BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 8 2014, pages 1185–1186
doi:10.1093/bioinformatics/btt745

Genetics and population analysis

Advance Access publication January 5, 2014

QCGWAS: A flexible R package for automated quality control of genome-wide association results

Peter J. van der Most[1], Ahmad Vaez[1], Bram P. Prins[1,2], M. Loretto Munoz[1], Harold Snieder[1], Behrooz Z. Alizadeh[1] and Ilja M. Nolte[1,*]

[1]Department of Epidemiology, University of Groningen, University Medical Center Groningen, P.O. box 30.001, 9700 RB Groningen, The Netherlands and [2]Cardiogenetics Lab, Human Genetics Research Centre, St. George's Hospital Medical School, London SW17 0RE, UK

Associate Editor: Martin Bishop

ABSTRACT

Summary: QCGWAS is an R package that automates the quality control of genome-wide association result files. Its main purpose is to facilitate the quality control of a large number of such files before meta-analysis. Alternatively, it can be used by individual cohorts to check their own result files. QCGWAS is flexible and has a wide range of options, allowing rapid generation of high-quality input files for meta-analysis of genome-wide association studies.

Availability: http://cran.r-project.org/web/packages/QCGWAS

Contact: i.m.nolte@umcg.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 23, 2013; revised on November 21, 2013; accepted on December 18, 2013

1 INTRODUCTION

The number of consortia aiming to identify genes for complex traits through meta-analysis of genome-wide association studies (GWAS) has mushroomed in the past 6 years. The advantage of this strategy is that large sample sizes can be reached, allowing detection of genetic variants with small effects. A downside is the lack of unified quality control (QC) on the GWAS analyses of the individual cohorts, as each cohort will typically perform their own analysis according to a standard analysis plan and share only summary statistics. GWAS result files are prone to errors due to the vast amount of data they contain and the different manner in which these data are generated by individual cohorts. Before combining data from individual studies in a meta-analysis, it is important to ensure that all data included are valid, of high quality and compatible between cohorts to reduce both the false-positive and the false-negative findings (de Bakker *et al.*, 2008). Because GWAS result files usually contain a standard set of variables, it is feasible to automate the QC of these files, thereby gaining speed, reliability, flexibility and the possibility to perform more elaborate checks.

To our knowledge, the only other software package currently available for QC of GWAS result files is GWAtoolbox (Fuchsberger *et al.*, 2012). However, GWAtoolbox does not produce cleaned results files, is less flexible regarding file format and uses a restrictive format for the QC log. This makes it less suited for processing (and comparing) large numbers of files in preparation of a meta-analysis. It also does not check allele information or allow for the retesting of individual QC steps. To address these shortcomings, we developed QCGWAS with the aim to automate QC and allow rapid generation of high-quality input files for GWAS meta-analyses.

*To whom correspondence should be addressed.

2 APPROACH

2.1 Implementation

QCGWAS is built as a package for R (R Development Core Team, 2012). The R platform was chosen because it is operating system-independent, commonly used, open source, can handle large datasets and is flexible regarding input file format. QCGWAS requires R version 3.0.1 or later (64-bit recommended) and can be downloaded from the Comprehensive R Archive Network Web site (http://cran.r-project.org).

2.2 Usage

The main QC by QCGWAS is executed by the QC_series(...) command. This function requires a minimum of two parameters: a list of filenames of GWAS result files and a translation table for the file headers. All other parameters are optional, allowing for a flexible and user-customized QC.

2.3 Approach

A standard QC consists of six steps (Fig. 1):

STAGE 1: a GWAS result file is inspected for missing and invalid data. Duplicated single nucleotide polymorphisms (SNPs) and SNPs lacking crucial variables are removed.

STAGE 2: alleles and strand information are checked and fixed by matching it to a given reference (e.g. HapMap). The SNPs can be removed when their alleles or allele frequencies do not match the reference. This harmonizes the alleles across result files. Next, it correlates the reported allele frequencies for all SNPs to those from the reference set and generates scatter plots to show deviations (Supplementary Fig. S1).

STAGE 3: QC plots are generated (see Supplementary Fig. S2–S4). These include histograms of the distribution of SNP quality parameters (allele frequencies, Hardy-Weinberg equilibrium *P*-values, call rates and imputation quality), a Manhattan plot and a series of Quantile-Quantile (QQ) plots filtered for SNP quality.

