# 2021 International Statistical Genetics Workshop

## DAY 3. TUTORIAL – PART 2. SOLUTIONS.

Key information can be found here (links accessed on May 2021).
https://www.cog-genomics.org/plink/1.9/assoc#linear
https://www.cog-genomics.org/plink/1.9/formats
https://www.cog-genomics.org/plink/1.9/index

See also the file day3_part2_cheatseet.docx

## 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.**

```
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.**

```
plink --bfile adclean.cc --logistic --covar adpc.txt --out 1.1_adclean.cc
```

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

PLINK detected 170 cases, 182 controls, and 4 covariates, as detailed in the log file.

**Check the results (stored in 1.1_adclean.cc.assoc.logistic) and that you know the content of each of the columns. Remember, when PLINK 1.9 uses bfiles, the effect allele is A1, which is the minor allele. So an OR > 1 means A1 is associated with an increased risk relative to A2.**

The description of the output can be found on: https://www.cog-genomics.org/plink/1.9/formats#assoc_linear

**.assoc.linear, .assoc.logistic (multi-covariate association analysis report)**

Produced by --linear/--logistic.

A text file with a header line, and **T** lines per variant typically with the following nine fields (where **T** is normally the number of terms, but the 'genotypic' and 'hethom' modifiers and the --tests flag can change this):

| | |
|---|---|
| *CHR* | Chromosome code. *Not present with 'no-snp' modifier.* |
| *SNP* | Variant identifier. *Not present with 'no-snp'.* |
| *BP* | Base-pair coordinate. *Not present with 'no-snp'.* |
| *A1* | Allele 1 (usually minor). *Not present with 'no-snp'.* |
| TEST | Test identifier |
| NMISS | Number of observations (nonmissing genotype, phenotype, and covariates) |
| 'BETA'/'OR' | Regression coefficient (--linear, "--logistic beta") or odds ratio (--logistic without 'beta') |
| STAT | T-statistic |
| P | Asymptotic p-value for t-statistic |

If --ci 0.*xy* has also been specified, the following three fields are inserted before 'STAT':

| | |
|---|---|
| SE | Standard error of beta (log-odds) estimate |
| L*xy* | Bottom of *xy*% symmetric approx. confidence interval |
| H*xy* | Top of *xy*% approx. confidence interval |

Refer to the PLINK 1.07 documentation for more details.

**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?**

By default, case/control phenotypes are expected to be encoded as 1=unaffected (control), 2=affected (case); 0 is accepted as an alternate missing value encoding. If you use the --1 flag, 0 is interpreted as unaffected status instead, while 1 maps to affected. This also forces phenotypes to be interpreted as case/control. See: https://www.cog-genomics.org/plink/1.9/input#pheno

**Phenotype encoding**

