CanDLE: Illuminating Biases in Transcriptomic Pan-Cancer Diagnosis

Gabriel Mejía(✉), Natasha Bloch, and Pablo Arbelaez

Center for Research and Formation in Artificial Intelligence,
Universidad de los Andes, Bogotá, Colombia
{gm.mejia,pa.arbleaez,n.blochm}@uniandes.edu.co

Abstract. Automatic cancer diagnosis based on RNA-Seq profiles is at the intersection of transcriptome analysis and machine learning. Methods developed for this task could be a valuable support in clinical practice and provide insights into the cancer causal mechanisms. To correctly approach this problem, the largest existing resource (The Cancer Genome Atlas) must be complemented with healthy tissue samples from the Genotype-Tissue Expression project. In this work, we empirically prove that previous approaches to joining these databases suffer from translation biases and correct them using batch z-score normalization. Moreover, we propose CanDLE, a multinomial logistic regression model that achieves state of the art performance in multilabel cancer/healthy tissue type classification (94.1% balanced accuracy) and all-vs-one cancer type detection (78.0% average max F_1).

Keywords: Cancer Classification · Cancer Detection · Machine Learning · Multinomial Logistic Regression · TCGA · GTEx

1 Introduction

Over the last decade, the fast advances in genome sequencing technologies have proven revolutionary, promoting improvements in experimental and high throughput techniques related to transcriptome analysis and bioinformatics [17]. This effort has led to an increase in public RNA-Seq data available to the cancer research community [2]. Consequently, large datasets like The Cancer Genome Atlas (TCGA) [1] have been established, serving as a valuable framework for obtaining standardized and curated genetic expression profiles. This data abundance, combined with the recent success of machine learning in medical applications, makes the automatic cancer diagnosis from transcriptomic samples more plausible than ever before.

There have been numerous approaches to this problem [2,10,12,3,13,15] using a wide range of techniques that go from classic machine learning algorithms (e.g., K-nearest neighbors [10]) to cutting edge deep learning models (e.g., graph convolutional neural networks [15]). However, most of these works are trained and tested exclusively in the TCGA, which has an extremely low number of healthy tissue samples ($\approx$ 7%). Due to this limitation, almost all methods aggregate

healthy samples from different tissues in a single class and solve a multilabel classification of 34 classes (33 cancer types and 1 small healthy class). These experimental conditions are far from what is observed in clinical practice and, consequently, lower the applicability of the results. Moreover, given the significant differences of gene expression profiles between tissues, a tissue classifier will perform adequately in this framework without learning to discriminate between cancer and healthy tissue.

One way to improve the practical usability of the models is to include a comparable and paired amount of healthy samples. For this task, the Genotype-Tissue Expression project (GTEx) is the natural choice since it has a similar scale to the TCGA and captures RNA-Seq samples from non-diseased tissue sites [11]. Aiming to provide a common database, recent works [18,19] apply standardized quantification and normalization to raw data and obtain joint TCGA and GTEx cohorts. Using these resources, novel research has been published performing cancer detection [14], or multi-task cancer type classification [7] and achieving outstanding results.

Although these advances are clear steps in the right direction, two main problems need to be addressed in the current state of the art: (1) there is no empirical proof of the absence of translation biases in the joint datasets, and (2) there is no clear evidence of an important metric improvement associated to the use of cutting edge deep learning techniques compared to simple algorithms. Of these, the first issue is of cornerstone importance because if the origin sources are linearly separable, then the problem would again be in its over-simplified form where it is enough to separate the GTEx and TCGA and then perform tissue classification to obtain great results.

To address both problems, in this work, we first empirically prove the existence of such bias and correct it using batch z-score normalization, which is widely adopted by the machine learning community [16]. And secondly, we propose a simple multinomial logistic regression method to perform multilabel cancer/healthy tissue type classification with sound performance. Our model, which we call CanDLE (Cancer Diagnosis Logistic Engine), cannot only be used for multilabel classification but also highly unbalanced specific cancer type detection.

