Bayesian Biomarker Discovery for RNAseq Data

Ali Foroughi pour
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, Ohio
foroughipour.1@osu.edu

Lori A. Dalton
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, Ohio
dalton@ece.osu.edu

ABSTRACT

RNAseq has become a popular technology for biomarker discovery. However, in many applications, such as single cell sequencing, zero counts comprise a considerable portion of data. Here we propose a new RNAseq model that explicitly models zero counts and solve a previously proposed feature selection framework, called Optimal Bayesian Filter (OBF), for this model and find the posterior probability of a feature having distributional differences across classes. As the posterior does not exist in closed form, we propose Sequence Approximation OBF (SA-OBF) as a closed form approximation which is based on log transformations of non-zero reads. We use SA-OBF to study two breast cancer RNAseq datasets.

CCS CONCEPTS

• **Theory of computation → Bayesian analysis**; • **Computing methodologies → Feature selection**;

KEYWORDS

biomarker discovery, feature selection, RNA sequencing

ACM Reference Format:
Ali Foroughi pour and Lori A. Dalton. 2018. Bayesian Biomarker Discovery for RNAseq Data. In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3233547.3233693

1 EXTENDED ABSTRACT

Biomarker discovery aims to find biological markers that differentiate between different groups, are involved in the biological mechanisms of the disease under study, and can be further utilized for diagnosis, prognosis, drug development, etc. [6] While current high throughput technologies provide a deluge of data per point, research is usually constrained to small samples impeding reliable and reproducible biomarker discovery [2, 6].

RNA sequencing (RNAseq) has become a popular technology for biomarker discovery. In many applications, such as single cell sequencing, zero reads comprise a large portion of data which poses a challenge for many popular algorithms and transformations used to study RNAseq data. For example voom transform [7] adds the constant 0.5 to all reads to avoid taking log of zero, genes with zero median expression are typically filtered out when methods such as DESeq and EdgeR are used [8], and cuffdiff2 removes genes with zero or low median expression [8]. Many current methods suffer low power for low expression genes [13], and the need for methods that better analyze genes with low average expression has been emphasized in [12]. Recently, models that directly account for zero/low reads have been proposed [16]. While the model of [16] explicitly models zero/low reads, it performs two separate sets of hypothesis tests, one to detect if the probability of a zero/low read is significantly different between the two classes, and another to detect if the mean expression of reads deviating from zero are significantly different.

Here we propose a new RNAseq model that explicitly models zero reads. We also propose an algorithm for biomarker discovery based on the proposed RNAseq model using a Bayesian framework that finds the sample conditioned probability of a feature having distributional differences across classes [5]. Optimal Bayesian filter (OBF) is the variation that assumes independent features and has been solved for Gaussian [5] and categorical features [4]. Extending OBF for the proposed RNAseq model we observe the posterior does not exist in closed form. Therefore, we propose an approximate posterior based on log transformations of non-zero reads and obtain Sequence Approximation OBF (SA-OBF). SA-OBF is fast and memory efficient, and can handle transcription per million (TPM), reads per kilobase million (RPKM), and fragments per kilobase million (FPKM) data as well. SA-OBF detects two modes of distributional differences across classes: (1) differences in the probabilities of observing zero reads and (2) distributional differences between non-zero reads. However, in contrast to the two phase analysis of [16] this is done at one step, combining information of both modes of distributional differences.

Data obtained in [1] and [15] are deposited on gene expression omnibus (GEO) [3] with accession numbers GSE47462 and GSE58135, containing 24 and 56 healthy, and 48 and 112 breast cancer points, respectively. GSE47462 provides read counts and GSE58135 reports FPKM. We use SA-OBF with a non-informative prior to select the top 1000 genes, and perform enrichment analysis using PANTHER [9, 10]. PANTHER pathways recognize 106 and 98 genes of GSE47462 and GSE58135, respectively. Top 10 genes and pathways are listed in Tabs. 1 and 2, respectively, of which many, such as the DVL1 gene [14] and p38 MAPK pathway [11], have been suggested to be affected in breast cancer.

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08.
https://doi.org/10.1145/3233547.3233693

REFERENCES

[1] A. Brunner, J. Li, X. Guo, et al. 2014. A shared transcriptional program in early breast neoplasias despite genetic and clinical distinctions. *Genome Biology* 15, 5 (2014), R71.

[2] E. Diamandis. 2010. Cancer biomarkers: can we turn recent failures into success? *J. of the Nat. Cancer Inst.* 102, 19 (2010), 1462–1467.