```
--missing-phenotype <integer>
--1
```

Missing phenotypes are normally expected to be encoded as -9. You can change this to another integer with **--missing-phenotype**. (This is a slight change from PLINK 1.07: floating point values are now disallowed due to rounding issues, and nonnumeric values such as 'NA' are rejected since they're treated as missing phenotypes no matter what. Note that --output-missing-phenotype can be given a nonnumeric string.)

Case/control phenotypes are expected to be encoded as 1=unaffected (control), 2=affected (case); 0 is accepted as an alternate missing value encoding. If you use the **--1** flag, 0 is interpreted as unaffected status instead, while 1 maps to affected. *This also forces phenotypes to be interpreted as case/control.*

**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.**

plink --bfile adclean.cc --logistic 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?**

The results in 1.1 include the tests on the covariates while the results in 1.2 have the covariate-specific lines removed.

**Where can we find the SNP-phenotype association after controlling for the covariates?**

In both files, the p-values on the first row per SNP (ADD) represent the test for the SNP-phenotype association after controlling for the covariate/s. The covariate rows show the test associated with each of the covariate-phenotype associations. See: https://zzz.bwh.harvard.edu/plink/faq.shtml#faq11

**When I include covariates with `--linear` or `--logistic`, what do the p-values mean?**

If one or more covariates are included (by `--covar`) when using `--linear` or `--logistic`, PLINK performs a multiple regression analysis and reports the coefficients and p-values for each term (i.e. SNP, covariates, any interaction terms). The only term omitted from the report is the intercept.

The p-values for the covariates **do not** represent the test for the SNP-phenotype association after controlling for the covariate. That is the first row (`ADD`). Rather, the covariate term is the test associated with the covariate-phenotype association. These p-values might be extremely significant (e.g. if one covaries for smoking in an analysis of heart disease, etc) but this does not mean that the SNP has a highly significant effect necessarily. For example:

| CHR | SNP | BP | A1 | TEST | NMISS | BETA | STAT | P |
|---|---|---|---|---|---|---|---|---|
| 1 | rs1234567 | 742429 | G | ADD | 1495 | -0.03335 | -0.1732 | 0.8625 |
| 1 | rs1234567 | 742429 | G | COV1 | 1495 | 0.1143 | 9.748 | 8.321e-022 |

suggests that the covariate is highly correlated with the outcome (which will often be already known, presumably), but there is no evidence that the SNP is in any way correlated with phenotype. These correspond to the partial regression coefficient terms of a muliple regression

```
Y ~ m + b1.ADD + b2.COV1 + e
```

where p=0.8625 is the Wald test for `b1`, p=8e-22 is the Wald test for `b2`, the covariate-phenotype relationship. To repeat: it does not mean that the SNP-phenotype test has a p=8e-22 after controlling for `COV1`.

**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 regression coefficients (betas) instead of odds ratios.**

plink --bfile adclean.cc --logistic hide-covar beta --covar adpc.txt --out 1.3_adclean.cc

**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.**

plink --bfile adclean.cc --logistic hide-covar --covar adpc.txt --freq --out 1.4_adclean.cc

**Explore the output files and note you have an extra one, 1.4_adclean.cc.frq. In this one, we can see A1, A2, and MAF. We know that A1 is the effect allele or the tested allele. A1 is the minor allele by default when we work with a bfile format. PLINK gives you the MAF, which is the allele frequency for A1. For bi-allelic variants, the allele frequency for A2, should we need it, could be calculated by subtraction (1-MAF).**

The description of the content of the freq file can be found on https://www.cog-genomics.org/plink/1.9/formats#frq

### .frq (basic allele frequency report)

Produced by --freq. Valid input for --read-freq.

A text file with a header line, and then one line per variant with the following six fields:

| | |
|---|---|
| CHR | Chromosome code |
| SNP | Variant identifier |
| A1 | Allele 1 (usually minor) |
| A2 | Allele 2 (usually major) |
| MAF | Allele 1 frequency |
| NCHROBS | Number of allele observations |

**Note that we could get the allele frequencies separately for cases and controls by adding the "case control" modifier to --freq; it'd look like:**

**plink --bfile adclean.cc --logistic hide-covar beta --covar adpc.txt --freq case-control --out 1.5_adclean.cc**

See https://www.cog-genomics.org/plink/1.9/basic_stats#freq

### Allele frequency

```
--freq [{counts | case-control}] ['gz']
--freqx ['gz']
  (alias: --frqx)
```

By itself, **--freq** writes a minor allele frequency report to `plink.frq`. If you add the '**counts**' modifier, an allele count report is written to `plink.frq.count` instead. Alternatively, you can use --freq with --within/--family to write a cluster-stratified frequency report to `plink.frq.strat`, or use the '**case-control**' modifier to write a case/control phenotype-stratified report to `plink.frq.cc`.

And https://www.cog-genomics.org/plink/1.9/formats#frq_cc

### .frq.cc (case/control phenotype-stratified allele frequency report)

Produced by "--freq case-control". *Not* valid input for --read-freq.

A text file with a header line, and then one line per variant with the following eight fields:

| | |
|---|---|
| CHR | Chromosome code |
| SNP | Variant identifier |
| A1 | Allele 1 (usually minor) |
| A2 | Allele 2 (usually major) |
| MAF_A | Allele 1 frequency in cases |
| MAF_U | Allele 1 frequency in controls |
| NCHROBS_A | Number of case allele observations |
| NCHROBS_U | Number of control allele observations |

**1.5. Plot the results from 1.4.**

**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.**

