Part 1: GCTA-GREML tutorial (L. Yengo – 10/06/2021)

Prepare folder for practical

```
mkdir GREML_practical
cd GREML_practical
```

Copy the data from Loic's faculty directory

```
cp /faculty/loic/2021/scriptGREML_practical.sh .
cp /faculty/loic/2021/GREML_practical/mydata.* .
ls -l
```

[the file "scriptGREML_practical.sh" contains all the command to run for the practical]

```
workshop-15:~/test/GREML_practical> ls -l
total 13979
-rw-r--r-- 1 loic workshop 18000003 Jun  9 19:04 mydata.bed
-rw-r--r-- 1 loic workshop   264201 Jun  9 19:04 mydata.bim
-rw-r--r-- 1 loic workshop   170000 Jun  9 19:04 mydata.fam
-rw-r--r-- 1 loic workshop   322007 Jun  9 19:04 mydata.phen
-rw-r--r-- 1 loic workshop     1384 Jun  9 19:04 scriptGREML_practical.sh
```

Exercise 1: Calculate the Genetic Relationship Matrix (GRM) using all SNPs

```
gcta64 --bfile mydata --make-grm-bin --out mydata_allSNPs
```

Questions:

1) *How many individuals/SNPs do you have in the sample?*
2) *What is the mean of the diagonal elements of the GRM? Is that expected?*

Exercise 2: Calculate GRM using SNPs with MAF>0.05 and MAF<0.05

```
gcta64 --bfile mydata --maf 0.05 --make-grm-bin --out mydata_maf_above_0.05
gcta64 --bfile mydata --max-maf 0.05 --make-grm-bin --out mydata_maf_below_0.05
```

Questions:

1) *How many SNPs have a MAF<0.05?*
2) *Why is the variance of diagonal elements larger in the GRM based on SNPs with MAF<0.05?*

Some more data processing…

```
## Creating mgrm.txt file (with the "prefix" of the MAF stratified GRMs)
echo Creating mgrm.txt file
echo mydata_maf_below_0.05 >  mgrm.txt
echo mydata_maf_above_0.05 >> mgrm.txt
```

[Note: in general it is preferable to use the full path for defining the mgrm.txt file]

Exercise 3: Identify (genetically) unrelated individuals

```
gcta64 --grm mydata_allSNPs --grm-singleton 0.05 --out unrelated
```

Questions:

1) *How many individuals are unrelated in this sample? GRM cut-off 0.05? cut-off 0.025?*

Exercise 4: GREML estimation of heritability with and without relatives

```
gcta64 --grm mydata_allSNPs --pheno mydata.phen --mpheno 1 \
     --reml --out trait1_with_relatives

gcta64 --grm mydata_allSNPs --pheno mydata.phen --mpheno 1 \
     --reml --out trait1_without_relatives --grm-cutoff 0.05
```

Questions:

1) *Compare heritability estimates from these two analyses. What could explain the difference that you observe?*

Exercise 5: MAF-stratified heritability estimation

(Note: this is using a different phenotype than in Exercise 4)

```
gcta64 --grm mydata_allSNPs --pheno mydata.phen --mpheno 2 \
     --reml --keep unrelated.singleton.txt --out trait2_allSNPs

gcta64 --mgrm mgrm.txt --pheno mydata.phen --mpheno 2 \
     --reml --keep unrelated.singleton.txt --out trait2_maf_stratified
```

Questions: Compare heritability estimates from these two analyses

1) *What could explain the difference that you observe?*
2) *Which one of these estimates should you trust more? How can you be sure?*

Part 2: LD score regression tutorial (S. Gazal – 10/06/2021)

Prepare folder for practical

```
mkdir LDSC_practical
cd LDSC_practical
```

Locate data to use for the practical (no need to copy them)

```
ls -l /data/LDSC_REF
```

```
workshop-29:~/test/LDSC> ls -l LDSC_REF/
total 81870
drwxr-xr-x 2 loic workshop      112 Jun  5 18:45 baselineLD_v2.2
drwxr-xr-x 2 loic workshop       90 Jun  5 18:44 ldscores
-rw-r--r-- 1 loic workshop 38756281 Jun  5 18:45 PASS_Bipolar_Disorder.sumstats
-rw-r--r-- 1 loic workshop 55730612 Jun  5 18:45 PASS_Schizophrenia.sumstats
drwxr-xr-x 2 loic workshop       90 Jun  5 18:45 plink_files
-rw-r--r-- 1 loic workshop 42255462 Jun  5 18:45 UKB_460K.body_HEIGHTz.sumstats
drwxr-xr-x 2 loic workshop       24 Jun  5 18:45 weights_hm3_no_MHC
```

In this tutorial we will apply LD score regression (LDSC) and Stratified LD score regression (S-LDSC) on summary statistics from height computed on UK Biobank.

