Genome-wide Association Studies

Practical 1: Cleaning

Data Use Agreement

- To maximize your learning experience, we will be working with genuine human genetic data

- Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

- Please email pctgadmin@imb.uq.edu.au to confirm that you agree with the following:

    - “I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

Cluster Access

- You have all been provided with login details to computing resources needed for the practical component

- An SSH terminal is needed to connect to the computing:

    - Windows: Install PuTTY
    - Hostname: as provided (203.101.228.xxx)
    - User: as provided
    - Check Connection > SSH > X11 > Enable X11 forwarding

    - Mac/Linux: Use the terminal
    - `ssh -X <user>@203.101.228.xxx`

PLINK

- PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale genotype/phenotype analyses in a computationally efficient manner.
- We are using PLINK 1.90.
- https://www.cog-genomics.org/plink/ ← comprehensive documentation of all options and file formats

```
[allan@analysis1 ~]$ plink
PLINK v1.90b6.26 64-bit (2 Apr 2022)           www.cog-genomics.org/plink/1.9/
(C) 2005-2022 Shaun Purcell, Christopher Chang   GNU General Public License v3

  plink <input flag(s)...> [command flag(s)...] [other flag(s)...]
  plink --help [flag name(s)...]

Commands include --make-bed, --recode, --flip-scan, --merge-list,
--write-snplist, --list-duplicate-vars, --freqx, --missing, --test-mishap,
--hardy, --mendel, --ibc, --impute-sex, --indep-pairphase, --r2, --show-tags,
--blocks, --distance, --genome, --homozyg, --make-rel, --make-grm-gz,
--rel-cutoff, --cluster, --pca, --neighbour, --ibs-test, --regress-distance,
--model, --bd, --gxe, --logistic, --dosage, --lasso, --test-missing,
--make-perm-pheno, --tdt, --qfam, --annotate, --clump, --gene-report,
--meta-analysis, --epistasis, --fast-epistasis, and --score.

"plink --help | more" describes all functions (warning: long).
```

Data

- Data for this practical is found in the directory:
  - `/data/module1/gwas/part1/`

- Three files:
  - `gwas.bed` → binary file containing all genotypes
    `gwas.bim` → information about SNP markers
    `gwas.fam` → information about individuals

Data

- `gwas.fam` → information about individuals

```
[allan@analysis1 ~]$ head /data/module1/gwas/part1/gwas.fam
7653762 7653762 0 0 2 -9
8144519 8144519 0 0 2 -9
2337680 2337680 0 0 2 -9
5219864 5219864 0 0 1 -9
1417721 1417721 0 0 1 -9
2371103 2371103 0 0 2 -9
472262 472262 0 0 1 -9
566177 566177 0 0 2 -9
8097907 8097907 0 0 2 -9
8738370 8738370 0 0 2 -9
```

Data

- `gwas.bim` → information about SNP markers

```
[allan@analysis1 ~]$ head /data/module1/gwas/part1/gwas.bim
1       rs3131972       0       752721  1       2
1       rs3115850       0       761147  1       2
1       rs12562034      0       768448  1       2
1       rs4040617       0       779322  2       1
1       rs4970383       0       838555  1       2
1       rs950122        0       846864  1       2
1       rs6657440       0       850780  2       1
1       rs13303101      0       862124  1       2
1       rs1110052       0       873558  2       1
1       rs3748592       0       880238  1       2
```

Using PLINK

- All commands are well documented on the website

- Basic command:
  - `plink --bfile <data prefix> --command`
  - `plink --bfile /data/module1/gwas/part1/gwas --...`

Per Individual Quality Control

- Reminder: there are five basic steps to removing “bad” individuals

    1) removal of individuals with excess missing genotypes
    2) removal of individuals with outlying homozygosity values
    3) remove of samples showing a discordant sex
    4) removal of related or duplicate samples, and
    5) removal of ancestry outliers

1) Excess Missing Genotypes

- Command:
  - plink --bfile /data/module1/gwas/part1/gwas --missing

```
[allan@analysis1 ~]$ plink --bfile /data/module1/gwas/part1/gwas --missing
PLINK v1.90b6.26 64-bit (2 Apr 2022)           www.cog-genomics.org/plink/1.9/
(C) 2005-2022 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plink.log.
Options in effect:
  --bfile /data/module1/gwas/part1/gwas
  --missing

64141 MB RAM detected; reserving 32070 MB for main workspace.
298255 variants loaded from .bim file.
11793 people (5351 males, 6442 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 11793 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.968674.
--missing: Sample missing data report written to plink.imiss, and variant-based
missing data report written to plink.lmiss.
```