```
awk '{print $1,$2,$3,$9}' 1.4_adclean.cc.assoc.logistic | grep -v NA >
plot.adclean.cc.logistic.txt
```

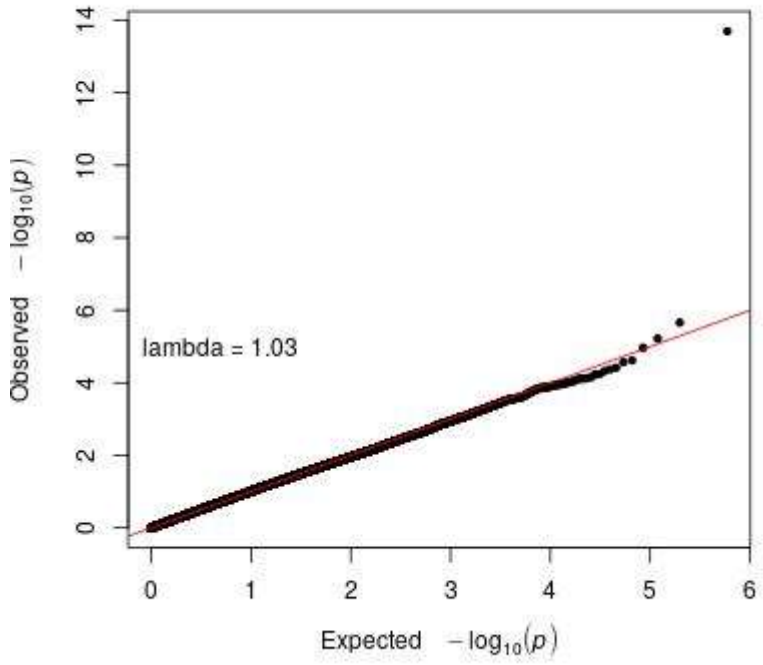
**What's the SNP with the lowest p-value (or top SNP)?**

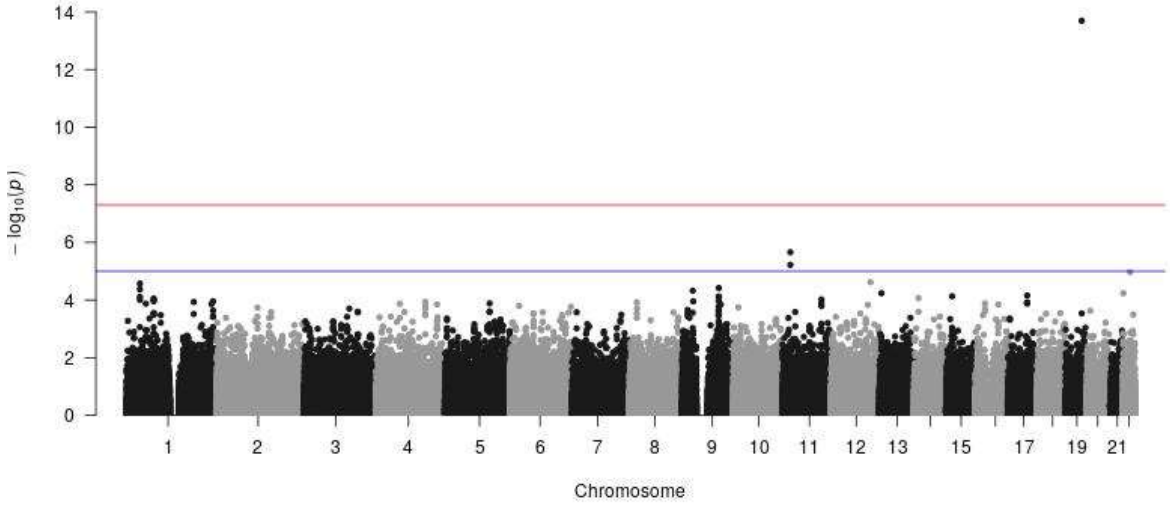
```
sort -k4 -g plot.adclean.cc.logistic.txt | head
```

The SNP with the lowest p-value is rs4420638 (or chr19:45422946).

**QQ and Manhattan plots**

**Open in RStudio (https://workshop.colorado.edu/rstudio/) the script Rscript_qqMan.R and plot the results. Set your working directory to the folder where you have your results (you can check first what your working directory is by typing pwd in the command line). Run the script. Explore the QQ plot and the Manhattan plot. What information do you gather from these plots and the lambda value? Do you detect any anomalies?**



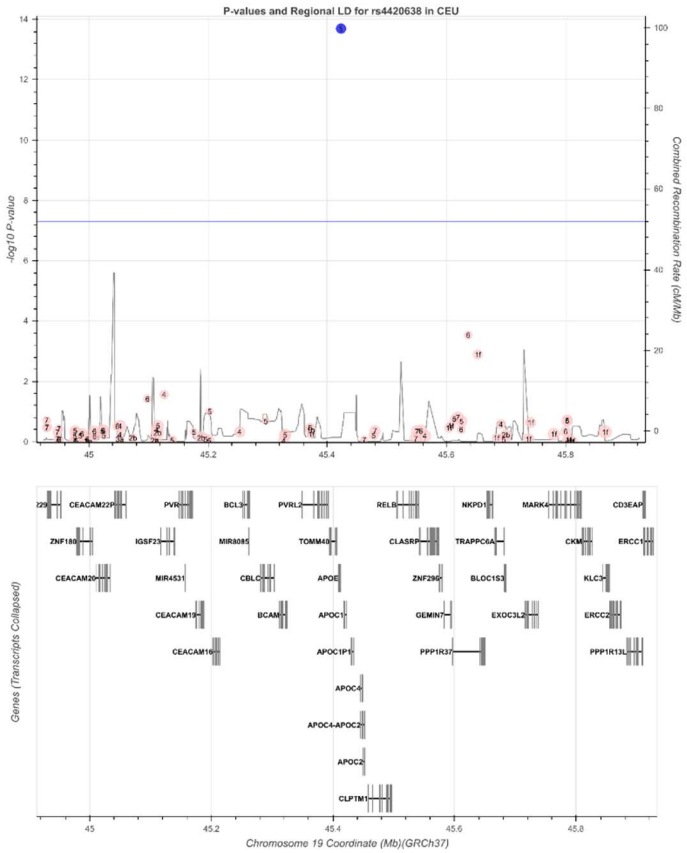



The QQ plot doesn't show signs of systematic bias or inflation (most of the points are on the diagonal) and the lambda value is close to 1, indicating no inflation. The Manhattan plot looks healthy too, although the SNP with the lowest p-value is "floating" and may not be a real association - we usually expect that some of its neighbouring SNPs in LD were also associated with the phenotype. Remember this is genotyped data and we are missing information about multiple SNPs, some of them could be on LD with this one.

**Regional plot**

**Still in RStudio, go to the "Files" section and check the file plot.adclean.cc.logistic.txt. Then go to More and then Export. The file will download to your local computer.**
**You can explore association p-value results and LD patterns by uploading your results (or a genomic region of interest) to LDassoc in LDlink: https://ldlink.nci.nih.gov/?tab=ldassoc**
**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.**
**Select the columns that contain the information on Chromosome, Position, and P-Value.**
**LDassoc has three options for visualising 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 (rs4420638 or chr19:45422946).**
**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.**

**Explore the interactive plot. Is the index variant in high LD with any of the nearby variants? Remember the R2 is a measure of linkage disequilibrium or correlation of alleles for two genetic variants.**




There are no variants in linkage disequilibrium observed in this plot; if there were, we would expect they would be also associated with the phenotype and therefore they would have a lower p-value.

## 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/
```

