GCTA Practical 2

Goal: To use GCTA to estimate h^2_{SNP} from whole genome sequence data & understand how MAF/LD patterns influence biases

GCTA practical: Real genotypes, simulated phenotypes

Genotype Data to Make the Genetic Relatedness Matrix (GRM)

Whole Genome Sequence used to make GRM

- 1,000 Genomes + UK10K sequence data
- All variants included (except singletons)
- Relatedness<0.05
- N = 3,363
- Plink used to create the GRM (--make-grm-bin)
  - **ALREADY DONE**

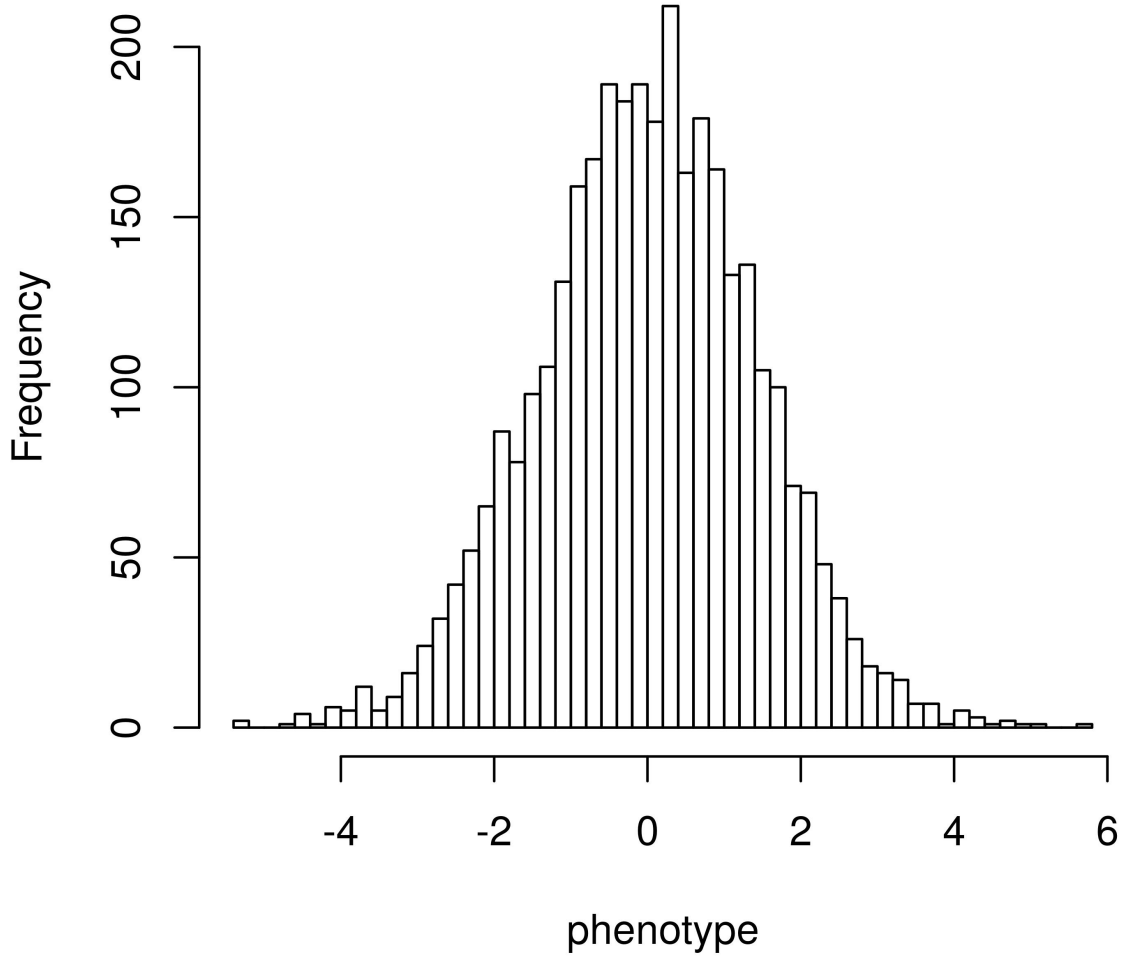
GCTA practical: Real genotypes, simulated phenotypes

