.

The “G” part in GREML…

Assuming that u_i’s are normally random variables with mean 0 and variance equal to σ_g^2.

$$\text{var}\left[\boldsymbol{g} = \begin{pmatrix} g_1 \\ \vdots \\ g_n \end{pmatrix}\right] = \sigma_g^2 \begin{pmatrix} \frac{1}{m}\sum_{i=1}^{m} z_{i1} z_{i1} & \cdots & \frac{1}{m}\sum_{i=1}^{m} z_{i1} z_{in} \\ & \underbrace{\frac{1}{m}\sum_{i=1}^{m} z_{ij} z_{ik}}_{\hat{\pi}_{jk}} & \\ & & \ddots \end{pmatrix} = \sigma_g^2\,\mathbf{GRM}$$

The “G” part in GREML…

So if we now assume that the residual terms (the “e” in Y=g+e) are independent normally distributed variables with mean 0 and variance σ_e^2.

Then we can write our final “G” model as

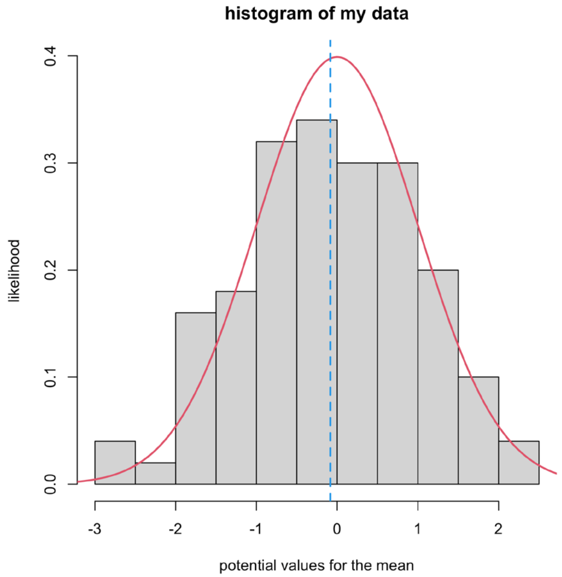
$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim N(\boldsymbol{0}, \sigma_g^2 \mathbf{GRM} + \sigma_e^2 I_n)$$

So if we estimate $\widehat{\sigma_e^2}$ and $\widehat{\sigma_g^2}$, we can deduce an estimator of h^2 as $\hat{h}^2_{\mathrm{GREML}} = \widehat{\sigma_g^2}/(\widehat{\sigma_g^2} + \widehat{\sigma_e^2})$.

The “REML” part in GREML…

REML = Restricted Maximum Likelihood.

Maximum Likelihood.



The "REML" part in GREML…

Why Restricted?

Y = experimental variables + g + e

experimental variables (a.k.a fixed effects)
recruitment centres, genotyping batches, etc.

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sigma_g^2 \mathbf{GRM} + \sigma_e^2 I_n)$$

This extended model is known as a **Linear Mixed Model (LMM)**

REML = ML on transformed data

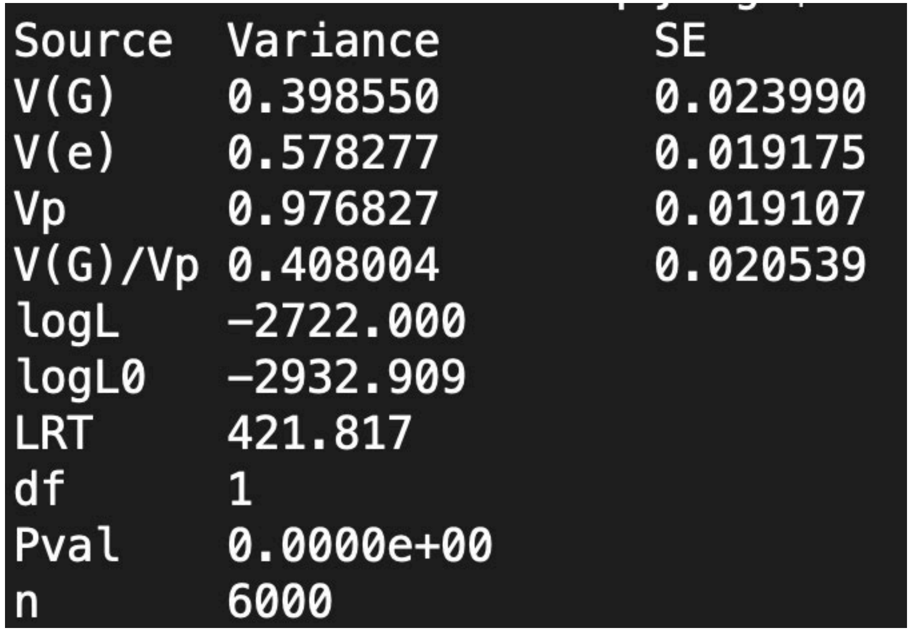
where the fixed effects are residualized.