**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 must still contain a phenotype Column 6.**

```
plink --bfile adclean.cont --linear hide-covar --pheno adclean.cont.txt --covar adpc.txt --out 2.1_adclean.cont
```

See https://www.cog-genomics.org/plink/1.9/input#pheno

### Phenotypes

**Loading from an alternate phenotype file**

```
--pheno <filename>
--mpheno <n>
--pheno-name <column name>
--all-pheno
--pheno-merge
```

**--pheno** causes phenotype values to be read from the 3rd column of the specified space- or tab-delimited file, instead of the .fam or .ped file. The first and second columns of that file must contain family and within-family IDs, respectively.

In combination with --pheno, **--mpheno** lets you use the (**n**+2)th column instead of the 3rd column, while **--pheno-name** lets you select a column by title. (In order to use --pheno-name, there must be a header row with first two entries 'FID' and 'IID'.) The new **--pheno-merge** flag tells PLINK to use the phenotype value in the .fam/.ped file when no value is present in the --pheno file; without it, the phenotype is always treated as missing in this case.

--allow-no-sex is now required if you want to retain phenotype values for missing-sex samples. This is a change from PLINK 1.07; we believe it would be more confusing to continue treating regular and --pheno phenotypes differently, and apologize for any temporary inconvenience we've caused.

**--all-pheno** causes all phenotypes present in the --pheno file to be subject to the association tests you've requested. (--pheno-merge then applies to every phenotype.) Note that, when dealing with a very large number of phenotypes, specialized software is usually more appropriate than --all-pheno; we recommend Matrix eQTL or FastQTL, which process thousands of phenotypes simultaneously and achieve a level of efficiency not possible with --all-pheno + --assoc/--linear. (**Update, 1 Apr 2019:** PLINK 2.0 also handles this case efficiently now.)

**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, and obtaining a standardised beta (to mean zero, unit variance).**

plink --bfile adclean.cont --linear  hide-covar standard-beta --pheno adclean.cont.txt --covar adpc.txt --covar-name PC1 --out 2.2_adclean.cont

or

plink --bfile adclean.cont --linear  hide-covar standard-beta --pheno adclean.cont.txt --covar adpc.txt --covar-number 1 --out 2.2_adclean.cont

With --linear, the 'standard-beta' modifier standardizes the phenotype and all predictors to zero mean and unit variance before regression.

See also: https://www.cog-genomics.org/plink/1.9/input#covar

## Covariates

```
--covar <filename> ['keep-pheno-on-missing-cov']
--covar-name <column ID(s)/range(s)...>
--covar-number <column number(s)/range(s)...>
--no-const-covar
--allow-no-covars
```

**--covar** designates the file to load covariates from. The file format is the same as for --pheno (optional header line, FID and IID in first two columns, covariates in remaining columns). By default, the main phenotype is set to missing if any covariate is missing; you can disable this with the '**keep-pheno-on-missing-cov**' modifier.

**--covar-name** lets you specify a subset of covariates to load, by column name; separate multiple column names with spaces or commas, and use dashes to designate ranges. (Spaces are not permitted immediately before or after a range-denoting dash.) **--covar-number** lets you use column numbers instead.

For example, if the first row of the covariate file is

```
FID IID SITE AGE DOB BMI ETH SMOKE STATUS ALC
```

then the following two expressions have the same effect:

```
--covar-name AGE, BMI-SMOKE, ALC
--covar-number 2, 4-6, 8
```

**--no-const-covar** excludes all constant covariates. PLINK normally errors out if this causes all covariates to be excluded (or if the --covar file contained no covariates in the first place), but you can use the **--allow-no-covars** flag to make it try to proceed.

