The Open Bioinformatics Journal
DOI: 10.2174/0118750362364938250520114456, 2025, 18, e18750362364938

ISSN: 1875-0362
1

RESEARCH ARTICLE **OPEN ACCESS**

Characterizing and Evaluating Cell Specialization Through the Gini Index of Gene Expression: A TCGA Normal Vs. Tumor Case Study

Fabio Cumbo[1] and Daniele Santoni[2,*]

[1]*Center for Computational Life Sciences, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, 44195OH, USA*
[2]*Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Via dei Taurini 19, Rome, 00185 RM, Italy*

Abstract:

Background: The Gini index, introduced by the Italian statistician and demographer *Corrado Gini* in the first decades of the 1900s, is commonly used as a measure of statistical dispersion to evaluate income inequality within a nation. However, it is a powerful and effective measure to characterize any sample distribution and evaluate how far it is from a uniform one.

Methods: In this work we used the Gini Index as an effective and reliable measurement of the specialization of cells, using it to evaluate and compare the specialization level of normal and tumor cells according to their gene expressions.

Results: It turned out that, on average, tumor cells tend to lose their specialization or, in other words, their capacity to be the cells they were intended to be due to cancer effects. This loss of specialization in tumor cells corresponds, in our analysis, to a lower Gini Index with respect to normal cells. This behavior was observed both at a single patient level comparing Gini Indexes of coupled samples (from the same patient) and at a global level comparing distributions of Gini Indexes in normal and tumor datasets.

Discussion: This work demonstrates that the Gini Index (GI) effectively captures the loss of transcriptional specialization in tumor cells compared to normal tissues, with statistically significant differences observed both within patients and across cancer types, despite some exceptions, such as KICH and THCA.

Conclusion: In conclusion we are confident that GI could be a valuable and effective parameter to evaluate cell specialization and could provide significant insights in the context of cancer studies.

Keywords: Gini index, Gene expression, Tumor biology, Computational biology, Statistical hypothesis tests.

* Address correspondence to this author at the Institute for Systems Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, *Via* dei Taurini 19, Rome, 00185 RM, Italy; E-mail: daniele.santoni@iasi.cnr.it

Cite as: Cumbo F, Santoni D. Characterizing and Evaluating Cell Specialization Through the Gini Index of Gene Expression: A TCGA Normal Vs. Tumor Case Study. Open Bioinform J, 2025; 18: e18750362364938.
http://dx.doi.org/10.2174/0118750362364938250520114456

Received: October 29, 2024
Revised: January 24, 2025
Accepted: January 29, 2025
Published: May 29, 2025

1. INTRODUCTION

The Gini index (GI) was introduced by the Italian statistician and demographer *Corrado Gini* in the first decades of the 1900s [1-3]. It is commonly used as a measure of statistical dispersion to evaluate income inequality within a nation. The general principle is based on the comparison between the portion of economic resources and the portion of the population that possesses those resources. In other words, the GI measures how the distribution of any source of data deviates from being

uniform, collapsing this information into a number ranging from 0 to 1. In a country where a small number of individuals are extremely wealthy while the vast majority are extremely poor, the Gini Index (GI) is very high and approaches 1. Conversely, the Gini Index (GI) is very low—approaching 0—in a country where the majority of people have similar or comparable incomes. Although the GI was introduced and commonly used in this context, it can be applied to any distribution to investigate how far it is from being a uniform distribution. The GI provides information that is associated with but also complementary to other measures, such as the standard deviation or entropy, but it has several advantages, such as the direct comparability of any set of data since it is a number in the range [0,1]. Moreover, it does not need any kind of assumption as in the case of entropy, where, in most cases, a binning supervised pre-process is necessary. Jiang and colleagues in 2016 developed *GiniClust*, a tool that uses the GI in a biological context to characterize rare cell types in single-cell experiments [4]. In 2018, Tsoucas and Yuan developed a new tool, *GiniClust2*, that improved the ability to detect and cluster different cell types in single-cell experiments [5]. In 2021, Nguyen and colleagues developed the Polar Gini Curve method to characterize cluster markers by analyzing single-cell RNA sequencing data [6]. The GI was also used to characterize and identify gene classes, for example, housekeeping as well as Transporter genes according to their expression variability across different cells [7, 8], to select genes for normalizing expression profiling data [9] or in combination with Support Vector Machines to select informative genes that improve the effectiveness of classification [10].

