Yuzuncu Yil University
Journal of the Institute of Natural & Applied Sciences
https://dergipark.org.tr/en/pub/yyufbed

Research Article

Bibliometric Analysis on Methods and Tools Developed for DGE Analysis: Current Trends and Future Perspectives

Necla KOÇHAN*

Izmir University of Economics, Faculty of Arts and Sciences, Department of Mathematics, 35330, Izmir, Türkiye
*Corresponding author e-mail: necla.kochan@ieu.edu.tr

Abstract: Differential gene expression (DGE) analysis has gained significant attention with the advent of next-generation sequencing technologies, leading to the development of a wide range of methods and tools for DGE analysis. We performed bibliometric analysis using Biblioshiny and VOSviewer software to investigate the trends over the investigated period. Relevant papers with differential gene expression related terms as the subjects from 2005 to 2023 were retrieved from the Web of Science database. Network maps were generated using Biblioshiny and VOSviewer software to illustrate the published trends over the investigated period. A total of 729 studies were examined to reveal trends in the DGE analysis methodologies, tools, and packages. In the analysis, co-authorship, bibliographic coupling, and co-occurrence analyses were conducted for country, institution, source, author, and keyword productivity. It was found that the output and citation numbers increased after 2005. During the study period, the USA and China emerged as the leading contributors to the field. The temporal study revealed a significant increase in publications at certain times, followed by period of slight decrease. The greatest fall was observed between 2008 and 2010. Despite these decreases, DGE analysis remains a critical topic in genomics due to its essential role in understanding the mechanisms of any disease, gene function, and therapeutic targets. This trend suggests that current methods and tools are considered sufficiently powerful for identifying key informative genes associated with diverse diseases.

Keywords: Bibliometric analysis, Differential expression analysis, Differential gene expression, Gene expression, RNA-seq

DGE Analizi için Geliştirilen Yöntemler ve Araçlar Üzerine Bibliyometrik Analiz: Güncel Eğilimler ve Gelecek Perspektifleri

Öz: Diferansiyel gen ekspresyonu (DGE) analizi, yeni nesil dizileme teknolojilerinin ortaya çıkışıyla önemli bir ilgi kazanmıştır. Bu durum, DGE analizi için çeşitli yöntemlerin ve araçların geliştirilmesine yol açmıştır. Bu çalışmada, Biblioshiny ve VOSviewer yazılımları kullanılarak, incelenen dönem boyunca eğilimleri araştırmak amacıyla bibliyometrik analiz yapılmıştır. 2005-2023 yılları arasında Web of Science veri tabanından, diferansiyel gen ekspresyonu ile ilgili terimleri konu alan ilgili makaleler taranmıştır. İncelenen dönem boyunca yayımlanan eğilimleri göstermek için Biblioshiny ve VOSviewer yazılımları kullanılarak ağ haritaları oluşturulmuştur. Toplamda 729 çalışma, DGE analizi metodolojilerindeki, araçlarındaki ve paketlerindeki eğilimleri ortaya koymak amacıyla incelenmiştir. Bu amaçla, ülke, kurum, kaynak, yazar ve anahtar kelime üretkenliği açısından eş-yazarlık, bibliyografik eşleşme ve eş-oluşum analizleri yapılmıştır. 2005 yılından sonra çıktı ve atıf sayılarında artış gözlenmiştir. Çalışma süresince ABD ve Çin, DGE analizine en çok katkı sağlayan ülkeler olarak öne çıkmıştır. Zamansal çalışmalar, belirli aralıklarla bir miktar azalma olmakla birlikte, zaman içinde yayınlarda önemli bir artış olduğunu ortaya koymuştur. En büyük düşüş, 2008 ile 2010 yılları arasında gözlenmiştir. Bu düşüşlere rağmen, DGE analizi, herhangi bir hastalığın mekanizmalarını, gen işlevlerini ve terapötik hedefleri anlamadaki temel rolü nedeniyle genomikte kritik bir konu olmaya devam etmektedir. Bu eğilim, mevcut yöntemlerin ve araçların, çeşitli hastalıklarla ilişkili anahtar bilgilendirici genleri tanımlamak için yeterince güçlü kabul edildiğini göstermektedir.

Anahtar Kelimeler: Bibliyometrik analiz, Diferansiyel dizileme analizi, Diferansiyel gen ekspresyonu, Gen ifadesi, RNA-dizileme

Received: 27.11.2024
Accepted: 24.02.2025

How to cite: Koçhan, N. (2025). Bibliometric analysis on methods and tools developed for DGE analysis: Current trends and future perspectives. *Yuzuncu Yil University Journal of the Institute of Natural and Applied Sciences*, *30*(1), 78-91. https://doi.org/10.53433/yyufbed.1591489

1. Introduction

Differential gene expression (DGE) analysis has become increasingly popular in recent decades due to the development of high-throughput sequencing technologies such as next-generation sequencing, microarrays, proteomics, and metabolomics. DGE analysis is a common method frequently used to compare gene expression profiles among various samples/groups, such as healthy vs disease or control vs treatment or cells subjected to different treatments (Kebschull et al., 2017). The primary goal is to identify the informative genes whose patterns vary across samples. These genes, differentially expressed, are called biomarkers. These offer insights into the underlying mechanism of disease and gene regulation, and can be used to develop further treatments and targeted therapies. Additionally, DGE analysis plays a crucial role in drug discovery by providing insights into how drugs influence gene activity and cellular pathways. By monitoring gene expression changes in response to drug treatments, researchers can identify specific genes and pathways affected by a drug, helping to uncover potential therapeutic targets and biomarkers (Bai et al., 2013; Melouane et al., 2018).

Despite its importance, differentially expressed gene discovery can be challenging since gene expression data consists of thousands of genes and a relatively small number of samples. Numerous techniques have been developed to identify the genes that differ statistically between comparable groups or samples. To this end, a number of software applications and/or R packages have been introduced. One of the most powerful and efficient tools for RNA-seq analysis is edgeR (Empirical Analysis of Digital Gene Expression data in R) (Robinson et al., 2009; 2010), which is widely used due to its low false discovery rate (Robinson et al., 2009; 2010). Another common tool is DESeq2 (Differential Expression analysis for Sequence Count data) (Love et al., 2014), which is frequently used for datasets produced from a large number of samples with low variability. Cephe et al. (2023) and Rosati et al. (2024) provide a list of tools and approaches offered thus far, as well as elaborating on the advantages and disadvantages of DEG analysis methods. Rosati et al. (2024) provided an in-depth analysis of the bioinformatic pipelines and computational methods developed for DGE analysis, highlighting their strengths and limitations in biomarker discovery. They emphasizethe importance of selecting appropriate analytical strategies to enhance DGE studies' accuracy and reliability, which are crucial in advancing personalized medicine and understanding complex biological processes. The review serves as a valuable resource for researchers aiming to navigate the complexities of DGE analysis and its applications in biomarker identification. Similarly, Clark & Lillard (2024) explored the significance of these tools in analyzing complex genomic data to identify biomarkers that can guide personalized treatment strategies in precision oncology.

