Estimation & interpretation of genetic variance

UQ Winter School, June 2022

Dr Kathryn Kemper

(with thanks to Loic, JZ and Jian Yang)

International Statistical Genetics Workshop • Boulder, Colorado •

IBG

Pruned and modified version of 5-part workshop given by Dr Loic Yengo@ISGW

“Heritability of individual level data”

e.g. https://www.youtube.com/watch?v=Cjn5AtNPjzE

Outline

- Definition of heritability
- Estimation
  - Relationship matrices
  - HE regression
  - REML
- Interpreting h^2 estimates

Definition

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

Definition

The value of the trait, or phenotype (P), can be modelled as

$$P = \underbrace{A}_{\text{(additive) genetic factors}} + \underbrace{E}_{\text{Non-genetic factors}}$$

then $h^2 = \frac{\sigma_A^2}{\sigma_P^2}$

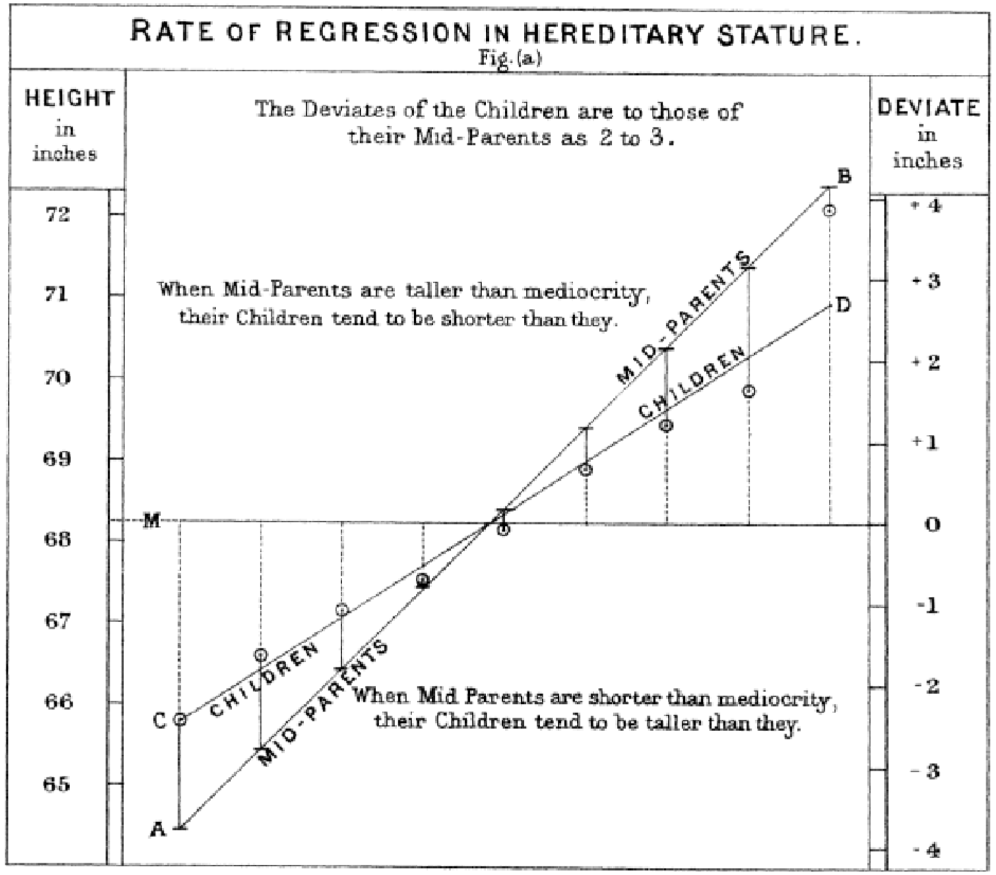
the heritability is the proportion of phenotypic variance (σ_P^2) attributable to additive genetic effects (σ_A^2)

Heritability ranges between 0 and 1

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

Definition

- How can we estimate ‘A’ when we can’t observe the true genetic effects?



Galton (1886)

What are ‘average’ relationships?

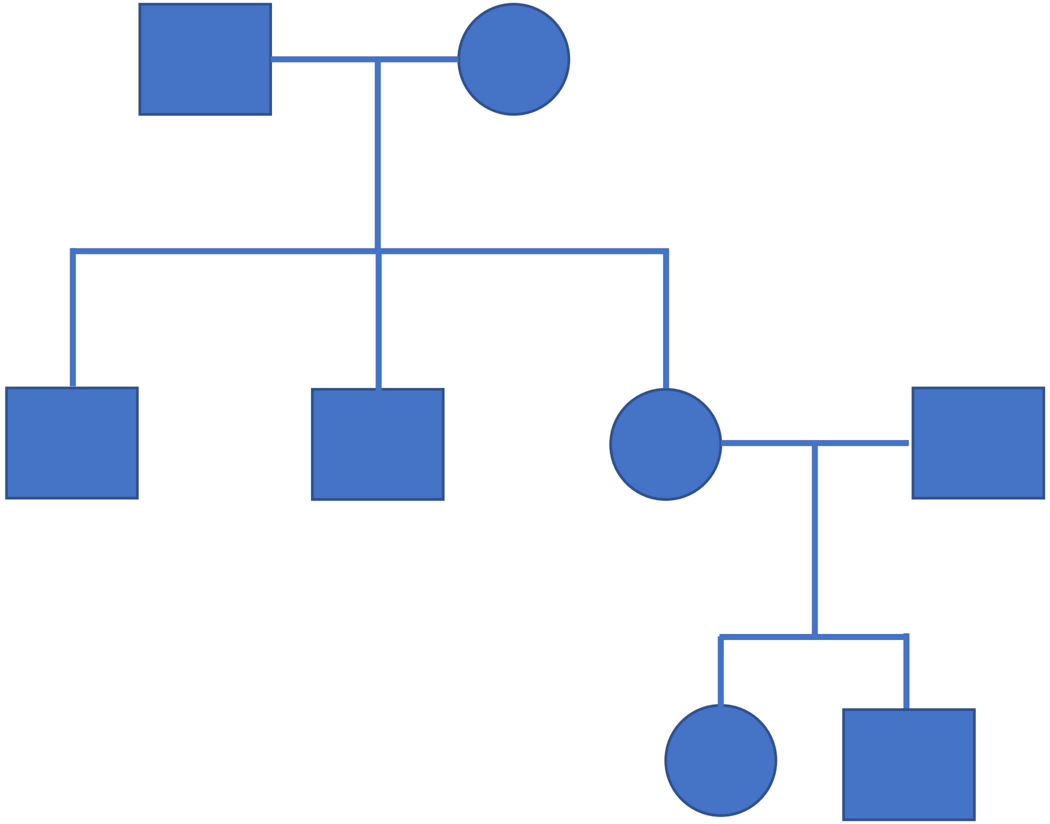
- Animal and plant breeders have used ‘average’ or pedigree relationships since 1950’s to drive genetic change

