Computer Session 5: Heritability estimates using GCTA

Felix Tropf & Maria Christodoulou

Heritability

Is the proportion of genetic variance over the total variance in a phenotype within a population:

$$h2 = V(G)/V(P)$$

nature genetics

Meta-analysis of the heritability of human traits based on fifty years of twin studies

Tinca J C Polderman[1,10], Beben Benyamin[2,10], Christiaan A de
Arjen van Bochoven[7], Peter M Visscher[2,8,11] & Danielle Posthu

Despite a century of research on complex traits in humans, the relative importance and specific nature of the influences of genes and environment on human traits remain controversial. We report a meta-analysis of twin correlations and reported variance components for 17,804 traits from 2,748 publications

Speci
genet
addit
Recei
(GW

across all traits, we showed that, on average, $2r_{DZ} - r_{MZ} = 0.042$ twins (**Supplementary Figs. 11** and **12**). The proportion of single

http://match.ctglab.nl/#/home

Recall Twins

Dizygotic/fraternal

Monozygotic/identical

Naive estimates

h2 = heritability

c2 = shared environmental influences

e2 = unique environmental influences/ measurement error

Naive estimates

$$h2 = 2*(r(MZ)-r(DZ))$$

$$c2 = 1-h2$$

$$e2 = 1-r(MZ)$$

Naive estimates

h2 = ?

c2 = ?

e2 = ?

Naive estimates

Why do we observe: r(MZ)?

Naive estimates

Why do we observe: r(MZ)?

MZ twins are more similar than other pairs of individuals, because they share:

1) Their genes (100%); A
2) Parts of their environment; C

r(MZ) = A + C

Naive estimates

Why do we observe: r(DZ)?

Naive estimates

Why do we observe: r(DZ)?
DZ twins are more similar than other pairs of individuals, because they share:

1) Their genes (50%); 0.50A
2) Parts of their environment; C

$$r(DZ) = 0.5A + C$$

Naive estimates

r(MZ)-r(DZ) = ?

Naive estimates

$$h2 = 2*(r(MZ) - r(DZ)) =$$

$$= 2((A+C) - (0.5*A+C)) =$$

$$= 2*(C - C + A - 0.5*A) =$$

$$2*(0.5*A) = A$$

Naive estimates

$$h2 = 2*(r(MZ)-r(DZ))$$

$$c2 = 1-h2$$

$$e2 = 1-r(MZ)$$

Example

You observe:

$$r(MZ) = 0.80$$

$$r(DZ) = 0.40$$

What are

$$e2 = ?$$
$$h2 = ?$$
$$c2 = ?$$

Example

You observe:

$$r(MZ) = 0.80$$

$$r(DZ) = 0.40$$

What are

$$e2 = 1-0.8 = 0.20$$
$$h2 = ?$$
$$c2 = ?$$

Example

You observe:

$$r(MZ) = 0.80$$
$$r(DZ) = 0.40$$

What are

$$e2 = 1-0.8 = 0.20$$
$$h2 = 2*(0.80-0.40) = 0.80$$
$$c2 = ?$$

Example

You observe:

$$r(MZ) = 0.80$$
$$r(DZ) = 0.40$$

What are

$$e2 = 1-0.8 = 0.20$$
$$h2 = 2*(0.80-0.40) = 0.80$$
$$c2 = 0.80-0.80 = 0$$

Example

You observe:

$$r(MZ) = 0.80$$
$$r(DZ) = 0.40$$

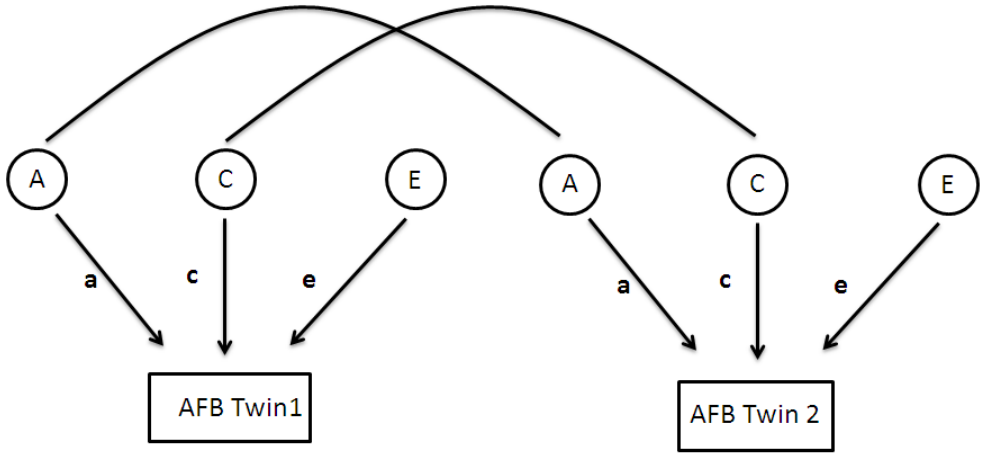
What are

$$e2 = 1-0.8 = 0.20$$
$$h2 = 2*(0.80-0.40) = 0.80$$
$$c2 = 0.80-0.80 = 0$$

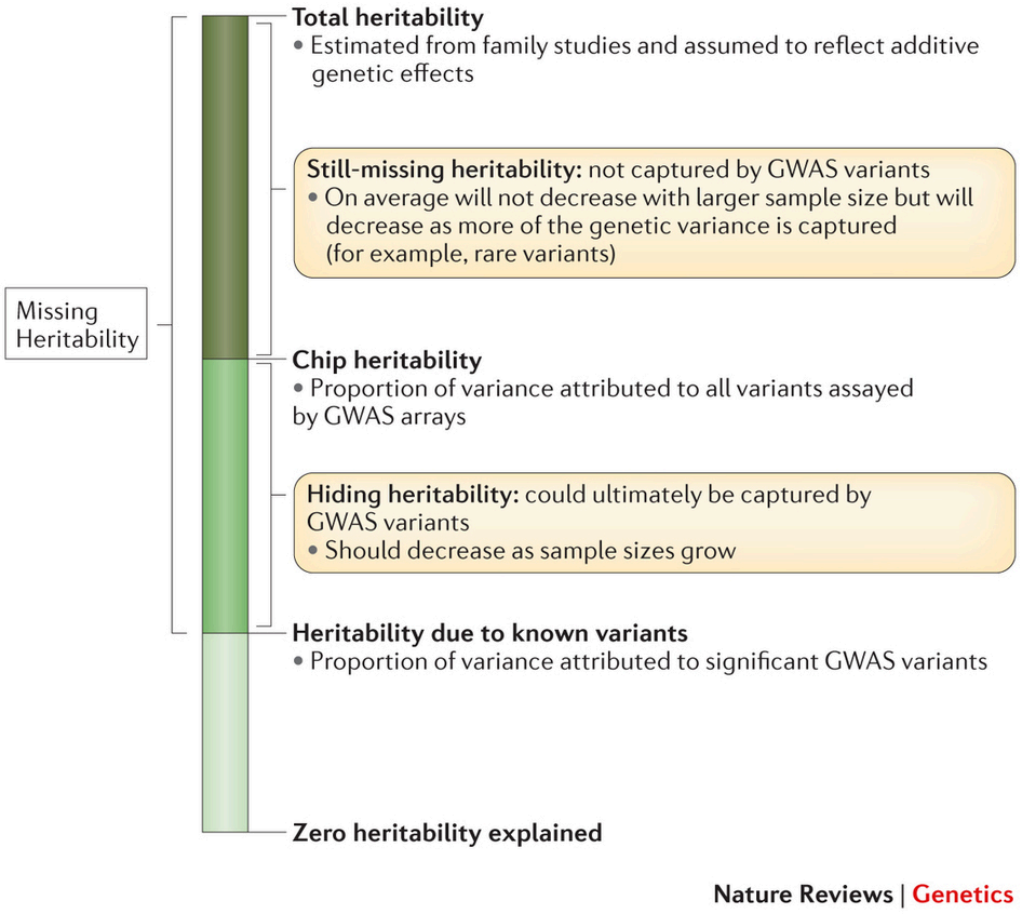
This is height!

