Quality Control for Genome-Wide Association Studies

Bart Baselmans & Meike Bartels **Boulder 2017**

Setting up files and directories

To perform a quality control protocol in a Genome-Wide Association Meta Analyses (GWAMA) project, we will introduce you to EasyQC. EasyQC is a freely available software tool based on the R language. EasyQC has many functions at the: 1) **study file level** 2) **meta-file level and** 3) **meta-analysis OUTPUT level.** Details about downloading and documentation are available at the reference section at the end of this document.

We will use the example files in

the **faculty/meike/2017/EasyQC** directory.

In this directory contains the following files:

1. Summary statistics test_GWAS. This file contains a set of 50K random SNPs of the Subjective Well-being GWAS of the Netherlands Twin Register

2. The EasyQC script test_GWAS.ecf to perform AND specify the actual data cleaning

3. A rs chromosome basepair identifier file rsTEST

4. An allele frequency reference set AFtest

For the purpose of this practicum, the identifier and allele frequency reference file are matched to the 50K GWAS summary statistics.

We will specify several important steps in the terminal to perform quality control at the study file level. The examples illustrate some important points in working with GWAS data, and the practical will hopefully lead to the understanding that quality control in large-scale GWAMA are essential for proper interpretation of results.

We will use the text editor, UNIX commands, RStudio, and a couple of other programs to help to interpreted results. You can copy and paste specific commands from this document to your UNIX terminal, or into the R window or ECF file.

The basics

Start by making a working folder and copying the practical files to it.

Open a new terminal window:

```
mkdir easyqc-practical

cp –r /faculty/meike/2017/EasyQQ/* easyqc-practical/

cd easyqc-practical
```

In your newly created folder you now have multiple files that are needed for the quality control procedure. *We will discuss all of them in a minute*. The most important file to work with is the so-called **test_GWAS.ecf**. Within this ecf file you have to specify the QC algorithm you want to apply to the GWAS summary statistics. It is best to take quite some time for this and think it thoroughly over, as this will be the QC steps that will be applied to all GWAS summary statistics that will be included in your meta-analysis.

ECF file

To open the test_GWAS.ecf file in your terminal we will use Nano (which is one of the many text-editors in the UNIX environment.

```
nano test_GWAS.ecf
```

The first thing that you need to specify is the path where you want to store the output of the QC algorithm. *I strongly suggest that for each cohort included in your meta-analysis you specify a separate folder*. In the output, you get the SNPs eliminated in every step, all your QC plots and an overview file of SNPs surviving QC and the number of SNPs eliminated in every step of the procedure. For the purpose of this practicum, we will create one OUTPUT folder:

1. Ctrl X -> to exit nano
2. Mkdir OUTPUT (to generate the OUTPUT folder
3. nano test_GWAS.ecf (to enter the QC protocol)

Specify OUTPUT path

Using: DEFINE –pathout /SPECIFY_YOUR_PATH_TO_OUTPUT_FOLDER

For example: /home/meike/2017/EasyQC/OUTPUT

Specify summary statistics properties

The next step is to specify the column names of your GWAS summary statistics (test_GWAS) using the

--acolIn command. To speed things up, this is already done for you, but keep in mind that most of the summary statistics you will use in your meta-analysis will come with different headers (yup, even when you specified fix column names in you SOP, because people do not always do what you ask them to do).

To overcome the problem that you need to change every script according to the column names specified by all different participating cohorts (which is very time consuming and prone to errors) you can convert the column names to **fix names**. This is done by the –acolNewName command. Again, this is already done for you. The large advantage of renaming all columns to fix names is that it allows you to use the majority of the QC script to all files you need to QC (in the SWB project this was up to 167 files, so you can imagine the advantage of fix column names).

Next, to speed up the cleaning process you need to specify the classes of the different columns (acolInClasses), how the missings of a summary statistic is defined (strMissing), and what kind of separator (TAB/SPACE etc) is used in the different summary statistics (strSeparator).

Specify the path of the input file

EASYIN –fileIn SPECIFY_YOUR_PATH_TO_test_GWAS

For example: /home/meike/2017/EasyQC/test_GWAS

You can also specify a short name that will be added to the CLEANED filenames (in this case we called it PRACTICAL. Also, you can specify missing columns over here (astrSetNumCol). We will get back to this in a bit.

When all paths are specified and linked to the right directories we can start building our QC algorithm using EasyQC

To tell the rpackage EASYQC to begin the quality control procedure, you need write down

START EASYQC (is already done for you)

Sanity Checks

Using CLEAN –rcdClean (r code to clean SNPs). You can tell EasyQC the criteria of which SNPs need to be removed from your summary statistics. *Note: You can specify anything you want. From missing alleles to missing imputed to missing SNP quality metrics!* For the purpose of this practical we specify a few basic Sanity Checks, but in real life you can decide to be much more stringent in cleaning (like we did in the SWB project).

In test_GWAS.ecf file you see that we will remove any SNP that has:

- Another chromosome than 1 to 22
- Missing A1 or A2
- Missing Pvalues, BETA, and SE
- Missing estimated effect allele frequency (EAF)
- Missing sample size (N)

The next set of SNPs we want to remove are SNPs with strange statistic properties:

- P-values below 0 and above 1
- SE smaller than zero
- Estimated allele frequencies below zero and above one

Again, note that you can specify more statistics over here

With the --strCleanName AND --blnWriteCleaned 1 commands you write the removed SNPs to a separated file and specify the name of that file. A Boolean value has the default of 1 (write to separate file). However, if you don't want to keep the removed SNPs you can set the boolean to 0.

Apply minimum thresholds

The next step in our QC protocol is to filter the file and apply minimum thresholds.