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

3 Different Phenotypes:
Vary the MAF of the Causal variants

- Random: drawn from all sequenced variants
- Rare: 0.0025 < MAF < 0.01
- Common: MAF > 0.05



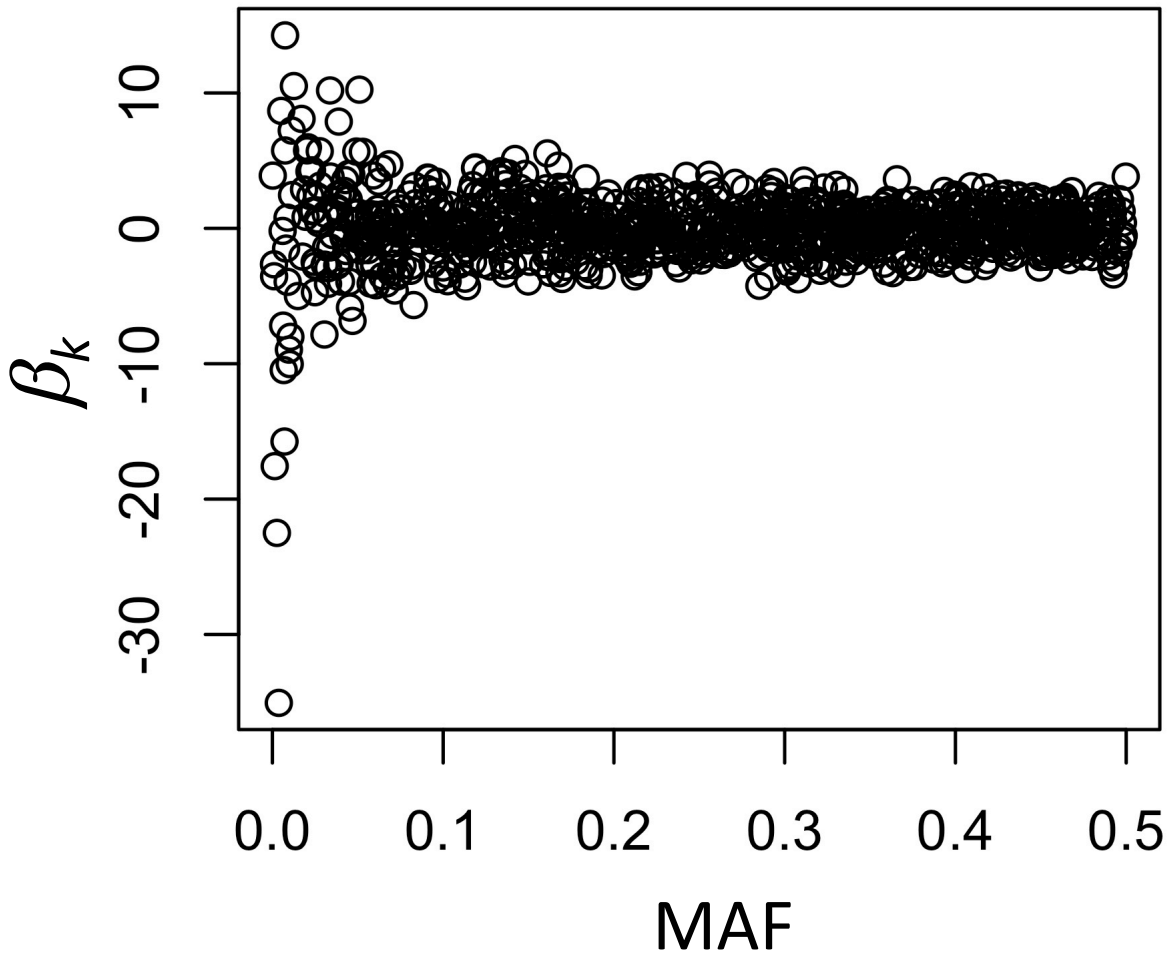
GCTA practical: Real genotypes, simulated phenotypes

