Computational and Structural Biotechnology Journal 23 (2024) 659–668

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

HiOmics: A cloud-based one-stop platform for the comprehensive analysis of large-scale omics data

Wen Li [a,b,c,1], Zhining Zhang [d,1], Bo Xie [a,1], Yunlin He [d], Kangming He [d], Hong Qiu [a,d], Zhiwei Lu [d], Chunlan Jiang [d], Xuanyu Pan [e], Yuxiao He [a], Wenyu Hu [d], Wenjian Liu [f], Tengcheng Que [f,g,h], Yanling Hu [a,b,c,d,f,*]

[a] Life Sciences Institute, Guangxi Medical University, Nanning, Guangxi, China
[b] Department of Biochemistry and Molecular Biology, School of Basic Medicine, Guangxi Medical University, Nanning, Guangxi, China
[c] Key Laboratory of Biological Molecular Medicine Research (Guangxi Medical University), Education Department of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, China
[d] Guangxi Henbio Biotechnology Co., Ltd., Nanning, Guangxi, China
[e] School of Basic Medicine, Guangxi Medical University, Nanning, Guangxi, China
[f] Faculty of Data Science, City University of Macau, Macau, China
[g] Youjiang Medical University for Nationalities, Baise, Guangxi, China
[h] Guangxi Zhuang Autonomous Terrestrial Wildlife Rescue Research and Epidemic Diseases Monitoring Center, Nanning, Guangxi, China

ARTICLE INFO

Keywords:
Artificial intelligence modeling
Cloud-based analysis
HiOmics
Large-scale omics data analysis
Multi-omics integration analysis

ABSTRACT

Analyzing the vast amount of omics data generated comprehensively by high-throughput sequencing technology is of utmost importance for scientists. In this context, we propose HiOmics, a cloud-based platform equipped with nearly 300 plugins designed for the comprehensive analysis and visualization of omics data. HiOmics utilizes the Element Plus framework to craft a user-friendly interface and harnesses Docker container technology to ensure the reliability and reproducibility of data analysis results. Furthermore, HiOmics employs the Workflow Description Language and Cromwell engine to construct workflows, ensuring the portability of data analysis and simplifying the examination of intricate data. Additionally, HiOmics has developed DataCheck, a tool based on Golang, which verifies and converts data formats. Finally, by leveraging the object storage technology and batch computing capabilities of public cloud platforms, HiOmics enables the storage and processing of large-scale data while maintaining resource independence among users.

1. Introduction

The advent of high-throughput sequencing technology has led to an exponential growth in omics data, encompassing transcriptomics, genomics, and metabolomics data [1–4]. The generation of large-scale omics data offers scientists valuable opportunities to gain comprehensive and profound insights into life processes [5–7]. However, it also presents remarkable challenges, demanding more intricate analysis methods/workflows and substantial computational resources. Although programming software, such as Python and R, offer advantages in data statistics and visualization, users must possess specific computer knowledge, including proficiency in at least one programming language and familiarity with the Linux operating system [8]. Traditional desktop software [9–12] can only implement one or a few functions within the entire analysis process, and different applications often have distinct requirements for input and output data formats. Furthermore, installing and configuring these desktop applications can pose a considerable challenge for users.

In recent years, numerous web-based bioinformatics tools have emerged [13,14]. These tools enable researchers to conduct analyses seamlessly, eliminating the need for mastering intricate programming skills or configuring complex analysis environments through interactive graphical interfaces. ImageGP [15] is a specialized visualization platform tailored for generating graphics related to biological and medical data. This platform offers a wide array of scientific graphic and analysis sub-functions. START App [16] and CANEapp [17] provide

* Corresponding author at: Life Sciences Institute, Guangxi Medical University, Nanning, Guangxi, China.
E-mail address: ylhupost@163.com (Y. Hu).
[1] Theoretically, the contributions of the three authors are equal and both are considered as the first author.

https://doi.org/10.1016/j.csbj.2024.01.002
Received 13 September 2023; Received in revised form 1 January 2024; Accepted 2 January 2024
Available online 5 January 2024

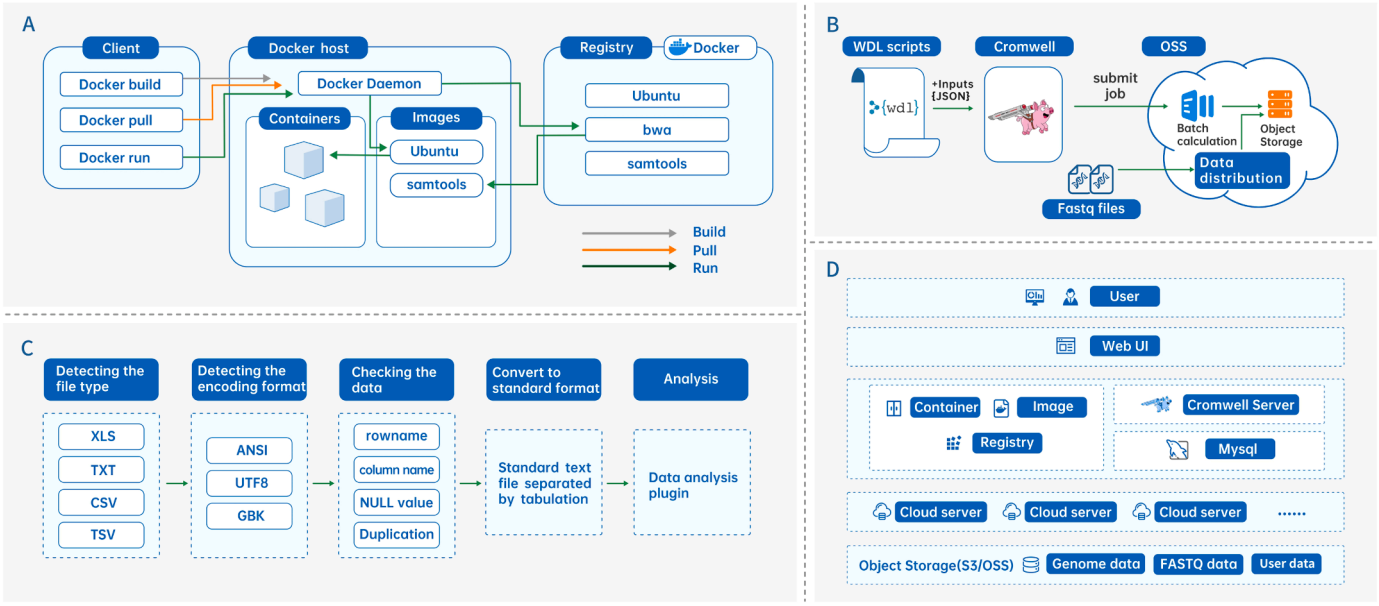


Fig. 1. Overview of the core advantages offered by the cloud-based HiOmics platform, enhancing development efficiency and user experience. (A) Leveraging Docker container technology to ensure the reliability and reproducibility of data analysis results. (B) HiOmics utilizes WDL + Cromwell to construct and manage intricate analysis workflows, integrating typical bioinformatics analysis processes. This diminishes the learning curve and exploration costs for users, meeting the deep data mining needs of non-programming users, such as clinical physicians. (C) Command-line tool DataCheck, based on Golang, automatically identifies and converts user-uploaded data into standard formats, thereby enhancing the user experience. (D) Utilization of object storage and batch computing technologies in public cloud platforms facilitates large-scale data storage and processing. Additionally, it ensures relative resource independence among different users.

transcriptomics data analysis and visualization services from read count data. MetaboAnalyst [18] specializes in the analysis and visualization of metabolomics data. IMG/M [19] focuses on metagenomics data analysis and visualization. Additionally, Metascape [20] provides functional enrichment, interaction component analysis, and gene annotation functions, catering to experimental biologists by offering comprehensive resources for gene list annotation and analysis. Although these individual web tools boast robust features, leveraging multiple platforms comprehensively is essential to formulating a complete system-level analysis workflow.

