Ranking of Survival-Related Gene Sets Through Integration of Single-Sample Gene Set Enrichment and Survival Analysis

Martin Špendl[1], Jaka Kokošar[1], Ela Praznik[1], Luka Ausec[2], and Blaž Zupan[1(✉)]

[1] Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia
{martin.spendl,jaka.kokosar,ela.praznik,blaz.zupan}@fri.uni-lj.si
[2] Genialis Inc., 177 Huntington Ave Ste 1703, Boston, Massachusetts, USA

Abstract. The onset and progression of a disease are often associated with changes in the expression of groups of genes from a particular molecular pathway. Gene set enrichment analysis has thus become a widely used tool in studying disease expression data; however, it has scarcely been utilized in the domain of survival analysis. Here we propose a computational approach to gene set enrichment analysis tailored to survival data. Our technique computes a single-sample gene set enrichment score for a particular gene set, separates the samples into an enriched and non-enriched cohort, and evaluates the separation according to the difference in survival of the cohorts. Using our method on the data from The Cancer Genome Atlas and Molecular Signatures Database Hallmark gene set collection, we successfully identified the gene sets whose enrichment is predictive of survival in particular cancer types. We show that the results of our method are supported by the empirical literature, where genes in the top-ranked gene sets are associated with survival prognosis. Our approach presents the potential of applying gene set enrichment to the domain of survival analysis, linking the disease-related changes in molecular pathways to survival prognosis.

Keywords: Gene set ranking · Survival analysis · Censored data · Survival curve · Gene expression · Single-sample gene set enrichment scoring

1 Introduction

The onset of diseases and the prediction of their progression are commonly associated with variations in the expression of genes that control specific molecular pathways. Such variations are often more informative and interpretable when

Supported by the Slovenian Research Agency grants P2-0209 and L2-3170.

J. M. Juarez et al. (Eds.): AIME 2023, LNAI 13897, pp. 328–337, 2023.
https://doi.org/10.1007/978-3-031-34344-5_39

considering groups of biologically related genes rather than individual genes alone [12]. Bioinformatics has developed various computational techniques that identify which gene sets are relevant to changes in phenotypes. One such technique is gene set enrichment analysis (GSEA) [22], which calculates an enrichment for each gene set based on the mRNA expression profile of samples and their binary phenotypes, such as tumour type, response, or exposure to a drug. In contrast to binary phenotypes, survival data includes information on the time until a target event and may include censored cases where the event has not yet occurred. We cannot use GSEA or similar enrichment approaches for such data.

Censored data is common in the clinical setting, and hence there is a need for approaches for survival-based gene set scoring methods, which are currently, at best, scarce. Those few reported in the literature rely on assigning gene set-specific scores to a single patient in the dataset. The gene set activity score (GSAS) algorithm [25] relates a score with the expression of transcription factors according to the BASE algorithm [4]. An immune-based prognostic score for ovarian cancer (IPSOV) [20] uses a similar approach to evaluate the association between characteristics and overall survival. Both were used for curating gene sets but not for ranking based on their prognostic power. Similar methods for single-cell RNA-seq were developed [6].

Here, we report on a technique that can rank gene sets based on their survival prognostic ability. Our method relies on calculating sample-based gene-set scores using a single-sample extension of the GSEA [3]. ssGSEA is a method that can assign a gene set enrichment score to each sample individually, thus not requiring a phenotype label. We use ssGSEA to order expression profiled samples based on the expression enrichment of a gene set. Using the median as a splitting criterion, we create two cohorts of equal size: an enriched and non-enriched cohort. The extent of enrichment corresponds to the overexpression of genes in the gene set compared to the average gene expression. Thus, enriched and non-enriched cohorts relate to mostly above and below-average expression of genes, respectively. We then evaluate the importance of a gene set for patient survival using a log-rank test between the cohorts on a Kaplan-Meier survival plot: the more significant the difference in survival characteristics, the higher the importance of a gene set. We rank gene sets according to their log-rank p-value and correct those using Benjamini-Hochberg FDR correction.

2 Methods

Given a gene set, our scoring method for survival data consists of three steps. First, we normalize gene expression values using a common normalization procedure. Second, we rank samples based on their gene set enrichment using a single-sample gene set scoring method [3]. And third, we split samples into two equal-sized cohorts and evaluate the difference in survival using the log-rank test. We repeat the procedure for all gene sets in the relevant gene set database and rank the gene sets according to their score. The ranked list is then subject to interpretation and further investigation by a molecular biologist. We showcase

the implemented method on cancer-related data sets and use a standard curated gene set database.

