CHAPTER 5

QCEWAS: A SOFTWARE PACKAGE FOR AUTOMATED QUALITY CONTROL OF RESULTS OF EPIGENOME-WIDE ASSOCIATION STUDIES

Peter J van der Most*, Leanne K Küpers*, Harold Snieder and Ilja Nolte,

* *Equal contribution as first author*

ABSTRACT

Epigenome-wide association studies are increasingly being conducted and several large international meta-analysis consortia have been established in the past years. With data originating from multiple sources, a thorough and centralized quality control is necessary. To facilitate this, we developed the QCEWAS R package. QCEWAS enables automated quality control of results files of epigenome-wide association studies prior to meta-analysis. QCEWAS produces cohort-specific statistics and graphs to interpret the quality of the results files, as well as cleaned input files ready for meta-analysis.

This script will soon be submitted as an open source software package at the *Comprehensive R Archive Network* website, and therefore it will be available to the wider research community.

INTRODUCTION

In recent years epigenome-wide association studies (EWAS) have gained increasing attention, resulting e.g. in two special issues in the International Journal of Epidemiology in 2012 and 2015. DNA methylation is one of the most studied and best understood mechanisms in epigenetics. It is often measured using the 450k BeadChip (Illumina Inc., San Diego, USA), which quantifies methylation levels of over 450,000 Cytosine-phosphate-Guanine (CpG) sites. This 450k chip is currently the standard platform because it offers a good balance of genome-wide coverage (>450K CpG sites), resolution (information on single base pairs), and throughput (12 samples per chip and up to 96 samples per run)[1].

Given the frequent use of the 450k chip, meta-analysis to combine results of methylation analyses from multiple cohorts is an obvious choice. As in traditional genome-wide association studies, this increases the sample size and thus the statistical power to find CpG sites that are associated with a disease or trait of interest.

However, before meta-analysing EWAS results originating from multiple sources, it is important to perform a thorough, centralized quality control (QC) in order to verify that cohort-specific results are valid, reliable, and of high quality, and to check whether results are comparable between cohorts. Because EWAS results files are often large and checking them by hand is cumbersome, automation of this process is desirable and will result in compatible and harmonized results from all sources.

To our knowledge, no other software packages are currently available for the QC of EWAS results files. Therefore, we developed the QCEWAS software package, allowing fast and easy assessment of the quality of EWAS results files through informative figures and statistics. Additionally, the package allows for quick generation of cleaned input files for the actual meta-analysis of EWASs.

APPROACH

QCEWAS was built as a package for R (R Development Core Team, 2015). The R platform was chosen as it is operating system independent, open source, can handle large datasets, and is flexible regarding input file format. In addition, most important software packages for analyzing EWAS data are also developed in R. QCEWAS requires R version

3 or later (64-bit recommended) and can be downloaded from the Comprehensive R Archive Network Website (http://cran.r-project.org).

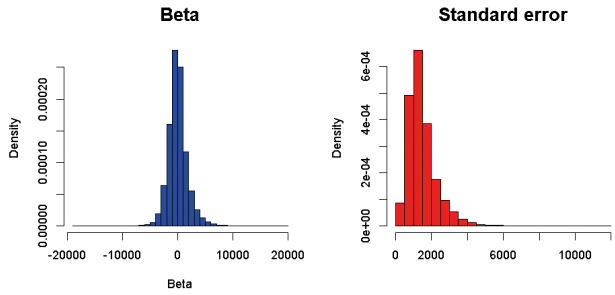
The QCEWAS package includes several functions, but the most important two are 'EWAS_QC' and 'EWAS_series'. The first performs a thorough QC on a single EWAS results file; the second can process a series of results files by calling 'EWAS_QC' for every file and additionally performs checks to compare the files.
For this package EWAS results from analyses using the 450k chip, as well as from the older 27k chip can be used.

