



An integrated approach to identify bimodal genes associated with prognosis in cancer

Josivan Ribeiro Justino^{1,2*}, Clovis Ferreira dos Reis^{1*} , Andre Luis Fonseca³, Sandro Jose de Souza^{1,4,5} and Beatriz Stransky^{1,6} 

¹Universidade Federal do Rio Grande do Norte (UFRN), Metrópole Digital, Centro Multiusuário de Bioinformática, Natal, RN, Brazil.

²Universidade Federal de Rondônia, Departamento de Matemática e Estatística, Ji-Paraná, RO, Brazil.

³Universidade de São Paulo, Departamento de Genética e Biologia Evolutiva, São Paulo, SP, Brazil.

⁴Universidade Federal do Rio Grande do Norte (UFRN), Instituto do Cérebro, Natal, RN, Brazil.

⁵Sichuan University, West China Hospital, Institutes for Systems Genetics, Chengdu, China.

⁶Universidade Federal do Rio Grande do Norte (UFRN), Centro de Tecnologia, Departamento de Engenharia Biomédica, Natal, RN, Brazil.

Abstract

Bimodal gene expression (where a gene expression distribution has two maxima) is associated with phenotypic diversity in different biological systems. A critical issue, thus, is the integration of expression and phenotype data to identify genuine associations. Here, we developed tools that allow both: i) the identification of genes with bimodal gene expression and ii) their association with prognosis in cancer patients from The Cancer Genome Atlas (TCGA). Bimodality was observed for 554 genes in expression data from 25 tumor types. Furthermore, 96 of these genes presented different prognosis when patients belonging to the two expression peaks were compared. The software to execute the method and the corresponding documentation are available at the Data access section.

Keywords: Cancer, gene expression, bimodal distribution, Gaussian Mixture Model, survival analysis.

Received: April 15, 2021; Accepted: July 08, 2021.

Introduction

Studies on gene expression and regulation have been directed towards a better understanding of a diverse range of biological processes, including initial differentiation in the embryonic stage and changes in health and disease that occur during life. These patterns of gene expression have been extensively used to establish associations between phenotypes and genetic/epigenetic information (Boyle *et al.*, 2017; Young *et al.*, 2019). The challenges for such studies are significant, however, and the identification of expression signatures enriched with genuine phenotypic associations is particularly welcome. In that aspect, bimodal gene expression is an interesting pattern since their identification capitalizes on the availability of genetic and phenotypic data from large cohorts of samples and each mode can, in theory, correspond to a phenotypic state of the system.

Few previous studies have searched for bimodality in large-scale gene expression data (Bessarabova *et al.*, 2010; Mason *et al.*, 2011; Shalek *et al.*, 2013) and causes for such bimodality have been discussed, including: i) differential action of transcription factors (Ochab-Marcinek and Tabaka; 2010), ii) regulation by microRNAs (Bosia *et al.*, 2017; Del Giudice *et al.*, 2018); iii) regulation by circular RNA (Hu and Zhou;

2018) and even iv) stochastic events (Samoilov *et al.*, 2005). For an extensive review of the different methods developed for detection of bimodality, please see Moody *et al.* (2019).

The identification of genes with a bimodal expression pattern, together with sample stratification, can be used to identify important clinical and therapeutic targets in different cancer types (Floristan *et al.*, 2020). Furthermore, this process can reveal molecular signatures that distinguish tumor subtypes, which would contribute to a better clinical understanding of the biological characteristics of cancer. To be clinically useful, a bimodal pattern must exhibit a clear separation between the two groups and have significant sample sizes (Han *et al.*, 2013).

The term “bimodal expression” is related in biology to two distinct groups of continuous values of gene expression for the same gene. As discussed by Moody *et al.* (2019), genes presenting a bimodal pattern present two modes of expression in the same population. Statistically, the set of continuous values of lower and higher expression has a more consistent definition as a mixture of Gaussian distributions.

Here, a computational protocol was developed to identify, in a genome-wide context, genes with bimodal expression patterns associated with prognosis in cancer samples. To prove the applicability and robustness of our method, we used this new tool to identify genes with bimodal expression in 25 tumor types whose expression data is available from The Cancer Genome Atlas (TCGA). Finally, we made use of the availability of clinical data from TCGA to find 96 genes, among the ones

Send to correspondence to Beatriz Stransky, UFRN, Departamento de Engenharia Biomédica, Av. Senador Salgado Filho, 300, Lagoa Nova, 59078-970, Natal, RN, Brasil. E-mail: beatriz.stransky@ufrn.br.

*These authors contributed equally to the article.

with bimodal gene expression, in which patients in the two expression peaks showed different prognosis. The software to execute the method and the corresponding documentation are available at the Data access section.

