A new genome-wide method to identify genes with bimodal gene expression

Josivan Ribeiro Justino^{1,2#}, Clovis F. Reis^{3#}, Andre Faustino Fonseca⁴, Sandro Jose de Souza^{3,5,6}, Beatriz Stransky^{3,7,*}

¹Bioinformatics Graduate Program, Digital Metropolis Institute, Federal University of Rio Grande do Norte (UFRN), Natal, Brazil

²Department of Mathematics and Statistics, Federal University of Rondônia, Ji-Parana, Brazil

³Bioinformatics Multidisciplinary Environment, Digital Metropolis Institute, UFRN, Natal, Brazil

⁴Genetics and Evolutionary Biology Department, São Paulo University, São Paulo, Brazil

⁵Brain Institute, UFRN, Natal, Brazil

⁶Institutes for Systems Genetics, West China Hospital, Sichuan University, Chengdu, China

⁷Biomedical Engineering Department, Center of Technology, UFRN, Brazil

* Corresponding author

E-mail: beatriz.stransky@ct.ufrn.br

These authors contributed equally to this work.

Abstract

A new method is presented to detect bimodality in gene expression data using the Gaussian Mixture Models to cluster samples in each mode. We have used the method to search for bimodal genes in data from 25 tumor types available from The Cancer Genome Atlas. The method identified 554 genes with bimodal gene expression, of which 46 were identified in more than one cancer type. To further illustrate the impact of the method, we show that 96 out of the 554 genes with bimodal expression patterns presented different prognosis when patients belonging to the two expression peaks are compared. The software to execute the method and the corresponding documentation are available at https://github.com/LabBiosystemUFRN/Bimodality_Genes.

Introduction

Studies on gene expression and regulation have been directed towards a better understanding of a diverse range of biological processes, including the 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 [1-2]. The challenges for such studies are significant, however, and the identification of expression signatures enriched with bona fide 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 clinical 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 [3-5]. Causes for such bimodality have been discussed, including: i) differential

action of transcription factors [6], ii) regulation by microRNAs [7, 8]; iii) regulation by circular RNA [9] and even iv) stochastic events [10].

The identification of genes with a bimodal expression pattern, together with sample stratification, can be used to identify important clinical and therapeutic targets [11]. 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 [12]. The term "bimodal expression" is related in biology to two distinct groups of continuous values of gene expression for the same gene. 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 genes with bimodal expression patterns using the Gaussian Mixture Models (GMM) in a genome-wide context. 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 with bimodal gene expression, in which patients in the two expression peaks showed different prognosis. Software to execute the method and the corresponding documentation are freely available at https://github.com/LabBiosystemUFRN/Bimodality_Genes.

Materials & 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 (<https://portal.gdc.cancer.gov>). 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, the clinical information was extracted from cBioPortal for Cancer Genomics (<https://www.cbioportal.org/>).

Detection of bimodality

The detection of bimodality involves a three-step process, configured by 7 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 Supplementary Figure 1 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 [13], 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 [13], 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 can be considered 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 [14].

To perform the classification based on GMM, we used the Mclust function of the R mclust package [15], which performs a clustering of data using the expectation maximization (EM) technique, performing successive grouping operations and comparing groups with a Gaussian distribution [16,17]. 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 [18-20], and an excessive rigor in the use of such reliability would cause large disposal of samples;

The result of the process is illustrated in Supplementary Figure 2. 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 Supplementary Figure 3A, that gene is classified as bimodal. Otherwise, the gene is discarded from the bimodal gene pool. Such a situation can be seen in Supplementary Figure 3B, 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 [21].

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 [22].

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 analyze 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 0.002 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 the 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 tumors were selected because they have a minimum number of 100 patients in their dataset. A total of 554 unique genes was identified as having a bimodal pattern for at least one tumor type (analysis performed without the logarithmic transformation). Table 1 shows the numbers of genes identified as bimodal for each tumor type (listed at Supplementary Table 1) and Figure 2 shows the bimodal pattern of expression for 25 genes, arbitrarily selected, one for each tumor type. Supplementary Figure 4 shows the same type of plot for all genes found to have a bimodal expression pattern.

Fig 1. Computational scheme for the identification of genes showing bimodal gene expression patterns. 1A) Stages performed to process the data. 1B) Schematic view of a hypothetical gene with bimodal expression with all important parameters used to define bimodality indicated. 1C) Schematic view of sample clustering process, which identifies

samples belonging to each peak in the bimodal distribution (see main text for more 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		

Figure 2: Expression plot showing bimodality for a selection of genes. A) Gene LRRC14 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 ARHGDIB for THYM and Z) Gene MLH1 for UCEC.

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.

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	ZDHC15	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

*Gene corresponds to the official gene name; *Number refers to the number of tumor types where that respective gene is bimodal and *Chr is the chromosome where the gene is mapped.

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 3 shows the respective Kaplan-Meier plots for few genes that showed significant differences in survival between patients belonging to peaks 1 and 2. Supplementary Figure 4 shows the Kaplan-Meier plot for all 96 genes associated with prognosis.

