Quantifying Intratumor Heterogeneity by Key Genes Selected Using Concrete Autoencoder

Raihanul Bari Tanvir(✉), Ricardo Ruiz, Samuel Ebert, Masrur Sobhan, Abdullah Al Mamun, and Ananda Mohan Mondal

Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA
{rtanv003,rruiz101,seber007,msobh002,mmamu009,amondal}@fiu.edu

Abstract. The tumor cell population in cancer tissue has distinct molecular characteristics and exhibits different phenotypes, thus, resulting in different subpopulations. This phenomenon is known as Intratumor Heterogeneity (ITH), a major contributor to drug resistance, poor prognosis, etc. Therefore, quantifying the levels of ITH in cancer patients is essential, and many algorithms do so in different ways, using different types of omics data. DEPTH2 algorithm utilizes transcriptomic data to assess ITH scores and exhibits promising performance. However, it quantifies ITH using all genes, limiting the identification of ITH-related prognostic genes. We hypothesize that a subset of key genes is sufficient to quantify the ITH level, and this subset of key genes could be ITH-related prognostic genes. To prove our hypothesis, we propose an unsupervised deep learning-based framework using Concrete Autoencoder (CAE) to select a subset of cancer-specific key genes for ITH evaluation. For the experiment, we used gene expression profile data of breast, kidney, and lung cancer tumor cohorts from the TCGA repository. Multi-run CAE identified three sets of key genes for each cancer cohort. Comparing ITH scores derived from all genes and CAE-selected key genes showed similar prognostic outcomes. Subtypes of lung cancer displayed consistent ITH distributions for both gene sets. Based on these observations, it can be concluded that a subset of key genes, instead of all, is sufficient for ITH quantification. Our results also showed that many key genes are prognostically significant and can be used as therapeutic targets.

Keywords: Concrete Autoencoder · Deep Learning · Gene Expression · Intratumor Heterogeneity · ITH

1 Introduction

Intratumor Heterogeneity (ITH) refers to different types of tumor cell subpopulations within a tumor [1]. Even though these cell subpopulations have the same origin (tumor tissue, patient), they exhibit different phenotypes and molecular characteristics. ITH is one of the main challenges for targeted cancer therapy, as the difference in tumor cells and their microenvironments makes it harder for targeted cancer therapy to eradicate cancer cells [2, 3]. Therefore, an accurate assessment of ITH is essential to understand

P. Maji et al. (Eds.): PReMI 2023, LNCS 14301, pp. 844–852, 2023.
https://doi.org/10.1007/978-3-031-45170-6_88

the tumor dynamics and the development of effective and durable therapeutic strategies. ITH causes can vary depending on different levels, such as the genome, epigenome, transcriptome, etc. [4]. For example, reduced DNA damage mechanisms, microenvironmental factors (hypoxia, acidosis, etc.) [5], subclonal evolution [2], etc., contribute to ITH at the genomic level. The methylation of tumor suppressor genes is an example of ITH at the epigenomic level [6]. Different gene expression patterns contribute to ITH at the transcriptome level, which is observed to mirror ITH at the genomic or epigenomic level or both [5, 7]. This makes transcriptomic data suitable for quantifying ITH.

Different algorithms for quantifying ITH exist, such as ABSOLUTE [8], MATH [9], EXPANDS [10], and PhyloWGS [11]. These algorithms use genomic data, such as – copy number alterations (CNA), somatic mutation profiles, etc. Some algorithms take advantage of transcriptome profile that mirrors ITH at the genomic and epigenomic level, such as – tITH [12], sITH [13], DEPTH [14], and DEPTH2 [15]. In contrast to other ITH evaluation techniques, such as DEPTH and others, the DEPTH2 method assesses ITH independently of normal controls. This implies that it can be utilized for all tumor gene expression profiles regardless of the availability of corresponding normal samples' gene expression data. tITH requires protein-protein interaction (PPI) network along with gene expression data. Unlike tITH, DEPTH2 calculates the ITH score using only gene expression data.

Though the DEPTH2 method is statistically sound, the drawbacks are- (i) it uses expression values of all genes (~20,000) in calculating the ITH score and (ii) it cannot guide finding the prognostically significant genes. We argue that not all genes are related to ITH, and a subset of key genes is sufficient to calculate the ITH score at the transcriptome level.

