Trends in Developmental Biology
Vol. 12, 2019

Original Article

The expression profile of genes required for the development of the mammary gland

Lissur A. Orsine[1], Glaura C. Franco[2] and J. Miguel Ortega[1,*]

[1]Department of Biochemistry and Immunology, ICB, UFMG, Belo Horizonte, MG, Brazil;
[2]Department of Statistics, ICEx, UFMG, Belo Horizonte, MG, Brazil.

ABSTRACT

The development of the mammary gland encompasses four key periods. They are controlled by a set of gene products, either expressed at the period or acting in it, although they have already been expressed before. Conducting text-mining analysis, we collected 406 genes involved in the four periods of breast development: (i) embryonic development, (ii) puberty, (iii) pregnancy and lactation and (iv) regression after lactation. Comparative profiles of gene expression are available from several sources, and one of them is the Genotype-Tissue Expression (GTEx) project which comprises RNAseq data from 53 tissues. Here we analyzed the expression of the set of text-mined mammary gland genes with a novel criterion which indicates if they are highly differentially expressed as outliers in the different tissues contained in the GTEx dataset. This approach aims to reveal in which tissues they are either "overactive" or "tissue-specific" and concomitantly being implicated in breast development. Data showed that, in most of the cases, genes of the mammary gland development, which are GOAT (Gene OverActive or Tissue-specific) in a subset of tissues, are not GOAT in breast. Thus, development of different tissues might require GOATs that will likewise act as controllers of breast development without being highly differentially expressed to play their role. Our analyses contribute to the understanding of the scenario of developmental control by gene expression, while providing a simple way to point out highly differentially expressed genes in different tissues.

KEYWORDS: expression profiles, mammary gland development, overactive genes, tissue-specific genes.

INTRODUCTION

In the recent past, when a researcher became interested in the comparative expression of a given gene, the source of information was UniGene, from NCBI [1, 2]. Partial sequences from mRNAs, known as Expressed Sequence Tags (ESTs) from several tissues were assembled into *contigs* and the composition of the consensus allowed a representation of tissue expression pattern as virtual hybridization spots, a procedure that, nowadays, does not bring any clue of what it is to our young researchers (Figure 1). However, the tissue expression pattern was there, promptly inspected. Besides, expert scientists, mostly on bioinformatics, could extract this information from several independent deposits in NCBI Gene Expression Omnibus (GEO) [3, 4] and other sources. The initiative of the EMBL-EBI Expression Atlas [5-7] became a popular way to obtain this information.

The EMBL-EBI Expression Atlas provides a good representation of comparative tissue expression for a gene of interest. The main output is a heat map table where one can move the cursor over

*Corresponding author: miguel@icb.ufmg.br

esophagus	0		0/20154
eye	14		3/208840
heart	0		0/89524
intestine	4		1/231981
kidney	71		15/210778
larynx	0		0/23466
liver	204		42/205291
lung	11		4/334815
lymph	0		0/44302
lymph node	0		0/89748
mammary gland	19		3/151230
mouth	0		0/66150
muscle	9		1/106371
nerve	0		0/15535
ovary	0		0/101488
pancreas	4		1/213440

Figure 1. Expression data from NCBI UniGene showing tissue, TPM, spot intensity based on TPM and ESTs per total ESTs in pool [1, 2]. Gene is SERPINF2. Figure was extracted from NCBI Handbook [13].

a particular cell to turn colors into numbers, with transcripts per million (TPM) as units. For comparative biology, the best sources of information are the "baseline experiments", which comprise expression rates for many tissues or cell lines. A great source of information found in EMBL-EBI Expression Atlas is a set of 53 distinct tissues coming from the Genotype-Tissue Expression (GTEx) project [8, 9]. Other sources are presently available, such as: expression across 530 tissues tested by GENENVESTIGATOR [10]; expression shown in Bgee [11], where low Raw score means that the gene is highly expressed across all gene collections, conditions and species while the expression score uses the minimum and maximum Rank of the species to normalize the expression to a value between 0 and 100; and GTEx data at old version v6, accessed within the UCSC Genome Browser [12].

These sources of comparative biological information, despite depicting which genes continue to be relevant for adult tissue functions, lack on information about the developmental patterns of gene expression. Frequently this kind of information has to be extracted from series, dedicated to a single tissue or cell type and often under physiological conditions of interest of the specific research. However, it is possible, with the information attained in these sources, to verify if genes that are highly differentially expressed in already developed tissues might have had been implicated in the history of development of a specific tissue. Here we developed a method to point out which genes can be characterized as highly differentially expressed, characterizing them as overactive in some tissues or tissue-specific, and indicating which of these would contribute to mammary gland development stages. Therefore we set out to investigate if genes that are implicated in breast development and highly differentially expressed in other tissues, present the same status in the breast.

