Estimation of SNP-heritability (h^2_{SNP}) using linear mixed models

Loic Yengo

University of Queensland

5th of March, 2019

ICQG6

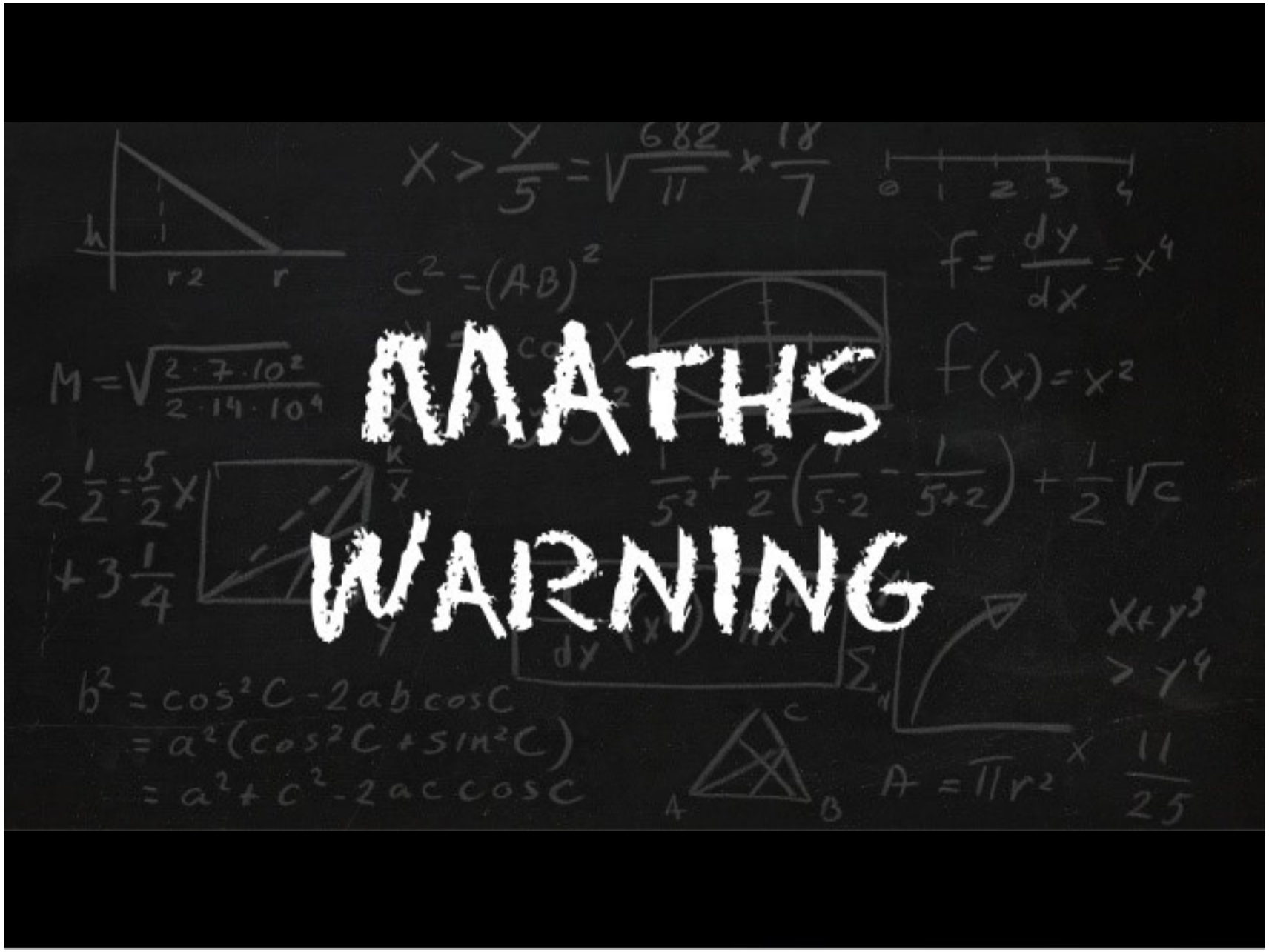
June 14th – 19th 2020, Brisbane, Australia

Outline

- Linear Mixed Models
- GREML (Genome-based Restricted Maximum Likelihood) estimation
- Computational challenges in GREML estimation
- Bias in SNP-heritability due to model misspecification
- Power to detect SNP heritability

Outline

- **Linear Mixed Models**
- GREML (Genome-based Restricted Maximum Likelihood) estimation
- Computational challenges in GREML estimation
- Bias in SNP-heritability due to model misspecification
- Power to detect SNP heritability


MATHS
WARNING

Linear Mixed Models?

