# GCTA Practical 1

Goal: To use GCTA to estimate $h^2_{SNP}$ from SNP array data for a single polygenic phenotype

# Simulation:

Useful to test assumptions, performance, understanding of reality.
Allows flexibility in altering conditions

**Two general approaches:**

- Simulate phenotype from real genotype data
  - Real patterns of LD, polymorphism, stratification, allele frequency
  - Often used to assess how methods perform
- Simulate both genotype and phenotype data
  - More control over the demography, allele frequency
  - Programs to do this in either forward-time (e.g., GeneEvolve [Tahmasbi & Keller 2016]) or coalescent (e.g., Hudson's ms to generate genotypes)

# **GCTA practical:** Real genotypes, simulated phenotypes

**Genotype Data to Make the Genetic Relatedness Matrix (GRM)**

Similar to what might be collected for a GWAS study with SNP array data

- 1,000 Genomes + UK10K sequence data
- Using Affymetrix Axiom Array positions

# **GCTA practical:** Real genotypes, simulated phenotypes

**GRM: Axiom Array Positions**

MAF > 0.05

HWE $p < 10^{-5}$

N = 3,363

Relatedness < 0.05

**GRM ELEMENTS:**

$$A_{ij} = \frac{1}{m}\sum_{k}^{m}\frac{(x_{ik}-2p_k)(x_{jk}-2p_k)}{2p_k(1-p_k)}$$

$$var(\boldsymbol{y}) = \boldsymbol{A}\sigma_v^2 + \boldsymbol{I}\sigma_e^2$$

$$h^2_{SNP} = \sigma^2_v / (\sigma^2_v + \sigma^2_e)$$

**GRM ALREADY CONSTRUCTED** (plink or gcta)

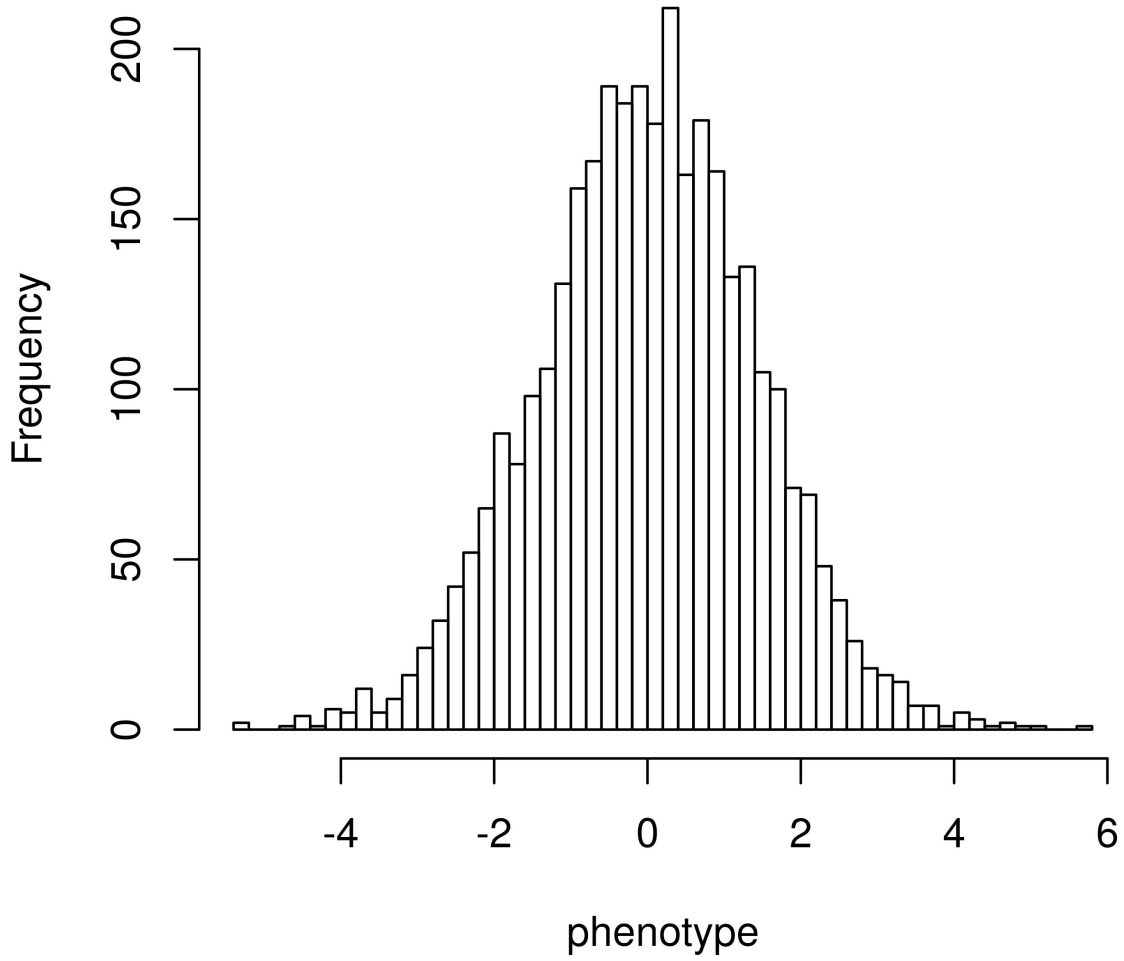
# **GCTA practical:** Real genotypes, simulated phenotypes