Since its launch in 2010, Galaxy [21] has paved the way for numerous cloud-based web tools that now play a pivotal role in high-throughput data analysis. Sangerbox [22] boasts powerful interactive plotting capabilities, allowing users to intuitively adjust parameters within the interface. However, its bioinformatics tools remain limited. Conversely, Hiplot [23] offers a wider array of interactive visualization tools, primarily tailored for lightweight data processing and plotting requirements. Qiita [24] stands out for its extensive collection of community resources, but its analysis functions are primarily focused on microbiome data. DolphinNext [25] empowers users to construct intricate data analysis workflows effortlessly using drag-and-drop operations, although its advanced analysis tools are somewhat limited in scope. Although Galaxy integrates numerous upstream and downstream analysis tools and facilitates customization of complex workflows through drag-and-drop functionality, non-bioinformatics users face a learning curve to familiarize themselves with the various tools and workflows [21]. Furthermore, issues, such as a non-user-friendly interface and frequent errors in input data formatting, significantly affect the overall user experience.

To address these challenges, we have developed HiOmics, a dedicated cloud-based web platform (https://www.henbio.com/en/tools) designed for the comprehensive analysis and visualization of multi-omics data. HiOmics utilizes the Element Plus framework [26], not only presenting an appealing and user-friendly interface but also incorporating interactive features to elevate the overall user experience. For consistent and reliable analysis results across different environments, HiOmics relies on Docker container technology [27], packaging the application and its dependencies into an independent container. Workflow Description Language (WDL) [28] streamlines the management and construction of data analysis workflows, enhancing efficiency and reproducibility. To cater to users without programming backgrounds, HiOmics seamlessly integrates typical bioinformatics pipelines. Additionally, we have developed DataCheck, a tool built on Golang [29], to automatically verify and convert input data formats, making it adaptable to various input types. The adoption of object storage and batch processing technologies on public cloud platforms empowers HiOmics to manage large-scale data storage and processing. This infrastructure ensures relative resource independence among users, facilitating efficient data management and processing.

2. Methods/implementation

2.1. Building user-friendly web interfaces with Elements Plus

HiOmics employs the UI library Element Plus [26], which is based on Vue.js 3, to create a sleek and user-friendly interface. Backend data communication is facilitated by utilizing the ThinkPHP framework and the MySQL database. All HiOmics applications are developed using open-source software like R and Python. They are abstracted as independent, parameterized, pluggable, and easily version-controlled plugins. With a simple mouse-click, users can seamlessly perform a wide array of tasks, ranging from basic visualization to complex omics analysis, without the need for any coding or programming expertise. All functionalities are accessible via the web interface provided by HiOmics.

2.2. Achieving reproducibility with container technology

In the field of bioinformatics, ensuring the reproducibility of data analysis results is of paramount importance [30,31]. To address this issue, HiOmics employs Docker container technology [32,33] to package each software tool, alongside its dependencies, analysis scripts, and runtime environment, into a singular, independent container image.

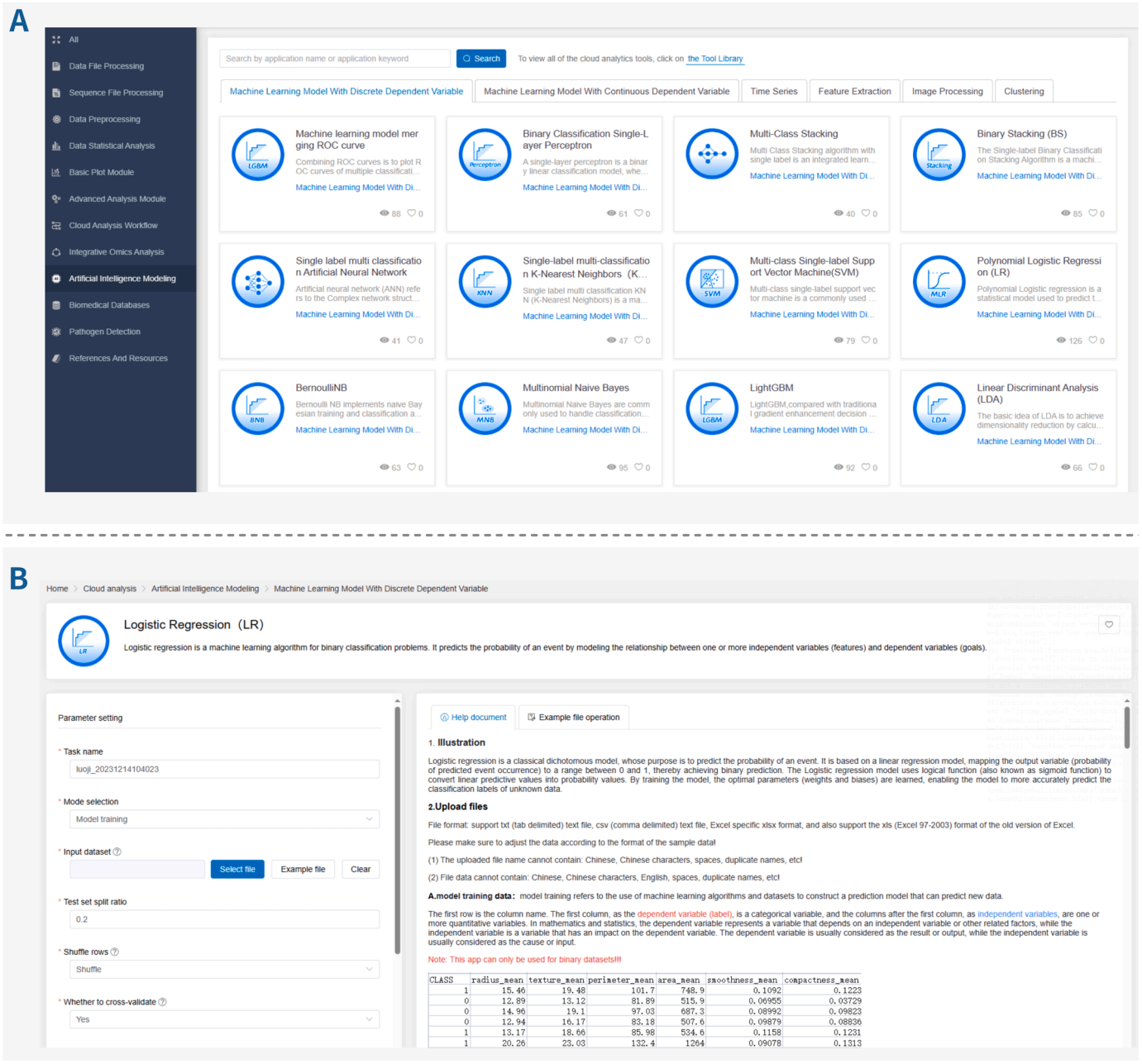


Fig. 2. HiOmics offers a user-friendly interface. (A) Fig. A showcases the main interface of the HiOmics Artificial intelligence (AI) Modeling module. Users can swiftly locate the desired tools using the search bar at the top or the navigation menu on the left. (B) Fig. B illustrates the interface of the "Logistic Regression" plugin within Artificial Intelligence module. The parameter panel is located on the left, while the explanation panel is situated on the right.

Subsequently, these images are uploaded to a unified container image repository. Each software tool corresponds to an individual Docker image, and collectively, multiple images form a comprehensive data analysis workflow. This approach facilitates the convenience of building once and running anywhere, circumventing redundant software and environment setups. It guarantees consistency in software, versioning, and runtime environments across diverse servers and operating systems (Fig. 1A). Ultimately, this strategy ensures the reproducibility and reliability of data analysis results.