MATERIALS AND METHODS

Developmental functional specifications of tissues are often connected to the expression of a set of genes, which control the pathways that culminate with the tissue development. Genes implicated in the development of mammary gland were selected with a combination of a query in PubMed and the

processing of this large result with a text mining tool named MedLine Ranker [14] which, with the help of a few abstracts that are of user interest, the *training set*, can rank a great amount of abstracts related to them. The procedure of building a pathway with this approach has been described in a manual [15]. Basically, selected abstracts are processed with another text mining tool called PESCADOR [16] to reveal interactions between genes. Here we analyzed the expression of 406 genes implicated in four periods of mammary gland development: (i) embryonic development; (ii) puberty; (iii) pregnancy and lactation; (iv) regression after lactation. Expression data were obtained from the EMBL-EBI Expression Atlas [5, 6, 7] and we focused on a set of 53 distinct tissues coming from the Genotype-Tissue Expression (GTEx) project [8, 9]. For each gene, the expression given in TPM, averaged from many samples per tissue, was obtained and sorted in ascending order. We determined the values of first and third Quartiles, Q1 and Q3, and the Interquartile Range, IQR = Q3 - Q1. In a boxplot graph [17, 18], outliers are determined by multiplying IQR per 1.5 and this product is added to Q3 to obtain a limit known as upper inner fence. Therefore, for upper inner fence determination, the calculation is Q3 + (1.5 × IQR). Samples above this fence and below upper outer fence are considered mild outliers. The upper outer fence is calculated by Q3 + (3 × IQR), and samples above this limit are known as far outliers. Manual inspection of GTEx profiles suggested that the farthest outliers could be classified by the factor that multiplies 1.5 in the first formula, and we named this factor o-score, after outlier score. Thus, we calculated fences as Q3 + (o-score × 1.5 × IQR). For a given gene, the o-score corresponding to a given tissue TPM is determined by:

$$\text{o-score} = (\text{TPM} - \text{Q3}) / (\text{IQR} \times 1.5)$$

However, for some profiles, gene expression assumes very small values for most tissues, thus Q1 and Q3 tend to zero; therefore the IQR also tends to zero, and o-score is infinite. These genes have been considered in the literature as tissue-specific genes (TS), since their expression is absent in most tissues. For characterizing tissues where gene is TS, we ordered the expression data, given in TPM, and calculated the cumulative expression from the less to the most expressed gene. We then determined 25% of the cumulative expression, inspected the ordered tissues to determine which tissue first accumulated 25% of total cumulative expression in the ranked list and filtered out tissues up to this one, saving, therefore, the tissues with the most expressed TS genes for comparative analysis.

For TS genes, instead of o-score, we considered ts-score as the cumulative expression in a ranked list of the 53 tissues. Thus, in this study we aimed to verify if genes with the highest o-scores or ts-scores in breast were enriched in the list of text-mined genes related to mammary gland development, and, if not, in which tissues these mined genes present high o-score and ts-score. For o-score we restricted the analysis to o-score > 5 and ts-score > 25% of the cumulative expression.

RESULTS

Overactive genes versus tissue-specific genes

Tissue-specific genes have been commented in the literature for years; however, it is possible to observe from manual inspection that some genes are broadly expressed across tissues, as Glyceraldehyde Dehydrogenase gene (GAPDH). As seen in Figure 2, besides being expressed in all tissues, GAPDH is more active in a couple of tissues and in EBV-transformed lymphocyte (Epstein-Barr virus-transformed lymphocyte), a cell line also present in the GTEx dataset. We have noticed that tissues below the upper inner fence of boxplots presented a profile that, if submitted to a statistical adherence test, fits the normal distribution. Therefore, we consider these tissues as having a Gaussian expression. The biological interpretation of this fact is that tissues aim to produce the average of 14 TPM (i.e. the average of the points below the first line, the inner fence, in Figure 2) and, like most biological processes, the result is a normal distribution around the mean. But we noticed that some tissues produce expression levels that reject the Gaussian distribution. In this report we will concentrate on those with o-score over five, which, in the GAPDH example, are the skeletal muscle tissue and the cell line EBV-transformed lymphocyte.

