PatternMarkers and Genome-Wide CoGAPS in Analysis in Parallel Sets (GWCoGAPS) for data-driven detection of novel biomarkers via whole transcriptome Non-negative matrix factorization (NMF)

Genevieve Stein-O'Brien[1,2], Jacob Carey[1], Wai-shing Lee[1], Michael Considine[1], Alexander Favorov[1,3,4], Emily Flam[1], Theresa Guo[1], Lucy Li[1], Luigi Marchionni[1], Thomas Sherman[1], Shawn Sivy[5], Daria Gaykalova[1], Ronald McKay[2], Michael Ochs[5], Carlo Colantuoni[1,2]*, & Elana Fertig[1]*
[1]Johns Hopkins Medical Institution, Baltimore, MD, USA
[2]Lieber Institute for Brain Development, Baltimore, MD USA
[3]Vavilov Institute of General Genetics, Moscow, Russia
[4]Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, Russia
[5]The College of New Jersey, Ewing Township, NJ USA

Abstract
Summary: NMF algorithms associate gene expression changes with biological processes (e.g., time-course dynamics or disease subtypes). Compared with univariate associations, the relative weights of NMF solutions can obscure biomarkers identification. Therefore, we developed a novel PatternMarkers statistic to extract unique genes for biological validation and enhanced visualization of NMF results. Finding novel and unbiased gene markers with PatternMarkers requires whole-genome data. However, NMF algorithms typically do not converge for the tens of thousands of genes in genome-wide profiling. Therefore, we also developed GWCoGAPS, the first robust Bayesian NMF technique for whole genome transcriptomics using the sparse, MCMC algorithm, CoGAPS. This software contains additional analytic and visualization tools including a Shiny web application, patternMatcher, which are generalized for any NMF. Using these tools, we find granular brain-region and cell-type specific signatures with corresponding biomarkers in GTex data, illustrating GWCoGAPS and patternMarkers unique ability to detect data-driven biomarkers from whole genome data.
Availability: PatternMarkers and GWCoGAPS are in the CoGAPS Bioconductor package as of version 3.5 under the GPL license.
Contact: CColantu@jhmi.edu; ejfertig@jhmi.edu
Supplementary information: Supplementary data is 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 (Trendafilov and Unkel, 2011; Kossenkov and Ochs, 2009; Ochs and Fertig, 2012), NMF algorithms have identified genes associated with yeast cell cycle and metabolism, cancer subtypes, and perturbations to cellular signaling in cancer(Li and Ngom, 2013; Brunet *et al.*, 2004; Mejía-Roa *et al.*, 2008; Fertig *et al.*, 2012; Ochs *et al.*, 2009; Fertig *et al.*, 2013; Kossenkov and Ochs, 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(1,000) 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 twelve different brain regions from seven post-mortem individuals from the Genotype-Tissue Expression Project (Consortium *et al.*, 2015), we concurrently parsed patterns of expression specific to brain regions and cell types to demonstrate the power of these algorithms for biomarker discovery.

2 **Methods**

NMF and CoGAPS

NMF decomposes a data matrix of **D** with N genes as rows and M samples as columns, into two matrices, the pattern matrix **P** with rows associated with BPs in samples and the amplitude matrix **A** with columns indicating the relative association of a given gene in each BP. CoGAPS is a Bayesian NMF that incorporates both non-negativity and sparsity in **A** and **P** as described in (Fertig *et al.*, 2010). The number of BPs (columns of **A** and rows of P, K) is an argument to the algorithm. Both the PatternMarkers statistic and GWCoGAPS algorithm are in the CoGAPS Bioconductor package as of version 3.5. Their code is generalized for other NMF algorithms.

