Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw349
Advance Access Publication Date: 13 June 2016
Applications Note

OXFORD

Gene expression

GSA-Lightning: ultra-fast permutation-based gene set analysis

Billy Heung Wing Chang[1,2] and Weidong Tian[1,*]

[1]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Department of Biostatistics and Computational Biology, School of Life Sciences, Fudan University, Shanghai 200436, People's Republic of China and [2]Jockey Club School of Public Health and Primary Care, the Chinese University of Hong Kong, Shatin, Hong Kong SAR

*To whom correspondence should be addressed.
Associate Editor: Janet Kelso

Received on October 27, 2015; revised on May 2, 2016; accepted on May 30, 2016

Abstract

Summary: The computational speed of many gene set analysis methods can be slow due to the computationally demanding permutation step. This article introduces GSA-Lightning, a fast implementation of permutation-based gene set analysis. GSA-Lightning achieves significant speedup compared with existing methods, particularly when the number of gene sets and permutations are large.
Availability and implementation: The GSA-Lightning R package is available on Github at https://github.com/billyhw/GSALightning and on R Bioconductor. The package also contains a comprehensive user's guide with a step-by-step tutorial vignette.
Contact: weidong.tian@fudan.edu.cn
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene set analysis embodies a range of techniques for analyzing the functional and biological properties of differentially-expressed genes. One common type of gene set analysis method, termed *self-contained* method (Goeman and Bühlmann, 2007), tests whether genes within a predefined gene-set are collectively differentially expressed across two different experimental conditions. The gene set's significance is typically assessed through permutation testing. Due to the computationally-demanding permutation tests, the speed of permutation-based gene set analysis methods is often compromised. There is a need to address this computation issue, particularly when performing large-scale analysis. As an example, for a genome-wide expression analysis of target genes of regulatory elements (Yao *et al.*, 2015), the number of gene sets, each set being the target genes of a regulatory element, can reach tens of thousands. If there are 30 000 gene sets, the significance level of 0.05 becomes 1.67×10^{-6} after Bonferroni correction. One million permutations are now required for accurate *P*-values estimation. Most existing methods are computationally infeasible in this case, due to the large number of gene sets and number of permutations required. This article introduces GSA-Lightning, a fast and memory-efficient implementation of permutation-based gene set analysis. Through a speed comparison of GSA-Lightning and other existing permutation-based gene set analysis methods, we demonstrate that GSA-Lightning is computationally much more efficient, particularly when the number of gene sets and permutations are large.

2 Implementation

GSA-Lightning is based on the GSA method (Efron and Tibshirani, 2007). GSA-Lightning computes the student-T statistics for each individual gene, and provides the 'maxmean', 'mean' and 'absolute-mean' options for combining the individual gene statistics into gene set statistics. Similar to GSA, GSA-Lightning provides the restandardization procedure and further supports independent and paired two-sample tests. GSA-Lightning takes advantage of R's efficiency in matrix computations, and achieves substantial speedup by calculating the gene set statistics and their permutation distributions through matrix computations. Indeed, the student-T statistics mainly depends only on the mean and standard deviation of the

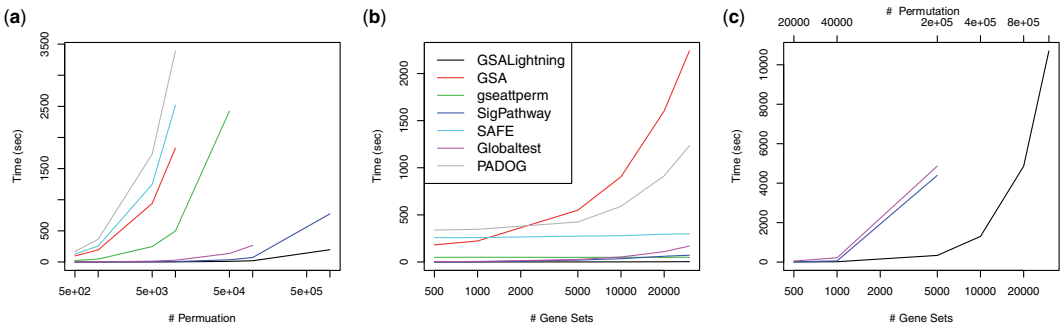

Fig. 1. Speed trial results. **(a)** Varying number of permutations when the number of gene sets is fixed at 500. **(b)** Varying number of gene sets when the number of permutations is fixed at 1000. **(c)** Varying number of gene sets and number of permutations. See body text for details. Note the *x*-axis of all three plots are on the log-scale

expression measurements, and the gene set statistics are functions of the mean of the individual statistics. All these quantities can be computed efficiently using matrix products. Also, it is actually possible to perform the permutations using matrix products. The online Supplementary Materials provide details on how this can be done. Speed and memory efficiency are further increased by using sparse matrix computations whenever possible. When the number of permutations is large, GSA-Lightning breaks the permutations into batches. Each batch is handled separately and then combined to obtain the final results. The online Supplementary Materials provide further details, and also contain a justification of the equivalence between GSA-Lightning and GSA by comparing the two methods' program outputs. This justifies that GSA-Lightning does not sacrifice GSA's accuracy in order to achieve the speedup.