RESULTS

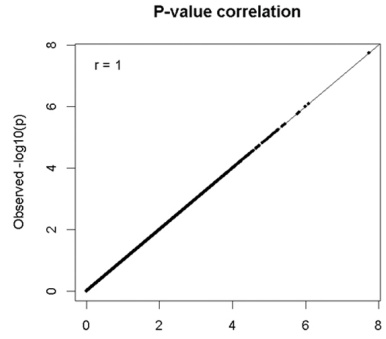
For each results file, QCEWAS carries out quality checks on:

1. data integrity: are the required data present and valid (e.g. no negative standard errors] or p-values)?;
2. outlier detection and removal (optional);
3. allosomal marker removal (optional);
4. effect size and SE distribution in histograms (**Figure 1A**);
5. reported p-values by correlating them with p-values calculated from effect size and SE (**Figure 1B**);
6. over-/under-significance of results via a QQ plot (**Figure 1C**);
7. the distribution of effect sizes versus p-values in a volcano plot (**Figure 1D**);
8. location of the signals in a Manhattan plot (**Figure 1E**).

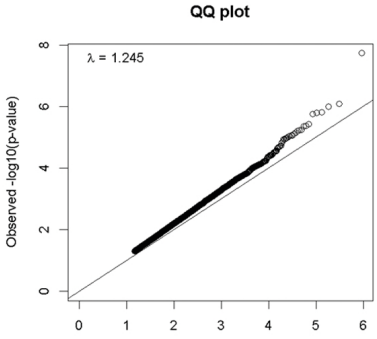
Additionally, two figures are produced to compare QC statistics of multiple EWAS results files: a precision plot, showing the distribution of the precision (1/median[SE]) against the square root of the sample size to check if precision increases proportionally with sample size (**Figure 2A**); and a boxplot showing the distributions of the effect sizes per file (**Figure 2B**). **Figure 2A** shows one cohort file (no. 3) with a precision that is higher than expected based on the trend of the other files and possibly another file (no. 10) with a lower precision. **Figure 2B** shows one outlying cohort (no. 1) with a wider spread of effect sizes than expected, suggesting use of a different measure or analysis model than the other cohorts. Finally, QCEWAS can produce cleaned EWAS results files that are ready for meta-analysis.



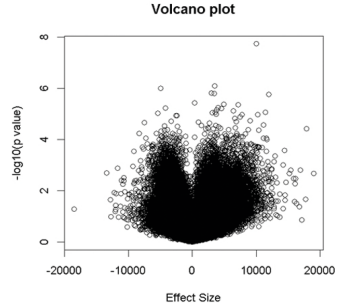
1A) Effect size and SE distribution



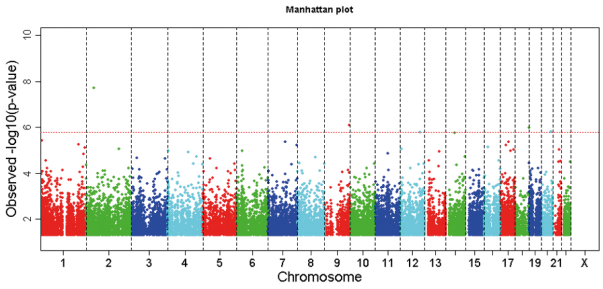
1B) P-value correlation



1C) QQ plot



1D) Volcano plot



1E) Manhattan plot

FIGURES 1A-1E. Cohort specific figures resulting from QCEWAS.