See the Appendix to install `ldsc` software and to download all the files used in this tutorial.

I/ LD score regression tutorial

I-a/ Some background

LD score regression (LDSC) is an approach to quantify confounding from genome wide association studies (GWAS) summary statistics (Bulik-Sullivan et al. 2015a *Nat Genet*). Briefly, under certain assumptions, we can write the expected χ_j^2 statistic of variant *j* as

$$E[\chi_j^2] = Nh^2 l(j)/\mathrm{M} + Nb + 1$$

where *N* is the GWAS sample size, *M* is the number of SNPs, such that h^2/M is the average heritability explained per SNP; *b* measures the contribution of confounding biases, such as cryptic relatedness and population stratification; and $l(j) = \sum_k r_{j,k}^2$ is the LD Score of variant *j*, which measures the amount of genetic variation tagged by *j*. As χ_j^2 are available through GWAS summary statistics, $l(j)$ can be computed using a reference sequencing dataset (here 1000 Genomes), and *N* and *M* are known, the parameters h^2 and *b* can be estimated through a simple linear regression.

LDSC analyses consider 3 sets of SNPs:

- Reference SNPs (M): this is the set of sequenced SNPs used to compute the LD scores. Here, we will use ~10M SNPs with an allele count ≥ 5 in the Europeans of 1000 Genomes.
- Regression SNPs: this is the set of GWAS SNPs used to perform the regression and to estimate the parameters h^2 and *b*. This set consists of ~1M SNPs that are in HapMap3 (used as a proxy for good imputation accuracy in the GWAS data), and not in the MHC region.
- Heritability SNPs: this is the set of reference SNPs used to estimate heritability. This set consists of ~6M SNPs with MAF ≥ 5%.

I-b/ Reference files

All the files requested for LDSC (and S-LDSC) analyses are in the `/data/LDSC_REF` folder.
Reference files are in `/data/LDSC_REF/plink_files`, and LD scores are in `/data/DSC_REF/ldscores`.
The `/data/LDSC_REF/plink_files` directory contains plink files (bed, bim, fam and frq files) for the reference genomes (~10M SNPs on ~500 individuals) on 22 chromosomes.
The `/data/LDSC_REF/ldscores` directory contains LD scores (in l2.ldscore.gz files), number of reference SNPs (in l2.M files) and number of heritability SNPs (in l2.M_5_50 files) on 22 chromosomes.

Type the following commands:

```
more /data/LDSC_REF/ldscores/LDscore.1.l2.M
more /data/LDSC_REF/ldscores/LDscore.1.l2.M_5_50
zcat /data/LDSC_REF/ldscores/LDscore.1.l2.ldscore.gz | head
zcat /data/LDSC_REF/ldscores/LDscore.1.l2.ldscore.gz | wc -l
```

Questions:

1) *What is the number of reference SNPs on chromosome 1? What is the number of heritability SNPs on chromosome 1?*
   [HINT – reference SNPs] Look here: `ls -l /data/LDSC_REF/ldscores/*.M`
   [HINT – heritability SNPs] Look here: `ls -l /data/LDSC_REF/ldscores/*.M_5_50`
2) *For how many SNPs do we have LD scores?*
   [HINT] LD score files: `ls -l /data/LDSC_REF/ldscores/*.ldscore.gz`

I-c/ Run LDSC

Type the following command:

```
ldsc  --h2 /data/LDSC_REF/UKB_460K.body_HEIGHTz.Neff.sumstats \
      --ref-ld-chr /data/LDSC_REF/ldscores/LDscore. \
      --w-ld-chr /data/LDSC_REF/weights_hm3_no_MHC/weights.hm3_noMHC. \
      --out UKB_460K.body_HEIGHTz.ldsc
```

This command takes as inputs summary statistics (though --h2), LD scores (through --ref-ld-chr) and regression weights (through --w-ld-chr; these weights correct for the fact that we are running a regression on non-independent observations).