2.1 Data

We collected cancer tissues from The Cancer Genome Atlas Program (TCGA) uploaded to the GEO portal (GSE62944) [17]. Samples are organised in data sets based on their tissue of origin. We collected mRNA sequencing data represented as a gene expression matrix. Different datasets vary in sample size; thus, we included only datasets with more than 100 samples (total of 20). We extracted the sample's survival time and event occurrence from clinical metadata. Survival time is the last known date when a patient is still alive. If a patient dies of cancer, we consider the event has occurred. Otherwise, if its status is unknown or it dies of unrelated death, its event status is censored. Datasets have varying sample sizes and ratios of censored data (Table 1).

Table 1. TCGA project statistics about censored data. The N is the number of samples in the dataset, and the Censored is the ratio of censored samples.

TCGA	CESC	HNSC	KIRC	LAML	LGG	LUAD	READ	SKCM
N	306	504	542	178	532	541	167	472
Censored	0.807	0.675	0.707	0.348	0.846	0.769	0.940	0.661

Gene expression is stored as transcripts per million (TPM); thus, all expression values for each sample sum up to a million. We use a standard procedure of log-transforming each gene expression with pseudo count 1 and z-score normalization for each gene across samples in a dataset. That is, we normalize the columns of the expression matrix.

Gene sets are sets of genes that encode proteins acting together in some biological process. Biologists create and curate them to better understand their function and interactions. We have considered a set of 50 curated gene sets called Hallmark gene sets from the Molecular Signature Database (MSigDB) [11,22], where gene sets represent states and processes in human cells (see Table 2).

Table 2. Example of three Hallmark gene sets and a few genes. N is the number of total genes, and gene names are from HUGO Gene Nomenclature Committee.

Hallmark gene set	N	corresponding genes
ANGIOGENESIS	36	APOH, FGRF1, ITGAV, LPL, VEGFA, ...
APOPTOSIS	161	BAX, BCL10, CASP1, ERBB2, MADD, ...
GYCOLYSIS	200	EGFR, G6PD, GALK1, LDHA, SOD1,

2.2 Single-Sample Gene Set Enrichment Analysis

Single-sample gene set enrichment analysis (ssGSEA) is a single-sample extension of the GSEA algorithm [3]. It assigns the enrichment score to a single sample based on the gene expression profile of a sample. This differs from the original GSEA algorithm, which computes the gene set's enrichment score based on the entire data set. The score represents the gene set's degree of enrichment in a sample in a given dataset. In simplified terms, gene sets with highly expressed genes will have a high enrichment score. Gene expression values of a sample are rank normalized, standardized, are sorted in decreasing order based on their rank r. Genes in the gene set form a probability mass function (PMF) with probabilities $|r|^{\alpha}$, while genes outside form a PMF with genes having equal probability. The enrichment score of a sample is then represented as the difference of cumulative density functions for those PMFs. In essence, the enrichment score of a sample describes the degree of above-average expression of genes in a gene set.

2.3 Gene Set Ranking for Survival Analysis

We aim to evaluate the utility of a gene set in separating samples into enriched and non-enriched cohorts based on their gene set enrichment score. We abstract the approach with the following procedure:

Algorithm 1. Gene Set Ranking for Survival Analysis

1: $data \leftarrow$ samples with normalized expression values
2: $geneSets \leftarrow$ Hallmark gene sets
3: $enrichemntScores \leftarrow$ ssGSEA(data, geneSets)
4: **for each** $score \in enrichemntScores$ **do**
5: $\quad cohorts \leftarrow$ split sample by median of enrichment score
6: $\quad p \leftarrow$ log-rank test between cohorts

The literature suggests the median as the least biased approach to split the data into two cohorts [2]. Our null hypothesis is that both cohorts have the same hazard function. We test the null hypothesis using a standard log-rank test, a form of χ^2 test with one degree of freedom (line 6). The 95% confidence intervals of the χ^2 test statistic are calculated using bootstrap without recalculating ssGSEA enrichment scores. We repeat the protocol for other gene sets and correct p-values for the false-discovery rate with the Benjamini-Hochberg procedure.

2.4 Robustness Estimation

We evaluate the robustness of the proposed enrichment scoring in three steps. First, we perform bootstrap sampling 1000 times to estimate the 95% confidence interval of the χ^2 test statistic without recalculating enrichment scores. Recalculating scores for samples on a bootstrapped dataset provides only marginally different 95% CI but requires much more computation. Our method estimates the CI using the same sample enrichment scores as in the original data set.