We will:

- filter out any monomorphic SNP
- apply a MAF threshold which is dependent on the number of people included in the GWAS you are qc'ing
- Based on the used imputation software we will remove SNPs with low imputation quality

If the cohort supplied us with the "Rsq" variable generated by MaCH23, we dropped SNPs with Rsq < 0.4. If they uploaded the "INFO" variable generated by IMPUTE24, we applied a threshold of <0.5. If PLINK's "info" variable was supplied, we applied a threshold of <0.8.

Create Uniform allele codes

To make sure that you specify the right alleles (effect allele and other allele) we will harmonize the alleles among all files in such a way that the output file alleles coded “A”, “C”, “G” and “T”.

```
HARMONIZEALLELES   --colInA1 EFFECT_ALLELE
                   --colInA2 OTHER_ALLELE
```

Remove all indells

```
CLEAN --rcdClean (EFFECT_ALLELE%in%c('I','D')) | (OTHER_ALLELE%in%c('I','D'))   --
strCleanName numDrop_INDEL --blnWriteCleaned 1
```

Harmonization of marker names

If you perform a GWAS meta-analysis (GWAMA) and receive multiple summary statistics, there will be no doubt that multiple cohorts us different markernames (rsnumbers or chr:bp, etc).

This step will make sure that your output file has a Unified markername so that in the meta-analysis phase you will not run into problems.

For this purpose, you need to specify the path in your ECF file to your directory in which the rs file is stored:

```
--fileMap /home/meike/2017/EasyQC/rsTEST
```

For the purpose of this practicum we created a file (rsTEST) based on the test_GWAS file. In real life you need to download the right reference set from the EasyQC website -> see the bottom of this document. BE AWARE THAT YOU USE THE RIGHT BUILD!!

To make sure that there are no duplicates, we will remove any duplicates that might be in your file.

```
CLEANDUPLICATES    --colInMarker cptid
                   --strMode removeall
```

Allele Frequencies

To check whether the allele frequencies provided by specific cohort are acceptable (e.g. non flipped SNPs etc), we need to merge the summary statistics with a reference allele frequencies file. Again, for the purpose of this practicum we adjusted the allele frequency file to the testGWAS file.

First you have to specify the path to where your AFtest file is stored.

home/meike/2017/EasyQC/AFtest

Second, the alleles need to be aligned with the reference file.

```
ADJUSTALLELES       --colInA1 EFFECT_ALLELE
                    --colInA2 OTHER_ALLELE
                    --colInFreq EAF
                    --colInBeta BETA
                    --colRefA1 a0.ref
                    --colRefA2 a1.ref
                    --blnMetalUseStrand 1
                    --blnRemoveMismatch 1
                    --blnRemoveInvalid 1
                    --blnWriteMismatch 1
                    --blnWriteInvalid 1
```

Using this step all allele frequencies mismatches will be removed, BUT stored in the output -> so you can have a look at it afterwards.

Ok, by now we have performed some essential QC steps (note, this practical is a simplification of a real-life situation). But a picture says more than a thousand words, so let's make some informative visualizations.

<u>VISUALIZATION</u>

There are a couple of plots that can be really informative during QC. One of the most frequently used plots are QQ plots, allele frequency plots, and so called PZ plots. These plots are easily made within EasyQC.

1) Allele Frequency plots

Within this plot, you can set the AF outlier by yourself. Usually AF deviations > 0.2 are considered as outliers.

```
## Plot frequencies against HapMap reference frequencies.
AFCHECK --colInFreq EAF
        --colRefFreq eaf.ref
        --numLimOutlier 0.2
        --blnPlotAll 1
```

```
## blnPlotAll 0 causes that only outlying SNPs with |Freq-Freq.ref|>0.2 will be plotted (way less computational time)
```

2) QQ plot -> check whether there is inflation in your data which might be indication of stratification, etc

```
## QQ plot

QQPLOT          --acolQQPlot PVAL
          --numPvalOffset 0.05
          --strMode subplot
```

3) Plot Z versus P -> This plot is to investigate whether the reported beta's and standard errors correspond to the reported p-values. By dividing beta by SE you will obtain the Z values. You can convert these to pvalues and this should be in correspondence to the actual reported p-values…

```
PZPLOT          --colBeta BETA
                --colSe SE
               --colPval PVAL
```

Show Z plots to show that this is not always the case

<u>FINAL STEP</u>

Save the columns you are interested in and let the Rpackage EasyQC know to stop

```
## Save cleaned file.

GETCOLS --acolOut
cptid;SNPID;CHR;POS;EFFECT_ALLELE;OTHER_ALLELE;EAF;INFO;SE;PVAL;BETA;N

WRITE  --strPrefix CLEANED.
          --strMissing  .
          --strMode gz
```

Let’s see if it works

1. Go to R Studio

2. Set the right working directory

```
Setwd()
```

3. Load EasyQC

```
Library (EasyQC)
```

4. run EasyQC

```
EasyQC(“test_GWAS.ecf”)
```

Note: you will get some warnings about NAs introduced by coercion. That is expected and can be ignored.

A. Check all output and answer the following questions

- How many indels are removed?
- How many badly imputed SNPs are removed?
- How many SNPs were not in the reference set?
- How many SNPs with MAF below threshold?

B. Check cleaned.

To open this file:

```
zless -S CLEANED.PRATICAL.gz
```

C. Check the plots. Anything unexpected?

D. Check the summary file of the QC procedure

- Take a look at the .rep file. This is a plain-text file with many colums with summary data on the cleaning. You can open this file in your favorite text-editor or a speadsheet program (your laptop has LibreOffice Calc for this).
- The most important numbers are in column 2 and 3, which note the number of SNPs in your input file and the number of SNPs that survied QC, respectively