Questions: [HINT] `cat UKB_460K.body_HEIGHTz.ldsc.log`

1) *How many SNPs have been used in the regression?*
2) *What is the intercept? Is it significantly greater than 1?*
3) *What is the heritability estimate? (Note that this model assumes the same per-SNP heritability for every trait, which lead to underestimate of heritability estimates; see next section)*

II/ Stratified LD score regression tutorial

II-a/ Some background

Stratified LD score regression (S-LDSC) is an approach to partition heritability across functional annotations from GWAS summary statistics (Finucane et al. 2015 *Nat Genet*). Briefly, under certain assumptions, we can write the expected χ_j^2 statistic of variant *j* as

$$E[\chi_j^2] = N\sum_c \tau_c l(j,c) + Nb + 1$$

where $l(j,c) = \sum_k a_c(k) r_{jk}^2$ is the LD score of SNP j with respect to continuous values $a_c(k)$ of annotation a_c, r_{jk} is the correlation between SNP j and k in a reference panel, and τ_c is the effect size of annotation a_c on per-SNP heritability (conditioned on all other annotations). The recommended set of annotations to use with S-LDSC is called the baseline-LD model (Gazal et al. 2017 *Nat Genet*).

II-b/ The baseline-LD model

The annotations and corresponding LD scores from the baseline-LD model are in `/data/LDSC_REF/baselineLD_v2.2`. This directory contains annotations (in annot.gz files), LD scores (in l2.ldscore.gz files), number of reference SNPs (in l2.M files) and number of heritability SNPs (in l2.M_5_50 files) on 22 chromosomes.

Type the following commands:

```
zcat /data/LDSC_REF/baselineLD_v2.2/baselineLD.1.annot.gz | head | less -S
zcat /data/LDSC_REF/baselineLD_v2.2/baselineLD.1.annot.gz | head -1 | sed s/\\t/\\n/g
zcat /data/LDSC_REF/baselineLD_v2.2/baselineLD.1.annot.gz | head -1 | sed s/\\t/\\n/g | wc -l
zcat /data/LDSC_REF/baselineLD_v2.2/baselineLD.1.annot.gz | wc -l
```

Questions:

1) *What is the number of annotations in the baseline-LD model?*
2) *What is the number of SNPs annotated on chromosome 1?*

II-c/ Run S-LDSC

Type the following command:

```
ldsc  --h2 /data/LDSC_REF/UKB_460K.body_HEIGHTz.Neff.sumstats \
      --ref-ld-chr /data/LDSC_REF/baselineLD_v2.2/baselineLD. \
      --w-ld-chr /data/LDSC_REF/weights_hm3_no_MHC/weights.hm3_noMHC. \
      --frqfile-chr /data/LDSC_REF/plink_files/1000G.EUR.QC. \
      --overlap-annot \
      --print-coefficients \
      --out UKB_460K.body_HEIGHTz.sldsc
```

This command takes as inputs summary statistics (though --h2), LD scores (through --ref-ld-chr), regression weights (through --w-ld-chr), frequency files (through –frqfile-chr; these files are used to partition heritability for SNPs with MAF >= 5%), and compulsory options (--overlap-annot and --print-coefficients).

Questions: [HINT] `less -S UKB_460K.body_HEIGHTz.ldsc.log` and `column -t UKB_460K.body_HEIGHTz.ldsc.results | less -S`

1) *How many SNPs have been used in the regression*
2) *What is the intercept? Is it significantly greater than 1?*
3) *What is the heritability? How does it compare to the one of LDSC? Which one should you trust more?*
4) *What annotation is most enriched in heritability? (Be careful, enrichment outputs make sense only for binary or probabilistic annotations)*
5) *What is the annotation with the most significant regression coefficient?*

Bonus - Commands to compute genetic correlation:

Using GCTA:
(Details here: https://cnsgenomics.com/software/gcta/#BivariateGREMLanalysis)

```
gcta64 --grm mydata_allSNPs --pheno mydata.phen --reml-bivar 1 2
```

Using LDSC:

```
ldsc \
--rg
/data/LDSC_REF/PASS_Schizophrenia.sumstats,/data/LDSC_REF/PASS_Bipolar_Disorder
.sumstats \
--ref-ld-chr /data/LDSC_REF/ldscores/LDscore. \
--w-ld-chr /data/LDSC_REF/weights_hm3_no_MHC/weights.hm3_noMHC. \
--out Schizophrenia_BipolarDisorder
```