1) Excess Missing Genotypes

- Output: plink.imiss:

```
[allan@analysis1 ~]$ head plink.imiss
      FID       IID MISS_PHENO   N_MISS   N_GENO   F_MISS
  7653762   7653762          Y    12517   298255  0.04197
  8144519   8144519          Y     8427   298255  0.02825
  2337680   2337680          Y    13300   298255  0.04459
  5219864   5219864          Y     9609   298255  0.03222
  1417721   1417721          Y     9415   298255  0.03157
  2371103   2371103          Y     9633   298255   0.0323
   472262    472262          Y     7739   298255  0.02595
   566177    566177          Y     9082   298255  0.03045
  8097907   8097907          Y     7707   298255  0.02584
```

2) Outlying Homozygosity Values

- Command: `--het`
- Output: plink.het

- Read the output file into R. Are there any outliers that should be removed?

3) Discordant Sex

- Command: `--check-sex`

```
[allan@analysis1 ~]$ plink --bfile /data/module1/gwas/part1/gwas --check-sex
PLINK v1.90b6.26 64-bit (2 Apr 2022)           www.cog-genomics.org/plink/1.9/
(C) 2005-2022 Shaun Purcell, Christopher Chang   GNU General Public License v3
Logging to plink.log.
Options in effect:
  --bfile /data/module1/gwas/part1/gwas
  --check-sex

64141 MB RAM detected; reserving 32070 MB for main workspace.
298255 variants loaded from .bim file.
11793 people (5351 males, 6442 females) loaded from .fam.
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 11793 founders and 0 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is 0.968674.
298255 variants and 11793 people pass filters and QC.
Note: No phenotypes present.
Error: --check-sex/--impute-sex requires at least one polymorphic X chromosome
locus.
```

4) Remove Related Samples

- **This takes a LONG time to run**

- Command: `--genome`
- Command: `--rel-cutoff`

5) Remove Ancestry Outliers

- **This takes a LONG time to run**

- Command: `--pca`
- Need a reference data set – e.g. 1000 Genomes

- A large number of protocols for this are available online.
- e.g. https://enigma.ini.usc.edu/protocols/genetics-protocols/ ← covers ancestry checks and imputation

Per Marker Quality Control

- Reminder: the four steps of marker quality control:

1) removal of SNPs with excess missing genotypes
2) removal of SNPs that deviate from Hardy-Weinberg equilibrium
3) removal of SNPs with low minor allele frequency
4) comparing allele frequency to known values

1) Excess Missingness

- Command: `--missing`
- Output: plink.lmiss

```
[allan@analysis1 ~]$ head plink.lmiss
 CHR          SNP   N_MISS   N_GENO   F_MISS
   1   rs3131972       59    11793 0.005003
   1   rs3115850      812    11793  0.06885
   1  rs12562034       19    11793 0.001611
   1   rs4040617       35    11793 0.002968
   1   rs4970383       15    11793 0.001272
   1    rs950122      143    11793  0.01213
   1   rs6657440       16    11793 0.001357
   1  rs13303101       13    11793 0.001102
   1   rs1110052       49    11793 0.004155
```

2) Hardy-Weinberg equilibrium

- Command: `--hardy`
- Output: plink.hwe

- Typical threshold is 10^{-6}. How many SNP will be removed?

3) Low Minor Allele Frequency

- Command: --freq
- Output: plink.frq

- How many SNP have minor allele frequency below 1%?

4) Comparing to Known Allele Frequencies

- Allele frequencies from a reference population are given in the file “reference_allele_frequencies.txt”
- Compare to frequencies calculated in previous step

- You will need to use an SFTP client to copy any generated figure across to your computer

Putting it all Together!

- Approach #1: Create files for individuals/SNPs you want to keep/remove

- Commands: `--keep / --remove` (individuals)
- Commands: `--extract / --exclude` (SNPs)
- Commands: `--make-bed --out <filename>` (output cleaned data to a new file)

Putting it all Together!

- Approach #2: Some PLINK commands allow you to do some of the filtering on the go
- Command example:

```
plink --bfile <path to data>
     --maf 0.01 --geno 0.05 --mind 0.05 --hwe 0.000001
     --out <new name>
```

- This does the cleaning of individuals with low genotyping rate, SNPs with low genotyping rate, HWE issues and low MAF in one go
- Is it a good idea to do both individual and SNP cleaning at the same time?
  -