Bioinformatics, 33(12), 2017, 1892–1894
doi: 10.1093/bioinformatics/btx058
Advance Access Publication Date: 8 February 2017
Applications Note

OXFORD

Gene expression

PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF

Genevieve L. Stein-O'Brien[1,2,*], Jacob L. Carey[3], Wai Shing Lee[3], Michael Considine[3], Alexander V. Favorov[3,4,5], Emily Flam[6], Theresa Guo[6], Sijia Li[6], Luigi Marchionni[6], Thomas Sherman[7], Shawn Sivy[7], Daria A. Gaykalova[6], Ronald D. McKay[2], Michael F. Ochs[7], Carlo Colantuoni[8,9,*] and Elana J. Fertig[3,*]

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA, [2]Lieber Institute for Brain Development, Baltimore, MD, USA, [3]Department of Oncology and Division of Biostatistics and Bioinformatics, Johns Hopkins School of Medicine, Baltimore, MD, USA, [4]Vavilov Institute of General Genetics, Moscow, Russia, [5]Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia, [6]Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins School of Medicine, Baltimore, MD, USA, [7]Department of Mathematics and Statistics, The College of New Jersey, Ewing Township, NJ, USA, [8]Department of Neurology and Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA and [9]Institute for Genome Sciences, University of Maryland School of Medicine

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on October 28, 2016; revised on January 6, 2017; editorial decision on January 26, 2017; accepted on January 27, 2017

Abstract

Summary: Non-negative Matrix Factorization (NMF) algorithms associate gene expression with biological processes (e.g. time-course dynamics or disease subtypes). Compared with univariate associations, the relative weights of NMF solutions can obscure biomarkers. Therefore, we developed a novel patternMarkers statistic to extract genes for biological validation and enhanced visualization of NMF results. Finding novel and unbiased gene markers with patternMarkers requires whole-genome data. Therefore, we also developed Genome-Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS), the first robust whole genome Bayesian NMF using the sparse, MCMC algorithm, CoGAPS. Additionally, a manual version of the GWCoGAPS algorithm contains analytic and visualization tools including patternMatcher, a Shiny web application. The decomposition in the manual pipeline can be replaced with any NMF algorithm, for further generalization of the software. Using these tools, we find granular brain-region and cell-type specific signatures with corresponding biomarkers in GTEx data, illustrating GWCoGAPS and patternMarkers ascertainment of data-driven biomarkers from whole-genome data.

Availability and Implementation: PatternMarkers & GWCoGAPS are in the CoGAPS Bioconductor package (3.5) under the GPL license.

Contact: gsteinobrien@jhmi.edu or ccolantu@jhmi.edu or ejfertig@jhmi.edu

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

1 Introduction

Numerous high-throughput studies link gene expression changes to biological processes (BPs) including regulatory networks and the cell signaling processes. Previously shown effective at deconvoluting multiplexed regulation and gene reuse in BPs (Kossenkov and Ochs, 2009; Ochs and Fertig, 2012; Trendafilov and Unkel, 2011), NMF algorithms have identified genes associated with yeast cell cycle and metabolism, cancer subtypes and perturbations to cellular signaling in cancer (Brunet *et al.*, 2004; Fertig *et al.*, 2012, 2013; Kossenkov and Ochs, 2009; Li and Ngom, 2013; Mejía-Roa *et al.*, 2008; Ochs *et al.*, 2009; Wang *et al.*, 2006). However, the continuous and interdependent nature of many NMF results can make biological inference challenging especially when searching for biomarkers or genetic drivers. A method to obtaining genes that uniquely identify NMF solutions would eliminate these challenges.