Appendix – installing `ldsc` and downloading LDSC files

*** This is NOT required for the practical as both GCTA and LDSC (and reference files) are already installed on the workshop server ***

The tutorial to install `ldsc` software is at https://github.com/bulik/ldsc. To download all the files used in this tutorial, copy the following lines:

```
URL=https://storage.googleapis.com/broad-alkesgroup-public/LDSCORE

#Download reference files
wget $URL/1000G_Phase3_frq.tgz
```

```
wget $URL/1000G_Phase3_plinkfiles.tgz
wget $URL/1000G_Phase3_weights_hm3_no_MHC.tgz
wget $URL/1000G_Phase3_ldscores.tgz
wget $URL/1000G_Phase3_baselineLD_v2.2_ldscores.tgz

#Untar the files
tar zxvf 1000G_Phase3_frq.tgz
tar zxvf 1000G_Phase3_plinkfiles.tgz
tar zxvf 1000G_Phase3_weights_hm3_no_MHC.tgz
tar zxvf 1000G_Phase3_ldscores.tgz
tar zxvf 1000G_Phase3_baselineLD_v2.2_ldscores.tgz
rm *tgz

#Organize and rename directories
mv 1000G_EUR_Phase3_plink plink_files
mv 1000G_Phase3_frq/* plink_files; rm -rf 1000G_Phase3_frq
mv 1000G_Phase3_weights_hm3_no_MHC weights_hm3_no_MHC
mv LDscore ldscores
mkdir baselineLD_v2.2; mv baselineLD.* baselineLD_v2.2

#Download summary statistics
wget $URL/all_sumstats/UKB_460K.body_HEIGHTz.sumstats
wget $URL/all_sumstats/PASS_Schizophrenia.sumstats
wget $URL/all_sumstats/PASS_Bipolar_Disorder.sumstats
```

ANSWERS TO QUESTIONS + MORE Q&A

Part 1: GCTA-GREML tutorial

Exercise 1: Calculate the Genetic Relationship Matrix (GRM) using all SNPs

3) *How many individuals/SNPs do you have in the sample*: N=6,000 individuals and M=12,000 SNPs
4) *What is the mean of the diagonal elements of the GRM? Is that expected?*
   Mean diagonal of GRM = 1.00012. Yes, it is expected to be around 1, as this measures the genetic relationship of an individual with themselves.

Exercise 2: Calculate GRM using SNPs with MAF>0.05 and MAF<0.05

3) *How many SNPs have a MAF<0.05?* There 797 SNPs with MAF<0.05.
4) *Why is the variance of diagonal elements larger in the GRM based on SNPs with MAF<0.05?* The variance of GRM terms depends on the number of SNPs used to calculate it. The variance is larger because fewer SNPs were used.

Exercise 3: Identify (genetically) unrelated individuals

2) *How many individuals are unrelated in this sample? GRM cut-off 0.05? cut-off 0.025?* There are N=4808 unrelated individuals at a 0.05 GRM cut-off and N=92 unrelated individuals at a 0.025 GRM cut-off.

Exercise 4: GREML estimation of heritability with and without relatives

2) *Compare heritability estimates from these two analyses. What could explain the difference that you observe?* The heritability estimate is ~0.4 when relatives are included and ~0.25 without relatives. This could be explained by 1) unaccounted shared environmental effects between relatives which are inflating estimates of heritability. In practice other explanation may include 1) non-additive genetic effects or 2) partially tagging of causal variants by SNPs used to calculate the GRM.

Exercise 5: MAF-stratified heritability estimation

Questions: Compare heritability estimates from these two analyses

3) *What could explain the difference that you observe?* The difference between these two estimates can be explained by the fact that the MAF distribution of causal variants is different from that of SNPs used to calculate the GRM (Lecture – part 4).
4) *Which one of these estimates should you trust more? How can you be sure?* The heritability estimate from the stratified analysis should be trusted more because it is designed to minimize the MAF and LD heterogeneity between SNPs and causal variants. One potential strategy to check if we've done enough correction is to stratify SNPs into further MAF or LD bins and repeat the analysis.