**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.**

plink --bfile adclean.cont --linear hide-covar --pheno adclean.cont.txt --covar adpc.txt --ci 0.95 --out 2.3_adclean.cont

**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.**

```
awk '{print $1,$2,$3,$12}' 2.3_adclean.cont.assoc.linear | grep -v NA >
plot.adclean.cont.linear.txt
```

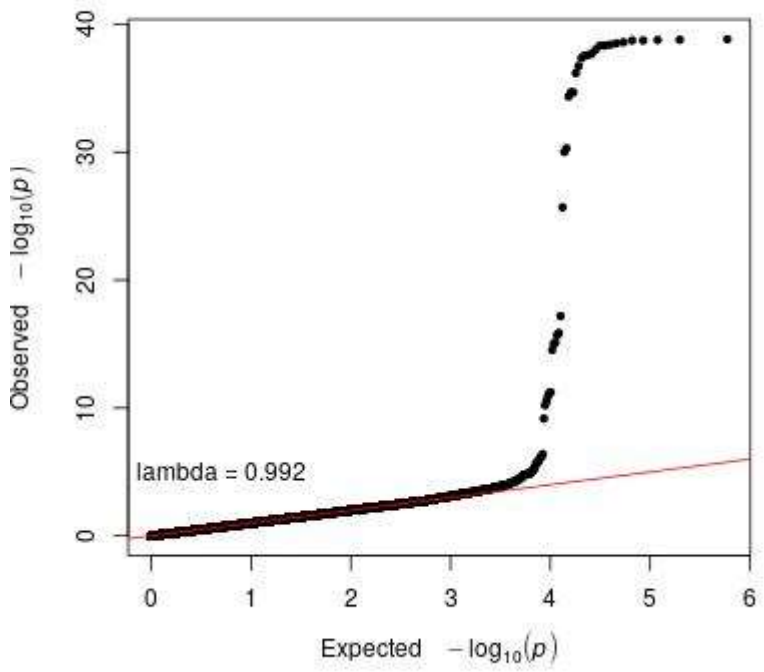
**What's the SNP with the lowest p-value (or top SNP)?**

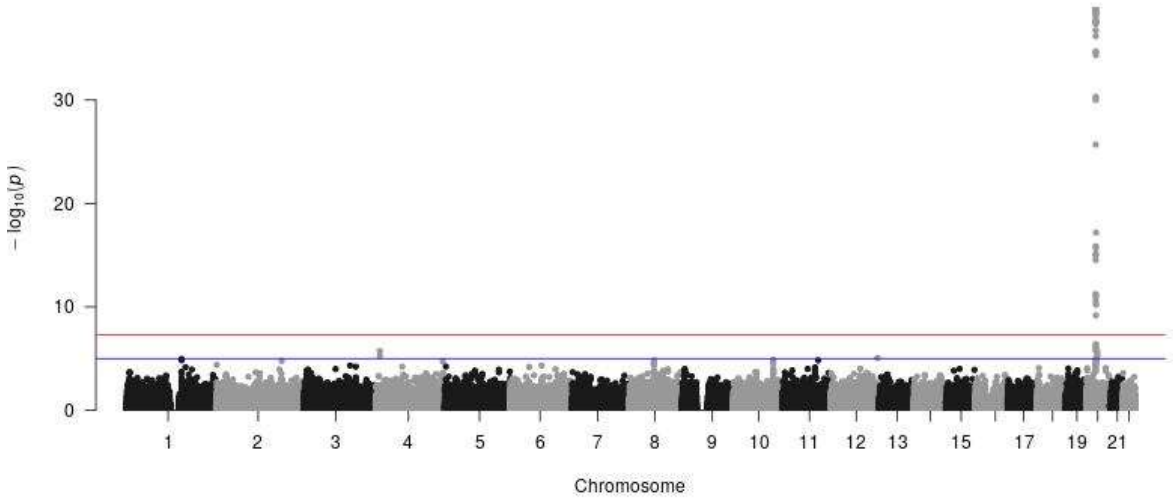
```
sort -k4 -g plot.adclean.cc.linear.txt | head
```

The SNP with the lowest p-value is rs6050598 (or chr20:25397257).