- Linear?

  A (quantitative) trait of interest (y) is model as a **linear** combination of multiple variables:

  (1) $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u_1 z_1 + \ldots + u_M z_M + e$

- Mixed?

  We distinguish between "K fixed effects" and "M random effects".

Linear Mixed Models?

- Linear?

  A (quantitative) trait of interest (y) is model as a **linear** combination of multiple variables:

  (1) $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u_1 z_1 + \ldots + u_M z_M + e$

- Mixed?

  We distinguish between "K fixed effects" and "M random effects".

- The difference between "fixed" and "random" is not strictly arbitrary.
- Fixed effects are parameters for which you actually want to know the value.
- Random effects are often considered as "nuisance" parameters.
- With genetic data, we often model SNPs effects as random because there are **too many of them**.

Assumptions

- In this model, random effects are
  - SNP effects: u_j (SNP j)
  - Environmental effects: e_i (individual i)

z_{ij} = scaled genotype.

$$\frac{x_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}$$

$$(1)\ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K + u_1 z_1 + \ldots + u_M z_M + e$$

Assumptions about random effects

- In this model, random effects are
  - SNP effects: u_j (SNP j)
  - Environmental effects: e_i (individual i)

- We often assume random effects to be **independent** and such as
  (1) $E[u_j] = 0$ and $var[u_j] = \sigma^2_g / M$ [σ^2_g: genetic variance / M: number of SNPs analyzed]
  (2) $var[e_i] = \sigma^2_e$ [environmental variance]

$var[y] = \sigma^2_g + \sigma^2_e$ => SNP heritability: $\boldsymbol{h^2_{SNP} = \sigma^2_g / [\sigma^2_g + \sigma^2_e]}$.

Assumptions about random effects

- In this model, random effects are
  - SNP effects: u_j (SNP j)
  - Environmental effects: e_i (individual i)

- We often assume random effects to be **independent** and such as
  (1) $E[u_j] = 0$ and $var[u_j] = \sigma^2_g / M$ [σ^2_g: genetic variance / M: number of SNPs analyzed]
  (2) $var[e_i] = \sigma^2_e$ [environmental variance]

$var[y] = \sigma^2_g + \sigma^2_e$ => SNP heritability: $\boldsymbol{h^2_{SNP} = \sigma^2_g / [\sigma^2_g + \sigma^2_e] = \sigma^2_g / var[y]}$.

Outline

- Linear Mixed Models
- **GREML (Genome-based Restricted Maximum Likelihood) estimation**
- Computational challenges in GREML estimation
- Bias in SNP-heritability due to model misspecification
- Power to detect SNP heritability

Maximum Likelihood Estimation (MLE)

Key idea 1: Observed data are generated by random process.

Key idea 2: Random process corresponds to probability distribution (Normal, Poisson, etc.)

Key idea 3: MLE looks for the probability distribution that is most likely to have generated the data.

Maximum Likelihood Estimation (MLE)

We classically assume **normal distribution** as $u_j \sim N(0, \sigma^2_g / M)$ and $e_i \sim N(0, \sigma^2_e)$ [M SNPs].

NxN **Genetic Relationship Matrix** (GRM)

$$y|X \sim N(X\beta, \sigma^2_g [ZZ'/M] + \sigma^2_e I_N)$$

Notations

y|X: distribution of y conditional on X

N(m,v): [multivariate] normal distribution with mean m and variance v.

I_N: NxN identity matrix (1's on the diagonal and 0's elsewhere).