Recent studies have demonstrated that machine learning and deep learning approaches have begun to play an important role in DGE analysis, enabling the investigation of complex diseases, as well as the identification of genes or biomarkers for prognosis and/or diagnosis (Liñares Blanco et al., 2019; Mahendran et al., 2020; Dhillon et al., 2023). These approaches have great potential in the processing of high-dimensional gene expression data and in improving prediction accuracy.

As highlighted by the studies above, the rapid advancement of high-throughput sequencing technologies has significantly expanded research on DGE analysis tools and methodologies. Despite this extensive research, there is a lack of bibliometric studies assessing trends, collaborations, and research focus in this field. To address this gap, we conducted a bibliometric analysis of DGE research from 2005 to 2023, examining publication trends, international collaborations, keyword patterns, and emerging research directions. This study's comprehensive overview of the field's evolution offers insights into future developments, which is crucial for identifying research priorities, guiding funding decisions, future studies, and fostering interdisciplinary collaboration in gene expression analysis. Given the pivotal role of DGE in biomarker discovery, disease classification, and personalized medicine, understanding research trends can help accelerate the development of targeted therapies and improve clinical outcomes.

2. Material and Methods

2.1. Search strategy, data collection and data analysis

Data were obtained from the Web of Science (WoS) database and the search period included the scientific outputs from January 1, 2005 to December 31, 2023. A comprehensive online search was

carried out using the following search strategy: (AB=(("differential-expression" OR "differential expression" OR "differentially expressed" OR "differential analysis" OR "differential analyses" OR "differentially gene expression" OR "differential gene expression analyses" OR "differential gene expression")) AND AB=(("gene-expression" OR "RNA-Sequencing" OR "RNA-Seq" OR "microarray" OR "RNA-seq" OR "gene expression" OR "RNA-sequencing" OR "RNA sequencing")) AND AB=(("method" OR "approach" OR "software" OR "algorithm" OR "statistical" OR "estimation" OR "modeling" OR "technique")) NOT TI=(("normalization" OR "transformation" OR "alignment" OR "outlier")) AND TI=(("differential-expression" OR "differential expression" OR "differentially expressed" OR "RNA-Seq" OR "differential analysis" OR "differential analyses" OR "differentially gene expression" OR "differential gene expression analyses" OR "differential gene expression"))) AND (DT==("ARTICLE" OR "REVIEW") AND LA==("ENGLISH")). All records obtained from the literature using this strategy were exported to a Plain Text File (.txt). This query resulted in a total of 1881 outputs, however, after titles and abstracts were screened for relevancy, only 729 of these were considered for biometric analysis. This filtering step was necessary prior to the bibliometric analysis given the large number of studies that employed only current approaches for DGE analysis.

2.2. Bibliometric analysis

The R-based **Bibliometrix** package (Biblioshiny v.4.1.4, www.bibliometrix.org) (Aria & Cuccurullo, 2017), the VOSviewer software (v.1.6.20) (van Eck & Waltman, 2010) and the R programming language (v.4.2.2) were used to conduct bibliometric analysis. The final dataset of 729 papers was analyzed after filtering by year and language using the **Bibliometrix** package. In line with the study's primary focus, the analyses focused on the annual scientific production of countries and institutions, most cited articles, most influential authors, trending topics, and common word analysis.

3. Results

3.1. General information about data

After year, language and document type filters were applied to the bibliometric analysis, a total of 729 differential expression analysis studies, consisting of 723 articles and 6 reviews, were included (Table 1). Figure 1 reports annual numbers of publications from 2005 to 2023, revealing fluctuations in the numbers. Peaks with noticeable increases are seen around 2014, and again in 2021, followed by a decline after 2022.

Table 1. Main data information

Main Information About Data	**Results**
Timespan	2005:2023
Sources (Journals, Books, etc)	184
Documents	729
Annual Growth Rate %	2.28
Document Average Age	8.83
Average citations per doc	205.6
References	14937
Document Contents	
Keywords Plus (ID)	1152
Author's Keywords (DE)	1187
Authors	
Authors	2781
Authors of single-authored docs	23

Table 1. Main data information (continued)

Main Information About Data	Results
Authors Collaboration	
Single-authored docs	26
Co-Authors per Doc	4.67
International co-authorships %	21.54
Document Types	
article	656
article; book chapter	17
article; data paper	1
article; early access	2
article; proceedings paper	47
review	5
review; book chapter	1

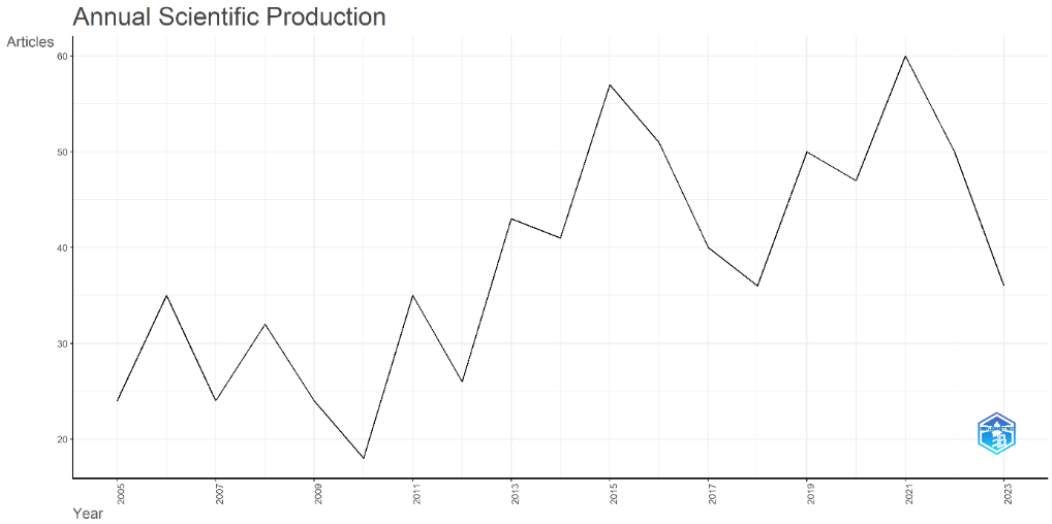

Figure 1. Annual number of publications on developing methods, tools or packages for differential expression analysis.

