PLOS COMPUTATIONAL BIOLOGY

PERSPECTIVE

Gene signatures for cancer research: A 25-year retrospective and future avenues

Wei Liu[1], **Huaqin He**[1], **Davide Chicco**[2,3]*

1 College of Life Sciences, Fujian Agriculture and Forestry University, Fuzhou, China, **2** Dipartimento di Informatica Sistemistica e Comunicazione, Università di Milano-Bicocca, Milan, Italy, **3** Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada

* davidechicco@davidechicco.it

OPEN ACCESS

Citation: Liu W, He H, Chicco D (2024) Gene signatures for cancer research: A 25-year retrospective and future avenues. PLoS Comput Biol 20(10): e1012512. https://doi.org/10.1371/journal.pcbi.1012512

Editor: Arturo Medrano-Soto, University of California San Diego, UNITED STATES OF AMERICA

Published: October 16, 2024

Funding: The work of D.C. was funded by the European Union–Next Generation EU programme, in the context of The National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 "Conseguenze e sfide dell'invecchiamento", Project Age-It (Ageing Well in an Ageing Society), and was supported by Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Over the past two decades, extensive studies, particularly in cancer analysis through large datasets like The Cancer Genome Atlas (TCGA), have aimed at improving patient therapies and precision medicine. However, limited overlap and inconsistencies among gene signatures across different cohorts pose challenges. The dynamic nature of the transcriptome, encompassing diverse RNA species and functional complexities at gene and isoform levels, introduces intricacies, and current gene signatures face reproducibility issues due to the unique transcriptomic landscape of each patient. In this context, discrepancies arising from diverse sequencing technologies, data analysis algorithms, and software tools further hinder consistency. While careful experimental design, analytical strategies, and standardized protocols could enhance reproducibility, future prospects lie in multiomics data integration, machine learning techniques, open science practices, and collaborative efforts. Standardized metrics, quality control measures, and advancements in single-cell RNA-seq will contribute to unbiased gene signature identification. In this perspective article, we outline some thoughts and insights addressing challenges, standardized practices, and advanced methodologies enhancing the reliability of gene signatures in disease transcriptomic research.

Perspective

The current scenario on gene signatures

The Cancer Genome Atlas (TCGA) is a comprehensive resource that provides a wealth of genomic and clinical data for various types of cancer [1]. Tens of thousands of "gene signature" scientific papers have been published since the proposal of gene signature by Eric S. Lander in 1999 [2].

If one searched "TCGA signature" on Google Scholar, about 90 thousand studies would pop up into the screen (Fig 1A). Currently, TCGA contains multiomics data from 69 primary cancer sites [1], which may mean that each cancer type has been analyzed over 1,000 times. A "gene expression signature" was originally defined as a single gene or a panel of altered genes with validated specificity in terms of diagnosis, prognosis, or prediction of therapeutic response [3]. Over the past 25 years, a large number of gene signatures have been generated for human diseases (Fig 1B) [4]. In 2024, gene expression signatures continue to play a pivotal

Competing interests: The authors have declared that no competing interests exist.

role in cancer research and clinical applications. These signatures, comprising patterns of gene activities, offer valuable insights into prognosis, treatment response, and patient stratification.

Recent studies, such as those evaluating the PAM50 gene expression signatures in older breast cancer patients [5], underscore their prognostic capacity. Additionally, the Tumor Inflammation Signature (TIS), consisting of 18 genes, developed by Merck in collaboration with NanoString, has demonstrated utility in identifying patients who may benefit from specific cancer treatments [6].

Several gene signatures are clinically available, such as Oncotype and Prosigna [7].

With integrated larger datasets and modern algorithms, better gene expression signature was proposed to improve these commercial platforms [7,8]. Furthermore, advancements like patient-derived gene expression signatures induced by cancer treatment are contributing to a deeper understanding of treatment effects at the molecular level [9].

As technology and research progress, gene expression signatures remain indispensable tools, providing actionable information for personalized medicine and enhancing our comprehension of cancer biology. Key characteristics of those recent gene signatures contain tens of genes that represent the different biological functions of diseases. For example, the 18 genes of TIS cover four areas of biology related to T and natural killer cells, antigen-presenting cells, interferon gamma biology, and T cell exhaustion [6].