2.3. *Building standard workflows with WDL*

Bioinformatics analysis tasks typically involve multiple steps and necessitate the combined use of various software and scripts, such as shell, Python, Perl, and command-line software. The utilization of WDL [28] enables the automation of bioinformatics analysis workflows. This automation standardizes and unifies the inputs, outputs, and running environments for each step, thereby greatly improving the reproducibility and reliability of the analysis. HiOmics employs WDL + Cromwell [34] to construct and manage workflows, providing a web-based interface for task submission (Fig. 1B). Currently, HiOmics has launched typical omics analysis workflows, including transcriptomics, metagenomics, and whole-genome sequencing. Additionally, it has integrated fundamental data, such as reference genomes and indexes for commonly studied species. Through the web-based interface, users can easily complete complex omics analysis tasks by uploading data files, and specifying a few parameters.

Table 1
Comparison of Web Services between Henbio and other Cloud-based Platforms.

Cloud-based platforms	HiOmics	Hiplot	ImageGP	Galaxy
number of analysis tools	290 +	330 +	34	1600 +
Data file processing	8	small	no	full
Sequence file processing	43	no	small	full
Data preprocessing	14	no	no	full
Data statistical analysis	35	small	no	small
Basic plotting	60	full	small	moderate
Advanced analysis	28	full	small	moderate
Integrative omics analysis	13	no	no	no
Cloud-based analysis Workflow	49	no	no	constructed by users
Artificial intelligence modeling	42	no	no	no
Pathogen detection	7	no	no	small
Biomedical Databases	726	no	no	small

2.4. Achieving diverse input data formats with DataCheck

Many data analysis software applications impose strict requirements on the input data format, often mandating data to be in comma-separated CSV files. Uploading a table-type XLSX file, for instance, would trigger an error. To circumvent such issues and minimize user inconvenience regarding format conversion, we have developed a command-line tool called DataCheck using Golang [29] (Fig. 1C). This tool is integrated into the analysis script and automatically validates data before executing specific analysis tasks. It can automatically identify the type, encoding format, and delimiter of user-uploaded files and then convert them into the standard format required by the plugin. Users no longer need to concern themselves with whether they should upload a text file or a spreadsheet file. Additionally, this tool conducts automatic data validation, detecting common data errors, such as non-numeric values, irregularly formatted matrices, inconsistent row or column names, as well as duplicate rows and columns. It alerts users through log messages, empowering them to perform targeted checks and rectify the data accordingly.

2.5. Implementing large-scale data storage and processing with cloud platform architecture

The rapid advancement of sequencing technology, coupled with decreasing costs, has resulted in a swift increase in high-throughput sequencing data [35–37]. Individual data files often reach gigabyte levels, thereby posing unprecedented challenges in data storage and computation [38,39]. To counter this challenge, HiOmics embraces public cloud object storage technology (Alibaba OSS, similar to Amazon S3), offering a scalable storage solution for large-scale high-throughput

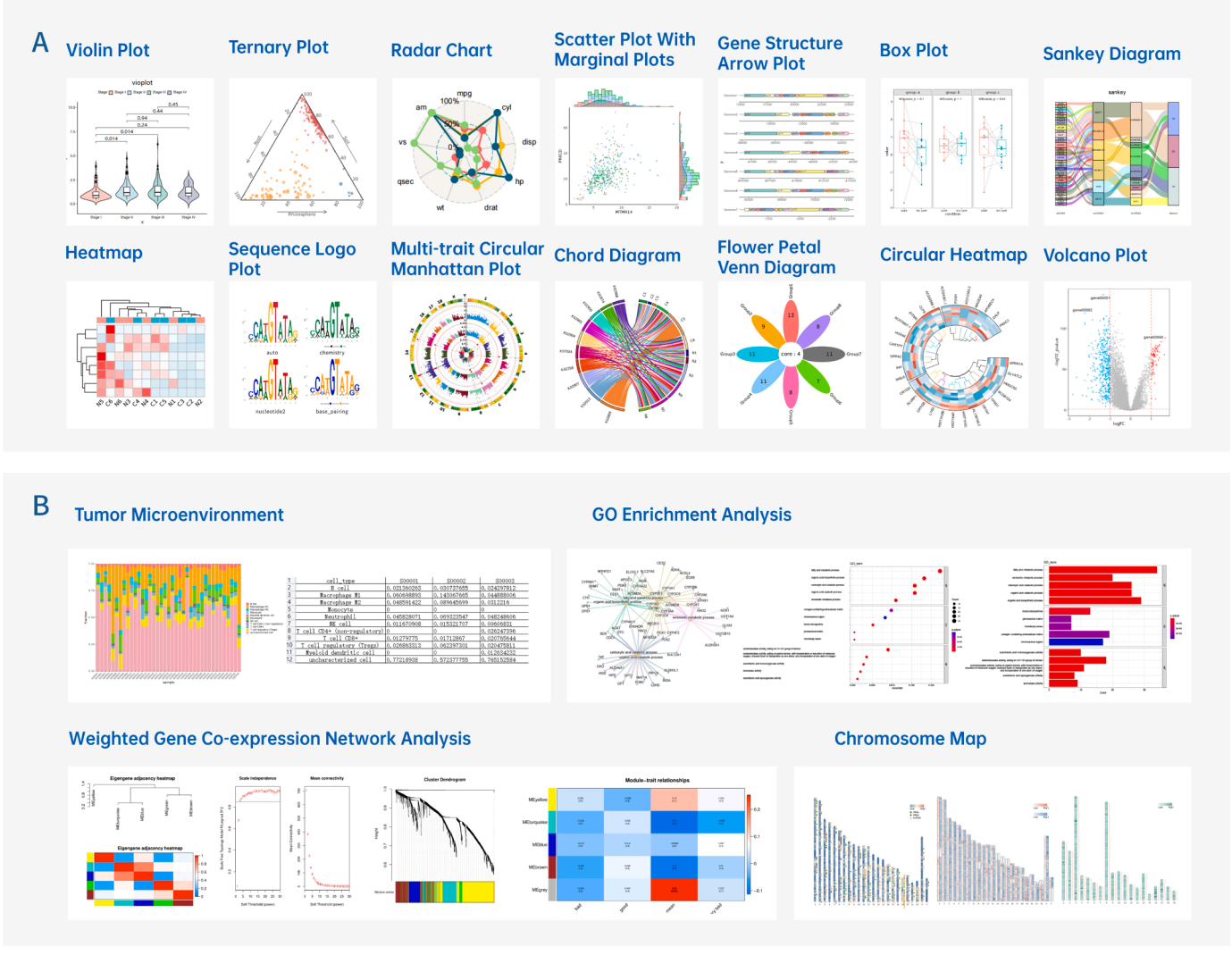


Fig. 3. HiOmics offers a variety of high-throughput data analysis and visualization tools to meet common requirements. (A) Partial visualization results from the Basic Plotting Module. (B) Partial visualization results from the Advanced Analysis Module.