In 2024, Furth and colleagues evaluated epigenetic heterogeneity using GI to demonstrate how oncogenic **IDH1^{mut}** drives the loss of histone acetylation and increases chromatin heterogeneity [11]. In immunoinformatics, GI has been shown to be an effective measure for evaluating diversity in single-cell T-cell and B-cell receptor sequencing experiments [12-14]. Additionally, GI has been applied in various biological contexts as an attribute selection metric in decision trees and random forest algorithms [15-17].

To the best of our knowledge, this work is the first attempt to apply the GI to gene expression in the context of tumors, comparing the GI of normal and tumor cells and evaluating their cell specialization.

In this work, we introduce the GI as an effective and reliable measurement of the specialization of cells, using it to evaluate and compare the specialization level of normal and tumor cells according to their gene expressions. The statistical significance of the differences between tumor and normal GI values was evaluated through hypothesis tests both at a single-patient level comparing GIs of coupled samples (from the same patient) and at a global level comparing distributions of GIs in normal and tumor datasets.

2. METHOD

We focus on the public gene expression (FPKM - *Fragments Per Kilobase of transcript per Million mapped reads*) quantification experiments of the TCGA program available on the open-access OpenGDC repository [18, 19] for running our analyses. Here, every kind of experimental data and metadata is first extracted from the Genomic Data Commons portal [20, 21], and then standardized into the free-BED format whose structure is described in the OpenGDC Format Definition documentation available at http://geco.deib.polimi.it/opengdc/.

The gene expression quantification data contain the list of genes involved in the experiments with their genomic coordinates (defined as chromosome, start position, end position, and strand) and their quantitative information like the htseq-count (number of reads mapping to a specific gene), FPKM (*Fragments Per Kilobase of transcript per Million mapped reads* – it normalizes the read count based on gene length and the total number of mapped reads), and FPKM-UQ (same as FPKM but considering the upper quartile only).

In this study, we focus on 17 out of 33 different tumor types (see Abbreviations section for the complete list of considered tumor types) available in the OpenGDC database, considering a number of paired normal-tumor samples ranging from a minimum of 9 (ESCA) to a maximum of 112 (BRCA). The number of samples for each tumor type is reported in Table **1** and **2**.

For each cancer type and each patient, we compare paired normal and tumor gene expression GIs to assess whether they are significantly different. The actual difference between normal and tumor GIs is compared with a distribution of 1,000 artificial GI differences obtained by randomizing gene expressions of the two samples.

For each cancer type and each patient, we compared paired normal and tumor gene expression GIs to establish whether they are significantly different. For a given cancer type, we build the set of samples $P=\{p_1, p_2,.. p_n\}$ for which both normal and tumor gene expressions are available. The GIs of gene expression associated with each p_i are indicated with GI_i^T and GI_i^H for normal and tumor samples, respectively. We define it as $GID(p_i) = GI_i^H - GI_i^T$: the difference between normal and tumor GIs. We then determine whether $GID\ (p_i)$ value is statistically significant or, in other words, whether the observed value can attest that the two conditions show significantly different gene expression distributions when evaluated through the GI. The $GID(p_i)$ is a pure number that, in general, depends on the two distributions, so we have to compare it with an expected value obtained by taking the two distributions into consideration. We generate artificial pairs of random gene expression vectors by shuffling the two gene expression vectors. For each gene, we randomly associate to the first vector one out of the two gene expression values and we assign the other value to the other vector. Once the two randomized vectors are obtained, we compute the correspondent GIs, namely GI_i^{HR} and = GI_i^{TR}

Table 1. Summary table of the Gini indices of normal and tumor samples for each of the 17 tumor types from the TCGA program (1st column), reporting the number of subject for which both normal and tumor samples are available (paired – 2nd column), the average Gini index with its standard deviation (3rd and 4th columns), followed by the minimum and maximum Gini indices (5th and 6th columns).

