# 2023 International Statistical Genetics Workshop

## Day 2. Optional morning GWAS practical.<br>INSTRUCTIONS

March 7, 2023
Luke Evans & Wei Zhou
Practical originally developed by Katrina Grasby and Lucía Colodro Conde

**These INSTRUCTIONS contains the information for the GWAS practical.**

**Learning Goals:**
1. Run a GWAS using real data
2. Familiarize yourself with the standard types of input and output for a GWAS
3. Practice using the command line
4. Improve your familiarity plink2, a common tool in genetics (beyond GWAS, too)
5. Interpret why peaks form in a Manhattan plot, and how to explore the genome around associated loci

This tutorial aims to make you familiar with genome-wide association analysis, the way in which PLINK2 works (and some of the useful options it provides!), and some tools to visualize your results.

It is important that you explore the outputs and the log files and that you understand what they mean. Please ask if you have any questions, and check the plink file formats page.

**Things to note during this practical:**
A. There is a plain text file of the instructions alone on the server. You can directly copy and paste from that into the terminal or R. You can also type in the commands directly.
B. There are questions throughout these instructions. These are meant to have you look through the files and output, and think about the commands, what they're doing and how they're operating.

```
Commands in plink are highlighted in blue
```

C. Most of the options that you will be using work the same for logistic and linear regression.
D. *Some of the commands below are missing (indicated with XXX).* You will have to determine the correct syntax or command to run the analyses, and discuss the questions proposed.
E. You will find the commands and their output in Day2_GWAS_Practical_RUBRIC.pdf.
Feel free to take a look through it during the practical, but a good strategy is to
1. try the commands yourself or with your group
2. troubleshoot things yourself and ask your group
3. check the rubric
4. ask others & faculty

The rubric has all the expected answers and figures, but the process of doing the GWAS and exploring the output is what will help you learn.

---

**DAY 2. GWAS TUTORIAL**

Key information about PLINK2 can be found here:
https://www.cog-genomics.org/plink/2.0/
https://www.cog-genomics.org/plink/2.0/assoc
https://www.cog-genomics.org/plink/2.0/formats

**FIRST: Copy the files to your working directory & navigate to that directory:**

```
cp -r /faculty/luke/2023/gwas2/ ./
cd gwas2/
```

## EXERCISE 1. LOGISTIC REGRESSION (BINARY TRAIT).

**1.0. Go to the case-control folder and check the files you have there.** The phenotype is Alzheimer's Disease (AD) status.
Source: https://med.miami.edu/faculty/amanda-myers-phd &
https://xzmxbgsv808roffneicreq.on.drv.tw/www.lfun/LFUN/LFUN/DATA.html

```
cd casecontrol
```

**1.1. Run a logistic regression for the phenotype (AD status) including the principal components in file adpc.txt as covariates to correct for genetic ancestry.**

- [hint: you can find the flags and modifiers that you can include in the logistic and linear regressions on Association analysis > Regression with covariates, https://www.cog-genomics.org/plink/2.0/assoc#glm in the PLINK 2.0 website]

```
plink2 --bfile adclean.cc --glm sex --covar adpc.txt --out 1.1_logistic.ad.cc --threads 4
```

**Check the log file.** How many cases and controls were detected? How many covariates?

```
less -S 1.1_logistic.ad.cc.log
```

**Check the results (stored in 1.1_logistic.ad.cc.PHENO1.glm.logistic.hybrid)** and be sure that you understand the content of each of the columns.

How many lines are in the file?

```
wc -l 1.1_logistic.ad.cc.PHENO1.glm.logistic.hybrid
```

Look at the first few lines of the output:

```
head 1.1_logistic.ad.cc.PHENO1.glm.logistic.hybrid
```

What is the OR of the second variant and the corresponding p-value?
What are the different values in the “TEST” column, and what do they tell you? Which is the one you’re interested in for a GWAS?
[hint: you can check File formats > .glm.logistic, https://www.cog-genomics.org/plink/2.0/formats#glm_logistic, on the PLINK2.0 website].

**Did any SNP associations reach genome-wide significance?**

```
grep ADD 1.1_logistic.ad.cc.PHENO1.glm.logistic.hybrid | awk '$13<5e-8'
or
sort -k13 -g 1.1_logistic.ad.cc.PHENO1.glm.logistic.hybrid | head ### This just sorts it by p-value & prints out the first few lines
```

What is the grep command doing here?
What does the awk command do?
What is "|" doing?

**What if cases and controls had been coded as 1 and 0,** respectively? What could have we done to make PLINK interpret this coding appropriately?