In the second step, we evaluate how individual genes in the gene set influence the results compared to random genes. We incrementally remove a subset of genes and compare the statistic with the original gene set. Additionally, we remove genes from the gene set and replace them with randomly selected genes. By repeating the procedure 100 times, we calculate 95% confidence intervals of evaluation. The third approach evaluates the robustness in terms of sample size. We downsample the original dataset incrementally to 50% of the original size and compare the χ^2 test statistic. Each downsampling is performed 100 times to estimate the 95% confidence intervals.

3 Results

With the proposed approach, we could find Hallmark genesets that characterize cohorts with significantly different survival characteristics for six of our study's twenty TCGA cancer datasets. In Table 3, we report each data set's top three gene sets and their corresponding test statistics. We would find a gene set significant if the FDR corrected p-value is below 0.01. The table also includes a reference for each gene set that confirms our findings in the existing literature; for brevity, we only include the most relevant articles that have already reported the relation between genes in a Hallmark gene set and their prognostic power in a cancer type.

Table 3. Up to three significant top-ranked Hallmark gene sets for each of the six TCGA cancer types. We report the most relevant reference if p-values are below 0.01. A complete list of literature references is available on our GitHub repository (see Conclusion). CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma, HNSC - Head and Neck squamous cell carcinoma, KIRC - Kidney renal clear cell carcinoma, LGG - Brain Lower Grade Glioma, LUAD - Lung adenocarcinoma, SKCM - Skin Cutaneous Melanoma.

TCGA	Hallmark	χ^2	p_{value}	References
CESC	UV_RESPONSE_DN	16.2	2.63e-03	[8]
	ANGIOGENESIS	14.7	2.63e-03	[24]
	PROTEIN_SECRETION	14.3	2.63e-03	[15]
HNSC	GLYCOLYSIS	26.0	1.70e-05	[10]
	MTORC1_SIGNALING	17.5	7.25e-04	[21]
	XENOBIOTIC_METABOLISM	15.0	1.78e-03	[14]
KIRC	HEME_METABOLISM	26.1	1.63e-05	[7]
	FATTY_ACID_METABOLISM	18.1	5.28e-04	[5]
	ANDROGEN_RESPONSE	16.7	7.35e-04	[27]
LGG	EMT	19.4	2.87e-04	[23]
	ANGIOGENESIS	18.6	2.87e-04	[16]
	COAGULATION	18.3	2.87e-04	[18]
LUAD	MTORC1_SIGNALING	14.4	4.02e-03	[13]
	HYPOXIA	14.2	4.02e-03	[19]
	GLYCOLYSIS	12.1	8.29e-03	[26]
SKCM	INTERFERON_GAMMA_RESPONSE	15.0	5.36e-03	[1]
	INTERFERON_ALPHA_RESPONSE	13.3	6.78e-03	[9]

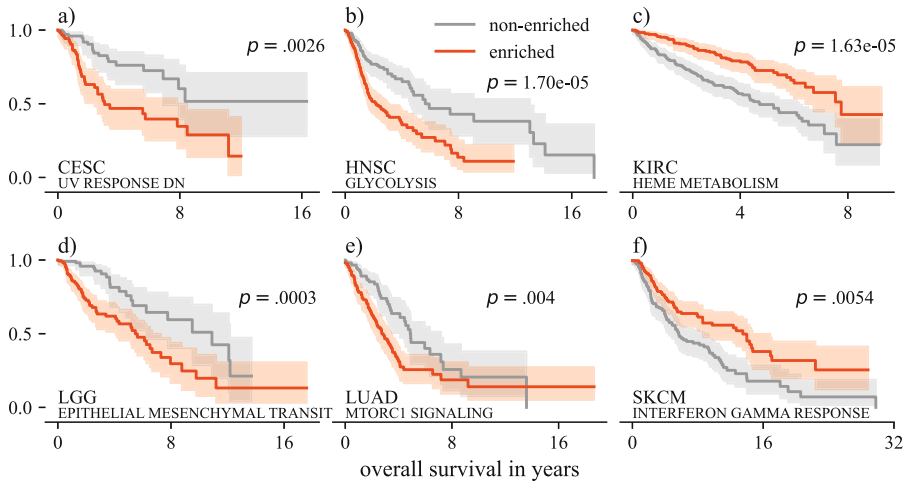

Fig. 1. Kaplan-Meier survival curves for the best-performing gene set. a) CESC-UV response down, b) HNSC-Gylcolysis, c) KIRC-Heme metabolism, d) LGG-Epithelial Mesenchymal Transition, e) LUAD - MTORC1 signalling, f) SKCM-IFN-γ response.