Here, we develop patternMarkers, a statistic to take the relative gene weights output from NMF algorithms and to return only those genes that are strongly associated with a particular pattern or with a linear combination of patterns. Identifying unbiased biomarkers using patternMarkers requires genome-wide transcriptional data. To maximize the potential for novel marker detection, we set out to expand the O(1000) gene limit, which is typical to achieve convergence in NMF, to the O(10 000) genes comprising the entire human transcriptome. Currently, NMF methods are highly dependent upon the genes selected or compaction methods to limit the size of the data matrices used for analysis (de Campos *et al.*, 2013). Therefore, we developed GWCoGAPS, a whole genome implementation of CoGAPS (Fertig et al, 2010), a Markov chain Monte Carlo (MCMC) NMF that encodes sparsity in the decomposed matrices with an atomic prior (Sibisi and Skilling, 1997). Previously, we demonstrated that CoGAPS analysis of datasets containing representative subsets of the genes converge with similar patterns. These patterns can then be fixed to a consensus pattern across the datasets to provide a robust whole-genome NMF, without the prohibitively large computational cost of NMF factorization of a single matrix containing the entire genome. GWCoGAPS takes advantage of parallel computing to massively cut runtime and ensure genome-wide convergence. We also include a Shiny web application, patternMatcher, to compare patterns across parallel runs to increase robustness and interpretability of the resulting patterns. Using patternMarkers with GWCoGAPS to analyze tissues from the Genotype-Tissue Expression Project (GTEx Consortium, 2015), we parsed patterns of expression specific to brain regions and cell types to demonstrate the power of these algorithms for biomarker discovery.

2 Materials and methods

NMF decomposes a data matrix of D with N genes as rows and M samples as columns, into two matrices, as **D ~ AP**. The pattern matrix P has rows associated with BPs in samples and the amplitude matrix A has columns indicating the relative association of a given gene, where the total number of BPs (k) is an input parameter. CoGAPS is a Bayesian NMF that incorporates both non-negativity and sparsity in A and P as described in (Fertig *et al.*, 2010). Both patternMarkers and GWCoGAPS are in the CoGAPS Bioconductor package as of version 3.5 and are generalized for other NMF algorithms.

The patternMarkers statistic (s_{ij}) scores the association of the *i*th gene's values in the amplitude matrix (A_i) with the *j*th pattern or linear combination of patterns by computing

$$s_{ij}(\overline{w}_j) = \sqrt{\sum_k \left(\frac{A_{ik}}{\max A_i} - \overline{w}_{jk}\right)^2} \quad (1)$$

where *i* indices all the genes in the original data matrix, k indices all the patterns in the NMF solution, and $\overline{w}_j$ is a vector of components specifying the *j*th linear combination of patterns that is constrained to sum to 1, and *j* indices the total number of linear combinations for which patternMarkers statistics are computed. The default setting for Eq. (1) sets $j = \{1, \ldots, k\}$, such that $\overline{w}$ is a set containing a unit vector for each pattern and $s_{ij}(\overline{w}_j)$ is an l_2 norm indicating the exclusivity of the contribution of gene *i* to the pattern *j* and the corresponding BP. Scaling by the maximum value of each gene in the NMF solution ($\max A_i$) decouples the effect of overall gene expression level without impacting the quality of the factorization. Genes are ranked by increasing $s_{ij}(\overline{w}_j)$ such that the higher the rank of the gene, the less it is associated with the considered pattern. Users can output a list of data frames containing the scores and ranks for every gene using the 'All' option of the 'threshold' argument. Alternatively, unique gene sets can be generated by either subsetting each gene by its lowest ranking $s_{ij}(\overline{w}_j)$ In the case where $j > 1$, the ranked list for each pattern can also be thresholded by the highest value for which $s_i(\overline{w}_j)$ is the lowest.

The GWCoGAPS function automates and parallelizes the whole-genome CoGAPS analysis from Fertig *et al.* (2013) in a single R function. GWCoGAPS has three parameters: the number of sets for partitioning the whole genome data, the seed for each Markov Chain, and the method for determining the consensus patterns. A new modification to CoGAPS, setting the seed both ensures that each set of genes is run with a different set of random numbers and that runs on any dataset are reproducible. A default pattern matching function is provided along with a Shiny-based web application patternMatcher for recompiling the parallelized results (Supplementary Fig. S1A). Additional runtime options, input and manual implementations are described in the GWCoGAPS vignette.

RPKM RNAseq data for the seven samples with most brain regions was downloaded from dbGaP. GWCoGAPS was run for a range of k patterns with k = 10 selected and uncertainty as 10% of the data (Fertig *et al.*, 2013). The code to reproduce this analysis and the GWCoGAPS results are in Supplementary Files S3 and S4.