Material and Methods

Data samples

Expression and clinical data from 25 different tumor types were obtained from The Cancer Genome Atlas (TCGA) project through the Genomic Data Commons Data Portal. Expression data for 24,456 genes were evaluated to identify genes with a bimodal distribution, using Fragments by Exon Kilobase per Millions of Mapped Fragments (FPKM) values. For survival analysis, clinical information was extracted from cBioPortal for Cancer Genomics.

Detection of bimodality

The detection of bimodality involves a three-step process, configured by seven parameters, listed below:

- a) **minExpression** - defines the minimum expression value in the analysis. It prevents noise in readings of low expression value from influencing the correct detection of peaks, particularly at values close to zero. This parameter must be appropriate to the type of measurement unit of expression to be used. Its default value is 0.02 FPKM;
- b) **minSampleSize** - defines the minimum sample size in the analysis. Datasets with a number of samples smaller than this value do not undergo any processing, being immediately discarded. Its default value is 50 samples.
- c) **MinClusterSize** - defines the minimum size, in relation to the total number of samples, that a cluster must have to be considered as one of the bimodal clusters. This aims to discard groups of relatively small sample populations composed of outliers, capable of altering the density profile to the point of being mistaken as a peak, especially when they occur in the upper tail of the distribution. Its default value is 10% of the total samples considered.
- d) **Threshold Up** - defines the minimum difference between the points detected as adjacent peaks and valleys on the density curve. If the difference between them is less than the parameter value, this oscillation in the density graph will be disregarded in the detection process. Its default value is 10% of the maximum density value.
- e) **Threshold Down** - peaks whose density values are below this threshold will be discarded. This aims to rule out small fluctuations in the expression values that normally occur in the upper tail of the distribution, which cause the density to fluctuate widely in this region. Its default value is 20% of the maximum density value.
- f) **Smoothing factor** – this parameter mitigates the variations in the derivatives curve to make detection less sensitive. Its default value is ‘true’.
- g) **useLog** – this parameter defines whether the expression values will be considered in their original form or whether they should be transformed into a base 2 or base 10 logarithm before analysis. This helps to improve

the sensitivity of the algorithm, particularly when the range of expression values is quite wide, which causes the upper tail of the density curve to flatten, making the peak detection process more difficult. An example of this difference in the density profile can be seen in Figure S1 where the same dataset has its density curve plotted with and without the \log_{10} transformation of the expression values. Its default value is “none”.

The three steps are:

a) Peak detection:

In this step, the initial screening of candidate genes for bimodality is performed using the density derivative. First, the density of the expression distribution of each gene is calculated using the density function of the R stats package R Core Team (2021), with the “nrd0” method to calculate the smoothing bandwidth. This method was chosen specifically because it is less precise than methods like the Sheater Jones bandwidth, guaranteeing only the detection of large fluctuations in density.

Next, the first density derivative is calculated, which undergoes a smoothing process designed to decrease the sensitivity of peak detection. For this purpose, the smooth.spline function of the R stats package was used (R Core Team, 2021), with the parameter defined by the smoothing factor. Derivative values tending to zero indicate a peak or valley. The threshold Up and Down parameters are then applied, which will define which peaks are relevant. As a result, this process returns the estimated number of peaks, which will become variable k in the subsequent step.

b) Clustering:

A data model that presents a characteristic of bimodality can be considered as the overlap of probabilistic models that represent two distinct subpopulations. In this way, we can consider bimodal distributions as a model of mixing Gaussian data (Gaussian Mixture Models - GMM) and use their specific algorithms to perform the identification and separation of these subpopulations (Titterton *et al.*, 1986).

To perform the classification based on GMM, we used the Mclust function of the R mclust package (Scrucca *et al.*, 2016), which performs a clustering of data using the expectation maximization (EM) technique, performing successive grouping operations and comparing groups with a Gaussian distribution (Hasselblad, 1966; Gelman *et al.*, 2013). This process can either infer the number of clusters expected in the distribution or start from a ‘ k ’ parameter that will designate the number of desired clusters. In our case, we already have such information, the number of clusters will be equal to the number of peaks, estimated in the previous step, plus one. Consequently, the algorithm will group the data in k effective clusters, plus an additional cluster ($k + 1$), that will contain all data points with low affinity to the main clusters.