3.2. Countries and institutions analysis

In total, 51 countries and 810 institutions were encompassed in differential-expression studies research. The 10 countries display a global distribution, including the USA, China, Austria, England, and Germany (Table 2). The USA recorded the highest number of publications (n=358), followed by China (n=133), and Australia (n=46). Co-authorship country analysis was carried out to shed light on international collaborations. Table 2 shows each country's co-authorship relationships with other countries, as well as its overall strength. Notably, the USA leads with 358 publications and a total link strength of 113. In the network representation, each circle represents an author's country, and font size denotes the frequency of collaboration (Figure 2). The size of the nodes in Figure 2 reflects the number of collaborations from each country, while the connecting lines indicate the strength of collaboration. Similar research fields for countries are represented by the same colors. For instance, the USA has many strong collaborations with China (Figure 2).

Table 2. The topmost 10 strong co-authorship linked document-productive countries

Rank	Countries	Clusters	Links	Total link strength	Documents	Citations
1	USA	4	21	113	358	104591
2	China	2	8	56	133	2322
3	Australia	2	12	35	46	62694
4	England	1	14	35	40	7830
5	Germany	5	15	33	38	51625
6	South Korea	4	2	10	28	329
7	Japan	2	3	6	24	399
8	India	4	3	10	22	771
9	Canada	1	7	12	20	271
10	France	1	5	9	20	1141

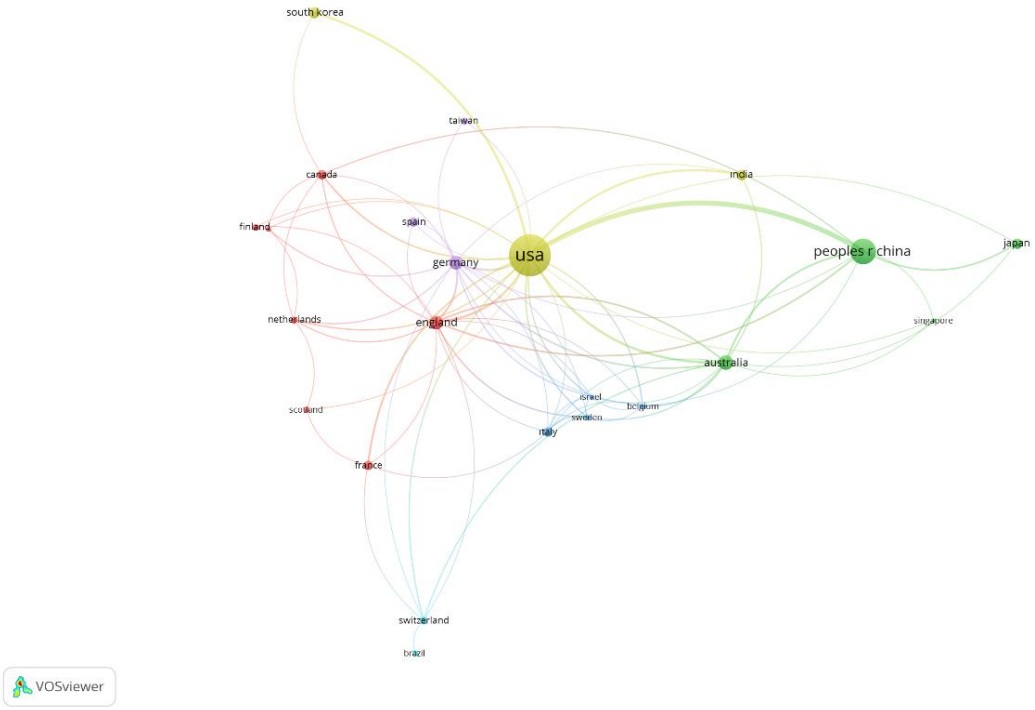

Figure 2. Map of visualization of countries on research of differential expression analysis methods.

The analysis identified 6 clusters, with a total of 76 links and 202 total link strength. Cluster 1 (red) shows significant co-authorship for, England (n=40), Canada (n=20), France (n=20), Finland (n=11), Netherlands (n=10), Russia (n=6), and Scotland (n=8). Cluster 2 (green) is characterized by a high level of cooperation among Australia (n=46), Japan (n=24), China (n=133), and Singapore. Differential expression analysis research reveals a close relationship among Belgium (n=6), Israel (n=5), Italy (n=18), and Sweden (n=9) in cluster 3 (Blue). Cluster 4 (yellow) is characterized by strong ties among the United States (n=358), South Korea (n=28), and India (n=22). In cluster 5 (purple), Germany (n=38), Spain (n=18), and Taiwan (10) have significant co-authorship. In addition, Brazil (n=9) and Switzerland (13) collaborate closely in cluster 6 (light blue).

According to the most frequent affiliations stated by the corresponding authors, as shown in Figure 3**,** University of California System has the highest number of papers (n=63), closely followed by University of Texas System (55 papers), Harvard University (54 papers), and John Hopkins University (47 papers).

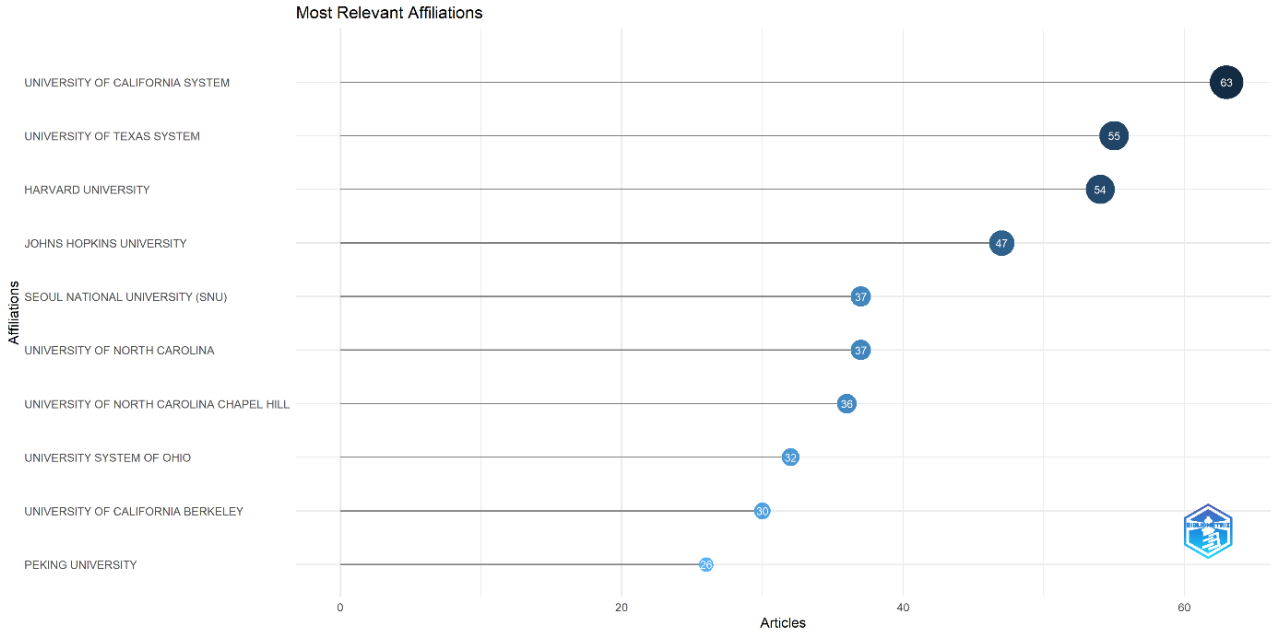

Figure 3. Most relevant institutions.

