



Van Der Most, P. J., Küpers, L. K., Snieder, H., & Nolte, I. (2017). QCEWAS: automated quality control of results of epigenome-wide association studies. *Bioinformatics*, 33(8), 1243-1245. <https://doi.org/10.1093/bioinformatics/btw766>

Peer reviewed version

Link to published version (if available):
[10.1093/bioinformatics/btw766](https://doi.org/10.1093/bioinformatics/btw766)

[Link to publication record on the Bristol Research Portal](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Oxford University Press at <https://academic.oup.com/bioinformatics/article/33/8/1243/2949506> . Please refer to any applicable terms of use of the publisher.

University of Bristol – Bristol Research Portal

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/brp-terms/>

QCEWAS: automated quality control of results of epigenome-wide association studies

Peter J. van der Most^{1†}, Leanne K. Küpers^{1†}, Harold Snieder¹, Ilja Nolte^{1*}

¹ Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

† Equal contributions as first authors

*To whom correspondence should be addressed.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

ABSTRACT

Summary: The increasing popularity of epigenome-wide association studies (EWAS) has led to the establishment of several large international meta-analysis consortia. However, when using data originating from multiple sources, a thorough and centralized quality control is essential. To facilitate this, we developed the QCEWAS R package. QCEWAS enables automated quality control of results files of EWAS. QCEWAS produces cohort-specific statistics and graphs to interpret the quality of the results files, graphs comparing results of multiple cohorts, as well as cleaned input files ready for meta-analysis.

Availability and Implementation: <https://cran.r-project.org/web/packages/QCEWAS>

Contact: i.m.nolte@umcg.nl

1 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* (February 2012 41:1 and August 2015 44:4). DNA methylation is one of the most studied and best understood mechanisms in epigenetics. It is often measured using the Infinium HumanMethylation27 or HumanMethylation450 BeadChip or the MethylationEPIC kit (Illumina Inc., San Diego, USA)..

Given the frequent use of these chips, 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 Cytosine-phosphate-Guanine sites that are associated with a disease or trait of interest and indeed the first meta-EWAS studies were recently published (Joubert et al. 2015; Gruzieva et al. 2016).

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.

Therefore, we developed the QCEWAS software package, allowing fast and easy assessment of the quality of EWAS results files through informative figures and statistics. Compared to other software packages for the QC of EWAS results (Assenov et al., 2014; Kilaru et al., 2012), QCEWAS allows for more extensive QC of individual cohort's results as well as for a side-by-side comparison of multiple EWAS results. Additionally, the

QCEWAS package can generate cleaned input files for use in existing meta-analysis programs.

2 APPROACH

2.1 Implementation

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 (<https://cran.r-project.org>).

2.2 Usage

The QCEWAS package includes several functions, the most important ones being 'EWAS_QC' and 'EWAS_series'. The first performs a thorough QC on a single EWAS results file; the second processes a series of results files by calling 'EWAS_QC' for every file and additionally produces graphs to compare data from the results files to one another.

The package can be used to process EWAS results from analyses using the 450k chip, the older 27k chip, as well as the newly developed MethylationEPIC chip (>850k CpGs).

3 RESULTS

3.1 Features

For each results file, QCEWAS carries out the following checks:

- data integrity: are the required data present and valid (e.g. no negative standard errors [SE] or p-values)?;
- cohort-specific outlier detection and removal (optional);
- allosomal sites removal (optional);
- removal of questionable sites, e.g. polymorphic and cross-reactive sites (Chen et al. 2013) (optional);
- effect size and SE distribution in histograms;
- checking reported p-values by comparing them to p-values calculated from effect size and SE;
- a QQ plot to assess over-/under-significance of results (Figure 1A);
- the distribution of effect sizes versus p-values in a volcano plot (Figure 1B);
- the location of the signals in a Manhattan plot (Figure 1C).