This process returns, in addition to the clustering of samples in $k + 1$ clusters, an uncertainty index related to such classification. Arbitrarily, only the samples whose reliability in the classification received an index higher than 46% are maintained. This low rigor in the use of uncertainty values is justified because a distribution of expression indexes tends to be closer to a Poisson than to a Gaussian (Marioni *et al.*, 2008;

Langmead *et al.*, 2010; Wang *et al.*, 2010), and an excessive rigor in the use of such reliability would cause large disposal of samples. The result of the process is illustrated in Figure S2. After discarding clusters smaller than `MinClusterSize` (m) and samples contained in cluster $k + 1$, the remaining samples are passed on to the third phase of the process.

c) Peak confirmation

The samples contained in the largest k - m clusters are subjected to a new peak detection process, identical to step 1, to confirm the initial screening. If the Peak Detection process continues to identify a bimodality, as is the case shown in Figure S3A, that gene is classified as bimodal. Otherwise, the gene is discarded from the bimodal gene pool. Such a situation can be seen in Figure S3B, where the bimodality existing in the original dataset no longer can be identified when the filtered samples are used.

Survival Analysis

To verify if individuals belonging to the two different peaks of expression in a bimodal gene presented a significant difference in survival curves, we performed an analysis using the clinical data from CBioPortal, obtained as indicated above, and the Survival package in R (Therneau, 2021).

The samples identified as peak 1 and peak 2 from 554 genes with bimodal distribution were selected and Kaplan-Meier curves were evaluated with a significance level of 5% and 1% using the log-rank test. Kaplan-Meier curves were plotted using the `ggplot2` package (Wickham, 2016). All computational analysis were done using RStudio IDE.

Data availability

The computational pipeline to execute the method is freely available at https://github.com/LabBiosystemUFRN/Bimodality_Genes.

Results

Development of a method to identify genes with bimodality in gene expression data

In our method, described in Figure 1A, the identification of gene expression bimodality involves a density function, which can be used to analyse the expression values (in FPKM) for all human genes in any set of samples. Using a computational algorithm in R, the maximum and minimum points in the expression density curve of each gene is defined by identifying where the values of its derivative curve change its value signal (Figure 1B). To avoid possible noises in the stratification of samples, data points below 0.02 FPKM were excluded. To identify robust distributions concerning the difference in bimodality peaks, two thresholds were established: (i) a maximum value of 10% of density, used specifically to eliminate small ripples in the upper tail of the distribution curve, which could indicate irrelevant peaks; (ii) a 5% difference between the peak and the density valley, guaranteeing significant differences in the bimodality peaks. All these parameters are shown in Figure 1B in a schematic bimodal distribution of a hypothetical gene.

After the identification of genes with a bimodal expression pattern, we next performed sample stratification to identify

samples belonging to the first and second peaks. Assuming that each peak of the density curve represents the fashion of a subpopulation with different expression features, we can consider that the distribution of expression values, of genes with bimodality characteristics, constitutes a mixture model. Therefore, to ensure the reliable selection of samples according to the selected bimodal genes, an analysis based on the Gaussian Mixture Models (GMM) probabilistic model was introduced in our computational protocol. A schematic view of the sample stratification step is shown in Figure 1C. Finally, stratified samples were submitted to the step of bimodality detection again. Only genes that remained with a bimodal pattern after sample stratification are listed in our final results.

Identification of bimodal genes using data from 25 different tumor types from TCGA

To illustrate the use of our method, we have collected all gene expression and clinical data from TCGA for the following tumors: BLCA, BRCA, CESC, COADREAD, ESCA, GBM, HNSC, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, OV, PAAD, PCPG, PRAD, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC. These tumor types were selected because they have a minimum number of 100 patients in their respective cohorts. A total of 554 unique genes was identified as having a bimodal pattern for at least one tumor type. Table 1 shows the numbers of genes identified as bimodal for each tumor type (listed at Table S1) and Figure 2 shows the bimodal pattern of expression for 25 genes, arbitrarily selected, one for each tumor type.

We found 46 genes showing bimodal expression in more than one tumor type (Table 2). The ones most frequently found were SLC35E2, EIF1AY and RPS27, which have a bimodal pattern in 19, 10 and 10 tumor types, respectively. In this list of genes, chromosomal distribution is significantly biased toward the Y chromosome ($p < 10^{-5}$), as observed in Table 2.

Patients in different expression peaks have different prognosis

We wondered whether patients belonging to the two different peaks of the bimodal distribution would present different prognosis, as evaluated by survival curves in a Kaplan-Meier plot. All genes identified as having a bimodal distribution (Table 1) were tested. A total of 96 genes were identified as having their bimodal pattern significantly ($p < 0.01$) associated with prognosis (samples belonging to the first peak having either a better or worse prognosis when compared to samples belonging to the second peak). If a threshold of $p < 0.05$ is used, 176 genes are identified as associated with prognosis. Figure S4 shows the expression plots, reporting the bimodality, for all 96 genes found to have a bimodal expression pattern.

Figure 3 shows the respective Kaplan-Meier plots for few genes that showed significant differences in survival between patients belonging to peaks 1 and 2. Figure S5 shows the Kaplan-Meier plot for all 96 genes associated with prognosis.