Structural equation modelling

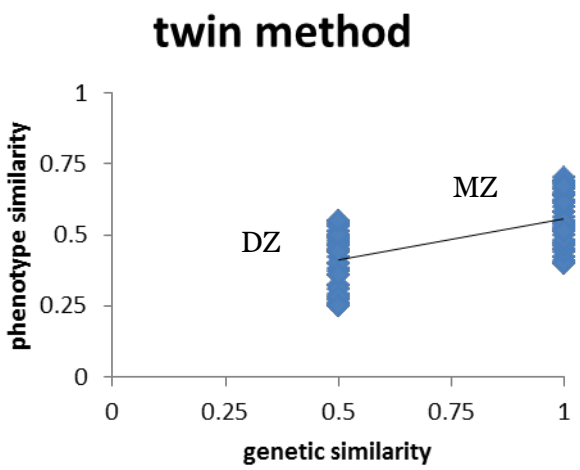
1) Explicit assumptions
2) Goodness of fit tests
3) Confidence intervals of variance components are given
4) Alternative models a fitted
5) Multivariate analysis integrated
6) Software: OpenMx (R-package), twinlm (R-package), etc

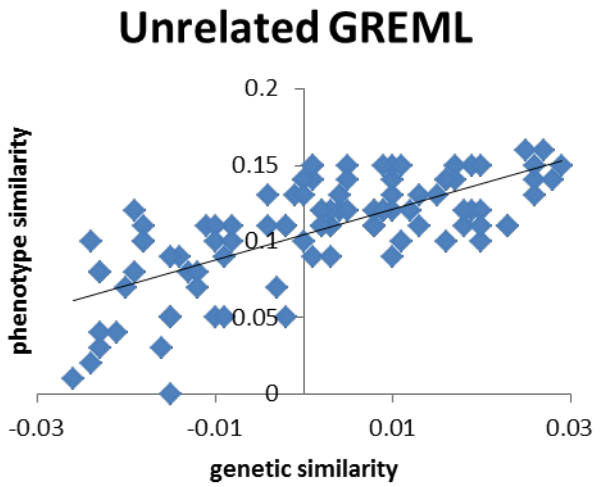


Witte et al (2014)



Twin and GREML method

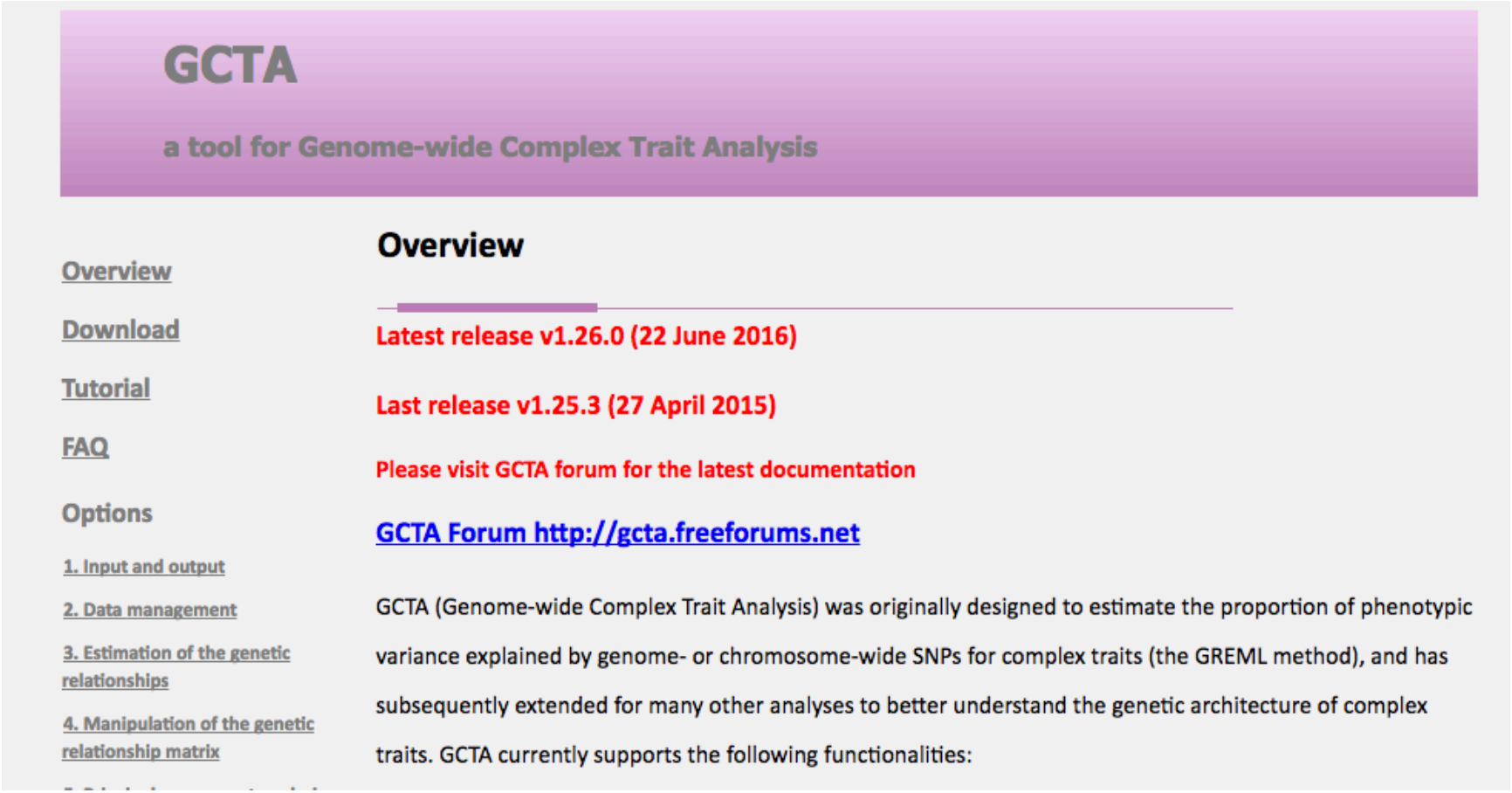




GREML analysis

$$\mathbf{V} = \mathbf{A}\sigma^2_{\mathbf{G}} + \mathbf{I}\sigma^2_{\mathbf{E}},$$

http://cnsgenomics.com/software/gcta/



Yang et al 2011

REPORT

GCTA: A Tool for Genome-wide Complex Trait Analysis

Jian Yang,[1,*] S. Hong Lee,[1] Michael E. Goddard,[2,3] and Peter M. Visscher[1]

For most human complex diseases and traits, SNPs identified by genome-wide association studies (GWAS) explain only a small fraction of the heritability. Here we report a user-friendly software tool called *g*enome-wide *c*omplex *t*rait *a*nalysis (GCTA), which was developed based on a method we recently developed to address the "missing heritability" problem. GCTA estimates the variance explained by all the SNPs on a chromosome or on the whole genome for a complex trait rather than testing the association of any particular SNP to the trait. We introduce GCTA's five main functions: data management, estimation of the genetic relationships from SNPs, mixed linear model analysis of variance explained by the SNPs, estimation of the linkage disequilibrium structure, and GWAS simulation. We focus on the function of estimating the variance explained by all the SNPs on the X chromosome and testing the hypotheses of dosage compensation. The GCTA software is a versatile tool to estimate and partition complex trait variation with large GWAS data sets.