- Human geneticists have mostly relied on comparing MZ and DZ twins

- These approaches rely on the average genetic relationship between relatives

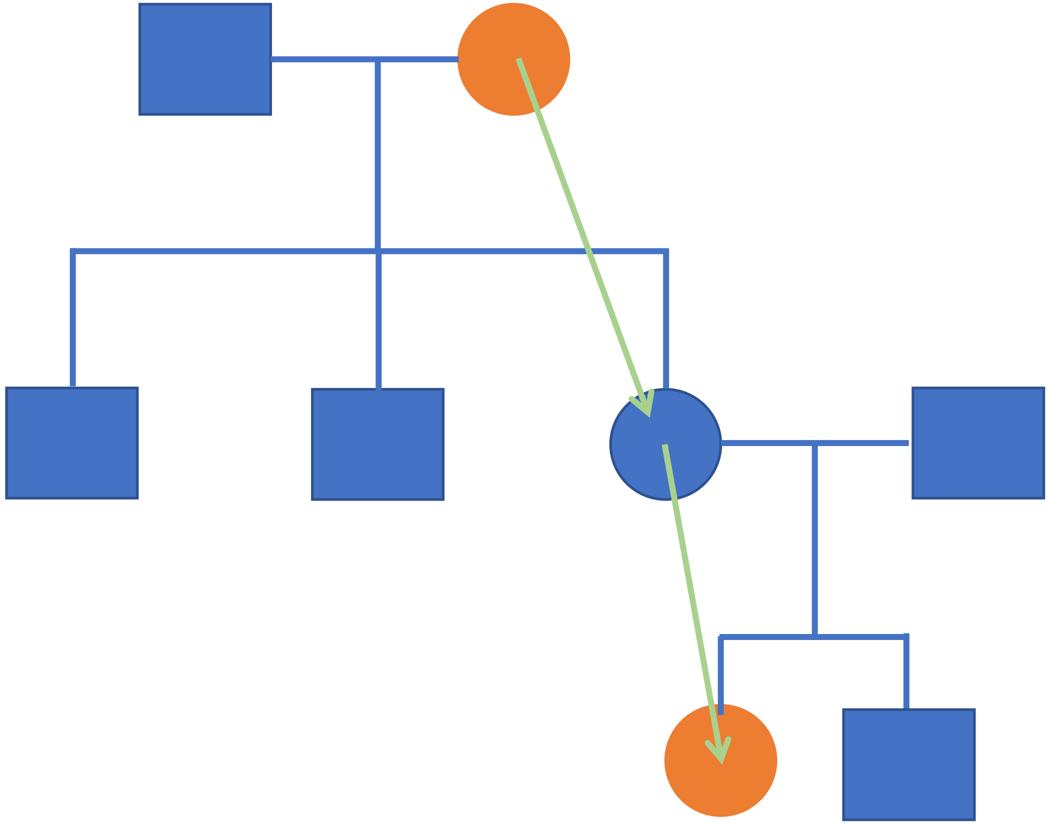
What are ‘average’ relationships?

Relationship	Relationship co-efficient (π)
MZ twins	1.0 = 0.5^0
Parent-child	0.5 = 0.5^1
Full-sibs/DZ twins	0.5 = $2x0.5^2$
Half-sibs	$0.25 = 0.5^2$
Avuncular	$0.25 = 2x0.5^3$
Grandparent-child	$0.25 = 0.5^2$
1st cousins	$0.125 = 2x0.5^4$



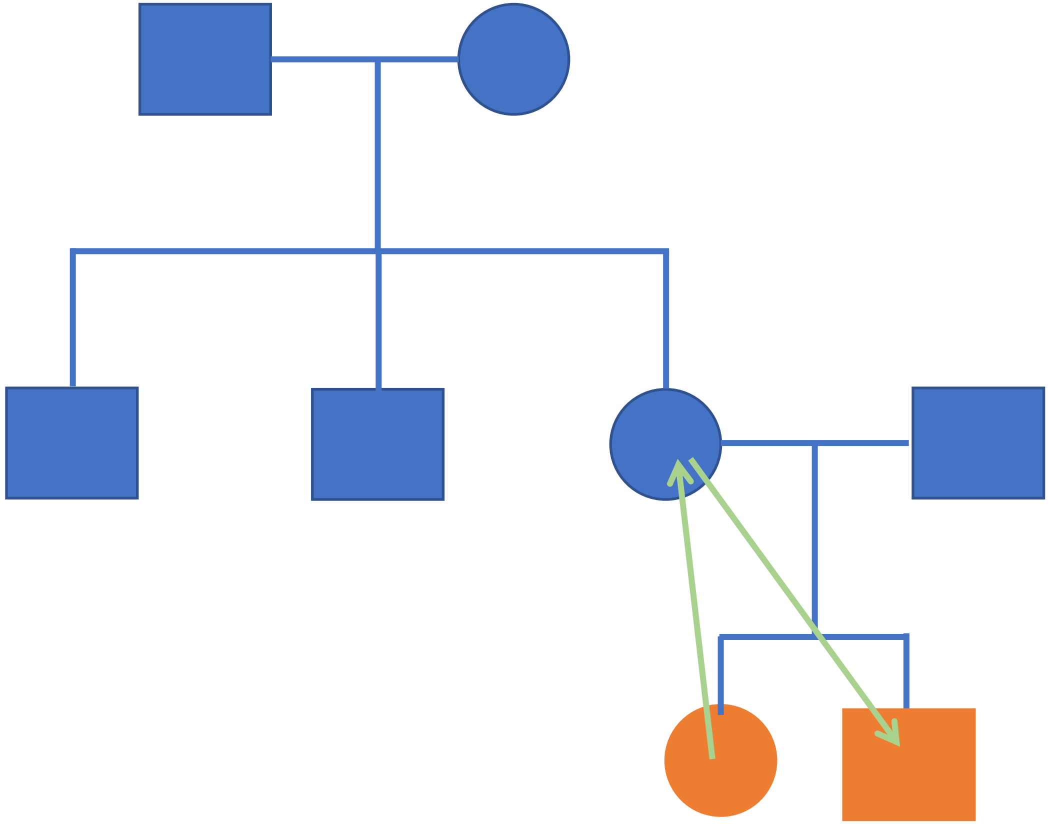
What are ‘average’ relationships?

Relationship	Relationship co-efficient (π)
MZ twins	1.0 = 0.5^0
Parent-child	0.5 = 0.5^1
Full-sibs/DZ twins	0.5 = $2x0.5^2$
Half-sibs	$0.25 = 0.5^2$
Avuncular	$0.25 = 2x0.5^3$
Grandparent-child	$0.25 = 0.5^2$
1st cousins	$0.125 = 2x0.5^4$