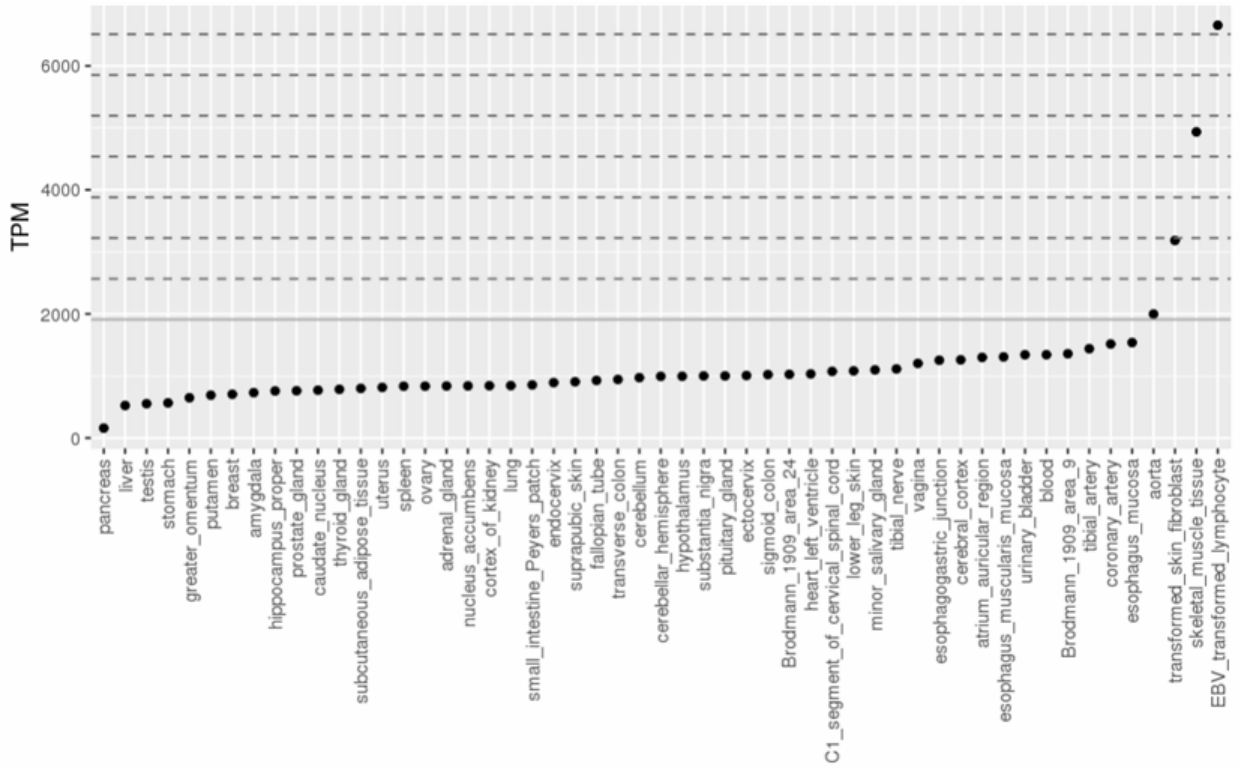

Figure 2. Expression profile for Glyceraldehyde Dehydrogenase (GAPDH) gene. Lines are fences drawn with discrete values of o-score.

This gene is indeed considered expressed in all tissues; however, it has been depicted with distinctive expression as a far outlier in some tissues by our approach.

Evidently, GAPDH has been chosen as an example of a gene being overactive in skeletal muscle tissue besides being expressed in all tissues, but as a component of glycolysis it is not within our list of relevant genes in the mammary gland development. However, when we observed the overactive genes that form this list in all tissues, remarkably only nine are overactive in breast (Figure 3). Actually, from the genes of interest, 11 tissues bear more than nine overactive genes: EBV-transformed lymphocyte (13), blood (10), cerebellar hemisphere (10), cortex of kidney (10), liver (11), lower leg skin (14), small intestine Peyers patch (10), spleen (14), suprapubic skin (14), testis (11) and transformed skin fibroblast (16), showing that genes that are important for mammary gland development do not necessarily have to be overactive in developed breast, and that some other tissues, but not all, may have more genes that contribute with high impact in their regulation and also participate in mammary gland development. As for the tissue-specific genes, breast shows only two genes, CSN1S1 (Alpha-S1-casein) and ELF5 (ETS-related transcription factor Elf-5). Figure 4 presents an example of the tissue-specific expression of Alpha-S1-casein, which plays an important role in the capacity of milk to transport calcium phosphate. It can be noticed that the highest expression is not in breast, but in subcutaneous adipose tissue.

