Bioinformatics, 35(11), 2019, 1958–1959
doi: 10.1093/bioinformatics/bty892
Advance Access Publication Date: 22 October 2018
Applications Note

Genome analysis

methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing

Xu Ren and Pei Fen Kuan*

Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on July 17, 2018; revised on August 31, 2018; editorial decision on October 16, 2018; accepted on October 19, 2018

Abstract

Motivation: An important downstream analysis following differential expression from RNA sequencing (RNA-Seq) or DNA methylation analysis is the gene set testing to relate significant genes or CpGs to known biological properties. However, the traditional gene set testing approaches result in biased *P*-values due to the difference in gene length. Existing methods accounting for length bias were primarily developed for RNA-Seq data. For DNA methylation data profiled using the Illumina arrays, separate methods adjusting for the number of CpGs instead of gene length are necessary.

Results: We developed methylGSA, a Bioconductor package for gene set testing in DNA methylation data. Our accompanying Shiny app provides an interactive way of accessing functions and visualizing the results in methylGSA package.

Availability and implementation: methylGSA is available at Bioconductor repository: https://bioconductor.org/packages/methylGSA and Shiny app is available at: http://www.ams.sunysb.edu/%7epfkuan/softwares.html#methylGSA.

Contact: peifen.kuan@stonybrook.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput experiments including RNA-Seq and DNA methylation assays are widely used to identify differentially expressed (DE) genes or CpGs. A popular approach to decipher the list of DE genes or CpGs is via the gene set testing, which can be divided into two types, namely over-representation analysis (ORA) and functional class scoring (FCS) (Khatri et al., 2012). ORA works by first dividing the genes to DE and non-DE genes based on a pre-specified P-value threshold. The level of over-representation is quantified by hypergeometric, Fisher exact or chi-square tests. ORA can lead to information loss because the procedure ignores the genes marginally above the P-value threshold, which may also provide valuable information. In addition, genes within the same gene set are usually correlated but this feature is not considered in ORA. FCS is developed to overcome the drawbacks of ORA by first computing a DE test P-value for each gene. Instead of using only DE genes, FCS ranks all genes by the DE test P-values and utilizes the entire gene list to compute a gene set level P-value via permutation. Thus, FCS is able to detect small but coordinated change in gene expression and incorporates the dependence structure among the genes in a gene set.

Traditional gene set testing approaches have been shown to yield biased results due to gene length difference in both RNA-Seq and DNA methylation data (Geeleher *et al.*, 2013; Young *et al.*, 2012). In particular, longer genes are more likely to be identified as DE and gene sets containing a higher fraction of long genes tend to be declared significant. To adjust for length bias in RNA-Seq data, Young *et al.* (2012) proposed GOSeq based on weighted resampling and Wallenius non-central hypergeometric approximation for ORA; whereas Gao *et al.* (2011) proposed a Wald type test statistics, adjusting for length bias in gene ranking for FCS. On the other hand, Mi *et al.* (2012) introduced GOglm for FCS by incorporating gene length as a covariate in a logistic regression model; whereas

Ren *et al.* (2017) modified the standardization step of GSA (Efron and Tibshirani, 2007) by including gene length adjustment in SeqGSA. For DNA methylation data, Phipson *et al.* (2015) proposed missMethyl by modifying GOSeq; whereas Li *et al.* (2017) proposed MethylSet using logistic kernel machine model followed by logistic regression with gene length as a covariate. Both missMethyl and MethylSet were developed for ORA.

For DNA methylation data measured using the Illumina's 450K or EPIC arrays, the above mentioned methods for ORA may lead to inaccurate adjustment because the number of CpGs profiled and gene length is not of a one-to-one correspondence. As shown in Supplementary Figure S1, the number of CpGs vary among the genes of similar gene length. In addition, to date there is no length bias adjustment method for FCS in DNA methylation data. In addition, EWAS of DNA methylation gives rise to multiple CpG association *P*-values per gene. Therefore, methods adjusting for the number of CpGs instead of gene length, as well as methods for FCS are required. We introduce two methods, namely methylRRA and methylglm to address these limitations for DNA methylation data.