We find literature support for all top-ranked gene set-cancer type pairs. For example, in the case of cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), the cancer occurrence is highly linked to the infection with human papillomavirus (HPV) infection. This virus produces proteins that transform a human cell into a cancer cell by degrading the main tumour suppressor protein, p53. They also inhibit DNA damage repair in response to UV exposure, leading to extensive mutations. The tumour becomes more invasive by inducing angiogenesis and forming new blood vessels. Our method finds those samples with enriched scores in Hallmark gene sets related to those pathways have lower survival curves.

We observe the survival difference in cohorts suggested by the top-ranked gene sets for TCGA datasets. We observe that the enriched cohort is linked to worse survival compared to non-enriched in 4 out of 6 cases (Fig. 1a,b,d,e), whereas linked to a better prognosis for the other two (Fig. 1c, f).

4 Discussion

The results from the TCGA datasets suggest that our proposed method can pinpoint the relevant gene sets and that the ranking can identify those best related to the phenomena represented in the corresponding dataset. Gene set scoring produces a ranked list with FDR-corrected p-values, but the process is just a hypothesis generation. We should consider significant results with caution. However, a large body of literature confirming our case studies findings suggested that our results are not a result of chance.

There are also Hallmark gene sets that we expect to be enriched in some cancer types but did not appear to be significant. These missing results could stem from our data collection process or assumptions of our proposed method. Namely, we did not consider prior treatment, genetic predispositions, or other diseases when modelling survival time. On the other hand, one of the assumptions is that the ratio between the enriched and non-enriched cohorts is equal. As we show below, this is a broad overstatement, but the literature suggests it is the least biased [2].

4.1 Analysis of the Method's Robustness

We comment on the robustness of the approach by showing an example of the highest ranked gene set HALLMARK_GYLCOLYSIS on the Head and Neck squamous cell carcinoma (HNSC) dataset (Fig. 2). We use the bootstrap method to evaluate the 95% CI of the test statistic. Bootstrap confidence intervals for higher test statistic values are wider and normally distributed, while lower values have a more skewed distribution towards 0. We observe the number of unique samples in a dataset does not affect the size of 95% CI.

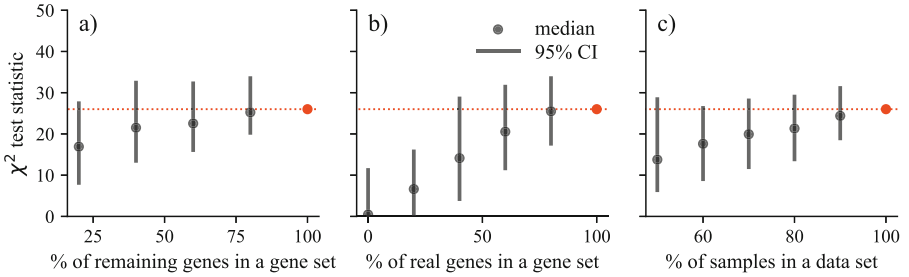

Fig. 2. Robustness of the method. We used the TCGA-HNSC dataset and HALLMARK_GLYCOLYSIS gene set while comparing χ^2 test statistic. a) Reducing the number of genes in the gene set, b) Replacing gene set genes with random ones, c) Reducing the number of samples in a data set.

We observe how the number of genes in a gene set affects performance (Fig. 2a). Removing as much as half of genes from the gene set has a marginal effect on calculated test statistics. The following shows how scoring is not dependent on any single gene, but their effect is combined in sample ranking. The redundancy of genes in biological pathways and gene sets is known. In contrast, when replacing genes in a gene set with random genes, we observe a clear shift of the test statistic towards lower values (Fig. 2b). Adding noise to the enrichment calculation impacts sample ranking and, thus, the method's performance. Confidence intervals of the mean over multiple runs also become smaller due to the relative distance from zero.

Removing samples from the dataset results in slowly decreasing values in the χ^2 test statistic. Confidence intervals become wider due to the variation in possible cohort combinations. Even when removing 50% of all samples, the gene set is enriched with statistical significance. This suggests that we can use this method even on smaller sample sizes.