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.

Discussion

A new genome-wide method is presented to identify genes with bimodal patterns of expression using GMM analysis [14] for the stratification of samples. GMM has been previously used in the analysis of gene expression data [23-25] 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 genes unique 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, had been found by [4] to have bimodal gene expression in human skeletal muscle. The same report from [4] found that GSTM1 has a

bimodal expression pattern in muscle tissue. Other genes include RPS27, found by [11] to have bimodal expression in several tumor types, and USP9AY, found to show bimodality in endometrium [26]. 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 [27], 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 [28], FOXJ3 [29, 30] and CDC25 [31, 32], among many others. However, the great majority of these reports only associate expression with prognosis. On the other hand, we associate the bimodal expression pattern with prognosis. To our knowledge, only [11] have associated the bimodal expression pattern of RPS27 with clinical outcome in several tumor types, a gene 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 computational pipeline to execute the method is freely available at https://github.com/LabBiosystemUFRN/Bimodality_Genes.

REFERENCES

1. Young AI, Benonisdottir S, Przeworski M, Kong A. Deconstructing the sources of genotype/phenotype associations in humans. *Science*. 2019 Sep 27;365(6460):1396-1400. doi: 10.1126/science.aax3710.
2. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017, 169(7):1177-1186. doi: 10.1016/j.cell.2017.05.038. PMID: 28622505; PMCID: PMC5536862.
3. Bessarabova M, Kirillov E, Shi W, Bugrim A, Nikolsky Y, Nikolskaya T. Bimodal gene expression patterns in breast cancer. *BMC Genomics*. 2010;11 Suppl 1(Suppl 1):S8. Published 2010 Feb 10. doi:10.1186/1471-2164-11-S1-S8
4. Mason CC, Hanson RL, Ossowski V, Bian L, Baier LJ, Krakoff J, et al. Bimodal distribution of RNA expression levels in human skeletal muscle tissue. *BMC genomics*, v. 12, n. 1, p. 98, (2011). doi.org/10.1186/1471-2164-12-98
5. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, v. 498, n. 7453, p. 236–240, 2013. doi.org/10.1038/nature12172
6. Ochab-Marcinek A, Tabaka M. Bimodal gene expression in noncooperative regulatory systems. *Proceedings of the National Academy of Sciences*, v. 107, n. 51, p. 22096-22101, 2010. doi.org/10.1073/pnas.1008965107
7. Bosia C, Sgrò F, Conti L, Baldassi C, Brusa D, Carvalho F, et al. RNAs competing for microRNAs mutually influence their fluctuations in a highly non-linear microRNA-

- 335 dependent manner in single cells. *Genome Biol* 18, 37 (2017). doi.org/10.1186/s13059-
336 017-1162-x
- 337 **8.** Giudice MD, Bo S, Grigolon S, Bosia C. On the role of extrinsic noise in microRNA-
338 mediated bimodal gene expression. *PLoS Comput Biol*, v. 14, n. 4, p. e1006063, 2018.
339 doi.org/10.1371/journal.pcbi.1006063
- 340 **9.** Hu Q, Zhou T. ElciRNA-mediated gene expression: tunability and bimodality. *Febs*
341 *Letters*, v. 592, n. 20, p. 3460-3471, 2018. doi.org/10.1002/1873-3468.13253
- 342 **10.** Samoilov M, Plyasunov S, Arkin AP. Stochastic amplification and signaling in
343 enzymatic futile cycles through noise-induced bistability with oscillations. *Proceedings of*
344 *the National Academy of Sciences*, v. 102, n. 7, p. 2310-2315, 2005.
345 doi.org/10.1073/pnas.0406841102
- 346 **11.** Floristán A, Morales L, Hanniford D, Martinez C, Castellano-Sanz E, Dolgalev I, et al.
347 Functional analysis of RPS27 mutations and expression in melanoma. *Pigment Cell*
348 *Melanoma Res*, v. 33, n. 3, p. 466-479, 2020. doi: 10.1111/pcmr.12841.
- 349 **12.** Han F, Faraco J, Dong XS, Ollila HM, Lin L, Li J, An P, et al. "Genome Wide Analysis
350 of Narcolepsy in China Implicates Novel Immune Loci and Reveals Changes in
351 Association prior to versus after the 2009 H1N1 Influenza Pandemic." *PLoS Genetics* 9
352 (10): e1003880, 2013. doi.org/10.1371/journal.pgen.1003880
- 353 **13.** R Core Team (2020). R: A language and environment for statistical computing.
- 354 **14.** Titterton DM, Smith AFM, Makov HE "Statistical Analysis of Finite Mixture
355 Distributions," John Wiley & Sons, (1985). ISBN 0471907634
- 356 **15.** Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and
357 density estimation using Gaussian finite mixture models. *The R Journal*, v. 8/1, pp. 289-
358 317, 2016.
- 359 **16.** Gelman A, Carlin JB, Stern HS, Rubin D. *Bayesian Data Analysis*. New York:
360 Chapman and Hall/CRC, (1995), p. 552. doi.org/10.1201/9780429258411