Discussion

A new genome-wide method is presented to identify genes with bimodal patterns of expression using GMM analysis (Titterton *et al.*, 1986) for the stratification of

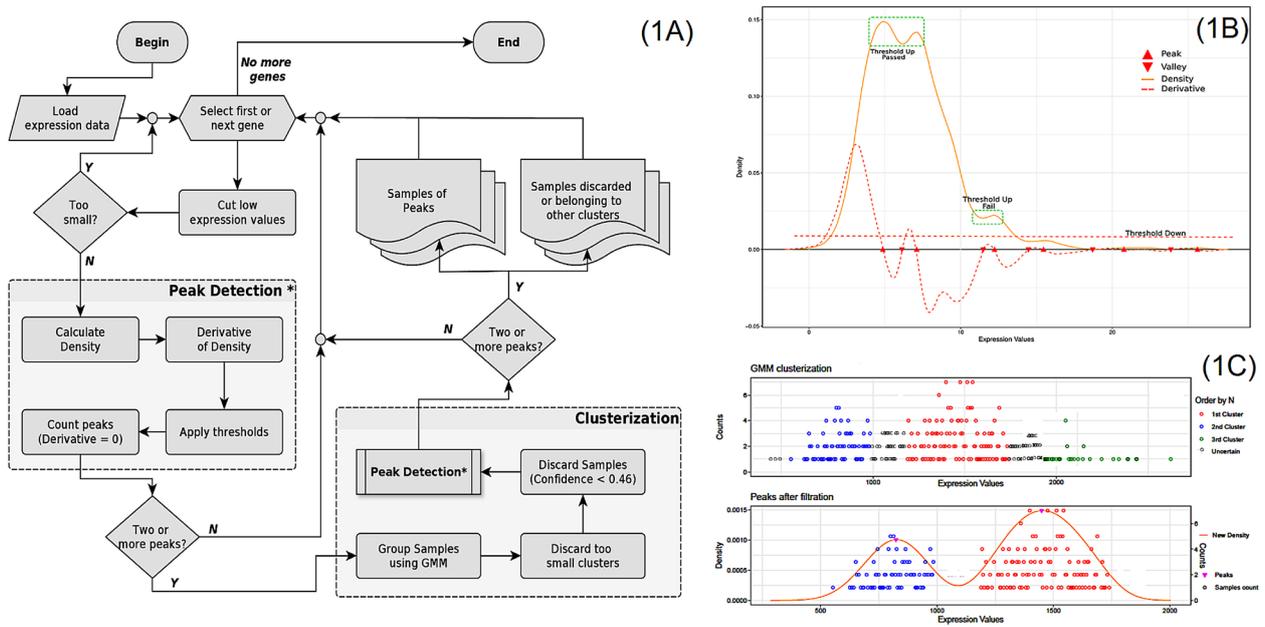


Figure 1 – Computational scheme for the identification of genes showing bimodal gene expression patterns. (A) Stages performed to process the data. (B) Schematic view of a hypothetical gene with bimodal expression with all important parameters used to define bimodality indicated. (C) Schematic view of sample clustering process, which identifies samples belonging to each peak in the bimodal distribution (see main text for details).

Table 1 – Number of genes showing bimodality for each tumor type.

Tumor	Bimodal genes	Tumor	Bimodal genes	Tumor	Bimodal genes
BLCA	8	LAML	18	SARC	6
BRCA	13	LGG	90	SKCM	14
CESC	14	LIHC	11	STAD	3
COADREAD	22	LUAD	7	TGCT	77
ESCA	8	LUSC	6	THCA	55
GBM	29	OV	5	THYM	181
HNSC	5	PAAD	12	UCEC	11
KIRC	15	PCPG	53		
KIRP	11	PRAD	9		

samples. GMM has been previously used in the analysis of gene expression data (Ficklin *et al.*, 2017; Golumbeanu *et al.*, 2019; Mirzal, 2020) but to our knowledge this is the first application of such a method for the identification of genes with bimodal expression patterns.

The applicability of the method is shown by using gene expression and clinical data for 25 tumor types available from TCGA. We identified 554 unique genes with bimodal gene expression (Table 1). Forty-six of them were identified as bimodal in more than one tumor type. Several of them have been reported previously as having a bimodal expression pattern. One of them, ERAP2, has been found by Mason *et al.* (2011) to have bimodal gene expression in human skeletal muscle. The same report (Mason *et al.*, 2011) found that GSTM1 has a bimodal expression pattern in muscle tissue. Other genes include RPS27, found by Floristan *et al.* (2020) to have bimodal expression in several tumor types, and USP9AY, found to show bimodality in endometrium (Bhat *et al.*, 2019). Interestingly, among the 46 genes with bimodality in more than one tumor type, 12 are mapped to the Y chromosome

($p < 10^{-5}$), an unexpected observation due to the low gene density in this chromosome. As reviewed by Lau (2020), some genes on the Y chromosome have dosage-sensitive functions, which might be related to a bimodal expression pattern. This remains to be further explored.

