frontiers | Frontiers in Applied Mathematics and Statistics

EDITORIAL
published: 18 July 2022
doi: 10.3389/fams.2022.967585

# Editorial: Statistical Data Science - Theory and Applications in Analyzing Omics Data

*Li Xing [1], Xuekui Zhang [2]* and Liangliang Wang [3]*

[1] Department of Mathematics and Statistics, University of Saskatchewan, Saskatoon, SK, Canada, [2] Department of Mathematics and Statistics, University of Victoria, Victoria, BC, Canada, [3] Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada

**Keywords: data science, machine learning, omics, genomics, health outcome, statistics, high-dimensional data, network analysis**

**Editorial on the Research Topic**

**Statistical Data Science - Theory and Applications in Analyzing Omics Data**

OPEN ACCESS

***Edited and reviewed by:***
*Xiaoming Huo,*
*Georgia Institute of Technology,*
*United States*

****Correspondence:***
*Xuekui Zhang*
*xuekui@uvic.ca*

***Specialty section:***
*This article was submitted to Statistics and Probability, a section of the journal Frontiers in Applied Mathematics and Statistics*

***Received:*** *13 June 2022*
***Accepted:*** *24 June 2022*
***Published:*** *18 July 2022*

***Citation:***
*Xing L, Zhang X and Wang L (2022) Editorial: Statistical Data Science - Theory and Applications in Analyzing Omics Data. Front. Appl. Math. Stat. 8:967585. doi: 10.3389/fams.2022.967585*

Recent advancement in statistical data science empowers researchers to extract insights from high-dimensional omics data and connect them with health outcomes to prevent diseases and improve treatments. More and more research studies have been conducted focusing on statistics methodological development and its application to omics data. During the editorial process, we collected a number of research submissions and supported publications of eight excellent papers. Those papers cover a whole spectrum of omics data, including DNA count data, bulk and single-cell RNA sequencing data, DNA methylation data, and microbiome data, and a wide range of disease outcomes, such as COVID-19 disease, Crohn's disease, Hepatocellular Carcinoma, and Low-Grade Glioma. The types of research in those papers are also diversified, including theoretical development in novel statistical testing for DNA counts, a systematic literature review of statistical methods in microbiome data, a benchmark study of single-cell auto-annotation methods to provide guidelines to users, and all sorts of constructive data science method applications.

Those papers demonstrate that data science methods have significantly contributed to the growth of research involving omics data analysis. Since the sizes of omics data are usually big and their intrinsic structures are rather complex, those challenges in the data prohibit the use of traditional statistical methods. On the other side, data science methods have certain advantages in dealing with big and complex data. We would like to emphasize the importance of data science methods in omics data analysis from the following three aspects.

First, omics data usually contain big data matrices with massive columns of omics information but relatively fewer samples (the small-n-big-p problem). For example, Human has 20,000 protein-coding genes and 60,000 pseudogenes and non-coding genes [1]. So, for RNA sequencing data, the number of columns/genes is counted in thousands. In a Genome-wide association study, participants are genotyped for the most common known single-nucleotide polymorphisms, typically one million or more [2]. As for the microbiome, which includes bacteria, viruses and fungi living on and in humans, their estimated number is 10–100 trillion [3]. Such a large data size could also easily be gotten squared. Especially nowadays, the sample size keeps expanding. For example, the UK biobank study has recruited half-million participants. Data science methods provide our handful of tools for dealing with big data. Since the neural network gets the most recognition in the industry, it is also popular for omics data analysis [4]. In addition to the neural network, a black box type tool, the statistical learning methods, such as penalized regression, random forest, and boosting methods, are also capable of dealing with big data and making reasonable interpretations and predictions [5].

Second, structures and connections among omics measurements are complex and cannot be exploited by traditional regression models. Network-based pathway analysis can illustrate the intrinsic relationships between omics measurements, help find pathways, and conduct causal inference [6]. Intensive research has been undertaken in the areas of network and pathway analysis. For example, based on the gene co-expression network, we can find the gene modules containing coregulated genes to aid causal inference [7, 8]. We can also generate new research hypotheses for diseases based on network analysis via their intensive communication sub-community [9].

Third, the volume of omics data continues to expand at a faster and faster speed due to the technologies' advancements [10]. At the same time, industrial development also pushes data science methods to grow at a fantastic rate. Therefore, our statisticians' tasks are twofold: (1) developing novel methods mainly designed for omics data; and (2) innovative borrowing methods from other fields and adapting them to omics data analysis. Both are feasible and beneficial to move the omics data research field forward.

The next era's targets in data science method development and application to omics data should be aligned with the following two interconnected aims: (1) integrated analysis of multi-omics data. We request novel analysis pipelines, which could employ the collected measurements from all levels of the human body, incorporate multi-omics information, and link them with health outcomes. It could provide helpful and comprehensive information to assist disease diagnosis and improve treatment. (2) Building a user-friendly platform to facilitate auto extraction of data from their storage and the online use of the analysis pipeline. Since the data often requires cloud storage, the platform should be designed to easily access data in the cloud server and perform online analysis by pushing buttons. In summary, we need more advanced integrated methods and a more convenient application platform for omics data to move the fields forwards.

## AUTHOR CONTRIBUTIONS

LX and XZ jointly drafted the manuscript. LW revised the manuscript. All authors contributed to the article and approved the submitted version.

## FUNDING

LX is funded by Natural Sciences and Engineering Research Council of Canada (Grant No. RGPIN-2021-03530). XZ is a Tier 2 Canada Research Chairs (Grant No. 950-231363) and funded by Natural Sciences and Engineering Research Council of Canada (Grant No. RGPIN-2017-04722). LW is funded by Natural Sciences and Engineering Research Council of Canada (Grant No. RGPIN-2019-06131).

## REFERENCES

1. Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. *Nucleic Acids Res*. (2021) 50:D988–95. doi: 10.1093/nar/gkab1049
2. Auton A, Abecasis GR. A global reference for human genetic variation. *Nature*. (2015) 526:68–74. doi: 10.1038/nature15393
3. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. *Nutr Rev*. (2012) 70:S38–S44. doi: 10.1111/j.1753-4887.2012.00493.x
4. Yu H, Samuels DC, Zhao YY, Guo Y. Architectures and accuracy of artificial neural network for disease classification from omics data. *BMC Genomics*. (2019) 20:167. doi: 10.1186/s12864-019-5546-z
5. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer New York (2021).
6. Rezola A, Pey J, Tobalina L, Rubio n, Beasley JE, Planes FJ. Advances in network-based metabolic pathway analysis and gene expression data integration. *Brief. Bioinform*. (2014) 16:265–79. doi: 10.1093/bib/bbu009
7. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform*. (2017) 19:575–92. doi: 10.1093/bib/bbw139
8. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*. (2003) 34:166–76. doi: 10.1038/ng1165
9. Huang X, Shao X, Xing L, Hu Y, Sin DD, Zhang X. The impact of lockdown timing on COVID-19 transmission across US counties. *eClinicalMedicine*. (2021) 38:101035. doi: 10.1016/j.eclinm.2021.101035
10. Perez-Riverol Y, Zorin A, Dass G, Vu MT, Xu P, Glont M, et al. Quantifying the impact of public omics data. *Nat Commun*. (2019) 10:3512. doi: 10.1038/s41467-019-11461-w

**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.