This study presents a deep learning-based computational framework that utilizes an unsupervised concrete autoencoder (CAE) to identify key genes for quantifying Intratumor Heterogeneity (ITH). The framework selects a subset of key genes from Breast Invasive Carcinoma (BRCA), Kidney Renal Carcinoma (KIRC), and Lung Adenocarcinoma (LUAD) using expression profile data from the TCGA repository. The ITH scores are then calculated using all genes and the selected key genes. The results demonstrate that using the subset of 100 key genes outperforms all ~20,000 genes in terms of survival and prognostic outcomes for the three cancer types. The key genes exhibit consistent levels of ITH across cancer subtypes and show potential as prognostic markers and therapeutic targets. This study highlights the effectiveness of a reduced set of key genes in quantifying ITH at the transcriptome level. The overall framework is depicted in Supplementary Fig. S1.

2 Materials and Methods

2.1 Dataset Collection

We collected gene expression datasets of BRCA,LUAD, and KIRC cancers from the UCSC Xena Browser database [16]. Each dataset contains expression profiles of 20,530 mRNAs. The number of tumor samples for each cancer type was as follows: BRCA (1097 samples), LUAD (533 samples), and KIRC (517 samples).

2.2 Concrete Autoencoder to Select Cancer-Specific Key Genes

Concrete Autoencoder (CAE) [17], an unsupervised deep learning approach is used to identify cancer-specific key genes. CAE identifies features most informative for a given dataset [18–23]. CAE differs from the standard Autoencoder in the encoder part, where CAE employs a concrete selector layer (See Fig. S2). This selector layer is based on Concrete distribution [24], a relaxed variant of discrete distribution. Unlike the encoder part of the CAE, the decoder part resembles closely with the standard Autoencoder. The selector layer is used to incorporate discrete distribution into deep learning algorithms. For example, CAE uses it to learn a subset of the most informative features and produce minimum reconstruction error. In the learning phase, the selector layer learns a subset of features, which depends on a hyperparameter called Temperature (T), which is gradually lowered during the training phase to a low value using a simple annealing scheduling. This gradual decrease in temperature helps the concrete distribution to learn and select a definite subset of features [17]. In the selector layer, each unit selects a unique feature with the highest probability from the original feature space. Thus, CAE selects the most informative subset of features, and the reconstruction of the original feature space using the selected subset of features produces minimum reconstruction error. In the original Autoencoder, the features learned at the encoder part are latent features, whereas those learned at CAE are actual features. CAE was trained on each gene expression data of BRCA, KIRC, and LUAD, and 100 features were selected in each run. While training CAE, the dataset was divided randomly into 80/20 split for training and testing. Details of hyperparameter tuning are in Table S1.

2.3 Training CAE

Figure S3 shows the characteristics curve for CAE or an instance of the training behaviors of CAE for the LUAD dataset. The hyperparameter, Temperature (T), was reduced using a simple annealing schedule from 10 to 0.1 from the start epoch to the last. The reconstruction errors (loss) for the training and validation sets are plotted using blue and red curves, respectively. It shows that both errors were relatively high during the early training phase, as expected, and both reached a minimum plateau at the end. Also, the mean-max probability finally approaches 1.0 (yellow curve). The CAE was implemented using Keras (https://keras.io/). Experiments were conducted on the high-performance cluster with NVIDIA Quatro K620 GPU with 384 cores and 2 GB memory devices.

2.4 ITH Level Estimation Method

To calculate the Intratumor Heterogeneity (ITH) score, we used a scoring method named - Deviating Gene Expression Profiling Tumor Heterogeneity, or DEPTH2 in short [15], defined as –

$$\sqrt{\frac{\sum_{i=1}^{m}\left(z(ex(G_i,T)) - \frac{1}{m}\sum_{j=1}^{m} z(ex(G_j,T))\right)^2}{m-1}} \tag{1}$$

where,

$$z(ex(G_i, T) = \frac{\left|(ex(G_i, T) - \frac{1}{t}\sum_{j=1}^{t} ex(G_i, TS_j)\right|}{SD_i} \tag{2}$$