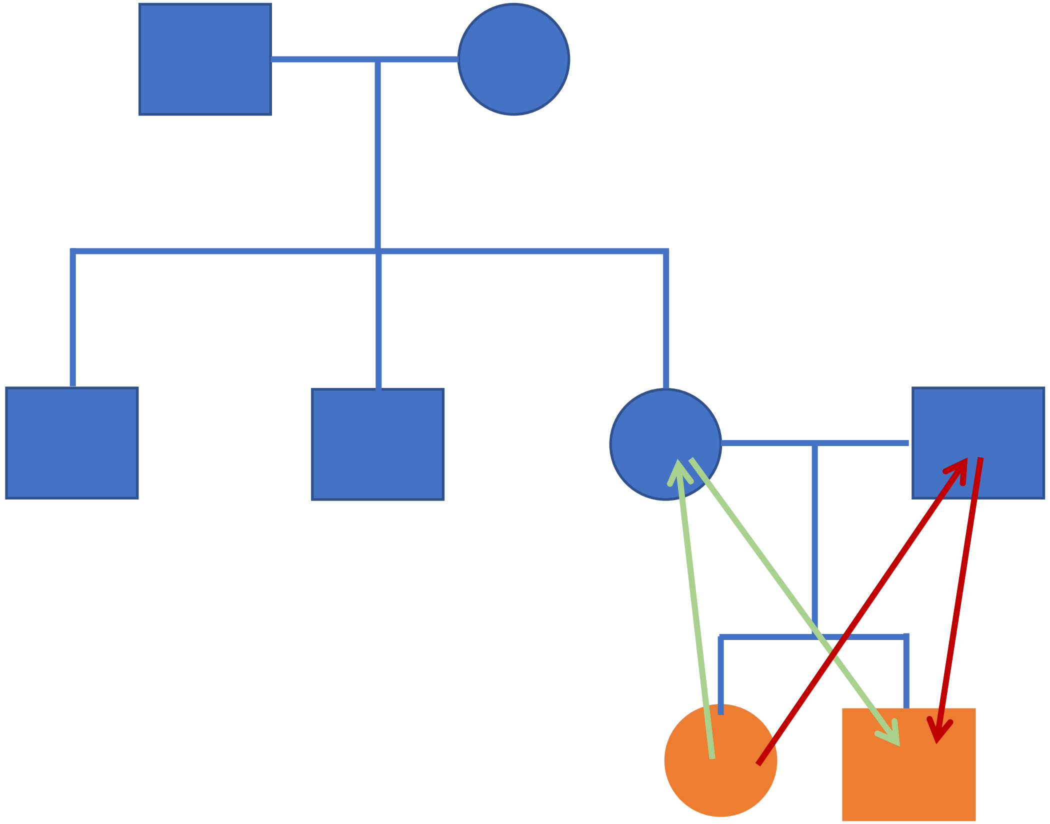
What are ‘average’ relationships?

Relationship	Relationship co-efficient (π)
MZ twins	1.0 = 0.5^0
Parent-child	0.5 = 0.5^1
Full-sibs/DZ twins	0.5 = $2x0.5^2$
Half-sibs	$0.25 = 0.5^2$
Avuncular	$0.25 = 2x0.5^3$
Grandparent-child	$0.25 = 0.5^2$
1st cousins	$0.125 = 2x0.5^4$



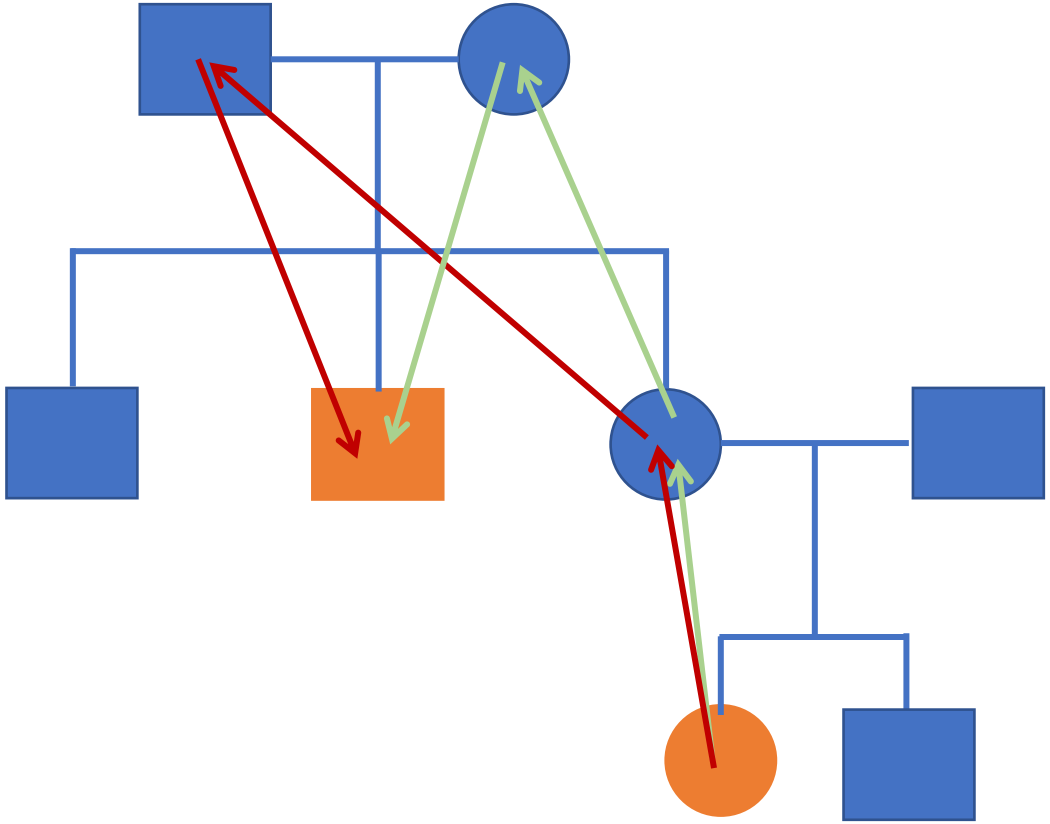
What are ‘average’ relationships?

Relationship	Relationship co-efficient (π)
MZ twins	1.0 = 0.5^0
Parent-child	0.5 = 0.5^1
Full-sibs/DZ twins	0.5 = $2x0.5^2$
Half-sibs	$0.25 = 0.5^2$
Avuncular	$0.25 = 2x0.5^3$
Grandparent-child	$0.25 = 0.5^2$
1st cousins	$0.125 = 2x0.5^4$



What are ‘average’ relationships?

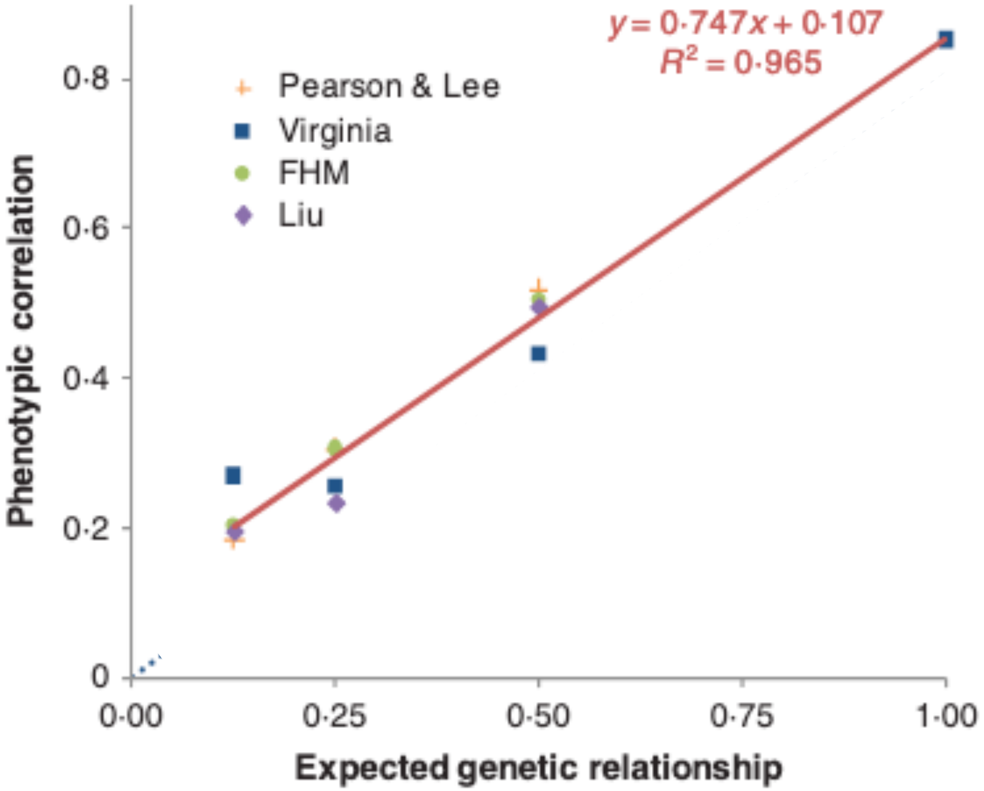
Relationship	Relationship co-efficient (π)
MZ twins	1.0 = 0.5^0
Parent-child	0.5 = 0.5^1
Full-sibs/DZ twins	0.5 = $2x0.5^2$
Half-sibs	$0.25 = 0.5^2$
Avuncular	$0.25 = 2x0.5^3$
Grandparent-child	$0.25 = 0.5^2$
1st cousins	$0.125 = 2x0.5^4$