Figure 4 depicts the connection between different institutions, countries, and publications. The height of the rectangular boxes indicates the frequency of the appearance of a particular institution, country, or journal within the collaborative network. The findings demonstrated that the United States is the leading country, with all ten of the highest ranking institutions. Harvard University, John Hopkins University, and the University of California System are among the most notable instances. China and Australia, respectively, are the next most prominent countries.

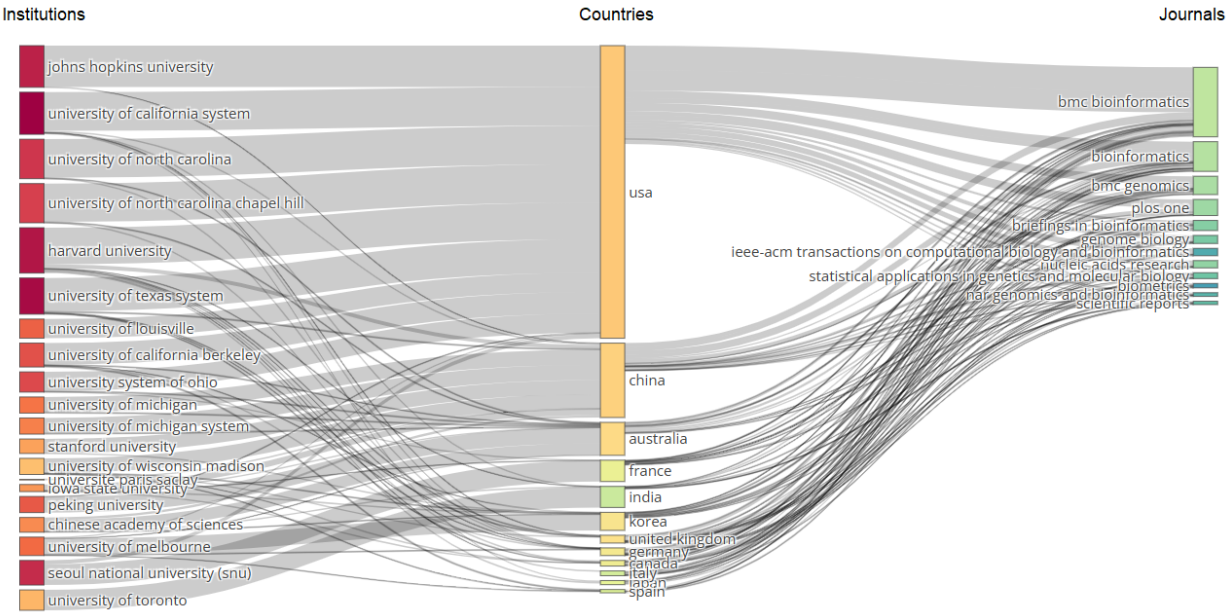

Figure 4. Three-field plot displaying the network between institutions (left), countries (middle), and journals (right).

3.3. Bibliographic coupling with sources

In total, 184 sources were created from scientific outputs and the complete counting method was applied with a minimum threshold of 5. Only 27 outputs were able to meet these criteria, and for each

of these, the overall strength of bibliographic linkages to other sources was calculated (see Table 3, Figure 5). The analysis showed 351 links and a total connection strength of 130101, forming 4 clusters of 27 items. The clusters consisted of 14, 9, 3, and 1 elements, respectively.

Sources in the first cluster include Annals of Applied Statistics (n=5), Bioinformatics (n=84), Biometrics (n=9), Biostatistics (n=8), BMC Bioinformatics (n=102), Cancer Informatics (n=6), Computational Biology and Chemistry (n=6), Computational Statistics & Data Analysis (n=9), Genomics (7), IEEE-ACM Transactions on Computational Biology and Bioinformatics (n=10), Journal of Computational Biology (n=9), Plos One (n=39), Statistical Applications in Genetics and Molecular Biology (n=14), and Statistics in Biosciences (n=7). Sources in the second cluster include Briefings in Bioinformatics (n=19), Frontiers in Genetics (n=9), Genome Biology (n=14), Methods (n=8), Nar Genomics and Bioinformatics (n=11), Nature Communications (n=8), Peerj (n=9), Plos Computational Biology (n=8), and Scientific Reports (n=13). In the third cluster, sources include BMC Genomics (n=46), Journal of Bioinformatics and Computational Biology (n=6), and Nucleic Acids Research (n=28), while in the fourth cluster, Current Bioinformatics (n=5) has the highest overall link strength.

Table 3. The topmost 10 strong bibliographic coupling with sources

Rank	Sources	Clusters	Links	Total link strength	Documents	Citations
1	BMC Bioinformatics	1	26	43745	102	3686
2	Bioinformatics	1	26	36866	84	33561
3	BMC Genomics	3	26	24039	46	1028
4	Plos One	1	26	22249	39	1791
5	Nucleic Acids Research	3	26	19500	28	28342
6	Briefings in Bioinformatics	2	26	9732	19	543
7	Genome Biology	2	26	10233	14	44638
8	Statistical Applications in Genetics and Molecular Biology	1	26	7769	14	460
9	Scientific Reports	2	26	7300	13	383
10	Nar Genomics and Bioinformatics	2	26	6361	11	629

