


# Sangerbox: A comprehensive, interaction-friendly clinical bioinformatics analysis platform

Weitao Shen<sup>1</sup> | Ziguang Song<sup>2,3,4</sup> | Xiao Zhong<sup>2,3</sup> | Mei Huang<sup>1</sup> |  
Danting Shen<sup>4</sup> | Pingping Gao<sup>2</sup> | Xiaoqian Qian<sup>5</sup> | Mengmeng Wang<sup>6</sup> |  
Xiubin He<sup>1</sup> | Tonglian Wang<sup>1</sup> | Shuang Li<sup>1</sup> | Xiang Song<sup>2,4</sup> 

<sup>1</sup>Bioinformatics R&D Department, Hangzhou Mugu Technology Co., Ltd, Hangzhou, China

<sup>2</sup>Department of Cardiovascular Medicine, Shanghai University of Medicine & Health Sciences Affiliated Zhoupu Hospital, Shanghai, China

<sup>3</sup>Cardiovascular Center, The Fourth Affiliated Hospital of Harbin Medical University, Harbin, Heilongjiang, China

<sup>4</sup>College of Basic Medical, Harbin Medical University, Harbin, Heilongjiang, China

<sup>5</sup>Renal Division, Department of Internal Medicine, Xinhua Hospital Affiliated to Shanghai Jiao Tong University of Medicine, Shanghai, China

<sup>6</sup>Oncology Research Center, Beidahuang Industry Group General Hospital, Harbin, Heilongjiang, China

## Correspondence

Xiang Song, Department of Cardiovascular Medicine, Shanghai University of Medicine & Health Sciences Affiliated Zhoupu Hospital, Shanghai, China.

Email: [song761231@sina.com](mailto:song761231@sina.com)

Shuang Li, Hangzhou Mugu Technology Co., Ltd, Hangzhou, China.

Email: [lishuang@cqu.edu.cn](mailto:lishuang@cqu.edu.cn)

## Funding information

Shanghai Pudong New District Zhoupu Hospital, Grant/Award Number: ZP-XK-2021B-1; Health and Family Planning Committee of Pudong New Area, Grant/Award Number: PWRI2021-08

## SUMMARY

In recent decades, the continuous development of high-throughput sequencing technology has increased data volume in medical research [1]. At the same time, almost all clinical researchers have their own independent omics data, which provided a better condition for data mining and a deeper understanding of gene functions. However, due to the larger amount of data, many common and cutting-edge effective bioinformatics research methods still cannot be widely used. This has encouraged the establishment of many analysis platforms and databases to accommodate the

demands of users, for instance, Qiita for amplicon data and analysis [2], ImageGP for plotting [3], QIIME for microbiome analysis and visualization, and Majorbio cloud for multiomics [4], have been developed for omics research. Some databases or servers provide solutions for special problems. Metascape [5] was designed to provide functional annotations of genes and function enrichment analysis; metaOrigin [6] supports metabolome original analysis from microbiome; Gene2vec for m6A prediction [7]; iNAP for network analysis [8]; PsRobot [9] for small RNA meta-analysis; DeepKla [10] for protein lysine lactylation site prediction. Additionally, some web services are outdated, and inefficient

Weitao Shen, Ziguang song, and Xiao Zhong contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *iMeta* published by John Wiley & Sons Australia, Ltd on behalf of *iMeta* Science.

interaction often fails to meet the current needs of researchers. Therefore, the demand to complete massive data processing tasks urgently requires a comprehensive bioinformatics analysis platform.