PatternMarkers

The patternMarkers statistic finds the genes most uniquely associated with a given pattern or linear combination of patterns by computing $\sqrt{(A_{i\cdot} - lp)^t (A_{i\cdot} - lp)}$, where $A_{i\cdot}$ are the elements of the A matrix for the i^{th} gene scaled to have a maximum of one and l is the p^{th} user specified norm. Paternmarkers defaults to $p=k$, such that l is the identity vector and the associated distance is computed separately for each of the k patterns. Unique sets are generated by ranking a genes associated distances from each norm such that the higher the rank of the gene, the less it is associated with the considered pattern. Genes are subset by their lowest ranking pattern or thresholded using the first gene to have a lower ranking in another patterns.

GWCoGAPS

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 (Fig 1) for recompiling the parallelized results. Additional runtime options, input, and manual implementations are described in the GWCoGAPS vignette.

GTeX Data

RPKM level data for the seven samples with most brain regions available was downloaded from dbGaP. GWCoGAPS was run for a range of k patterns with k=10 selected and uncertainty as 10% of the data as previously described in Fertig et al. (2013). The code to reproduce this analyses as well as the GWCoGAPS results are in Supplemental Files 1 and 2.

3 **Results / Discussion**

We apply GWCoGAPS to analyze patterns related to biological processes from distinct brain regions for different individuals in GTeX. The GWCoGAPS solutions for the initial parallel runs of on of the patterns is used to illustrate the strong association between patterns identified from the data subsets using the patternMarker Shiny App in Figure 1A. Two of the ten GWCoGAPS patterns for the GTex data are illustrated in Figure 1B. The first pattern highlights GWCoGAPS ability to deconvolute tissue specific signatures. This pattern uniquely identifies the cerebellum, determined to be the most distinct region by the consortium (Consortium *et al.*, 2015). GTEx found that strong individual specific effects increases with tissue relatedness as illustrated by their inability to achieve tissue specific clusters of the different brain regions by expression alone (Melé *et al.*, 2015; Consortium *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 enrichment in cerebellum development (GO:0021549 p=2.1E-04) and cerebellum morphogenesis (GO:0021587 p=3.4E-03).

The second pattern in Figure 1B illustrates PatternMarkers power as inference is difficult from the GWCoGAPS result alone. This pattern depicts subpopulations of cells residing in multiple brain regions derived from common precursors in the dorsal pallium. Progeny of the dorsal pallium are specified by the transcription factors TBr1 and Emx1 (Remedios *et al.*, 2007) ranked second and fourth by the PatternMarker statistic for this pattern. Gene set enrichment tests further confirms the signature as being enriched for pallium development (GO:0021543 p=1.6E-08. The output of the plotPatternMarkers function for both cerebellum and dorsal pallium patternMarkers is given Figure 1C.

Deconvolution of cell type and tissue specific signatures from aggregate transcriptomics 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 accommodate a wide variety of NMF techniques. Finally, the PatternMarkers statistic to derives gene sets uniquely representative of biological processes from the continuous gene weights of NMF solutions. Together, PatternMarkers and GWCoGAPS represent a major advance in bioinformatic approaches to find data-driven biomarkers and genetic drivers in whole genome transcriptomic data.

ACKNOWLEDGEMENTS

Funding: This work was supported by the National Institutes of Health [NCI R01CA177669 and K25CA141053 to E.J.F., NLM R01LM001100 to M.F.O. and NCI P30 CA006973] and the Cleveland Foundation and Johns Hopkins University Discovery Awards to E.J.F.

Conflicts 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.

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

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.

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.

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, pp. 59–77.

Li,Y. and Ngom,A. (2013) The non-negative matrix factorization toolbox for biological data mining. *Source code for biology and medicine*.

Mejía-Roa,E. *et al.* (2008) bioNMF: a web-based tool for nonnegative matrix factorization in biology. *Nucl. Acids Res.*, **36**, W523–8.

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. 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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 217–235.

Trendafilov,N.T. and Unkel,S. (2011) Exploratory factor analysis of data matrices with more variables than observations. *Journal of Computational and Graphical Statistics*, **20**, 874–891.

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