STAGE 4: various QC statistics are calculated, of which the most important are the genomic-control lambda to check for population stratification

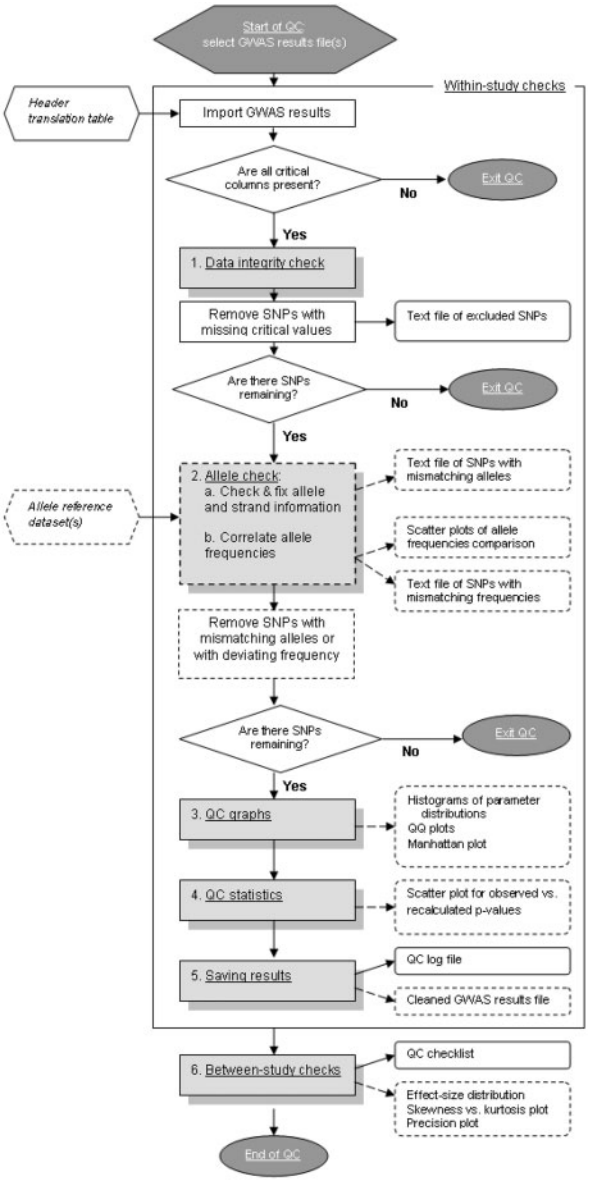

Fig. 1. Flow diagram of the six steps (marked by light grey-shaded rectangles) comprising the default QC performed by QCGWAS. Input files are indicated by hexagons and the created output files by rounded rectangles. Dashed lines indicate that the check is optional

(Devlin and Roeder, 1999), Visscher's statistic (Yang *et al.*, 2012) to determine whether the standard errors are in line with the sample size reported, the skewness and kurtosis of the effect-size distribution and the correlation between the reported *P*-values and those calculated from the effect size and standard error.

STAGE 5: the cleaned GWAS result file is saved and extensive QC information is written to a log file. The cleaned file can be saved in different formats, ensuring compatibility for immediate meta-analysis by GWAMA (Mägi and Morris, 2010), META (Liu *et al.*, 2010), MetABEL (Aulchenko *et al.*, 2007), METAL (Willer *et al.*, 2010) or PLINK (Purcell *et al.*, 2007).

STAGE 6: several between-study checks are performed, including a comparison of skewness and kurtosis, of sample sizes and standard errors and of effect-size range to identify incorrect units and/or trait transformations (Supplementary Fig. S5). A checklist of QC statistics is also created.

Each of the steps of the QC can be enabled or disabled by the user, allowing for a flexible QC pipeline, and quick retests of particular steps. Finally, independent functions are provided for the creation of histograms or QQ plots using combinations of filter parameters and regional association plots.

2.4 Performance

On a Windows 7 computer with 2.4 GHz and 48 GB RAM, a QC of a HapMap-imputed GWAS result file (2.5 million SNPs) takes between 5 and 15 min/file. Memory usage is between 2 and 3 GB, depending on the number of graphs to be created. Sequence-imputed results files, such as 1000 Genomes-based data (The 1000 Genomes Project Consortium, 2012) take ~40 min and 20 GB of RAM.

3 CONCLUSION

QCGWAS is a flexible and comprehensive package for automated QC of GWAS result files. It can handle a large number of files within reasonable time and is therefore particularly useful for a centralized QC preceding a GWAS meta-analysis. It can also be used by individual cohorts to inspect the quality of their results. Currently it is geared toward quantitative traits, but case-control results can also be used with proper transformations. Future versions of the package are under development to accommodate non-SNP variants, such as used in sequence-based GWAS data.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Josée Dupuis for constructive discussions and for sharing her QC procedure and scripts at an early stage. They are also grateful to Nicola Barban and Jornt Mandemakers for their useful feedback on the use of QCGWAS.

Conflict of Interest: none declared.

REFERENCES

Aulchenko,Y.S. *et al.* (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics*, **23**, 1294–1296.

De Bakker,P.I.W. *et al.* (2008) Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.*, **17**, R122–R128.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Fuchsberger,C. *et al.* (2012) GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. *Bioinformatics*, **28**, 444–445.

Liu,J.Z. *et al.* (2010) Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat. Genet.*, **42**, 436–440.

Mägi,R. and Morris,A.P. (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.

R Development Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Willer,C.J. *et al.* (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.

Yang,J. *et al.* (2012) FTO genotype is associated with phenotypic variability of body mass index. *Nature*, **490**, 267–272.