**Simulated phenotypes with a standard polygenic model**

- 1,000 causal variants
- Randomly from whole genome sequence data
- Realistic LD & MAF with respect to SNP array data used to create the GRM

- Phenotypes
  - $y_i = g_i + e_i$

$g_i$



# Real genotypes, simulated phenotypes

**Simulated phenotypes with a standard polygenic model**

- $y_i = g_i + e_i$
- $g_i = \sum w_{ik} \beta_k$
  - $w_{ik}$ = 0/1/2 genotype
  - $\beta_k$ = allelic effect size ~N(0,1/[2$p_k$(1-$p_k$)])
  - $p_k$ = MAF
  - $g_i$'s normalized

$w_i$ = [0, 1, 2, 1]

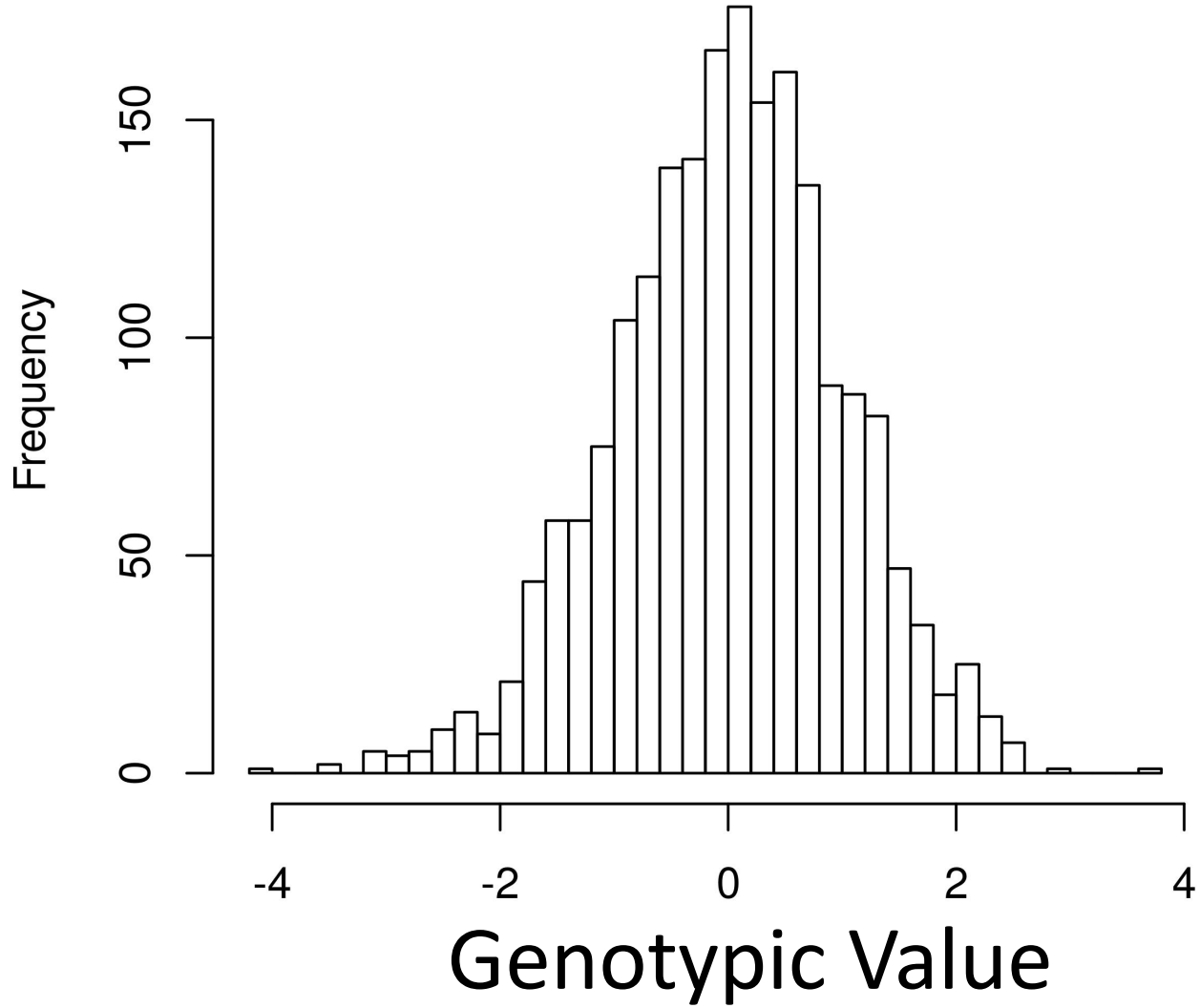
$\beta$ = [2.2, -1.16, 4.1, -0.01]

