Package 'QCEWAS'

February 7, 2023

Type Package

Title Fast and Easy Quality Control of EWAS Results Files

Version 1.2-3

Date 2023-02-03

Author Peter J. van der Most,
Leanne K. Kupers,
Ilja Nolte

Maintainer Peter J. van der Most <p.j.van.der.most@umcg.nl>

Depends R (>= 4.0.0), methods

Description Tools for (automated and manual) quality control of the results of Epigenome-Wide Association Studies.

License GPL (>= 3)

NeedsCompilation no

Repository CRAN

Date/Publication 2023-02-07 12:22:32 UTC

R topics documented:

QCEWAS-package 2
EWAS_plots 2
EWAS_QC 4
EWAS_series 8
P_correlation 10
P_lambda 11
translate_header 12

Index **15**

QCEWAS-package *Quality Control of Epigenome-Wide Association Study Results*

Description

Functions for automated and manual quality control of Epigenome-Wide Association Study results.

Details

Package:	QCEWAS
Type:	Package
Version:	1.2-3
Date:	2023-02-03
License:	GPL (>= 3)

When combining the results of multiple Epigenome-Wide Association Studies (EWAS) into a meta-analysis, it is recommended to run a quality check (QC) over the individual files first. This is not only to check if the individual results are valid, reliable, and of high quality, but also to ensure that they are comparable between cohorts. The `QCEWAS` package was developed to automate this process.

The core function of `QCEWAS` is `EWAS_QC`. This function accepts the filename of a single EWAS results file and performs a thorough quality check. For more information, see the `EWAS_QC` help page. The function `EWAS_series` is a wrapper function that accepts multiple file names, and then runs `EWAS_QC` for all of them. It also generates a few additional plots to compare the results of the files, allowing the user to spot differences in effect-size distribution or standard errors.

The functions `EWAS_plots` and `P_correlation` are subroutines of `EWAS_QC` that can also be called by the user to perform specific QC tasks.

QCEWAS also includes a Quick-Start Guide. A link to the guide is provided when the package is loaded into R.

Author(s)

P.J. van der Most, Leanne K. Kupers and Ilja M. Nolte

Maintainer: P.J. van der Most <p.j.van.der.most@umcg.nl>

EWAS_plots *Manhattan and Quantile-Quantile plots for EWAS results files*

Description

This function is used by `EWAS_QC` to generate quantile-quantile (QQ) and Manhattan plots. It can also be called by users. Note that it does not generate the histogram or volcano plot - 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,
           high_quality_plots = FALSE,
           save_name = "dataset",
           header_translations)
```

Arguments

`dataset` either a vector of p-values, or a data frame containing the columns `CHR` (chromosome number), `MAPINFO` (base-pair position), and `P_VAL` (p-value). If different columnnames are used, the `header_translations` argument can be used to translate these. `CHR` and `MAPINFO` are only required for generating a Manhattan plot. Note that, unlike `EWAS_QC`, this function 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 and 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.

`high_quality_plots`
logical. Setting this to TRUE will save the graphs as high-resolution tiff images.

`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 of `dataset` to the standard names. 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 (when chromosome and position data are included) 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.

EWAS_QC *Automated Quality Control of EWAS results files*

Description

The main function of the QCEWAS package. EWAS_QC accepts a single EWAS results file and runs a thorough quality check (QC), 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),
        markers_to_exclude,
        exclude_outliers = FALSE,
        exclude_X = FALSE, exclude_Y = FALSE,
        save_final_dataset = TRUE, gzip_final_dataset = TRUE,
        header_final_dataset = "standard",
        high_quality_plots = FALSE,
        return_beta = FALSE, N_return_beta = 500000L,
        ...)
```

Arguments

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 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 TARGETID, CHR (chromosome), and MAPINFO (position), using those exact names. Other columns may be included but will be ignored. 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.

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 argument to c(NA, NA).

markers_to_exclude
: Either a vector or data frame containing a list of CpG IDs that need to be excluded before starting the QC (in case of a data frame only the first column will be processed), or the name of a file containing the same. This argument is optional: if not specified, no exclusions are made. Note that when a single value (a vector of length 1) is passed to this argument, EWAS_QC will treat it as a filename even when no such file can be found. If you want to remove a single CpG, either pass it to this argument via a file, or add a dummy value to the vector to give it length 2 (e.g. c("cg02198983", "dummy")).

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 providing a map to EWAS_QC via the map argument.

save_final_dataset
: logical determining whether the cleaned dataset will be saved.

gzip_final_dataset
: logical determining whether the saved dataset will be compressed in the .gz format.