Ninety six, out of 554 genes with bimodal gene expression in all tumor types analyzed here, were identified as having differential prognosis when patients belonging to the two different modes were compared. Expression of several genes identified by us are known predictors of clinical outcome in different tumor types including ANXA1 (Gibbs and Vishwanatha, 2017), FOXJ3 (Ban *et al.*, 2013; Ma *et al.*, 2016) and CDC25 (Liu *et al.*, 2019, 2020), among many others. However, the great majority of these reports only associate overall expression with prognosis. Here, on the other hand, we associate the bimodal expression pattern with prognosis. To our knowledge, only Floristan *et al.* (2020) have associated the bimodal expression pattern of RPS27 with clinical outcome in several tumor types, a gene

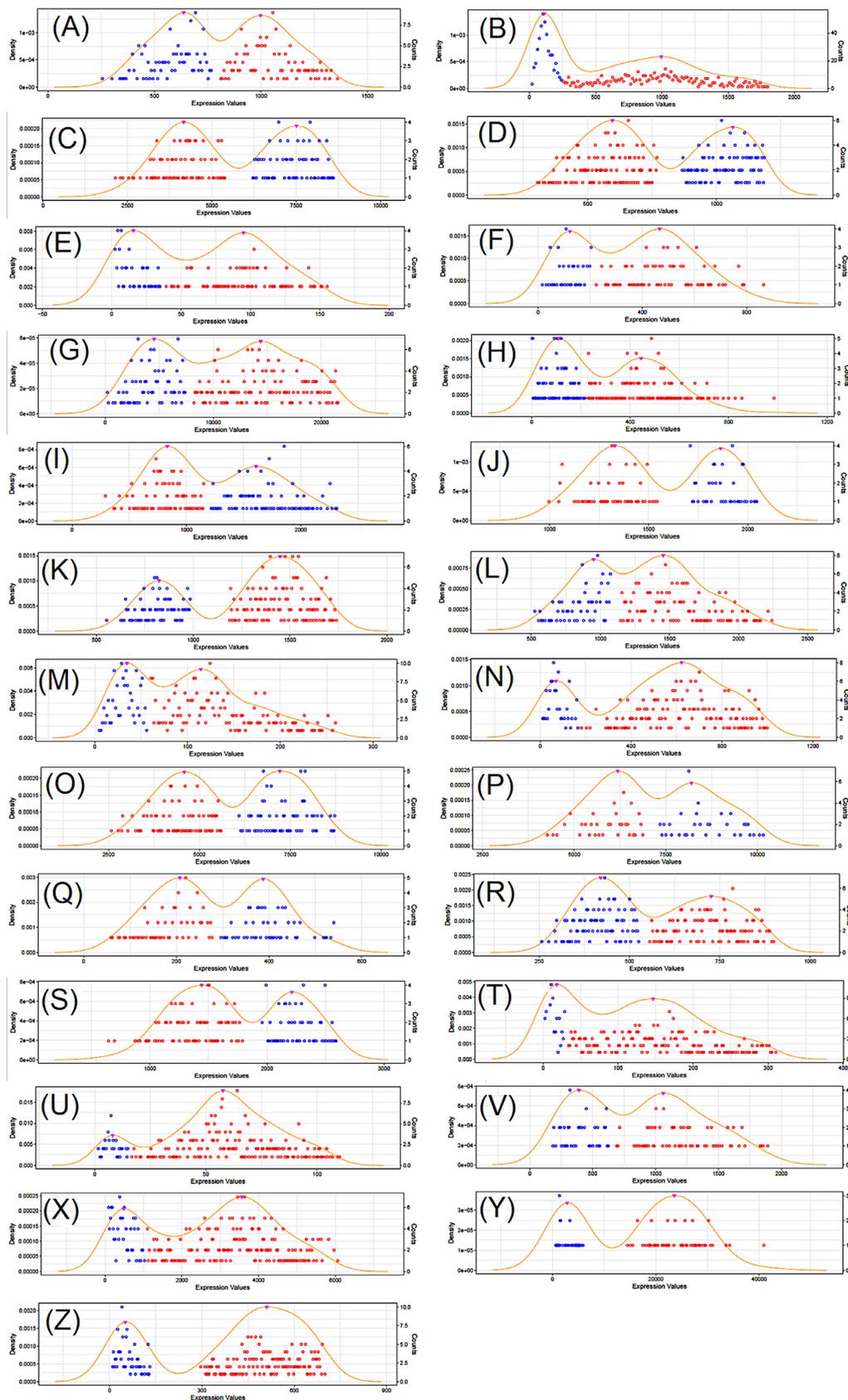
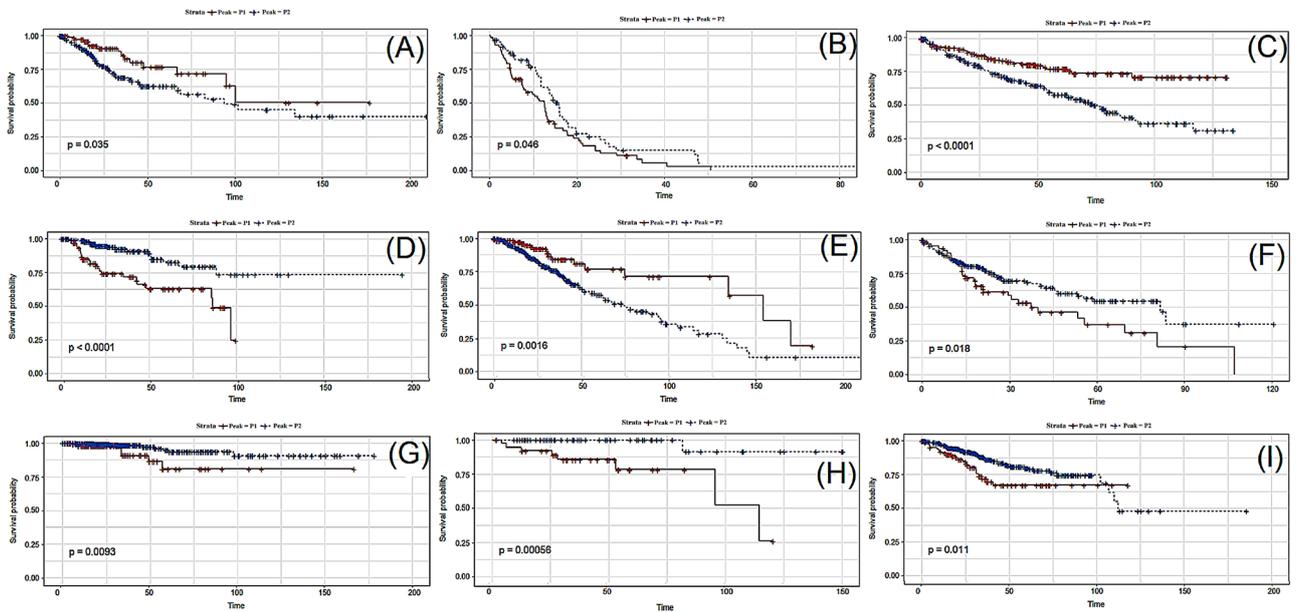


