GWAS association testing + genotype extraction

Connecting to the cluster

connect to the HPC:

```
# remember to use your actual username ssh
USERNAME@kennedy.st-andrews.ac.uk
```

(or use PuTTY on windows)

Workshop materials

we have installed software under:

```
/gpfs1/scratch/bioinf/BL4273/miniforge3/envs/gd5302/bin/
```

dataset available under:

```
/scratch/bioinf/gd5302/$USER/data/p1/01_dataset/
```

scripts for association test:

```
/scratch/bioinf/gd5302/data/p2/04_association_test
```

scripts for genotype construction:

```
/scratch/bioinf/gd5302/data/p2/05_genotype
```

Workshop materials

Update dataset by:

```
cd /scratch/bioinf/gd5302
cp data/p1/01_dataset/1kgeas_binary.txt ./USERNAME/data/p1/01_dataset/
```

Add new scripts by:

```
cd /scratch/bioinf/gd5302
cp -r data/p2 ./USERNAME/data/
```

Change `**USERNAME**` to your own!

[submit] 04_association_test/run_association_test.sh

Submit the script for execution:

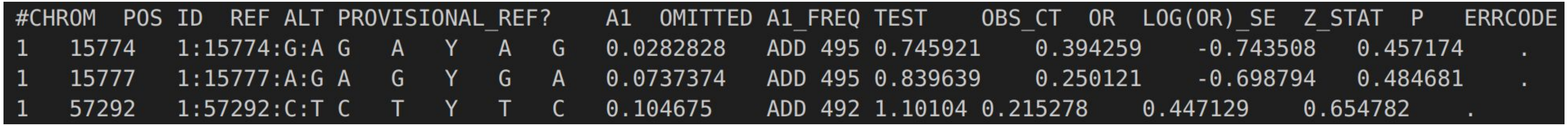
```
cd ./USERNAME/data/p2/04_association_test
sbatch run_association_test.sh
```

Change `**USERNAME**` to your own!

(make sure the previous script has finished first)

[output] 04_association_test/1kgeas.B1.glm.firth

head -n4 **1kgeas.B1.glm.firth**



```
#CHROM  POS ID  REF ALT PROVISIONAL_REF?    A1  OMITTED A1_FREQ TEST    OBS_CT  OR  LOG(OR)_SE  Z_STAT  P   ERRCODE
1   15774   1:15774:G:A G   A   Y   A   G   0.0282828   ADD 495 0.745921    0.394259    -0.743508   0.457174    .
1   15777   1:15777:A:G A   G   Y   G   A   0.0737374   ADD 495 0.839639    0.250121    -0.698794   0.484681    .
1   57292   1:57292:C:T C   T   Y   T   C   0.104675    ADD 492 1.10104 0.215278    0.447129    0.654782    .
```

This will be the input for the Manhattan plot

[submit] 04_association_test/run_manhattan-plot.sh

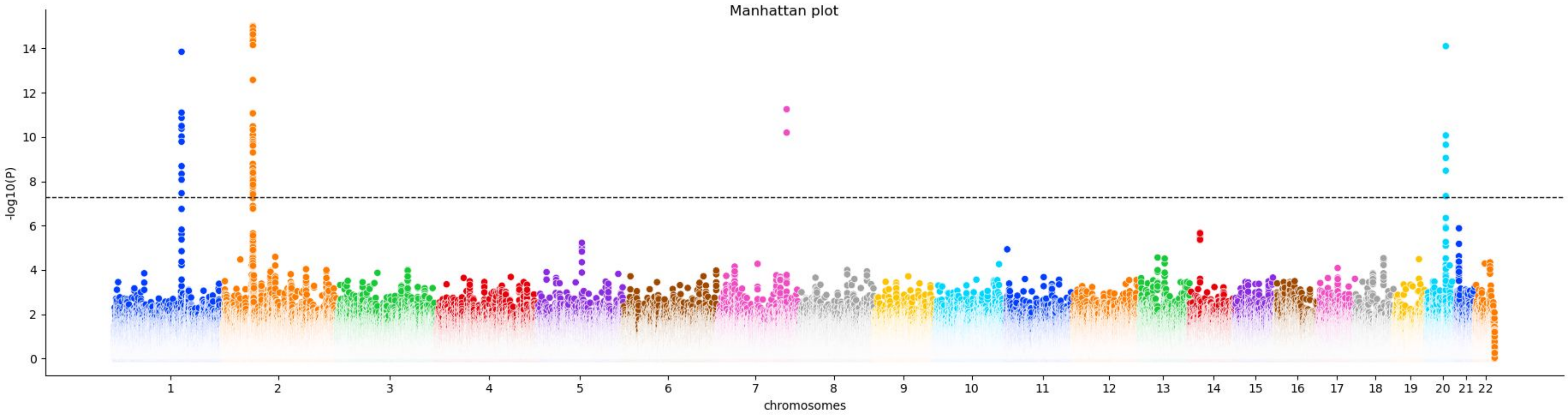
Submit the script for execution:

```
cd ./USERNAME/data/p2/04_association_test
sbatch run_manhattan-plot.sh
```

Change `**USERNAME**` to your own!