3 Results

Only methods with R implementations were compared to ensure hardware-independent results. Focusing on permutation-based two-sample testing, non-permutation-based methods were not compared. GSA-Lightning (with the 'maxmean' option and restandardization) was compared with six other methods using their default options unless stated otherwise. These methods were GSA (with the 'maxmean' option and restandardization) (Efron and Tibshirani, 2007), gseattperm from the R Bioconductor package Category, the self-contained version of SigPathway (Tian *et al.*, 2005), SAFE (Barry *et al.*, 2005), Globaltest (Goeman *et al.*, 2004) and PADOG (Tarca *et al.*, 2012). All analyses were performed using a MacBook Pro with a 2.3 GHz Intel Core i7 processor and 16 GB RAM.

The gene sets used were the target genes of 36 381 distal regulatory elements from (Lu *et al.*, 2013) with three or more target genes. The breast cancer data set from The Cancer Genome Atlas (The Cancer Genome Atlas Research Network, 2013) downloaded using the ELMER R Bioconductor package (Yao *et al.*, 2015) was used as the gene expression data. The data set contained gene expression measures for 114 controls and 1104 patients. After removing genes without matching gene symbols and/or with 0 sample variance, 20 038 genes remained for further analysis.

Three speed trials were performed to examine the speed of the various methods. First, the number of gene sets was fixed at 500, and the number of permutations varied from 500 to 10^6. Second, the number of permutations was fixed at 1000, and the number of gene sets varied from 500 to 30 000. The third trial mimics a more realistic scenario: the numbers of gene sets were the same as in the second trial, but the number of permutations varied according to $2 \times$ (no. genesets)/0.05. This number of permutations ensured accurate *P*-value estimation even after multiple correction at significance level 0.05. Only GSA-Lightning, SigPathway, and Globaltest were tested in this trial as the other methods were too slow here. Figure 1 presents the results. GSA, SAFE and PADOG were not run for number of permutations $>10\,000$, and gseattperm was not run for $>50\,000$ permutations since these methods were too slow. Also, SigPathway and Globaltest encountered memory issues when the numbers of permutations were respectively $>10^6$ and 10^5 in the first trial, and when the number of gene sets was >5000 in the third trial, and hence were not performed there. GSA-Lightning was faster than all other methods. In particular, Figure 1c suggests that GSA-Lightning was the only method that could analyze large amounts of gene sets with a large number of permutations within reasonable time. In the experiments above, the maximum RAM usage for GSA-Lightning was 4 GB.

A speed comparison between the different gene set statistics, with or without restandardization, for GSA-Lightning and GSA is provided in the online Supplementary Materials. Note that all the methods compared earlier are based on different test statistics. The results and sensitivity of these methods will therefore be different, depending on the data set and the methods' assumptions. Nevertheless, a recent review considered GSA with the 'maxmean' statistics one of the better gene set analysis methods (Tarca *et al.*, 2013). By virtue of GSA's strengths, a fast version of GSA is worth developing. Also, a detailed speed comparison would involve comparing separately the speed for calculating the individual gene statistics, the gene set statistics, and the permutation. Such comparison is beyond the scope of this article. Still, for large-scale gene set analysis, the results above suggest that GSA-Lightning is the only possible option among the methods compared. Altogether, GSA-Lightning is a tool capable of performing large-scale gene set analysis with statistical power guarantee.

Acknowledgements

We thank Xinran Dong for providing earlier speed benchmarking results, and the anonymous reviewers for helpful comments on the article and the R package.

Funding

National Natural Science Foundation of China (31471245, 91231116, 31071113, 30971643, 8117078); the National Basic Research Program of China (2012CB316505); Chinese Hi-tech Research and Development Project (863) (2014AA021104); the Specialized Research Fund for the Doctoral

Program of Higher Education of China (20120071110018); the Innovation Program of Shanghai Municipal Education Commission (13ZZ006); Shanghai Municipal Science and Technology Commission (12431900100).

Conflict of Interest: none declared.

References

Barry,W. *et al.* (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Goeman,J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Goeman,J. *et al.* (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Lu,Y. *et al.* (2013) Combining Hi-C data with phylogenetic correlation to predict the target genes of distal regulatory elements in human genome. *Nucleic Acid Res.*, **41**, 10391–10402.

Tarca,A. *et al.* (2012) Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, **13**,

Tarca,A. *et al.* (2013) A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS One*, **8**, e79217.

The Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Tian,L. *et al.* (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci. USA*, **102**, 13544–13549.

Yao,L. *et al.* (2015) Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.*, **16**, 105.