Definition

We can estimate the heritability of a trait using average relationships, e.g.

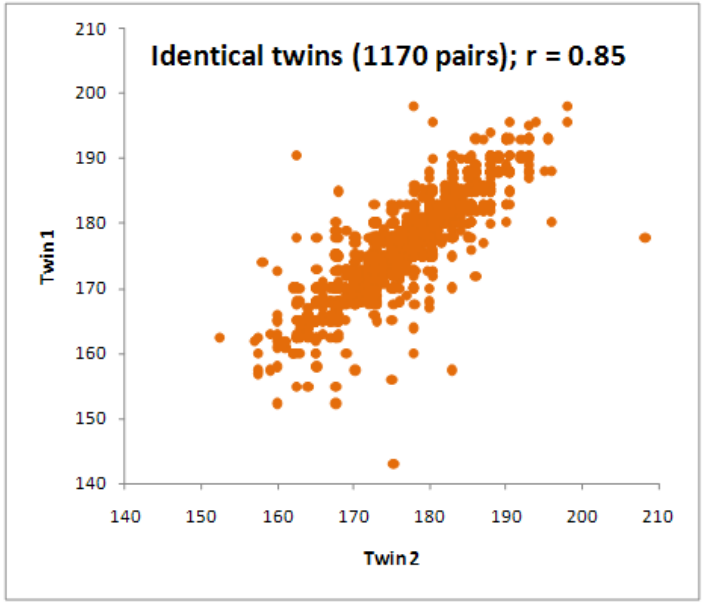
$$\mathrm{corr}(Y_i,Y_j) = h^2\boldsymbol{\pi}_{ij} + \text{residual}$$

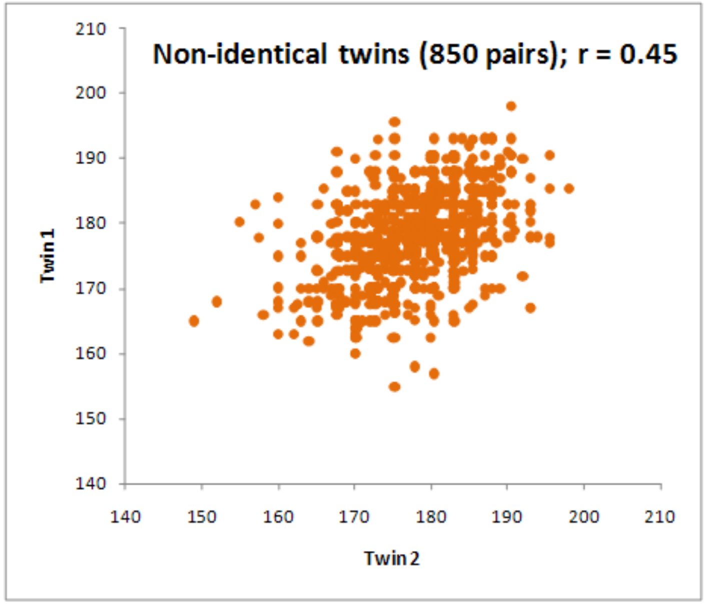


Visscher, McEvoy & Yang (2010) *Genet. Res.* **92**:371-379.

Definition

We can estimate the heritability of a trait using average relationships, e.g.


Identical twins (1170 pairs); r = 0.85


Non-identical twins (850 pairs); r = 0.45

<u>Twin-based estimate heritability:</u>

r(MZ) = [var(G) + var(E)]/ var(P)

r(DZ) = [0.5 var(G) + var(E)] / var(P)

2 [r(MZ) – r(DZ)] = 2[0.5.var(G)] / var(P)

= h^2

Why all the fuss about h^2?

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/artificial) 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)

Misconceptions about heritability

Heritability is an estimate, based on many assumptions. Beware.

Heritability is a property of a trait, in a population, at a given time. It is not fixed.

A low heritability does not necessarily mean the trait is not genetically determined.

- It suggests that non-genetic factors account for more variation.
- Is there phenotypic variation? (e.g. number of fingers)

Methods to estimate heritability

$$\boxed{\text{corr}(Y_i,Y_j) = h^2\boldsymbol{\pi}_{ij} + \text{residual}}$$

Covariance between ‘relatives’ is fundamental to h^2 estimation.

Methods differ in the approaches to combine corr(Y_i,Y_j) and $\boldsymbol{\pi}_{ij}$, e.g.

- ➢we can estimation relationships using pedigrees or genetic markers
- ➢we can use a regression, ANOVA or REML framework for parameter estimates

Relationship matrices

A relationship matrix (of dimension nxn, where n = number of individuals) is a square symmetrical matrix where each element defines the relationship between two individuals.

e.g. for pedigree relationships $A_{ij} = 0.5$ for full-sibs i and j

Relationship matrices – average relationship

A relationship matrix (of dimension nxn, where n = number of individuals) is a square symmetrical matrix where each element defines the relationship between two individuals.

e.g.

ID	Mother	Father
1	-	-
2	-	-
3	1	2
4	1	2
5	1	2
6	-	-
7	5	6

Lower triangle A matrix:

1.0						
0	1.0					
0.5	0.5	1.0				
0.5	0.5	0.5	1.0			
0.5	0.5	0.5	0.5	1.0		
0	0	0	0	0	1.0	
0.25	0.25	0.25	0.25	0.5	0.5	1.0

Relationship matrices

A relationship matrix (of dimension nxn, where n = number of individuals) is a square symmetrical matrix where each element defines the relationship between two individuals.

e.g. for pedigree relationships $A_{ij} = 0.5$ for full-sibs i and j

There are many ways to calculate a relationship matrices when using SNP data. We will focus on a standard estimator implemented in the software **GCTA**.