Table 1: Top breast cancer genes

GSE47462				GSE58135			
Rank	Gene	Rank	Gene	Rank	Gene	Rank	Gene
1	COL10A1	6	SYT9	1	DVL1	6	ZBTB7A
2	MYL2	7	PIGR	2	NCRNA00306	7	COX7A1
3	HS6ST3	8	EMX1	3	SIRT6	8	RP13-824C8.2
4	NPPA	9	OLR1	4	FAM129C	9	CDH20
5	LOC100127888	10	LOC643650	5	AMELX	10	PTBP1

Table 2: Over-represented breast cancer pathways

GSE47462		GSE58135	
Pathway name	P-value	Pathway name	P-value
Huntington disease	4.27E-03	Glutamine glutamate conversion	9.28E-03
Alzheimer disease-presenilin p.w.	7.22E-03	p38 MAPK p.w.	1.85E-02
Metabotropic glutamate receptor group III	2.95E-02	Angiogenesis	2.40E-02
Allantoin degradation	4.15E-02	CCKR sig. map	2.49E-02
Plasminogen activating cascade	4.21E-02	DPP sig. p.w.	2.65E-02
Ionotropic glutamate receptor p.w.	6.36E-02	BMP/activin sig. p.w.-drosophila	2.65E-02
Androgen/estrogene/progesterone biosyn.	6.79E-02	Adenine & hypoxanthine salvage	2.65E-02
Cytoskeletal regulation by Rho GTPase	6.94E-02	ATP synthesis	2.65E-02
Heterotri. G-prot. sig. rod outer seg. phototr.	7.41E-02	Endothelin sig. p.w.	3.51E-02
Metabotropic glutamate receptor group I	9.97E-02	Alzheimer disease-amyloid sec.	3.95E-02

[3] R. Edgar, M. Domrachev, and A. Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 1 (2002), 207–210.

[4] A. Foroughi pour and L. Dalton. 2017. Optimal Bayesian feature filtering for single-nucleotide polymorphism data. In *Proc. 2017 IEEE Int. Conf. on Bioinformatics and Biomedicine.* 2290–2292.

[5] A. Foroughi pour and L. Dalton. 2018. Heuristic algorithms for feature selection under Bayesian models with block-diagonal covariance structure. *BMC Bioinformatics* 19, 3 (2018), 70.

[6] S. Ilyin, S. Belkowski, and C. Plata-Salamán. 2004. Biomarker discovery and validation: technologies and integrative approaches. *Trends in Biotechnology* 22, 8 (2004), 411–416.

[7] C. Law, Y. Chen, W. Shi, et al. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology* 15, 2 (2014), R29.

[8] N. Leng, J. Dawson, J. Thomson, et al. 2013. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 29, 8 (2013), 1035–1043.

[9] H. Mi, X. Huang, A. Muruganujan, et al. 2017. PANTHER version 11: expanded annotation data from gene ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.* 45, 1 (2017), 183–189.

[10] H. Mi and P. Thomas. 2009. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. In *Protein Networks and Pathway Anal.*, Y. Nikolsky and J. Bryant (Eds.). Humana Press, Totowa, NJ, 123–140.

[11] R. Neve, T. Holbro, and N. Hynes. 2002. Distinct roles for phosphoinositide 3-kinase, mitogen-activated protein kinase and p38 MAPK in mediating cell cycle progression of breast cancer cells. *Oncogene* 21, 29 (2002), 4567.

[12] A. Oshlack, M. Robinson, and M. Young. 2010. From RNA-seq reads to differential expression results. *Genome Biology* 11, 12 (2010), 220.

[13] Y. Sha, J. Phan, and M. Wang. 2015. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. In *37th Annual Int. Conf. Engineering in Medicine and Biology Society.* 6461–6464.

[14] G. Turashvili, J. Bouchal, K. Baumforth, et al. 2007. Novel markers for differentiation of lobular and ductal invasive breast carcinomas by laser microdissection and microarray analysis. *BMC cancer* 7, 1 (2007), 55.

[15] K. Varley, J. Gertz, B. Roberts, et al. 2014. Recurrent read-through fusion transcripts in breast cancer. *Breast Cancer Research and Treatment* 146, 2 (2014), 287–297.

[16] Zhijin Wu, Yi Zhang, Michael L Stitzel, and Hao Wu. 2018. Two-phase differential expression analysis for single cell RNA-seq. *Bioinformatics* 1 (2018), 9.