Figure 2 – Expression plot showing bimodality for a selection of genes. (A) Gene LRR14 for BLCA; (B) Gene RPS27 for BRCA; (C) Gene SEP2 for CESC; (D) Gene CHMP7 for COADREAD; (E) Gene ZNF502 for ESCA; (F) Gene GPX8 for GBM; (G) Gene CDH3 for HNSC; (H) Gene UTY for KIRC; (I) Gene FUK for KIRP; (J) Gene ADSL for LAML; (K) Gene FOXJ3 for LGG; (L) Gene ALG8 for LIHC; (M) Gene TMLHE for LUAD; (N) Gene MTAP for LUSC; (O) Gene RCC2 for OV; (P) Gene CD164 for PAAD; (Q) Gene GMIP for PCPG; (R) Gene XRR1 for PRAD; (S) Gene EI24 for SARC; (T) Gene PPAPDC3 for SKCM; (U) Gene ZNF597 for STAD; (V) Gene PCMTD1 for TGCT; (X) Gene PLCD3 for THCA; (Y) Gene ARHGDI1 for THYM and (Z) Gene MLH1 for UCEC.

Table 2 – List of genes showing bimodality for more than one tumor type.

Gene	Number	Chr	Gene	Number	Chr
RPS27	10	1	TP53	3	17
GSTM1	4	1	PLCD3	2	17
LQK1	3	1	RPS28	4	19
KHDRBS1	2	1	C19orf46	2	19
RPF1	2	1	FKBP1AP1	2	19
SLC35E2	19	1	ZNF304	2	19
C2orf43	2	2	GSTT1	2	22
SPAG16	2	2	ZDHHC15	2	X
MLH1	2	3	AWAT1	2	X
ZNF502	2	3	MAGEA6	2	X
RPL9	4	4	RPS26P11	2	X
ERAP2	3	5	CYorf15A	9	Y
PDCD2	2	6	EIF1AY	10	Y
CHMP7	2	8	DDX3Y	8	Y
MTAP	2	9	UTY	8	Y
CSNK2A1P	2	11	KDM5D	7	Y
XRRA1	2	11	RPS4Y1	7	Y
SCNN1A	3	12	ZFY	8	Y
CEP290	2	12	USP9Y	4	Y
CHFR	2	12	PRKY	3	Y
SNRPN	2	15	TMSB4Y	3	Y
TUBGCP4	2	15	NLGN4Y	2	Y
ZNF597	2	16	TTY15	3	Y