Despite the great success of genome-wide association studies (GWAS), which have identified hundreds of SNPs conferring the genetic variation of human complex diseases and traits,[1] the genetic architecture of human complex traits still remains largely unexplained. For most traits, the associated SNPs from GWAS only explain a small fraction of the heritability.[2,3] There has not been any consensus on the explanation of the "missing heritability." Possible explanations include a large number of common variants with small effects, rare variants with large effects, and DNA structural variation.[2,4] We recently proposed a method of estimating the total amount of phenotypic variance captured by all SNPs on the current generation of commercial genotyping arrays and estimated that ~45% of the phenotypic variance for human height can be explained by all common SNPs.[5] Thus, most of the heritability for height is hiding rather than missing because of many SNPs with small effects.[5,6] In contrast to single-SNP association analysis, the basic concept behind

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \boldsymbol{\varepsilon} \text{ with } \mathbf{V} = \mathbf{A}\sigma_g^2 + \mathbf{I}\sigma_\varepsilon^2, \qquad \text{(Equation 2)}$$

where $\mathbf{g}$ is an $n \times 1$ vector of the total genetic effects of the individuals with $\mathbf{g} \sim N(0, \mathbf{A}\sigma_g^2)$, and $\mathbf{A}$ is interpreted as the genetic relationship matrix (GRM) between individuals. We can therefore estimate σ_g^2 by the restricted maximum likelihood (REML) approach,[10] relying on the GRM estimated from all the SNPs. Here we report a versatile tool called *g*enome-wide *c*omplex *t*rait *a*nalysis (GCTA), which implements the method of estimating variance explained by all SNPs, and extend the method to partition the genetic variance onto each of the chromosomes and also to estimate the variance explained by the X chromosome and test for dosage compensation in females. We developed GCTA in five function domains: data management, estimation of the GRM from a set of SNPs, estimation of the variance explained by all the SNPs on a single chromosome or the whole genome, estimation of linkage disequilibrium (LD) structure, and simulation.

Visscher et al (2010)

A Commentary on 'Common SNPs Explain a Large Proportion of the Heritability for Human Height' by Yang et al. (2010)

Peter M. Visscher,[1] Jian Yang[1] and Michael E. Goddard[2,3]

[1] *Queensland Statistical Genetics Laboratory, Queensland Institute of Medical Research, Brisbane, Australia*
[2] *Department of Food and Agricultural Systems, University of Melbourne, Australia*
[3] *Biosciences Research Division, Department of Primary Industries, Victoria, Melbourne Australia*

Recently a paper authored by ourselves and a number of co-authors about the proportion of phenotypic variation in height that is explained by common SNPs was published in *Nature Genetics* (Yang et al., 2010). Common SNPs explain a large proportion of the heritability for human height (Yang et al.). During the refereeing process (the paper was rejected by two other journals before publication in

hypotheses that could explain this missing heritability. It could be that the SNPs used in GWAS explain some or all of the additive genetic variance but most of them have such a small effect that they are not significant and therefore not reported. Alternatively, it could be that some or all of the mutations causing variation in height are not in perfect linkage disequilibrium (LD) with any of the SNPs and therefore part

Practical

Check

Do you have on your VM a folder containing:

1) plink
2) plink_mac
3) gcta
4) gcta_mac
5) test (.bed,.bim,.fam)
6) test.phen

Can you navigate to the folder with the command line? (cd ../; ls)

What we do now

1) Estimate a genetic relatedness matrix
2) Estimate principal components of this matrix
3) Estimate SNP-based heritability for four phenotypes
4) Estimate the genetic correlation across phenotypes

What is the first thing we do?

Cleaning data!

Clean data

Type:

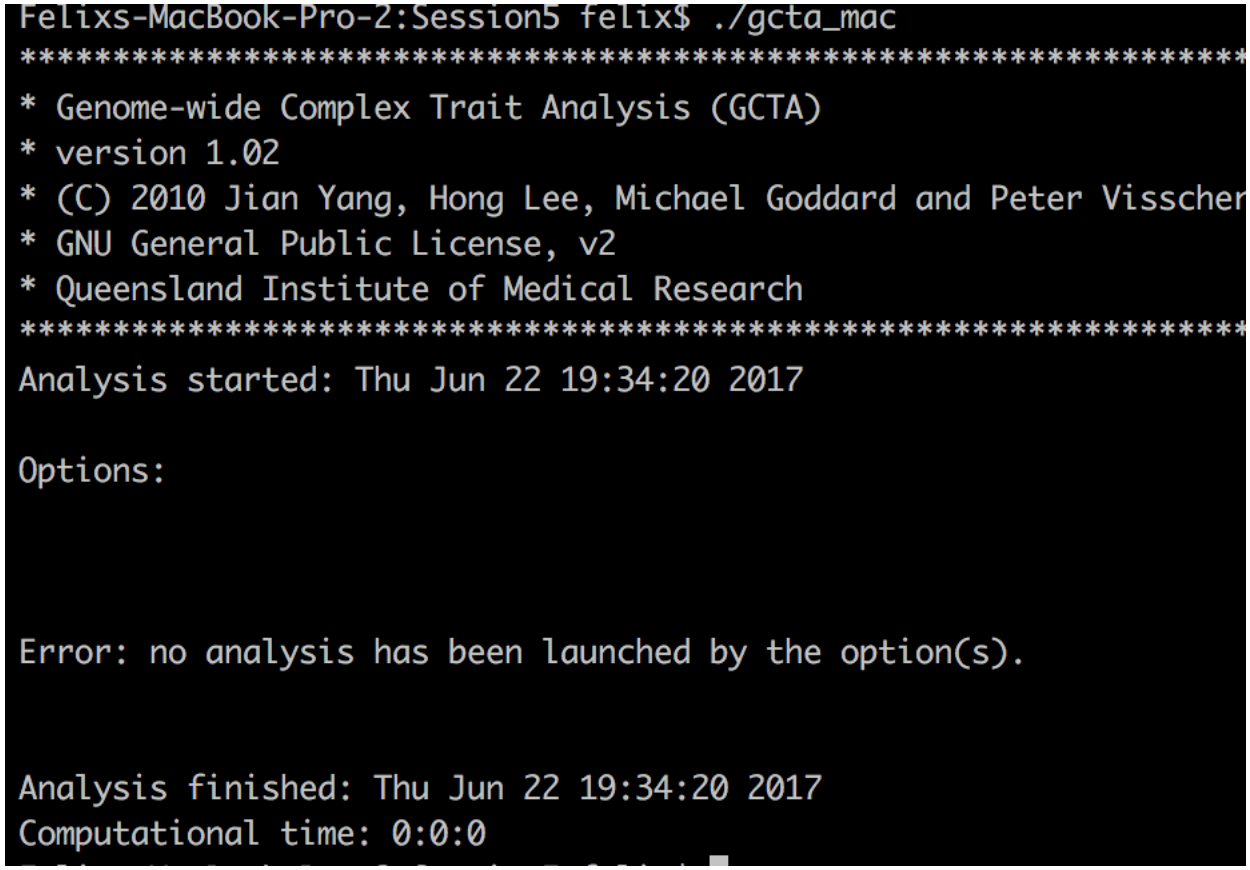
```
./plink_mac --bfile test \
--maf 0.01 \
--geno 0.1 \
--mind 0.1 \
--hwe 0.001 \
--out data_clean \
--make-bed
```

```
PLINK v1.90b4.4 64-bit (21 May 2017)
Options in effect:
  --bfile test
  --geno 0.1
  --hwe 0.001
  --maf 0.01
  --make-bed
  --mind 0.1
  --out data_clean

Hostname: login12
Working directory: /panfs/pan01/vol037/data/sfos-reprogene/gwas/Felix/SummerSchool/5practise
Start time: Tue Jun 20 14:05:25 2017

Random number seed: 1497963925
128814 MB RAM detected; reserving 64407 MB for main workspace.
1000 variants loaded from .bim file.
3925 people (1643 males, 2282 females) loaded from .fam.
25 people removed due to missing genotype data (--mind).
IDs written to data_clean.irem .
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 3900 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate in remaining samples is 0.990473.
15 variants removed due to missing genotype data (--geno).
--hwe: 4 variants removed due to Hardy-Weinberg exact test.
0 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
981 variants and 3900 people pass filters and QC.
Note: No phenotypes present.
--make-bed to data_clean.bed + data_clean.bim + data_clean.fam ... done.

End time: Tue Jun 20 14:05:26 2017
```