GREML estimation with GCTA

Run GCTA to estimate heritability of trait 1 using GREML

```
gcta64 --grm myData --pheno phenotype.txt --mpheno 1 --reml --out myGREML_estimates
```

[generates 2 files: myGREML_estimates.log, myGREML_estimates.hsq]



```
Source  Variance        SE
V(G)    0.398550        0.023990
V(e)    0.578277        0.019175
Vp      0.976827        0.019107
V(G)/Vp 0.408004        0.020539
logL    -2722.000
logL0   -2932.909
LRT     421.817
df      1
Pval    0.0000e+00
n       6000
```

Summary and next part

We introduced two ways to estimate h^2 in sample of individuals knowing their <u>phenotypes</u> and the <u>GRM</u>.

1) HE regression
2) GREML model

Next part will discuss

1) interpretation of h^2 estimated from SNPs.
2) some of the biases affecting these methods.

Interpretation of heritability estimates from SNPs

Part 4

Outline

- Missing heritability

- Biases in heritability estimation

Missing heritability

The case of the missing heritability

Relatedness varies across the genome

Therefore, if SNPs used to estimate heritability do not reflect well genetic resemblance at **causal variants**, then we have a "bias".

$$h^2_{\mathrm{GWAS}} \leq h^2_{\mathrm{SNP}} \leq h^2$$

h^2-h^2_{GWAS} is often denoted the "missing" heritability (e.g., 5% vs 80%).
h^2_{SNP}-h^2_{GWAS} is often denoted the "hidden/hiding" heritability.
h^2-h^2_{SNP} is denoted the (still) missing heritability.

Biases in heritability estimates

1) Shared environmental effects.

2) Population stratification

3) Distribution of MAF and linkage disequilibrium (LD) of SNPs used to calculate GRMs

Shared environmental effects

Phenotypic resemblance between relatives is not all due to genetic factors.

How much shared non-genetic factors vary with genetic relatedness is unknown.

Diagnosis: use different types of relatives in the inference and compare estimates.