From the list of genes that are implicated in mammary gland development, tissue-specific genes were more frequent in cerebellar hemisphere (3), cerebellum (3), lower leg skin (3), lung (5), minor salivary gland (3), suprapubic skin (3), testis (11) and urinary bladder (3) than in breast. Therefore, the tissue-specific expression of genes implicated in mammary gland development, as for overactive

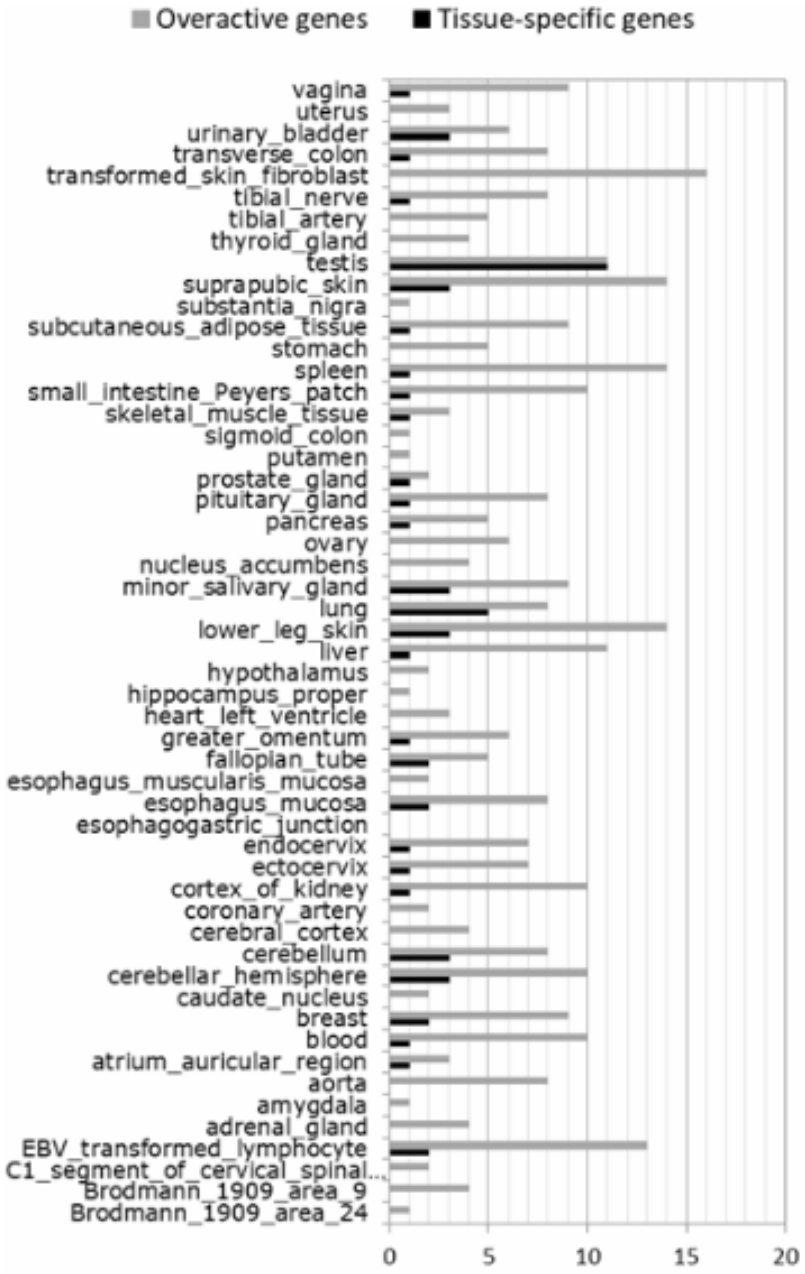

Figure 3. Overactive (o-score > 5) and tissue-specific genes (ts-score > 25%) in the Genotype-Tissue Expression (GTEx) project tissues.

genes, is also not restricted to breast. Summarizing, genes of interest show the highest expression in 19 tissues more frequently as in breast: 11 tissues for overactive and eight for tissue-specific genes, with concordance in four tissues: cerebellar hemisphere, lower leg skin, suprapubic skin and testis.

Gene expression in breast

From the list of genes of interest, only one is highly expressed in breast, the Homeobox protein aristaless-like 4, ALX4, with o-score 6.2 (Figure 5). Thus, contrary to what was expected, distinctive expression in breast was not the rule.