- 361 **17.** Hasselblad V. Estimation of parameters for a mixture of normal distributions. *Technometrics*,
362 v. 8, n. 3, p. 431-444, 1966. doi.org/10.1080/00401706.1966.10490375
- 363 **18.** Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of
364 technical reproducibility and comparison with gene expression arrays. *Genome Res.*
365 2008 Sep;18(9):1509-17. doi: 10.1101/gr.079558.108. Epub 2008 Jun 11. PMID:
366 18550803; PMCID: PMC2527709.
- 367 **19.** Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying
368 differentially expressed genes from RNA-seq data. *Bioinformatics*. 2010 Jan 1;26(1):136-
369 8. doi: 10.1093/bioinformatics/btp612. Epub 2009 Oct 24. PMID: 19855105.
- 370 **20.** Langmead B, Hansen KD, Leek JT. Cloud-scale RNA-sequencing differential
371 expression analysis with Myrna. *Genome Biol* **11**, R83 (2010). doi.org/10.1186/gb-2010-
372 11-8-r83
- 373 **21.** Therneau T (2020). A Package for Survival Analysis in R. R package version 3.2-7,
374 <https://CRAN.R-project.org/package=survival>
- 375 **22.** Wilkinson L. “ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H.”
376 *Biometrics*. 2011. doi.org/10.1111/j.1541-0420.2011.01616.x.
- 377 **23.** Ficklin SP, Dunwoodie LJ, Poehlman WL, Watson C, Roche KE, Feltus
378 FA. Discovering Condition-Specific Gene Co-Expression Patterns Using Gaussian
379 Mixture Models: A Cancer Case Study. *Scientific Reports*, v. **7**, 8617 (2017).
380 doi.org/10.1038/s41598-017-09094-4
- 381 **24.** Mirzal A. Statistical Analysis of Microarray Data Clustering using NMF, Spectral
382 Clustering, Kmeans, and GMM. *IEEE/ACM Transactions on Computational Biology and*
383 *Bioinformatics*, 2020. doi: 10.1109/TCBB.2020.3025486.
- 384 **25.** Golumbeanu M, Desfarges S, Hernandez C, Quadroni M, Rato S, Mohammadi P, et
385 al. Proteo-Transcriptomic Dynamics of Cellular Response to HIV-1 Infection. *Sci Rep*.
386 2019 Jan 18;9(1):213. doi: 10.1038/s41598-018-36135-3.
- 387 **26.** Bhat MA, Sharma JB, Roy KK, Sengupta J, Ghosh D. Genomic evidence of Y

- 388 chromosome microchimerism in the endometrium during endometriosis and in cases of
389 infertility. *Reprod Biol Endocrinol* **17**, 22 (2019). doi.org/10.1186/s12958-019-0465-z
- 390 **27.** Lau YFC. Y chromosome in health and diseases. *Cell & Bioscience*, v. 10, n. 1, p. 1-
391 10, 2020. doi.org/10.1186/s13578-020-00452-w
- 392 **28.** Gibbs LD, Vishwanatha JK. Prognostic impact of AnxA1 and AnxA2 gene expression
393 in triple-negative breast cancer. *Oncotarget*. 2017;9(2):2697-2704. Published 2017 Dec
394 23. doi:10.18632/oncotarget.23627
- 395 **29.** Ban JY, Park HJ, Kim SK, Kim JW, Lee YA, Choi IA, et al. Association of forkhead
396 box J3 (FOXJ3) polymorphisms with rheumatoid arthritis. *Mol Med Rep*. 2013
397 Oct;8(4):1235-41. doi: 10.3892/mmr.2013.1623. Epub 2013 Aug 9. PMID: 23933633.
- 398 **30.** Ma W, Yu Q, Jiang J, DU X, Huang L, Zhao L, et al. miR-517a is an independent
399 prognostic marker and contributes to cell migration and invasion in human colorectal
400 cancer. *Oncol Lett*. 2016 Apr;11(4):2583-2589. doi: 10.3892/ol.2016.4269. Epub 2016
401 Feb 24. PMID: 27073521; PMCID: PMC4812169.
- 402 **31.** Liu K, Zheng M, Lu R, Du J, Zhao Q, Li Z, et al. The role of CDC25C in cell cycle
403 regulation and clinical cancer therapy: a systematic review. *Cancer Cell Int*. 2020 Jun
404 3;20:213. doi: 10.1186/s12935-020-01304-w. PMID: 32518522; PMCID: PMC7268735.
- 405 **32.** Liu K, Lu R, Zhao Q, Du J, Li Y, Zheng M, et al. Association and clinicopathologic
406 significance of p38MAPK-ERK-JNK-CDC25C with polyploid giant cancer cell formation.
407 *Med Oncol*. 2019 Nov 16;37(1):6. doi: 10.1007/s12032-019-1330-9. PMID: 31734829.

Figure 1

Figure 2

Figure 3

Supp. Figure 1

Supp. Figure 2

Supp. Figure 3

Supp. Figure 4

Supp. Figure 5