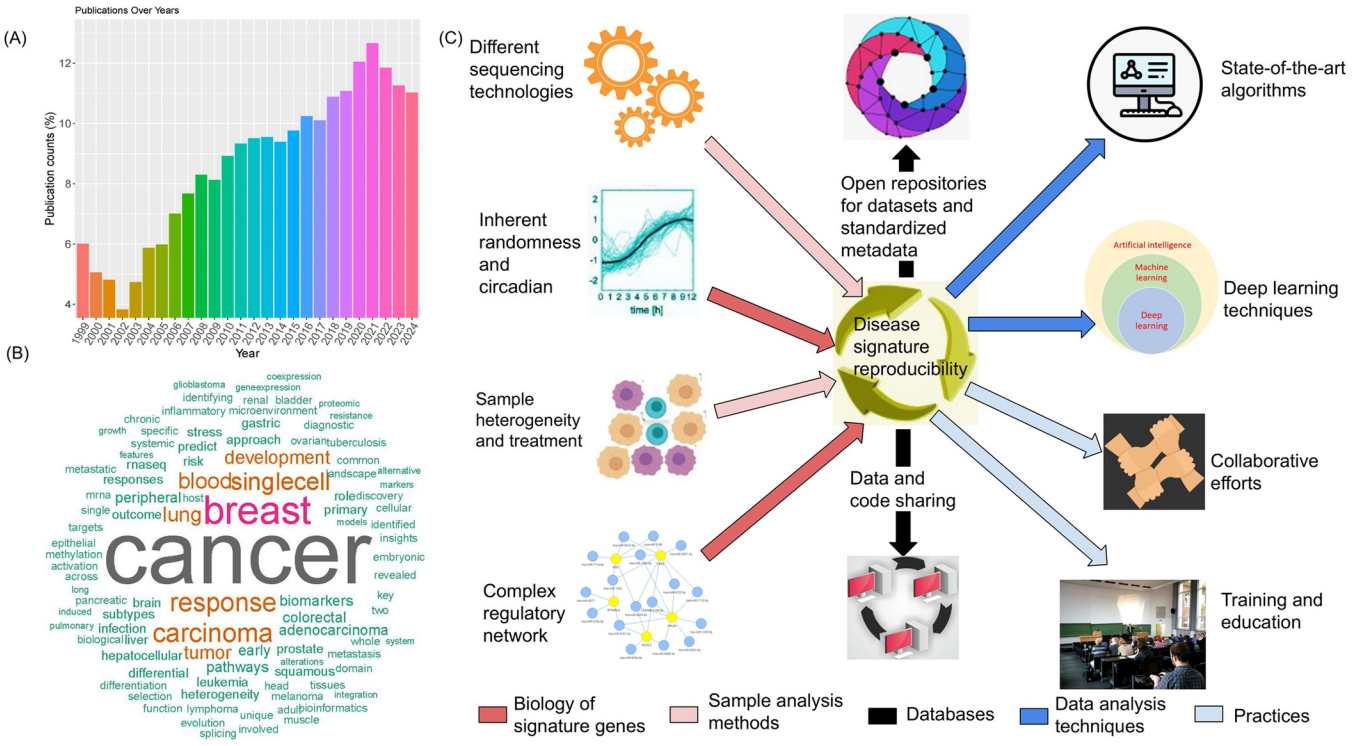

Fig 1. Factors influencing the discovery of new genetic signatures. Thousands of publications about cancer gene signatures have been published in the last 25 years; however, reproducibility is not always possible. (**A**) Publication counts for years 1999–2023. (**B**) Wordcloud shows the frequently mentioned words in publications. (**C**) Factors that can influence and improve the signature reproducibility. The inward arrows indicate factors that may influence the reproducibility of gene signatures, including different sequencing technologies, inherent randomness, sample heterogeneity, and complex regulatory networks. The outward arrows indicate strategies that may help improve the reproducibility of gene signatures, including state-of-the-art algorithms, deep learning techniques, open repositories for datasets and metadata, data and code sharing, collaborative efforts, and training and education.

https://doi.org/10.1371/journal.pcbi.1012512.g001

Still today, many biological process-related gene signatures have been proposed, such as cuproptosis, necroptosis, ferroptosis, inflammation, epithelial–mesenchymal transition (EMT), interferon gamma–related gene signatures [10–14].

The ultimate aim of gene signatures is to help improve patient therapies and to accelerate precision medicine [15]. However, there are only a small number of overlapping genes across gene signatures, and sometimes even contradictions are observed [16], hindering clinical applications [15]. Similar issues were raised recently in single-cell RNA-seq differential expression analysis [17]. The high rate of false discoveries in single-cell RNA-seq differential expression analysis, as discussed in recent literature [17], poses significant challenges for the clinical application of gene signatures.

The variability and noise intrinsic to single-cell RNA-seq data can lead to differential expression results that are inconsistent with bulk RNA-seq data, which is commonly used in clinical practice. This discrepancy might result in gene signatures that are less reliable and reproducible when applied to heterogeneous tissue samples, thereby limiting their clinical effectiveness.

