EDITORIAL
published: 26 May 2022
doi: 10.3389/fgene.2022.928724

Editorial: Advancement in Gene Set Analysis: Gaining Insight From High-Throughput Data

Farhad Maleki[1]*, *Sorin Draghici*[2], *Renee Menezes*[3] and *Anthony Kusalik*[4]

[1]Augmented Intelligence & Precision Health Laboratory, Department of Radiology and Research Institute of the McGill University Health Centre, Montreal, QC, Canada, [2]Department of Computer Science, Wayne State University, Detroit, MI, United States, [3]Biostatistics Centre and Department of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, Netherlands, [4]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

Keywords: gene set analysis, pathway analysis, enrichment analysis, gene expression, next-generation sequencing

Editorial on the Research Topic

Advancement in Gene Set Analysis: Gaining Insight from High-Throughput Data

The existence of high-throughput technologies allows for the study of a large number of genes in a single experiment. However, analyzing such high-throughput data and interpreting the results are challenging (Draghici, 2016).

Phenotypes or biological conditions often result from the coordinated activity of a group of genes or biomolecules. Consequently, the study of the coordinated expression pattern of biologically related genes is essential for understanding the mechanisms underlying these conditions or phenotypes. Knowledge bases such as GO (Consortium, 2004) and KEGG (Kanehisa and Goto, 2000) aim to capture knowledge about the roles that genes play in various biological processes and locations. Such resources can be generally divided into: 1) gene set databases (e.g., GO), which include only associations between genes and annotations such as biological processes; and 2) pathway databases (e.g., KEGG), which also capture knowledge related to the interactions between the genes.

Various categories of methods have been developed over time to extract knowledge from such resources (Maleki et al., 2020). The very first methods used a simple approach to identify the gene sets that are enriched in differentially expressed genes (Khatri et al., 2002; Dennis et al., 2003; Draghici et al., 2003b). This approach has various limitations including the fact that it ignores the magnitude of the measured gene expressions. This was addressed by the second generation of methods, pioneered by GSEA (Subramanian et al., 2005), and called functional class scoring (FCS). FCS methods use the correlation between gene expression and the phenotype but still ignore all the interactions between genes. This was addressed by the third generation of methods, called topology-based, or pathway analysis methods. The first such method, impact analysis (Draghici et al., 2007; Tarca et al., 2009), was soon followed by a plethora of over 20 other approaches (Khatri et al., 2012; Mitrea et al., 2013; Nguyen et al., 2018). Many of these methods have been bench-marked recently (Nguyen et al., 2019).

Even though pathway analysis methods are very different from enrichment and FCS methods, we will use "gene set analysis" to generically refer to the entire family of methods aimed at understanding the coordinated expression pattern of known gene sets or pathways. Despite the widespread use of gene set analysis, little consensus exists in the research community regarding best practices. This Research Topic is aimed at highlighting methodological advances as well as applications of gene set analysis to improve the utility of these methods in gaining insight from high-throughput expression studies. Highlights are as follows.

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

****Correspondence:***
Farhad Maleki
farhad.maleki@mail.mcgill.ca

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *26 April 2022*
Accepted: *06 May 2022*
Published: *26 May 2022*

Citation:
Maleki F, Draghici S, Menezes R and Kusalik A (2022) Editorial: Advancement in Gene Set Analysis: Gaining Insight From High-Throughput Data. Front. Genet. 13:928724. doi: 10.3389/fgene.2022.928724

Testing for case-control gene expression differences between two groups is a common approach in studies in which researchers are interested in the "difference of differences". Weiner et al. describe a frequent methodological error in using and interpreting gene set analysis methods for such studies. The error occurs when researchers test for differential expression separately in each group and consider genes with significant expression differences in only one comparison—i.e., one group—specific to that group. Based on this assumption, a gene set enrichment analysis is used to find gene sets/ pathways specific to only one group. Weiner et al. empirically show that such an approach could report differentially enriched gene sets even for scenarios with no statistically significant differences between the groups.

Marczyk et al. evaluate the effect of incorporating different approaches for integrating single-nucleotide polymorphism (SNP) information and linkage disequilibrium correction on the performance of several gene set analysis methods. They suggest that linkage disequilibrium correction and Stouffer integration could improve the performance of gene set analysis for genome-wide association studies.

Several articles focus on gene set analysis for cancer research. Luo et al. use GSEA (Subramanian et al., 2005) to study the pathways associated with DNA methylation-derived differentially expressed genes in patients with prostate cancer. Song et al. also identify a ubiquitin-related gene signature for prostate cancer prognosis. Li et al. study the association of S100 genes with well-known tumor-related pathways. Xu et al. utilize gene set analysis to identify biological functions and pathways associated with the ferroptosis-related genes in patients with skin cutaneous melanoma. Tan et al. use GSEA to identify gene sets associated with genes co-expressed with the SBSN gene. He et al. find genes differentially expressed in patients with renal cell carcinoma to be associated with autophagy-related pathways. They suggest a prognosis risk score for renal cell carcinoma based on autophagy-related genes that are differentially expressed in patients with the cancer.

The applications of gene set analysis are not limited to cancer research. Yousef et al. employ gene set analysis to validate the biological relevance of the results of their algorithm for miRNA-mRNA regulatory module detection. Du et al. identify hub genes and pathways implicated in osteoporosis. Wu et al. explore potential hub genes in non-alcoholic fatty liver disease and gene sets associated with these genes.

