Dear Editor,

Please find enclosed a manuscript entitled: "Efficient management and analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr." which I am submitting for exclusive consideration of publication as an original paper in Bioinformatics. Please note that this paper has already been uploaded as a bioRxiv preprint (https://www.biorxiv.org/content/early/2017/09/19/190926.full.pdf+html).

Genome-wide datasets produced for association studies have dramatically increased in size over the past years. A range of software and data formats have been developed to perform essential pre-processing steps and data analysis, often optimizing each of these steps within a dedicated implementation. This diverse and extremely rich software environment has been of tremendous benefit for the genetic community. However, it has two limitations: analysis pipelines become very complex and researchers have limited access to diverse analysis tools due to growing data sizes.

The paper describes two new R packages that provide a broad range of tools for analyzing large-scale genome-wide data in a single comprehensive framework. Doing all the analysis steps in R makes it more straightforward to implement a complex pipeline that combines several methods. We show how to perform Genome-Wide Association Study (GWAS) or computation of Polygenic Risk Scores (PRS) with these two packages in only a few lines of code.

In order to adapt to the large modern dataset produced for association studies, the R packages use memory-mapping, a technique that enables fast and seamless access to data stored on disk. We demonstrate the scalability of the two R packages by analyzing a biobank-scale simulated dataset. Moreover, we show that functions of these packages are very fast; for example Principal Component Analysis is 8 times as fast as other software.

Finally, the R packages could be used to develop new statistical methods for genomic studies. For instance, in this paper, several functions (imputation, PCA) make use of other R packages. Users can experiment new data analysis algorithms using only a few lines of R code. Moreover, the first package, bigstatsr, can be used to analyze any data in matrix-like format. As such this paper should be of interest to a broad readership interested in data analysis in diverse and related genomic fields.

Thank you for your consideration of my work.

Sincerely,
Florian Privé (florian.prive@univ-grenoble-alpes.fr)