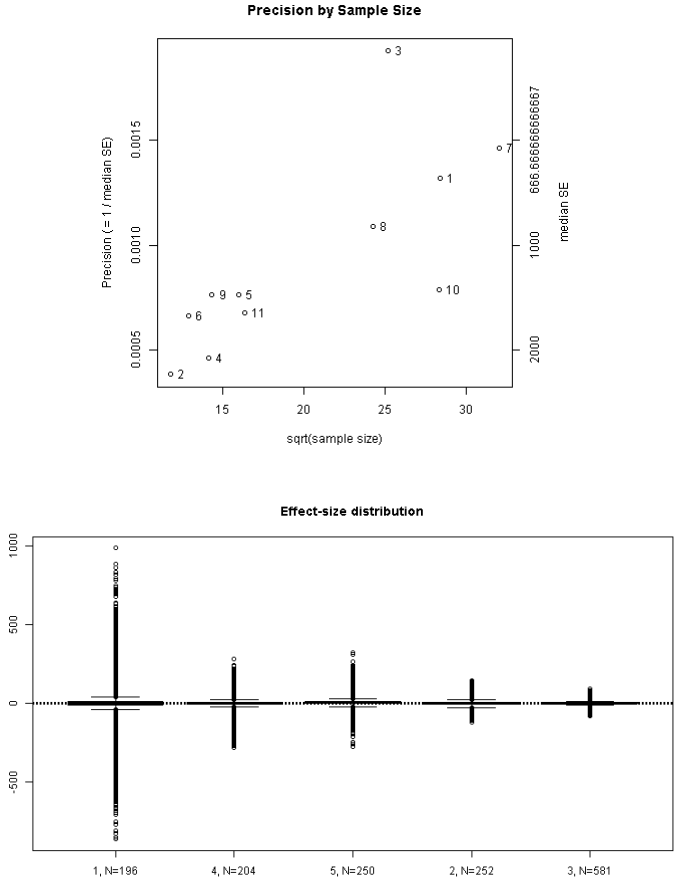


FIGURE 2. Comparison of QC statistics of EWAS results files using (a) a plot of distribution of precision against the √N and (b) a boxplot of the effect sizes per cohort file.

CONCLUSION

With QCEWAS we developed a flexible and easy-to-use software package for a complete QC of EWAS results files, allowing users to detect errors that would have biased the subsequent meta-analysis.

REFERENCES

1 Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011; 12: 529-41.

CHAPTER 5

SUPPLEMENTARY MATERIAL

QCEWAS USER MANUAL: *QUICK START GUIDE*

EWAS_QC

This is the main function of the QCEWAS package. EWAS_QC accepts a single EWAS results file and runs a thorough quality check (including data integrity), optionally applies various filters and generates QQ, Volcano and Manhattan plots. The function EWAS_series can be used to process multiple results files sequentially.

Usage

```
EWAS_QC(data,
        map,
        outputname,
        header_translations,
        threshold_outliers = c(NA, NA),
        exclude_outliers = FALSE,
        exclude_X = FALSE, exclude_Y = FALSE,
        save_final_dataset = TRUE, gzip_final_dataset = TRUE,
        header_final_dataset = "standard",
        return_beta = FALSE, N_return_beta = 500000L,
        ...)
```

- data = a data frame with EWAS results, or the name of a file containing the same. The table must include the columns PROBEID, BETA, SE, and P_VAL. Other columns may be included but will be ignored. If the column names differ from the above, the argument header_translations can be used to translate them. If a filename is entered in this argument, it will be imported via the read.table function. read.table can handle a variety of formats, including files compressed in the .gz format. EWAS_QC will pass any named, unknown arguments to read.table, so you can specify the column separator and NA string in EWAS_QC with the usual read.table arguments. (Note that this only applied to importing the EWAS results, and not the map or translation files.)
- map = a data frame with chromosome and position values of the probes, or the name of a file containing the same. This argument is optional: if no map is specified, EWAS_QC will skip the Manhattan plot and chromosome filters. map must include the columns PROBEID, CHR (chromosome), and POS (position), using those exact names. Other columns may be included but will be ignored. f a filename is entered in this argument, it will be imported via the read.table function. read.table can handle a variety of formats, including files compressed in the .gz format.
- outputname = a character string specifying the intended filename for the output. This includes not only the cleaned results file and the log, but also any graphs created. Do not include an extension; EWAS_QC adds these automatically.
- header_translations = a translation table for the column names of the input file, or the name of a file containing the same. This argument is optional: if not specified, EWAS_QC assumes the default column names are used. See translate_header for information on the format.
- threshold_outliers = a numeric string of length two. This defines which effect sizes will be treated as outliers. The first value specifies the lower limit (i.e. markers with effect sizes below this value are considered outliers), the second the upper limit. The check for low or high outliers is skipped if the respective value is set to NA. To skip the check entirely, set this to c(NA, NA).
- exclude_outliers = a logical value determining how outliers are treated. If TRUE, they are excluded from the final dataset. If FALSE, they are merely counted.