Simulated phenotypes with a standard polygenic model

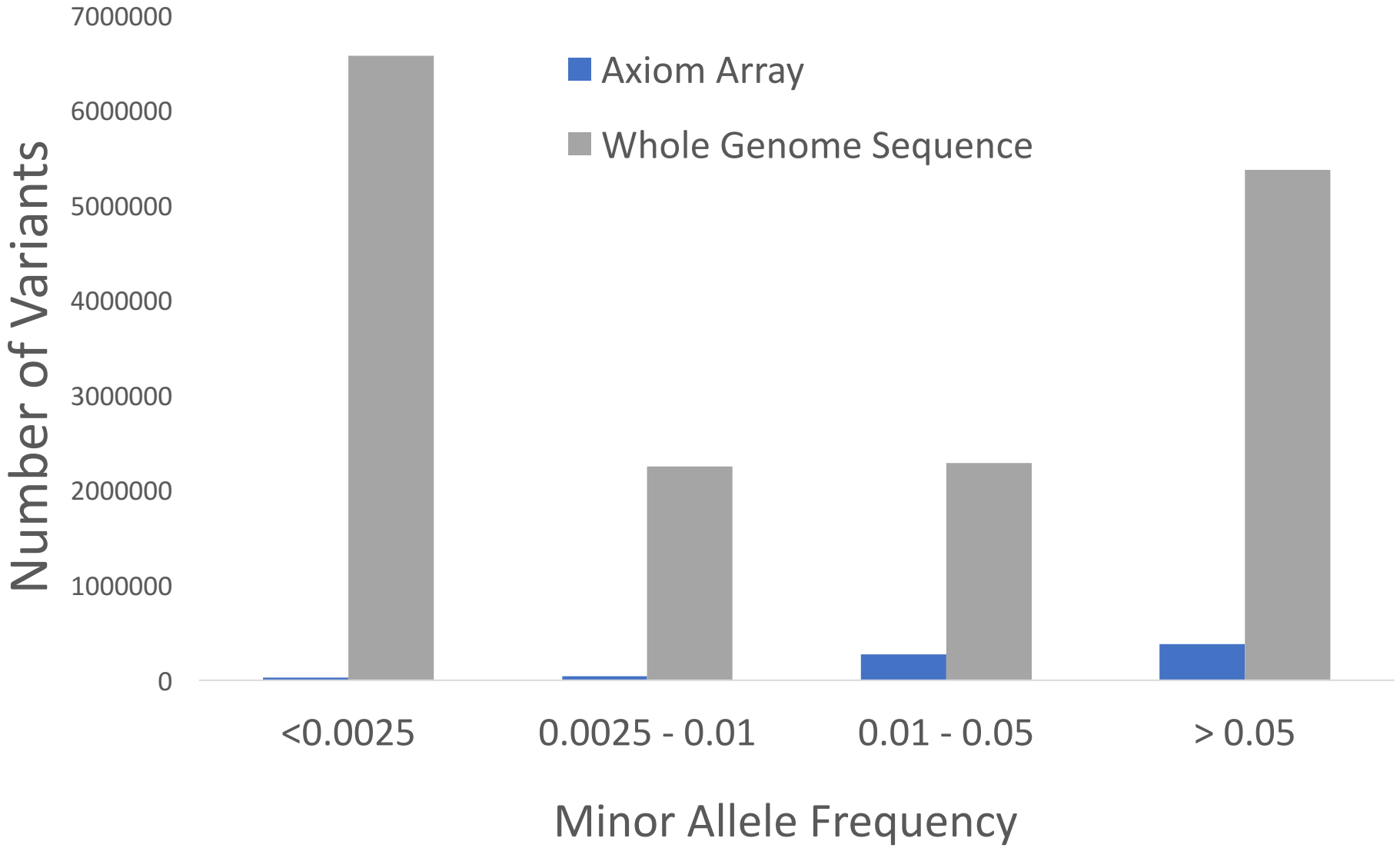
- 1,000 causal variants
- Randomly from whole genome sequence data
- Realistic LD & MAF with respect to SNP array data used to create the GRM