In this perspective article, we take a snapshot of the overall landscape of gene signatures in bioinformatics, outlining the main challenges and resources within this scientific domain.

The biology of gene signatures

The inherent dynamic nature of the transcriptome underscores the complexity and continuous fluctuations in the abundance of RNA molecules within a cell or tissue (Fig 1C). Various factors, including cellular composition, tissue origin, and developmental stages, could influence gene expression patterns. Therefore, gene signatures identified from one disease might not work in other diseases.

Gene signatures should not only be functionally annotated but also be bundled with detailed metadata, such as how sampling was performed and which age, sex, and tissue are in the origin cohort [18]. A large fraction of differential gene lists were found to be nonspecific, reflecting shared biology rather than technical artifacts or ascertainment biases [19]. The constant interplay of transcription, RNA processing, and degradation processes adds layers of intricacy to the transcriptomic landscape. Moreover, the diversity of RNA species, such as messenger RNA (mRNA), noncoding RNA (ncRNA), and various splice variants, adds another layer of complexity to the detection process [20].

Moreover, most of current gene signatures are mainly at the gene level, and not at the isoform level. Gene functional complexity includes its diverse roles across tissues, interactions with partners, and multifaceted functions. An example is TGF-β that functions as a tumor suppressor during the initial stages of epithelial carcinogenesis [21].

However, in advanced stages, TGF-β transforms into a tumor promoter, as cancer cells develop resistance to its growth-inhibitory effects through various mechanisms, including alterations in TGF-β signaling components [16]. The impact of TGF-β on cancer cells can be detrimental or beneficial depending on the cellular context [22,23].

However, there are many multifaceted genes within the genome. The current gene signature panel may not consider the effect of combinations of different genes on prognostic performance [24].

It is worth noting that multiple current studies rely on gene signatures from the same biology process. The single biology process-related genes may contain redundancy. An example is the gene coexpression analysis, which focuses on sets of genes but not individual genes and helps to reduce the redundancy [25]. To obtain a better inference of differential gene expression for lowly expressed genes, additional gene coexpression data were integrated to enhance

the power of differential analysis [26]. Compared to individual genes, gene coexpression modules have been identified as stable units in cancer cell lines [25]. This may be due to the complexity of gene interactions and gene redundancy.

The dysfunction of different genes may induce the same disease, but the same disease does not always involve dysfunction of the same group of genes. For personalized treatments, a drug that works effectively for a patient might not demonstrate similar efficacy for another patient with the same disease, as the two patients might not share the same dysregulated transcriptome. This aspect might explain why the behavior of a gene signature in one cohort is not always reproducible in another cohort. It has been recognized that each patient is unique [27]: unique life history, unique lifestyle, unique genetics, unique health habits, and so on.

Precision medicine emphasizes tailoring medical care to the specific characteristics of each patient, including genetic makeup, lifestyle, and environmental factors. This approach recognizes that individuals with the same disease may respond differently to treatments based on their unique health profile [28]. Therefore, it is not surprising to observe that gene signature identified in one dataset demonstrates lower accuracy in new datasets [29]. This may partially explain the fact that more than 90% of drug candidates fail during clinical trials, which may take effect only in patients with the exact signature [30].

Technologies, methods, and best practices

Different sequencing technologies can contribute to discrepancies in the observed gene expression profiles. This variability may arise from differences in sequencing chemistry, sequencing depth, read length, error rates, and other platform-specific features [31]. Consequently, the lack of uniformity in data generation across platforms poses a challenge in achieving consistent and reproducible gene signatures. Researchers need to consider these platform-related factors critically when interpreting and comparing transcriptome data. This issue emphasizes the importance of careful experimental design and analytical strategies to enhance the reliability of findings across different sequencing platforms. New data analysis methods were also developed to perform platform-independent analysis [1–3,32,33].

Variability in gene signatures may be worsened by the utilization of diverse data analysis algorithms and software tools. The choice of different analytical methods, parameter settings, and statistical approaches in fact can introduce significant differences in the interpretation of transcriptomic data: Factors such as normalization techniques [34], differential expression criteria [35], and the handling of batch effects [36,37] contribute to the disparities observed in gene signatures.

Additionally, software-specific features, updates, and algorithmic improvements over time can impact the consistency of results. To enhance the reproducibility of gene signatures, researchers should carefully consider the selection of data analysis tools, adhere to best practices, and conduct robust validation [38–40]. Standardized protocols and benchmarking exercises can aid in evaluating the performance and reliability of different algorithms, ultimately contributing to more consistent and meaningful outcomes in transcriptomic analyses [41–45].

