Development of bioinformatic tools and application of novel statistical methods in genome-wide analysis

van der Most, Peter Johannes

Chapter 3

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

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

* Authors contributed equally

Bioinformatics, 2017, 33 (8): 1243-1245

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.

The package is available at:
https://cran.r-project.org/web/packages/QCEWAS

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 (CpG) sites that are associated with a disease or trait of interest and indeed the first meta-EWAS studies were recently published[1,2].

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[3,4], 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.

Approach

Implementation

QCEWAS was built as a package for R[5]. 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).

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).

Results

Features

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

a. data integrity: are the required data present and valid (e.g. no negative standard errors [SE] or p-values)?;
b. cohort-specific outlier detection and removal (optional);
c. allosomal sites removal (optional);
d. removal of questionable sites, e.g. polymorphic and crossreactive sites[6] (optional);
e. effect size and SE distribution in histograms;
f. checking reported p-values by comparing them to p-values calculated from effect size and SE;
g. a QQ plot to assess over-/under-significance of results (Figure 1A);
h. the distribution of effect sizes versus p-values in a volcano plot (Figure 1B);
i. the location of the signals in a Manhattan plot (Figure 1C).

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/median[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[7], META[8], METAL[9], or PLINK[10].

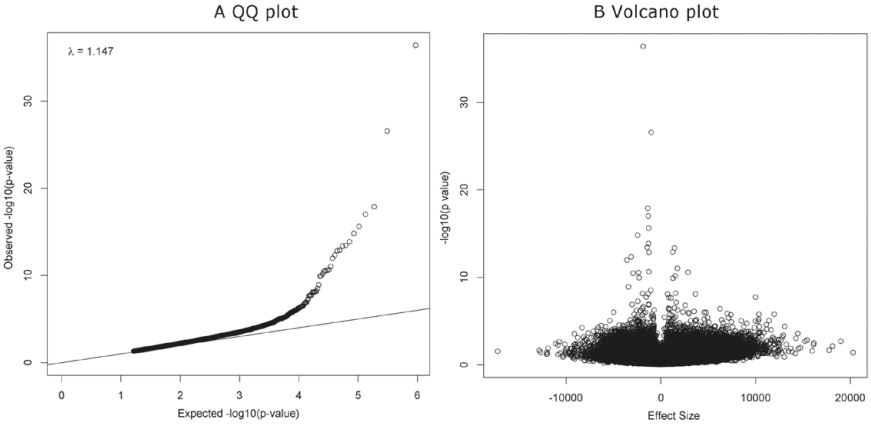
Performance

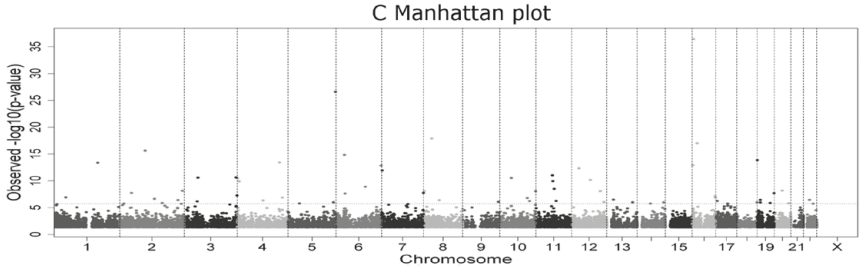
On a 64-bit Windows 7 computer with 2.4GHz and 48GB RAM, a QC of an EWAS results file with ~470,000 markers takes less than a minute and 400 MB of RAM.

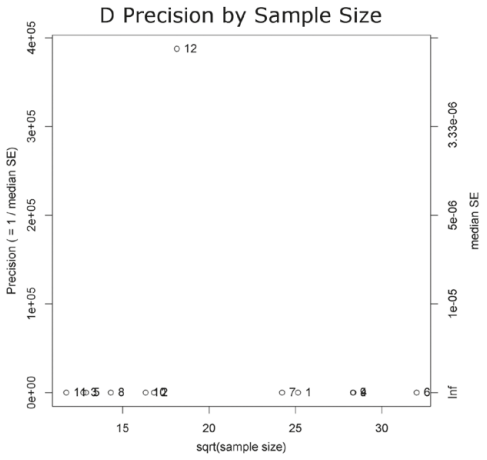
Conclusion

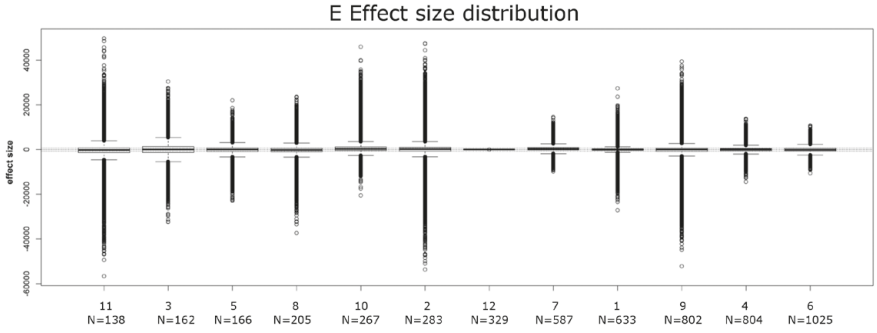
With QCEWAS we developed a flexible and easy-to-use software package for a complete QC of EWAS results files, allowing users to detect errors that would have biased the subsequent meta-analysis.

FIGURE 1 | A selection of diagnostic plots showing cohort-specific (A–C) and between-cohort QC characteristics (D,E) produced by QCEWAS.


A QQ plot
λ = 1.147
Observed -log10(p-value)
Expected -log10(p-value)
B Volcano plot
-log10(p value)
Effect Size


C Manhattan plot
Observed -log10(p-value)
Chromosome


D Precision by Sample Size
Precision (= 1 / median SE)
median SE
sqrt(sample size)


E Effect size distribution
effect size
11 N=138
3 N=162
5 N=166
8 N=205
10 N=267
2 N=283
12 N=329
7 N=587
1 N=633
9 N=802
4 N=804
6 N=1025

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.

Funding

This work was funded by the department of epidemiology, University of Groningen, Medical Center Groningen, the Netherlands.

Conflict of Interest: none declared.

References

1. Joubert BR, Felix JF, Yousefi P *et al:* DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. *Am J Hum Genet* 2016; **98:** 680-696.
2. Gruzieva O, Xu C, Breton CV *et al:* Epigenome-Wide Meta-Analysis of Methylation in Children Related to Prenatal NO2 Air Pollution Exposure. *Environ Health Perspect* 2017; **125:** 104-110.
3. Assenov Y, Mueller F, Lutsik P, Walter J, Lengauer T, Bock C: Comprehensive analysis of DNA methylation data with RnBeads. *Nature Methods* 2014; **11:** 1138-1140.
4. Kilaru V, Barfield RT, Schroeder JW, Smith AK, Conneely KN: MethLAB A graphical user interface package for the analysis of array-based DNA methylation data. *Epigenetics* 2012; **7:** 225-229.
5. R Core Team: R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing, 2016.
6. Chen Y, Lemire M, Choufani S *et al:* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013; **8:** 203-209.
7. Magi R, Morris AP: GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010; **11:** 288.
8. Liu JZ, Tozzi F, Waterworth DM *et al:* Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet* 2010; **42:** 436-440.
9. Willer CJ, Li Y, Abecasis GR: METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26:** 2190-2191.
10. Purcell S, Neale B, Todd-Brown K *et al:* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81:** 559-575.