Reply to reviewers comments

"GCPBayes: An R package for studying Cross-Phenotype Genetic Associations with Group-level Bayesian Meta-Analysis"

(First round)

by Baghfalaki, Sugier, Asgari, Truong, and Liquet

November 9, 2022

Dr. Dianne Cook
Editor
The R Journal

November 9, 2022
Dear Dr. Dianne Cook,
Dear reviewers,
We are writing to submit the revised version of the manuscript titled: "GCPBayes: An R package for studying Cross-Phenotype Genetic Associations with Group-level Bayesian Meta-Analysis" co-authored by Baghfalaki, Sugier, Asgari, Truong, and Liquet for possible publication in *The R Journal* journal.
We greatly appreciated the careful reading of our paper and the many helpful suggestions made to improve our manuscript. We think that this revised version reflects all the concerns raised by the reviewers. We give in the following our point-by-point responses to all comments. To help the review process, we showed reviewers' comments in grey color and our responses in black color.

Yours faithfully,
Taban Baghfalaki

Reply to comments of Reviewer 1

Overview

In this paper by Baghfalaki et al., the authors present the R package GCPBayes. A package for performing Group-level Bayesian Meta-Analysis for studying Cross-Phenotype Genetic Associations. The package can perform Bayesian meta-analysis using summary statistics and also individual-level data. Furthermore, it can perform meta-analysis across multiple phenotypes to detect pleiotropy at group-level and within groups (gene and SNP level, respectively). The ability to detect pleiotropy at both levels is interesting since most currently available R packages can only work at the SNP level.

Statistical Methodology review:

The methods implemented in the package for detecting group-level pleiotropy are based on multivariate spike and slap priors. The package also can apply group and within-group selection by using a hierarchical sparsity prior. The statistical methods used are sound and up to date. However, I miss some explanation of the methods in the paper's main text. Of course, the technical details of the different priors and the simulated examples fit perfectly in the appendix. However, most of the explanations in "Appendix B: Statistical inference" should be included in the main text because they are needed to understand and interpret the results from the main functions of the package.
I don't find the term "rejection of hypothesis" appropriate in the context of Bayesian statistics. Overall, it seems too stringent to set specific thresholds on lBFDR and log10BF to decide if there is pleiotropy. It would be more beneficial to the readers and users of the package to understand what these values mean and let them decide how to interpret their results. The same applies to the theta values for determining if there is pleiotropy at the group level.

Response: Thank you for these remarks about the technical details. Some of the explanations of the appendix B are added in the Guidelines section of the main text and, in the new version of the paper, Appendix B is removed. We rephrased some part of the main text concerning the way to investigate non null effects and pleiotropy effects. Also, the notions as lBFDR and log10BF more carefully are defined to help the user to interpret their results, when suggesting a convenient threshold.

Software review

Overall, the code of the different functions is clean and easy to follow. Though complete, the documentation is sparse. Some descriptions and function parameters should be described in more detail. The output of the functions should be tidier. Maybe the authors should add a print() or summary() function for each main function in the package.
Examples are provided for using each of the various functions in the package. However, the paper is hard to follow with the provided code since no random seed has been set, and results vary every time the script runs. Comments for specific code in the package: In the CS(), DS() and HS() functions there is a lot of redundant code. For example: if

```
(nchains > 1) {
  ts <- sample(1:nchains, 2)
  for (k in 1:K) {
    AA <- list(matrix(RES1[[ts[1]]]$mcmcchain$Beta[[k]]),
```

```
              matrix(RES1[[ts[2]]]$mcmcchain$Beta[[k]]))
    Tab <- data.frame(snpnames, mean(RES1[[1]]$mcmcchain$Beta[[k]]),
            sd(RES1[[1]]$mcmcchain$Beta[[k]]),
            t(quantile(RES1[[1]]$mcmcchain$Beta[[k]],

  } }
if (nchains == 1) { for (k in 1:K) {
  Tab <- data.frame(snpnames, mean(RES1[[1]]$mcmcchain$Beta[[k]]),
                    sd(RES1[[1]]$mcmcchain$Beta[[k]]),
                    t(quantile(RES1[[1]]$mcmcchain$Beta[[k]],

  colnames(Tab) <- cbind("Name of SNP", "Mean",
             "SD", "val2.5pc", "Median", "val97.5pc") Summary$Beta[[k]] <- Tab
} }
```