Reproducible gene signature identification faces several challenges, but advancements in methodologies and practices offer promising future perspectives: The establishment of standardized protocols and guidelines for experimental design, sample processing, and data analysis is crucial (Fig 1A). While consensus within the scientific community on best practices will contribute to increased reproducibility [46], continued efforts in benchmarking different algorithms and software tools will provide valuable insights into their strengths and limitations. Comparative studies across multiple platforms and algorithms will aid researchers in making informed choices, promoting transparency and reproducibility, and the integration of

multiomics data [47], combining information from genomics, transcriptomics, proteomics, and other layers, can enhance the robustness of gene signature identification [48,49]. ENCODE [50] is an example of a successful multiomics project. Regarding open source software packages, mixOmics [51] in R and INTEGRATE [52] in Python are examples of effective tools for this scope.

This holistic approach might capture a more comprehensive view of molecular changes and increase the reproducibility of identified signatures (Fig 1C).

Moreover, advancements in machine learning and deep learning techniques hold promise for improving the accuracy and reproducibility of gene signature identification [53]. These approaches have the potential to identify complex patterns and interactions within large-scale omics datasets. Embracing open science practices, such as open data sharing, open source code sharing, and transparent reporting, can enhance reproducibility. In this context, the choice of open source programming languages and software packages results being a key pillar of any reliable bioinformatics project: Using open source software code such as R, Python, Rust, or Julia, in fact, can guarantee the free, unrestricted reproducibility of the computational experiments by anyone in the world [54]. Open source popular computational biology projects such as Bioconductor [55], Bioconda [56], and Galaxy [57] deserve special attention for bioinformaticians.

Of course, reproducibility can be impacted by software-specific updates, which might improve the precision of the results but require efforts and energy from researchers to be installed [58,58]. In the just-mentioned platforms (Bioconda, Bioconductor, and Galaxy), these problems are mitigated by special focus and attention on documentation [59].

On the other hand, open repositories for datasets and standardized metadata can facilitate result validation and comparison across studies: Open public online bioinformatics resources can help researchers both find and release new datasets for signature identification.

Some online resources and search engines for open, unrestricted, deidentified biomedical data are the following:

- Gene Expression Omnibus (GEO) [60]
- ArrayExpress [61]
- Sequence Read Archive (SRA) [62]
- Zenodo [63,64]
- Kaggle [65]
- University of California Irvine Machine Learning Repository [66]
- Figshare [67,68]
- PhysioNet [69,70]
- Google Dataset Search [71]
- re3data.org [72]

If data availability is pivotal, also the integration of different data formats is a relevant aspect in this scenario. Collaborative efforts within the scientific community to harmonize data formats, processing pipelines, and analysis workflows and the integration of rigorous quality control measures at various stages of the experimental and analytical processes is pivotal (Fig 1C). Also, standardized metrics for assessing data quality, normalization effectiveness, and batch effect correction can enhance reproducibility, while the utilization of the state-of-the-art

single-cell RNA-seq and structured ontology information [73–77] with improved statistical methods can help to identify unbiased gene signatures [78].

Ongoing free training and education initiatives to keep researchers updated on the latest methodologies and best practices, such as Software Carpentry [79], can contribute to forge skilled researchers [80,81]. Moreover, initiatives such as the open access *Education* collection of the *PLOS Computational Biology* journal [82] and the free online bioinformatics video courses on Coursera [83] can be useful for students and researchers worldwide, especially in developing countries.

Biases for reporting genes associated with higher fold changes should be considered [84], and novel algorithms for detecting gene signatures should be developed [85]. Comprehensive and updated gene signature databases should be established, so that researchers can upload their own gene signatures and compare their results with the published data [86], by using different measures that calculate the degree of overlap between gene signatures [4].

Finally, an evidence-based approach is required to translate gene signatures from the laboratory to clinical practice [87]. All these improvements would help to assess the reliability of newly identified gene signatures, which can ultimately influence the discovery of better therapies and drugs, which, in turn, can impact positively the lives of patients in the hospitals (Fig 1C).

In the future, we expect a more frequent adoption of the just-mentioned best practices to discover novel, more robust, and effective gene signatures for cancer research.

Author Contributions

Conceptualization: Wei Liu, Davide Chicco.

Data curation: Davide Chicco.

Formal analysis: Wei Liu, Huaqin He.

Investigation: Wei Liu, Davide Chicco.

Methodology: Wei Liu, Huaqin He, Davide Chicco.

Project administration: Wei Liu.

Supervision: Huaqin He, Davide Chicco.

Validation: Davide Chicco.

Visualization: Wei Liu, Davide Chicco.

Writing – original draft: Wei Liu, Huaqin He, Davide Chicco.

Writing – review & editing: Wei Liu.