Tumor Type	Subjects	Summary of Gini Indices			
		Normal (Average ± Stdev)	Tumor (Average ± Stdev)	Normal (Min – Max)	Tumor (Min – Max)
BLCA	19	0.926 ± 0.010	0.916 ± 0.014	0.905 – 0.945	0.889 – 0.934
BRCA	112	0.906 ± 0.019	0.907 ± 0.016	0.873 – 0.966	0.878 – 0.971
CHOL	9	0.953 ± 0.003	0.917 ± 0.015	0.964 – 0.964	0.902 – 0.953
COAD	41	0.926 ± 0.008	0.917 ± 0.013	0.910 – 0.944	0.892 – 0.962
ESCA	8	0.937 ± 0.016	0.912 ± 0.004	0.911 – 0.965	0.901 – 0.920
HNSC	43	0.943 ± 0.042	0.922 ± 0.014	0.907 – 0.972	0.887 – 0.960
KICH	23	0.908 ± 0.019	0.948 ± 0.015	0.880 – 0.932	0.909 – 0.974
KIRC	72	0.926 ± 0.027	0.916 ± 0.015	0.891 – 0.955	0.878 – 0.949
KIRP	31	0.924 ± 0.026	0.925 ± 0.020	0.890 – 0.946	0.888 – 0.958
LIHC	50	0.968 ± 0.065	0.951 ± 0.012	0.953 – 0.983	0.914 – 0.975
LUAD	57	0.914 ± 0.021	0.909 ± 0.013	0.893 – 0.939	0.870 – 0.939
LUSC	49	0.913 ± 0.021	0.905 ± 0.016	0.892 – 0.935	0.854 – 0.938
PRAD	52	0.913 ± 0.020	0.913 ± 0.016	0.871 – 0.956	0.879 – 0.978
READ	9	0.919 ± 0.024	0.918 ± 0.008	0.906 – 0.933	0.895 – 0.933
STAD	27	0.945 ± 0.043	0.924 ± 0.016	0.915 – 0.969	0.894 – 0.957
THCA	58	0.908 ± 0.019	0.918 ± 0.017	0.891 – 0.950	0.887 – 0.974
UCEC	23	0.911 ± 0.020	0.919 ± 0.02	0.893 – 0.924	0.881 – 0.965

Table 2. Summary table of the statistical analysis performed on the number of normal-tumor samples (second column) for each of the involved 17 tumor types (first column): (Z-scores) columns 3–5 report the number of paired samples (and their percentage) for which the *p*-value, computed considering the z-score of the actual pair and those of the 1,000 Gini indices on the randomized gene expression profiles, is smaller than 0.01. In particular, the column "positive" reports the number of significant *p*-values of positive z-scores, and *vice-versa* for the "negative" column. Instead, the "not significant" column contains the number of paired samples for which the *p*-value is not significant, regardless of the positive or negative sign of their z-scores. Results are color coded according to the values reported under the 3rd, 4th and 5th columns: green if "positive" sample pairs (3rd column) are the majority (percentage higher than 50%) and red, on the other way around, if "negative" sample pairs (5th column) are the majority (percentage higher than 50%). (Wilcoxon) the presence/absence of the + and - symbols near the tumor type represents the statistical significance according to the Wilcoxon rank-sum test.

Tumor Type	Samples Normal – Tumor (Paired)	Comparison of Gini Indices for each Patient Through z-score		
		Positive ($p<0.01$)	Not Significant	Negative ($p<0.01$)
BLCA	19 – 408 (19)	11 (57.9%)	6 (31.6%)	2 (10.5%)
BRCA	113 – 1090 (112)	34 (30.4%)	34 (30.4%)	44 (39.3%)
CHOL +	9 – 36 (9)	9 (100%)	0 (0%)	0 (0%)
COAD +	41 – 456 (41)	22 (53.7%)	15 (36.6%)	4 (9.8%)
ESCA +	11 – 161 (8)	6 (75.0%)	2 (25.0%)	0 (0%)
HNSC +	44 – 500 (43)	32 (74.4%)	8 (18.6%)	3 (7.0%)
KICH -	24 – 65 (23)	0 (0%)	1 (4.3%)	22 (95.7%)
KIRC +	72 – 530 (72)	35 (48.6%)	20 (27.8%)	17 (23.6%)
KIRP	32 – 288 (31)	9 (29.0%)	9 (29.0%)	13 (41.9%)
LIHC +	50 – 371 (50)	43 (86.0%)	5 (10.0%)	2 (4.0%)
LUAD	59 – 513 (57)	19 (33.3%)	30 (52.6%)	8 (14.1%)
LUSC +	49 – 501 (49)	19 (38.8%)	23 (46.9%)	7 (14.3%)
PRAD	52 – 495 (52)	9 (17.3%)	27 (51.9%)	16 (30.8%)
READ	10 – 166 (9)	1 (11.1%)	8 (88.9%)	0 (0%)
STAD +	32 – 375 (27)	18 (66.7%)	9 (33.3%)	0 (0%)
THCA -	58 – 502 (58)	6 (10.3%)	20 (34.5%)	32 (55.2%)
UCEC	35 – 543 (23)	7 (30.4%)	5 (21.7%)	11 (47.8%)