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.

high_quality_plots
: logical. Setting this to TRUE will save the graphs as high-resolution tiff images.

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 value includes a vector of effect sizes. N_return_beta defines the length of the vector.

... arguments passed to read.table for importing the EWAS results file.

Details

QCEWAS includes a Quick-Start guide in the doc folder of the library. This guide will explain how to run a QC and how to interpret the results. The start-up message when loading QCEWAS will indicate where it can be found on your computer. In brief, the QC consists of the following 5 stages:

- Checking data integrity:

  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 sets them to NA.
- Filter for outliers and sex-chromosomes (optional)

  Counts the number of outlying markers, as well as chromosome X and Y markers, and deletes them if specified. The markers named in markers_to_exclude are removed here as well.
- Generating QC plots

  A histogram of beta and standard error distribution is plotted.

  The p-values are checked by correlating and plotting them against p-values calculated from the effect size and standard error.

  A QQ plot is generated to test for over/undersignificance.

  A Manhattan plot is generated to see where the signals (if any) are located.

  A Volcano plot is generated to check the distribution of effect sizes vs. p values.
- Creating a QC log

  The log contains notes about any problems encountered during the QC, as well as several tables describing the data.
- 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 effect size 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.

mean_methylation
a NULL: this functionality has not been implemented yet.

effect_size if return_beta is TRUE, this is a numeric vector of length N_return_beta, containing a random selection of effect sizes from the filtered dataset. If FALSE, this will be NULL.

Note

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.

See Also

See EWAS_series for running a QC over multiple files.

See EWAS_plots and P_correlation for carrying out specific steps of the QC.

Examples

```
# For use in this example, the 2 sample files in the
# extdata folder of the QCEWAS library will be copied
# to your current R working directory. Running the QC
# generates 7 new files in your working directory:
# a cleaned, post-QC dataset, a log file, and 5 graphs.
# Consult the Quick-Start guide for more information on
# how to interpret these.
## Not run:
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "sample_map.txt.gz"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "sample1.txt.gz"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)

QC_results <- EWAS_QC(data = "sample1.txt.gz",
                      map = "sample_map.txt.gz",
                      outputname = "sample_output",
                      threshold_outliers = c(-20, 20),
                      exclude_outliers = FALSE,
                      exclude_X = TRUE, exclude_Y = FALSE,
                      save_final_dataset = TRUE, gzip_final_dataset = FALSE)

## End(Not run)
```

EWAS_series *Quality Control and Comparison of multiple EWAS results files*

Description

This function runs a QC (via the function EWAS_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,
            high_quality_plots = FALSE,
            N_plot_beta = 500000L,
            ...)
```

Arguments

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 (as listed in the EWAS_files argument) and sample sizes of the datasets, or the name of a file containing the same. The data frame must contain the columns file and N, with those exact names. All files listed in the EWAS_files argument must be included in the file column. This argument is optional: if not specified, EWAS_series will not generate a precision plot.

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 details.

save_final_dataset, gzip_final_dataset, high_quality_plots
logical values. See EWAS_QC for details.

N_plot_beta integer specifying how many beta values per file should be used in the effect-size comparison plot. Set this to a value larger than the number of markers in the datasets to include all markers.

... arguments passed to EWAS_QC.

Details

QCEWAS includes a Quick-Start guide in the doc folder of the library. This guide will explain how to run a QC and how to interpret the results. The start-up message when loading QCEWAS will indicate where it can be found on your computer. In brief, 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.

As of version 1.2-0, the effect-size distribution plot shows a random (rather than proportional) selection of effect-sizes from the cohort. As a consequence, rerunning QC over a dataset may result in a slightly different distribution plot in each run. This is only a cosmetic issue (as the default sample size is sufficiently large to include the majority of a normally-sized EWAS dataset) and can be averted entirely by changing the N_plot_beta argument to a value exceeding the number of markers in the dataset(s).

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 (also saved as EWAS_QC_legend.txt), 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.

See Also

EWAS_QC

Examples

```
# For use in this example, the 4 sample files in the
# extdata folder of the QCEWAS library will be copied
# to your current R working directory. Running the QC
# generates several files in your working directory:
# consult the Quick-Start Guide for more information
# on how to interpret these.
## Not run:
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "sample_map.txt.gz"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "sample1.txt.gz"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "sample2.txt.gz"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "translation_table.txt"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)

sample_list <- c("sample1.txt.gz", "sample2.txt.gz")
sample_N <- data.frame(file = sample_list,
                       N = c(77, 79),
                       stringsAsFactors = FALSE)