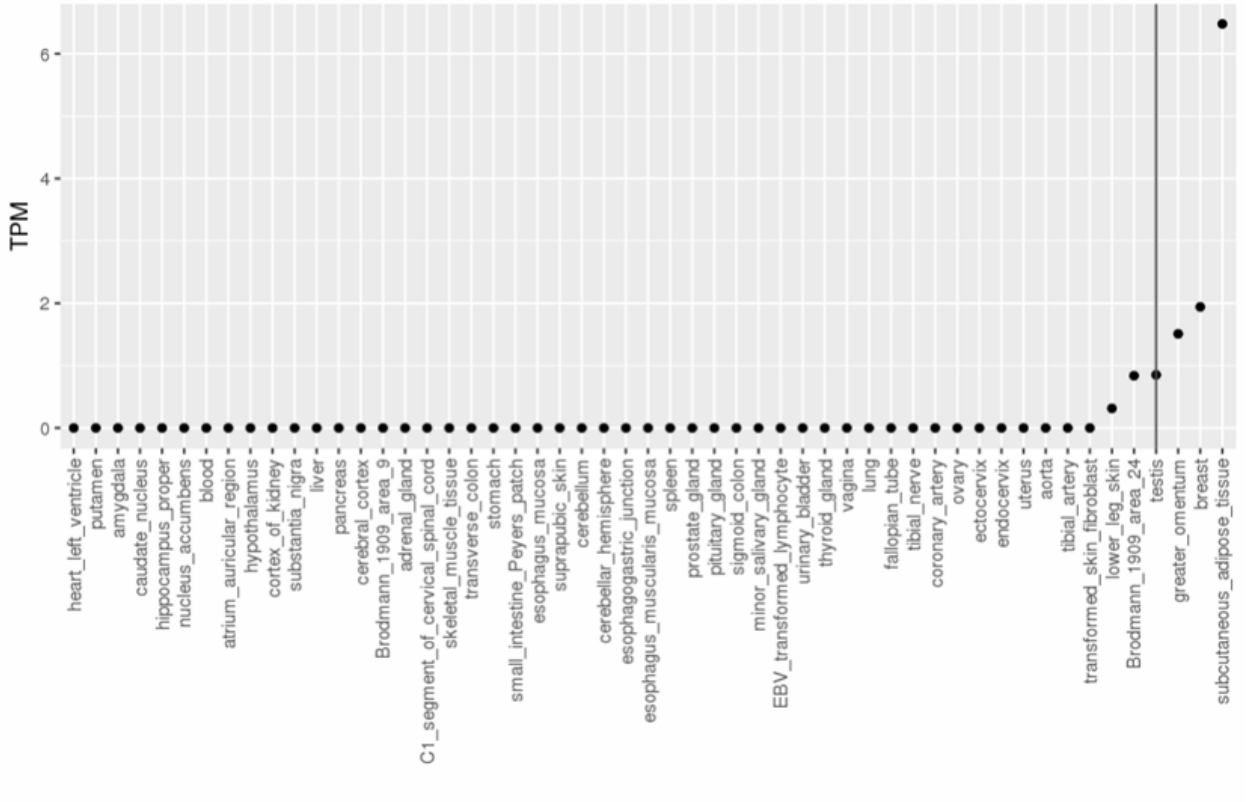

Figure 4. Expression profile for Alpha-S1-casein, CSN1S1 gene, Tissue-specific in greater omentum, breast and subcutaneous adipose tissue (ts-score > 25%). Vertical lane filters out tissues below threshold.

Genes from the mammary gland development list that were overactive in breast comprise: TFAP2C (transcription factor AP-2 gamma), TBX15 (T-box transcription factor TBX15), LPL (lipoprotein lipase), GPAM (mitochondrial glycerol-3-phosphate acyltransferase 1), CD36 (platelet glycoprotein 4), TIMP4 (metalloproteinase inhibitor 4), the already mentioned ALX4 (Homeobox protein aristaless-like 4), MMP3 (Stromelysin-1) and LTF (Lactotransferrin). Tissues that share more overactive genes with breast are: subcutaneous adipose tissue and greater omentum, both sharing the same four genes, LPL, GPAM, CD36 and TIMP4, while subcutaneous adipose tissue also shares with breast TBX15.

On the other hand, two genes from the mammary gland development list have shown tissue-specific expressions: CSN1S1 (Figure 4) and ELF5. CSN1S1 is also TS in greater omentum and subcutaneous adipose tissue and ELF5, in urinary bladder and minor salivary gland.

