GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data

Christian Fuchsberger1,∗,†, Daniel Taliun2,∗,†, Peter P. Pramstaller2 and Cristian Pattaro2; on behalf of the CKDGen consortium

1Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and 2Center for Biomedicine, European Academy of Bolzano (EURAC), 39100 Bolzano, Italy

ABSTRACT
Summary: The GWAtoolbox is an R package that standardizes and accelerates the handling of data from genome-wide association studies (GWAS), particularly in the context of large-scale GWAS meta-analyses. A key feature of GWAtoolbox is its ability to perform quality control (QC) of any number of files in a matter of minutes. The implemented workflow has been structured to check three particular data quality aspects: (i) data formatting, (ii) quality of the GWAS results and (iii) data consistency across studies. Output consists of an extensive list of quality statistics and plots which allow inspection of individual files and between-study comparison to identify systematic bias.
Availability: http://www.eurac.edu/GWAtoolbox
Contact: cfuchsb@umich.edu; daniel.taliun@eurac.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

Received on June 22, 2011; revised on November 13, 2011; accepted on December 2, 2011

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 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 a large number of 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 (cryptic) relatedness can be suspected in case of an inflated 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 SNPs around the same point at (sk50 = 0, ku50) from all studies. Points should cluster 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("GWAQCScript.txt") or pgwasqc("GWAQCScript.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 is independent of the number of studies and increases linearly with the number of markers being analyzed.

### Table 1. Run time performance under Mac OS X 10.6.8 on a 2.7 GHz Intel Core i7 CPU with 8 GB of RAM using real GWA data

<table>
<thead>
<tr>
<th>No. of studies</th>
<th>Run time, one (two) processes (min) No. of SNPs Memory, one (two) processes (GB) No. of SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>4.3 (2.5) 18 (10) 67 0.3 (0.6) 1.5 (2.5) 4.8</td>
</tr>
<tr>
<td>50</td>
<td>22 (13) 88 (52) 333 0.3 (0.6) 1.5 (2.5) 4.8</td>
</tr>
<tr>
<td>100</td>
<td>43 (25) 175 (104) 669 0.3 (0.6) 1.5 (2.5) 4.8</td>
</tr>
</tbody>
</table>

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

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; Köttgen 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