Our main contributions can be summarized as follows:

1. We empirically prove that previous approaches to joining the GTEx and TCGA databases suffer from significant biases and use a simple batch normalization technique to correct them.
2. We show that a simple method such as multinomial logistic regression can obtain state of the art performance (94.1% balanced accuracy) in multilabel classification of cancerous and healthy tissue types.
3. We demonstrate that CanDLE can detect specific cancer types in highly unbalanced scenarios with state of the art performance (78.0% average max F_1).
4. We exploit the simplicity of our method to perform intuitive and direct gene relevance interpretation in pan-cancer classification.

To ensure the reproducibility of our results, all the resources of this paper are publicly available in https://github.com/g27182818/CanDLE.

2 Related Work

2.1 Joining the TCGA and GTEx Databases

Both the TCGA and GTEx are the gold standard databases for publicly available RNA-Seq profiles of cancerous and healthy tissue, respectively. The TCGA has processed more than $10,000$ samples spanning 33 cancer types, and healthy tissue controls [1], and the GTEx project has collected samples from 54 non-diseased tissue sites across nearly $1,000$ individuals [11]. However, differences in alignment, quantification, and normalization protocols had prevented the use of both databases in joint transcriptomic analyses.

Vivian et al. [18] were the first to propose a unified GTEx-TCGA dataset. They performed standardized alignment with STAR [4] and standardized quantification with RSEM [9]. They finally obtained $18,354$ samples with $60,498$ genes. Later on, Wang et al. [19] expanded this work by imposing more rigid requirements on the input data. They also used STAR alignment and RSEM quantification but applied a quality control stage between the two and added a batch correction method to the quantification output. This last processing step was meant to eliminate the non-biological effects of data sources. They also eliminated the categories that did not have a counterpart in the other database and ended up having $10,366$ valid samples with approximately $19,000$ genes depending on the tissue.

2.2 Classification/Detection Methods

A handful of works have proposed classification algorithms for cancer using transcriptomic data [2,10,12,3,13,15], however, only the studies by Quinn et al. [14] and Hong et al. [7] have taken into account the necessity to add healthy samples from the GTEx project using the Wang et al. dataset. Quinn et al. fit an anomaly detector to GTEx samples, predict any out-of-distribution TCGA sample as cancerous, and report an accuracy of $> 90\%$ in 5/6 of the used tissues. Hong et al. trained two multi-task multilayer perceptrons that classified decease stage, tissue of origin, and neoplastic subclassification in a hierarchical fashion. They used the first $2,000$ principal components of the data as input to their algorithm achieving 99% accuracy in decease state classification, 97% accuracy in tissue of origin classification, and 92% accuracy in neoplastic subclassification.

We compare our classification results with two re-implementations of the Hong et al. model [7]. The original version trained over the first $2,000$ principal components and a second version trained using all genes. To adequate the model to our framework, we implemented just 2 classes (cancer/healthy) in the decease state classification head and lowered the learning rate of the complete feature model by a factor of 100 (for convergence).

We also benchmark our detection results against an adaptation of the original Quinn et al. [14] source code. This detector was trained to detect cancer types instead of healthy tissues in an all-vs-one fashion.

3 Correcting Bias in the Input Data

To formally test for translation biases, we trained a linear support vector classifier (SVC) in both available datasets to predict the data source. The SVCs were trained in 80% of the samples and tested in the remaining 20%. The results can be observed in Table 1. Surprisingly, both data sources (GTEx and TCGA) can be linearly separated in both unified datasets. This observation was expected in the Vivian et al. dataset, as [19] demonstrated the existence of batch effects after standardized quantification, but given that Wang et al. performed batch correction, one would expect the data not to be biased in that case. With these results, the metrics of both the Quinn et al. anomaly detector and the Hong et al. multilayer perceptron lose clinical utility.

Table 1. Weighted average results of a linear support vector classifier predicting the origin of the data (GTEx or TCGA).

Dataset	Normalization	Precision	Recall	F1	Support
Vivian et al. [18]	-	1.00	1.00	1.00	3671
Wang et al. [19]	-	0.99	0.99	0.99	1822
Vivian et al. [18]	Batch	**0.14**	**0.13**	**0.13**	3671

To correct this bias, we performed a simple z-score standardization for each "batch" or origin database. This method centers the data at the origin and imposes a mean of 0 and a standard deviation of 1 for every gene. Considering that the batch correction in Wang et al. did little to un-bias the data and the fact that the Vivian et al. dataset has substantially more samples, we set the former as our working dataframe and perform batch normalization over RSEM $\log_2(TPM + 0.001)$ values. With these changes, a trained SVC (Table 1) could not separate the data sources linearly.