During mammary gland development [19], in embryonic development there is participation of the overactive genes TBX15 and CD36. Curiously, TBX15 is inhibited by BMP4 in embryo, to dorsoventrally positioning the gland development. CD36 is an important lipid transporter acting at the embryonic stage. In puberty stage four overactive genes participate: TFAP2C, ALX4, TIMP4 and LPL. TFAP2C codifies a transcription factor regulated by estrogen and progesterone which stimulates cellular proliferation in terminal end buds. TIMP4 is a metallopeptidase inhibitor implicated in tissue remodeling. LPL, lipoprotein lipase, acts in the capture of lipids. Moreover in pregnancy and lactation period, LPL and CD36 remain important, with the participation of GPAM, overactive, which catalyzes the conversion of fat acid to lysophosphatidic acid. Furthermore, in Involution period, LTF plays a role against bacterial infection while MMP3 cleaves glycoproteins, being important for tissue remodeling. Both LTF and MMP3 are overactive genes.

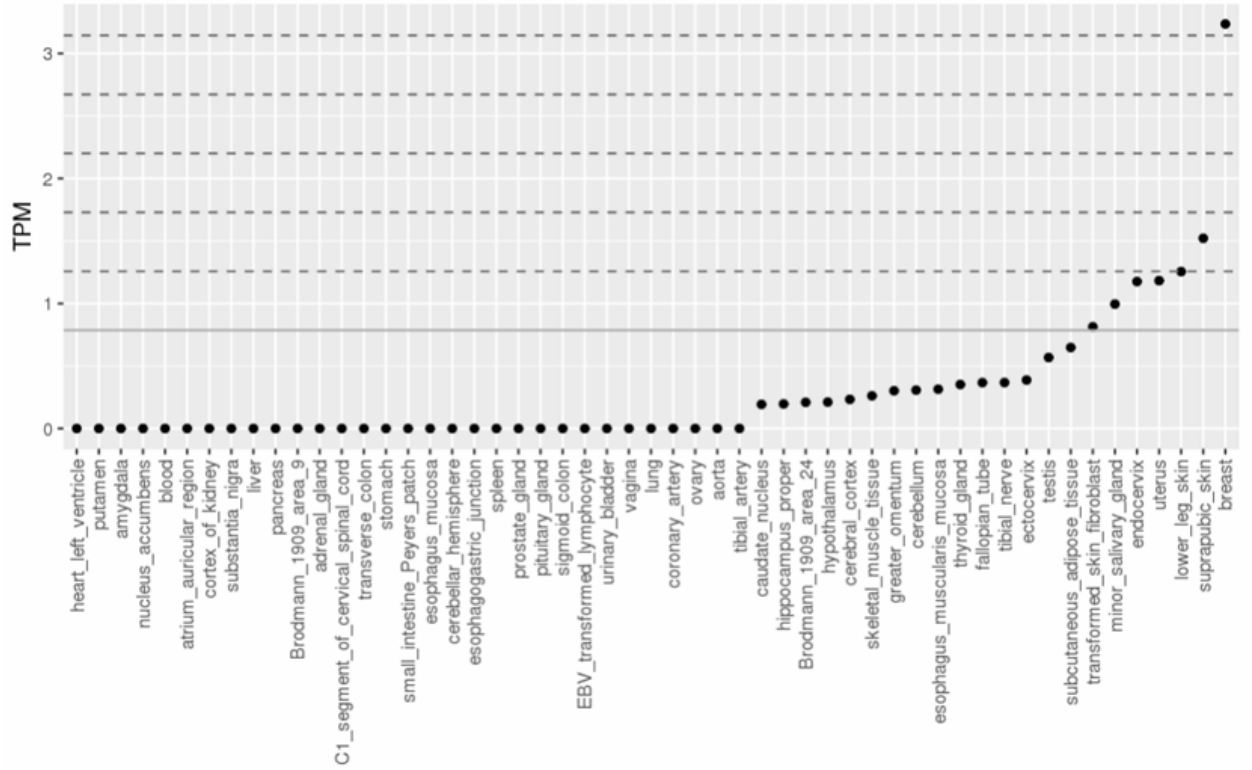

Figure 5. Expression profile for Homeobox protein aristaless-like 4, ALX4 gene. Lines are fences drawn with discrete values of o-score.

The analysis presented here is rather preliminary to explain the reasons for these overactive and tissue-specific genes to be highly differentially expressed in developed breast, but probably the answer comprises sharing of binding sites for transcription factors acting in the specific periods of development, what is currently being investigated.