Fig. 4. Representative use case of advanced analysis module. (A) Screenshot of the "Each-sample Immune Infiltration Analysis" plugin. (B) and (C) Display the input files, including the gene expression matrix file and the grouping file. (D) Expression heatmaps of various cell types across different samples, with colors indicating different sample groups, where blue represents the high-risk group and red represents the low-risk group. (E) File summarizing enrichment scores of various cells across different samples. (F) Boxplots comparing cell enrichment scores between high-risk and low-risk groups, with cell types labeled below and significance levels indicated above (* for $P < 0.05$, ** for $P < 0.01$, *** for $P < 0.001$).

sequencing data. This approach bypasses local storage capacity limitations on a single server and effortlessly expands storage in tandem with user growth (currently, individual users possess a default permanent storage space of 5 GB, extendable as required). Simultaneously, HiOmics utilizes public cloud batch computing (Alibaba Cloud Batch Compute, similar to Amazon Batch) as the core computing resource. This system supports the parallel processing of numerous jobs, automatically manages resources, schedules tasks, and loads data based on task requirements (Fig. 1D). Therefore, resource utilization among HiOmics users remains relatively independent, averting resource contention and impact. This setup enhances task processing speed and eliminates the necessity to restrict the number of submitted tasks.

3. Results

3.1. Simple and user-friendly web page

HiOmics, a user-friendly web platform, does not require downloading or installing any software. The main interface features a left navigation bar and a right plugin bar, enabling users to swiftly locate the desired plugin based on their requirements (Fig. 2A). Moreover, users can use the search bar positioned at the top left corner of the page to find plugins quickly by entering keywords. Each plugin page adopts a two-column layout, as depicted in Fig. 2B. On the left side, a section is dedicated for data uploading and parameter configuration. Users can simply click the "Select File" option to choose files from the Cloud storage file manager or upload directly from their local device. To minimize input errors, most basic parameters are presented as drop-down lists for easy selection. The right side of the page contains a documentation and sample file toolbar, offering users a comprehensive understanding of the functionality and effective usage of the plugin. With only three simple steps—uploading data, configuring parameters, and submitting the task—users can effortlessly obtain the desired results without writing any code.

3.2. Numerous analysis and visualization tools

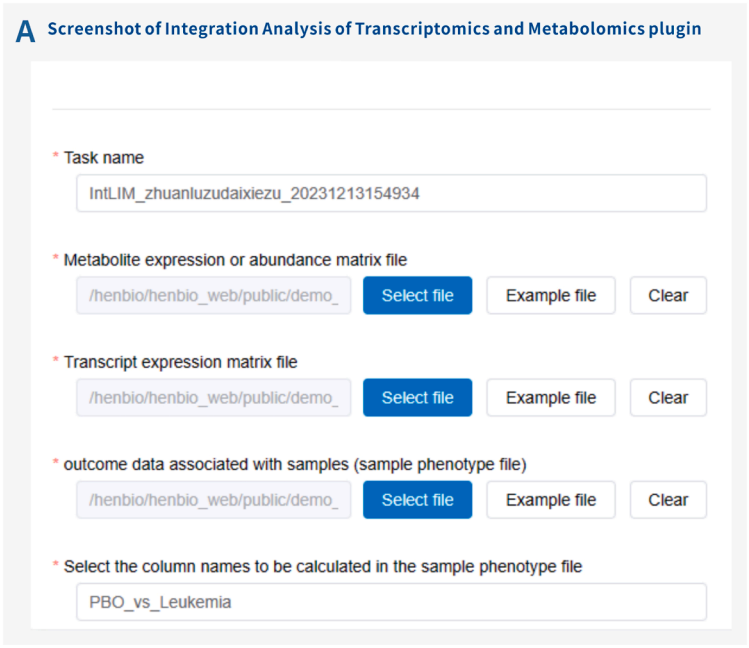
Since 2022, HiOmics has developed and launched over 290 interactive data analysis and visualization tools, all conveniently accessible through a user-friendly web interface. These tools encompass a comprehensive range of classic software and methods, covering the entire process from upstream sequence processing to downstream data visualization (Table 1).

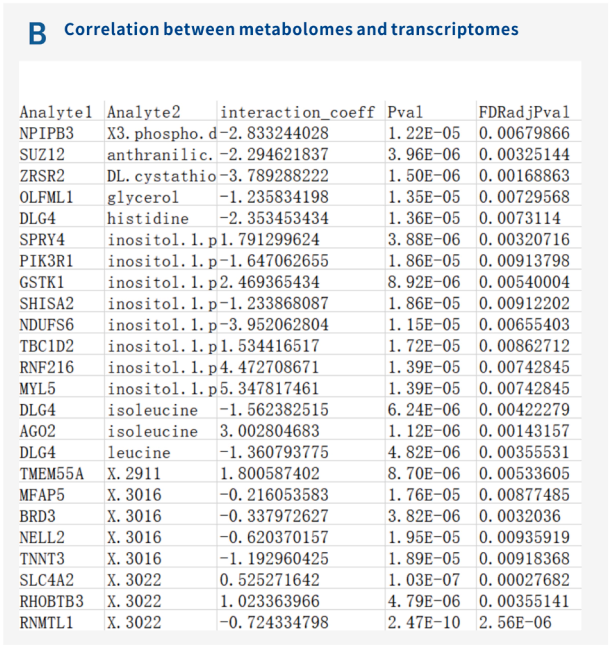
To accommodate article length constraints, this paper primarily focuses on introducing a few select functional plugins. These plugins are based on demo datasets, representing various feature categories and offering readers a fundamental understanding of their usage on this platform.

3.2.1. Use Case 1: Basic plotting module

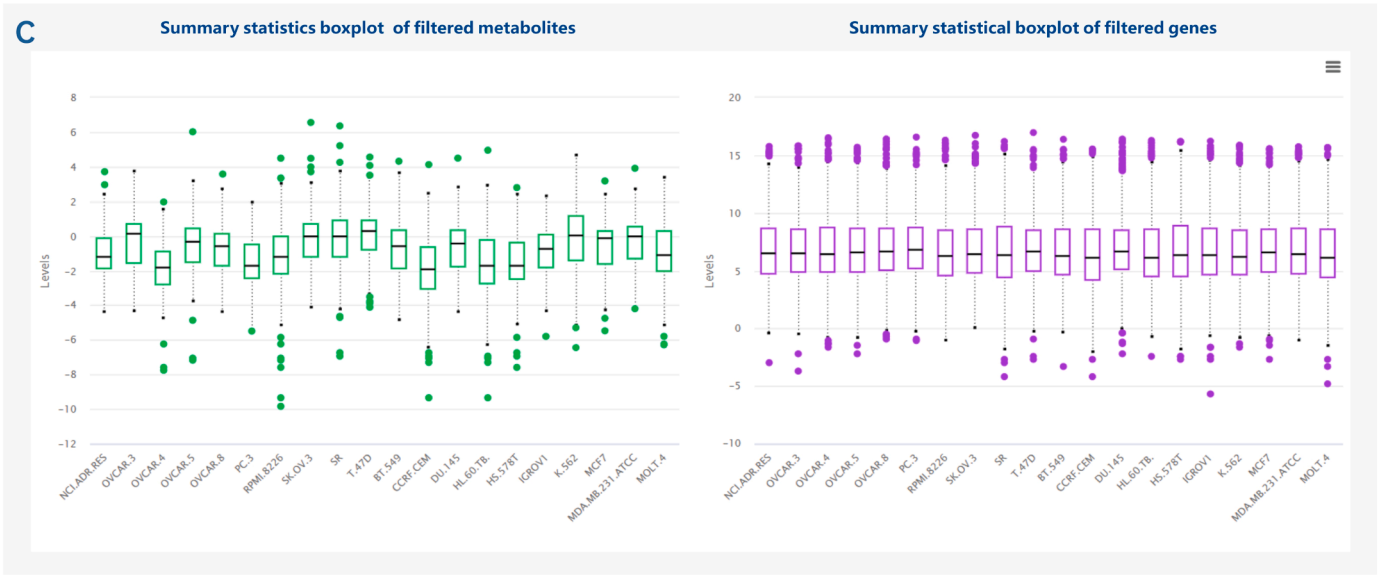
The basic plotting module offers 61 versatile plotting tools, including heatmap, Violin plot, volcano plot, Sankey diagram, Manhattan plot, Scatter plots, and box plot (Fig. 3A). These tools cater to diverse visualization needs, ensuring that the resulting images meet high-quality standards suitable for publication.

