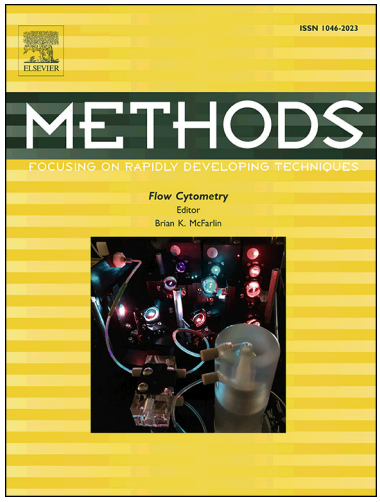
Accepted Manuscript



POST: a framework for set-based association analysis in high-dimensional data

Xueyuan Cao, E. Olusegun George, Mingjuan Wang, Dale B. Armstrong, Cheng Cheng, Susana Raimondi, Jeffrey E. Rubnitz, James R. Downing, Mondira Kundu, Stanley B. Pounds

PII: S1046-2023(17)30497-8
DOI: https://doi.org/10.1016/j.ymeth.2018.05.011
Reference: YMETH 4484

To appear in: *Methods*

Received Date: 9 January 2018
Revised Date: 11 May 2018
Accepted Date: 13 May 2018

Please cite this article as: X. Cao, E.O. George, M. Wang, D.B. Armstrong, C. Cheng, S. Raimondi, J.E. Rubnitz, J.R. Downing, M. Kundu, S.B. Pounds, POST: a framework for set-based association analysis in high-dimensional data, *Methods* (2018), doi: https://doi.org/10.1016/j.ymeth.2018.05.011

POST: a framework for set-based association analysis in high-dimensional data

1st Xueyuan Cao
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, USA
Department of Acute and Tertiary Care
University of Tennessee Health Science Center
Memphis, USA
xcao12@uthsc.edu

2nd E. Olusegun George
Department of Mathematics
University of Memphis
Memphis, USA
eogeorge@memphis.edu

3rd Mingjuan Wang
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, USA
mingjuan.wang@stjude.org

4th Dale B. Armstrong
Department of Mathematics
University of Memphis
Memphis, USA
ddbowman@memphis.edu

5th Cheng Cheng
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, USA
cheng.cheng@stjude.org

6th Susana Raimondi
Department of Pathology
St. Jude Children's Research Hospital
Memphis, USA
susana.raimondi@stjude.org

7th Jeffrey E. Rubnitz
Department of Oncology
St. Jude Children's Research Hospital
Memphis, USA
jeffrey.rubnitz@stjude.org

8th James R. Downing
Department of Pathology
St. Jude Children's Research Hospital
Memphis, USA
james.downing@stjude.org

9th Mondira Kundu
Department of Pathology
St. Jude Children's Research Hospital
Memphis, USA
mondira.kundu@stjude.org

10th Stanley B. Pounds
Department of Biostatistics
St. Jude Children's Research Hospital
Memphis, USA
stanley.pounds@stjude.org

***Abstract*—Evaluating the differential expression of a set of genes belonging to a common biological process or ontology has proven to be a very useful tool for biological discovery. However, existing gene-set association methods are limited to applications that evaluate differential expression across $k \geq 2$ treatment groups or biological categories. This limitation precludes researchers from most effectively evaluating the association with other phenotypes that may be more clinically meaningful, such as quantitative variables or censored survival time variables. Projection onto the Orthogonal Space Testing (POST) is proposed as a general procedure that can robustly evaluate the association of a gene-set with several different types of phenotypic data (categorical, ordinal, continuous, or censored). For each gene-set, POST transforms the gene profiles into a set of eigenvectors and then uses statistical modeling to compute a set of z-statistics that measure the association of each eigenvector with the phenotype. The overall gene-set statistic is the sum of squared z-statistics weighted by the corresponding eigenvalues. Finally, bootstrapping is used to compute a p-value. POST may evaluate associations with or without adjustment for covariates. In simulation studies, it is shown that the performance of POST in evaluating the association with a categorical phenotype is similar to or exceeds that of existing methods. In evaluating the association of 875 biological processes with the time to relapse of pediatric acute myeloid leukemia, POST identified the well-known oncogenic WNT signaling pathway as its top hit. These results indicate that POST can be a very useful tool for evaluating the association of a gene-set with a variety of different phenotypes. We have developed an R package named POST which is freely available in Bioconductor.**

***Index Terms*—Gene profiling, Gene network, Orthogonal projection, Data integration**

This work was supported by American Lebanese Syrian Associated Charities (ALSAC)

I. Introduction

Microarrays and high-throughput sequencing have empowered investigators to simultaneously measure the expressions or other genomic features of thousands of genes in biological specimens. This results in data matrix with thousands to million rows, a form of high dimensional data in which number of features greatly exceeds the number of observations. Subsequently, statistical analysis is used to test the association of the expression of individual genes with a phenotype. As thousands to million of tests are performed simultaneously, multiple testing should be addressed before declaring a list