Genetic Relationship Matrix (GRM)

$$y|X \sim N(X\beta, \sigma^2_g\,[GRM] + \sigma^2_e I_N)$$

.99	-0.01	.01
-.01	1.07	.03
-.03	.001	1.01

1	0	0
0	1	0
0	0	1

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
```

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

(1) Tells us something about relatedness in the sample (**GRM ≈ 2x Kinship**).
(2) Computationally costly…but can be re-used multiple times.
(3) Eigen decomposition of that matrix yields principal components.

REstricted? **REML = MLE on residualised data!**

- When our focus in on estimating variance components (σ^2_g and σ^2_e), we prefer to use a REstricted Maximum Likelihood estimation (REML).

- The principle is simply to remove (residualise) the fixed effects from the inference.

$$y^* | X \sim N(0, \sigma^2_g [ZZ'/M] + \sigma^2_e I_N)$$

- **When sample size is large, MLE and REML are equivalent**. [distinguish REML from GREML]

Outline

- Linear Mixed Models
- GREML (Genome-based Restricted Maximum Likelihood) estimation
- **Computational challenges in GREML estimation**
- Bias in SNP-heritability due to model misspecification
- Power to detect SNP heritability

Computational challenge in GREML

- A number of algorithms have been developed to estimate variance components under GREML (at least 3 implemented in GCTA, Yang *et al.* 2011 `--reml-alg` option).

- Computational complexity is however **cubic** with respect to sample size $\sim O(N^3)$, i.e. if you increase your sample size by a factor k, computational time is increased by a factor $\sim k^3$.

- Alternatives are approximate inference as implemented in BOLT-LMM software (Loh *et al.* 2015) or *"divide and conquer"* approaches.

Computational challenge in GREML

- A number of algorithms have been developed to estimate variance components under GREML (at least 3 implemented in GCTA, Yang *et al.* 2011 `--reml-alg` option).

- Computational complexity is however **cubic** with respect to sample size $\sim O(N^3)$, i.e. if you increase your sample size by a factor k, computational time is increased by a factor $\sim k^3$.

- Alternatives are approximate inference as implemented in BOLT-LMM software (Loh *et al.* 2015) or *"divide and conquer"* approaches.

Computational challenge in GREML

- A number of algorithms have been developed to estimate variance components under GREML (at least 3 implemented in GCTA, Yang *et al.* 2011 `--reml-alg` option).

- Computational complexity is however **cubic** with respect to sample size $\sim O(N^3)$, i.e. if you increase your sample size by a factor k, computational time is increased by a factor $\sim k^3$.

- Alternatives are approximate inference as implemented in BOLT-LMM software (Loh *et al.* 2015) or "*divide and conquer*" approaches.

Outline

- Linear Mixed Models
- GREML (Genome-based Restricted Maximum Likelihood) estimation
- Computational challenges in GREML estimation
- **Bias in SNP-heritability due to model misspecification**
- Power to detect SNP heritability

Few things to be aware of…

- $h^2_{SNP} < h^2$ in general when SNPs poorly tag causal variants.

- However, even if causal variants are "captured", biases may still arise because of specific features of the genetic architecture of traits (e.g. LD pruning). [Can be fixed with MAF or LD stratification]

- Also, under assortative mating interpretation of estimates of SNP heritability can be challenging.

- It is important to use unrelated individuals to estimate SNP heritability, otherwise estimates can be biased by shared environment (as for family-based estimates). Classical GRM thresholds are >0.05 or >0.025.

Few things to be aware of…

- $h^2_{SNP} < h^2$ in general when SNPs poorly tag causal variants.

- However, even if causal variants are "captured", biases may still arise because of specific features of the genetic architecture of traits (e.g. LD pruning). [Can be fixed with MAF or LD stratification]

- Also, under assortative mating interpretation of estimates of SNP heritability can be challenging.

- It is important to use unrelated individuals to estimate SNP heritability, otherwise estimates can be biased by shared environment (as for family-based estimates). Classical GRM thresholds are >0.05 or >0.025.

Few things to be aware of…

- $h^2_{SNP} < h^2$ in general when SNPs poorly tag causal variants.
- However, even if causal variants are "captured", biases may still arise because of specific features of the genetic architecture of traits (e.g. LD pruning). [Can be fixed with MAF or LD stratification]
