Chapter 18
Analyzing Multi-Omic Data with Integrative Platforms

Yan Zou

Abstract The exponential growth of molecular data put up a new challenge to the biologists. The difficulty in data storage, processing, transmission, connection, and the demand for multi-omic data analysis motivates scientists to set up integrative platforms and workflows. Here we introduce some prominent integrated bioinformatics platforms. Among them, Galaxy will be carefully discussed for its development, core values, flexible workflows, and relevant framework applications.

Keywords Multi-omic · Integrative platform · Galaxy

18.1 Integrating Diverse Tools into Bioinformatics Platforms

The availability of high-throughput sequencing technologies and high-resolution mass spectrometry in genomics, transcriptomics, proteomics, metabolomics, and phenomics promotes a large-scale multi-omic data complex. The information content is higher in integrated analysis, which requires connecting and comparing data in different omics, than in any of the molecular levels studied separately. The exponential growths of molecular data, however, put up a new challenge for biologists. The storage, processing, transmission, connection, and analysis of these data complex demand the use of disparate software programs and require computational resources beyond the capacity of many biological laboratories (Chen and Hofestädt 2014; Boekel et al. 2015). Furthermore, disparate software requires extra training time when a new analysis is conducted and standardizing diverse data formats into the identical one that program requests bring further inconvenience. For these reasons, multi-omic platforms are emerged to embrace the complexity that is associated with the exponentially increasing amounts of data.

Y. Zou (✉)
College of Life Sciences, Zhejiang University, Hangzhou, P. R. China
e-mail: 3160102154@zju.edu.cn

M. Chen, R. Hofestädt (eds.), *Integrative Bioinformatics*,
https://doi.org/10.1007/978-981-16-6795-4_18

Some prominent software platforms have already shown their merits in coping with these problems. Some are compatible with all data regions, and some are specially designed for specific regions, such as cancer, plant cells, and viruses.

18.1.1 General Integrative Platforms and Workflows

Most bioinformatics tools used in genomics, transcriptomics, and proteomics are set in programming and command-line environments, which can be time-consuming and complex for researchers to get started. Ideal platforms and workflows such as Galaxy, Taverna, and Snakemake are created to meet the urgent need for the interactive analyses of big biological data.

18.1.1.1 Galaxy

Galaxy is a scientific workflow, data integration, and data and analysis publishing web-based platform established in 2005 (Blankenberg et al. 2011). Its graphical query interface combined with customized data storage can simplify the process Schatz (2010). Its development, core values, and flexible workflows will be carefully discussed in Sect. 18.2.

18.1.1.2 Snakemake

Snakemake (available in https://snakemake.readthedocs.io/en/stable/) is a Python-based scalable bioinformatics workflow engine published in 2012. It can scale from single-core workstations to compute clusters without modifying the workflow. It is the first system to support the use of automatically inferred multiple named wildcards (or variables) in input and output filenames (Köster and Rahmann 2018). Interaction between Snakemake and those installed in local or web-based tools is also available when both support the input and output data formats. In recent years, some Snakemake extensions, such as RASflow (Zhang and Jonassen 2020) and Sequanix (Desvillechabrol et al. 2018), build modular analysis workflow and establish graphical interfaces to help Snakemake be more flexible and user-friendly.

18.1.1.3 Taverna

Taverna (available in https://incubator.apache.org/projects/taverna.html) is a tool for the composition and enactment of bioinformatics workflows. Taverna includes a workbench application that provides a graphical user interface for the composition of workflows (Oinn et al. 2004). Scientists can organize their workflows in a new language called the simple conceptual unified flow language (Scufl). It can integrate with the bioinformatics resource shared as Web services among the community. Taverna and Galaxy, two workflow systems widely accepted and applied by the bioinformatics community, can also be integrated into a single environment, Tavaxy (available in https://www.tavaxy.org/) (Abouelhoda et al. 2012).

18.1.2 Integrative Platforms for Specialized Data

Some integrative platforms are specially established for analyzing data from specific regions, such as cancer, virus, plants, and fungi. Combined with gene and protein expression with signaling pathways and cell characteristics, these platforms contribute professional and accessible means for biologists to process data.

18.1.2.1 Combine Integrative Platforms with Clinical Data for Cancer Research

Distinct signaling pathways and altered molecular functions in cancer cells and clusters are displayed in the integrated analyses of molecular data. The platforms built specifically for cancer cell data analysis allow cancer researchers to interactively explore altered gene sets and signaling pathways (Gao et al. 2013). What makes the platforms driven by cancer cell studies distinguished is their strong connection with clinical outcomes and potential. Genomic, metabonomics, and clinical data might be used to identify novel patient subgroups. Clinical therapy can be tailored for each patient when statistical models are produced and treatment strategies are evaluated based on stratified patient groups (Kristensen et al. 2014).

The cBioPortal for Cancer Genomics (cBioPortal, http://cbioportal.org) is one of the widely used integrative platforms specialized for analyzing cancer-related data. The cBioPortal is established for integrative analysis of cancer genomics and clinical patient profiles. With 15 provisional TCGA (The Cancer Genome Atlas) datasets and other open datasets contained, the web-based cBioPortal is uniquely designed to store every single data in the gene level and combine these data with available de-identified clinical data. The fundamental abstraction of this platform is the concept of altered genes (Cerami et al. 2012), which is used to help users simplify the mixed and complicated current datasets and develop genomic hypotheses proceeding from genetic alterations across samples, genes, and pathways.

Among other platforms targeted in cancer cell research, Web-TCGA can uniquely provide methylation analyses (Deng et al. 2016); Firebrowse (http://firebrowse.org/) can also characterize and identify genomic patterns in human cancer models through visual and programmatical tools.

