Development of bioinformatic tools and application of novel statistical methods in genome-wide analysis
van der Most, Peter Johannes

Publication date:
2017

Chapter 2

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

Peter J. van der Most, Ahmad Vaez, Bram P. Prins, M. Loretto Munoz, Harold Snieder, Behrooz Z. Alizadeh, Ilja M. Nolte

2

Abstract

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.

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

Supplementary information can be found at Bioinformatics Online:
https://academic.oup.com/bioinformatics

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[1]. 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[2]. 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.

Approach

Implementation

QCGWAS is built as a package for R[3]. 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 Website (http://cran.r-project.org).

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.

Approach

A standard QC consists of six steps (Figure 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 Figure S1).

Stage 3: QC plots are generated (see Supplementary Figure S2-4). 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[4], Visscher's statistic[5] 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[6], META[7], MetABEL[8], METAL[9] or PLINK[10].
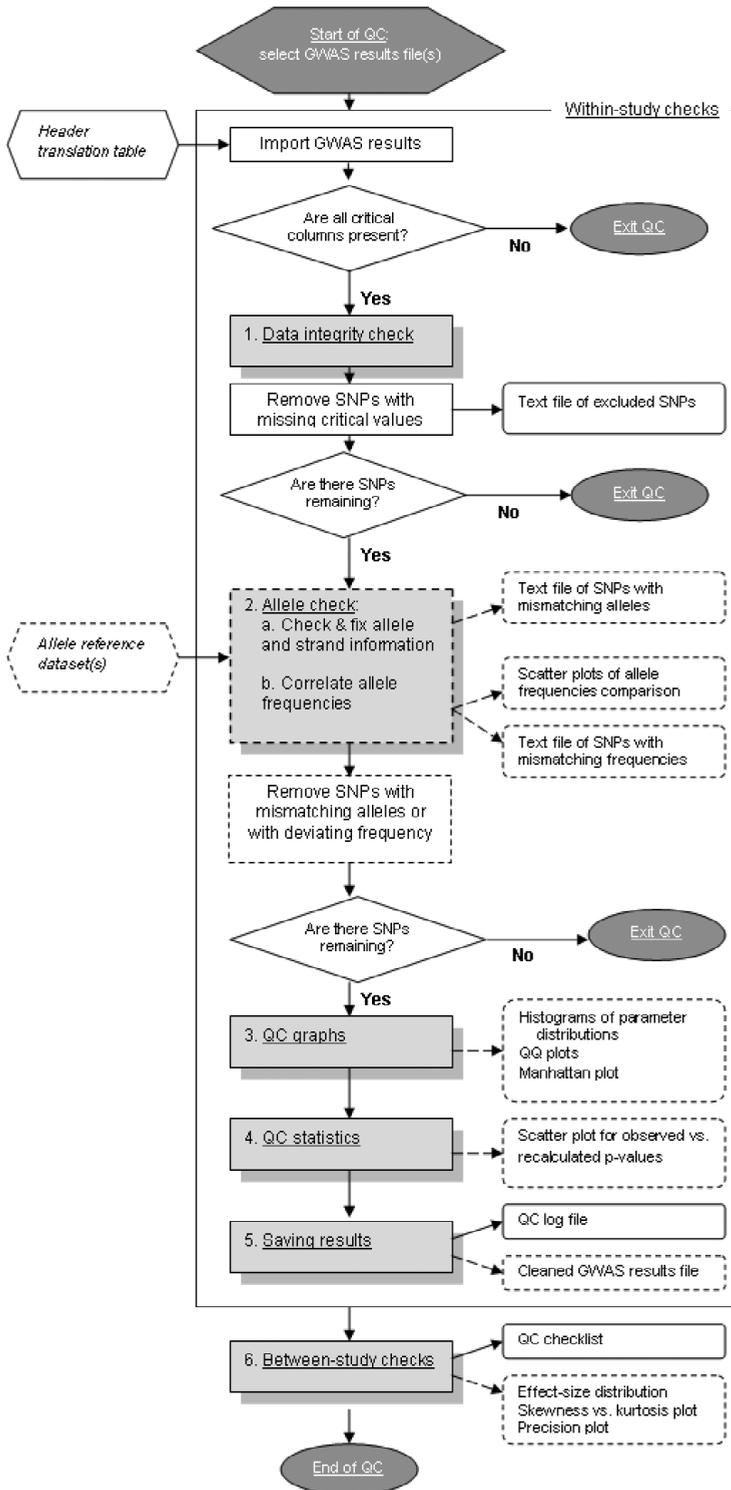
Stage 6: several between-study checks are performed, including a comparison of skewness and kurtosis, of sample sizes and standard errors and of effect-size range to identify incorrect units and/or trait transformations (Supplementary Figure S5). A checklist of QC statistics is also created.

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.

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[11] take ~40 minutes and 20 GB of RAM.

FIGURE 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.

2

```
                    ┌─────────────────────┐
                    │   Start of QC:       │
                    │ select GWAS results  │
                    │      file(s)         │
                    └─────────────────────┘
```

Within-study checks

Header translation table ──▶ Import GWAS results

Are all critical columns present? ── No ──▶ Exit QC

Yes

1. Data integrity check

Remove SNPs with missing critical values ──▶ Text file of excluded SNPs

Are there SNPs remaining? ── No ──▶ Exit QC

Yes

Allele reference dataset(s) ──▶ 2. Allele check:
a. Check & fix allele and strand information

b. Correlate allele frequencies

──▶ Text file of SNPs with mismatching alleles

──▶ Scatter plots of allele frequencies comparison

──▶ Text file of SNPs with mismatching frequencies

Remove SNPs with mismatching alleles or with deviating frequency

Are there SNPs remaining? ── No ──▶ Exit QC

Yes

3. QC graphs ──▶ Histograms of parameter distributions
QQ plots
Manhattan plot

4. QC statistics ──▶ Scatter plot for observed vs. recalculated p-values

5. Saving results ──▶ QC log file
──▶ Cleaned GWAS results file

6. Between-study checks ──▶ QC checklist
──▶ Effect-size distribution
Skewness vs. kurtosis plot
Precision plot

End of QC

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

References

1. de Bakker PIW, Ferreira MAR, Jia X, Neale BM, Raychaudhuri S, Voight BF: Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008; **17:** R122-R128.
2. Fuchsberger C, Taliun D, Pramstaller PP, Pattaro C, CKDGen Consortium: GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. *Bioinformatics* 2012; **28:** 444-445.
3. R Core Team: R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing, 2012.
4. Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55:** 997-1004.
5. Yang J, Loos RJF, Powell JE *et al:* FTO genotype is associated with phenotypic variability of body mass index. *Nature* 2012; **490:** 267-272.
6. Magi R, Morris AP: GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 2010; **11:** 288.
7. 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.
8. Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM: GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 2007; **23:** 1294-1296.
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.
11. Altshuler DM, Durbin RM, Abecasis GR *et al:* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012; **491:** 56-65.