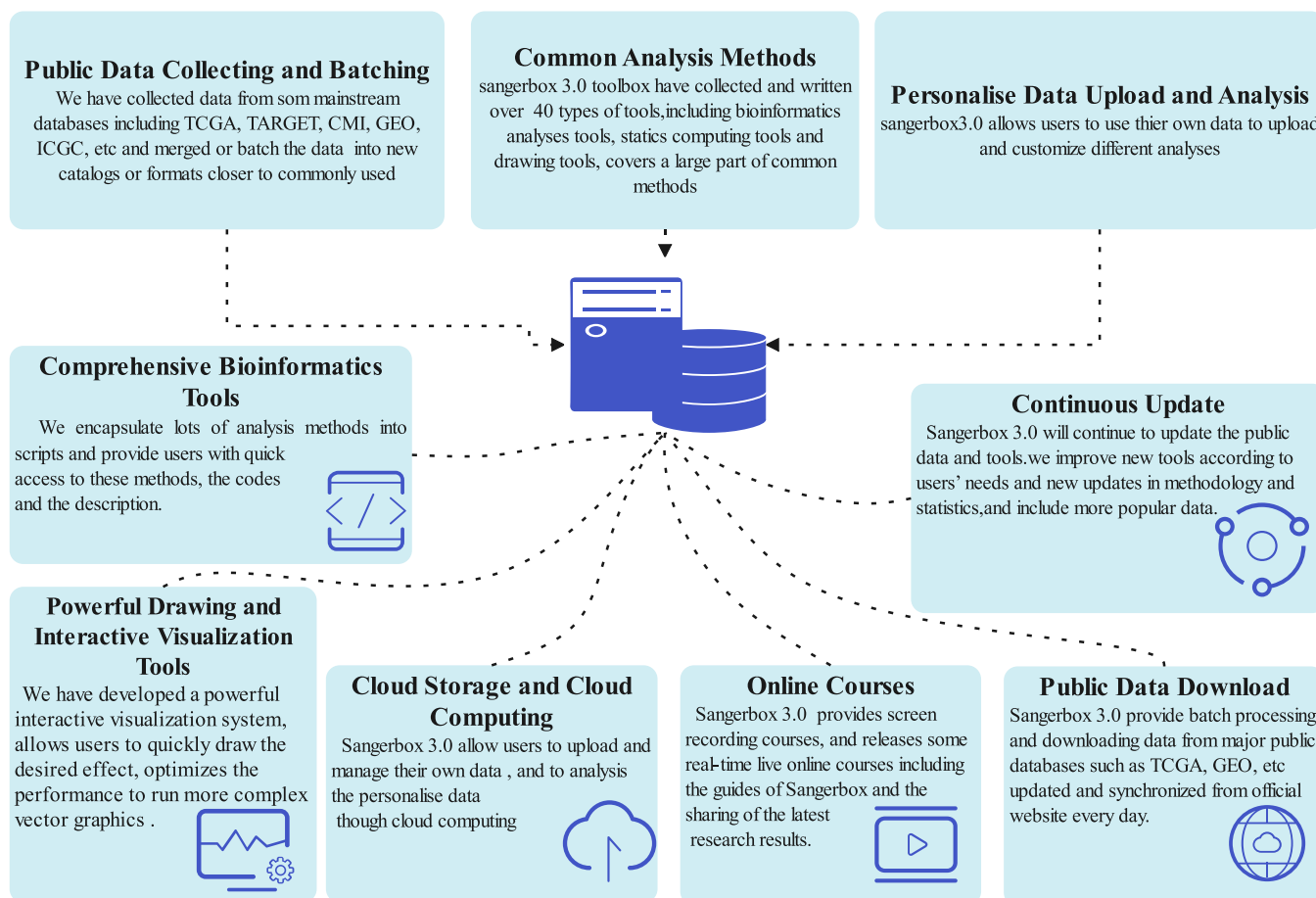
Hence, we have developed a website platform, Sangerbox (<http://vip.sangerbox.com/>). The platform as a user-friendly interface supports differential analysis and provides interactive customizable analysis tools, including various kinds of correlation analyses, pathway enrichment analysis, weighted correlation network analysis (WGCNA) [11] as well as some other common tools and functions. Users only need to upload their own corresponding data into Sangerbox, select required parameters, submit, and collect the results after the task has been completed. We have also established a new interactive plotting system that allows users to adjust the parameters in the image; moreover, optimized plotting performance enables users to adjust large-capacity vector maps on the website. At the same time, Sangerbox has integrated GEO, TCGA, ICGC and other databases and processed the data in batches, which greatly reduces the difficulty to obtain data and improves the efficiency of bioinformatics analysis for users. Additionally, we also provide users with rich sources of

bioinformatics analysis courses, creating a platform for researchers to share and exchange knowledge.