Due to the complex nature of gene set analysis, developing tools that conduct gene set analysis and facilitate interpreting its results is valuable. Among tools commonly used for gene set analysis are DAVID (Dennis et al., 2003), Enrichr (Kuleshov et al., 2016), WebGestalt (Liao et al., 2019), iPathwayGuide (Ahsan and Draghici, 2017), and Onto-Tools (Draghici et al., 2003a). In this Research Topic, Yue et al. present "PAGER Web APP" as an interactive web-based application supporting online R scripting of integrative gene set analysis, and Odom et al. develop an R Package for integrative analysis of multi-omics datasets offering the functionality to work with matched or non-matched samples.

Despite the existence of a large number of gene set analysis methods, there is little consistency among different methods when analyzing the same gene expression dataset (Maleki et al., 2019b; Nguyen et al., 2019). Although gene set overlap is a common phenomenon in gene set databases, most gene set analysis methods disregard such an overlap. This results in a lack of specificity of these methods (Maleki et al., 2020).

Evaluating gene set analysis methods is extremely important (Zyla et al., 2016, 2019) However, most gene set analysis methods have been evaluated either based on oversimplified data—which do not represent real expression datasets and real gene set knowledge bases—or based on real expression datasets with presumed enrichment status for gene sets. Maleki et al. (2021) developed Silver as a methodology for evaluating such methods without relying on oversimplifying assumptions. Besides a thorough evaluation, new gene set analysis methods need to be systematically assessed to find the minimum number of samples required to achieve reproducible results (Maleki et al., 2019a).

The papers published in this Research Topic indicate that the development of gene set analysis methods and tools remains an active research area.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

REFERENCES

Ahsan, S., and Draghici, S. (2017). Identifying Significantly Impacted Pathways and Putative Mechanisms with iPathwayGuide. *Curr. Protoc. Bioinforma.* 57, 7–30. doi:10.1002/cpbi.24

Consortium, G. O. (2004). The Gene Ontology (GO) Database and Informatics Resource. *Nucleic Acids Res.* 32, D258–D261. doi:10.1093/nar/gkh036

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* 4, P3. doi:10.1186/gb-2003-4-5-p3

Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., and Krawetz, S. A. (2003b). Global Functional Profiling of Gene Expression. *Genomics* 81, 98–104. doi:10.1016/s0888-7543(02)00021-6

Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S. A., and Tainsky, M. A. (2003a). Onto-Tools, the Toolkit of the Modern Biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.* 31, 3775–3781. doi:10.1093/nar/gkg624

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., et al. (2007). A Systems Biology Approach for Pathway Level Analysis. *Genome Res.* 17, 1537–1545. doi:10.1101/gr.6202607

Draghici, S. (2016). *Statistics and Data Analysis for Microarrays Using R and Bioconductor*. Boca Raton, FL: CRC Press.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Comput. Biol.* 8, e1002375. doi:10.1371/journal.pcbi.1002375

Khatri, P., Draghici, S., Ostermeier, G. C., and Krawetz, S. A. (2002). Profiling Gene Expression Using Onto-Express. *Genomics* 79, 266–270. doi:10.1006/geno.2002.6698

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Res.* 44, W90–W97. doi:10.1093/nar/gkw377

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: Gene Set Analysis Toolkit with Revamped UIs and APIs. *Nucleic Acids Res.* 47, W199–W205. doi:10.1093/nar/gkz401

Maleki, F., Ovens, K., McQuillan, I., and Kusalik, A. J. (2019a). Size Matters: How Sample Size Affects the Reproducibility and Specificity of Gene Set Analysis. *Hum. Genomics* 13, 42. doi:10.1186/s40246-019-0226-2

Maleki, F., Ovens, K., Hogan, D. J., and Kusalik, A. J. (2020). Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front. Genet.* 11, 654. doi:10.3389/fgene.2020.00654

Maleki, F., Ovens, K. L., Hogan, D. J., Rezaei, E., Rosenberg, A. M., and Kusalik, A. J. (2019b). Measuring Consistency Among Gene Set Analysis Methods: A Systematic Study. *J. Bioinform. Comput. Biol.* 17, 1940010. doi:10.1142/s0219720019400109

Maleki, F., Ovens, K., McQuillan, I., and Kusalik, A. J. (2021). Silver: Forging Almost Gold Standard Datasets. *Genes* 12, 1523. doi:10.3390/genes12101523

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., et al. (2013). Methods and Approaches in the Topology-Based Analysis of Biological Pathways. *Front. Physiol.* 4, 278. doi:10.3389/fphys.2013.00278

Nguyen, T., Mitrea, C., and Draghici, S. (2018). Network-Based Approaches for Pathway Level Analysis. *Curr. Protoc. Bioinforma.* 61, 8–24. doi:10.1002/cpbi.42

Nguyen, T. M., Shafi, A., Nguyen, T., and Draghici, S. (2019). Correction to: Identifying Significantly Impacted Pathways: a Comprehensive Review and Assessment. *Genome Biol.* 20, 234. doi:10.1186/s13059-019-1882-1

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J.-s., et al. (2009). A Novel Signaling Pathway Impact Analysis. *Bioinformatics* 25, 75–82. doi:10.1093/bioinformatics/btn577

Zyla, J., Marczyk, M., Domaszewska, T., Kaufmann, S. H. E., Polanska, J., and Weiner, J., 3rd (2019). Gene Set Enrichment for Reproducible Science: Comparison of CERNO and Eight Other Algorithms. *Bioinformatics* 35, 5146–5154. doi:10.1093/bioinformatics/btz447

Zyla, J., Marczyk, M., and Polanska, J. (2016). "Sensitivity, Specificity and Prioritization of Gene Set Analysis When Applying Different Ranking Metrics," in International Conference on Practical Applications of Computational Biology & Bioinformatics, June 13, 2016, Seville, Spain. (Berlin: Springer), 61–69. doi:10.1007/978-3-319-40126-3_7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.