References

1. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45(10):1113–1120. https://doi.org/10.1038/ng.2764 PMID: 24071849
2. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286 (5439):531–537. https://doi.org/10.1126/science.286.5439.531 PMID: 10521349
3. Chibon F. Cancer gene expression signatures–the rise and fall? Eur J Cancer. 2013; 49(8):2000–2009. https://doi.org/10.1016/j.ejca.2013.02.021 PMID: 23498875
4. Shi X, Yi H, Ma S. Measures for the degree of overlap of gene signatures and applications to TCGA. Brief Bioinform. 2015; 16(5):735–744. https://doi.org/10.1093/bib/bbu049 PMID: 25552438

5. Castresana-Aguirre M, Johansson A, Matikas A, Foukakis T, Lindström LS, Tobin NP. Clinically relevant gene signatures provide independent prognostic information in older breast cancer patients. Breast Cancer Res. 2024; 26(1):1–11. https://doi.org/10.1186/s13058-023-01758-6 PMID: 38167446
6. Damotte D, Warren S, Arrondeau J, Boudou-Rouquette P, Mansuet-Lupo A, Biton J, et al. The tumor inflammation signature (TIS) is associated with anti-PD-1 treatment benefit in the CERTIM pan-cancer cohort. J Transl Med. 2019; 17:1–10.
7. Bueno-Fortes S, Berral-Gonzalez A, Sánchez-Santos JM, Martin-Merino M, De Las Rivas J. Identification of a gene expression signature associated with breast cancer survival and risk that improves clinical genomic platforms. Bioinformatics Adv Dermatol. 2023; 3(1):vbad037. https://doi.org/10.1093/bioadv/vbad037 PMID: 37096121
8. Brayer KJ, Kang H, El-Naggar AK, Andreasen S, Homøe P, Kiss K, et al. Dominant gene expression profiles define adenoid cystic carcinoma (ACC) from different tissues: validation of a gene signature classifier for poor survival in salivary gland ACC. Cancer. 2023; 15(5):1390. https://doi.org/10.3390/cancers15051390 PMID: 36900183
9. Liu Z, Chen R, Yang L, Jiang J, Ma S, Chen L, et al. CDS-DB, an omnibus for patient-derived gene expression signatures induced by cancer treatment. Nucleic Acids Res. 2024; 52(D1):D1163–D1179. https://doi.org/10.1093/nar/gkad888 PMID: 37889038
10. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. Nat Rev Genet. 2007; 8(8):601–609. https://doi.org/10.1038/nrg2137 PMID: 17607306
11. Solé X, Bonifaci N, López-Bigas N, Berenguer A, Hernandez P, Reina O, et al. Biological convergence of cancer signatures. PLoS ONE. 2009; 4(2):e4544. https://doi.org/10.1371/journal.pone.0004544 PMID: 19229342
12. Manjang K, Tripathi S, Yli-Harja O, Dehmer M, Glazko G, Emmert-Streib F. Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. Sci Rep. 2021; 11(1):156. https://doi.org/10.1038/s41598-020-79375-y PMID: 33420139
13. Yamaguchi K, Mandai M, Oura T, Matsumura N, Hamanishi J, Baba T, et al. Identification of an ovarian clear cell carcinoma gene signature that reflects inherent disease biology and the carcinogenic processes. Oncogene. 2010; 29(12):1741–1752. https://doi.org/10.1038/onc.2009.470 PMID: 20062075
14. Okuzono Y, Hoshino T. Comprehensive biological interpretation of gene signatures using semantic distributed representation. BioRxiv 846691 [Preprint]. 2019. Available from: https://www.biorxiv.org/content/10.1101/846691v1.full.
15. Sotiriou C, Piccart MJ. Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? Nat Rev Cancer. 2007; 7(7):545–553. https://doi.org/10.1038/nrc2173 PMID: 17585334
16. Santibanez JF, Obradović H, Kukolj T, Krstić J. Transforming growth factor-β, matrix metalloproteinases, and urokinase-type plasminogen activator interaction in the cancer epithelial to mesenchymal transition. Dev Dyn. 2018; 247(3):382–395. https://doi.org/10.1002/dvdy.24554 PMID: 28722327
17. Squair JW, Gautier M, Kathe C, Anderson MA, James ND, Hutson TH, et al. Confronting false discoveries in single-cell differential expression. Nat Commun. 2021; 12(1):5692. https://doi.org/10.1038/s41467-021-25960-2 PMID: 34584091
18. Hippen AA, Greene CS. Expanding and remixing the metadata landscape. Trends in Cancer. 2021; 7(4):276–278. https://doi.org/10.1016/j.trecan.2020.10.011 PMID: 33229213
19. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. Proc Natl Acad Sci. 2019; 116(13):6491–6500. https://doi.org/10.1073/pnas.1802973116 PMID: 30846554
20. Collins LJ. The RNA infrastructure: an introduction to ncRNA networks. RNA Infrastructure and Networks. 2011, p. 1–19. https://doi.org/10.1007/978-1-4614-0332-6_1 PMID: 21915779
21. Derynck R, Akhurst RJ, Balmain A. TGF-β signaling in tumor suppression and cancer progression. Nat Genet. 2001; 29(2):117–129.
22. Wu F, Weigel KJ, Zhou H, Wang XJ. Paradoxical roles of TGF-β signaling in suppressing and promoting squamous cell carcinoma. Acta Biochim Biophys Sin. 2018; 50(1):98–105.
23. Kuburich NA, Sabapathy T, Demestichas BR, Maddela JJ, den Hollander P, Mani SA. Proactive and reactive roles of TGF-β in EMT-induced plasticity. Semin Cancer Biol. 2023; 95:120–139.
24. Abend M, Amundson SA, Badie C, Brzoska K, Kriehuber R, Lacombe J, et al. RENEB inter-laboratory comparison 2021: the gene expression assay. Radiat Res. 2023; 199(6):598–615. https://doi.org/10.1667/RADE-22-00206.1 PMID: 37057982
25. Liu W, Li L, Li W. Gene co-expression analysis identifies common modules related to prognosis and drug resistance in cancer cell lines. Int J Cancer. 2014; 135(12):2795–2803. https://doi.org/10.1002/ijc.28935 PMID: 24771271

