GWAtoolbox: An R Package for Time Efficient Quality Control of Multiple GWAS Data Files

Daniel Taliun[1,2,*], Christian Fuchsberger[3], Peter P. Pramstaller[1], Cristian Pattaro[1] on behalf of the CKDGen consortium

1. Institute of Genetic Medicine, European Academy Bozen-Bolzano (EURAC), Bolzano, Italy
2. Free University of Bozen-Bolzano, Bolzano, Italy
3. Department of Biostatistics, University of Michigan, Ann Arbor, MI
*Contact author: daniel.taliun@eurac.edu

Keywords: genome-wide association study, quality control, visualization

In the recent years, Genome-Wide Association Studies (GWASs) have been proven to be a very powerful approach to uncover common genetic variants affecting human disease risk and quantitative outcome levels. To date, 1212 genetic loci were reported to be significantly associated with at least one of 210 traits (Hindorff et al. (2011)). To allow sufficient power to identify variants with small effects, GWAS sample size has been augmented by pooling results from dozens of individual GWASs into large meta-analyses efforts. However, combining results from a large number of GWASs, which differ in terms of study design, population structure, data management, and statistical analysis, poses several challenges regarding the consistency and quality of data which are usually difficult to be addressed systematically. This is mainly due to the GWAS file size, which typically includes 2.5 to 7 million rows (corresponding to genetic variants) and > 9 columns (attributes). Consequently, the data harmonization across studies usually takes several weeks or months.

While working in the CKDGen Consortium, aim to detect renal function genes (Köttgen et al. (2010)), we have been performing meta-analyses of several GWASs. To standardize and speed up the quality control (QC) process we developed the **GWAtoolbox**, an *R* package which lightens and accelerates the handling of huge amounts of data from GWASs. **GWAtoolbox** provides time efficient QC of data stored in dozens of files. Based on a simple configuration script, **GWAtoolbox** can process any number of files and produce QC reports in a matter of minutes. QC reports consist of an extensive list of quality statistics and graphical output presented using DHTML, which allow fast and easy inspection of individual data files. Additional statistics and graphs allow quick identification of studies that are systematically different from the other (outliers). The high time efficiency was achieved through the data reduction technique integrated into the visualization pipeline. In particular, instead of passing all the million data points to the *R* plotting routines, only a small part of data points is selected in a such way that preserves the quality of the final graphical output. Additionally, the implementation of all computationally intensive steps was transferred to *C++*.

Through its extensive use in several current GWAS meta-analyses, **GWAtoolbox** has been proven to significantly speed up the data management and to improve the overall meta-analysis quality. **GWAtoolbox** is open source and available for MacOS, Linux, and Windows OS, at http://www.eurac.edu/gwatoolbox.

References

Hindorff, L. et al. (2011). A catalog of published genome-wide association studies. Available at: www.genome.gov/gwastudies. Accessed March 23, 2011.

Köttgen, A. et al. (2010). New loci associated with kidney function and chronic kidney disease. *Nat Genet 42*(5), 376–84.