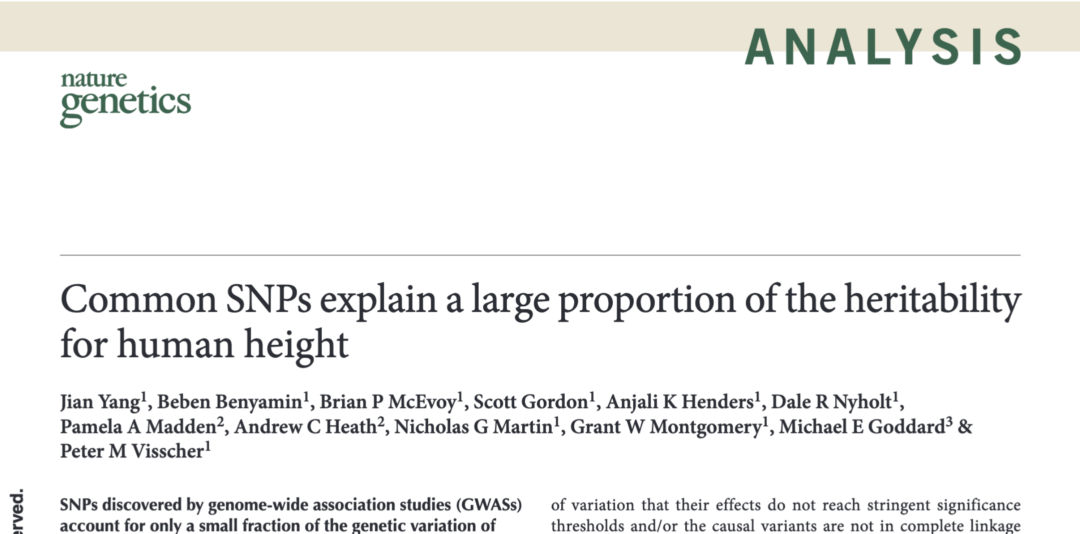
Standard GRM estimator

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

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
```



ANALYSIS

nature genetics

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

Jian Yang[1], Beben Benyamin[1], Brian P McEvoy[1], Scott Gordon[1], Anjali K Henders[1], Dale R Nyholt[1], Pamela A Madden[2], Andrew C Heath[2], Nicholas G Martin[1], Grant W Montgomery[1], Michael E Goddard[3] & Peter M Visscher[1]

SNPs discovered by genome-wide association studies (GWASs) account for only a small fraction of the genetic variation of

of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage

Estimating genetic variance

- Haseman-Elston (HE) regression, ‘method of moments’

$$\boxed{\mathrm{corr}(Y_i,Y_j) = h^2\boldsymbol{\pi}_{ij} + \mathrm{residual}}$$

- ANOVA for balanced designs

- REstricted Maximum Likelihood (REML)

Haseman-Elston (HE) regression

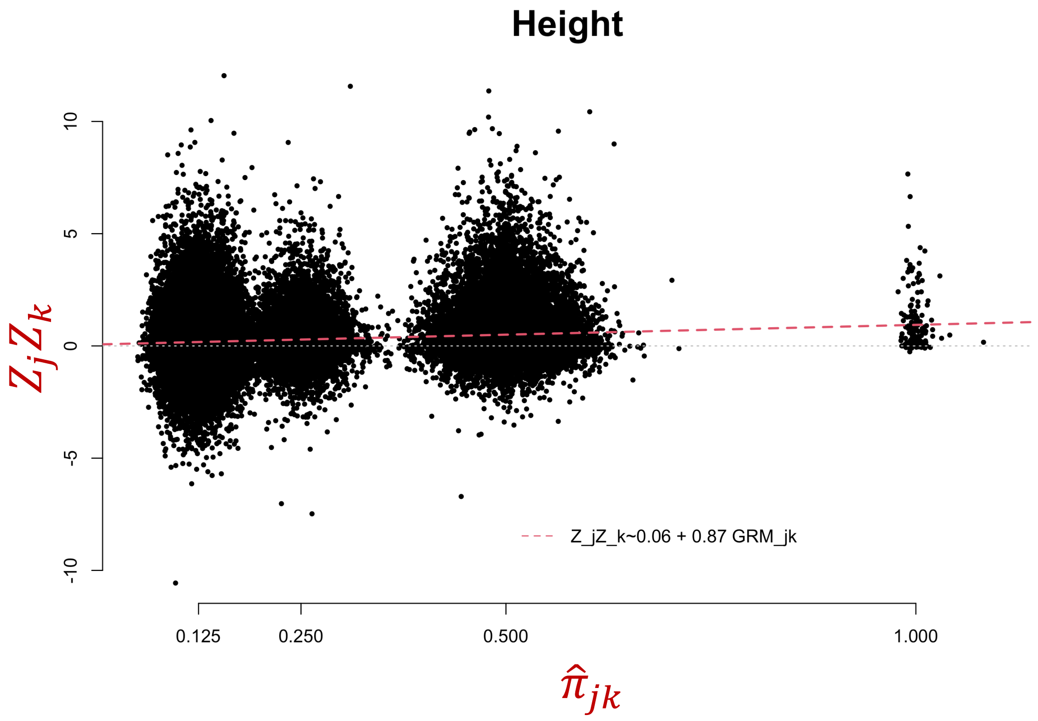
HE regression estimates h^2 by regressing $Z_j Z_k$ onto $\hat{\pi}_{jk}$,

where

$$Z_j = (Y_j - \mathrm{mean}(Y))/\mathrm{sd}(Y), \text{ and}$$
$$Z_k = (Y_k - \mathrm{mean}(Y))/\mathrm{sd}(Y)$$

Thus,

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



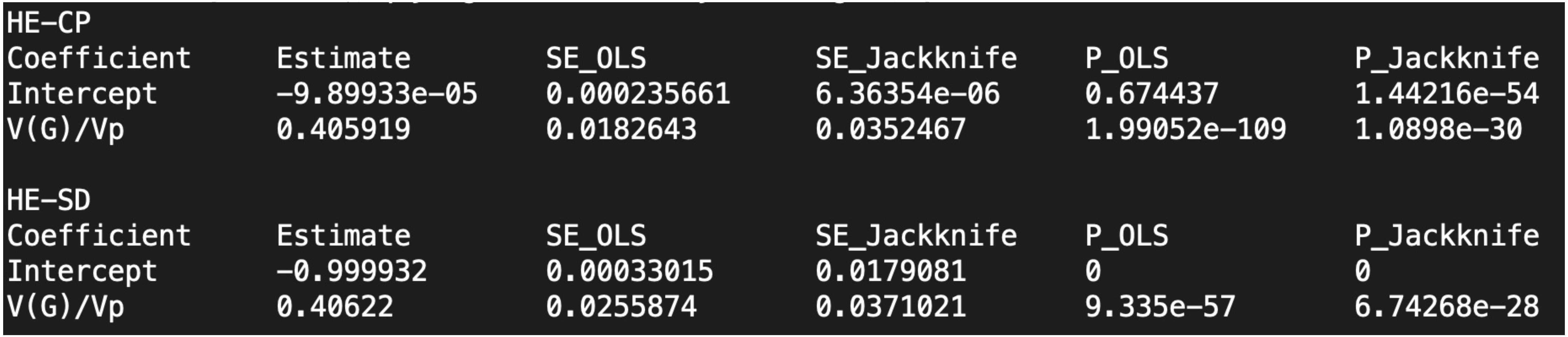
Data from UK Biobank participants
(Application number 12505)

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

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]

REML estimation

- Mixed model: $\boldsymbol{y} = \boldsymbol{X\beta} + \mathbf{Zu} + \mathbf{e}$,
  - where $\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \mathbf{A}\sigma_a^2)$ & $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, I_n\sigma_e^2)$

- Use REstricted Maximum Likelihood to estimate parameters
  - 'restricted' in that we also have fixed effects i.e. $\mathbf{y} \sim N(\boldsymbol{X\beta}, \sigma_a^2\boldsymbol{A} + \sigma_e^2 I_n)$
  - REML accounts for df lost due to estimation of fixed effects

- Specifically in human genetics, often called GREML (Genome-based REML)

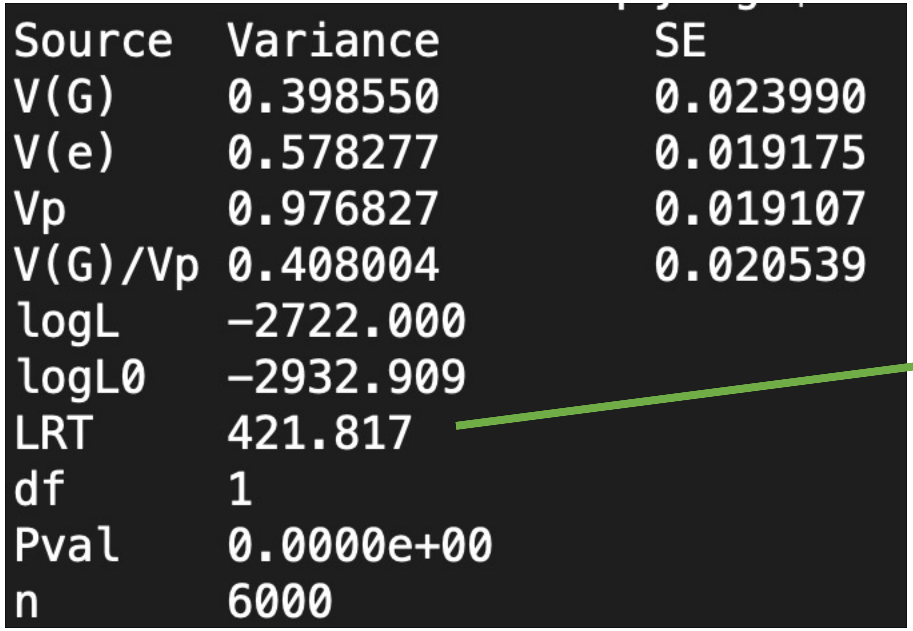
- Genome-based :-
  - using SNPs to determine relationships
  - assume that genetic effects are a linear combination of SNP effects
  - "big-p little-n" problem (where n = samples & p = predictors)

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
```