26. Yang EW, Girke T, Jiang T. Differential gene expression analysis using coexpression and RNA-seq data. Bioinformatics. 2013; 29(17):2153–2161. https://doi.org/10.1093/bioinformatics/btt363 PMID: 23793751

27. Ogino S, Fuchs CS, Giovannucci E. How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. Expert Rev Mol Diagn. 2012; 12(6):621–628. https://doi.org/10.1586/erm.12.46 PMID: 22845482

28. Juruena MF, Cleare AJ, Papadopoulos AS, Poon L, Lightman S, Pariante CM. Different responses to dexamethasone and prednisolone in the same depressed patients. Psychopharmacology (Berl). 2006; 189:225–235. https://doi.org/10.1007/s00213-006-0555-4 PMID: 17016711

29. Waldron L, Haibe-Kains B, Culhane AC, Riester M, Ding J, Wang XV, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst. 2014; 106(5):dju049. https://doi.org/10.1093/jnci/dju049 PMID: 24700801

30. Singh M, Ferrara N. Modeling and predicting clinical efficacy for drugs targeting the tumor milieu. Nat Biotechnol. 2012; 30(7):648–657. https://doi.org/10.1038/nbt.2286 PMID: 22781694

31. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012; 13(1):1–13. https://doi.org/10.1186/1471-2164-13-341 PMID: 22827831

32. He Y, Liu W. TissueSpace: a web tool for rank-based transcriptome representation and its applications in molecular medicine. Genes Genomics. 2022; 44(7):793–799. https://doi.org/10.1007/s13258-022-01245-w PMID: 35511320

33. Angel PW, Rajab N, Deng Y, Pacheco CM, Chen T, Lê Cao KA, et al. A simple, scalable approach to building a cross-platform transcriptome atlas. PLoS Comput Biol. 2020; 16(9):e1008219. https://doi.org/10.1371/journal.pcbi.1008219 PMID: 32986694

34. Yang Q, Hong J, Li Y, Xue W, Li S, Yang H, et al. A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies. Brief Bioinform. 2020; 21(6):2142–2152. https://doi.org/10.1093/bib/bbz137 PMID: 31776543

35. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013; 14(1):1–18. https://doi.org/10.1186/1471-2105-14-91 PMID: 23497356

36. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS ONE. 2011; 6(2):e17238. https://doi.org/10.1371/journal.pone.0017238 PMID: 21386892

37. Sprang M, Andrade-Navarro MA, Fontaine JF. Batch effect detection and correction in RNA-seq data using machine-learning-based automated assessment of quality. BMC Bioinformatics. 2022; 23(6):1–15. https://doi.org/10.1186/s12859-022-04775-y PMID: 35836114

38. List M, Ebert P, Albrecht F. Ten simple rules for developing usable software in computational biology. PLoS Comput Biol. 2017; 13(1):e1005265. https://doi.org/10.1371/journal.pcbi.1005265 PMID: 28056032

39. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013; 9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285 PMID: 24204232

40. Schwab S, Janiaud P, Dayan M, Amrhein V, Panczak R, Palagi PM, et al. Ten simple rules for good research practice. PLoS Comput Biol. 2022; 18(6):e1010139. https://doi.org/10.1371/journal.pcbi.1010139 PMID: 35737655

41. Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023; 24(8):550–572. https://doi.org/10.1038/s41576-023-00586-w PMID: 37002403

42. Pullin JM, McCarthy DJ. A comparison of marker gene selection methods for single-cell RNA sequencing data. Genome Biol. 2024; 25(1):56. https://doi.org/10.1186/s13059-024-03183-0 PMID: 38409056

43. Benayoun B, Singh PP. Considerations for reproducible omics in aging research. Nat Aging. 2023; 3(8). https://doi.org/10.1038/s43587-023-00448-4 PMID: 37386258

44. Lepetit M, Ilie MD, Chanal M, Raverot G, Bertolino P, Arpin C, et al. scAN1. 0: A reproducible and standardized pipeline for processing 10X single cell RNAseq data. In Silico Biol. 2023; 15(1–2):1–11. https://doi.org/10.3233/ISB-210240 PMID: 36278344

45. Cao Y, Chang TG, Sahni S, Ruppin E. Reusability report: Leveraging supervised learning to uncover phenotype-relevant biology from single-cell RNA sequencing data. Nat Mach Intell. 2024; 6:307–314.

46. Markowetz F. Five selfish reasons to work reproducibly. Genome Biol. 2015; 16(1):1–4.

47. Chicco D, Cumbo F, Angione C. Ten quick tips for avoiding pitfalls in multi-omics data integration analyses. PLoS Comput Biol. 2023; 19(7):e1011224. https://doi.org/10.1371/journal.pcbi.1011224 PMID: 37410704

48. Chicco D, Alameer A, Rahmati S, Jurman G. Towards a potential pan-cancer prognostic signature for gene expression based on probesets and ensemble machine learning. BioData Mining. 2022; 15(1):1–23.

49. Chicco D, Sanavia T, Jurman G. Signature literature review reveals AHCY, DPYSL3, and NME1 as the most recurrent prognostic genes for neuroblastoma. BioData Mining. 2023; 16(1):7. https://doi.org/10.1186/s13040-023-00325-1 PMID: 36870971

50. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011; 9(4):e1001046. https://doi.org/10.1371/journal.pbio.1001046 PMID: 21526222

51. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: an R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017; 13(11):e1005752. https://doi.org/10.1371/journal.pcbi.1005752 PMID: 29099853

52. Di Filippo M, Pescini D, Galuzzi BG, Bonanomi M, Gaglio D, Mangano E, et al. INTEGRATE: model-based multi-omics data integration to characterize multi-level metabolic regulation. PLoS Comput Biol. 2022; 18(2):e1009337. https://doi.org/10.1371/journal.pcbi.1009337 PMID: 35130273

53. Lee BD, Gitter A, Greene CS, Raschka S, Maguire F, Titus AJ, et al. Ten quick tips for deep learning in biology. PLoS Comput Biol. 2022; 18(3):e1009803. https://doi.org/10.1371/journal.pcbi.1009803 PMID: 35324884

54. AlNoamany Y, Borghi JA. Towards computational reproducibility: researcher perspectives on the use and sharing of software. PeerJ Comput Sci. 2018; 4:e163. https://doi.org/10.7717/peerj-cs.163 PMID: 33816816

55. Amezquita RA, Lun AT, Becht E, Carey VJ, Carpp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. Nat Methods. 2020; 17(2):137–145. https://doi.org/10.1038/s41592-019-0654-x PMID: 31792435

56. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018; 15(7):475–476. https://doi.org/10.1038/s41592-018-0046-7 PMID: 29967506

57. Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res. 2020; 48 (W1):W395–W402. https://doi.org/10.1093/nar/gkaa434 PMID: 32479607

58. Vaniea K, Rashidi Y. Tales of software updates: the process of updating software. Proceedings of CHI '16 –the 2016 CHI Conference on Human Factors in Computing Systems; 2016. p. 3215–3226.

59. Karimzadeh M, Hoffman MM. Top considerations for creating bioinformatics software documentation. Brief Bioinform. 2018; 19(4):693–699. https://doi.org/10.1093/bib/bbw134 PMID: 28088754

60. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30(1):207–210. https://doi.org/10.1093/nar/30.1.207 PMID: 11752295

61. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2014; 43(D1):D1113–D1116. https://doi.org/10.1093/nar/gku1057 PMID: 25361974

62. Kodama Y, Shumway M, Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012; 40(D1):D54–D56. https://doi.org/10.1093/nar/gkr854 PMID: 22009675

63. Zenodo. Research, shared; 2013 [cited 2023 Nov 15]. Available from: https://www.zenodo.org.

64. Sicilia MA, García-Barriocanal E, Sánchez-Alonso S. Community curation in open dataset repositories: insights from Zenodo. Procedia Comput Sci. 2017; 106:54–60.