To train and test the models, we filter out $4,892$ genes with no variation over both datasets (standard deviation of 0.0) to remove genes that express exactly the same in all samples. The class distribution of the resulting dataset can be seen in Table 2. Summarizing, we work with $18,354$ samples, $55,602$ genes and make a standard 60/20/20% train/validation/test data partition.

4 Method

4.1 CanDLE

The CanDLE architecture is the simplest approach to the problem using a gradient based method. It is a multinomial logistic regression that given an input

Table 2. Class distribution of the normalized Vivian et al. [18] dataset.

GTEx				TCGA					
Tissue	N	Tissue	N	Class	N	Class	N	Class	N
ADI	517	MUS	396	ACC	77	LIHC	371	UCEC	181
ADR-GLA	131	NER	278	BLCA	407	LUAD	515	UCS	57
BLA	28	OVA	88	BRCA	1098	LUSC	498	UVM	79
BLO	444	PAN	171	CESC	306	MESO	87		
BLO-VSL	606	PIT	107	CHOL	36	OV	427		
BRA	1152	PRO	152	COAD	288	PAAD	179		
BRE	292	SAL-GLA	55	DLBC	47	PCPG	182		
CER	13	SKI	859	ESCA	182	PRAD	496		
COL	359	SMA-INT	92	GBM	165	READ	92		
ESO	666	SPL	100	HNSC	520	SARC	262		
FAL-TUB	5	STO	210	KICH	66	SKCM	469		
HEA	377	TES	165	KIRC	531	STAD	414		
KID	157	THY	338	KIRP	289	TGCT	137		
LIV	169	UTE	91	LAML	243	THCA	512		
LUN	397	VAG	85	LGG	522	THYM	119		

$x \in \mathbb{R}^{n_g}$ computes a probability vector $p \in \mathbb{R}^c$ given by:

$$p = \text{softMax}(Wx), \tag{1}$$

where n_g is the number of considered genes, c is the number of classes of the problem and $W \in \mathbb{R}^{c \times n_g}$ is a learnable weight matrix. To perform multilabel classification c is set to the 63 classes available in the Vivian et al. normalized dataset, and to perform all-vs-one detection of an specific class, c is set to 2.

CanDLE is trained with a cross entropy loss. However, given the unbalanced nature of multilabel classification and all-vs-one detection, we weighted the penalization of each class c_i by a δ_i coefficient given by $\delta_i = B/N_i^2$. Where B is a constant set to 2.5×10^5, and N_i is the number of training samples of class c_i.

4.2 Interpretability

An advantage of the simple CanDLE architecture is its interpretation ease. Consider a logit $l_i = w_i x$ associated to the class c_i and computed for a sample x. Here, w_i is the i^{th} row of W and has one component w_{ij} per gene. Note that the predicted class of x will be the one with the maximum logit. The components of w_i with the highest absolute value are those that influence the most the computation of l_i and, therefore, the classification of x in c_i. To interpret our method, we train CanDLE 100 times with different random partitions and perform a Wald z test [5] for each weight w_{ij} in W ($\alpha = 1 \times 10^{-6}$). We discard non significant weights and use the absolute value of the mean $|\bar{w}_{ij}|$ as a relative importance measure of how much the j^{th} gene influences the prediction of the class c_i.

To obtain a unified list of genes for pan-cancer classification, we select the top $1,000$ genes in each class and order by the number of times that they were selected in any cancer class. We threshold this list at three repetitions (ensuring that chosen genes are important predictors for at least three cancer types) and perform Gene Ontology (GO) biological process enrichment analysis using ShinyGO [6].

Implementation Details: we train our method on an NVIDIA TITAN-X PascalGPU with a learning rate of 10^{-5}, 100 samples per batch and use an Adam [8] optimizer with standard parameters. CanDLE was trained for 20 epochs in multilabel classification and one epoch in all-vs-one detection as it yielded better results. For this same reason, genes with an original standard deviation lower than 0.1 where excluded in the detection task.