Since August 2021 when the Sangerbox cloud platform was established, it has accumulated more than 20,000 users, running 150,428 times for analysis, plotting, downloads, and other tasks, and previous versions of Sangerbox have been used 813,816 times. This proves that a comprehensive, interaction-friendly bioinformatics data analysis platform is greatly needed and welcomed by researchers in the field. The content and framework of Sangerbox are shown in Figure 1. Such a platform can greatly facilitate data mining, scientific discussion, and treatment discovery processes in a wide range of biological and clinical research areas.

## Convenient, powerful and interactive analyzing and plotting tools

The Sangerbox platform accelerates the analysis of researchers' data, improves the utilization of both public and personal data, and contributes to the development of clinical research. Bioinformatics analysis has long been difficult for clinical and specialized experimental



**FIGURE 1** Content and framework of Sangerbox



difference analysis, and so on. In addition, some common bioanalysis processes, such as weighted correlation network analysis (WGCNA) [11], principal component analysis (PCA) [18], survival analysis [19], Gene Set Enrichment Analysis (GSEA) [20], and Limma differential expression analysis [21], are accommodated as well.

Sangerbox will continue to improve new tools according to users' needs and new updates in methodology and statistics so that researchers using the Sangerbox platform can reduce the learning cost and more clinical bioinformatics data can be more efficiently processed, thereby contributing to the development of clinical research.

## Powerful interactive visualization interface

We have developed a new visual interaction system with the goal of “what you see is what you get” and without cumbersome programming code or complex parameter settings. The platform is established based on d3.js and jquery.js. The interactive visual interface, which was designed using JavaScript, allows users to intuitively

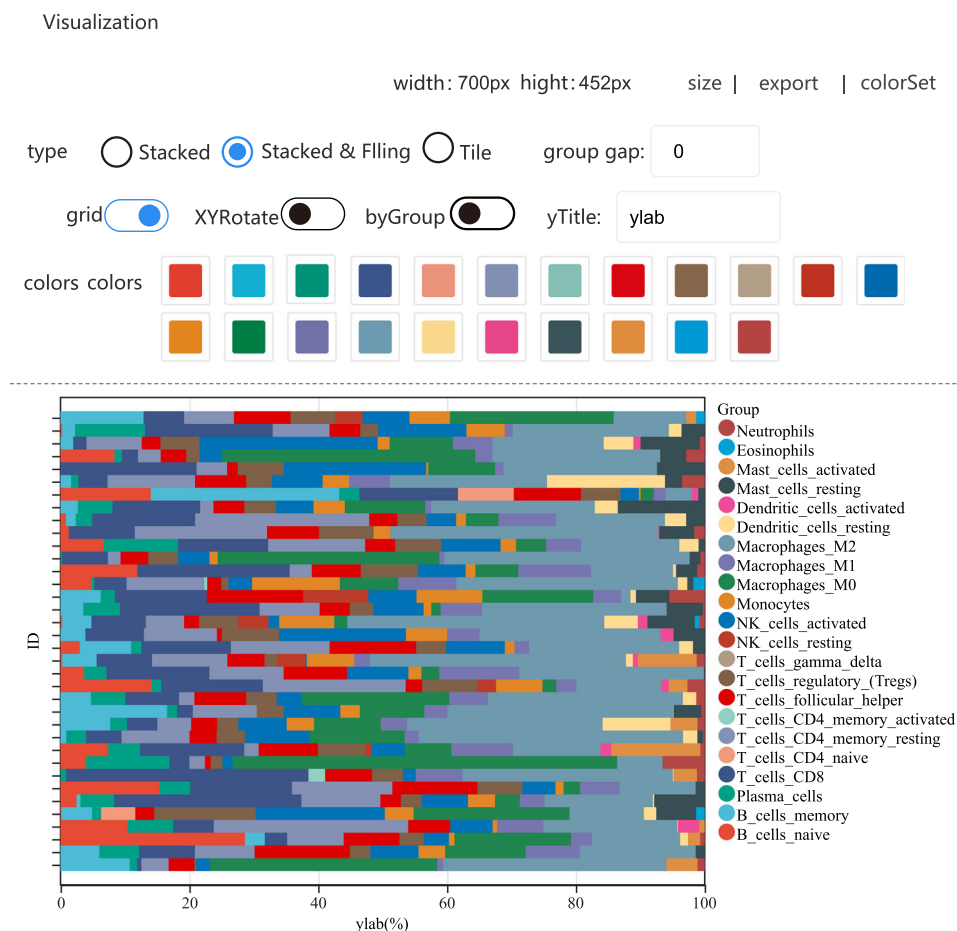
obtain vector graphics using mouse clicks and input to set parameters in web pages, thereby realizing our goal of “what you see is what you get.” Furthermore, Sangerbox also supports users to export bitmap or vector graphics in different formats to further support their research.