Do the gcta

Type:

./gcta64


Felixs-MacBook-Pro-2:Session5 felix$ ./gcta_mac
****
* Genome-wide Complex Trait Analysis (GCTA)
* version 1.02
* (C) 2010 Jian Yang, Hong Lee, Michael Goddard and Peter Visscher
* GNU General Public License, v2
* Queensland Institute of Medical Research
****
Analysis started: Thu Jun 22 19:34:20 2017
Options:
Error: no analysis has been launched by the option(s).
Analysis finished: Thu Jun 22 19:34:20 2017
Computational time: 0:0:0

Make a GRM

Type:

```
./gcta64 --bfile data_clean \
--make-grm \
--autosome \
--out grm
```

```
**********************************************************************
* Genome-wide Complex Trait Analysis (GCTA)
* version 1.02
* (C) 2010 Jian Yang, Hong Lee, Michael Goddard and Peter Visscher
* GNU General Public License, v2
* Queensland Institute of Medical Research
**********************************************************************
Analysis started: Thu Jun 22 19:35:01 2017

Options:
--bfile data_clean
--make-grm
--autosome
--out grm


Reading PLINK FAM file from [data_clean.fam].
3900 individuals to be included from [data_clean.fam].
Reading PLINK BIM file from [data_clean.bim].
981 SNPs to be included from [data_clean.bim].
981 SNPs from chromosome 1 to chromosome 22 are included in the analysis.
Reading PLINK BED file from [data_clean.bed] in SNP-major format ...
Genotype data for 3900 individuals and 981 SNPs to be included from [data_clean.bed].
Recoding genotypes (individual major mode) ...
Calculating allele frequencies ...

Calculating the genetic relationship matrix ... (NOTE: default speed-optimized mode, may use huge RAM)
3900 of 3900 individuals.

Saving the genetic relationship matrix to the file [grm.grm.gz] (in compressed text format).
The genetic relationship matrix has been saved in the file [grm.grm.gz] (in compressed text format).
IDs for the GRM file [grm.grm.gz] have been saved in the file [grm.grm.id].

Analysis finished: Thu Jun 22 19:35:39 2017
Computational time: 0:0:38
```

Check the log

Type:

ls

What do you see?

Type:

gunzip

Check files

Type:

head grm.grm

Type:

head grm.grm.id

```
1       1       981     9.732453e-01
2       1       975     -4.029516e-02
2       2       975     1.082984e+00
3       1       975     -5.887686e-03
3       2       969     -3.548041e-06
3       3       975     1.125246e+00
4       1       981     5.549567e-02
4       2       975     -5.122846e-02
4       3       975     1.144494e-01
4       4       981     1.046521e+00
```

```
1       11
2       21
3       31
4       41
5       51
6       61
7       71
8       81
9       91
10      101
```

Check the log

Type:

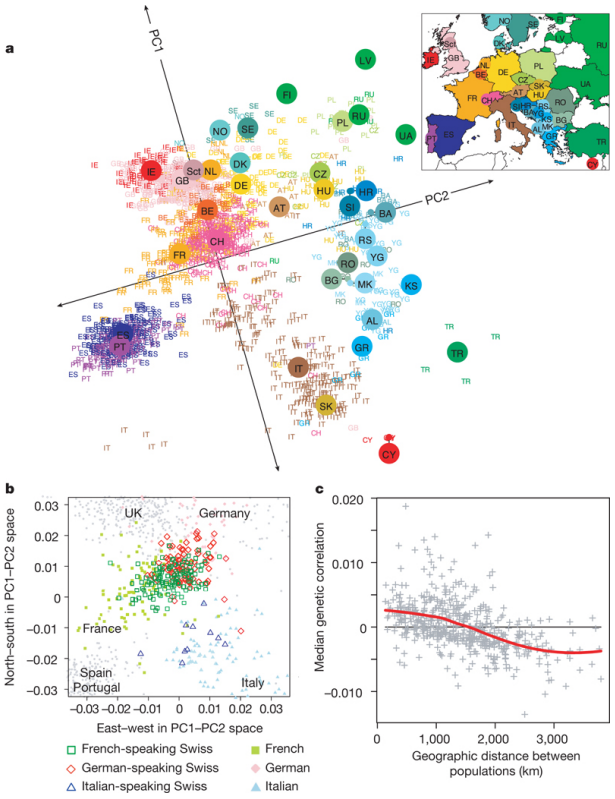
gzip grm.grm

Delete shared environmental influences/related individuals

Type:

./gcta64 --grm grm
--grm-cutoff X
--make-grm
--out grm_cutX

Population structure within Europe.



J Novembre *et al. Nature* **000**, 1-4 (2008) doi:10.1038/nature07331

nature

Generate principal components

Type:

```
./gcta64 --grm grm \
--pca 20 \
--out pca
```

```
* version 1.02
* (C) 2010 Jian Yang, Hong Lee, Michael Goddard and Peter Visscher
* GNU General Public License, v2
* Queensland Institute of Medical Research
**********************************************************************
Analysis started: Thu Jun 22 19:42:53 2017

Options:
--grm grm
--pca 20
--out pca


Reading IDs of the genetic relationship matrix (GRM) from [grm.grm.id].
3900 IDs read from [grm.grm.id].
Reading the GRM from [grm.grm.gz].
Pairwise genetic relationships between 3900 individuals are included from [grm.grm.gz].

Performing principal component analysis ...
Eigenvalues of 3900 individuals have been saved in [pca.eigenval].
```