5 Results and Discussion

5.1 Mutilabel Classification

Detailed results of cancer and healthy tissue multilabel classification are shown in Table 3. CanDLE is capable of achieving a state of the art performance of 94.1% (test) and 92.0% (validation) balanced accuracy outperforming both versions of the Hong et al. [7] model by a large margin. It obtained an absolute difference of +7.3% (test) and +5.6% (validation) in this metric when compared to the complete feature re-implementation. These results prove that a simple method can correctly discriminate the transcriptomic signatures of all healthy tissue and cancer types even when the biases are removed. Additionally, we observed that, although the Hong et al. model is more complex and flexible, the fact that it performs multiple predictions for a single sample makes the method prone to performing chained errors (i.e. correctly predicting cancerous and colon tissue but erroneously give a kidney cancer subtype classification). Also note that in a clinical setting the multiple predictions offer no benefit over a direct classification performed by CanDLE.

Table 3. Multilabel classification results on the normalized Vivian *et al* [18] dataset. mACC: balanced accuracy, ACC: accuracy, mAP: mean average precision, PCA: principal component analysis, CF: complete features.

	Validation			Test		
Model	mACC	ACC	mAP	mACC	ACC	mAP
Hong et al. PCA [7]	64.0	65.3	-	61.7	64.9	-
Hong et al. CF[7]	86.4	84.6	-	86.8	84.4	-
CanDLE w/o weights	91.2	**96.0**	**95.3**	90.9	**95.7**	**96.6**
CanDLE	**92.0**	95.6	94.2	**94.1**	95.6	95.0

As expected, adding a weighted loss function improved the results in terms of balanced accuracy (Table 3) by +3.2% (test) and +0.8% (validation). This behavior also implied a slight decrease in both total accuracy and mean average precision since the inclusion of weights prioritized a great performance in each class over bulk correct predictions. Such observations suggest that addressing class unbalance using loss weights is an effective technique in the current framework.

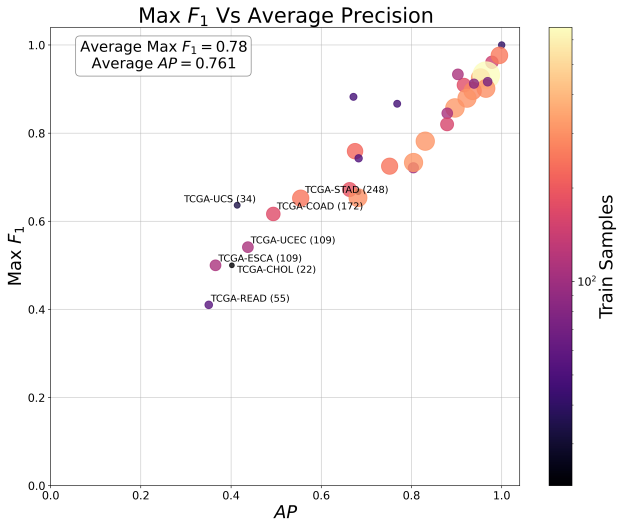

Fig. 1. CanDLE detection max F_1 Vs. AP summary plot for all cancer types over the normalized Vivian et al. [18] validation set. The color and size of each point correspond to the number of training samples available. Specific cancer identifiers are shown for all classes with max F_1 or AP below 0.6. Brackets contain the number of training samples.

5.2 Detection

A summary plot of the all-vs-one CanDLE detection experiments can be seen in Figure 1. Interestingly, digestive cancers were the hardest to detect. But, not surprisingly, the worst performance was observed in underrepresented cancer types (READ, CHOL, ESCA, UCEC, UCS, COAD, and STAD). This observation indicates that access to a bigger number of training samples generally helps to obtain better results. However, it is outstanding that CanDLE achieves a mean max F_1 and average AP of 78.0% and 76.1% respectively, considering that in some cases, the number of positive samples was extremely low (e.g., 22 training samples for CHOL). This is especially relevant in clinical practice, where highly unbalanced detection is common.