**QQ and Manhattan plots**

**Open in RStudio (https://workshop.colorado.edu/rstudio/) the script Rscript_qqMan.R and plot the results. Set your working directory to the folder where you have your results (you can check first what your working directory is by typing pwd in the command line). Run the script. Explore the QQ plot and the Manhattan plot. What information do you gather from these plots and the lambda value? Do you detect any anomalies?**





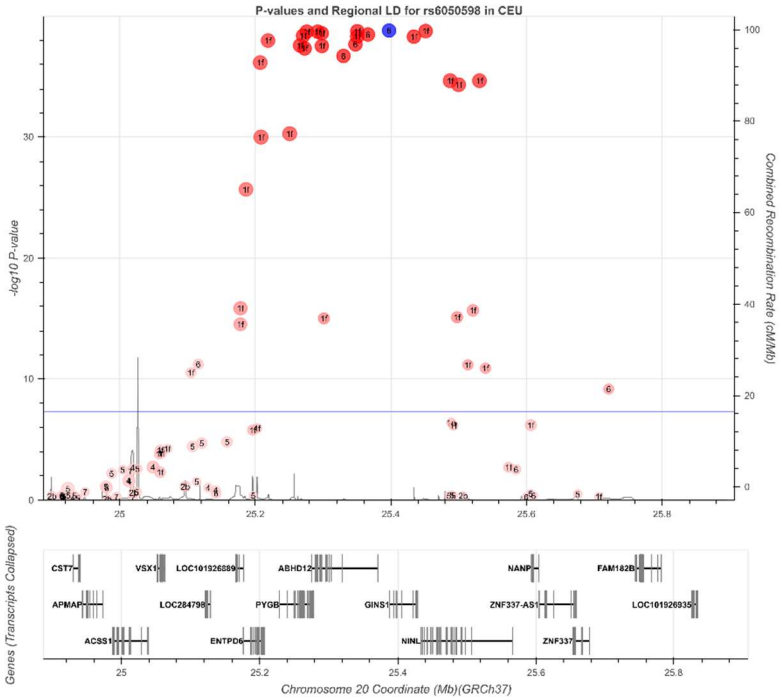

The QQ plot doesn't show signs of systematic bias or inflation (most of the dots fall in the diagonal, except those with the lowest p-values, probably associated with out trait) and the lambda value is close to 1, indicating no inflation. The Manhattan plot looks healthy too. In this case we can see SNPs in the same region being strongly associated with the phenotype (we see a "tower"). Remember this is genotyped data and we are missing information about multiple SNPs.

Note: It's very difficult that you see results like these in a sample of ~300 individuals… unless e.g. the phenotype you are analysing is a genetic variable, like this one.

**Regional plot**

**Still in RStudio, go to the "Files" section and check the file plot.adclean.cont.linear.txt. Then go to More and then Export. The file will download to your local computer.**
**You can explore association p-value results and LD patterns by uploading your results (or a genomic region of interest) to LDassoc in LDlink: https://ldlink.nci.nih.gov/?tab=ldassoc.**
**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.**
**Select the columns that contain the information on Chromosome, Position, and P-Value.**
**LDassoc has three options for visualising 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 (rs6050598 or chr20:25397257).**
**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.**

**Explore the interactive plot. Is the index variant in high LD with any of the nearby variants? Remember the R2 is a measure of linkage disequilibrium or correlation of alleles for two genetic variants.**




There are some variants in high linkage disequilibrium with the index variant and they are also associated with the phenotype.