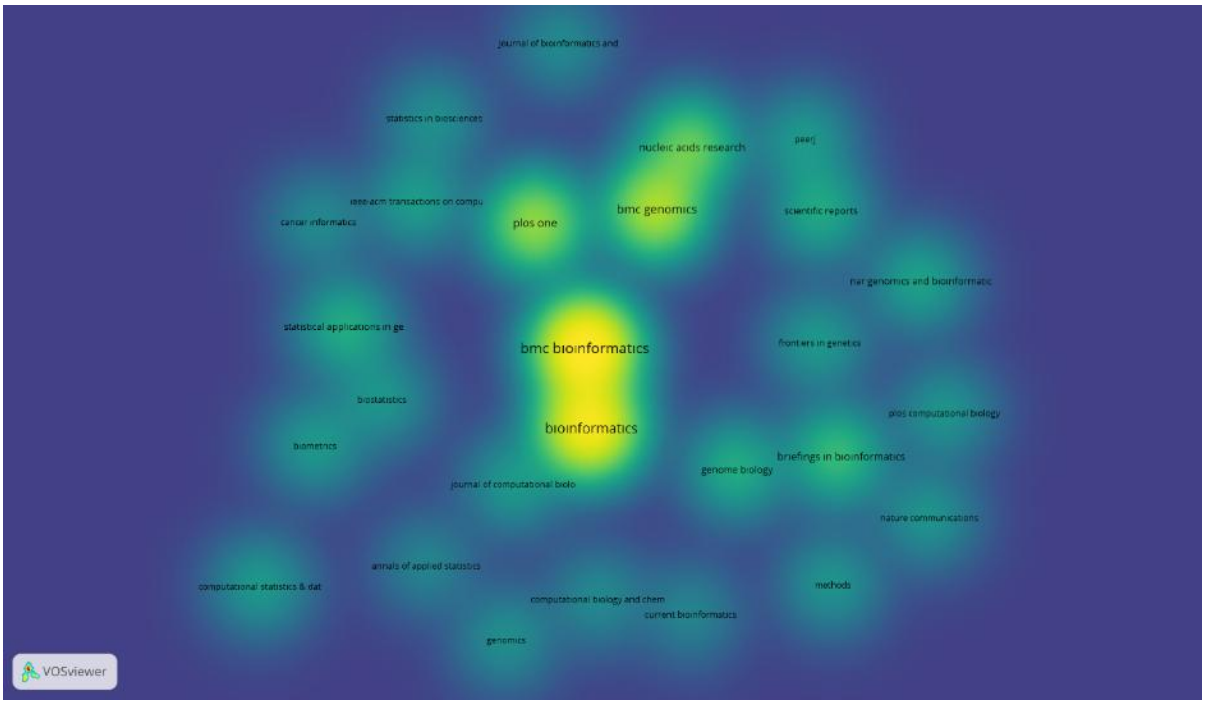

Figure 5. Bibliographic coupling with sources.

Among the 10 leading journals, as indicated in Table 4, “BMC Bioinformatics” is notable for its significant contribution to the categories of Biochemical Research Methods, Biotechnology and Applied Microbiology, Mathematical and Computational Biology. It is ranked first in the number of published documents (102) in these fields, and its Journal Impact Factor (JIF) percentile is 76.5, placing it in Q1. However, this journal's most recent impact factor is 2.9, considerably lower than journals such as "Nucleic Acids Research" and "Briefings in Bioinformatics", which have JIF percentiles of 98.2 and 95.9, respectively, positioning them as more authoritative.

Table 4. Top 10 journals with the highest number of published papers

Rank	Journals	Documents	JIF Percentile	JIF Quartile
1	BMC Bioinformatics	102	76.5	Q1
2	Bioinformatics	84	90.2	Q1
3	BMC Genomics	46	67.8	Q2
4	Plos One	39	76.5	Q1
5	Nucleic Acids Research	28	98.2	Q1
6	Briefings in Bioinformatics	19	95.9	Q1
7	Genome Biology	14	95.1	Q1
8	Statistical Applications in Genetics and Molecular Biology	14	34.8	Q3
9	Scientific Reports	13	81.7	Q1
10	Nar Genomics and Bioinformatics	11	88.6	Q1

JIF: Journal Impact Factor

The 10 journals with the most publications in the field of differential expression analysis are reported in Figure 6a. It was found that "BMC Bioinformatics" journal was the leader, 102 articles, followed by "Bioinformatics" with 84 articles. Figure 6b depicts the earliest scientific research on differential expression analysis conducted in the Bradford area. Journals in this core region are regarded as top-tier publications in the field of differential expression analysis, serving as the foundation for the development of most later articles in the field. This chart serves to identify the most productive journals in the field of differential expression analysis studies.