- exclude_X, exclude_Y = logical values determining whether markers at the X and Y chromosome respectively are excluded from the final dataset. This requires a map to be specified.
- save_final_dataset = logical determining whether the cleaned dataset will be saved.
- gzip_final_dataset = logical determining whether the saved dataset will be compressed as .gz file.
- save_standard_header = logical determining whether the saved dataset will use the standard column names (if TRUE) or the original ones (if FALSE).
- Header_final_dataset = either a character vector or a table determining the header names used in the final dataset, or the name of a file containing the same. If "original", the final dataset will use the same column names as the original input file. If "standard", it will use the default EWAS_QC column names. If a table, it will be passed to translate_header to convert the column names. If a table, the default column names (PROBEID, BETA, SE, and P_VAL) must be in the second column, and the desired column names in the first.
- return_beta, N_return_beta = arguments used by EWAS_series. These are not important for users and can be ignored. For the sake of completeness: return_beta is a logical value – if TRUE, the function return includes a vector of effect sizes. N_return_beta defines the length of the vector.
- ... = arguments passed to read.table when importing the EWAS results file.

Details

The Quality Control has the following 5 stages:

1. Checking data integrity
   a. The values inside the EWAS results are tested for validity. If impossible p-values, effect-sizes, etc. are encountered, EWAS_QC generates a warning in the R console and set them to NA.
2. Filter for outliers and sex-chromosomes (optional)
   a. Counts the number of outlying markers, as well as chromosome X and Y markers, and deletes them if specified.
3. Generating QC plots
   a. A histogram of beta and standard error distribution is plotted.
   b. The p-values are checked by correlating and plotting them against p-values calculated from the beta and standard error.
   c. A QQ plot is generated to test for over/undersignificance.
   d. A Manhattan plot is generated to see where the signals (if any) are located.
   e. A Volcano plot is generated to check the distribution of effect sizes vs. p values.
4. Creating a QC log
   a. The log contains notes about any problems encountered during the QC, as well as several tables describing the data.
5. Saving the cleaned dataset (optional)

Value

The main output of EWAS_QC are the cleaned results file, log file and QC graphs. However, the function also returns a list with 9 elements:

- data_input = the file name of the input file, if loaded from a file. If not, this will be an empty character string.

- file = the filename of the cleaned results file.
- QC_success = logical, indicates whether EWAS_QC was able to run a full QC on the file Note that a TRUE value does not mean that no problems where encountered, merely that the full QC was executed.
- lambda = the lambda value of reported p-values in the cleaned dataset.
- p_cor = the correlation between reported and expected (based on beta and standard error) p values.
- N = a named integer vector reporting how many markers were in the original dataset, how many had missing values, how many were on chromosomes X and Y, how many were outliers, how many were removed and how many are in the final, cleaned dataset. Has no relation to the N argument of EWAS_series.
- SE_median = a numeric value – the median of the standard errors in the cleaned dataset.
- effect_size = if return_beta is TRUE, this is a numeric vector of length N_return_beta, containing a representative selection of beta values from the filtered dataset. If FALSE, this will be NULL.

Notes

The function will return a warning if it encounters p-values < 1e-300, as this is close to the smallest number that R can process correctly. Various functions in the QCEWAS package will set these values to 1e-300 to ensure proper handling.

EWAS_SERIES

This function runs a QC over multiple files and generates additional graphs to comparing the results of these files.