- Also, under assortative mating interpretation of estimates of SNP heritability can be challenging.
- It is important to use unrelated individuals to estimate SNP heritability, otherwise estimates can be biased by shared environment (as for family-based estimates). Classical GRM thresholds are >0.05 or >0.025.

Few things to be aware of…

- $h^2_{SNP} < h^2$ in general when SNPs poorly tag causal variants.
- However, even if causal variants are "captured", biases may still arise because of specific features of the genetic architecture of traits (e.g. LD pruning). [Can be fixed with MAF or LD stratification]
- Also, under assortative mating interpretation of estimates of SNP heritability can be challenging.
- It is important to use unrelated individuals to estimate SNP heritability, otherwise estimates can be biased by shared environment (as for family-based estimates). Classical GRM thresholds are >0.05 or >0.025.

Few things to be aware of…

- $h^2_{SNP} < h^2$ in general when SNPs poorly tag causal variants.

- However, even if causal variants are "captured", biases may still arise because of specific features of the genetic architecture of traits (e.g. LD pruning). [Can be fixed with MAF or LD stratification]

- Also, under assortative mating interpretation of estimates of SNP heritability can be challenging.

- It is important to use unrelated individuals to estimate SNP heritability, otherwise estimates can be biased by shared environment (as for family-based estimates). Classical GRM thresholds are >0.05 or >0.025.

Outline

- Linear Mixed Models
- GREML (Genome-based Restricted Maximum Likelihood) estimation
- Computational challenges in GREML estimation
- Bias in SNP-heritability due to model misspecification
- **Power to detect SNP heritability**

How can I know if my sample is large enough?

- Visscher et al. (2014) derived the expectation of the standard error (SE) of estimates of SNP heritability as

$$SE(h^2_{SNP}) \approx 1/[N\ SD[GRM]]$$

In Europeans SD(GRM) ≈ 10^{-5}. Therefore $SE(h^2_{SNP}) \approx 316/N$.

.

- If $N=10^5$, then expected SE is ~0.003 [10 divide and conquer => SE~0.01].

- Consequence: e.g. need N> 3,500 to detect an heritability of ~0.2 (Expected SE ≈ 0.09).

Power calculator: https://cnsgenomics.shinyapps.io/gctaPower/

Summary (1)

- LMM are special cases of linear regression models that involves fixed and random effects.
- If we assuming all SNPs to have random effects on a trait, then the variance of these effects measures the SNP heritability.
- REML = Maximum likelihood estimation on “residualised” data wrt to fixed effects.
- GREML estimation is computationally costly when sample size (N) is large!

Summary (2)

- $h^2_{SNP} < h^2$ in general when SNPs poorly tag causal variants.

- However, even if causal variants are "captured", biases may still arise because of specific features of the genetic architecture of traits (e.g. LD pruning). [Can be fixed with MAF or LD stratification]

- Also, under assortative mating interpretation of estimates of SNP heritability can be challenging.

- It is important to use unrelated individuals to estimate SNP heritability, otherwise estimates can be biased by shared environment (as for family-based estimates). Classical GRM thresholds are >0.05 or >0.025.

References

- GCTA
  - Yang et al. (2010). Nat. Genet. **42**(7):565-9
  - Yang et al. (2011). AJHG. **88**(1):76-82

- BOLT-LMM
  - Loh et al. (2015). Nat. Genet. **47**:284–290
  - Loh et al. (2018). Nat. Genet. **50**(7):906-908

- Power to detect heritability
  - Visscher et al. (2014). Plos Genet. **10**(4):e1004269

- REML

Estimation of SNP-heritability (h^2_{SNP}) using linear mixed models – practical using GCTA

Loic Yengo

University of Queensland

5th of March, 2019

GCTA

a tool for Genome-wide Complex Trait Analysis

GCTA SMR GSMR OSCA GCTB Program in PCTG CTG forum

- A command line software (similar to PLINK)
- Version 1.92 available for Linux, Mac and Windows
- Initially developed for estimation of SNP heritability
- Has now multiple functionalities
  - Mixed-Model GWAS
  - LD score calculations
  - Principal Components Analysis
  - Gene-based test.
  - Etc.
- Many examples available on the website: http://cnsgenomics.com/software/gcta/#GREMLanalysis

Locate the data and software

Create a folder for the practical…**please type**

(replace "**yourFolder**" with the name of your own folder, e.g. mine is "yengo")