The significance of h^2_{SNP} is assessed by likelihood ratio test (LRT)

H_0: $h^2_{SNP} = 0$

H_1: $h^2_{SNP} \neq 0$

LRT = $2[L(H_1) – L(H_0)]$ is distributed as a half probability of 0 and a half probability of chi-squared with 1 d.f.

Interpreting h^2 estimates

- h^2 estimates are based on many assumptions, depending on your approach
  - ➢e.g. h^2-SNP depends on the SNP you use!
  - ➢most models assume random mating

- Bias from shared environment
  - ➢can be model explicitly, e.g. twin analysis (still biased?)
  - ➢use only unrelated individuals ($\boldsymbol{\pi}_{ij} < 0.05$)
  - ➢within-family estimates

- G-E confounding – population stratification or parental effects
  - ➢Fit PCs as covariates
  - ➢Within-family analyses

Missing heritability

The case of the missing heritability

Manolio et al. 2009 Nature

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

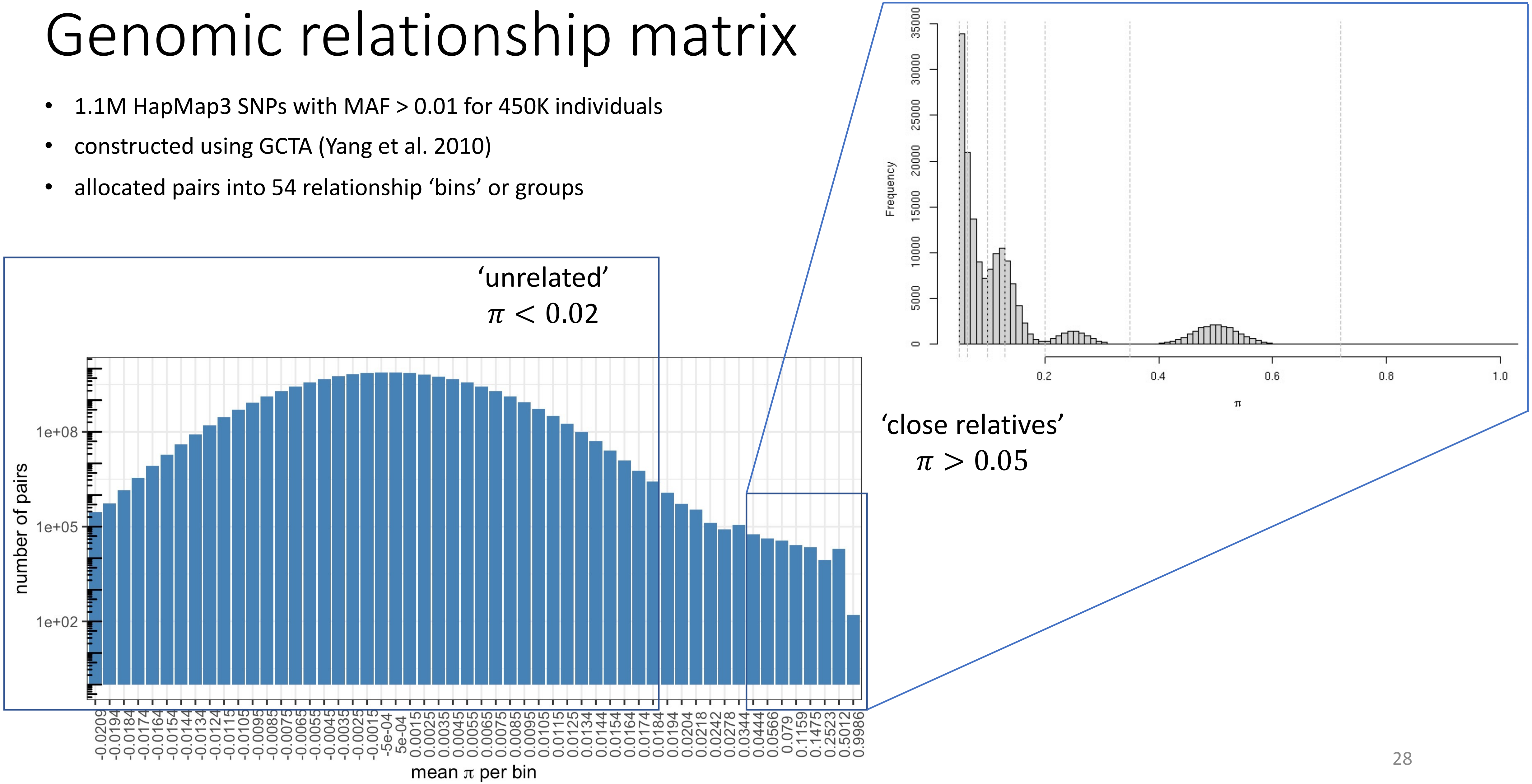
h^2_{PED}-h^2_{GWAS} is often denoted the “missing” heritability (e.g., 5% vs 80%).

h^2_{SNP}-h^2_{GWAS} is often denoted the “hidden/hiding” heritability.

h^2_{PED}-h^2_{SNP} is denoted the (still) missing heritability.