Run your gcta heritability analysis

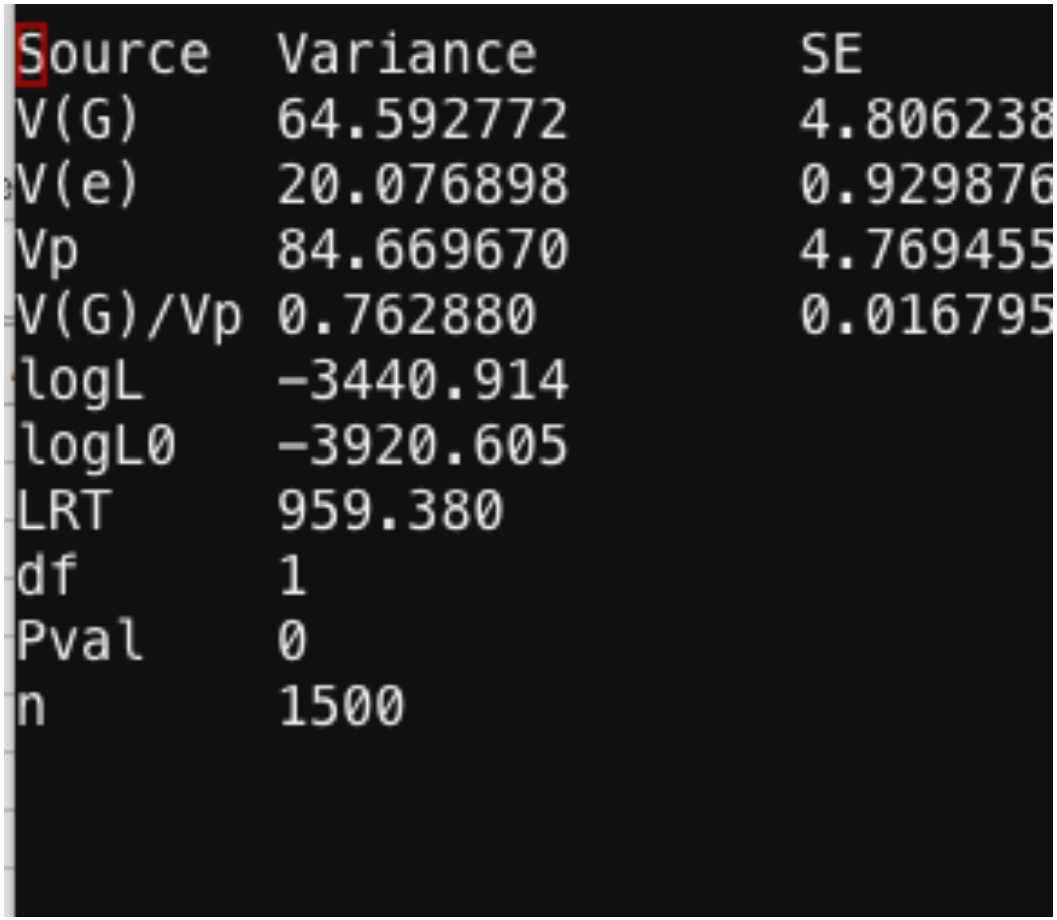
Type:

```
./gcta64 --grm grm \
--reml \
--qcovar pca.eigenvec \
--pheno test.phen \
--out results1
```

Look at protocol

```
Reading IDs of the GRM from [grm.grm.id].
3900 IDs read from [grm.grm.id].
Reading the GRM from [grm.grm.bin].
GRM for 3900 individuals are included from [grm.grm.bin].
Reading phenotypes from [test.phen].
There are 4 traits specified in the file [test.phen].
Trait #1 is included for analysis.
Non-missing phenotypes of 1500 individuals are included from [test.phen].
Reading quantitative covariates from [pca.eigenvec].
20 quantitative covariate(s) of 3900 individuals read from [pca.eigenvec].

20 quantitative variable(s) included as covariate(s).
1500 individuals are in common in these files.

Performing  REML analysis ... (Note: may take hours depending on sample size).
1500 observations, 21 fixed effect(s), and 2 variance component(s)(including residual variance).
Calculating prior values of variance components by EM-REML ...
Updated prior values: 40.2767 27.4837
logL: -3528.94
```

Check results1



Source	Variance	SE
V(G)	64.592772	4.806238
V(e)	20.076898	0.929876
Vp	84.669670	4.769455
V(G)/Vp	0.762880	0.016795
logL	-3440.914	
logL0	-3920.605	
LRT	959.380	
df	1	
Pval	0	
n	1500	

Run your gcta heritability analysis

Type:

```
./gcta64 --grm grm \
--reml \
--qcovar pca.eigenvec \
--pheno test.phen \
--mpheno 2 \
--out results2
```

Run your gcta heritability analysis

Type:

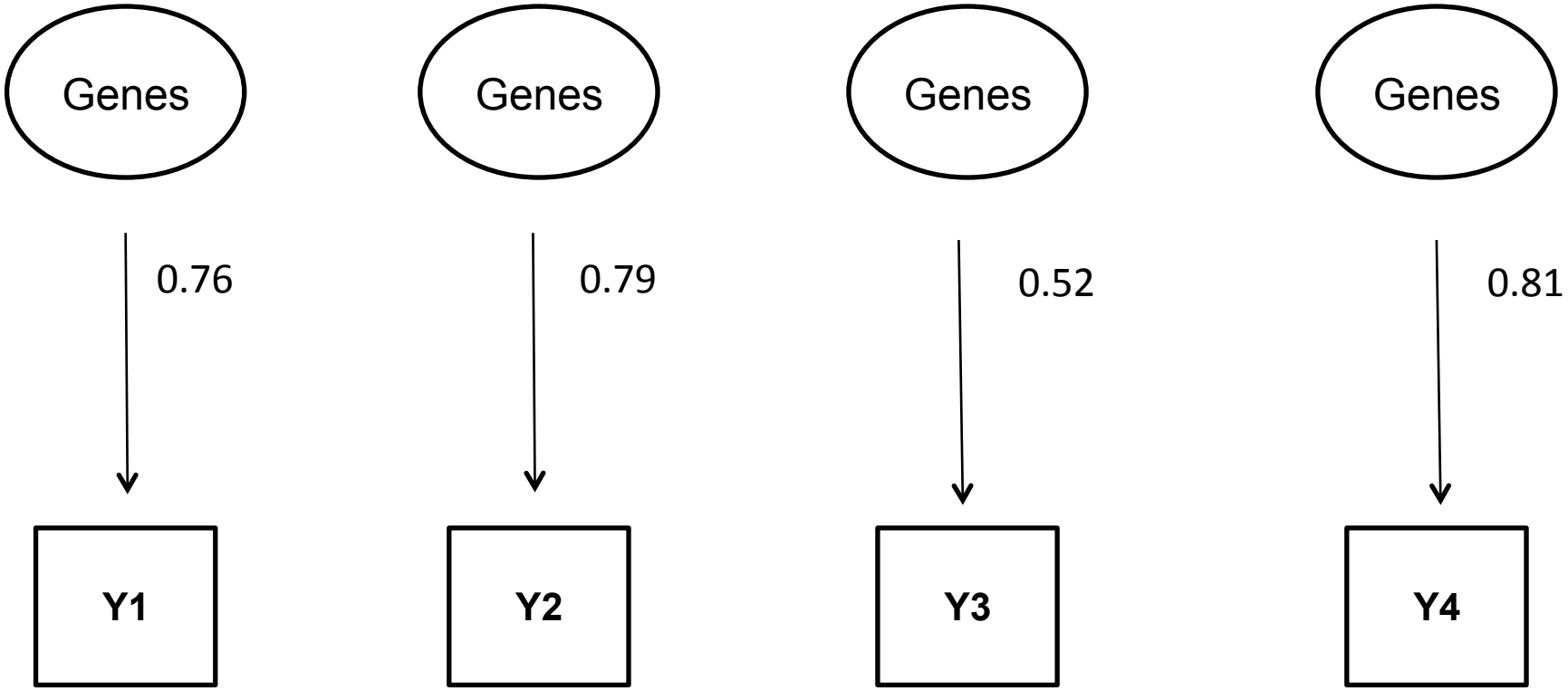
```
./gcta64 --grm grm \
--reml \
--qcovar pca.eigenvec \
--pheno test.phen \
--mpheno 3 \
--out results3
```

Run your gcta heritability analysis

Type:

```
./gcta64 --grm grm \
--reml \
--qcovar pca.eigenvec \
--pheno test.phen \
--mpheno 4 \
--out results4
```

All results



Bivariate genetic analysis

Type:

```
./gcta64 --grm grm \
--reml-bivar 1 2 \
--qcovar pca.eigenvec \
--pheno test.phen \
--out results_bivar1
```