This can be seen in many other parts of the functions CS(), DS(), and HS(). MCMCplot() should be improved. It doesn't provide any options to configure the plots. I miss another plot function for the representation of the inference results.

Response: Thanks for the reviewer's helpful comments. First of all, we have to mention that the new version of the GCPBayes package (GCPBayes 4.0.0) is now available in R CRAN. The following points are added to the new version of the paper/package:

1. About adding descriptions and function parameters, in the new version of the package, in "Details" section of the CS, DS and HS functions, the hierarchical set-up of these functions are added. Thanks to the reviewer's recommendation, by considering these parts, the function parameters are clearer now.

2. Three new functions as summaryCS(), summaryDS() and summaryHS() are added as generic functions to produce result summaries of the results of the CS(), DS() and HS() functions, respectively. The input of these functions are the result of a call to CS(), DS() or HS() functions, respectively and the outputs are important criteria and information in the called functions.
   Here are a summary of the new functions:
   Output for summaryCS: Name of Gene, Number of SNPs, Name of SNPs, log10BF, lBFDR, theta, and Significance based on CI
   Output for summaryDS: Name of Gene, Number of SNPs, Name of SNPs, log10BF, lBFDR, theta, Significance based on CI, and Significance based on median thresholding
   Output for summaryHS: Name of Gene, Number of SNPs, Name of SNPs, Significance based on CI, and Pleiotropic effect based on median thresholding

3. We try to remove the redundant line codes and the redundant variables. Then, we updated the package and ubmitted the new package on the CRAN. The new version of the package includes the desired changes.

4. Some new options are added to MCMCplot() function (betatype, acftype, dencol, denlty, and denbg). Using the following options, a user could perform some modifications to the plot output:
   Betatype: The type of plot desired. The following values are possible: "p" for points, "l" for lines (the default), "b" for both points and lines, "c" for empty points joined by lines, "o" for overplotted points and lines, "s" and "S" for stair steps and "h" for histogram-like vertical lines, "n" does not produce any points or lines.

acftype: String giving the type of acf to be computed. Allowed values are "correlation" (the default), "covariance" or "partial".
dencol: The color for filling the density plot. The default is "white", unless density color is specified.
denlty: The line type to be used in the density plot. The value could be 0=blank, 1=solid, 2=dashed, 3=dotted, 4=dotdash, 5=longdash, and 6=twodash. The default value is "1".
denbg: The color to be used for the background of the density plot. The default is "white", unless density color is specified.

5. Finally, the random seeds are added to the examples of the paper. So, now it is possible to follow the paper results since a random seed has been now set.

Article contents

The article devotes too much space to tasks performed by other packages (vif(), corrplot() function, glm(), bglm(), ...). This section comprises four pages and should be simplified just to show the different inputs that are accepted by the package's functions. I would find it more useful to provide more insight regarding the output of the package's functions. For example, there is no explanation for the meaning and interpretation of the posterior probability of association (PPA), the logarithm of the Bayes factor (log10BF), or the local false discovery rate (locFDF) or theta. As explained before, in my opinion, the authors should rethink the guidelines section.

Response: Thank you for these remarks. We still believe that this part is important for users who have access to individual data and wish to properly use the package as GCPBayes results are highly dependent on summary statistics used as inputs, but we agree that this section could be shorten. Hence, we reduced the section "Individual level data" that is mostly devoted to external function of the package. We also added more explanations for the meaning and interpretation of the posterior probability of association (PPA), the logarithm of the Bayes factor (log10BF), or the local false discovery rate (lBFDR) and θ in the Guidelines section.