and,

$$SD_i = \sqrt{\frac{\sum_j^t \left((ex(G_i, T) - \frac{1}{t}\sum_{j=1}^{t} ex(G_i, TS_j)\right)^2}{t-1}} \tag{3}$$

where T is the tumor sample for which the score is being calculated. G_i is the i-th gene, and m is the number of genes. $ex(G_i, T)$ expression of gene G_i in sample T. It assigns a score to each patient. It is based on standard deviations of the z-score of the gene expression value variations. If a tumor displays similar z-scored expression values across most genes, it will have a low DEPTH2 score and a lower ITH level. In contrast, if there is variation in gene expression alterations, the tumor will receive a higher DEPTH2 score. This score indicates how much the gene expressions deviate from the norm for all tumors and genes within the matrix. We calculated the ITH score for each cancer patient of BRCA, KIRC, and LUAD employing DEPTH2 using two sets of genes. One score uses all the genes, and the other uses only the key genes selected by multi-run CAE.

2.5 Survival Analysis

Survival Analysis was performed to check whether two groups of patients based on high and low ITH scores are significantly distinguishable in prognosis. In our analysis, the event of interest is the death of cancer patients.

Survival Analysis Based on ITH Scores: Samples were sorted in descending order of the ITH score, and then the top and bottom of the total samples were taken as two groups. This analysis compared the prognostic importance of ITH scores derived using all genes and key genes (our study).

Survival Analysis Based on Each Key Gene: The cohort was divided into two groups based on the median gene expression values. This survival analysis helped identify prognostically significant genes.

After forming two groups, the Kaplan-Meier curves were plotted, and the Log-rank test was performed to check the statistical significance of the difference in survival function.

3 Results and Discussion

3.1 Multi-run CAE to Select Key Genes

Due to the stochastic nature of CAE, the model was trained ten times, and in each run, 100 features were selected for each cancer cohort - BRCA, KIRC, and LUAD. Figure 1(a) shows the stochastic nature of CAE since only 16 genes are common between three

single-run CAE. In the case of 10-run CAE, the top 100 features were selected from the combined list sorted in descending order based on the frequency of appearance of a feature in 10 runs. It is clear from Fig. 1(b) that there are 53 genes common between three batches of 10 runs, which is more than the common genes (16 genes) in three single runs. Thus, a multi-run approach was adopted to select the robust set of features.

The top 100 frequent features were chosen to select the key features based on the assumption that the more frequent a feature in different runs, the more informative the feature is. The combined lists of features derived from 10-run CAE consist of 469, 527, and 435 genes for BRCA, KIRC, and LUAD, respectively. The frequency range of the top 100 features is 3 to 10 for each cancer cohort, which means that the most frequent features appeared in all ten runs, and the least frequent one appeared in 3 runs.

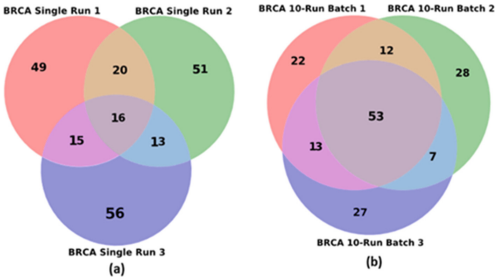

Fig. 1. Selecting the robust set of features. (a) Venn diagram of three sets of 100 genes from three single-run CAE; CAE produces only 16 features common between single runs. (b) three sets of most frequent 100 features from 10-run CAE. 10-run CAE produces more features (53 genes) common between three batches of runs. Thus, multi-run CAE produces a robust set of features.

3.2 Multi-run CAE Selects Cancer-specific Genes

We investigated whether there were any common genes between two sets or among the three sets of key genes derived from three cancers, shown by the Venn diagram in Fig. S4. It shows that there is no common gene between the three gene sets. However, a few genes are common between each pair of gene sets: 5 between BRCA and LUAD, 3 between KIRC and BRCA, and 3 between KIRC and LUAD. Since the size of each set is 100 and there are only a few genes common between two sets and none between the three sets, thus, the key gene sets are cancer-specific.