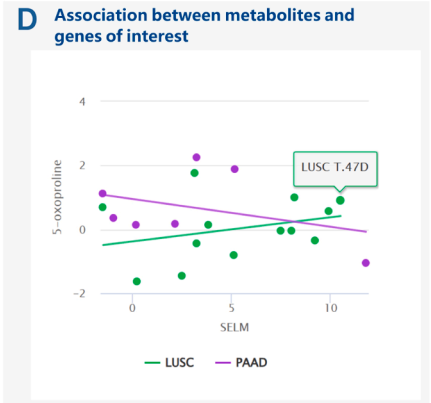
Next, let us take a heatmap as an example to introduce the usage of the basic plotting tool. Upon entering the heatmap plugin interface, you can easily select the gene expression matrix and group files from your local uploads or cloud storage file manager. After setting the parameters, simply click the "Run" button to generate the heatmap. Parameters for the heatmap include options for color scheme, row width, column height and normalization direction. Additional settings allow displaying specific genes of interest, clustering, row and column names, legends, and numerical values. Upon clicking the "Submit Task" button in the interface, the platform will automatically deliver the plotting task to Cromwell for scheduling and execution through the API interface. Once the task is completed, you can preview or download the heatmap from the result files. The heatmap can be saved in PDF, PNG, or SVG formats, and you may also choose to store it in the cloud storage file manager.





Analyte1	Analyte2	interaction_coeff	Pval	FDRadjPval
NPIPB3	X3. phospho. d	-2. 833244028	1. 22E-05	0. 00679866
SUZ12	anthranilic.	-2. 294621837	3. 96E-06	0. 00325144
ZRSR2	DL. cystathio	-3. 789288222	1. 50E-06	0. 00168863
OLFML1	glycerol	-1. 235834198	1. 35E-05	0. 00729568
DLG4	histidine	-2. 353453434	1. 36E-05	0. 0073114
SPRY4	inositol. 1. p	1. 791299624	3. 88E-06	0. 00320716
PIK3R1	inositol. 1. p	-1. 647062655	1. 86E-05	0. 00913798
GSTK1	inositol. 1. p	2. 469365434	8. 92E-06	0. 00540004
SHISA2	inositol. 1. p	-1. 233868087	1. 86E-05	0. 00912202
NDUFS6	inositol. 1. p	-3. 952062804	1. 15E-05	0. 00655403
TBC1D2	inositol. 1. p	1. 534416517	1. 72E-05	0. 00862712
RNF216	inositol. 1. p	4. 472708671	1. 39E-05	0. 00742845
MYL5	inositol. 1. p	5. 347817461	1. 39E-05	0. 00742845
DLG4	isoleucine	-1. 562382515	6. 24E-06	0. 00422279
AGO2	isoleucine	3. 002804683	1. 12E-06	0. 00143157
DLG4	leucine	-1. 360793775	4. 82E-06	0. 00355531
TMEM55A	X. 2911	1. 800587402	8. 70E-06	0. 00533605
MFAP5	X. 3016	-0. 216053583	1. 76E-05	0. 00877485
BRD3	X. 3016	-0. 337972627	3. 82E-06	0. 0032036
NELL2	X. 3016	-0. 620370157	1. 95E-05	0. 00935919
TNNT3	X. 3016	-1. 192960425	1. 89E-05	0. 00918368
SLC4A2	X. 3022	0. 525271642	1. 03E-07	0. 00027682
RHOBTB3	X. 3022	1. 023363966	4. 79E-06	0. 00355141
RNMTL1	X. 3022	-0. 724334798	2. 47E-10	2. 56E-06





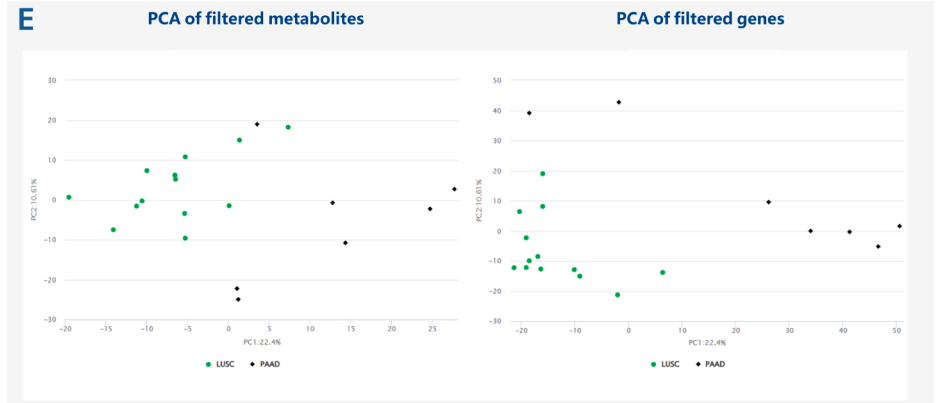


Fig. 5. Exemplification of the integrative omics analysis module's utility. (A) Screenshot of the "Integration Analysis of Transcriptomics and Metabolomics Data" plugin. (B) Results of transcriptome and metabolome correlation analysis are displayed in the excel table. (C) Summary statistics boxplot of filtered metabolites and genes. (D) Association between metabolites and genes of interest. (E) Principal Component Analysis (PCA) plots of filtered metabolites and genes.