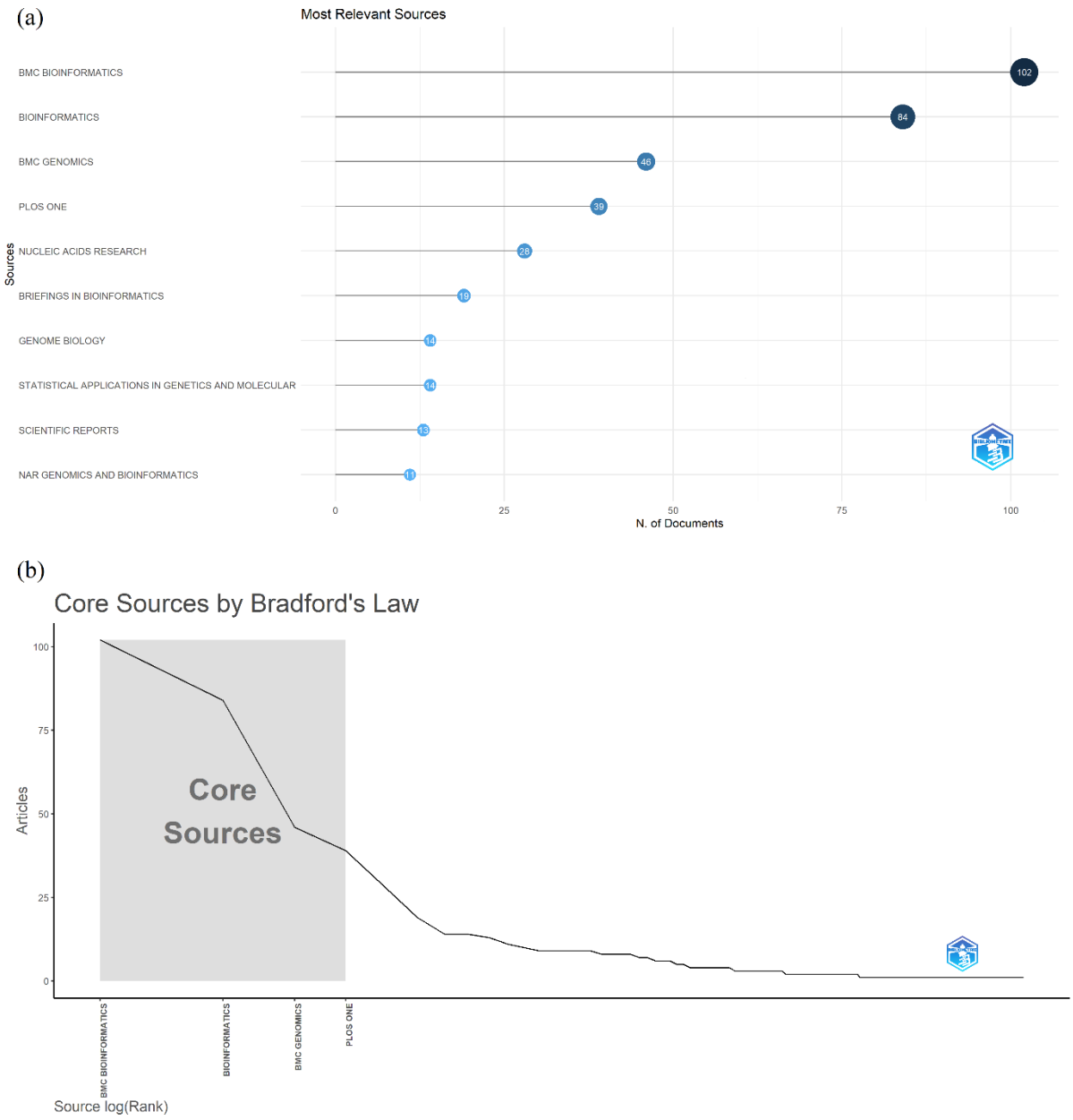

Figure 6. (a) Most relevant sources, (b) Core Sources by Bradford's Law.

3.4. Highly cited articles and most prolific authors

Figure 7a shows the most highly cited research articles. The paper receiving the most citations in the field was "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2" by Love et al. (2014), with a total of 43481 during the investigation period. This is followed by Robinson et al. (2010) 's publication "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", which has 26930 citations.

Figure 7b displays the most prolific authors in the field of differential-expression analysis. The density of circle in the visualization represents the author' number of citations for the year, while the size of the circle represents the number of published papers. As the number of citations and publications increases, the darkness and size of the circles increase in direct proportion. It is seen that GK Smyth constantly published articles from 2005 until 2017. The most trending year for GK Smyth is 2015. In 2015, there is a substantially larger and darker dot, indicating that GK Smyth authored many publications with a high citation impact (1000 or more). The second most prolific author is Nettleton D,

who consistently published articles from 2005 to 2020, followed by Zhang Y (2008-2022) and Elo LL (2009-2022).

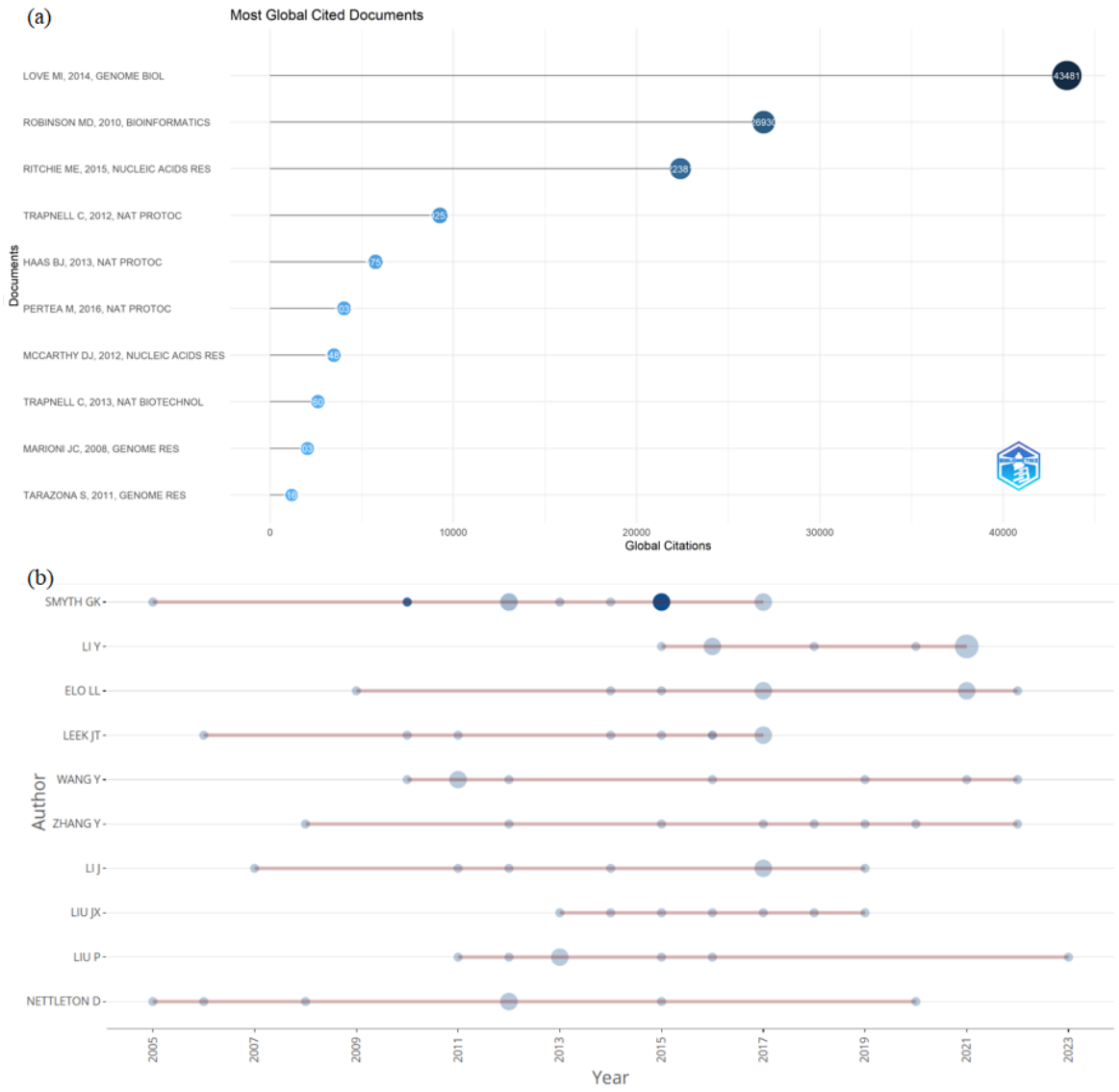

Figure 7. (a) Most global cited documents and (b) production of the most productive authors over time.

3.5. Keyword and trend topics analysis

Keyword analysis is critical for gaining insight into the key topics and trends suitable for further exploration in a research field. This type of analysis enables an understanding of the most discussed subjects and significant topics. Figure 8a shows a word cloud with the commonly used keywords in the relevant subject, including terms like "gene-expression," "microarray," "differential expression," and "rna-seq." This indicates the key areas of application/use of microarray and RNA-seq data in DGE analysis.