2 Materials and methods

2.1 methylRRA

Robust rank aggregation (RRA) is a parameter free model first introduced by Kolde *et al.* (2012) for data integration in meta-analysis. We adapt the RRA approach to adjust for number of CpGs in DNA methylation gene set testing as follows:

For gene i, let $P_1, P_2, \ldots, P_n$ be the P-values of individual CpGs. Under the null hypothesis, $P_1, P_2, \ldots, P_n \overset{iid}{\sim} Unif[0,1]$. Let $P_{(1)}, P_{(2)}, \ldots, P_{(n)}$ be the order statistics and $\rho = \min\{\Pr(P_{(1)} < P_{(1)\text{obs}}), \ldots, \Pr(P_{(n)} < P_{(n)\text{obs}})\}$. Similar to Kolde *et al.* (2012), ρ is converted into a P-value (P) by Bonferroni correction. This method automatically adjusts for number of CpGs and can be utilized for both the ORA and FCS. In ORA, the P's are further corrected for multiple testing via the false discover rate (FDR) approach (Benjamini and Hochberg, 1995) for detecting DE genes, followed by a hypergeometric test. In FCS, we convert P's into z-scores. We then apply the Gene Set Enrichment Analysis (Subramanian *et al.*, 2005) on the gene list ranked by the z-scores.

2.2 methylglm

methylglm is an extension of GOglm (Mi *et al.*, 2012) originally developed for RNA-Seq. Instead of adjusting for gene length, we adjust for the number of CpGs in the logistic regression model: $\log it(\pi_i) = \beta_0 + \beta_1 x_i + \beta_2 c_i$, where c_i denotes the number of CpGs associated with gene i, x_i denotes negative log transform of the minimum of the c_i P-values from DE analysis and π_i denotes the probability that gene i is in the gene set. The hypothesis test of interest is: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. A significant test result indicates gene set is enriched after adjusting for the number of CpGs. methylglm is an FCS method because it utilizes the information from the entire gene list to fit the logistic regression model without the need to preselect DE genes.

Both methylRRA and methylglm are implemented as a Bioconductor package methylGSA along with a shiny app (Fig. 1) which can be used for analyzing DNA methylation from the Illumina's 450K or EPIC arrays. The software also allows users to consider different genomic context including gene body and promoter. Simulation and case studies illustrating the superior performance of our proposed methods compared to existing methods are provided in Supplementary Material.

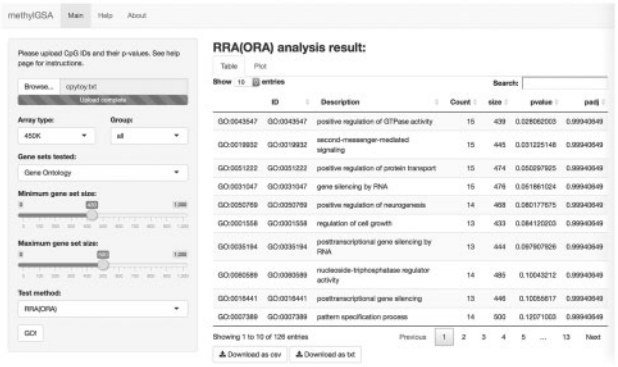

Fig. 1. Screen shot of the methylGSA shiny app

3 Discussion

methylGSA is a flexible and user friendly software for gene set testing with length bias adjustment in DNA methylation data. It allows users to identify enriched or over-represented gene sets or pathways from the Gene Ontology, KEGG and Reactome databases.

Funding

This work was supported in part by the CDC/NIOSH award U01 OH011478-01.

Conflict of Interest: none declared.

References

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

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

Gao,L. *et al.* (2011) Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics*, **27**, 662–669.

Geeleher,P. *et al.* (2013) Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics*, **29**, 1851–1857.

Khatri,P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.

Kolde,R. *et al.* (2012) Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, **28**, 573–580.

Li,S. *et al.* (2017) Correcting length-bias in gene set analysis for DNA methylation data. *Stat. Interface*, **10**, 279–289.

Mi,G. *et al.* (2012) Length bias correction in gene ontology enrichment analysis using logistic regression. *PLoS One*, **7**, e46128.

Phipson,B. *et al.* (2015) missMethyl: an R package for analyzing data from Illumina HumanMethylation450 platform. *Bioinformatics*, **32**, 286–288.

Ren,X. *et al.* (2017) Gene set analysis controlling for length bias in RNA-seq experiments. *BioData Min.*, **10**, 5.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Young,M.D. *et al.* (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.*, **11**, R14.