and finally $GIDR(p_i) = GI_i^{HR}\text{-}GI_i^{HT}$ We iterate this procedure 1,000 times, obtaining a collection of 1,000 $GIDR(p_i)$ and we then compute the z-score as:

$$Z_i = \frac{GID(p_i) - Average(GIDR(p_i))}{Stdv\ (GIDR(p_i))}$$

where $Average\ (GIDR(p_i))$ and $Stdv\ (GIDR(p_i))$ are the average and the standard deviation of the 1,000 obtained $GIDR(p_i)$. According to the Shapiro-Wilk normality test, performed on sample cases (data not shown), the distribution of $GIDR(p_i)$ can be considered to be extracted from a normal distribution, allowing us to compute the *P*-value from a given z-score. We compute *P*-values considering separately the two tails of the normal distribution, evaluating the *p*-value for both tails when the normal GI is greater than the tumor one and vice versa when the normal GI is smaller than the tumor one. We set a *P*-value threshold of 0.01, applying the Bonferroni correction with respect to the number of samples of the considered tumor type.

At the end of this procedure we obtain for each tumor type and each patient a *P*-value indicating whether the normal and tumor expression values are significantly different from a statistical point of view when evaluated through the Gini indexes. A significant *P*-value derived from a positive z-score is associated with a positive difference, indicating that normal GI is significantly greater than tumor GI. In this case, normal cells are more specialized than tumor cells. On the other hand, a significant *P*-value derived from a negative z-score is associated with a negative difference, indicating that normal GI is significantly smaller than tumor GI.

For each cancer type we also compare paired normal and tumor gene expression GI distributions. To establish whether they come from the same hypothetical distribution, we perform paired Wilcoxon tests. The Bonferroni adjustment is applied for multiple test corrections. Finally, we compare normal and tumor GI distributions for all available samples, including unpaired ones (samples for which only one condition is available). Unpaired Wilcoxon tests and the Bonferroni adjustment are applied in the same way.

The entire procedure is repeated using the standard deviation (STDEV) of gene expression values, instead of the Gini Index (GI), to enable comparison at both the individual patient level and the global level.

3. RESULTS

The main goal of this paper is to study cell specialization through the GI index of gene expression comparing cancer and normal cells. First, we present a global view of GI values associated with samples coming from patients with different cancer types (see Abbreviations section) for both normal and tumor cells. We then analyze and compare for each single patient, normal and tumor GIs, showing through z-score values that they are mostly significantly different. Then, we study and evaluate for each tumor type the statistical differences between normal and tumor GI distributions through Wilcoxon tests. We apply paired statistical tests to compare GI distributions of normal and tumor-coupled samples. Finally, we consider a broader dataset including all available samples, even if not coupled, by applying non-paired statistical tests.

Table **1** reports a global view of GI values for each tumor type. The first column indicates the tumor type, the second column indicates the number of subjects for which both normal and tumor samples are available. The other columns show other statistical parameters related to GI distributions. GI values are typically distributed around 0.9, with the average in normal samples ranging from 0.907 in BRCA to 0.969 in LICH, while in tumor samples, from 0.906 in LUSC to 0.951 in LICH.

Fig. (**1**) shows the comparison between GI distributions of tumor (orange) and normal (blue) cells for four different cancer types. Panel A – on the left side of the figure – (HNSC and LIHC) clearly shows higher GI values for normal samples compared to tumor samples. On the other hand, panel C – right side of the figure – (THCA) shows an opposite behavior with higher values for tumor samples. Panel B (PRAD) shows an intermediate case