Figure 8b depicts the findings of the keyword co-occurrence network analysis, demonstrating that frequently used keywords are divided into six clusters. The results suggest that DGE analysis incorporates concepts from the fields of statistics, biology, computational biology, and genetics, demonstrating its intrinsic multidisciplinary nature**.** Figure 8c displays trend topics across years (2004–2022) with topics on the y-axis and years on the x-axis. Each topic is shown by a horizontal line that

depicts the period of its relevance, with larger bubble sizes representing higher term frequency. The three most trending topics in the field are differentially expressed genes from 2011 to 2020, gene expression from 2009 to 2018 and differential gene expression from 2014 to 2021. Figure 8c indicates that the focus of research is shifting away from statistical approaches such as multiple tests towards more sophisticated computational tools such as machine learning approaches.

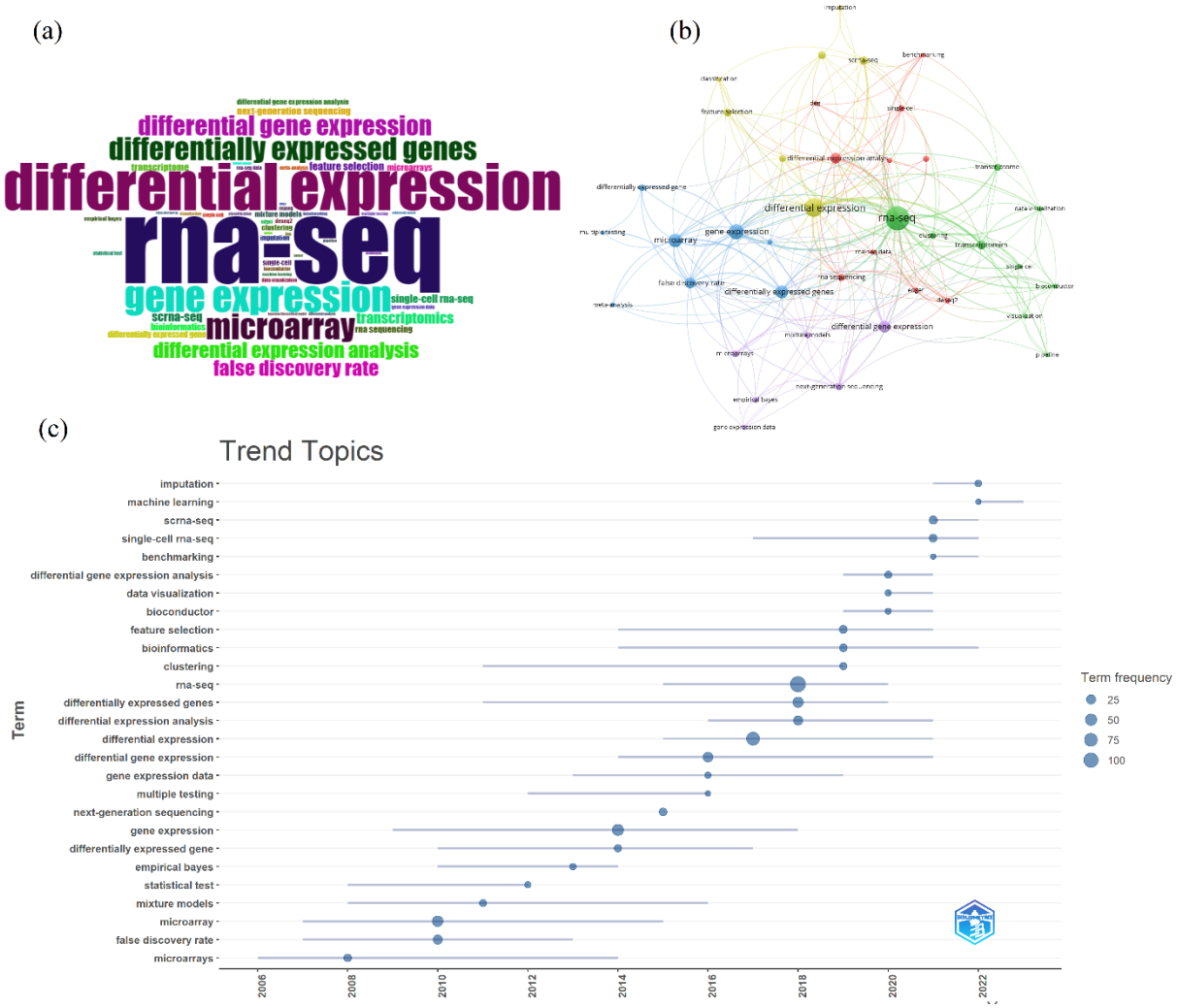

Figure 8. (a) Word cloud of the most frequent terms in the author's keywords of selected articles, (b) keyword co-occurrence network in differential-expression analysis research, and (c) trend topics across years.

4. Discussion and Conclusion

Differential gene expression analysis is a technique used in bioinformatics to better understand the complex mechanisms underlying various biological processes. Researchers can identify differentially expressed genes by comparing gene expression levels across various groups, such as healthy versus diseased tissues subjected to different treatments. This offers insight into fundamental biological processes and facilitates the identification of genes associated with specific biological processes, particularly in the context of identifying biomarkers for diagnosis or prognosis. As a result, many methods/tools have been developed for DGE analysis (Hardcastle & Kelly, 2010; Di et al., 2011; Kvam et al., 2012; Rapaport et al., 2013; Law et al., 2014; Seyednasrollah et al., 2015; Chowdhury et al., 2020; Costa-Silva et al., 2022). Understanding how these methodologies have evolved over time is crucial for identifying key contributors, emerging trends, and potential research gaps. This study presents a detailed analysis of articles proposing methodologies and tools for DGE analysis between 2005 and 2023, in terms of sources, authors, institutions, countries, keywords, and clusters. Furthermore,

we present an overview of the shifting trends in the literature on these topics. The data reveal that, despite fluctuations in interest in the development of methods and tools, DGE analysis remains critical in genomics due to its importance in disease processes, gene function, and therapeutic targets. This fact is not disproven by the relative fall in interest in developing tools and methods for DGE analysis, but rather, it implies that the current methods and tools are considered sufficiently powerful to identify important informative genes associated with specific disorders.

There was an increase in the number of studies between 2010 and 2015 (Figure 1), due to an increasing interest and demand for more advanced computational methods and tools for analyzing gene expression data. During this period, there was greater interest in high-throughput sequencing technologies like RNA-seq. This resulted in an exponential growth in the volume and complexity of transcriptomic data, prompting the development of novel algorithms, statistical approaches, and software tools for efficient analysis. There was a slight temporary decline in the number of studies after 2015, which was an unexpected trend. We hypothesize that this may be dure to the maturation of existing methods/tools, where current tools are deemed sufficiently robust for most applications. However, DGE analysis regained its importance after a few years. This increase peaked in 2021. The studies at this time were mostly published in top-tier journals, including BMC Bioinformatics, Bioinformatics, BMC Genomics, Nucleic Acids Research, Briefings in Bioinformatics, etc. Publication in these widely recognized journals in genetics and bioinformatics confirms the high scientific impact and credibility of the DGE analysis.