Genomic relationship matrix

- 1.1M HapMap3 SNPs with MAF > 0.01 for 450K individuals
- constructed using GCTA (Yang et al. 2010)
- allocated pairs into 54 relationship ‘bins’ or groups

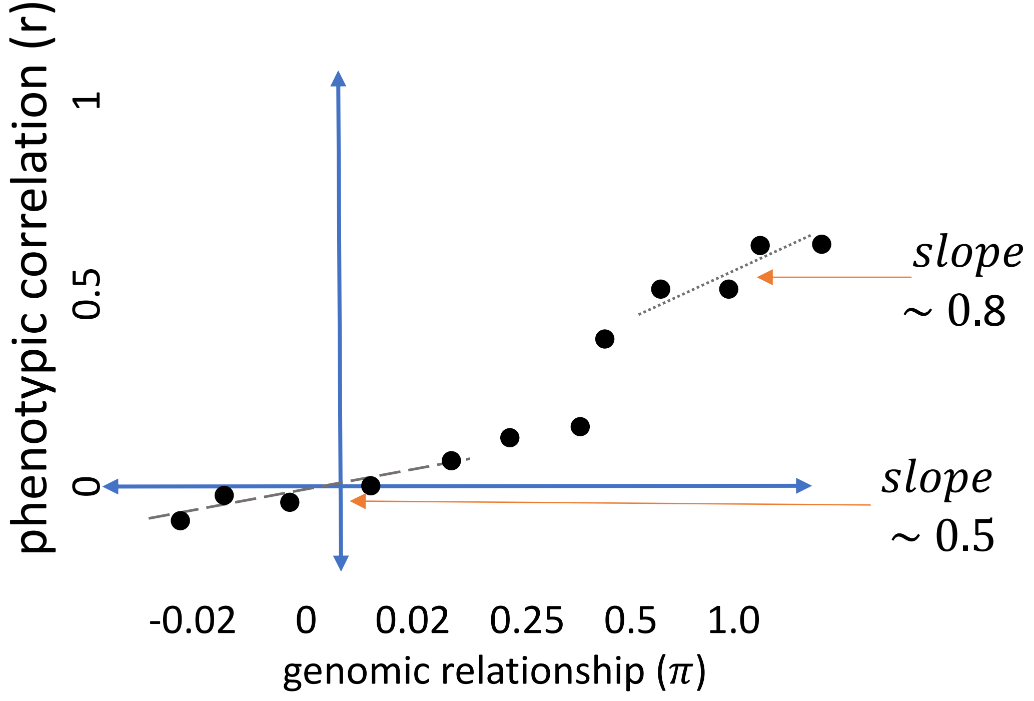


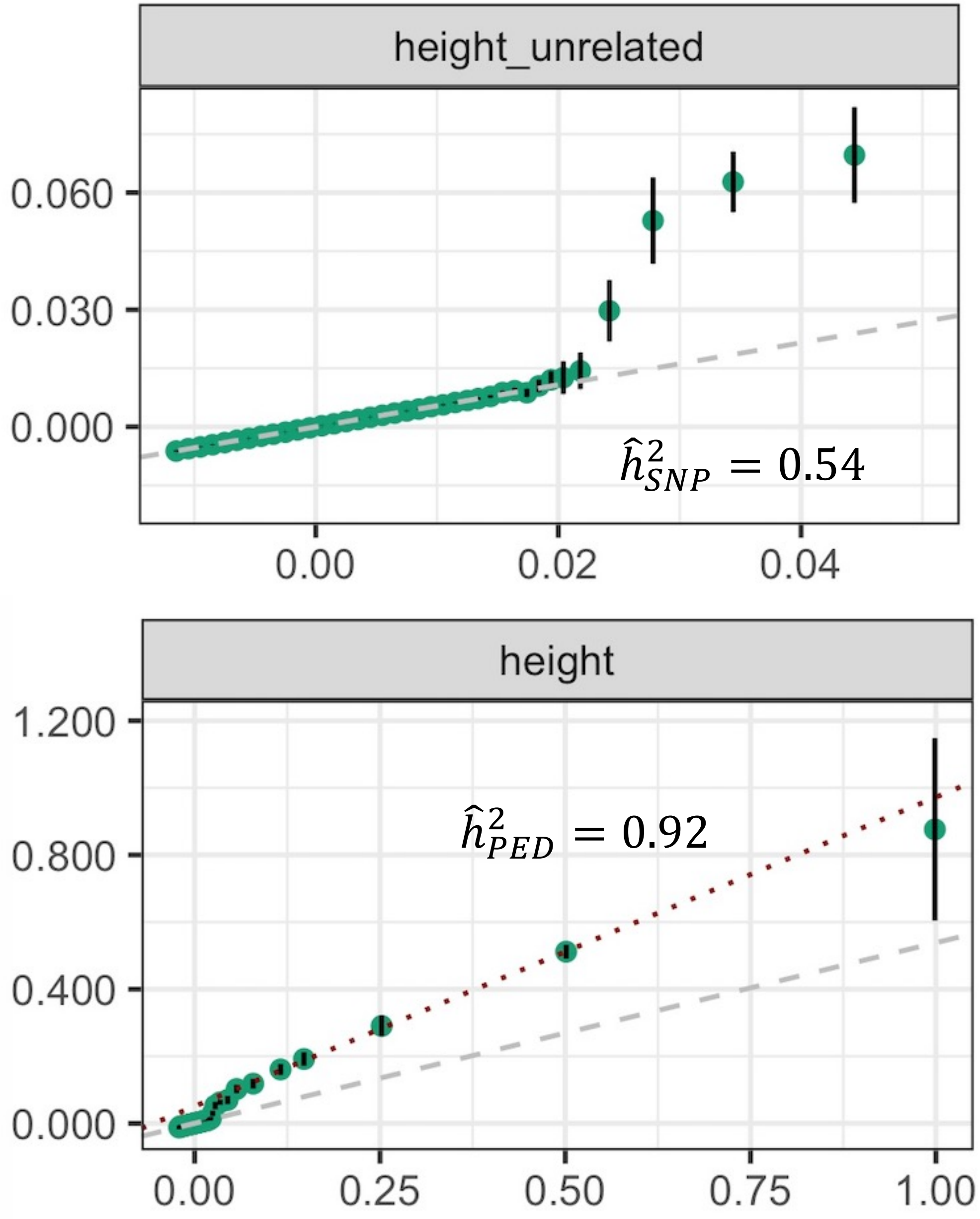
‘unrelated’
$\pi < 0.02$

‘close relatives’
$\pi > 0.05$

Missing heritability? – human height

$$\mathrm{corr}(Y_i,Y_j) = h^2\boldsymbol{\pi}_{ij} + \text{residual}$$

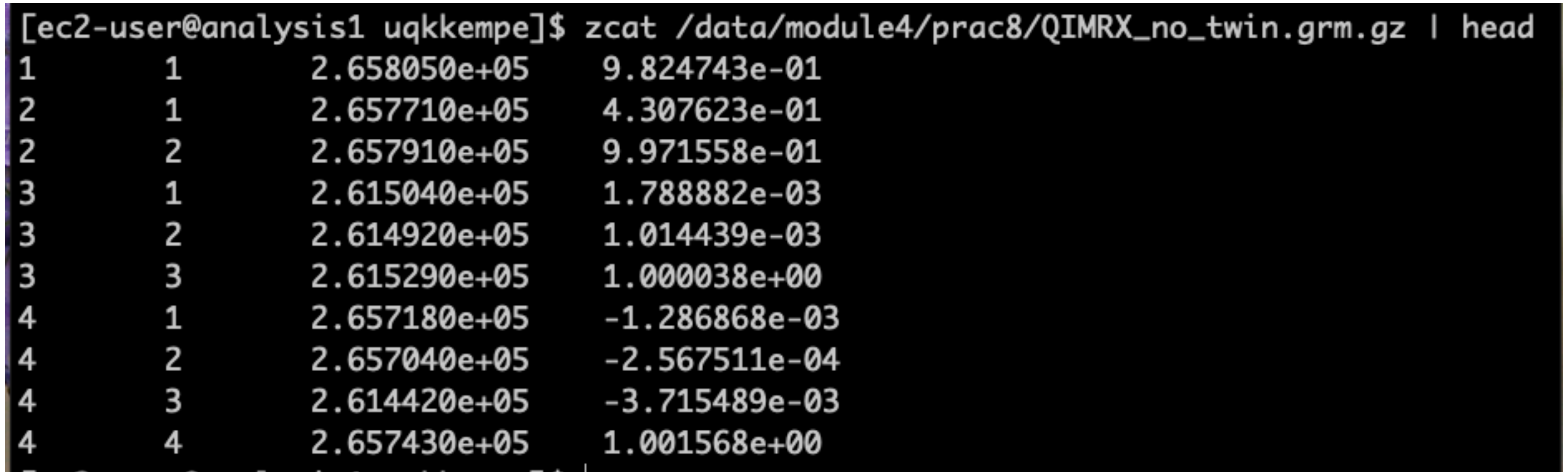




Practical – estimation of h^2 using GCTA

We will loosely follow a practical developed for STAT 3306-7306

- grm is located in /data/module4/prac8/QIMRX_no_twin.grm.gz
- Made with the following code in GCTA, e.g.
    - ➢gcta --bfile <plinkSNPFile> --make-grm.gz --out QIMRX_no_twin
- Look at the ‘gzip’ version of the grm using ‘zcat <file>.grm.gz | head` at the command line



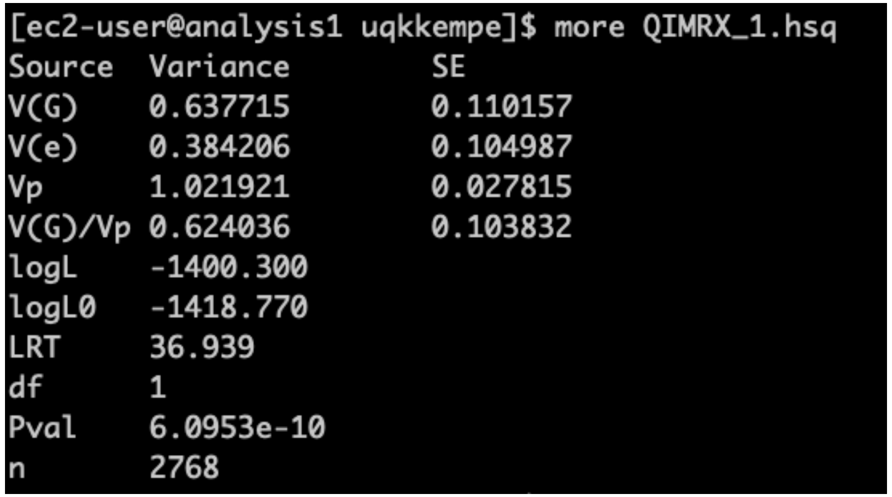
```
[ec2-user@analysis1 uqkkempe]$ zcat /data/module4/prac8/QIMRX_no_twin.grm.gz | head
1       1       2.658050e+05    9.824743e-01
2       1       2.657710e+05    4.307623e-01
2       2       2.657910e+05    9.971558e-01
3       1       2.615040e+05    1.788882e-03
3       2       2.614920e+05    1.014439e-03
3       3       2.615290e+05    1.000038e+00
4       1       2.657180e+05    -1.286868e-03
4       2       2.657040e+05    -2.567511e-04
4       3       2.614420e+05    -3.715489e-03
4       4       2.657430e+05    1.001568e+00
```

Practical – estimation of h^2 using GCTA

- Use GCTA to estimate the SNP-heritability for height, e.g.

```
gcta    --grm /data/module4/prac8/QIMRX_no_twin \
        --pheno /data/module4/prac8/HT_T_X.pheno \
        --mpheno 1 --reml --out QIMRX_1
```

- View the resulting file at the command line using ‘more’



```
[ec2-user@analysis1 uqkkempe]$ more QIMRX_1.hsq
Source  Variance        SE
V(G)    0.637715        0.110157
V(e)    0.384206        0.104987
Vp      1.021921        0.027815
V(G)/Vp 0.624036        0.103832
logL    -1400.300
logL0   -1418.770
LRT     36.939