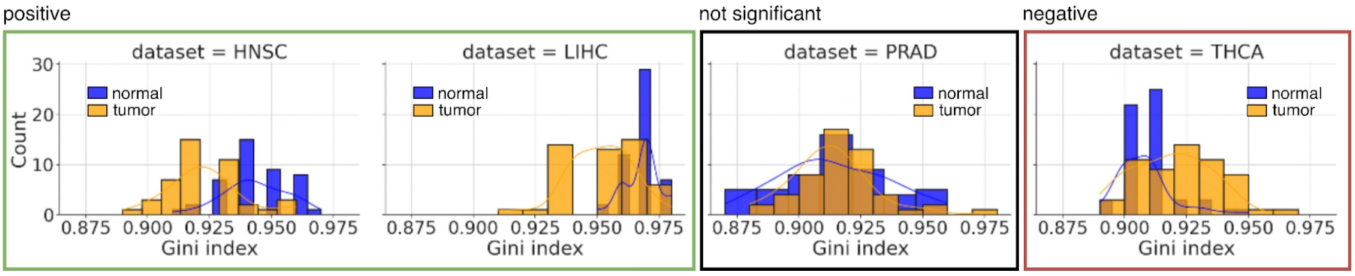

Fig. (1). Tumor and normal GI distributions for 4 different tumor types: HNSC and LIHC (panel A, green box), PRAD (panel B, black box), and THCA (panel C, red box). The plots show for each of the 4 considered tumor types how many samples (y-axis), tumors in pink and normal in cyan, have a GI falling in the corresponding bin (x-axis).

where there is no clear prevalence between tumor and normal samples. Most of the tumor types typically show GI values that are lower in cancer than in normal samples (see Supplementary Material **S1** – Supplementary Tables **S1.1-17** – for a complete view of GI distributions of all the cancer types).

In order to statistically evaluate the significance of this observed difference at a patient level, we compare the difference between tumor and normal GIs with the difference distribution between artificial gene expression arrays randomly generated from the actual ones obtaining a z-score value and the corresponding *P*-value (see Materials and Methods).

In Table **2**, the number and percentage of patients showing a significant positive difference between normal and cancer cells are reported in column 3 for each tumor type. In the same way, columns 4 and 5 report the numbers and percentages of non-significant and significant negative GI differences, respectively. The rows corresponding to a given tumor type are highlighted in green (red) when the majority of patients show a significant positive (negative) difference (*P*-value < 0.01). In the same Table **2**, **+** and **-** symbols indicate the statistical significance of the Wilcoxon rank-sum test performed on all the normal and tumor samples regardless of their pairing (see Method Section). A **+** symbol denotes that normal GI values are significantly higher than tumor GI values considering a Bonferroni adjusted *p*-value<0.01, while a **-** symbol denotes that normal GI values are significantly smaller than tumor GI values. The two analyses lead to, as expected, similar and consistent results, even if they provide a different view at a patient level and a global level.

Wilcoxon tests are also performed on the GIs distributions of paired samples for each tumor, obtaining the same results except for LUSC and ESCA, which are found not significant.

We always refer to the different tumor types with their abbreviations as reported on the Genomics Data Commons website (see Abbreviations section for the complete list of tumor types alongside their extended description).

Note that we also performed the same statistical analysis of z-scores and Wilcoxon rank-sum test based on the STDEVs instead of GIs, with the aim of proving the effectiveness of GIs over a more classical statistical approach. A summary table of the statistical analysis based on the STDEVs is reported in Supplementary Table **S2.1**. Also, note that we reported the distribution of the STDEVs alongside the GIs in Supplementary Material **S1** – Supplementary Tables **S1.1-17**. As can be observed by the comparison of Table **2** (GI) and Supplementary Table **S1.1-17** (STDEV), similar results are obtained at a global level through Wilcoxon tests (the only differences regard BRCA that is positive and THCA that is not negative for Stdev). On the contrary, the scenario is very different at a patient level; while 7 positives and 2 negatives are found by GI, only 1 positive and 1 negative are found by Stdev. Those results suggest that GI is able to capture the differences in specialization between normal and tumor cells in particular at a single patient level, where Stdev mostly fails.

4. DISCUSSION

As reported in the literature [22], the transcriptional specialization of a tumor is significantly less than the corresponding normal tissue. Consistently, the observed loss of specialization in tumor cells corresponds in our analysis to a lower GI with respect to normal cells. This behavior was observed both at a single patient level comparing GIs of coupled samples (from the same patient) through z-score analysis and at a global level comparing distributions of GIs in normal and tumor datasets.