3.3 All Genes vs. Key Genes in ITH Scoring: Whole Cancer Cohorts

We compared the ITH scores calculated for BRCA, KIRC, and LUAD cohorts using two different sets of genes: (i) DEPTH2 score calculated using all genes and (ii) DEPTH2 score calculated using only the key genes selected by the multi-run CAE system (our work). Survival analysis is used to compare the two ITH scores. Figure 2 presents the results of survival analyses, Kaplan Meier plots, for cancer cohort - BRCA based on ITH scores derived from all genes (Fig. 2a) and key genes (Fig. 2b).

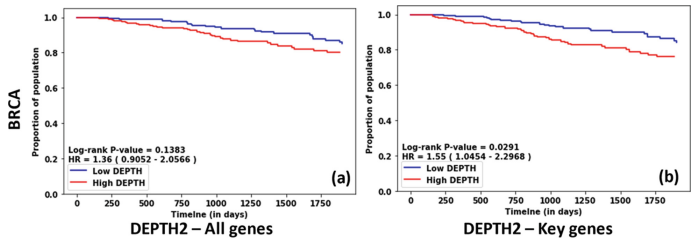

Fig. 2. Survival analysis of BRCA cohort. Kaplan Meier plots based on DEPTH2 score calculated using all genes (a) and key genes (b).

It is evident from Kaplan Meier plots that high DEPTH2 scores are related to poor prognosis, and low DEPTH2 scores have a higher chance of survival.

Survival analysis of BRCA showed a P-value of 0.1383 (not significant) and Hazard Ratio (HR) of 1.36 using all genes (Fig. 2a), while key genes produced a significant result with a P-value of 0.0291 and HR of 1.55 (Fig. 2b). The latter case is prognostically significant (P-value ≤ 0.05) compared to the former, thus validating our claim.

Similarly, better results were found using key genes than all genes both in LUAD (P-value: 0.0019 vs. 0.109; HR: 1.79 vs. 1.34) and KIRC (P-value: 5.18e−07 vs. 0.0018, HR: 2.67 vs. 1.78), as shown in Fig. S5a–b and S5c–d.

Our investigation showed that 100 key genes produced better results than all genes (~20,000) in three types of cancers. Thus, we do not need all genes to evaluate the ITH scores.

3.4 All Genes vs. Key Genes in ITH Scoring: LUAD Subtypes

In this section, we show the comparison of ITH scores (DEPTH2) for LUAD subtypes calculated using all genes versus key genes. Of 435 LUAD patients, 55, 34, and 54 are labeled as Terminal Respiratory Unit (TRU), Proximal Proliferation (PP), and Proximal Inflammation (PI), respectively. The remaining patients did not have any subtype-based labels. This molecular subtyping was done in [25]. It is evident from survival analysis that the TRU subtype is prognostically favorable and has a higher chance of survival than the PI and PP subtypes combined (Fig. S6).

Figure 3 shows the ITH score distribution for three subtypes, using all genes and key genes. Min-max normalization on DEPTH2 scores was performed to bring the distribution to the same scale. It is seen that the subtype TRU has comparatively lower values in ITH score than other subtypes, which supports the higher chance of survival for the TRU subtype than PI and PP combined (Fig. S6). It is also clear that the distribution of ITH scores for three subtypes remained the same for all genes and key genes.

We performed correlation analysis to compare the distribution of the DEPTH2 scores using all genes and key genes, and the results are shown in Table S2. It is observed that there is a relatively high correlation between DEPTH2 scores of each subtype of LUAD cancer using all and key genes. It is clear from the P-values (ns: not significant) in Fig. 3 that the two scores for each subtype derived using all genes and key genes are not significantly different. Both all genes and key genes produced the same level of

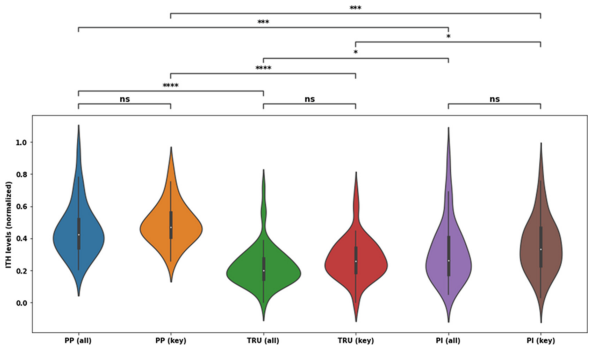

Fig. 3. Comparison of ITH scores of three molecular subtypes of LUAD using all genes (labeled as 'all') and key genes (labeled as 'key'). Distribution of Min-max normalized ITH score (DEPTH2) in three molecular subtypes of LUAD in violin plots. The Mann-Whitney-Wilcoxon test between two distributions was performed, and stars marked the p-value significance. P-value annotation legend: ns (not significant): $0.05 < p \leq 1$, *: $0.01 < p \leq 0.05$, **: $0.001 < p \leq 0.01$, ***: $0.0001 < p \leq 0.001$, ****: $p \leq 0.0001$