```
datPath=/scratch/201903/yengo
gctaPath=/opt/gcta_1.92.0beta3
plinkPath=/usr/bin
pracPath=/scratch/201903/yourFolder/practical_greml
mkdir –p $pracPath
```

Ideas of the practical

- Run GCTA to calculate GRM, identify related individuals

- Run --reml on two traits
  - A trait where causal variants are rare with some effect from shared environment
  - A trait where causal variants are poorly tagged.

- Run MAF stratified analysis

Exercise 1

Calculate GRM...

```
$gctaPath/gcta64 --bfile $datPath/mydata --make-grm-bin --out $pracPath/mydata_allSNPs
```

...and identify unrelated individuals

```
$gctaPath/gcta64 --grm $pracPath/mydata_allSNPs --grm-singleton 0.05 --out $pracPath/unrelated
```

Exercise 1

```
*******************************************************************
* Genome-wide Complex Trait Analysis (GCTA)
* version 1.92.0 beta2
* (C) 2010-2018, The University of Queensland
* Please report bugs to: Jian Yang <jian.yang@uq.edu.au>
*******************************************************************
Analysis started at 15:49:20 UTC on Mon Mar 04 2019.
Hostname: cocoa

Options:

--grm /scratch/201903/yengo/testPractical/mydata_allSNPs
--grm-singleton 0.05
--out /scratch/201903/yengo/testPractical/unrelated

The program will be running on 8 threads at most.
Pruning the GRM with a cutoff of 0.050000...
Total number of parts to proceed: 1
    Processing part 1
Related family pairs have been saved to /scratch/201903/yengo/testPractical/unrelated.family.txt
After pruning the GRM, there are 4808 individuals (1192 individuals removed).
    Pruned singleton IDs has been saved to /scratch/201903/yengo/testPractical/unrelated.singleton.txt

Analysis finished at 15:49:20 UTC on Mon Mar 04 2019
Computational time: 0.081524 second(s).
```

Exercise 2

- Estimate SNP heritability of trait 1 with and without relatives.

```
$gctaPath/gcta64 --grm $pracPath/mydata_allSNPs --pheno $datPath/mydata.phen --mpheno 1 --reml \
        --out $pracPath/trait1_with_relatives

$gctaPath/gcta64 --grm $pracPath/mydata_allSNPs --pheno $datPath/mydata.phen --mpheno 1 --reml \
        --out $pracPath/trait1_without_relatives --grm-cutoff 0.05
```

- What can you conclude?

Exercise 2

- Estimate SNP heritability of trait 1 with and without relatives.

```
$gctaPath/gcta64 --grm $pracPath/mydata_allSNPs --pheno $datPath/mydata.phen --mpheno 1 --reml \
        --out $pracPath/trait1_with_relatives

$gctaPath/gcta64 --grm $pracPath/mydata_allSNPs --pheno $datPath/mydata.phen --mpheno 1 --reml \
        --out $pracPath/trait1_without_relatives --grm-cutoff 0.05
```

- What can you conclude?

```
Summary result of REML analysis:
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
```

```
Summary result of REML analysis:
Source          Variance        SE
V(G)            0.253065        0.027906
V(e)            0.731259        0.027626
Vp              0.984324        0.020507
V(G)/Vp         0.257096        0.026772
```

heritability

Exercise 3

Calculate MAF stratified GRMs…

```
$gctaPath/gcta64 --grm $pracPath/mydata_allSNPs --pheno $datPath/mydata.phen --mpheno 2 --reml \
        --keep $pracPath/unrelated.singleton.txt --out $pracPath/trait2_allSNPs
```

```
$gctaPath/gcta64 --mgrm $pracPath/mgrm.txt --pheno $datPath/mydata.phen --mpheno 2 --reml \
        --keep $pracPath/unrelated.singleton.txt  --out $pracPath/trait2_maf_stratified
```

```
Summary result of REML analysis:
Source          Variance        SE
V(G)            0.479330        0.031773
V(e)            0.531696        0.024740
Vp              1.011026        0.022118
V(G)/Vp         0.474102        0.026153
```

```
Summary result of REML analysis:
Source          Variance        SE
V(G1)           0.618051        0.035290
V(G2)           0.007838        0.011030
V(e)            0.385462        0.013896
Vp              1.011350        0.035904
V(G1)/Vp        0.611114        0.014835
V(G2)/Vp        0.007750        0.010906

Sum of V(G)/Vp  0.618864        0.018175
```

Extensions

- Estimate heritability of diseases
- Add covariates explain structure of phen file.

- Haseman-Elston (HE) regression is implemented in GCTA.

  e.g.: `gcta64 --HEreg --grm test --pheno test.phen --out test`

- Estimation of genetic correlation between traits