Interestingly, despite this being the overall typical behavior, few patients show an unexpected increase in their GIs. Similarly, not all cancer types display the same behavior. Some of them, in particular KICH and THCA, show an unexpected increase of specialization in tumor cells (in 95% and 55% of samples, respectively). This astonishing result could suggest that there are peculiar shared patterns between these two tumor types, as reported in a study [23]. One possible reason to investigate further could be a lower tumor mutational burden in THCA and KICH compared to other cancer types, which may affect GI.

It is worth noting that the differences in GI values between normal and cancer cells are comparable to or smaller than the differences among different tissues. Thus, we conclude that the tissue of origin remains more relevant than the tumor or normal condition in determining GI. While one might expect much smaller GI values in tumor samples compared with normal ones, the observed results and statistical analyses (with a significance threshold of $P<0.01$ and often much smaller *P*-values) demonstrate that the differences between normal and tumor GIs are highly significant in most cancer types and patients.

The impact and significance of this work may further increase as more data becomes available in TCGA, providing greater statistical robustness and allowing for deeper insights.

CONCLUSION

The GI characterizes a distribution by assessing its deviation from a uniform distribution. It provides information related to, but also complementary to, other statistical measures such as STDEV. In this view it seems particularly suitable to be applied in the context of computational biology. To the best of our knowledge, this work is the first attempt to apply GI to gene expression in the context of tumors, comparing the GI of normal and tumor cells.

We are confident that GI could be a valuable and effective parameter to evaluate cell specialization and could provide significant insights in the context of cancer studies.

AUTHORS' CONTRIBUTIONS

It is hereby acknowledged that all authors have accepted responsibility for the manuscript's content and consented to its submission. They have meticulously reviewed all results and unanimously approved the final version of the manuscript.

LIST OF ABBREVIATIONS

The 17 different tumor types considered in this study in the form of study abbreviations are all reported below with their extended name as reported on the official Genomics Data Commons portal at https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations.

BLCA = Bladder Urothelial Carcinoma

BRCA = Breast Invasive Carcinoma

CHOL = Cholangiocarcinoma

COAD = Colon Adenocarcinoma

ESCA = Esophageal Carcinoma

HNSC = Head and Neck Squamous Cell Carcinoma

KICH = Kidney Chromophobe

KIRC = Kidney Renal Clear Cell Carcinoma

KIRP = Kidney Renal Papillary Cell Carcinoma

LIHC = Liver Hepatocellular Carcinoma

LUAD = Lung Adenocarcinoma

LUSC = Lung Squamous Cell Carcinoma

PRAD = Prostate Adenocarcinoma

READ = Rectum Adenocarcinoma

STAD = Stomach Adenocarcinoma

THCA = Thyroid Carcinoma

UCEC = Uterine Corpus Endometrial Carcinoma

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used in this research.

CONSENT FOR PUBLICATION

Not applicable.

STANDARDS OF REPORTING

STROBE guidelines were followed.

AVAILABILITY OF DATA AND MATERIALS

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

FUNDING

This work has been partially supported by CNR-IASI Project BIOSYS3 – Optimization, models and Algorithms for Bioinformatics and System Science – DIT.AD021.128.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Gini C. Concentration and dependency ratios. Riv Polit Econ 1997; 87: 769-90.

[2] Gini C. Variability and mutability: Contribution to the study of distributions and statistical relationships. Typography by Paolo Cuppin 1912.

[3] Mukhopadhyay N, Sengupta PP. Gini Inequality Index: Methods and Applications. CRC Press 2021. http://dx.doi.org/10.1201/9781003143642

[4] Jiang L, Chen H, Pinello L, Yuan GC. GiniClust: Detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol 2016; 17(1): 144. http://dx.doi.org/10.1186/s13059-016-1010-4 PMID: 27368803

[5] Tsoucas D, Yuan GC. GiniClust2: A cluster-aware, weighted ensemble clustering method for cell-type detection. Genome Biol 2018; 19(1): 58. http://dx.doi.org/10.1186/s13059-018-1431-3 PMID: 29747686

[6] Nguyen TM, Jeevan JJ, Xu N, Chen JY. Polar Gini curve: A technique to discover gene expression spatial patterns from single-cell RNA-Seq data. Genom Proteom Bioinform 2021; 19(3): 493-503. http://dx.doi.org/10.1016/j.gpb.2020.09.006 PMID: 34958962

[7] O'Hagan S, Wright Muelas M, Day PJ, Lundberg E, Kell DB. GeneGini: Assessment via the Gini coefficient of reference "Housekeeping" genes and diverse human transporter expression profiles. Cell Syst 2018; 6(2): 230-244.e1. http://dx.doi.org/10.1016/j.cels.2018.01.003 PMID: 29428416