**Figure 3** – Kaplan-Meier plots of representative genes for each tumor type (one gene per tumor, arbitrarily selected). P1 and P2 correspond to the two modes of the bimodal distribution. (A) Gene ZNF304 for CESC; (B) Gene ZBTB45 for GBM; (C) Gene XRRA1 for KIRC; (D) Gene DYNC2LI1 for KIRP; (E) Gene CDC42 for LGG; (F) Gene KDM5D for LIHC; (G) Gene ANXA1 for THCA; (H) Gene LAIR1 for THYM and (I) Gene HNF1B for UCEC.

also observed in our data. This makes our analysis the first one, to our knowledge, to explore the association between the modes of gene expression distribution with prognosis in a genome-wide context.

Several issues should be considered in the interpretation of our results. For example, cellular heterogeneity within samples in a given cohort is a factor that can generate genes with bimodal expression. In our case, this is minimized by the fact that TCGA samples are selected for high tumor cell content but this issue should be critically considered when more heterogeneous cohorts are analyzed. Furthermore, clinical and/or biological features should be considered when interpreting data from our method. For example, in cancer studies one should be careful with cohort heterogeneity regarding staging and progression, among many other clinical features.

We envisage that our method will be a useful tool for the genome-wide identification of genes with bimodal pattern of expression. The software to execute the method and the corresponding documentation are available at the Data access section.

Acknowledgements

This work was supported by Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (grant 23038.004629/2014-19 to SJS and BS), by Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq (grant 305233/2015-7) and by Center of Excellence - International Cooperation InitiativGrant, West China Hospital, Sichuan University (serial number 139200032), both to SJS.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author Contributions

JRJ, CFR, SJS conceived the study; JRJ, CFR, ALF analyzed the data; JRJ, CFR, SJS, BS wrote the manuscript, and all authors read and approved the final version.