The University of California System is the leading associated institution, and the United States, China, and Australia are the main contributors. With the most publications and citations, the USA is clearly leading in the field in developing methods/tools, driven by the country's robust research infrastructure, prestigious universities/institutions such as the University of California System, along with substantial funding for genomics and bioinformatics. The USA's number of publications and citations reflects its major contributions to developing and applying computational tools to innovative research.

The paper "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2" by Love et al. (2014) received the most citations in the field of differential-expression analysis, with 43481 citations), followed by Robinson et al. (2010)'s publication "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data", with 26930 citations. GK Smyth, among the authors of the second and third most cited papers, was the most prolific contributor in the field, highlighting his role in promoting and increasing the efficiency of the methods and tools for DGE analysis. This demonstrates Smyth contributions' substantial impact on advancing the field, which is in line with our expectations.

The simplest DE detection techniques employ a test statistic to determine whether genes exhibit a statistically significant change in gene expression under various conditions. While it is possible to use non-parametric methods (Li & Tibshirani, 2013) for this purpose, due to the small number of replicates, non-parametric methods may not be sufficiently efficient to detect those genes that are differentially expressed. To overcome this, parametric methods are utilized (Robles et al., 2012; Seyednasrollah et al., 2015). The most frequently used parametric methods in analysis tools are based on the Poisson and Negative Binomial (NB) distributions. For instance, edgeR uses NB distribution for DGE analysis, since NB distribution accounts for the overdispersion commonly observed in RNA-Seq data. However, trend topic analysis shows a shift towards machine learning and/or deep learning techniques in recent years (Figure 8c). These more flexible and powerful approaches have the advantage of being able to learn complex patterns in the data without any assumptions about the underlying distribution. Machine learning models can more effectively process high-dimensional, noisy data, offering more accurate and robust predictions for DE detection, especially with a small number of replicates, or highly complex data.

Despite its enormous potential use, differential gene expression analysis poses several challenges, including data normalization, batch effects, and the need for robust/strong statistical approaches (Chowdhury et al., 2020; Costa-Silva et al., 2022). Development including integrating multi-omics data and developing sophisticated computational tools will further enhance the accuracy, reliability and biological interpretation of differential gene expression analysis.

Acknowledgements

I would like to thank to Dr. Ahu Cephe for her invaluable recommendations.

References

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

Bai, J. P. F., Alekseyenko, A. V., Statnikov, A., Wang, I. M., & Wong, P. H. (2013). Strategic applications of gene expression: From drug discovery/development to bedside. *The AAPS Journal*, *15*(2), 427-437. https://doi.org/10.1208/s12248-012-9447-1

Cephe, A., Koçhan, N., Ertürk Zararsız, G., Eldem, V., & Zararsız, G. (2023). Class discovery, comparison, and prediction methods for RNA-Seq data. In J. Wang (Ed.), *Encyclopedia of Data Science and Machine Learning* (pp. 2060-2084). IGI Global. https://doi.org/10.4018/978-1-7998-9220-5.ch123

Chowdhury, H. A., Bhattacharyya, D. K., & Kalita, J. K. (2020). (Differential) Co-expression analysis of gene expression: A survey of best practices. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17*(4), 1154-1173. https://doi.org/10.1109/TCBB.2019.2893170

Clark, A. J., & Lillard, J. W., Jr. (2024). A comprehensive review of bioinformatics tools for genomic biomarker discovery driving precision oncology. *Genes, 15*(8), 1036. https://doi.org/10.3390/genes15081036

Costa-Silva, J., Domingues, D. S., Menotti, D., Hungria, M., & Lopes, F. M. (2022). Temporal progress of gene expression analysis with RNA-Seq data: A review on the relationship between computational methods. *Computational and Structural Biotechnology Journal*, *21*, 86-98. https://doi.org/10.1016/j.csbj.2022.11.051

Dhillon, A., Singh, A., & Bhalla, V. K. (2023). A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: From computational needs to machine learning and deep learning. *Archives of Computational Methods in Engineering, 30*, 917-949. https://doi.org/10.1007/s11831-022-09821-9

Di, Y., Schafer, D. W., Cumbie, J. S., & Chang, J. H. (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology, 10*(1). https://doi.org/10.2202/1544-6115.1637

Hardcastle, T. J., & Kelly, K. A. (2010). baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics, 11*, 422. https://doi.org/10.1186/1471-2105-11-422

Kebschull, M., Fittler, M. J., Demmer, R. T., & Papapanou, P. N. (2017). Differential expression and functional analysis of high-throughput-omics data using open source tools. *Methods in Molecular Biology, 1537*, 327-345. https://doi.org/10.1007/978-1-4939-6685-1_19

Kvam, V. M., Liu, P., & Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *American Journal of Botany, 99*(2), 248-256. https://doi.org/10.3732/ajb.1100340

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology, 15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Li, J., & Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research, 22*(5), 519-536. https://doi.org/10.1177/0962280211428386

Liñares Blanco, J., Gestal, M., Dorado, J., & Fernandez-Lozano, C. (2019). Differential gene expression analysis of RNA-seq data using machine learning for cancer research. In G. A. Tsihrintzis, M. Virvou, E. Sakkopoulos, & L. Jain (Eds.), *Machine Learning Paradigms* (Vol. 1, pp. 43–63). Springer, Cham. https://doi.org/10.1007/978-3-030-15628-2_3

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*, 550. https://doi.org/10.1186/s13059-014-0550-8

Mahendran, N., Vincent, P. M. D. R., Srinivasan, K., & Chang, C. (2020). Machine learning based computational gene selection models: a survey, performance evaluation, open issues, and future research directions. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.603808

Melouane, A., Ghanemi, A., Aubé, S., Yoshioka, M., & St-Amand, J. (2018). Differential gene expression analysis in ageing muscle and drug discovery perspectives. *Ageing Research Reviews*, *41*, 53-63. https://doi.org/10.1016/j.arr.2017.10.006

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., & Betel, D. (2013). Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, *14*, 3158. https://doi.org/10.1186/gb-2013-14-9-r95

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). *edgeR: Empirical analysis of digital gene expression data in R*. Bioconductor. https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616

Robles, J. A., Qureshi, S. E., Stephen, S. J., Wilson, S. R., Burden, C. J., & Taylor, J. M. (2012). Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, *13*(1), 484. https://doi.org/10.1186/1471-2164-13-484

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and Structural Biotechnology Journal*, *23*, 1154-1168. https://doi.org/10.1016/j.csbj.2024.02.018

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2015). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, *16*(1), 59-70. https://doi.org/10.1093/bib/bbt086

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3