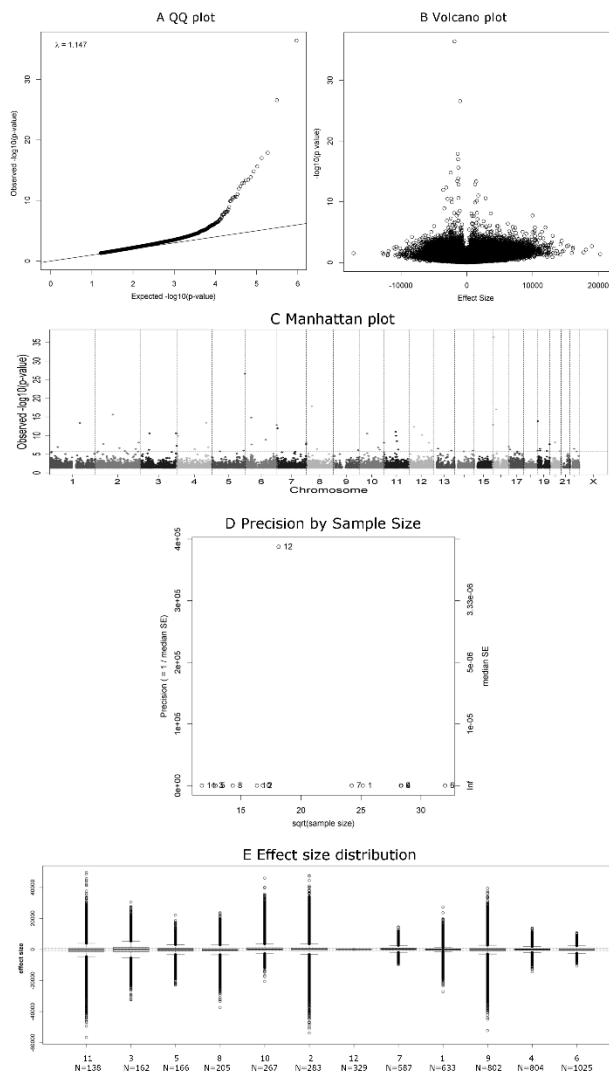


Figure 1. Diagnostic plots showing cohort-specific (A-C) and between-cohort QC characteristics (D-E) produced by QCEWAS.

Additionally, two figures are produced using the `EWAS_series` function to compare QC statistics of multiple EWAS results files assuming that the range of phenotypic values analyzed by the different cohorts is similar: a precision plot, to test if precision ($1/\text{median}[\text{SE}]$) increases proportionally with sample size (Figure

1D); and a boxplot showing the distributions of the effect sizes per file (Figure 1E). Figure 1D shows one cohort file (no. 12) with a precision that is higher than expected. It is expected that all studies are on a diagonal line. Much deviation from this line may indicate phenotype scaling issues and that needs to be checked by the respective cohort. Figure 1E shows one clear outlying cohort (no. 12) with a much smaller spread of effect sizes than expected based on sample size, suggesting use of a different measure or analysis model.

Finally, QCEWAS can produce cleaned EWAS results files that are ready for meta-analysis by GWAMA (Mägi and Morris, 2010), META (Liu et al. 2010), METAL (Willer et al. 2010), or PLINK (Purcell et al. 2007).

3.2 Performance

On a 64-bit Windows 7 computer with 2.4GHz and 48GB RAM, a QC of an EWAS results file with ~470,000 markers takes less than a minute and 400 MB of RAM.

4 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.

ACKNOWLEDGEMENTS

We would like to acknowledge the members of the Pregnancy and Child Epigenetics (PACE) consortium for providing multiple results files to use as input for testing the QCEWAS R package.

Conflict of Interest: none declared.

REFERENCES

- Assenov Y, et al. (2014) Comprehensive analysis of DNA methylation data with RnBeads. *Nat Methods* 11 : 1138–1140
- Chen YA, et al. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 8(2) : 203-209
- Gruzieva O, et al. Epigenome-wide meta-analysis of methylation in children related to prenatal NO₂ air pollution exposure. *Environ Health Perspect.* 2016 [Epub ahead of print].
- Joubert BR, et al. (2016) DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. *Am J Hum Genet.* 98(4) : 680-696.
- Kilaru V, et al. (2012). MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics* 7 : 225-229.
- Liu JZ, et al. (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 42 : 436-440.
- Mägi R and Morris AP (2010) GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 11 : 288.
- Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81 : 559-575.
- Rakyan VK, et al. Epigenome-wide association studies for common human diseases. *Nat Rev Genet* 2011(12) : 529-541.
- Willer CJ, et al. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26 : 2190-2191.