[hint: you can check Standard data input > Phenotypes > Phenotype encoding, https://www.cog-genomics.org/plink/2.0/input#pheno, in the PLINK 2.0 website]

**What if sex had not been coded in the fam file, or was in both your covariate and fam files**?

What could have we done to make PLINK interpret this coding appropriately?

[hint: you can check --glm modifiers in https://www.cog-genomics.org/plink/2.0/assoc#glm, in the PLINK 2.0 website]

**1.2. Run a logistic regression for the case-control variable AD including the principal components as covariates and hiding the results of the covariates.**

```
plink2 --bfile adclean.cc --glm sex hide-covar --covar adpc.txt --out 1.2_adclean.cc
```

What's the difference between the sets of results generated in 1.1 and 1.2?

**1.3. Run a logistic regression for the case-control variable AD including the principal components as covariates, hiding the results of the covariates, and getting 95% Confidence intervals of odds ratios.**

What is a confidence interval?

[hint: you can check again https://www.cog-genomics.org/plink/2.0/assoc#glm, in the PLINK 2.0 website]

```
plink2 --bfile adclean.cc --glm sex hide-covar --ci 0.95 --covar adpc.txt --out 1.3_adclean.cc --threads 4
```

**1.4. Run a logistic regression for the case-control variable AD including the principal components as covariates, hiding the results of the covariates, and getting the allele frequencies**.

[hint: you can check Main functions > Basic statistics > --freq or Allele frequency, https://www.cog-genomics.org/plink/1.9/basic_stats#freq, in the PLINK 1.9 website].

```
plink2 --bfile adclean.cc --glm sex hide-covar --covar adpc.txt --freq --out 1.4_adclean.cc --threads 4
```

## EXERCISE 2. LINEAR REGRESSION (CONTINUOUS TRAIT)

**2.0. Go to the continuous folder and check the files you have there.** The phenotype is a transcript probe (gene expression). Pay attention to the file adclean.cont.txt.

```
cd ../continuous/
head adclean.cont.txt
```

**2.1. Run a linear regression for the continuous trait including the genetic principal components as covariates, hiding the results of the covariates, and using the --pheno option.** The advantage of using an extra file for phenotypes (and the --pheno option) is that if there were several phenotypes, it would be possible to run analyses on them at the same time (something not possible with ped or fam files). Note that when using the --pheno option, the original ped or fam files can still contain a phenotype in the phenotypc column, but it can be missing (NA/nan/-9). See Standard data input > Phenotypes, https://www.cog-genomics.org/plink/2.0/input#pheno.

```
plink2 --bfile adclean.cont --glm sex hide-covar --pheno adclean.cont.txt --covar adpc.txt --out
2.1_adclean.cont --threads 4
```

**2.2. Run a linear regression for the continuous trait including only PC1 as covariate, hiding the results of the covariate, using the --pheno option.**

# [hint: again, you can check the commands you can use to run logisic or linear regressions on data input > Covariates,  https://www.cog-genomics.org/plink/2.0/input#pheno. For options related to how parameters are handled in other models (e.g., estimating non-additive effects), you can seehttps://www.cog-genomics.org/plink/2.0/assoc

```
plink2 --bfile adclean.cont --glm hide-covar --pheno adclean.cont.txt --covar adpc.txt --covar-name PC1
--out 2.2_adclean.cont --threads 4
```

**2.3. Run a linear regression for the continuous trait including the principal components as covariates, hiding the results of the covariates, using the --pheno option, and getting 95% confidence intervals for the beta.**

```
plink2 --bfile adclean.cont --glm sex hide-covar --pheno adclean.cont.txt --covar adpc.txt --ci 0.95 --out 2.3_adclean.cont --threads 4
```

What is different about these results than 2.1 results? How might you use the added information?

**2.4. Plot the results from 2.3.**

We will create three plots to explore the results. For that, we will need at least three columns: chromosome, base pair position, and p-value; having a SNP column (containing the rs number) will allow extra options in our plots. We'll **first create a file containing this information, excluding the markers with no results**.

What is "|" doing in the next line? What does the grep command do, and why would you want to use that?

```
awk '{print $1,$2,$3,$14}' 2.3_adclean.cont.GI_34147330-S.glm.linear | grep -v NA > plot.adclean.cont.linear.txt
```

**What's the SNP with the lowest p-value (or top SNP)?**

```
sort -k4 -g plot.adclean.cont.linear.txt | head
```

What is the "-g" doing in the line above? Hint: use "man sort" to check the documentation of commands.

**EXERCISE 3. PLOT OUT YOUR RESULTS**

You'll next make your own Q-Q and Manhattan plots from these data, then look at the region around the strongest association.