In addition, it is well acknowledged that the vector plotting adjustment process consumes great arithmetic power. Thus, we have optimized the plotting performance of vectors on the web site, allowing researchers in different working environments to adjust their vector images according to their own needs.

As shown in Figure 3, a stacked histogram was displayed as an example. Sangerbox supports data input to obtain initial pictures of histogram form, group spacing, grid switch, and so on, and then users could select parameters by mouse click or direct input. The same interface is also provided for plotting charts on the web page.

## Public data download and processing

The Sangerbox platform supports easier download of public data and simple but fast data preprocessing, which helps



**FIGURE 3** The example of the histogram visualization panel.

researchers to obtain and apply data more quickly. These data are derived from The Cancer Genome Atlas (TCGA) [22], International Cancer Genome Consortium (ICGC) [23], Gene Expression Omnibus (GEO) [24], and other databases with clinical follow-up information, clinical data, mutation data, and expression profile data. Sangerbox also provides preprocessing functions for the gene expression profile matrix in NCBI's GEO database, supporting a direct use for reannotation, standardization, and other steps. At the same time, a new category for TCGA, ICGC, TARGET (therapeutically applicable research to generate effective treatments) and other databases have been integrated to better accommodate the using habits of general researchers and help users obtain public data more simply and quickly.

The Sangerbox platform has built a complete course sharing platform. While providing screen recording courses, we also release some real-time live online courses. This can not only help scientific research users get familiar with how to use the platform but also informs them about the cutting-edge research trends and research methods in certain fields.

## Users and publications

From August 2021, Sangerbox has accumulated more than 20,000 scientific research users during about half a year and has completed 150,428 data analyses, mining, and mapping tasks. In the last few years, over 300 documented journal articles have cited the use of Sangerbox in their methodology. The platform will constantly upgrade and update the content.

## Methods and techniques

We employed the SpringCloud framework, java 11,10.2.30-MariaDB to build the website, website backstage, and database. Web pages were constructed using the layui framework, html 5, and JavaScript languages. R 3.6.3, R 4.0.1, Perl, and JavaScript languages were applied to write and convert into arithmetic and analysis scripts.

## AUTHOR CONTRIBUTIONS

Xiang Song and Shuang Li conceived the idea of developing the Sangerbox platform. Weitao Shen completed the technical construction of the platform and the realization of each function and wrote the manuscript. Ziguang Song and Xiao Zhong are responsible for collecting and collating data. Mei Huang was responsible for editing and revising the manuscript. Danting Shen, Xiaoqian Qian, Pingping Gao, Mengmeng Wang, Xiubin He, and Tonglian Wang completed the collection and

arrangement of methods and materials for the platform. All authors contributed to the development of Sangerbox.

## ACKNOWLEDGMENTS

The authors acknowledge Dr. Yong-Xin Liu for the professional advice on this manuscript. This study was funded by the Shanghai Pudong New District Zhoupu Hospital Foundation for Talent Introduction Program (Grant/Award Numbers: ZP-XK-2021B-1) and the Leading Personnel Training Program of Pudong New District Health and Family Planning Commission of Shanghai, China (Grant/Award Numbers: PWRI2021-08). Song Xiang is the host of the aforementioned projects.