$g_i$ = (2.2*0) + (-1.16*1) + (4.1*2) + (-0.01*1)

# Real genotypes, simulated phenotypes

**Simulated phenotypes with a standard polygenic model**

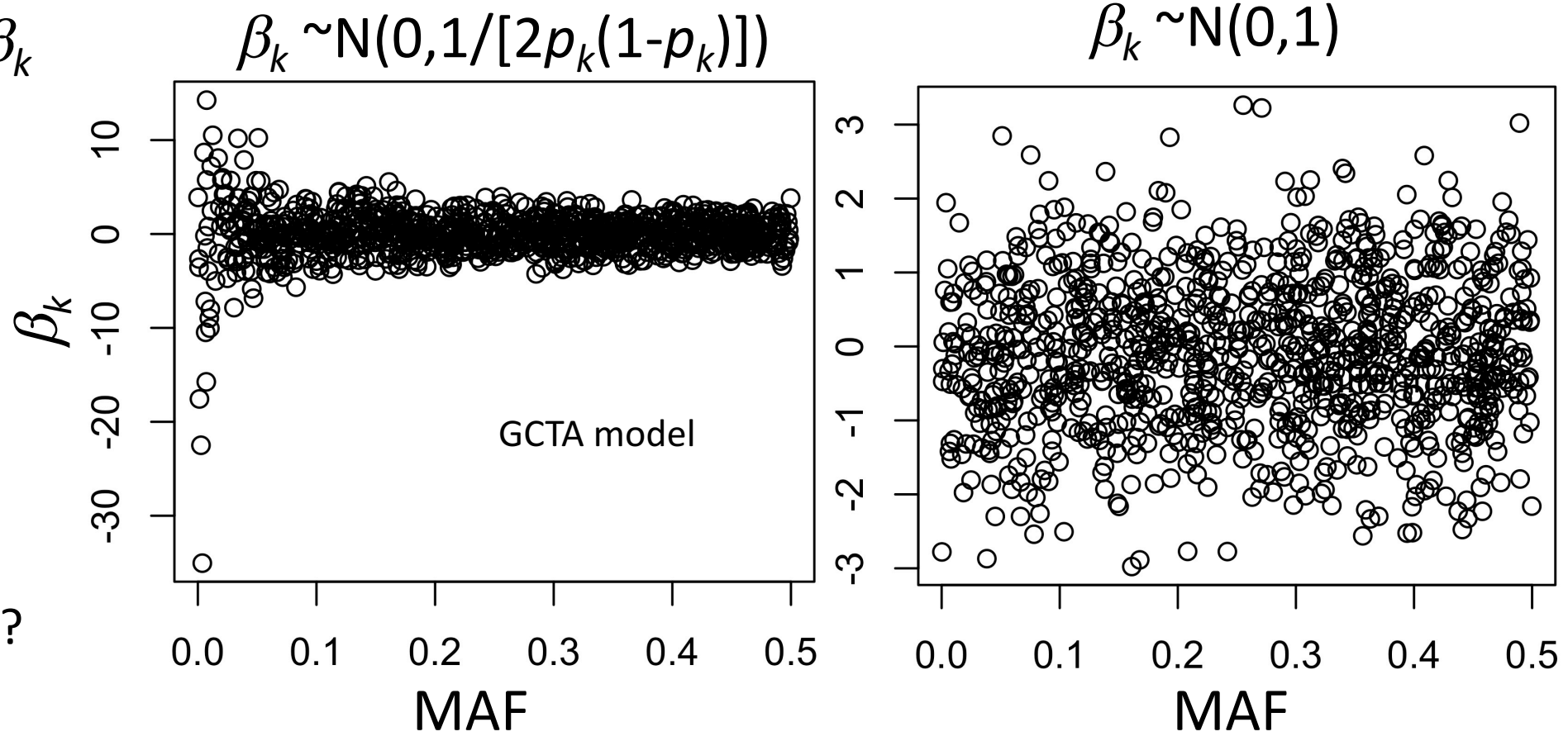
- $y_i = g_i + e_i$
- $g_i = \sum w_{ik} \beta_k$
  - $w_{ik}$ = 0/1/2 genotype
  - $\beta_k$ = allelic effect size ~N(0,1/[2$p_k$(1-$p_k$)])
  - $p_k$ = MAF
  - $g_i$'s normalized