The adaptation of the Quinn et al. [14] anomaly detector obtained a mean max F_1 of 77.7% and an average AP of 77.2%, making the performance of Can-

DLE state of the art in mean max F_1 and competitive in average AP. These results are particularly important considering that the Quinn et al. model was explicitly designed for detection tasks while CanDLE is a flexible and simple architecture that can also perform multilabel classification. Moreover, CanDLE has the advantage of providing more transparent and direct interpretability of detection models when compared with the Quinn et al. algorithm.

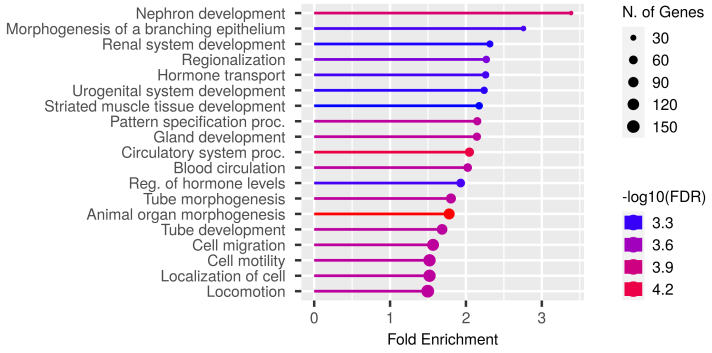

Fig. 2. Biological processes gene ontology results of the top 1,000 predictor genes shared by at least 3 TCGA classes. Developmental and morphogenesis processes appear to have a mayor role.

5.3 Interpretability

After performing our interpretability protocol, we obtained an ordered list of 1,982 genes. Notably, the majority of the weights (71.3%) resulted significant in the Wald z test highlighting the reliability of CanDLE. The genes YWHAQ, AC073578, AC013417, and RNU6-1207P were found to be important predictors in more than 10 cancer types. Figure 2 shows the results of GO biological process enrichment analysis where it is clear that CanDLE recognizes developmental and morphogenesis pathways as essential to perform pan-cancer classification.

6 Conclusion

In this work, we empirically prove that previous approaches to joining the TCGA and GTEx databases have significant biases and correct them with a z-score batch standardization. Additionally, we present CanDLE, a simple multinomial logistic regression method that can perform both cancer/healthy tissue type multilabel classification and all-vs-one detection with state of the art performance. Finally, we leveraged the simplicity of CanDLE to interpret gene relevance in pan-cancer classification which recognized developmental and morphogenesis pathways as important predictors.

Acknowledgements GM acknowledges the support of a UniAndes-DeepMind Scholarship 2022. We also acknowledge the valuable help of Camilo Becerra in graphics and tables preparation, and Danniel Moreno for useful discussions and feedback.

References

1. The Cancer Genome Atlas Program - National Cancer Institute, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
2. Ahn, T., Goo, T., hee Lee, C., Kim, S., Han, K., Park, S., Park, T.: Deep learning-based identification of cancer or normal tissue using gene expression data. pp. 1748–1752. IEEE (12 2018). https://doi.org/10.1109/BIBM.2018.8621108, https://ieeexplore.ieee.org/document/8621108/
3. Chen, H.I.H., Chiu, Y.C., Zhang, T., Zhang, S., Huang, Y., Chen, Y.: Gsae: an autoencoder with embedded gene-set nodes for genomics functional characterization. BMC Systems Biology 2018 12:8 **12**, 45–57 (12 2018). https://doi.org/10.1186/S12918-018-0642-2, https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-018-0642-2
4. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: Star: ultrafast universal rna-seq aligner. Bioinformatics **29**, 15–21 (1 2013). https://doi.org/10.1093/bioinformatics/bts635, https://academic.oup.com/bioinformatics/article-lookup/doi/10.109
5. Fávero, L.P., Belfiore, P.: Binary and multinomial logistic regression models (2019). https://doi.org/10.1016/B978-0-12-811216-8.00014-8, https://linkinghub.elsevier.com/retrieve/pii/B9780128112168000148
6. Ge, S.X., Jung, D., Yao, R.: Shinygo: a graphical gene-set enrichment tool for animals and plants. Bioinformatics **36**, 2628–2629 (4 2020). https://doi.org/10.1093/bioinformatics/btz931, https://academic.oup.com/bioinformatics/article/36/8/2628/5688742
7. Hong, J., Hachem, L.D., Fehlings, M.G.: A deep learning model to classify neoplastic state and tissue origin from transcriptomic data. Scientific Reports **12**, 9669 (12 2022). https://doi.org/10.1038/s41598-022-13665-5, https://www.nature.com/articles/s41598-022-13665-5
8. Kingma, D.P., Ba, J.L.: Adam: A Method for Stochastic Optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings (dec 2014), https://arxiv.org/abs/1412.6980v9
9. Li, B., Dewey, C.N.: Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC Bioinformatics **12**, 323 (12 2011). https://doi.org/10.1186/1471-2105-12-323, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323
10. Li, Y., Kang, K., Krahn, J.M., Croutwater, N., Lee, K., Umbach, D.M., Li, L.: A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. BMC Genomics 2017 18:1 **18**, 1–13 (7 2017). https://doi.org/10.1186/S12864-017-3906-0, https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3906-0
11. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim,