Summary of the review

This paper is very interesting and provides a complete overview of the functionalities implemented in the GCPBayes package. However, the statistical methods are poorly explained to the reader in the main text. In addition, the guidelines section seems too stringent, recommending specific thresholds with no justification, which would lead a non-technical reader to follow them blindly with the potential risks in doing so. Authors should add a section devoted to explaining how to interpret each of the values obtained by their functions. Regarding the code, the main inference functions work as expected. However, their output should be made tidier, and some parts of their inner code are redundant. The MCMCplot() function should be improved; it is very simple and offers no options to configure the plots.

Response: Thank you for giving a summary of the review. The response to these comments are given in details by the previous responses.

Reply to comments of Reviewer 2

In this paper, the authors introduced and implemented GCPBayes as an R package for studying cross-phenotype genetic associations with group-level Bayesian meta-analysis, based on their previously developed method (Baghfalaki et al. 2021). The proposed package has been implemented in order to consider and detect pleiotropy at both variable-level and group-level. The inputs of the developed functions are SNP-level summary statistics data derived from GWAS. Overall, the implemented functions are based on well-established approaches. The examples in the paper are helpful and clear with sufficient details. The authors included background as well as practical guidelines of using GCPBayes package. The methods used for producing results (in real data application/simulations) and data interpretations are justified. I do have following comments for further improving the manuscript. Since there is no line number in the submitted manuscript, I have highlighted the corresponding parts in the PDF file.

Main comments:

1. Page 4. The term "multicollinearity" is a pure statistical terminology. In (statistical) genetics literature, it refers to "linkage disequilibrium (LD)" which is a very well-known concept. There are many approaches that can properly account for SNPs in LD besides the mentioned ridge regression (for example, LD score regression). I wouldn't expect the authors to add additional literatures, but the authors are recommended to connect "linkage disequilibrium (LD)" when introducing "multicollinearity" to better represent the underlying context (genetics).

Response: Thank you for this useful comment. In the new version of the paper page 4, we have changed the mentioned part to "Though, another strategy should be considered in the case of multicollinearity that is the existence of near-linear relationships among variables of a group (Malo, Libiger, and Schork 2008). This phenomenon is widely spread in genetic data were non-random association of alleles at different loci in a given population are frequent, introducing large structures of correlation between SNPs. It is known as linkage disequilibrium (LD). This can create inaccurate estimates of the regression coefficients, and also inflate the standard errors of the regression coefficients (Saleh, Arashi, and Kibria 2019). A primary procedure for the visualization of this phenomenon is to draw pairwise scatter plots of variables or to consider the correlation matrix."

2. Following #1, in page 6, I fully agree with the authors' argument that "statistically speaking, we recommend to use individual level data to increase the power of the method by using non-diagonal covariance matrix of the effects as inputs of GCPBayes, when available". However, I do not see Example 9 as well as the arguments in page 14 informative: "In the next example, we show that if the sample size is large enough, using the diagonal covariance matrix (summary statistics level data) gives the same results as those with the non-diagonal one (individual level data)." "So, if the sample size is large enough, a user could use diagonal covariance matrix (summary statistics level data) as input and could be able to detect potential pleiotropic signals without any problem at group and/or individual level." If the true correlations between different SNPs are non-zero, then any model using only summary statistics (diagonal covariance matrix) would still be mis-specified from the statistical perspective, and the above conclusions based on specific examples should not be made. Hence, the authors are recommended to remove Example 9 from the revised manuscript (or put the entire Example 9 in the Appendix with very conservative arguments being made).

Response: Thank you for your very careful review. We have removed Example 9 in the new version of the paper.