Check results

V(G)_tr1	62.317691	4.263853
V(G)_tr2	65.070172	4.435115
C(G)_tr12	63.570985	4.132618
V(e)_tr1	20.410120	0.906862
V(e)_tr2	19.260815	0.873884
C(e)_tr12	4.515476	0.631877
Vp_tr1	82.727811	4.318916
Vp_tr2	84.330987	4.454698
V(G)/Vp_tr1	0.753286	0.015472
V(G)/Vp_tr2	0.771605	0.014927
rG	0.998303	0.005234
logL	-6453.297	

Bivariate genetic analysis

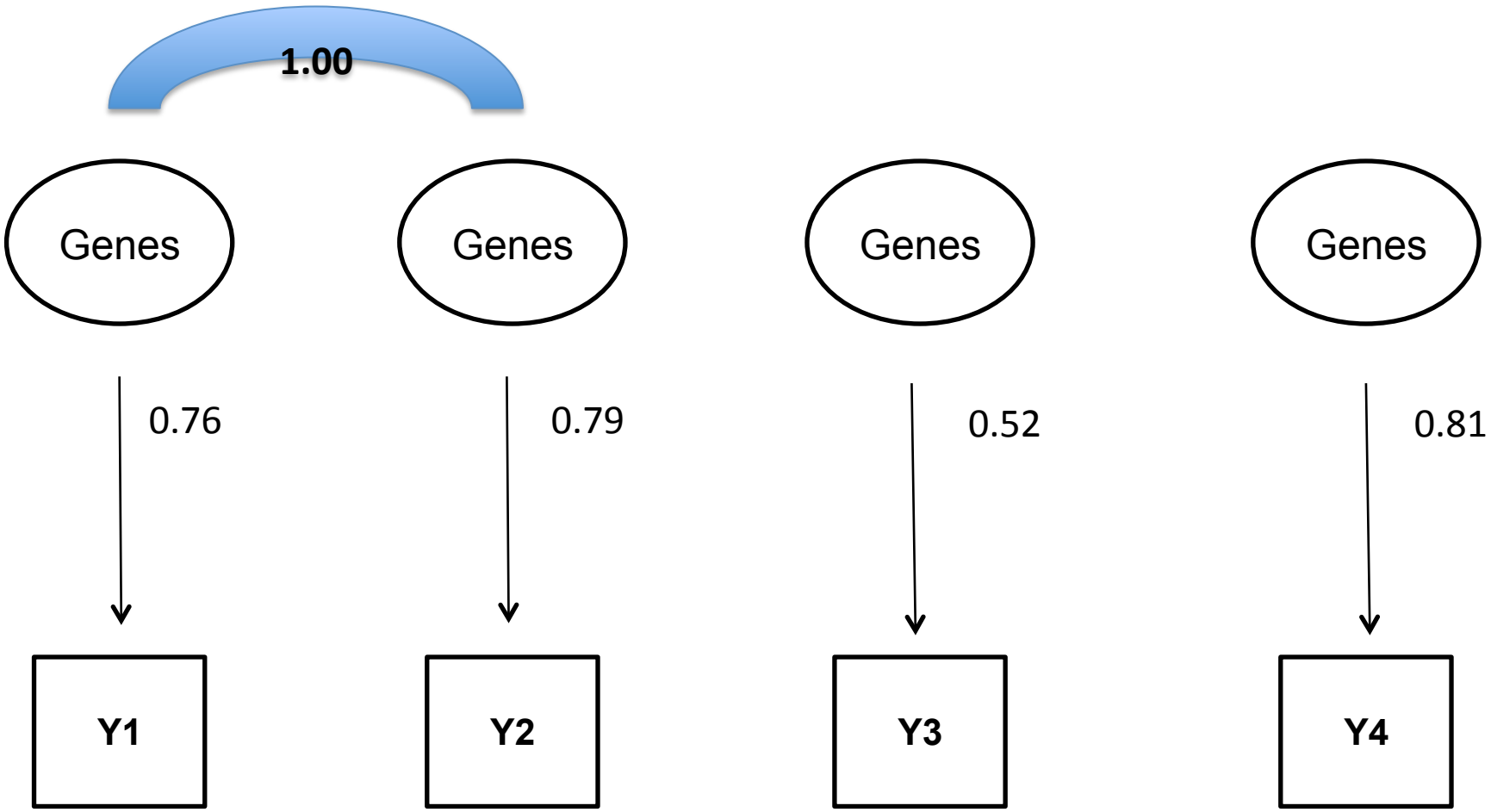
Type:

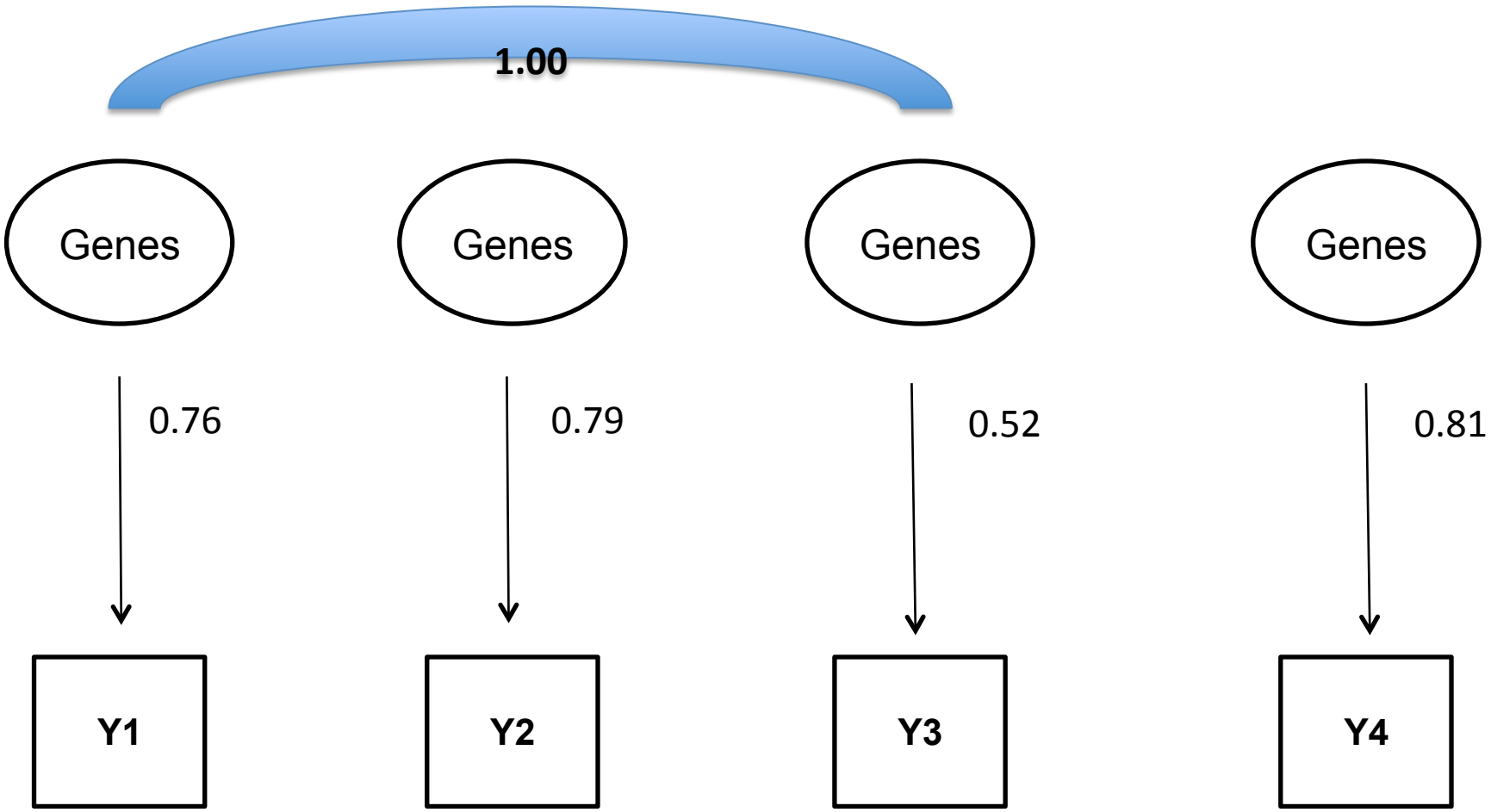
```
./gcta64 --grm grm \
--reml-bivar 1 3 \
--qcovar pca.eigenvec \
--pheno test.phen \
--out results_bivar2
```