# Real genotypes, simulated phenotypes

**Simulated phenotypes with a standard polygenic model**

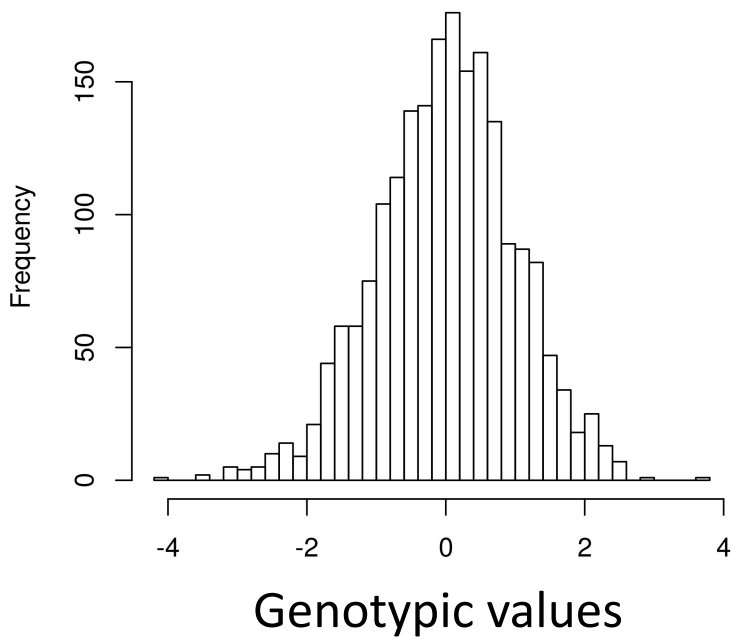
- $y_i = g_i + e_i$
- $g_i = \sum w_{ik}\beta_k$