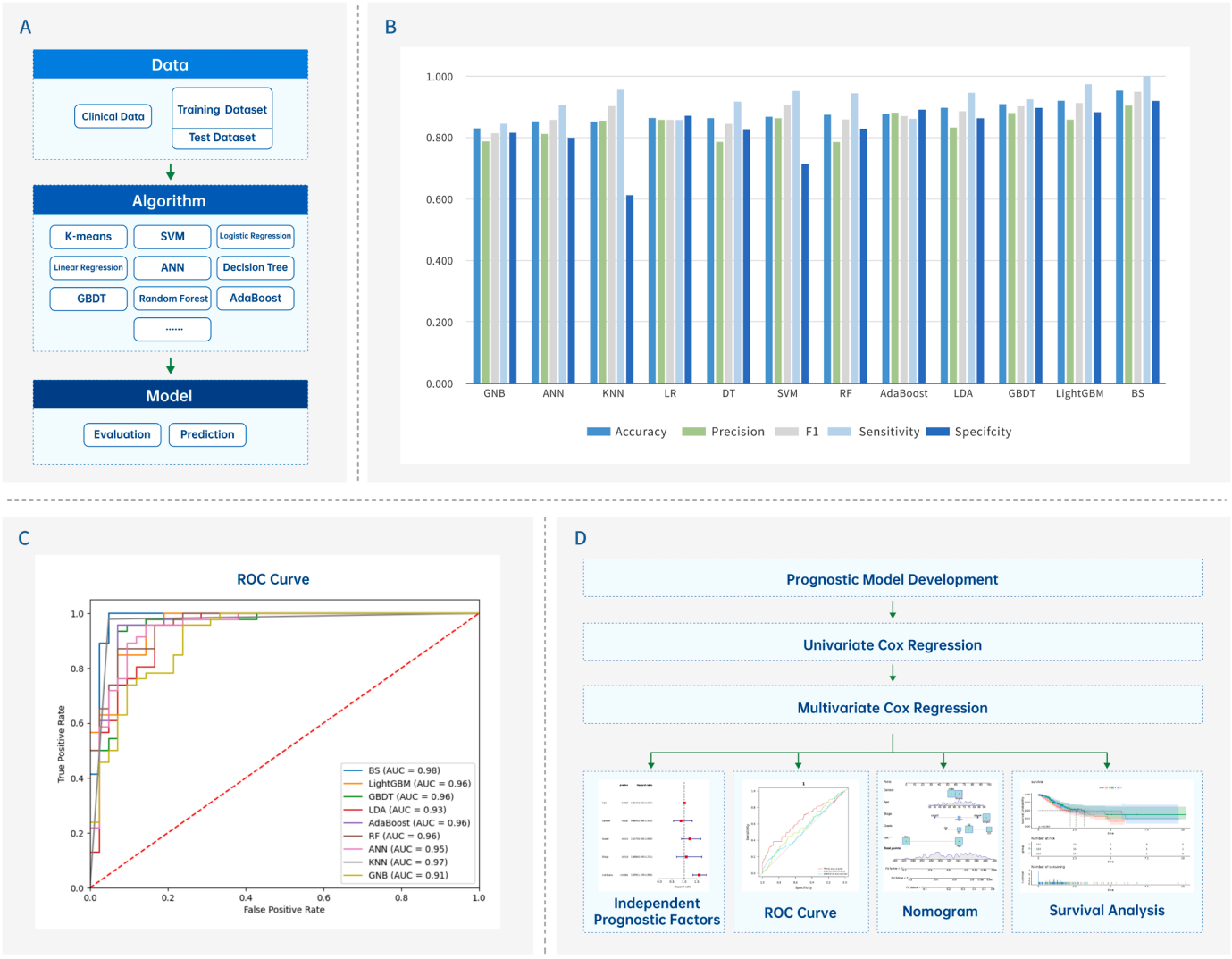


Fig. 6. Overview of Artificial Intelligence Modules. (A) This figure illustrates the machine learning models supported by the HiOmics platform and the modeling process. (B) Performance comparison of 12 different binary classification machine learning algorithms in terms of accuracy, precision, sensitivity, and specificity. (C) The ROC curves and AUC values for nine of these machine learning models. (D) This figure demonstrates the construction and evaluation process of clinical prognosis models on the HiOmics platform.

3.2.2. *Use case 2: advanced analysis module*

The advanced analysis module includes 29 plugins, covering various functionalities, such as prognosis model construction, immune infiltration analysis, Mendelian randomization, GO enrichment analysis, co-expression analysis, and phylogenetic tree analysis (Fig. 3B).

In this work, we utilize TCGA's bladder urothelial carcinoma dataset [40] to demonstrate the functionality of the "Each-sample Immune Infiltration Analysis" plugin. Users accessing the plugin webpage are required to input the gene expression matrix and grouping files, and then set the relevant parameters before initiating the analysis. The generated output comprises enrichment scores and expression heatmaps representing various cell types across different samples, along with boxplots displaying scores of different cells across distinct groups (with and without p-values), as depicted in Fig. 4.

3.2.3. *Use case 3: multi-omics integration analysis module*

Beyond single omics analysis, HiOmics prioritizes the integration of multiple omics data. Presently, the Integration Omics Analysis module provides 13 distinct plugins tailored for joint multi-omics analysis. Here, the NCI-60 cancer cell line metabolomics and gene expression data [41,42] were retrieved to demonstrate the functionality of the "Integration Analysis of Transcriptomics and Metabolomics Data" plugin, depicted in Fig. 5. To initiate the process, start by clicking on "Select File" to input the metabolite expression (abundance) matrix, transcriptome expression matrix, and the sample phenotype file, which contains outcome data associated with each sample. Next, manually enter or select from a dropdown menu the column name in the sample phenotype file that necessitates calculation, the data type of this column (discrete or continuous), and specify the names of the metabolites and genes to be plotted on the correlation graph. Once configured, click "Run" to submit the data for background processing. The generated output comprises five graphs and a table, as illustrated in Fig. 5.

3.2.4. *Use Case 4: Artificial intelligence (AI) modeling module*

3.2.4.1. Machine learning models. Numerous biological problems can be reformulated as data analysis and pattern recognition problems. The AI modeling module of HiOmics has incorporated interactive plugins for 32 widely used machine learning algorithms, encompassing artificial neural networks, support vector machines, random forests, and AdaBoost. These plugins enable users to effectively manage and analyze substantial volumes of biomedical data.

Table 2
Performance Comparison of Different Binary Classification Machine Learning Models.

APP Name	Accuracy	Precision	Sensitivity	Specificity	F1
k-Nearest Neighbors (KNN)	0.852	0.813	0.907	0.800	0.857
Artificial Neural Network (ANN)	0.852	0.854	0.953	0.611	0.901
Gaussian Naive Bayes (GNB)	0.864	0.833	0.909	0.818	0.870
Logistic Regression (LR)	0.864	0.857	0.857	0.870	0.857
Decision Tree (DT)	0.864	0.786	0.917	0.827	0.846
Support Vector Machine (SVM)	0.869	0.864	0.950	0.714	0.905
Random Forest (RF)	0.875	0.786	0.943	0.830	0.857
AdaBoost	0.875	0.881	0.860	0.889	0.871
Linear Discriminant Analysis (LDA)	0.898	0.833	0.946	0.863	0.886
Gradient Boosting Decision Tree (GBDT)	0.909	0.881	0.925	0.896	0.902
Light GBM	0.920	0.857	0.973	0.882	0.911
Binary Stacking (BS)	0.955	0.905	1.000	0.920	0.950

To showcase the functionality of the AI modeling tools of HiOmics, we chose a publicly available heart dataset [43], utilizing twelve binary classification machine learning algorithms to assess their performance within the HiOmics data analysis framework. Notably, our objective was to showcase the application of various machine learning plugins, rather than aiming for superior predictive performance compared to state-of-the-art methods. The algorithm parameters were set as follows: a test set split ratio of 0.2, non-shuffled data rows, utilization of fivefold cross-validation, and no custom model parameters were defined. Fig. 6A depicts a simplified machine learning workflow. The performance metrics, including accuracy, precision, sensitivity, and specificity for different algorithms, are presented in Table 2 and Fig. 6B, while Fig. 6C illustrates the ROC curves for nine out of the twelve algorithms.

3.2.4.2. Clinical prognostic models. The Prognosis Model module offers clinical doctors a variety of visual plugins, such as Lasso Regression and Cox Regression Model, to investigate the factors influencing different disease outcomes. These plugins aid in assessing the probability of individual outcome events by utilizing multiple predictive factors. Fig. 6D provides a simplified example illustrating the functionality of this module. Initially, the "Univariate COX Model" plugin is employed to identify substantial survival-related factors. Subsequently, statistically significant factors are chosen for analysis using the ".

Multivariable COX Model" plugin, generating forest plots and column line graphs. Moreover, the "Independent Prognostic Analysis" plugin evaluates whether specific factors are independent of other clinical characteristics, serving as independent prognostic factors. The " ROC Curve Analysis " plugin assesses the predictive accuracy of various clinical factors and risk scores for survival time. Additionally, the "Survival Curve and Risk Curve" plugin evaluates the accuracy of the prognosis prediction model while identifying substantial differences in overall survival time between high- and low-risk groups classified by the model.

3.3. Convenient workflow

HiOmics has developed a collection of over 10 bioinformatics data processing workflows (Fig. 7), encompassing transcriptome analysis, metagenomic analysis, and genetic variation analysis, all built upon the standardized workflow language, WDL. These workflows address users' essential requirements for executing complex data analysis. For example, the microbiome amplicon analysis workflow integrates vital software and databases for various steps, including paired-end sequence merging, barcode and primer removal, quality control, denoising for acquiring ASV representative sequences, taxonomic annotation, diversity analysis, functional prediction, and differential analysis using LEfSe [44,45]. All these steps are conveniently presented on a single page, allowing users to seamlessly navigate through the workflow and obtain desired results effortlessly. These workflows substantially streamline comprehensive data mining, particularly benefiting users without scripting language or workflow construction expertise.