DISCUSSION

Development drives modifications in cells and tissues by several mechanisms, amongst which gene expression is the main force. Distinct tissues will develop under the expression of key genes. A simple and fast way to list developmental-associated genes is to conduct text mining directed by thematic queries. Here we analyzed a list of 406 genes collected with the focus on participation in mammary gland development. We have introduced the concept of overactive gene as the gene expressed as a far outlier in a boxplot graph. Usually boxplots are presented with a whisker, a bar that goes from the third Quartile (Q3) until the last point beneath the inner fence. Inner fence is calculated by adding to the third Quartile (Q3) a product of 1.5 times the Interquartile Range (IQR). Observations beyond the inner fence are classified as outliers in the statistical literature. Manual inspection of gene expression profiles pointed out that, for some tissues, the expression corresponds to a value far beyond the inner fence. Therefore, we defined a parameter called o-score, which was built to multiply IQR by 1.5 in fence calculation (see 'Materials and Methods' section). This parameter is able to identify several fences and can position the outliers qualitatively. Moreover, the o-score can be calculated for each sample. This parameter allows for characterization of gene expression activities that differ from the majority of other tissues. We named the far outlier as overactive, after the concept of overexpressed gene that is when high expression is produced ectopically. This classification allows for the depicting of tissues in which the expression is highly differentially expressed, although being expressed in all tissues.

Furthermore, genes that are not expressed in most of tissues have been classically called tissue-specific. It is expected that both classes of genes might participate in tissue development. Thus, one can ask, would a list of genes implicated in a developmental biology phenomenon be enriched by overactive and tissue-specific type of genes? Would it be possible that a significant number of genes from this list are more overactive or TS in distinct tissues than the one studied? Would those tissues be functionally related to the tissue that originated the list? These questions have been addressed here, and data points to a different direction. There are tissues in which genes from the list are GOAT (Gene OverActive or Tissue-specific), i.e., either overactive or TS. Sometimes distinct tissues show more GOAT genes than the original tissue of interest. Reasons for this effect must be further investigated and probably involve sharing of transcriptional binding sites in promoters. However, the scenario drawn here points to a combinatorial consortium of specially expressed genes that may be important for other tissues, either being tissue-specific of a tissue and contributing less intensively to the tissue of interest or participating in the category proposed here, the overactive genes, highly differentially expressed in other tissues and acting without high expression in the mammary gland development.

CONCLUSION

A set of gene products is required for the development of the mammary gland. We selected with text-mining tools 406 genes involved in the four periods of breast development: (i) embryonic development, (ii) puberty, (iii) pregnancy and lactation and (iv) regression after lactation. Applying some criteria on the analysis of tissue expression profiles we classified a gene as GOAT (Gene OverActive or Tissue-specific) when it is highly differentially expressed in some tissues. We observed that in most of the cases, genes involved in the mammary gland development, which are GOAT in a subset of tissues, are not GOAT in breast. Thus, development of different tissues might require GOATs that will likewise act as controllers of breast development without being highly differentially expressed to play their role. Our analyses contribute to the understanding of the scenario of developmental control by gene expression, while providing a simple way to point out highly differentially expressed genes in different tissues.

SUPPLEMENTARY DATA

Supplementary figures showing expression profiles of mentioned genes and tables with o-scores and ts-scores for mammary gland development genes are available at our website [20].

ACKNOWLEDGEMENTS

Authors are thankful to the help provided by the members of the EMBL-EBI Expression Atlas: Alvis Brazma, Irene Papatheodorou, Jonathan Manning, Pablo Moreno and Suhaib Mohammed. We also thank Dr. Darren Natale from PIR USA for critically reviewing this manuscript.

FUNDING

This work was supported by Capes Computational Biology BSC network and FAPEMIG.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ABBREVIATIONS

ALX4	:	Homeobox protein aristaless-like 4
CD36	:	Platelet glycoprotein 4
CSN1S1	:	Alpha-S1-casein
EBV	:	Epstein-Barr Virus
ELF5	:	ETS-related transcription factor Elf-5
ESTs	:	Expressed Sequence Tags
GAPDH	:	Glyceraldehyde Dehydrogenase
GOAT	:	Gene OverActive or Tissue-specific
GTEx	:	The Genotype-Tissue Expression
GPAM	:	Mitochondrial glycerol-3-phosphate acyltransferase 1
IQR	:	Interquartile Range
LPL	:	Lipoprotein lipase
LTF	:	Lactotransferrin
MMP3	:	Stromelysin-1
GEO	:	NCBI Gene Expression Omnibus
OA	:	Overactive genes
TS	:	Tissue-Specific genes
Q1	:	First quartile