**QQ & Manhattan Plots**

**Open in RStudio ([https://workshop.colorado.edu/rstudio/](https://workshop.colorado.edu/rstudio/)).**

**Open the script Rscript_qqMan.R (in the folder you copied) and plot the results.**

1. Set your working directory to the folder where you have your.
2. Run the script to create both plots.

What is the expected distribution of the p-values? Which plot is comparing the expected to the observed?
Can you locate the strongest association from your work above in both plots?
What is the y-axis of the Manhattan plot and why is it plotted in those units?
What information do you gather from these plots and the lambda value? Do you detect any anomalies?

> You need to set your working directory first. This is either by specifying the path in the script or by using Session>Set Working Directory>Choose Directory in the drop-down menu at the top of RStudio.
> You also need to tell it which file to use, here, "plot.adclean.cont.linear.text'.
> Note that while this file does have a header column, it starts with a "#" which comments out the line and which RStudio doesn't recognize. So, you then have to set the names of the column.
> Checking the structure uses str(Data), and will output information for each column in the "Data" object.

**Make a regional plot of the strongest association.**

**1.** Still in RStudio, go to the "Files" section (bottom right quadrant of the browser) and check the file plot.adclean.cont.linear.txt and the two jpg images you just created.

**2.** Go to More > Export. The file will download to your local computer, and you will need to unzip it.

> There is also a copy of this same file on the shared files section on the server if you have any trouble exporting this from Rstudio: Day2_plot.adclean.cont.linear.txt

**3.** Upload your results to LDassoc in LDlink: https://ldlink.nci.nih.gov/?tab=ldassoc.

> You can explore association p-value results and LD patterns using this tool.

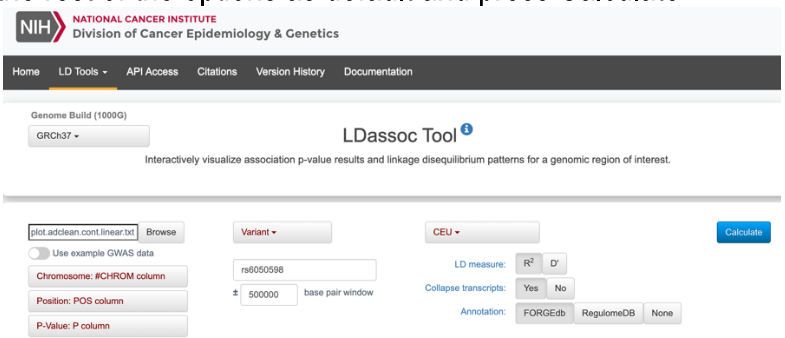
Upload your results using Browser (preferred: Chrome, Safari, Firefox 36+ or Internet Explorer). In real scenarios, you may want to upload only the information of a region of interest that you want to plot or only one chromosome because the upload time will be slow with very large files.

**4.** Select the columns that contain the information on Chromosome, Position, and P-Value.

**5.** Select the variant you want to center the plot on, and select the 1000 Genomes sub-population to plot recombination rates surrounding that variant.

LDassoc has three options for visualizing regions of association: by gene, by region or by variant. We will select visualising our results by using our lowest p-value variant as the index variant (rs449370). Select the CEU population (European, Utah Residents from North and West Europe) as the 1000 Genomes sub-population of interest. LDlink will calculate measures of linkage disequilibrium according to this population, which is the one that best matches the ancestry of our study population.

Leave the rest of the options as default and press Calculate.



Genome Build (1000G): GRCh37

LDassoc Tool

Interactively visualize association p-value results and linkage disequilibrium patterns for a genomic region of interest.

plot.adclean.cont.linear.txt

Use example GWAS data

Chromosome: #CHROM column

Position: POS column

P-Value: P column

Variant

rs6050598

± 500000 base pair window

CEU

LD measure: $R^2$ D'

Collapse transcripts: Yes No

Annotation: FORGEdb RegulomeDB None

**5.** Explore the interactive plot.

Is the index variant in high LD with any of the nearby variants? Is it around areas of high or low recombination?

If LD were much stronger and widespread, how would it change this plot?

What if the strongest locus was not in LD with any other loci? How would that change your plot?

Remember the R2 is a measure of linkage disequilibrium or correlation of alleles for two genetic variants.

**If you have gotten this far with time left, here are a couple of other challenges:**

**1.** Go back to your logistic regression results and plot them:

1.5. Plot the results from 1.4. You will need to use the commands you ran in section two to generate the intermediate files for RStudio and the interactive browser plots. Your challenge will be to adapt those commands to the right files in the right sequence, run the RStudio script again, and create the plots.

**2.** What does including the PC covariates do? Try running it without any PC covariates.

Why would you want to include PCs in your analysis?