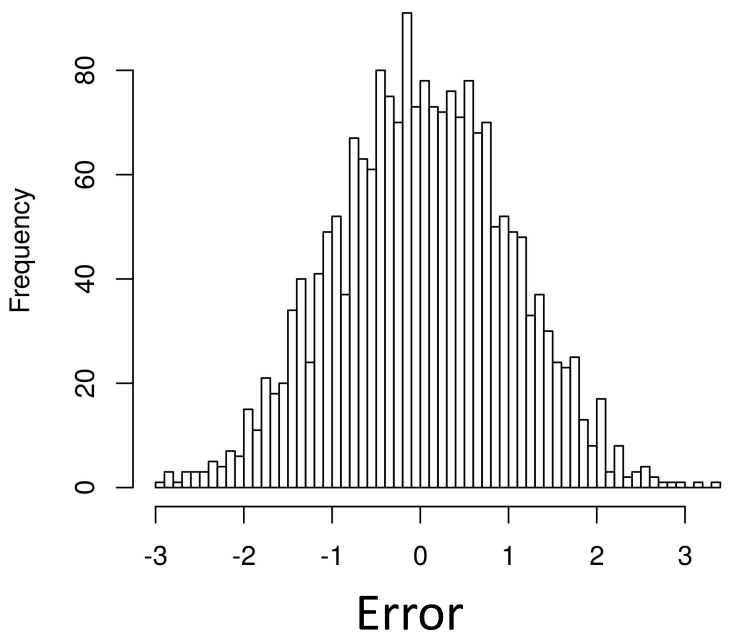


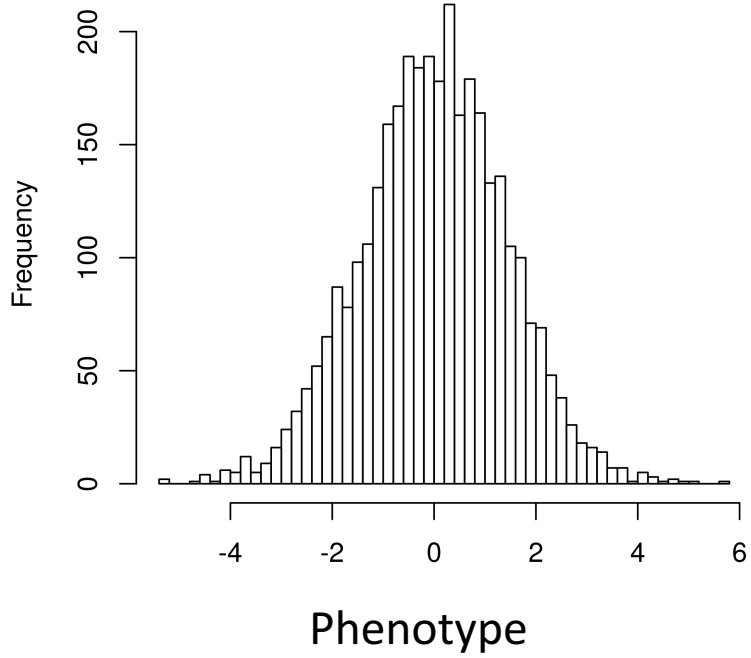

Assumptions?
Model (GRM)?
Causal variants?

# **GCTA practical:** Real genotypes, simulated phenotypes

- $y_i = g_i + e_i$
  - Add error ~N(0, (1- $h^2$)/$h^2$)
  - Simulated $h^2$ = 0.5



# GCTA Practical

## Data already loaded on local drives

- LOCATION: /faculty/luke/2017/Wednesday_practical_1

- GET DATA:
  - Open terminal
  - **TYPE:** cp –r /faculty/luke/2017/Wednesday_practical_1 /YOUR/HOME/DIRECTORY/HERE/
  - **TYPE:** cd /YOUR/HOME/DIRECTORY/HERE/Wednesday_practical_1

# GCTA Practical

- **TYPE:** ls

- GRM:
  - SNPs.rel05.grm.bin (binary file with GRM elements)
  - SNPs.rel05.grm.N.bin (binary file with the number of SNPs used to create the GRM)
  - SNPs.rel05.grm.id (id file with family ID and individual ID listed)
- Phenotype:
  - pheno_randomCVs.txt

# GCTA Practical

- **TYPE:** head SNPs.rel05.grm.id

# GCTA Practical

- **TYPE:** head SNPs.rel05.grm.small.txt
  - Example of the information in the GRM

$$A_{ij} = \frac{1}{m}\sum_{k}^{m}\frac{(x_{ik} - 2p_k)(x_{jk} - 2p_k)}{2p_k(1 - p_k)}$$

# GCTA Practical

- **TYPE:** head pheno_randomCVs.txt

# GCTA Practical: RUN GCTA

- GCTA COMMAND LINE:
  - Very similar to plink, uses -- flags
  - You must give it:
    - GRM prefix
    - Phenotype file
    - Analysis to perform