Usage

```
EWAS_series(EWAS_files,
            output_files,
            map,
            N,
            header_translations,
            save_final_dataset = TRUE, gzip_final_dataset = TRUE,
            ...)
```

- EWAS_files = a character vector containing the filenames of the EWAS results to be QC'ed.
- output_files = a character vector containing the filenames of the output files. Do not add an extension – EWAS_QC does so automatically.
- map = a data frame with chromosome and position values of the CpGs in data, or the name of a file containing the same. See EWAS_QC for details.
  - This argument is optional: if not specified, EWAS_QC will not generate a Manhattan plot and no filter for X and Y markers can be performed.
- N = a data frame containing the filenames (the same as in the EWAS_files argument) and the sample sizes of the datasets, or the name of a file containing the same.
  - This argument is optional: if not specified, EWAS_series will not generate a precision plot.
  - It must contain the columns 'file' and 'N', with those exact names.
  - The column 'file' must contain all filenames specified in EWAS_files.

- header_translations = a translation table for the column names of the EWAS files, or the name of a file containing the same. See translate_header for information on the format.
- save_final_dataset, gzip_final_dataset = logical values. See EWAS_QC for details.
- ... = arguments passed to EWAS_QC.

Details

EWAS_series works by calling EWAS_QC for every filename given in EWAS_files. After all files have been processed, it will generate two additional graphs: a precision plot (provided N was specified) and a beta-distribution plot. The former shows the distribution of precision (1 / median standard error) against the square root of the sample size of the results file. Normally, one expects to see a roughly positive correlation (i.e. the cohorts ought to cluster around the linear diagonal from the lower left to the upper right). The presence of outliers means that the outlying cohort(s) have a far higher/lower uncertainty in their estimates that can be expected from their sample size. This could indicate a different method, a different measure (check the effect-size distribution plot) or possibly over- or undersignificance of their estimates (check the QQ plot and lambda value).
The effect-size distribution plot allows comparison of the effect-size scale of different files. One expects the distribution to become somewhat narrower as sample size increases. However, large differences in scale suggest that the files used different units for their measurements.
Both plots use numbers rather than names to identify files. The full filenames and corresponding numbers are listed in the EWAS_QC_legend.txt file that is generated after EWAS_series completes.

Value

The main output of EWAS_series are the cleaned results files, logs and graphs. The function also returns an invisible data frame, listing the input file names, file numbers, whether they passed a complete QC (note that this merely indicates that the QC was completed, not that there were no problems), the standard error and, if specified, the sample size (this is the same table as was saved in EWAS_QC_legend.txt).

EWAS_PLOTS

This is a sub function of EWAS_QC that generates quantile-quantile (QQ) and Manhattan plots from the dataset. It can also be called by users (note that it does not generate the histogram or volcano plots – this is done by EWAS_QC itself).

Usage

```
EWAS_plots(dataset,
           plot_QQ = TRUE, plot_Man = TRUE,
           plot_cutoff_p = 0.05,
           plot_QQ_bands = FALSE,
           save_name = "dataset",
           header_translations)
```

- dataset = a data frame containing the columns "CHR" (chromosome), "POS" (position) and "P_VAL" (p-value), or a numeric vector of p values.
    - Chr(omosome) and position are only required when a Manhattan plot is made.

  - The order of columns, or the presence of other columns, does not matter.
  - Alternative column names can be translated via the header_translations argument.
  - Note that, unlike EWAS_QC, this argument does not accept filenames, only data frames or vectors.
- plot_QQ, plot_Man = logicals determining whether a QQ and Manhattan plot are made.
- plot_cutoff_p = numeric, the threshold of p-values to be shown in the QQ & Manhattan plots. Higher (less significant) p-values are excluded from the plot. The default setting is 0.05, which excludes 95% of data-points. It's not recommended to increase the value above 0.05, as this may dramatically increase running time and memory usage.
- plot_QQ_bands = logical, if TRUE, probability bands are added to the QQ plot.
- save_name = character string, the name used for the plot files (do not add an extension: EWAS_plots will do this automatically).
- header_translations = a table that translates the column names in dataset to the standard names (see dataset). See translate_header for details.