df      1
Pval    6.0953e-10
n       2768
```

Practical – closer look @ the GRM

- Open R & use
  - x = read.table("<file>.grm.gz") to load the data into R
  - lower triangle GRM; columns are row #, column #, # SNPs, relationship value

- Make a true/false vector if the relationship value is a diagonal
  - diagElement = x[,1] == x[,2]
  - sum(diagElement) ; head(diagElement)

- Plot results
  - hist(x[diagElement,4], breaks=2500, xlab="GRM diagonals")
  - hist(x[!diagElement,4], breaks=2500, xlab="GRM off-diagonals")

Practical – removing relatives

- In the off-diagonal plot there were some large relationships
- In R use “sum(x[!diagElement,4]>0.05)” to find out how many pairs

- Now we are going to use GCTA to remove one member of the close relative pairs

```
gcta --grm /data/module4/prac8/QIMRX_no_twin \
        --grm-cutoff 0.05 --make-grm --out QIMRX_nr
```

- Re-run SNP-heritability estimate with your new pruned matrix

```
gcta --grm QIMRX_nr --pheno /data/module4/prac8/HT_T_X.pheno \
        --mpheno 1 --reml --out QIMRX_nr_1
```

Practical – Haseman-Elston regression

- He regression can be run in GCTA using --Hereg, e.g.

```
gcta --grm QIMRX_nr --pheno /data/module4/prac8/HT_T_X.pheno --mpheno 1 \
    --HEreg --out QIMRX_nr_1b
```

- However today we’re going to to it by-hand, in R!

- Open R:

```
Hereg = read.table("/data/module4/prac8/he.grm.txt")
names(HEreg) <- c("ID1", "ID2", "SNPs", "REL","PROD")
```

- Do the regression and plot:

```
lm1 = lm(HEreg$PROD~HEreg$REL)
png(file="HEreg_all.png")
plot(HEreg$PROD~HEreg$REL,pch=".")
abline(lm1,col="orange",lwd=1.5)
dev.off()
```

$$\boxed{\mathrm{corr}(Y_i,Y_j) = h^2\boldsymbol{\pi}_{ij} + \mathrm{residual}}$$

Practical – Haseman-Elston regression

- Now let’s remove the relatives & retry

```
HEreg$unrel = HEreg$REL<0.05 #make a T/F vector on the relationship value
lm2 = lm(HEreg$PROD[HEreg$unrel]~HEreg$REL[HEreg$unrel])

plot(HEreg$PROD[HEreg$unrel]~HEreg$REL[HEreg$unrel],pch=".")
abline(lm2,col="orange",lwd=1.5)
dev.off()
```