Q3 : Third quartile
TBX15 : T-box transcription factor TBX15
TFAP2C : Transcription factor AP-2 gamma
TIMP4 : Metalloproteinase inhibitor 4
TPM : Transcripts Per Million

REFERENCES

1. NCBI Resource Coordinators. 2014, Nucleic Acids Res., 42 (Database issue), D7.
2. Schuler, G. D. 1997, J. Mol. Med. (Berl.), 75(10), 694.
3. Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. and Soboleva, A. 2013, Nucleic Acids Res., 41 (Database issue), D991.
4. Edgar, R., Domrachev, M. and Lash, A. E. 2002, Nucleic Acids Res., 30(1), 207.
5. Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A. M., George, N., Fexova, S., Fonseca, N. A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A. F., Jupp, S., Marioni, J., Meyer, K., Petryszak, R., Prada Medina, C. A., Talavera-López, C., Teichmann, S., Vizcaino, J. A. and Brazma, A. 2020, Nucleic Acids Res., 48(D1), D77.
6. Petryszak, R., Fonseca, N. A., Füllgrabe, A., Huerta, L., Keays, M., Tang, Y. A. and Brazma, A. 2017, Bioinformatics, 33(14), 2218.
7. Fonseca, N. A., Marioni, J. and Brazma, A. 2014, PLoS One, 9(9), e107026.
8. Aguet, F., Barbeira, N. A., Bonazzola, R., Brown, A., Castel, E. S., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., Parsana, E. P., Flynn, E., Fresard, L., Gaamzon, R. E., Hamel, R. A., He, Y., Hormozdiari, F., Mohammadi, P., Muñoz-Aguirre, M., Park, Y., Saha, A., Segrć, V. A., Strober, J. B., Wen, X., Wucher, V., Das, S., Garrido-Martín, D., Gay, R. N., Handsaker, E. R., Hoffman, J. P., Kashin, S., Kwong, A., Li, X., MacArthur, D., Rouhana, M. J., Stephens, M., Todres, E., Viñuela, A., Wang, G., Zou, Y., Brown, D. C., Cox, N., Dermitzakis, E., Engelhardt, E. B., Getz, G., Guigo, R., Montgomery, B. S., Stranger, E. B., Im, K. H., Battle, A., Ardlie, G. K. and Lappalainen, T. 2019, The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv, doi:https://doi.org/10.1101/787903.
9. GTEx Consortium. 2013, Nat. Genet., 45(6), 580.
10. Hruz, T., Laule, O., Szabo, G., Wessendorp, F., Bleuler, S., Oertle, L., Widmayer, P., Gruissem, W. and Zimmermann, P. 2008, Adv. Bioinformatics, 2008, 420747.
11. Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V. and Robinson-Rechavi, M. 2008, Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. A. Bairoch, S. Cohen-Boulakia and C. Froidevaux (Eds.), Data Integration in the Life Sciences. DILS 2008, Lecture Notes in Computer Science, Vol. 5109, Springer, Berlin, Heidelberg.
12. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. 2002, Genome Res., 12(6), 996.
13. Wagner, L. and Agarwala, R. 2013, The NCBI Handbook [Internet] 2nd Edition, M. Hoeppner and J. Ostell (Ed.), National Center for Biotechnology Information, Bethesda, 309.
14. Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M. and Andrade-Navarro, M. A. 2009, Nucleic Acids Res., 37 (Web Server issue), W141.
15. Trindade, D., Orsine, L. A., Barbosa-Silva, A., Donnard, E. R. and Ortega, J. M. 2015, Methods, 74, 16.
16. Barbosa-Silva, A., Fontaine, J. F., Donnard, E. R., Stussi, F., Ortega, J. M. and Andrade-Navarro, M. A. 2011, BMC Bioinformatics, 12, 435.
17. Tukey, J. W. 1977, Exploratory Data Analysis, Addison-Wesley, Reading, 39.
18. McGill, R., Tukey, J. W. and Larsen, W. A. 1978, The American Statistician, 32(1), 12.
19. Orsine, L. A. 2016, Elaboration of the mammary gland development pathway and

estimation of the origin of its genes and the system (unpublished master dissertation). Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. http://www.pgbioinfo.icb.ufmg.br/defesas/157M.Pdf

20. Additional data is available at http://biodados.icb.ufmg.br/mammaryglandprofiles/