3. Page 16. For computational time of GCPBayes, I only see that the authors "consider nine genes with different numbers of variables (SNPs)". What would be the computation cost if one wants to scan the whole genome for all 20,000 genes using GCPBayes? Most frequentist state-of-the-art scalable methods would consider a score-test based regression framework by first fitting a null model with only covariates and without any genotype data, then use the null model fit to test single SNP associations or SNP-set associations. Could the authors benchmark and report the computation cost for a genome-wide analysis in addition to the current candidate gene approach?

Response: Thanks for the reviewer's helpful comment. Actually, we are in the middle of application of GCPBayes on real datasets (in the whole genome). So, we could provide an estimation of running time for applying GCPBayes on all protein coding genes on the genome-wide scale. We have added a paragraph in the "Computational time for GCPBayes" section which explains the running time for a genome-wide analysis: "We have also applied GCPBayes on whole genome data of breast and ovarian cancers (summary statistics data). The computational time for running GCPBayes over all coding genes (number of genes: 18,244 and number of SNPs: 1 to 9,595) was about 17 days (on a PC with Intel® Core(TM) i7-1165G7 @2.80 GHz, 32 GB RAM). However, if a user applies parallelized loops with R for running GCPBayes on more than one gene at the same time (using "foreach" command), the computational time would be reduced. For instance, the computational time was about 40 hours using the same PC but running 6 genes in each loop (i.e. using 6 CPUs instead of 1)."

4. Page 7. The implemented GCPBayes package seems to depend on several other R packages (e.g. BhGLM for the bglm() function). However, I could not successfully install the BhGLM package for my environment since it was only hosted on GitHub but not hosted on CRAN. For improved usability, is it possible for the authors to incorporate the bglm() and other necessary functions from other packages within the GCPBayes package?

Response: Thanks for the reviewer's useful comment. We agree with the reviewer that the process would be easier if we incorporate other outsource functions inside our package. However, we wanted to give the opportunity to a user to use any desired function for the fitting process. Therefore, we have just used "bglm" function as an example and a user could use any other function to do that. In addition, due to a copyright issue in submission of a package to the CRAN host, we think it is better to call "bglm" function from its original source.

5. Page 18. The authors noted that "Besides, the GCPBayes package is designed for uncorrelated studies (no overlapping samples between studies)". Would the method/ package be able to handle "correlated samples in uncorrelated studies"? If not, the authors are recommended to mention this as a potential limitation as well.

Response: Thank you for your helpful comment. In the new version of the paper, the mentioned part is changed to "Besides, the GCPBayes package is designed for uncorrelated studies (no overlapping samples between studies). So, an improvement of the package for correlated studies could be considered for the future. Also, GCPBayes can deal with correlated samples in uncorrelated studies as it uses summary statistics. Though, a user should be careful that summary statistics has been calculated using methods taking into account for such correlations between samples."

Other comments:

6. Page 1. The authors are recommended to include the following cross-phenotype association detection method (Liu and Lin, 2018) for completeness.

Response: The mentioned reference has been added to the mentioned page of the new version of the paper.

7. Page 2. The authors are recommended to include the following paper to better support this following argument: By incorporating prior biological information in GWAS such as group structure information (gene or pathway), our approaches could uncover new pleiotropic signals (Baghfalaki et al., 2021; Li et al., 2020).

Response: The mentioned reference has been added to the mentioned page of the new version of the paper.

8. Several typos in spelling/typesetting (highlighted) could be fixed.

Response: Thanks for the reviewer's careful comments. We have fixed the mentioned typos in the new version of the paper.

References

Baghfalaki, T., Sugier, P. E., Truong, T., Pettitt, A. N., Mengersen, K., & Liquet, B. (2021). Bayesian meta-analysis models for cross cancer genomic investigation of pleiotropic effects using group structure. Statistics in Medicine, 40(6), 1498-1518.

Liu, Z., & Lin, X. (2018). Multiple phenotype association tests using summary statistics in genome-wide association studies. Biometrics, 74(1), 165-175.

Li, X., Li, Z., Zhou, H., Gaynor, S. M., Liu, Y., Chen, H., ... & Lin, X. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nature genetics, 52(9), 969-983.