D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C., Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F.: The genotype-tissue expression (gtex) project. Nature Genetics **45**, 580–585 (6 2013). https://doi.org/10.1038/ng.2653, http://www.nature.com/articles/ng.2653

12. Lyu, B., Haque, A.: Deep learning based tumor type classification using gene expression data. bioRxiv p. 364323 (7 2018). https://doi.org/10.1101/364323, https://www.biorxiv.org/content/10.1101/364323v1
13. Mostavi, M., Chiu, Y.C., Huang, Y., Chen, Y.: Convolutional neural network models for cancer type prediction based on gene expression. BMC Medical Genomics 2020 13:5 **13**, 1–13 (4 2020). https://doi.org/10.1186/S12920-020-0677-2, https://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-020-0677-2
14. Quinn, T.P., Nguyen, T., Lee, S.C., Venkatesh, S.: Cancer as a tissue anomaly: Classifying tumor transcriptomes based only on healthy data. Frontiers in Genetics **10**, 599 (7 2019). https://doi.org/10.3389/fgene.2019.00599, https://www.frontiersin.org/article/10.3389/fgene.2019.00599/full
15. Ramirez, R., Chiu, Y.C., Hererra, A., Mostavi, M., Ramirez, J., Chen, Y., Huang, Y., Jin, Y.F.: Classification of cancer types using graph convolutional neural networks. Frontiers in Physics **8**, 1–14 (2020). https://doi.org/10.3389/fphy.2020.00203
16. Singh, D., Singh, B.: Investigating the impact of data normalization on classification performance. Applied Soft Computing **97**, 105524 (12 2020). https://doi.org/10.1016/j.asoc.2019.105524, https://linkinghub.elsevier.com/retrieve/pii/S1568494619302947
17. Tripathi, R., Sharma, P., Chakraborty, P., Varadwaj, P.K.: Next-generation sequencing revolution through big data analytics. Frontiers in Life Science **9**, 119–149 (4 2016). https://doi.org/10.1080/21553769.2016.1178180, http://www.tandfonline.com/doi/full/10.1080/21553769.2016.1178180
18. Vivian, J., Rao, A.A., Nothaft, F.A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A.D., Musselman-Brown, A., Schmidt, H., Amstutz, P., Craft, B., Goldman, M., Rosenbloom, K., Cline, M., O'Connor, B., Hanna, M., Birger, C., Kent, W.J., Patterson, D.A., Joseph, A.D., Zhu, J., Zaranek, S., Getz, G., Haussler, D., Paten, B.: Toil enables reproducible, open source, big biomedical data analyses. Nature Biotechnology **35**, 314–316 (4 2017). https://doi.org/10.1038/nbt.3772, http://www.nature.com/articles/nbt.3772
19. Wang, Q., Armenia, J., Zhang, C., Penson, A.V., Reznik, E., Zhang, L., Minet, T., Ochoa, A., Gross, B.E., Iacobuzio-Donahue, C.A., Betel, D., Taylor, B.S., Gao, J.,

Schultz, N.: Unifying cancer and normal rna sequencing data from different sources. Scientific Data **5**, 180061 (12 2018). https://doi.org/10.1038/sdata.2018.61, http://www.nature.com/articles/sdata201861