## CONFLICTS OF INTEREST

Weitao Shen is the developer of the platform. Xiang Song and Shuang Li are cofounders of the project. Ziguang Song, Xiao Zhong, Danting Shen, Xiaoqian Qian, Pingping Gao, Mengmeng Wang, Mei Huang, Xiubin He, and Tonglian Wang are employees and researchers of Hangzhou Mugu Technology Co., Ltd.

## DATA AVAILABILITY STATEMENT

We will view or use the personal uploaded data of the users with the corresponding permission of the users. The public data we collected and batched is available online. Supplementary materials (figures, tables, scripts, graphical abstract, slides, videos, Chinese translated version, and update materials) may be found in the online DOI or iMeta Science <http://www.imeta.science/>.

## ORCID

Xiang Song  <http://orcid.org/0000-0002-1575-2307>

## REFERENCES

1. Chen, Tingting, Xu Chen, Sisi Zhang, Junwei Zhu, Bixia Tang, Anke Wang, Lili Dong, et al. 2021. "The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types." *Genomics, Proteomics & Bioinformatics* 19: 578–583. <https://doi.org/10.1016/j.gpb.2021.08.001>
2. Gonzalez, Antonio, Jose A. Navas-Molina, Tomasz Kosciolk, Daniel McDonald, Yoshiki Vázquez-Baeza, Gail Ackermann, Jeff DeReus, et al. 2018 "Qiita: Rapid, Web-Enabled Microbiome Meta-Analysis." *Nature Methods* 15: 796–798. <https://doi.org/10.1038/s41592-018-0141-9>
3. Chen, Tong, and Luqi Huang Yong-Xin Liu. 2022. "ImageGP: An Easy-To-Use Data Visualization Web Server for Scientific Researchers." *iMeta* 1: e5. <https://doi.org/10.1002/imt2.5>
4. Ren, Yi, Guo Yu, Caiping Shi, Linmeng Liu, Quan Guo, Chang Han, Dan Zhang, et al. 2022. "Majorbio Cloud: A One-Stop, Comprehensive Bioinformatic Platform for Multiomics Analyses." *iMeta* 1: e12. <https://doi.org/10.1002/imt2.12>
5. Zhou, Yingyao, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K. Chanda. 2019. "Metascape Provides a