(make sure the previous script has finished first)

[output] 04_association_test/manhattan_plot.png



# How many SNP(ID) have P < 5 × 10E-8?

Transfer Manhattan plot from HPC (windows)

- launch pscp.exe (installed with PuTTY)

```
# remember to use your own username instead of `USERNAME`

pscp
USERNAME@kennedy:/scratch/bioinf/gd5302/USERNAME/data/p2/04_association_test/manhattan_plot.png%USERPROFILE%\ Documents\manhattan_plot.png
```

[submit] 05_genotype/extract_genotypes.sh

Submit the script for execution:

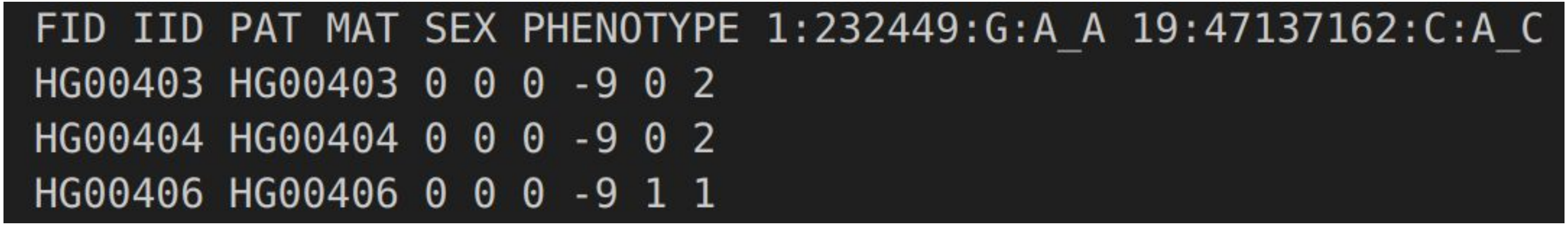
```
cd ./USERNAME/data/p2/05_genotype
sbatch extract_genotypes.sh
```

Change `**USERNAME**` to your own!

(make sure the previous script has finished first)

[output] 05_genotype/snp_genotypes.raw

head -n4 **snp_genotypes.raw**



```
FID IID PAT MAT SEX PHENOTYPE 1:232449:G:A_A 19:47137162:C:A_C
HG00403 HG00403 0 0 0 -9 0 2
HG00404 HG00404 0 0 0 -9 0 2
HG00406 HG00406 0 0 0 -9 1 1
```

This will be the input for the Distribution plot

[submit] 05_genotype/run_plot_genotypes.sh

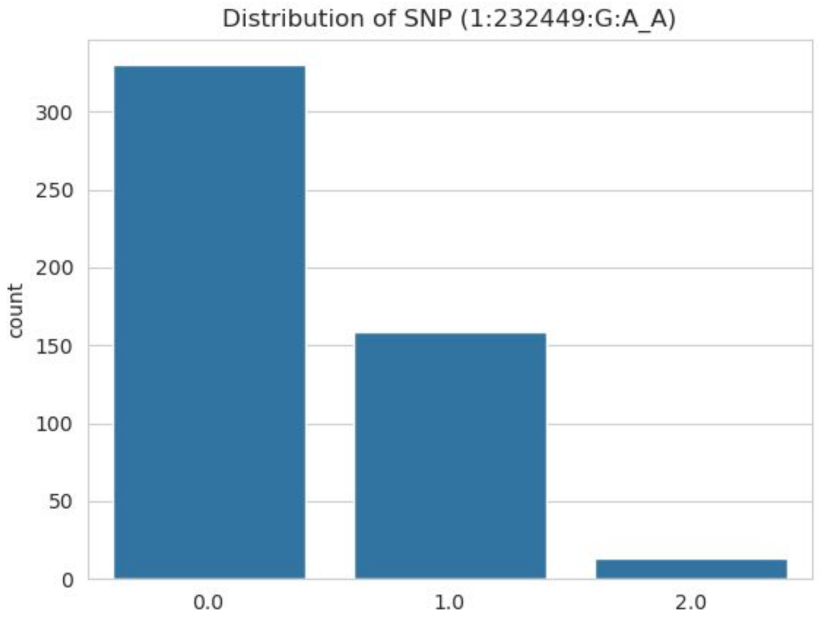
Submit the script for execution:

```
cd ./USERNAME/data/p2/05_genotype
sbatch run_plot_genotypes.sh
```

Change `**USERNAME**` to your own!

(make sure the previous script has finished first)

[output] 05_genotype/distribution_plot.2.png



# What's the counts for each genotype?

Transfer Distribution plot from HPC (windows)

- launch pscp.exe (installed with PuTTY)

```
# remember to use your own username instead of `USERNAME`

pscp USERNAME@kennedy:/scratch/bioinf/gd5302/USERNAME/data/p2/05_genotype/distribution_plot.2.png%USERPROFILE%\ Documents\distribution_plot.2.png
```