QC_results <- EWAS_series(EWAS_files = sample_list,
                          output_files = c("sample_output1", "sample_output2"),
                          map = "sample_map.txt.gz",
                          N = sample_N,
                          header_translations = "translation_table.txt",
                          save_final_dataset = FALSE,
                          threshold_outliers = c(-20, 20),
                          exclude_outliers = FALSE,
                          exclude_X = TRUE, exclude_Y = FALSE)

## End(Not run)
```

P_correlation *Testing P-value distribution*

Description

A sub-function of EWAS_QC that tests if the reported p-values match the p-value that can be derived from the effect-size 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 correlation.

Usage

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

Arguments

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.

high_quality_plots
logical. Setting this to TRUE will save the graph as a high-resolution tiff image.

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 plot function.

Details

P_correlation is primarily 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 *Calculation of the Lambda value*

Description

The Lambda value represents the inflation of p-values compared to a normal distribution of p.

Usage

```
P_lambda(p)
```

Arguments

p a numeric vector of p-values

Details

The function removes any missing values from p, and then returns:

`median(qchisq(p, df=1, lower.tail=FALSE)) / qchisq(0.5, 1)`

The lambda value represents the inflation of the p-values compared to a normal distribution. In a genome-wide study, one would expect the results for the vast majority of CpG sites to accord with the null hypothesis, i.e. the p-values are random, and have a normal distribution. Only sites that are significantly associated with the phenotype of interest should lie outside of the normal distribution.

Ideally the lambda value should be 1. Lambda represents the *overall* difference with the expected distribution - so the presence of a few significant results (i.e. p-values that do not follow the normal distribution) does not bias it.

However, if lambda is 2 or higher, it means that a substantial portion of your dataset is more significant than expected for a genome-wide study (i.e. oversignificance). This could mean your dataset has been filtered for low-significance markers. If this is not the case, you should consider doing a genomic control correction on the p-values, to correct the oversignificance.

Similary, values of 0.8 or lower indicate that your results are less significant than would be expected from a random distribution of p-values.

Value

A single numeric value, the lambda value.

Examples

```
pvector <- ppoints(10000)
P_lambda(pvector)
# The lambda of a random distribution of p-values equals 1

pvector[pvector > 0.9 & pvector < 0.91] <- NA
P_lambda(pvector)
# If low-significance results are removed (i.e. there are more
# significant results than expected) lambda increases
```

`translate_header` *Translate column names into standard names*

Description

This function is used to translate non-standard column names into the standard ones used by `EWAS_QC` and other functions.

Usage

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

Arguments

`header` character vector; the header to be translated.

`standard` character vector; the names `header` should be translated into.

`alternative` translation table; see below for more information.

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.

Value

`translate_header` returns an object of class 'list' with 6 components:

`header_h` character vector; the translated header. Unknown columns are included under their old names.

`missing_h` character vector; the standard column names that were not found. If none, this returns `NULL`.

`unknown_h` character vector; column names that could not be converted to a standard name. Note that these columns are also included in `header_h`. If none, this returns `NULL`.

`header_N`, `missing_N`, `unknown_N`
integer; the lengths of the above three vectors

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 be in a separate row.
- The alternatives must be capitalized.
- No duplicate alternatives are allowed.
- A header line is not required, and will be ignored if present.

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.

Examples

```
# For use in this example, the 2 sample files in the
# extdata folder of the QCEWAS library will be copied
# to your current R working directory
## Not run:
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "sample2.txt.gz"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)
file.copy(from = file.path(system.file("extdata", package = "QCEWAS"),
                           "translation_table.txt"),
          to = getwd(), overwrite = FALSE, recursive = FALSE)

sample_ewas <- read.table("sample2.txt.gz", header = TRUE,
                          stringsAsFactors = FALSE, nrow = 10)
colnames(sample_ewas)


translation_table <- read.table("translation_table.txt", header = TRUE,
                                stringsAsFactors = FALSE)
sample_translation <- translate_header(header = colnames(sample_ewas),
                                       alternative = translation_table)
sample_translation

colnames(sample_ewas) <- sample_translation$header_h

colnames(sample_ewas)

## End(Not run)
```

Index

∗ **package**
    QCEWAS-package, 2

EWAS_plots, *2*, 2, *7*
EWAS_QC, *2*, 4, *8–12*
EWAS_series, *2*, *4–7*, 8

P_correlation, *2*, *7*, 10
P_lambda, 11
plot, *11*

QCEWAS, *4*
QCEWAS (QCEWAS-package), 2
QCEWAS-package, 2

read.table, *4*, *6*

translate_header, *3*, *5*, *8*, *11*, 12