- Biologist-Oriented Resource for the Analysis of Systems-Level Datasets.” *Nature Communications* 10: 1523. <https://doi.org/10.1038/s41467-019-09234-6>
6. Yu, Gang, Cuifang Xu, Danni Zhang, Feng Ju, and Yan Ni. 2022. “MetOrigin: Discriminating the Origins Of Microbial Metabolites for Integrative Analysis of the Gut Microbiome and Metabolome.” *iMeta* 1: e10. <https://doi.org/10.1002/imt2.10>
  7. Zou, Quan, Pengwei Xing, Leyi Wei, and Bin Liu. 2019. “Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA.” *RNA* 25: 205–218. <https://doi.org/10.1261/rna.069112.118>
  8. Feng, Kai, Xi Peng, Zheng Zhang, Songsong Gu, Qing He, Wenli Shen, Zhujun Wang, et al. 2022. “iNAP: An Integrated Network Analysis Pipeline for Microbiome Studies.” *iMeta* 1: e13. <https://doi.org/10.1002/imt2.13>
  9. Wu, Hua-Jun, Tong Chen, Ying-Ke Ma, Meng Wang, and Xiu-Jie Wang. 2012. “PsRobot: A Web-Based Plant Small RNA Meta-Analysis Toolbox.” *Nucleic Acids Research* 40: W22–W28. <https://doi.org/10.1093/nar/gks554>
  10. Lv, Hao, Fu-Ying Dao, and Hao Lin. 2022. “DeepKla: An Attention Mechanism-Based Deep Neural Network for Protein Lysine Lactylation Site Prediction.” *iMeta* 1: e11. <https://doi.org/10.1002/imt2.11>
  11. Langfelder, Peter, and Steve Horvath. 2008. “WGCNA: An R Package for Weighted Correlation Network Analysis.” *BMC Bioinformatics* 9: 559. <https://doi.org/10.1186/1471-2105-9-559>
  12. Gu, Zuguang, and Matthias Schlesner Roland Eils. 2016. “Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data.” *Bioinformatics* 32: 2847–2849. <https://doi.org/10.1093/bioinformatics/btw313>
  13. Spitzer, Michaela, Jan Wildenhain, Juri Rappsilber, and Mike Tyers. 2014. “BoxPlotR: A Web Tool for Generation of Box Plots.” *Nature Methods* 11: 121–122. <https://doi.org/10.1038/nmeth.2811>
  14. Chen, Tong, Haiyan Zhang, Yu Liu, Yong-Xin Liu, and Luqi Huang. 2021. “EVENN: Easy to Create Repeatable and Editable Venn Diagrams and Venn Networks Online.” *Journal of Genetics and Genomics* 48: 863–866. <https://doi.org/10.1016/j.jgg.2021.07.007>
  15. Gu, Zuguang, Lei Gu, Roland Eils, Matthias Schlesner, and Benedikt Brors. 2014. “Circlize Implements and Enhances Circular Visualization in R.” *Bioinformatics* 30: 2811–2812. <https://doi.org/10.1093/bioinformatics/btu393>
  16. Wickham, Hadley. 2011. “ggplot2.” *WIREs Computational Statistics* 3: 180–185. <https://doi.org/10.1002/wics.147>
  17. Liu, Yong-Xin, Yuan Qin, Tong Chen, Meiping Lu, Xubo Qian, Xiaoxuan Guo, and Yang Bai. 2021. “A Practical Guide To Amplicon and Metagenomic Analysis Of Microbiome Data.” *Protein & Cell* 12: 315–330. <https://doi.org/10.1007/s13238-020-00724-8>
  18. Abdi, Hervé, and Lynne J. Williams. 2010. “Principal Component Analysis.” *WIREs Computational Statistics* 2: 433–459. <https://doi.org/10.1002/wics.101>
  19. Goel, Manish Kishore, Pardeep Khanna, and Jugal Kishore. 2010. “Understanding Survival Analysis: Kaplan-Meier Estimate.” *International Journal of Ayurveda research* 1: 274–278. <https://doi.org/10.4103/0974-7788.76794>
  20. Subramanian, Aravind, Pablo Tamayo, Vasmsi K. Mootha, Sayan Mukherjee, L. Ebert Benjamin, A. Gillette Michael, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-based Approach For Interpreting Genome-wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102: 15545–15550. <https://doi.org/10.1073/pnas.0506580102>
  21. Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-sequencing and Microarray Studies.” *Nucleic Acids Research* 43: e47–e47. <https://doi.org/10.1093/nar/gkv007>
  22. Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz. 2015. “Review the Cancer Genome Atlas (TCGA): An Immeasurable Source of Knowledge.” *Contemporary Oncology/Współczesna Onkologia* 1A: 68–77. <https://doi.org/10.5114/wo.2014.47136>
  23. Zhang, Junjun, Joachim Baran, A. Cros, Jonathan M. Guberman, Syed Haider, Jack Hsu, Yong Liang, et al. 2011. “International Cancer Genome Consortium Data Portal—A One-stop Shop For Cancer Genomics Data.” *Database* 2011: bar026. <https://doi.org/10.1093/database/bar026>
  24. Barrett, Tanya, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, et al. 2013. “NCBI GEO: Archive for Functional Genomics Data Sets—Update.” *Nucleic Acids Research* 41: D991–D995. <https://doi.org/10.1093/nar/gks1193>

**How to cite this article:** Shen, Weitao, Ziguang Song, Xiao Zhong, Mei Huang, Danting Shen, Pingping Gao, Xiaoqian Qian, et al. 2022. “Sangerbox: A Comprehensive, Interaction-Friendly Clinical Bioinformatics Analysis Platform.” *iMeta* 1, e36. <https://doi.org/10.1002/imt2.36>