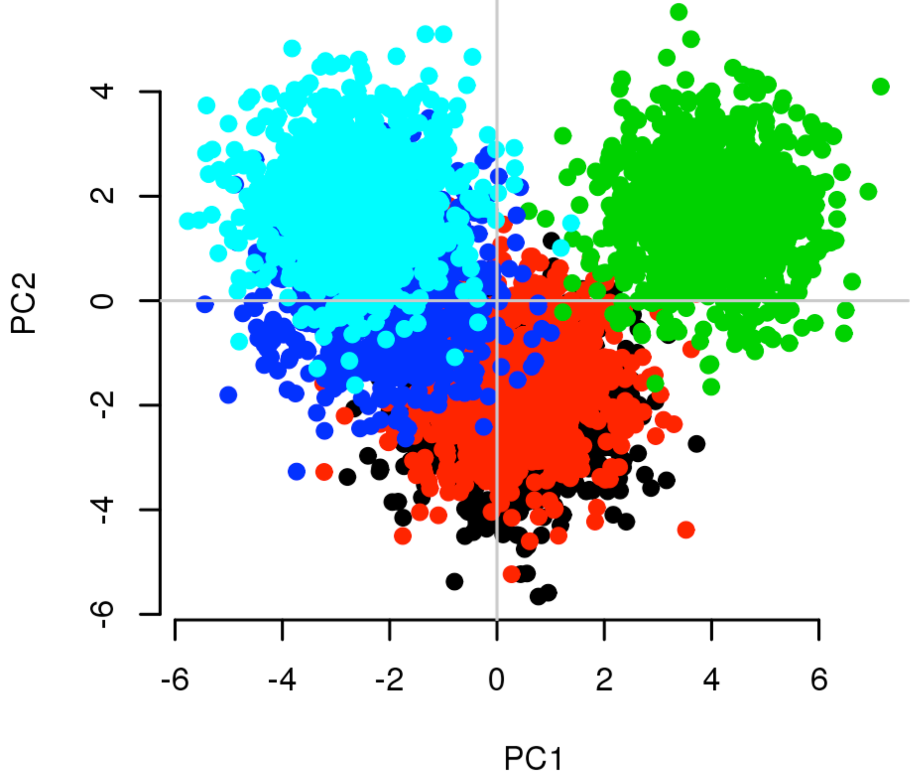
Solution 1: model (i.e. make assumptions about) shared environmental effects. E.g., $h^2 = 2(\text{corrMZ}-\text{corrDZ})$.

Solution 2: only analyse distance relatives GRM<0.05 (or <0.025)

Population stratification

Solution 1: If ancestry differences, then restrict analyses to ancestry homogeneous samples.

Solution 2: Use genetic principal components as fixed effects.



MAF and LD heterogeneity - issue

Speed et al. (2013) AJHG 91(6):1011-21 showed that estimation of h^2_{SNP} is robust to many violations of assumptions.

Strongest biases observed when there is a difference (heterogeneity) between the MAF and LD distribution of causal variants and that of SNPs used to calculate the GRM.

MAF and LD heterogeneity - solution

Solution 1: model the relationship between SNP effects (u_i in Part 3) and minor allele frequency and LD

(Speed et al. 2012, 2017; Zeng et al. 2018, etc.).

Solution 2 (LDMS): minimise heterogeneity by stratifying SNPs based on MAF and LD distribution (Yang et al. 2015, Evans et al. 2018).

LDMS method

Step 1: Calculate SNP attributes: MAF (p_i) and LD score (ℓ_i)

$\ell_i = \sum_{k=1}^{m} r_{ik}^2$ (r_{ik}^2: squared correlation of allele counts between SNP i and SNP k)

Step 2: Groups SNPs based on MAF and LD

e.g., 6 MAF groups:]1%-5%],]5%-10%],]10%-20%],]20%-30%],]30%-40%] and]40%-50%]
+ 4 LD score groups (quartile) with each MAF group => K=6x4=24 LDMS groups.

Step 3: Calculate a GRM for each group of SNPs.

Step 4: Estimate jointly the "σ_g^2" for each SNP group.

$$\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \sum_{k=1}^{K} \sigma_{g,k}^2 \mathbf{GRM}_k + \sigma_e^2 I_n) \Rightarrow \sigma_g^2 = \sum_{k=1}^{K} \sigma_{g,k}^2$$

Summary

Shared environmental effects and population stratification can bias heritability estimates (fix: PC + unrelated individuals)

MAF and LD distribution of SNPs can bias estimation of h^2_{SNP} (many solutions: in particular LDMS).

Examples of active research in heritability estimation

Part 5

What are people researching in this area (individual-level data)?

- Computational efficiency (biobank scale methods)
- Incorporate biological information (functional annotation)
- Estimation of non-additive genetic variance (e.g., dominance)
- Estimation of SNP-based heritability using whole-genome sequence data (integrating rare and ultra-rare variants)
- The impact of assortative mating on heritability estimates