4.2 Varying Splitting Threshold

Our cohort formation assumes equally-sized cohorts. Instead of using the median score for splitting, we could search for the score threshold that maximizes the log-rank statistics and find gene set enriched and non-enriched cohorts of different sizes. Figure 3 shows that such threshold search for the HALLMARK_ADIPOGENESIS gene set on the KIRC dataset improves the results. When using the default median value as a threshold, the log-rank statistic of 8.48 is substantially smaller than 25.37 for the split where 75% of the samples are placed in the enriched cohort. We have observed similar benefits of threshold search for other gene sets and data sets.

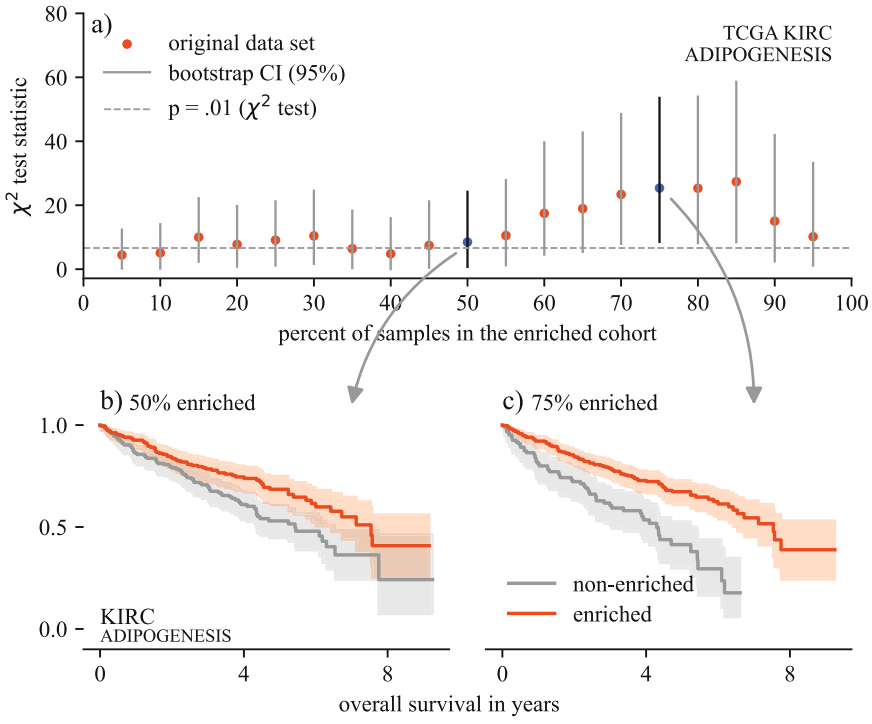

Fig. 3. Varying cohort split selection on HALLMARK_ADIPOGENESIS and TCGA KIRC dataset. a) Varying split threshold between 5 - 95%, b,c) Comparison of Kaplan-Meier plots for 50% and 75% per cent of samples in an enriched cohort.

5 Conclusions

The abundance of censored data and molecular fingerprints in clinical settings encourages the development of methods that can shed light on biological processes that govern disease progression. We propose a survival-related gene set ranking method based on single-sample enrichment scoring. An application of our method on publicly available data sets where the results match those from the literature confirms that our approach produces meaningful results with relevant implications to prognosis. The simplicity of the proposed method also leaves room for additional improvements, such as choosing different splitting criteria for cohort formation. The code and datasets used are available on GitHub (https://github.com/biolab/AIME-2023-paper) and archived on Zenodo [28].