4. Discussion

After the completion of the Human Genome Project, scientists expanded their focus from the genome to other 'omics' data types, —such as the transcriptome for gene expression, the epigenome for epigenetic markers, the proteome for protein production, and the metabolome for metabolic functional products. Although single omics data analysis usually involves correlation [46], researchers are increasingly collecting and analyzing multiple omics datasets

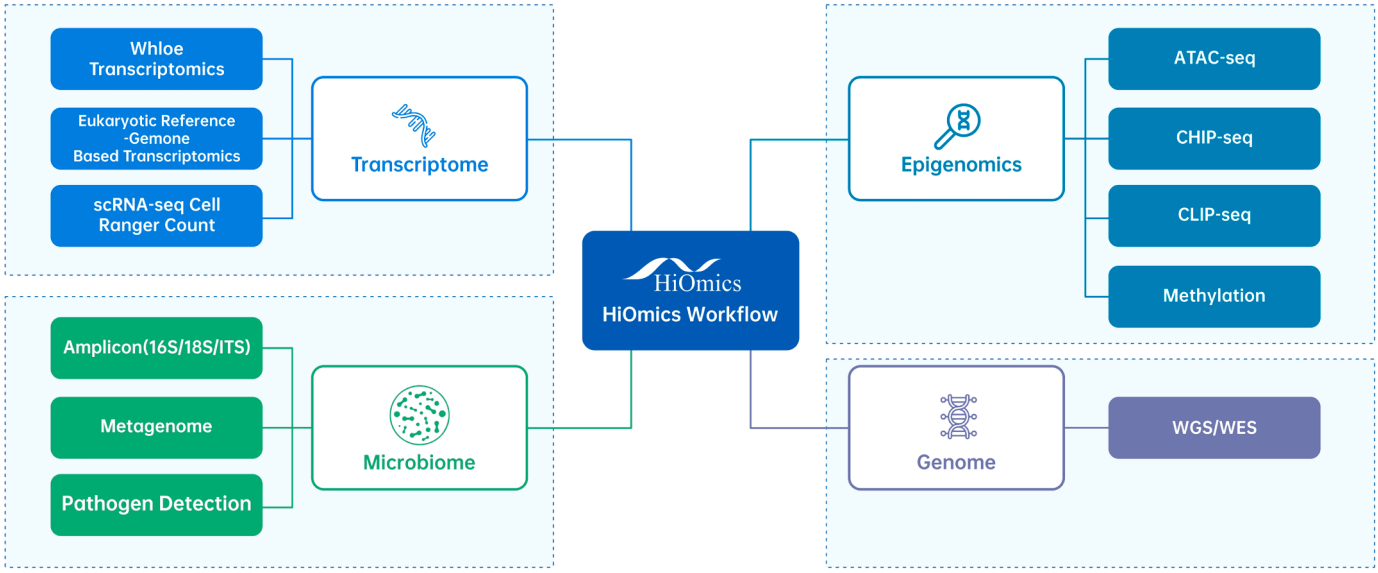


Fig. 7. All workflows of HiOmics.

simultaneously to attain a more comprehensive understanding. Computer science plays a critical role in handling large-scale multi-omics data. Desktop applications, programming languages, and server-based network tools are frequently utilized in bioinformatics but have inherent limitations. Some tools lack comprehensive data analysis capabilities, while others demand users to possess computer knowledge and familiarity with programming languages and Linux. Additionally, physical server constraints hinder large-scale data processing. Fortunately, cloud-based bioinformatics data analysis websites have emerged to counter these challenges. RAP [47] is a web-based tool tailored specifically for RNA-seq analysis, offering standard and customized workflows. Hiplot [23] excels in interactive analysis and visualization of lightweight single-omics data. Galaxy [48] empowers users to construct and manage their data analysis workflows, albeit requiring proficiency in multiple tools.

In this work, we introduce HiOmics, a comprehensive cloud-based bioinformatics data analysis platform leveraging state-of-the-art technologies, such as Docker containers, the WDL language, and the Cromwell engine. Tailored to meet the demands of multi-omics data processing and visualization, HiOmics boasts an extensive suite of nearly 300 data processing and visualization tools. These tools encompass diverse analysis requirements, featuring specialized toolsets for specific purposes and rigorously tested, reproducible analysis workflows utilizing multiple software packages. HiOmics caters to upstream analysis of raw sequencing data and downstream analysis, ensuring the delivery of high-quality visual outcomes. Its capabilities span from single omics analysis to intricate multi-omics integrative analysis, efficiently managing datasets of varying sizes, from small-scale to large-scale multi-sample data. Designed with a user-friendly graphical interface, HiOmics simplifies the processing of vast biomedical data, accommodating users without programming experience. With only three simple steps, researchers can effectively analyze and explore omics data, thereby eliminating barriers and empowering in-depth analysis to uncover new insights. For reliability, portability, and efficiency in bioinformatics data analysis workflows, HiOmics integrates WDL + Cromwell and Docker container technology. WDL + Cromwell offers robust workflow description and management capabilities, while Docker containers ensure independent and dependable workflow deployment and execution environments. This integration resolves compatibility issues across diverse environments, allowing for complex data analysis workflows while maintaining consistency and standardization. Additionally, it minimizes maintenance costs and workflow-related risks. To manage the growing volume of omics data, HiOmics incorporates public cloud object storage technology and batch computing techniques, ensuring users' resource independence and enabling efficient processing of vast datasets. Furthermore, standardized data format validation enhances the user experience, improving data analysis reliability and effectiveness within the platform.

Despite these advancements, HiOmics acknowledges the ongoing demand for personalized analysis plugins tailored to specific research fields. Hence, HiOmics is dedicated to developing additional plugins to cater to the diverse needs of its users. Concurrently, enhancing the speed and responsiveness of real-time interactions remains a focal point for HiOmics. Furthermore, recognizing the trend toward deep learning in biomedical data analysis [49–53], we aim to introduce more deep learning plugins, expanding our users' options for artificial intelligence modeling. Through continuous efforts, we aim to enhance HiOmics, providing users with more efficient, flexible, and convenient data analysis and visualization capabilities.

5. Conclusion

In general, HiOmics offers an integrated biomedical data analysis and visualization service, streamlining researcher workflows and enabling more convenient scientific research. Currently, users can freely access all of HiOmics functions by simply registering an account. However, for large data computations involving cloud computing and storage costs, we plan to introduce charges for this specific service in the future. This decision aims to cover the infrastructure costs associated with cloud service providers. Nevertheless, other tools and services will remain free to support a broader range of researchers in their studies.

Funding

This work is supported by the National Natural Science Foundation of China (82160537) and the Key Research and Development Program of Guangxi (2021AB12032).

CRediT authorship contribution statement

Wen Li: Writing – original draft. **Zhining Zhang:** Methodology, Supervision, Software. **Bo Xie:** Software, Validation. **Yunlin He:** Software. **Kangming He:** Software. **Hong Qiu:** Investigation, Resources. **Zhiwei Lu:** Visualization. **Chunlan Jiang:** Validation. X**uanyu Pan:** Software. **Yuxiao He:** Validation. **Wenyu Hu:** Validation. **Wenjian Liu:** Writing – review & editing. **Tengcheng Que:** Writing – review & editing. **Yanling Hu:** Supervision, Project administration, Funding acquisition.

Declaration of Competing interest

The author declares no competing interests.

Availability

The website can be freely accessed at https://henbio.com/en/tools and https://henbio.com/tools. Furthermore, the available open-source code of the website is located at https://github.com/yongkangning/HiOmics.