References

- Ban JY, Park HJ, Kim SK, Kim JW, Lee YA, Choi IA, Chung JH and Hong SJ (2013) Association of forkhead box J3 (FOXJ3) polymorphisms with rheumatoid arthritis. *Mol Med Rep* 8:1235-1241.
- Bessarabova M, Kirillov E, Shi W, Bugrim A, Nikolsky Y and Nikolskaya T (2010) Bimodal gene expression patterns in breast cancer. *BMC Genomics* 11 Suppl 1:S8.
- Bhat MA, Sharma JB, Roy KK, Sengupta J and Ghosh D (2019) Genomic evidence of Y chromosome microchimerism in the endometrium during endometriosis and in cases of infertility. *Reprod Biol Endocrinol* 17:22.
- Bosia C, Sgrò F, Conti L, Baldassi C, Brusa D, Cavallo F, Cunto FD, Turco E, Pagnani A and Zecchina R (2017) RNAs competing for microRNAs mutually influence their fluctuations in a highly non-linear microRNA-dependent manner in single cells. *Genome Biol* 18:37.
- Boyle EA, Li YI and Pritchard JK (2017) An expanded view of complex traits: From polygenic to omnigenic. *Cell* 169:1177-1186.
- Del Giudice M, Bo S, Grigolon S and Bosia C (2018) On the role of extrinsic noise in microRNA-mediated bimodal gene expression. *PLoS Comput Biol* 14:e1006063.
- Ficklin SP, Dunwoodie LJ, Poehlman WL, Watson C, Roche KE and Feltus FA (2017) Discovering condition-specific gene co-expression patterns using Gaussian mixture models: A cancer case study. *Sci Rep* 7:8617.
- Floristan A, Morales L, Hanniford D, Martinez C, Castellano-Sanz E, Dolgalev I, Ulloa-Morales A, Vega-Saenz ME, Moran U, Darvishian F *et al.* (2020) Functional analysis of RPS27 mutations and expression in melanoma. *Pigment Cell Melanoma Res* 33:466-479.
- Gelman A, Rubin DB, Carlin JB and Stern HS (2013) *Bayesian Data Analysis*, 3rd edition. CRC Press, Boca Raton, 675 pp.
- Gibbs LD and Vishwanatha JK (2017) Prognostic impact of AnxA1 and AnxA2 gene expression in triple-negative breast cancer. *Oncotarget* 9:2697-2704.
- Golumbeanu M, Desfarges S, Hernandez C, Quadroni M, Rato S, Mohammadi P, Telenti A, Beerwinkler N and Ciuffi A (2019) Proteo-transcriptomic dynamics of cellular response to HIV-1 infection. *Sci Rep* 9:213.
- Han F, Faraco J, Dong XS, Ollila HM, Lin L, Li J, An P, Wang S, Jiang KW, Gao ZC *et al.* (2013) Genome wide analysis of narcolepsy in China implicates novel immune loci and reveals changes in association prior to versus after the 2009 H1N1 influenza pandemic. *PLoS Genet* 9:e1003880.
- Hasselblad V (1966) Estimation of parameters for a mixture of normal distributions. *Technometrics* 8:431-444.
- Hu Q and Zhou T (2018) ElciRNA-mediated gene expression: tunability and bimodality. *FEBS Lett* 592:3460-3471.
- Langmead B, Hansen KD and Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol* 11:R83.
- Lau YC (2020) Y chromosome in health and diseases. *Cell Biosci* 10:97.
- Liu K, Lu R, Zhao Q, Du J, Li Y, Zheng M and Zhang S (2019) Association and clinicopathologic significance of p38MAPK-ERK-JNK-CDC25C with polyploid giant cancer cell formation. *Med Oncol* 37:6.
- Liu K, Zheng M, Lu R, Du J, Zhao Q, Li Z, Li Y and Zhang S (2020) The role of CDC25C in cell cycle regulation and clinical cancer therapy: a systematic review. *Cancer Cell Int* 20:213.
- Ma W, Yu Q, Jiang J, DU X, Huang L, Zhao L and Zhou QI (2016) miR-517a is an independent prognostic marker and contributes to cell migration and invasion in human colorectal cancer. *Oncol Lett* 11:2583-2589
- Marioni JC, Mason CE, Mane SM, Stephens M and Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18:1509-1517.
- Mason CC, Hanson RL, Ossowski V, Bian L, Baier LJ, Krakoff J and Bogardus C (2011) Bimodal distribution of RNA expression levels in human skeletal muscle tissue. *BMC Genomics* 12:98.
- Mirzal A (2020) Statistical Analysis of Microarray Data Clustering using NMF, Spectral Clustering, Kmeans, and GMM. *IEEE/ACM Trans Comput Biol Bioinform* Sep 21:1.
- Moody L, Mantha S, Chen H and Pan Y (2019) Computational methods to identify bimodal gene expression and facilitate personalized treatment in cancer patients. *J Biomed Informatics* 100:100001.
- Ochab-Marcinek A and Tabaka M (2010) Bimodal gene expression in noncooperative regulatory systems. *Proc Natl Acad Sci U S A* 107:22096-22101.
- Samoilov M, Plyasunov S and Arkin AP (2005) Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proc Natl Acad Sci U S A* 102:2310-2315.

- Scrucca L, Fop M, Murphy TB and Raftery AE (2016) mclust 5: Clustering, classification and density estimation using gaussian finite mixture models. *R Journal* 8:289-317.
- Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublotte JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D *et al.* (2013) Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature* 498:236-240.
- Titterton DM, Smith AFM and Makov UE (1986) *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 254 pp.
- Wang L, Feng Z, Wang X, Wang X and Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* 26:136-138.
- Wickham H (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, 260 pp.
- Young AI, Benonisdotir S, Przeworski M and Kong A (2019) Deconstructing the sources of genotype-phenotype associations in humans. *Science* 365:1396-1400.

Internet Resources

- cBioPortal for Cancer Genomics, <https://www.cbioportal.org/> (accessed 1 March 2021).
- Genomic Data Commons Data Portal, <https://portal.gdc.cancer.gov/> (accessed 1 March 2021).
- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/> (accessed 1 March 2021).
- RStudio, <https://www.rstudio.com/> (accessed 1 March 2021).

- Therneau T (2021). *A Package for Survival Analysis in R*. R package version 3.2-11, <https://CRAN.R-project.org/package=survival> (accessed 1 March 2021).

Supplementary material

The following online material is available for this article:

- Table S1 – List of bimodal genes estimated in each tumor.
- Figure S1 – Detection of bimodality for the RPS27 gene using gene expression data for breast adenocarcinoma (BRCA) from TCGA applying (upper graph), or not (lower graph), the \log_{10} transformation to the data.
- Figure S2 – Clustering of samples in $k + 1$ clusters.
- Figure S3 – Third phase of the bimodality detection process.
- Figure S4 – Expression plot showing bimodality for all 96 genes associated with prognosis.
- Figure S5 – Survival charts for the 96 genes associated with prognosis ($p < 0.01$) for the 426 following tumor types: KIRC, KIRP, LGG, SKCM, THCA and THYM.

Associate Editor: Guilherme Correa de Oliveira

License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License (type CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original article is properly cited.