3 Different Phenotypes:
Vary the MAF of the Causal variants

- Random: drawn from all sequenced variants
- Rare: 0.0025 < MAF < 0.01
- Common: MAF > 0.05



MAF of SNP array markers and whole genome sequence



GCTA Practical

Data already loaded on local drives

- LOCATION: /faculty/luke/2017/Wednesday_practical_2

- GET DATA:
  - Open terminal
  - **TYPE:** cp –r /faculty/luke/2017/Wednesday_practical_2 /YOUR/HOME/DIRECTORY/HERE/
  - **TYPE:** cd /YOUR/HOME/DIRECTORY/HERE/Wednesday_practical_2

GCTA Practical

- **TYPE:** ls

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

GCTA Practical: RUN GCTA

COMMAND:

Randomly chosen CVs:

gcta --grm-bin WGS.rel05 --pheno pheno_randomCVs.txt --reml --out WGSgrm.random --thread-num 4

GCTA Practical: GCTA OUTPUT

Randomly Drawn CVs:

TYPE: cat WGSgrm.random.hsq

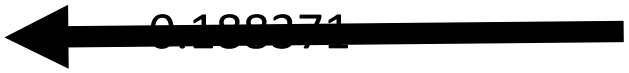
Source	Variance	SE
V(G)	1.040127	0.379450
V(e)	0.976229	0.381714
Vp	2.016356	0.049236
V(G)/Vp	0.515845	0.188371
logL	-2873.277	
logL0	-2877.338	
LRT	8.121	
Df	1	
Pval	0.002188	
n	3363	

GCTA Practical: GCTA OUTPUT

Randomly Drawn CVs:

TYPE: cat WGSgrm.random.hsq

Source	Variance	SE
V(G)	1.040127	0.379450
V(e)	0.976229	0.381714
Vp	2.016356	0.049236
V(G)/Vp	0.515845	0.188271
logL	-2873.277	
logL0	-2877.338	
LRT	8.121	
Df	1	
Pval	0.002188	
n	3363	



h^2_{SNP}

TRUE h^2 = 0.5

GCTA Practical: GCTA OUTPUT

Randomly Drawn CVs:

TYPE: cat WGSgrm.random.hsq

Source	Variance	SE
V(G)	1.040127	0.379450
V(e)	0.976229	0.381714
Vp	2.016356	0.049236
V(G)/Vp	0.515845	0.188371
logL	-2873.277	
logL0	-2877.338	
LRT	8.121	
Df	1	
Pval	0.002188	
n	3363	

h^2_{SNP}

TRUE h^2 = 0.5

95% CI: 0.51-1.96*0.188 = 0.146

0.51+1.96*0.188 = 0.885

Unbiased estimate

GCTA Practical: GCTA OUTPUT

Randomly Drawn CVs:

TYPE: cat WGSgrm.random.hsq

Source	Variance	SE
V(G)	1.040127	0.379450
V(e)	0.976229	0.381714
Vp	2.016356	0.049236
V(G)/Vp	0.515845	0.188371
logL	-2873.277	
logL0	-2877.338	
LRT	8.121	
Df	1	
Pval	0.002188	
n	3363	

Likelihood Ratio Test
Testing if V(G) > 0
2*(-2873.277--2877.338) = 8.1
X^2 test, 1 df

GCTA Practical: RUN GCTA

COMMANDS:

Rare CVs:

gcta --grm-bin /path/to/data/WGS.rel05 --pheno path/to/data/pheno_rareCVs.txt --reml --out WGSgrm.rare --thread-num 4

Common CVs:

gcta --grm-bin /path/to/data/WGS.rel05 --pheno path/to/data/pheno_commonCVs.txt --reml --out WGSgrm.common --thread-num 4

GCTA Practical: GCTA OUTPUT

Randomly Drawn CVs:

TYPE: cat WGSgrm.random.hsq

Source	Variance	SE
V(G)	1.040127	0.379450
V(e)	0.976229	0.381714
Vp	2.016356	0.049236
V(G)/Vp	0.515845	0.188371
logL	-2873.277	
logL0	-2877.338	
LRT	8.121	
Df	1	
Pval	0.002188	
n	3363	

Rare CVs (0.0025 < MAF < 0.01):

TYPE: cat WGSgrm.rare.hsq

Source	Variance	SE
V(G)	0.423509	0.358354
V(e)	1.535181	0.364253
Vp	1.958689	0.047950
V(G)/Vp	0.216220	0.183335
logL	-2820.059	
logL0	-2820.765	
LRT	1.412	
Df	1	
Pval	0.1173	
n	3363	

h^2_{SNP}

TRUE h^2 = 0.5

Downward bias

(small N = large SE)

GCTA Practical: GCTA OUTPUT

Randomly Drawn CVs:

TYPE: cat WGSgrm.random.hsq

Source	Variance	SE
V(G)	1.040127	0.379450
V(e)	0.976229	0.381714
Vp	2.016356	0.049236
V(G)/Vp	0.515845	0.188371
logL	-2873.277	
logL0	-2877.338	
LRT	8.121	
Df	1	
Pval	0.002188	
n	3363	

Rare CVs (MAF> 0.05):

TYPE: cat WGSgrm.common.hsq

Source	Variance	SE
V(G)	1.670281	0.382601
V(e)	0.322239	0.380868
Vp	1.992520	0.048599
V(G)/Vp	0.838276	0.191079
logL	-2855.224	
logL0	-2865.809	
LRT	21.171	
Df	1	
Pval	2.101e-06	
n	3363	

← h^2_{SNP}

TRUE h^2 = 0.5

Upward bias

(small N = large SE)

GCTA Practical – MAF-stratified

- MAF-stratified GREML – partition variance among MAF bins
  - Multiple GRMs included in the model, same otherwise
- Data:
  - Change to MGRM directory.
  - **TYPE:** cd MGRM

GCTA Practical – MAF-stratified

- GRMs:
  - WGS.rel05.common.* (all variants with MAF > 0.05)
  - WGS.rel05.uncommon.* (0.0025 < MAF < 0.05)
  - WGS.rel05.rare.* (MAF<0.0025)
- Phenotype:
  - pheno_randomCVs.txt (CVs randomly drawn from all sequence variants)
- MGRM list
  - List of GRM paths and prefixes
  - **TYPE:** cat mgrm.list.txt

GCTA Practical: RUN GCTA with multiple GRMs

COMMAND:

Randomly Drawn CVs:

gcta --mgrm mgrm.list.txt --pheno pheno_randomCVs.txt --reml --reml-no-lrt --out mgrm.randomCV --thread-num 4

GCTA Practical: GCTA OUTPUT MGRM MODEL

TYPE: cat mgrm.randomCV.hsq

Source	Variance	SE
V(G1)	0.303900	0.184182
V(G2)	0.127654	0.309142
V(G3)	0.653199	0.328909
V(e)	0.926493	0.435653
Vp	2.011246	0.049641
V(G1)/Vp	0.151100	0.091277
V(G2)/Vp	0.063470	0.153765
V(G3)/Vp	0.324773	0.164408
Sum of V(G)/Vp	0.539344	0.215105
logL	-2872.894	
N	3363	

GCTA Practical: RUN GCTA with multiple GRMs

Try running with different phenotypes

- **Common CVs**
- **Rare CVs**