difference in ITH between two subtypes. For example, PP and TRU (****), TRU and PI (*), and PP and PI (***).

Based on these observations, we do not need ~20,000 genes to calculate the ITH score; only 100 key genes will suffice.

3.5 Survival Analysis of Key Genes

Survival analysis was performed on each key gene from their respective cancer cohort to identify whether they possess prognostic capabilities. Figure S5 shows the forest plot of the prognostically significant genes and the summary of survival analyses in terms of Logrank P-value and Hazard Ratio with a 95% confidence interval. The thresholds for prognostically significant genes are Logrank P-value ≤ 0.05 and Hazard Ratio, HR $\neq 1$. Of 100 key genes for BRCA, 15 were prognostically significant, as shown in the forest plot in Fig. S7. Similarly, for KIRC and LUAD, 30 and 61 genes were prognostically significant. The list of genes with prognostically significant genes marked as bold is given in Table S3.

4 Conclusion and Future Direction

This study proposes that a subset of key genes instead of all genes (~20,000) is adequate for evaluating the ITH scores of individual tumors. To test this hypothesis, a computational framework was developed using a multi-run concrete autoencoder to select the key genes from gene expression profile data. Results showed that using only the selected 100 key genes instead of all ~20,000 genes produced better survival and prognostic outcomes for three cancers (BRCA, KIRC, and LUAD). Our investigation showed that key genes produce the same levels of ITH at the cancer subtype levels. We also showed that many of these key genes are prognostically significant, which can be investigated further as

possible therapeutic targets. This study concludes that a subset of key genes is sufficient to quantify the ITH at the transcriptome level.

However, this study has its limitations. The intratumor heterogeneity (ITH) is determined by genetic and epigenetic variation within an individual's tumor. The transcriptome reflects both types of heterogeneity, meaning that a unique set of genes may dictate ITH for each patient. However, our study used the same key genes to assess ITH across all patients for a specific type of tumor, which presents a limitation. The selection of 10 runs in multi-run CAE was arbitrary and may not be optimal for identifying a stable set of features for BRCA, KIRC, and LUAD cohorts. Despite these limitations, the study demonstrated that a short list of key genes is effective in assessing ITH levels. In future research, we will extend this study to determine the ideal number of runs needed to select a reliable feature set across different cohorts using multi-run CAE. Additionally, we aim to create an approach that identifies patient-specific key genes for evaluating ITH.

Acknowledgment. This work has been partially supported by the NSF CAREER Award #1901628.

Supplementary Materials. The supplementary materials are available at GitHub. https://github.com/mldag/ITH-Key-Genes-mrCAE/blob/main/supplementary.pdf.