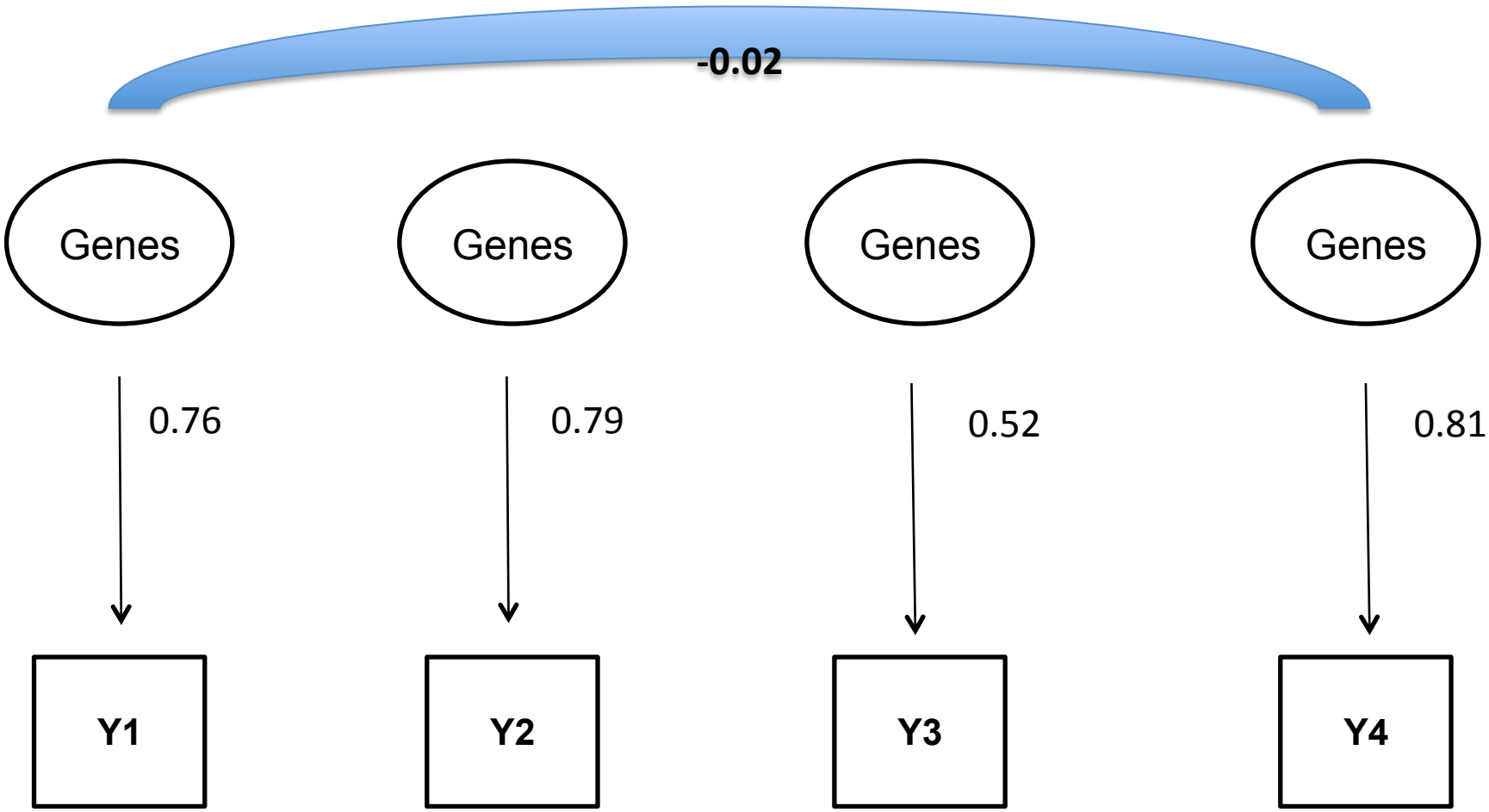
Bivariate genetic analysis

Type:

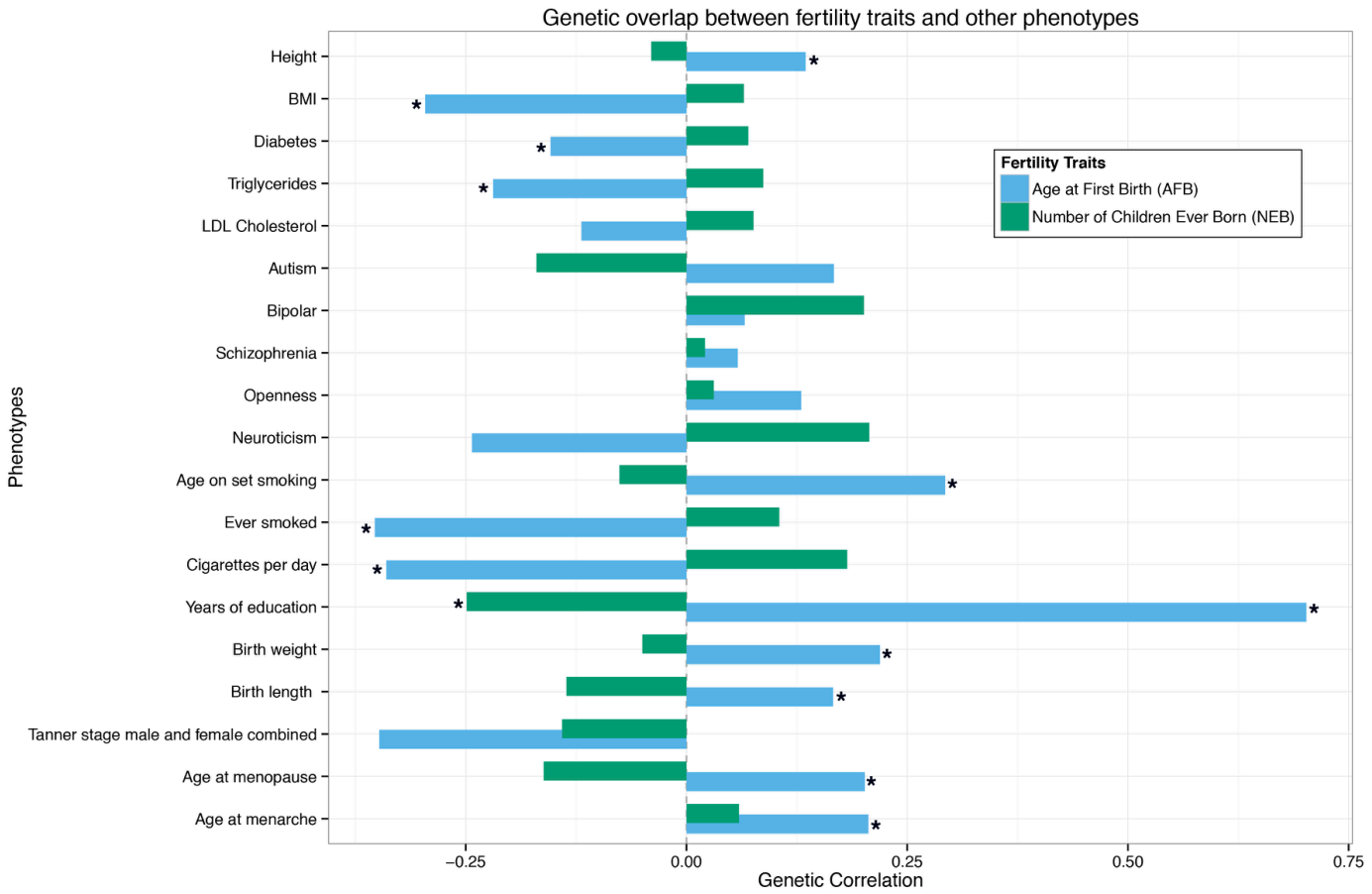
```
./gcta64 --grm grm \
--reml-bivar 1 4 \
--qcovar pca.eigenvec \
--pheno test.phen \
--out results_bivar3
```