Details

EWAS_plots is a fairly straightforward function. It accepts a data table or a vector of p-values, and generates QQ and Manhattan plots from these.

Value

EWAS_plots' most important output are the two graphs. However, it also returns a single, invisible, numeric value, representing the lambda calculated over the p-values.

P_CORRELATION

A sub function of EWAS_QC, P_correlation tests if the reported p-values match the p-value that can be derived from the beta and standard error values. Aberrations between these indicate that the p-values have been adjusted, or that there is some other problem with the data. It also creates a plot of reported vs. expected p-values that shows the aberration.

Usage

```
P_correlation(dataset,
        plot_correlation = TRUE, plot_if_threshold = FALSE, threshold_r = 0.99,
        save_name = "dataset",
        header_translations,
        ...)
```

- dataset = a data frame with the columns BETA (effect size), SE (standard error), and P_VAL (p value). If the column names differ from the above, the argument header_translations can be used to translate them.
- plot_correlation = logical, determines whether a graph is made of reported vs. expected p values.
- plot_if_threshold = logical. If TRUE, the plot is only generated if the p-value correlation is below the specified threshold.
- threshold_r = numeric. If the p-value correlation is below this, a warning is generated.

- save_name = character string used for the output file. Do not add an extension – P_correlation will do so automatically.
- header_translations = a translation table for the header of dataset. See translate_header for details.
- ... = arguments passed to the generic R plot function.

Details

P_correlation is primarly a subfunction of EWAS_QC, but it can be used separately.

Value

P_correlation returns a single numeric value, representing the correlation between reported and expected p-values.

P_LAMBDA

P_lambda calculates the lambda value from a vector of p-values.

Usage

```
P_lambda(p)
```

p = a numeric vector of pvalues

Details

The function removes any missing values from p, and then returns:
median(qchisq(p, df=1, lower.tail=FALSE)) / qchisq(0.5, 1)

Value

A single numeric value representing lambda.

TRANSLATE_HEADER

In our experience, cohorts will not always follow the analysis plan when assigning column (header) names in their results files. Checking and correcting these manually is cumbersome, so QCEWAS uses this function instead to convert the header to standard names by means of a translation table. It is, effectively, a subroutine of EWAS_QC, EWAS_plots and P_correlation, so the details are not relevant to most users. They merely need to include a translation table (the format is explained below), and the functions will automatically call translate_header.

Usage

```
translate_header(header,
  standard = c("PROBEID","CHR","POS","BETA","SE","P_VAL"),
  alternative)
```

- header = a character vector containing the header of the results file.
- standard = a character vector containing the standard column names (i.e. the ones that are required by various QCEWAS functions to run correctly).
- alternative = a translation table. See below for the format.

Details

The function takes the entries in standard one by one, and checks them against the translation table for alternatives. It will report any missing standard headers, as well as duplicate ones.

Note

The function will automatically capitalize the elements of the header argument (so the alternatives in the translation table must also be capitalized). Also, elements that are not in standard will not be translated, even if they are present in the translation table.

Value

The function returns a list with 6 elements

- header_N = integer, the length of header.
- header_h = character vector: header after translation.
- missing_N = integer, the number of elements in standard that could not be found in header.
- missing_h = character vector, the elements in standard that could not be found in header.
- unknown_N = integer, the number of elements in header that were not translated.
- unknown_h = character vector, the elements in header that were not translated.

Translation table

The translation table must meet the following requirements:

- 2 columns, with the default column names (i.e. the ones in the standard argument) in the first column, and the alternatives in the second.
- Multiple alternatives are allowed for a single standard name, but every alternative name must have its row.
- The alternatives must be capitalized.
- No duplicate alternatives are allowed.
- A header line is not required, and will be ignored if present.

PART 2

INFANT LIFESTYLE AND CHILDHOOD WEIGHT STATUS