Part 2: LD score regression tutorial

Command lines to answer the in the case of the UKBB height summary statistics 458K questions:

What is the number of reference SNPs? What is the number of heritability SNPs?

```
wc -l /data/LDSC_REF/plink_files/*bim # 9997231
awk '{if($5>=0.05) {print $0} }' /data/LDSC_REF/plink_files/1000G.EUR.QC.*.frq
| wc -l # 5961181-22
```

How many reference SNPs there are on chromosome 1? For how many SNPs do we have LD scores?

```
wc -l /data/LDSC_REF/plink_files/1000G.EUR.QC.1.bim # 779354
zcat /data/LDSC_REF/ldscores/LDscore.1.l2.ldscore.gz | wc -l # 98643-1
```

How many SNPs have been used in the regression?

```
1187056
```

What is the intercept? Is it significantly greater than 1?

```
1.8704 (0.0433). Yes, as the confidence interval (1.8704+/-1.96*0.0433) does
not include 1.
```

What is the heritability estimate? (Note that this model assumes the same per-SNP heritability for every trait, which lead to underestimate of heritability estimates; see next section)

```
0.3811 (0.0191)
```

What is the number of annotations in the baseline-LD model?

```
zcat /data/LDSC_REF/baselineLD_v2.2/baselineLD.1.annot.gz | head -1 | sed
s/\\t/\\n/g | wc -l # 101-4=97
```

What is the number of SNPs annotated on chromosome 1?

```
zcat /data/LDSC_REF/baselineLD_v2.2/baselineLD.1.annot.gz | wc -l # 779355-1
```

How many SNPs have been used in the regression

```
1186575
```

What is the intercept? Is it significantly greater than 1?

```
1.4527 (0.0595). Yes, as the confidence interval does not include 1.
```

What is the heritability? How does it compare to the one of LDSC? Which one should you trust more?

```
0.471 (0.0188). Higher. This one.
```

What annotation is most enriched in heritability? (Be careful, enrichment outputs make sense only for binary or probabilistic annotations)

```
column -t UKB_460K.body_HEIGHTz.sldsc.results | less -S > non_synonymous
(Enrichment=17.06)
```

What is the annotation with the most significant regression coefficient?

```
MAFbin7 (Coefficient_z-score=5.71)
```

More Q&A about LDSC

Why are we performing regression on HapMap 3 SNPs?
GWAS used to release summary statistics for imputed SNPs without information about imputation accuracy. The early LD score papers made the hypothesis that SNPs available in HapMap3 are likely to have been genotyped in the GWAS study, or imputed with good accuracy.

Why do we have LD scores only for ~1M SNPs?
As it is recommended to perform a regression on HapMap3 SNPs (~1M SNPs), it is not necessary to store LD scores for all reference SNPs (~10M SNPs). However, LD scores of these 1M SNPs are computed using the 10M reference SNPs.

What are the weights used in the regression?
Regressions need to be performed on independent observations. As regression SNPs are in LD with each other, weights are needed to correct this effect, and are computed here as the inverse LD score computed on regression SNPs.

What does Neff mean?
Some summary statistics are computed with linear mixed model, which increases the association power and thus association statistics. The sample size N used in LD score regression should thus not be the sample size that has been used to compute the summary statistics (in the case of the UKBB height summary statistics 458K) but the effective sample size (in the case of the UKBB height summary statistics 674K).

Can we use heritability estimates from ldsc?
The authors of ldsc do not recommend to use heritability estimates outputted by vanilla ldsc, as these estimates are based under an unrealistic model leading to downward biases of heritability. However, using the baseline-LD model correct for most of these biases and can be used with extreme caution.

Should we expect to have the same heritability estimates between GCTA and ldsc?

No. GCTA estimates the heritability tagged by SNPs used in the analyses, whereas ldsc estimates the heritability causally explained by common reference SNPs (i.e. removes the effect of low-frequency variants tagged by common SNPs).

Why some numbers are extremely high in the .results file?
S-LDSC can include continuous annotations that improve model fit (i.e. LD-related annotations to model LD architectures). For these annotations, the outputs in some of the columns in the .results file does not make sense (i.e. Prop._SNPs, Prop._h2, or Enrichment), unless the annotation is probabilistic (i.e. values between 0 and 1). However, the columns related to the regression coefficients make sense.

Why some annotations have significant enrichment but non-significant regression coefficients?
Some annotations have significant enrichment but non-significant regression coefficients, for example coding. This can be explained by the fact that coding regions are enriched in heritability, but that this enrichment is due to its high overlap with other annotations, such as conserved or non-synonymous variants. This is useful do determine how informative is an enriched annotation compared to the other ones.