References

1. Jamal-Hanjani, M., Quezada, S.A., Larkin, J., Swanton, C.: Translational implications of tumor heterogeneity. Clin. Cancer Res. **21**, 1258–1266 (2015)
2. Qazi, M.A., et al.: Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. Ann. Oncol. **28**(7), 1448–1456 (2017). https://doi.org/10.1093/annonc/mdx169
3. Reinartz, R., et al.: Functional Subclone profiling for prediction of treatment-induced intratumor population shifts and discovery of rational drug combinations in human glioblastoma. Clin. Cancer Res. **23**(2), 562–574 (2017). https://doi.org/10.1158/1078-0432.CCR-15-2089
4. Grzywa, T.M., Paskal, W., Włodarski, P.K.: Intratumor and intertumor heterogeneity in melanoma. Transl. Oncol. **10**(6), 956–975 (2017). https://doi.org/10.1016/j.tranon.2017.09.007
5. Gillies, R.J., Verduzco, D., Gatenby, R.A.: Evolutionary dynamics of carcinogenesis and why targeted therapy does not work. Nat. Rev. Cancer **12**(7), 487–493 (2012). https://doi.org/10.1038/nrc3298
6. Takamizawa, J., et al.: Reduced expression of the let-7 MicroRNAs in human lung cancers in association with shortened postoperative survival. Cancer Res. **64**(11), 3753–3756 (2004)
7. Sigalotti, L., et al.: Intratumor heterogeneity of cancer/testis antigens expression in human cutaneous melanoma is methylation-regulated and functionally reverted by 5-Aza-2′-deoxycytidine. Cancer Res. **64**(24), 9167–9171 (2004). https://doi.org/10.1158/0008-5472.CAN-04-1442
8. Carter, S.L., et al.: Absolute quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. **30**(5), 413–421 (2012). https://doi.org/10.1038/nbt.2203
9. Mroz, E.A., Rocco, J.W.: MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. Oral Oncol. **49**(3), 211–215 (2013). https://doi.org/10.1016/j.oraloncology.2012.09.007

10. Andor, N., Harness, J.V., Müller, S., Mewes, H.W., Petritsch, C.: Expands: expanding ploidy and allele frequency on nested subpopulations. Bioinformatics **30**(1), 50–60 (2014). https://doi.org/10.1093/bioinformatics/btt622
11. Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., Morris, Q.: PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. **16**, 1–20 (2015)
12. Park, Y., Lim, S., Nam, J.-W., Kim, S.: Measuring intratumor heterogeneity by network entropy using RNA-seq data. Sci. Rep. **6**(1), 37767 (2016). https://doi.org/10.1038/srep37767
13. Kim, M., Lee, S., Lim, S., Kim, S.: SpliceHetero: an information theoretic approach for measuring spliceomic intratumor heterogeneity from bulk tumor RNA-seq. PLoS ONE **14**(10), e0223520 (2019). https://doi.org/10.1371/journal.pone.0223520
14. Li, M., Zhang, Z., Li, L., Wang, X.: An algorithm to quantify intratumor heterogeneity based on alterations of gene expression profiles. Commun. Biol. **3**(1), 505 (2020). https://doi.org/10.1038/s42003-020-01230-7
15. Song, D., Wang, X.: DEPTH2: an mRNA-based algorithm to evaluate intratumor heterogeneity without reference to normal controls. J. Transl. Med. **20**(1), 150 (2022)
16. Goldman, M., et al.: The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. bioRxiv (2018)
17. Abid, A., Balin, M.F., Zou, J.: Concrete autoencoders: differentiable feature selection and reconstruction. In: 36th International Conference on Machine Learning, ICML 2019 (2019)
18. Tanvir, R.B., Sobhan, M., Mondal, A.M.: An autoencoder based bioinformatics framework for predicting prognosis of breast cancer patients. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 3160–3166 (2022)
19. Sobhan, M., Al Mamun, A., Tanvir, R.B., Alfonso, M.J., Valle, P., Mondal, A.M.: Deep learning to discover genomic signatures for racial disparity in lung cancer. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 2990–2992 (2020)
20. Sobhan, M., Kalie, K., Al Mamun, A., Godavarty, A., Mondal, A.M.: Skin tone benchmark dataset for diabetic foot ulcers and machine learning to discover the salient features. In: International Conference on Image Processing, Computer Vision, & Pattern Recognition (2022)
21. Al Mamun, A., et al.: Multi-run concrete autoencoder to identify prognostic lncRNAs for 12 cancers. Int. J. Mol. Sci. **22**, 11919 (2021)
22. Al Mamun, A., Sobhan, M., Tanvir, R.B., Dimitroff, C.J., Mondal, A.M.: Deep learning to discover cancer glycome genes signifying the origins of cancer. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020 (2020)
23. Al Mamun, A., Duan, W., Mondal, A.M.: Pan-cancer feature selection and classification reveals important long non-coding RNAs. In: Proceedings of the 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020, pp. 2417–2424 (2020)
24. Maddison, C.J., Mnih, A., Teh, Y.W.: The concrete distribution: a continuous relaxation of discrete random variables. In: 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (2017)
25. Comprehensive molecular profiling of lung adenocarcinoma. Nature (2014)