References

[1] Pandey D, Onkara Perumal P. A scoping review on deep learning for next-generation RNA-Seq. Data analysis. Funct Integr Genom 2023;23:134.
[2] Sathyanarayanan A, Mueller TT, Ali Moni M, Schueler K, Baune BT, et al. Multi-omics data integration methods and their applications in psychiatric disorders. Eur Neuropsychopharmacol 2023;69:26–46.
[3] Sucre A, Martinez M, Garin-Muga A. OmicSDK-transcriptomics: a web platform for transcriptomics data analysis. Stud Health Technol Inform 2023;302:1042–6.
[4] Leite ML, de Loiola Costa LS, Cunha VA, Kreniski V, de Oliveira Braga Filho M, et al. , Artificial intelligence and the future of life sciences. Drug Discov Today 2021;26:2515–26.
[5] Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform 2022;23.
[6] Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res 2018;46:D956–d963.
[7] Qian Y, Li L, Sun Z, Liu J, Yuan W, et al. A multi-omics view of the complex mechanism of vascular calcification. Biomed Pharmacother = Biomedecine Pharmacother 2021;135:111192.
[8] Pittard WS, Li S. The essential toolbox of data science: python, R, git, and docker. Methods Mol Biol (Clifton, N J) 2020;2104:265–311.
[9] Procter JB, Carstairs GM, Soares B, Mourão K, Ofoegbu TC, et al. Alignment of biological sequences with jalview. Mult Seq Alignment 2021:203–24.
[10] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35:1547–9.
[11] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
[12] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.
[13] Kern F, Fehlmann T, Keller A. On the lifetime of bioinformatics web services. Nucleic Acids Res 2020;48:12523–33.
[14] Zhang Z, Li H, Jiang S, Li R, Li W, et al. A survey and evaluation of Web-based tools/databases for variant analysis of TCGA data. Brief Bioinform 2019;20:1524–41.
[15] Chen T, Liu YX, Huang L. ImageGP: An easy-to-use data visualization web server for scientific researchers. iMeta 2022;1.
[16] Nelson JW, Sklenar J, Barnes AP, Minnier J. The START App: a web-based RNAseq analysis and visualization resource. Bioinformatics 2017;33:447–9.

[17] Velmeshev D, Lally P, Magistri M, Faghihi MA. CANEapp: a user-friendly application for automated next generation transcriptomic data analysis. BMC Genom 2016;17.
[18] Chong J, Soufan O, Li C, Caraus I, Li S, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 2018;46:W486–94.
[19] Chen IA, Chu K, Palaniappan K, Pillay M, Ratner A, et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. Nucleic Acids Res 2019;47:D666–77.
[20] Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10:1523.
[21] Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, et al. Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol 2010;89.
[22] Sangerbox. Available from: ⟨http://sangerbox.com/⟩, access date: June 20, 2023.
[23] Li J, Miao B, Wang S, Dong W, Xu H, et al. Hiplot: a comprehensive and easy-to-use web service for boosting publication-ready biomedical data visualization. Brief Bioinforma 2022;(23 ().
[24] Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vazquez-Baeza Y, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 2018;15:796–8.
[25] Yukselen O, Turkyilmaz O, Ozturk AR, Garber M, Kucukural A. DolphinNext: a distributed data processing platform for high throughput genomics. BMC Genom 2020;21:310.
[26] Element Plus: A Vue 3 UI framework. Available from: ⟨https://element-plus.org/zh-CN/⟩, Access date: June 20, 2023.
[27] Docker: Accelerated, Containerized Application Development. Available from: ⟨https://www.docker.com⟩, Access date: June 20, 2023.
[28] OpenWDL: Community Driven Open-development Workflow Language. Available from: ⟨https://openwdl.org/⟩, Access date: June 20, 2023.
[29] The Go Programming Language. Available from: ⟨https://golang.org/⟩, Access date: June 20, 2023.
[30] Suetake H, Fukusato T, Igarashi T, Ohta T. A workflow reproducibility scale for automatic validation of biological interpretation results. Gigascience 2022;12.
[31] Baker M. 1,500 scientists lift the lid on reproducibility. Nature 2016;533:452–4.
[32] Matelsky J, Kiar G, Johnson E, Rivera C, Toma M, et al. Container-based clinical solutions for portable and reproducible image analysis. J Digit Imaging 2018;31:315–20.
[33] You L, Sun H. Research and design of docker technology based authority management system. Comput Intell Neurosci 2022;2022:5325694.
[34] Cromwell: A. Workflow Management System. Available from: ⟨https://cromwell.readthedocs.io/en/stable/⟩, access date: June 20, 2023.
[35] Cao Y, Li L, Xu M, Feng Z, Sun X, et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res 2020;30:717–31.
[36] Tang Z, Fan W, Li Q, Wang D, Wen M, et al. MVIP: multi-omics portal of viral infection. Nucleic Acids Res 2022;50:D817–d827.
[37] Doricchi A, Platnich CM, Gimpel A, Horn F, Earle M, et al. Emerging approaches to DNA data storage: challenges and prospects. ACS nano 2022;16:17552–71.
[38] Dotan E, Alburquerque M, Wygoda E, Huchon D, Pupko T. GenomeFLTR: filtering reads made easy. Nucleic Acids Res 2023.
[39] Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet 2018;19:325.
[40] [dataset] TCGA Bladder Urothelial Carcinoma data, phs000178.v11.p8. Available from: ⟨https://portal.gdc.cancer.gov/projects/TCGA-BLCA⟩, Access date: June 20, 2023.
[41] [dataset] Molecular Target Data - NCI DTP Data - NCI Wiki. Available from: ⟨https://wiki.nci.nih.gov/display/ncidtpdata/molecular+target+data⟩, Access date: June 20, 2023.
[42] Siddiqui JK, Baskin E, Liu M, Cantemir-Stone CZ, Zhang B, et al. IntLIM: integration using linear models of metabolomics and gene expression data. BMC Bioinforma 2018;(19 ().
[43] [dataset] UCI Machine Learning Repository. Cleveland Heart Disease Database. Available from: ⟨https://archive.ics.uci.edu/dataset/45/heart+disease⟩, Access date: June 20, 2023.
[44] Liu YX, Qin Y, Chen T, Lu M, Qian X, et al. A practical guide to amplicon and metagenomic analysis of microbiome data. Protein Cell 2021;12:315–30.
[45] Liu YX, Chen L, Ma T, Li X, Zheng M, et al. EasyAmplicon: an easy-to-use, open-source, reproducible, and community-based pipeline for amplicon data analysis in microbiome research. iMeta 2023;2.
[46] Picard M, Scott-Boyer MP, Bodein A, Perin O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J 2021;19:3735–46.
[47] D'Antonio M, D'Onorio De Meo P, Pallocca M, Picardi E, D'Erchia AM, et al. RAP: RNA-Seq analysis pipeline, a new cloud-based NGS web application. BMC Genom 2015;16:S3.
[48] Nekrutenko A, Taylor J, Goecks J, Blankenberg D, Clements D, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res 2020;48:W395–402.
[49] Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, et al. Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 2021;13:152.
[50] Berrar D, Dubitzky W. Deep learning in bioinformatics and biomedicine. Brief Bioinform 2021;22:1513–4.
[51] Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet 2019;20:389–403.
[52] Sen P, Lamichhane S, Mathema VB, McGlinchey A, Dickens AM, et al. Deep learning meets metabolomics: a methodological perspective. Brief Bioinform 2021;22:1531–42.
[53] Wen B, Zeng WF, Liao Y, Shi Z, Savage SR, et al. Deep learning in proteomics. Proteomics 2020;20:e1900335.