65. Kaggle. Kaggle datasets–Explore, analyze, and share quality data. 2022 [cited 2023 Nov 15]. Available from: https://www.kaggle.com/datasets.

66. University of California Irvine. Machine Learning Repository. 1987 [cited 2023 Nov 15]. Available from: https://archive.ics.uci.edu/.

67. Figshare. Store, share, discover research; 2011 [cited 2023 Nov 15]. Available from: https://www.figshare.com.

68. Thelwall M, Kousha K. Figshare: a universal repository for academic resource sharing? Online Information Review. 2016; 40(3):333–346.

69. PhysioNet. The research resource for complex physiologic signals. 2011 [cited 2024 Mar 13]. Available from: https://www.physionet.org/.

70. Moody GB, Mark RG, Goldberger AL. PhysioNet: a web-based resource for the study of physiologic signals. IEEE Eng Med Biol Mag. 2001; 20(3):70–75. https://doi.org/10.1109/51.932728 PMID: 11446213

71. Google. Dataset search. 2023 [cited 2023 Nov 17]. Available from: https://datasetsearch.research.google.com/.

72. re3data. Registry of research data repositories. 2023 [cited 2023 Nov 13]. Available from: https://www.re3data.org.

73. Lopez C, Tucker S, Salameh T, Tucker C. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. J Biomed Inform. 2018; 85:30–39. https://doi.org/10.1016/j.jbi.2018.07.004 PMID: 30016722

74. Meehan TF, Vasilevsky NA, Mungall CJ, Dougall DS, Haendel MA, Blake JA, et al. Ontology based molecular signatures for immune cell types via gene expression analysis. BMC Bioinformatics. 2013; 14 (1):1–15. https://doi.org/10.1186/1471-2105-14-263 PMID: 24004649

75. Pinoli P, Chicco D, Masseroli M. Enhanced probabilistic latent semantic analysis with weighting schemes to predict genomic annotations. Proceedings of IEEE BIBE 2013 –the 13th IEEE International Conference on BioInformatics and BioEngineering. IEEE; 2013, p. 1–4.

76. Chicco D, Masseroli M. Software suite for gene and protein annotation prediction and similarity search. IEEE/ACM Trans Comput Biol Bioinform. 2014; 12(4):837–843.

77. Chicco D, Masseroli M. Ontology-based prediction and prioritization of gene functional annotations. IEEE/ACM Trans Comput Biol Bioinform. 2015; 13(2):248–260.

78. Levitin HM, Yuan J, Cheng YL, Ruiz FJ, Bush EC, Bruce JN, et al. De novo gene signature identification from single-cell RNA-seq with hierarchical Poisson factorization. Mol Syst Biol. 2019; 15(2):e8557. https://doi.org/10.15252/msb.20188557 PMID: 30796088

79. Wilson G. Software Carpentry: Lessons learned. F1000Res. 2014; 3. https://doi.org/10.12688/f1000research.3-62.v2 PMID: 24715981

80. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinform. 2019; 20(6):2044–2054.

81. Marangoni R, Bevilacqua V, Cannataro M, Mele BH, Mauri G, Marabotti A, et al. An overview of bioinformatics courses delivered at the academic level in Italy: reflections and recommendations from BITS. PLoS Comput Biol. 2023; 19(2):e1010846. https://doi.org/10.1371/journal.pcbi.1010846 PMID: 36780436

82. PLOS Computational Biology. Education collection. 2024 [cited 2024 Aug 13]. Available from: https://collections.plos.org/collection/compbiol-education/.

83. Coursera. Top bioinformatics courses. 2024 [cited 2024 Aug 13]. Available from: https://www.coursera.org/search?query=bioinformatics/.

84. Rodriguez-Esteban R, Jiang X. Differential gene expression in disease: a comparison between high-throughput studies and the literature. BMC Med Genomics. 2017; 10:1–10.

85. Bickel DR. Degrees of differential gene expression: detecting biologically significant expression differences and estimating their magnitudes. Bioinformatics. 2004; 20(5):682–688. https://doi.org/10.1093/bioinformatics/btg468 PMID: 15033875

86. Mizuno H, Kitada K, Nakai K, Sarai A. PrognoScan: a new database for meta-analysis of the prognostic value of genes. BMC Med Genomics. 2009; 2(18):1–11.

87. Michiels S, Ternès N, Rotolo F. Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice. Ann Oncol. 2016; 27(12):2160–2167. https://doi.org/10.1093/annonc/mdw307 PMID: 27634691