3 Results

We apply GWCoGAPS to analyze patterns related to brain regions for different individuals in GTEx. The GWCoGAPS solutions for the initial parallel runs of the patterns are used to illustrate the strong association between patterns identified from the subsets using patternMatcher (Supplementary Fig. S1A). The first pattern highlights GWCoGAPS' ability to deconvolute tissue specific signatures (Supplementary Fig. S1B). This pattern uniquely identifies the cerebellum, determined to be the most distinct region by the consortium (GTEx Consortium, 2015). GTEx found that strong individual specific effects increase with tissue relatedness as illustrated by their inability to achieve tissue specific clusters of the different brain regions by expression alone (GTEx Consortium, 2015; Melé *et al.*, 2015). By allowing for gene reuse across different patterns, GWCoGAPS is able to overcome these effects to isolate the cerebellums signature as confirmed by gene set enrichment (Subramanian *et al.*, 2005) in cerebellum development and morphogensis (GO:0021549 and GO:0021587 FWER *P*-value < 1.0E-03 and 2.6E-03, respectively, described in Supplementary File S5) on these patternMarkers scores.

The second pattern illustrates patternMarkers' power as inference is difficult from the GWCoGAPS result alone (Supplementary Fig. S1B). This pattern depicts subpopulations of cells in multiple brain regions derived from common pallium precursors. Progeny of the pallium are specified by transcription factors TBr1 and Emx1 (Remedios *et al.*, 2007) ranked second and fourth by the patternMarkers statistic. Gene set analysis on these patternMarkers scores confirms enrichment for pallium development (GO:0021543 FWER P-value $<$ 1.0E-03, Supplementary File S5).

Deconvolution of cell type and tissue specific signatures from aggregate data represent a major technical challenge. We have illustrated the unique ability of GWCoGAPS, the first whole genome Bayesian NMF, to accomplish this. The manual pipeline and Shiny App, patternMatcher, also expanded this methodology to a variety of NMF techniques. Finally, the patternMarkers statistic derives gene sets uniquely representative of BPs from the continuous weights of NMF solutions. Together, patternMarkers and GWCoGAPS find data-driven biomarkers and genetic drivers in whole genome transcriptomic data.

Funding

This work was supported by the National Institutes of Health [NCI R01CA177669 and K25CA141053 to E.J.F., NLM R01LM011000 to M.F.O. and NCI P30 CA006973], the Lieber Institute for Brain Development, and the Cleveland Foundation and Johns Hopkins University Discovery Awards to E.J.F.

Conflict of Interest: none declared.

References

Brunet,J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 4164–4169.

de Campos,C.P. *et al.* (2013) Discovering subgroups of patients from DNA copy number data using NMF on compacted matrices. *PLoS ONE*, **8**, e79720.

Fertig,E.J. *et al.* (2010) CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics*, **26**, 2792–2793.

Fertig,E.J. *et al.* (2012) Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. *BMC Genomics*, **13**, 160.

Fertig,E.J. *et al.* (2013) Preferential activation of the Hedgehog pathway by epigenetic modulations in HPV negative HNSCC identified with meta-pathway analysis. *PLoS ONE*, **8**, e78127.

GTEx Consortium. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.

Kossenkov,A.V. and Ochs,M.F. (2009) Chapter 3 Matrix factorization for recovery of biological processes from microarray data. In: *Methods in Enzymology*. Elsevier, San Diego, CA, pp. 59–77.

Li,Y. and Ngom,A. (2013) The non-negative matrix factorization toolbox for biological data mining. *Source Code Biol. Med.*, **8**, 10.

Mejía-Roa,E. *et al.* (2008) bioNMF: a web-based tool for nonnegative matrix factorization in biology. *Nucl. Acids Res.*, **36**, W523–W528.

Melé,M. *et al.* (2015) The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

Ochs,M.F. and Fertig,E.J. (2012) Matrix factorization for transcriptional regulatory network inference. In *IEEE Symp Comput Intell Bioinforma Comput Biol Proc.*, pp. 387–396.

Ochs,M.F. *et al.* (2009) Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res.*, **69**, 9125–9132.

Remedios,R. *et al.* (2007) A stream of cells migrating from the caudal telencephalon reveals a link between the amygdala and neocortex. *Nat. Neurosci.*, **10**, 1141–1150.

Sibisi,S. and Skilling,J. (1997) Prior distributions on measure space. *J. R. Stat. Soc. Ser. B (Statistical Methodology)*, **59**, 217–235.

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

Trendafilov,N.T. and Unkel,S. (2011) Exploratory factor analysis of data matrices with more variables than observations. *J. Comput. Graph. Stat.*, **20**, 874–891.

Wang,G. *et al.* (2006) LS-NMF: A modified non-negative matrix factorization algorithm utilizing uncertainty estimates. *BMC Bioinformatics*, **7**, 175.