1 INTRODUCTION

Meta-analyses of genome-wide association studies (GWAS) have proven to be powerful tools to uncover novel loci associated with a variety of complex traits (Manolio et al., 2009). Current GWAS meta-analyses often involve large numbers of studies (from dozens to >100 in large collaborative efforts such as those based on the Metabochip). Typically, each study provides summary statistics on the association between the study phenotype and >25 to ∼37 million single nucleotide polymorphisms (SNPs), depending on whether HapMap or 1000 Genomes Project data are used as reference sets. As a result, analysts dealing with GWAS meta-analyses need to process many files each of size >200 Mb, even when only minimal information is shared.

Given that GWAS involved in such meta-analyses may differ by P-value inflation, SNP filtering) during the meta-analysis process. However, these tools perform only basic sanity checks before meta-analyzing the data and do not allow extensive QC of GWAS result files (for an extensive comparison of popular tools, see Supplementary Table S1). Therefore, in-depth QC and handling of GWAS result files are frequently performed using in-house scripts. These scripts are often not very efficient in handling dozens of large files, and issues may be addressed in a non-systematic way. In the absence of a rigorous and efficient QC, undetected bias at the individual study level can introduce spurious heterogeneity, which in turn can increase the false positive rate and decrease power of the meta-analysis (de Bakker et al., 2008).

Here we introduce the GWAtoolbox, an R package that standardizes and accelerates the handling and QC of post-GWAS data. Its purposes are to make data handling easier, to support a more systematic QC and to facilitate the inspection of data quality prior to use of the GWAS result files in a meta-analysis (Supplementary Fig. S2).

2 APPROACH

2.1 Three-step QC

Inspired by the analysis protocols of several consortia (Köttgen et al., 2010; Pfeufer et al., 2009; Teslovich et al., 2010) and by well-accepted GWAS guidelines (de Bakker et al., 2008), we identified three main steps that should precede any GWAS meta-analysis:

(i) File-format check. Adherence of GWAS result files to prescribed formatting guidelines. Deviations from given guidelines could cause the failure of meta-analysis software or the inclusion of wrong or mis-matched variables in the data analysis.

(ii) Quality check at the individual study level. This can reveal that (a) the underlying association analysis was performed incorrectly or (b) the GWAS results are of poor quality. These problems can be identified as follows:

(a) The presence of unexpected values (e.g. P-value outside of the range [0–1]) or inconsistent data (e.g. availability of the P-value when the estimated effect size is missing) indicate that errors occurred during analysis or at the file management level.

(b) Poor-quality data can be identified by specific indices (e.g. low genotype imputation quality) or by the distribution of specific statistics [e.g. unmodeled population stratification or (cryptic) relatedness can be suspected in case of an inflated P-value distribution]. GWAtoolbox provides summary statistics and plots describing imputation quality, P-value distribution and additional features.
The GWAtoolbox can handle tens of studies with millions of phenotype of interest and so the distribution of z should follow the brief, we can assume that these SNPs are not associated with the phenotype of interest and so the distribution of z should follow the null distribution of no association. The kusk_check() function makes a scatterplot of sk50 and ku50 from all studies. Points should cluster around the same point at (sk50 = 0, ku50 ≥ 0): under the null, z should be symmetric and have the same dispersion across all studies. Departures from the cluster in terms of sk50 indicate asymmetric effect estimates and may reflect errors in the model fitting process, such as incorrect assumptions about the phenotype distribution. Departures in terms of ku50 may reflect incorrect phenotype transformation (e.g. no transformation when logarithm was required) or low genotype imputation quality. The transformation (e.g. no transformation when logarithm was required) or low genotype imputation quality. The dispersion_check() function plots individual study sample size versus mean SE(b) and is meant to identify studies with larger/smaller SE(b) than expected given their sample size (for more details, see Supplementary Tutorial S3 in Supplementary Material).

2.2 Implementation

Our implementation strategy was driven by two needs: (i) to provide maximum performance while keeping low system requirements and (ii) to provide an easy-to-use software producing self-explanatory and high-quality output. Therefore, all computationally intensive data processing steps were written in C++ and made accessible via an R package. Furthermore, GWAtoolbox takes advantage of built-in parallel computing support on modern multicore desktops by performing QC in parallel across studies. To submit multiple files to the QC workflow, we relied on the script format used in METAL, which allows the specification of custom headers and delimiters. We then added specific commands for QC checking.

2.3 Usage

A minimal pre-formatting of individual-study results to adhere to consortium guidelines is assumed. After setting up a simple METAL-like script, where all input files are listed and thresholds for QC parameters are defined, the QC process is initiated by a simple R command: gwasqc("GWASQC_script.txt") or pgwasqc("GWASQC_script.txt", number of processes) for the parallel version. The core output contains a set of HTML documents summarizing the quality of each input file with graphical and textual output. Additionally, for each study all key summary statistics are saved to a text file. Statistics include mean, SD, minimum, maximum, median, skewness and kurtosis of the following parameters: effect estimate, standard error, P-value, minor allele frequency and imputation quality.

2.4 Performance

The GWAtoolbox can handle tens of studies with millions of markers on a desktop computer (Table 1). Memory consumption

<table>
<thead>
<tr>
<th>No. of</th>
<th>Run time, one (two) processes</th>
<th>Memory, one (two) processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>studies</td>
<td>(min) No. of SNPs</td>
<td>2.5M</td>
</tr>
<tr>
<td>10</td>
<td>4.3 (2.5)</td>
<td>18 (10)</td>
</tr>
<tr>
<td>50</td>
<td>22 (13)</td>
<td>88 (52)</td>
</tr>
<tr>
<td>100</td>
<td>43 (25)</td>
<td>175 (104)</td>
</tr>
</tbody>
</table>

Processing the 37 million SNPs imputed GWAS datasets in parallel requires >8 GB of RAM. therefore, no results are reported.

is independent of the number of studies and increases linearly with the number of markers being analyzed.

3 CONCLUSIONS

An earlier version of GWAtoolbox was used successfully by the CKDGen consortium, which performed meta-analyses of >25 GWAS of renal function traits (Böger et al. 2011; Kötten et al. 2010). The fast QC process enabled a quick turn-around so that individual-study analysts could fix problems without causing major delays to the consortium. The total time dedicated to data QC decreased from months to a few weeks. At the time of writing, other GWAS consortia are integrating the GWAtoolbox into their QC meta-analysis workflow.

ACKNOWLEDGEMENTS

We thank Clemens Egger for the IT support. We are grateful to the CKDGen analysis group and to members of the Center for Statistical Genetics, University of Michigan, for testing the software and providing useful feedback. We thank Michael Boehnke and Tanya Teslovich for their critical reading of the manuscript.

Conflict of Interest: none declared.

REFERENCES