1.00
Genes
Genes
Genes
Genes
0.76
0.79
0.52
0.81
Y1
Y2
Y3
Y4


1.00
Genes
Genes
Genes
Genes
0.76
0.79
0.52
0.81
Y1
Y2
Y3
Y4

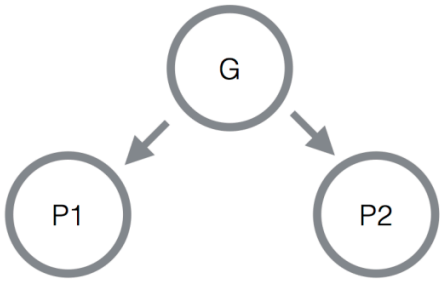

-0.02
Genes
Genes
Genes
Genes
0.76
0.79
0.52
0.81
Y1
Y2
Y3
Y4

r(G) across traits



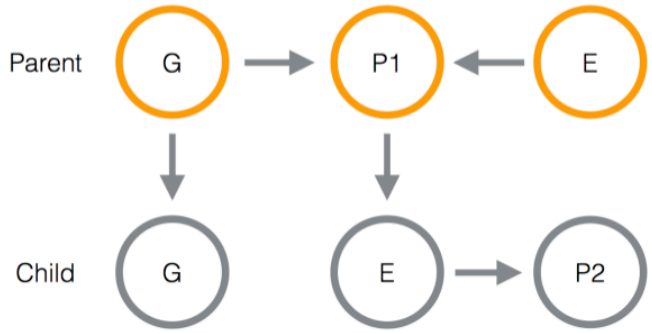
What is a genetic correlation?

1 “Biological” pleiotropy

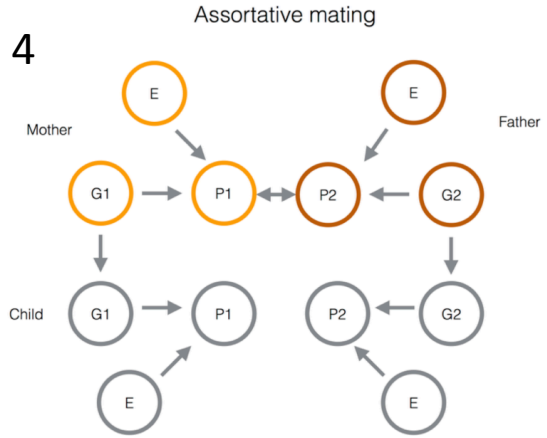


2 “Mediated” pleiotropy

3 Parental effects

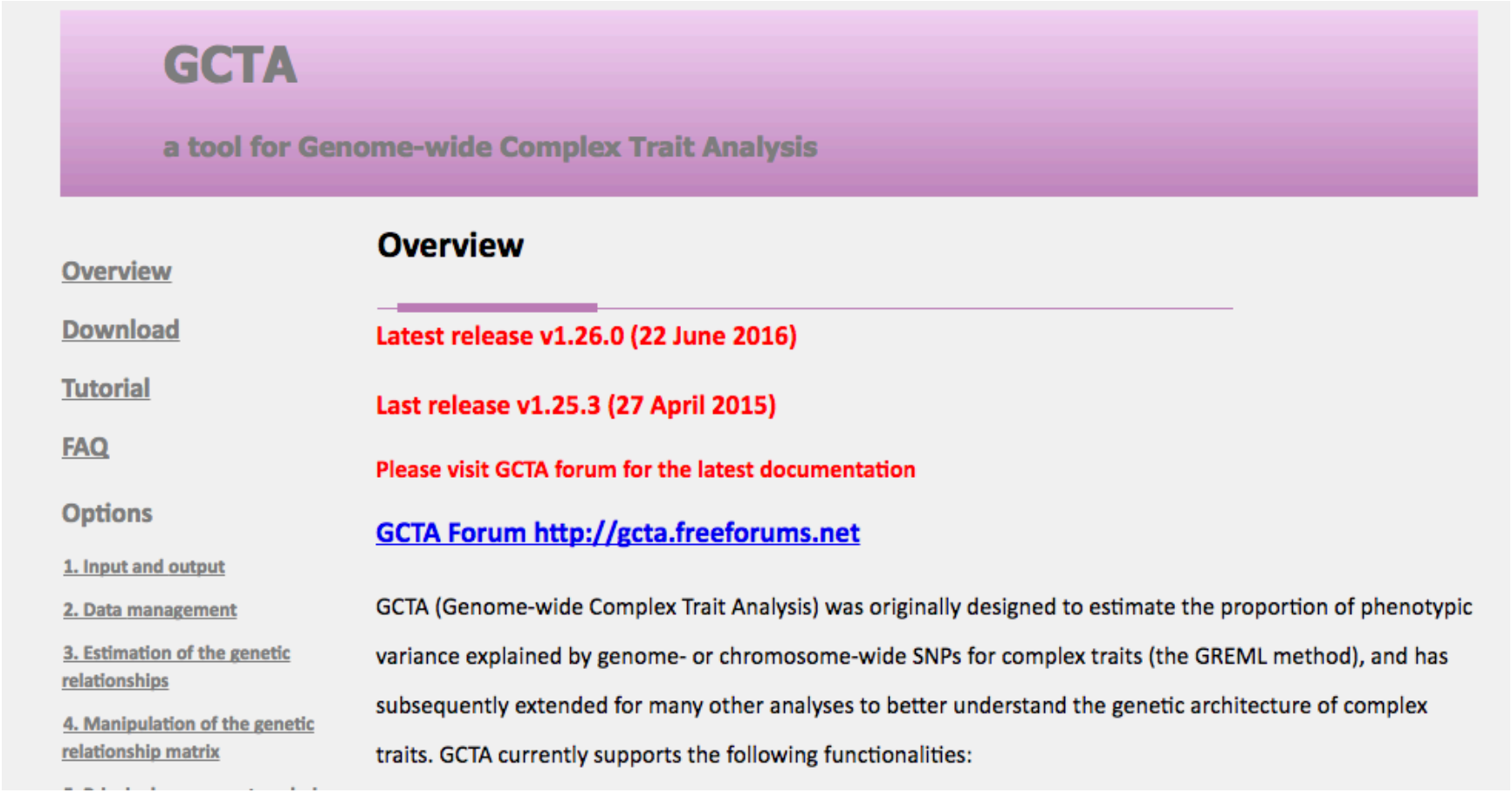


4



By Joe Pickrell

Check the website!



Thanks for your attention!

Questions?