# GCTA Practical: RUN GCTA

**COMMAND:**

**gcta --grm-bin SNPs.rel05 --pheno pheno_commonCVs.txt --reml --out SNPgrm --thread-num 4**

# GCTA Practical: RUN GCTA

COMMAND:

gcta --grm-bin SNPs.rel05 --pheno pheno_commonCVs.txt --reml --out SNPgrm --thread-num 4

## OUTPUT:

**TYPE:** cat SNPgrm.hsq

| Source | Variance | SE |
|---|---|---|
| V(G | 0.024886 | 0.220215 |
| V(e) | 0.991848 | 0.214227 |
| Vp | 2.016735 | 0.049553 |
| V(G)/Vp | 0.508191 | 0.106992 |
| logL | -2851.499 | |
| logL0 | -2865.440 | |
| LRT | 27.881 | |
| df | 1 | |
| Pval | 6.449e-08 | |
| n | 3362 | |

# GCTA Practical: RUN GCTA

COMMAND:

gcta --grm-bin SNPs.rel05 --pheno pheno_commonCVs.txt --reml --out SNPgrm --thread-num 4

OUTPUT:

**TYPE:** cat SNPgrm.hsq

| Source | Variance | SE |
|---|---|---|
| V(G | 0.024886 | 0.220215 |
| V(e) | 0.991848 | 0.214227 |
| Vp | 2.016735 | 0.049553 |
| V(G)/Vp | 0.508191 | 0.106093 |
| logL | -2851.499 | |
| logL0 | -2865.440 | |
| LRT | 27.881 | |
| df | 1 | |
| Pval | 6.449e-08 | |
| n | 3362 | |

$h^2_{SNP}$

**TRUE $h^2$ = 0.5**

# GCTA Practical: RUN GCTA

COMMAND:

gcta --grm-bin SNPs.rel05 --pheno pheno_commonCVs.txt --reml --out SNPgrm --thread-num 4

OUTPUT:

**TYPE:** cat SNPgrm.hsq

| Source | Variance | SE |
|---|---|---|
| V(G | 0.024886 | 0.220215 |
| V(e) | 0.991848 | 0.214227 |
| Vp | 2.016735 | 0.049553 |
| V(G)/Vp | 0.508191 | 0.106992 |
| logL | -2851.499 | |
| logL0 | -2865.440 | |
| LRT | 27.881 | |
| df | 1 | |
| Pval | 6.449e-08 | |
| n | 3362 | |

$h^2_{SNP}$

**TRUE $h^2$ = 0.5**

**95% CI: 0.508-1.96*0.107 = 0.3**

**0.508-1.96*0.107 = 0.72**

**Unbiased**

# GCTA Practical: RUN GCTA

COMMAND:

gcta --grm-bin SNPs.rel05 --pheno pheno_commonCVs.txt --reml --out SNPgrm --thread-num 4

OUTPUT:

**TYPE:** cat SNPgrm.hsq

| Source | Variance | SE |
|---|---|---|
| V(G | 0.024886 | 0.220215 |
| V(e) | 0.991848 | 0.214227 |
| Vp | 2.016735 | 0.049553 |
| V(G)/Vp | 0.508191 | 0.106992 |
| logL | -2851.499 | |
| logL0 | -2865.440 | |
| LRT | 27.881 | |
| df | 1 | |
| Pval | 6.449e-08 | |
| n | 3362 | |

Likelihood Ratio Test

Testing if V(G) > 0

2*(-2851.499--2865.44) = 27.88

$X^2$ test, 1 df

# GCTA Practical: RUN GCTA

COMMAND:

gcta --grm-bin SNPs.rel05 --pheno pheno_commonCVs.txt --reml --out SNPgrm --thread-num 4

OUTPUT:

**TYPE:** cat SNPgrm.hsq

| Source | Variance | SE |
|---|---|---|
| V(G | 0.024886 | 0.220215 |
| V(e) | 0.991848 | 0.214227 |
| Vp | 2.016735 | 0.049553 |
| V(G)/Vp | 0.508191 | 0.106992 |
| logL | -2851.499 | |
| logL0 | -2865.440 | |
| LRT | 27.881 | |
| df | 1 | |
| Pval | 6.449e-08 | |
| n | 3362 | |

**WHAT FACTORS INFLUENCE THE ESTIMATE?**