[8] Tung KF, Pan CY, Lin W. Housekeeping protein-coding genes interrogated with tissue and individual variations. Sci Rep 2024; 14(1): 12454. http://dx.doi.org/10.1038/s41598-024-63269-4 PMID: 38816574

[9] Wright Muelas M, Mughal F, O'Hagan S, Day PJ, Kell DB. The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. Sci Rep 2019; 9(1): 17960. http://dx.doi.org/10.1038/s41598-019-54288-7 PMID: 31784565

[10] Almutiri T, Saeed F. A hybrid feature selection method combining Gini index and support vector machine with recursive feature elimination for gene expression classification. IJDMMM 2022; 14(1): 1. http://dx.doi.org/10.1504/IJDMMM.2022.122038

[11] Furth N, Cohen N, Spitzer A, *et al*. Oncogenic $IDH1^{mut}$ drives robust loss of histone acetylation and increases chromatin heterogeneity. Proc Natl Acad Sci USA 2025; 122(1): e2403862122. http://dx.doi.org/10.1073/pnas.2403862122 PMID: 39793065

[12] Tuong ZK, van der Merwe R, Canete PF, Roco JA. Computational estimation of clonal diversity in autoimmunity. Immunol Cell Biol 2024; 102(8): 692-701. http://dx.doi.org/10.1111/imcb.12801 PMID: 39010261

[13] Suo C, Polanski K, Dann E, *et al*. Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. Nat Biotechnol 2024; 42(1): 40-51.

http://dx.doi.org/10.1038/s41587-023-01734-7 PMID: 37055623
[14] Stephenson E, Reynolds G, Botting RA, *et al*. Single-cell multi-omics analysis of the immune response in COVID-19. Nat Med 2021; 27(5): 904-16. http://dx.doi.org/10.1038/s41591-021-01329-2 PMID: 33879890
[15] Pallavi M, Tejaswi C, Srilakshmi R, Swarup C. Deciphering the Complexities of Breast Cancer Genomics at the Nexus of AI, Computer Vision, and Machine Learning. John Wiley & Sons, Ltd 2024; pp. 109-32.
[16] Ren J, Gao Q, Zhou X, *et al*. Identification of key gene expression associated with quality of life after recovery from COVID-19. Med Biol Eng Comput 2024; 62(4): 1031-48. http://dx.doi.org/10.1007/s11517-023-02988-8 PMID: 38123886
[17] Lokeswaran S, Manikandan P, Rajakumar R, Marimuthu M. Brain tumor classification from gene expression dataset through supervised machine learning algorithms. Conference: 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS). Chennai, India, 2017, pp. 318-323 http://dx.doi.org/10.1109/SSPS.2017.8071613.
[18] Cumbo F, Fiscon G, Ceri S, Masseroli M, Weitschek E. TCGA2BED: Extracting, extending, integrating, and querying The Cancer Genome Atlas. BMC Bioinformatics 2017; 18(1): 6. http://dx.doi.org/10.1186/s12859-016-1419-5 PMID: 28049410
[19] Cappelli E, Cumbo F, Bernasconi A, *et al*. OpenGDC: Unifying, modeling, integrating cancer genomic data and clinical metadata. Appl Sci 2020; 10(18): 6367. http://dx.doi.org/10.3390/app10186367
[20] Heath AP, Ferretti V, Agrawal S, *et al*. The NCI genomic data commons. Nat Genet 2021; 53(3): 257-62. http://dx.doi.org/10.1038/s41588-021-00791-5 PMID: 33619384
[21] Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI genomic data commons as an engine for precision medicine. Blood 2017; 130(4): 453-9. http://dx.doi.org/10.1182/blood-2017-03-735654 PMID: 28600341
[22] Martínez O, Reyes-Valdés MH, Herrera-Estrella L. Cancer reduces transcriptome specialization. PLoS One 2010; 5(5): e10398. http://dx.doi.org/10.1371/journal.pone.0010398 PMID: 20454660
[23] Bellini MI, Lori E, Forte F, *et al*. Thyroid and renal cancers: A bidirectional association. Front Oncol 2022; 12: 951976. http://dx.doi.org/10.3389/fonc.2022.951976 PMID: 36212468