References

1. Alavi, S., Stewart, A.J., Kefford, R.F., Lim, S.Y., Shklovskaya, E., Rizos, H.: Interferon signaling is frequently downregulated in melanoma. Front. Immunol. **9**, 1414 (2018)
2. Altman, D.G.: Prognostic models: a methodological framework and review of models for breast cancer. Cancer Invest. **27**(3), 235–243 (2009)
3. Barbie, D.A., et al.: Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature **462**(7269), 108–112 (2009)
4. Cheng, C., Yan, X., Sun, F., Li, L.M.: Inferring activity changes of transcription factors by binding association with sorted expression profiles. BMC Bioinf. **8**(1), 1–12 (2007)
5. Du, W., et al.: HIF drives lipid deposition and cancer in CCRCC via repression of fatty acid metabolism. Nat. Commun. **8**(1), 1–12 (2017)
6. Dwivedi, B., Mumme, H., Satpathy, S., Bhasin, S.S., Bhasin, M.: Survival genie, a web platform for survival analysis across pediatric and adult cancers. Sci. Rep. **12**(1), 3069 (2022)
7. Frezza, C., et al.: Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. Nature **477**(7363), 225–228 (2011)
8. Jackson, S., Storey, A.: E6 proteins from diverse cutaneous HPV types inhibit apoptosis in response to UV damage. Oncogene **19**(4), 592–598 (2000)
9. Kirkwood, J.M., Strawderman, M.H., Ernstoff, M.S., Smith, T.J., Borden, E.C., Blum, R.H.: Interferon alfa-2b adjuvant therapy of high-risk resected cutaneous melanoma: the eastern cooperative oncology group trial EST 1684. J. Clin. Oncol. **14**(1), 7–17 (1996)
10. Kumar, D.: Regulation of glycolysis in head and neck squamous cell carcinoma. Postdoc J.: J. Postdoctoral Res. Postdoctoral Affairs **5**(1), 14 (2017)
11. Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., Mesirov, J.P.: Molecular signatures database (MSigdDB) 3.0. Bioinformatics **27**(12), 1739–1740 (2011)
12. Maleki, F., Ovens, K., Hogan, D.J., Kusalik, A.J.: Gene set analysis: challenges, opportunities, and future research. Front. Genet. **11**, 654 (2020)
13. Marinov, M., Fischer, B., Arcaro, A.: Targeting mTOR signaling in lung cancer. Crit. Rev. Oncol. Hematol. **63**(2), 172–182 (2007)

14. Namani, A., Rahaman, M., Chen, M., Tang, X., et al.: Gene-expression signature regulated by the keap1-nrf2-cul3 axis is associated with a poor prognosis in head and neck squamous cell cancer. BMC Cancer **18**(1), 1–11 (2018)
15. Noordhuis, M.G., et al.: Expression of epidermal growth factor receptor (EGFR) and activated EGFR predict poor response to (chemo) radiation and survival in cervical cancerthe EGFR pathway in advanced-stage cervical cancer. Clin. Cancer Res. **15**(23), 7389–7397 (2009)
16. Plate, K.H., Risau, W.: Angiogenesis in malignant gliomas. Glia **15**(3), 339–347 (1995)
17. Rahman, M., Jackson, L.K., Johnson, W.E., Li, D.Y., Bild, A.H., Piccolo, S.R.: Alternative preprocessing of RNA-sequencing data in the cancer genome atlas leads to improved analysis results. Bioinformatics **31**(22), 3666–3672 (2015)
18. Rong, Y., Post, D.E., Pieper, R.O., Durden, D.L., Van Meir, E.G., Brat, D.J.: PTEN and hypoxia regulate tissue factor expression and plasma coagulation by glioblastoma. Can. Res. **65**(4), 1406–1413 (2005)
19. Salem, A., et al.:: Targeting hypoxia to improve non-small cell lung cancer outcome. JNCI: J. Nat. Cancer Instit. **110**(1), 14–30 (2018)
20. Shen, S., et al.: Development and validation of an immune gene-set based prognostic signature in ovarian cancer. EBioMedicine **40**, 318–326 (2019)
21. Simpson, D.R., Mell, L.K., Cohen, E.E.: Targeting the PI3K/AKT/mTOR pathway in squamous cell carcinoma of the head and neck. Oral Oncol. **51**(4), 291–298 (2015)
22. Subramanian, A., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. **102**(43), 15545–15550 (2005)
23. Tao, C., Huang, K., Shi, J., Hu, Q., Li, K., Zhu, X.: Genomics and prognosis analysis of epithelial-mesenchymal transition in glioma. Front. Oncol. **10**, 183 (2020)
24. Tomao, F., et al.: Angiogenesis and antiangiogenic agents in cervical cancer. Onco. Targets. Ther. **7**, 2237 (2014)
25. Varn, F.S., Ung, M.H., Lou, S.K., Cheng, C.: Integrative analysis of survival-associated gene sets in breast cancer. BMC Med. Genomics **8**(1), 1–16 (2015)
26. Zhang, L., Zhang, Z., Yu, Z.: Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. J. Transl. Med. **17**(1), 1–13 (2019)
27. Zhao, H., Leppert, J.T., Peehl, D.M.: A protective role for androgen receptor in clear cell renal cell carcinoma based on mining TCGA data. PLoS ONE **11**(1), e0146505 (2016)
28. Špendl, M., Kokošar, J.: biolab/aime-2023-paper: Version 1.0 (2023). https://doi.org/10.5281/zenodo.7572951