18.1.2.2 Specialized Integrative Platforms in Other Fields

Integrative platforms that focus on particular fields integrate with the other database resources in their domains for further research convenience. Integrating bioinformatics resource for fungi and oomycetes, FungiDB (fungidb.org) is a free online platform for data mining and functional genomics analysis which combines Eukaryotic Pathogen Genomics Database Resource (Basenko et al. 2018).

Scientists also use bioinformatics platforms as a tool for international cooperation in urgent issues. For the pharmaceutical development and antiviral drug prediction for the COVID-19 virus, Virus-CKB (https://www.cbligand.org/g/virus-ckb) is developed as a viral-associated disease-specific chemogenomics knowledgebase (Virus-CKB), which describes the chemical molecules, genes, and proteins involved in viral-associated diseases regulation (Feng et al. 2021).

18.2 Galaxy: A Widely Accepted General Bioinformatics System

18.2.1 Introduction

As has been mentioned in Sect. 18.1.1.1, Galaxy is a bioinformatics scientific workflow and data analysis platform, which is created in 2005. It is developed by the Galaxy team at Penn State, Johns Hopkins University, Oregon Health and Science University, and the Galaxy Community using Python language.

Galaxy was initially set up for genomics and transcriptomics data analysis from the very beginning. Nevertheless, with the maturation of proteomic and metabolomic technologies, multi-omic applications started to emerge after a few years since the Galaxy was created. Now it has assembled tools in multiple domains, such as gene expression, proteomics, epigenomics, and transcriptomics. It also contains cross-domain tools, including ecology, climate science, and computational chemistry. More than 7500 tools (Jalili et al. 2020) have been contributed to the Galaxy ToolShed till January 2020.

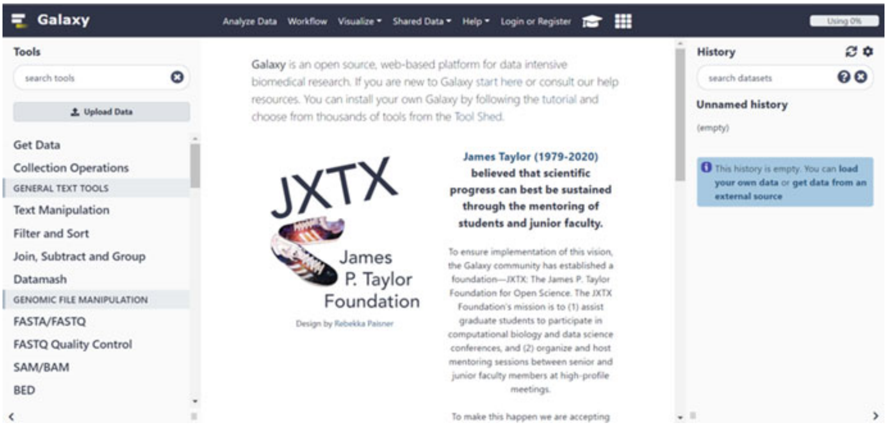
Galaxy now has a prosperous scientific community. The community keeps organizing conferences and meetings with Galaxy-related content and sharing tool tutorials in The Galaxy Training Network. With more than 9000 total publications, including over 7500 journal articles, 500 books, and 400 conference papers by 2020 (Jalili et al. 2020), this free and open-source platform has created a community for biology researchers.

18.2.2 How to Use Galaxy

Researchers can directly use the workspace of the Galaxy platform to conduct data analysis. Galaxy can be operated both on the web and locally. Between them, the webpage is more welcomed for its convenience.

As shown in Fig. 18.1, the webpage workspace of Galaxy is separated into four parts. Right on the top of the website is the navigation bar, which provides users with easy access to data processing, including analyzing, workflows, visualization, and user function for sharing data and tutorials. On the left is the analysis tool panel in which users can apply tools to their data. The component in the middle is the detail interface, where users set up and adjust different datasets, features, and filters for analysis. The history panel on the right of the workspace shows the status of the generation of the datasets. Users can also search the analysis history and extract workflows from the histories.

Users can create workflows based on their data analysis process. A workflow is a series of tools and dataset actions that run in sequence as a batch operation (Goecks et al. 2010). In Galaxy, workflows can be created from scratch using the workflow editor or generated from the analysis already completed in history Blankenberg et al. (2010). A successfully designed workflow can be continuously reused for the future analysis, enhancing reproducibility by applying the same methods to all of the users' data.



Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

James Taylor (1979-2020) believed that scientific progress can best be sustained through the mentoring of students and junior faculty.

To ensure implementation of this vision, the Galaxy community has established a foundation—JXTX: The James P. Taylor Foundation for Open Science. The JXTX Foundation's mission is to (1) assist graduate students to participate in computational biology and data science conferences, and (2) organize and host mentoring sessions between senior and junior faculty members at high-profile meetings.

To make this happen we are accepting

This history is empty. You can load your own data or get data from an external source

Fig. 18.1 The web-based workspace of the Galaxy platform. In the left column, users can choose the genomic tools. The detailed analysis content will be shown in the middle. The analysis history is in the right column, and users can search datasets using the search bar at the top of the panel

18.2.3 Key Requirements in Designing Galaxy

For most researchers, the use of disparate software programs and extra software training demands time and effort. The required computational resources sometimes will reach out of the capacity of most biological research laboratories. The ideal platform should meet these scientists' basic needs and keep the platform vibrant and easy to use. Five characteristics are crucial for building a thriving platform, including the flexibility to accommodate constantly evolving data types and emerging software across omics domains, reproducibility, open and free access, and long-term sustainability (Boekel et al. 2015).

Flexibility is the first general need that the developers plan to meet. Specifically, the platform needs to be open, extendable, and amenable to heterogeneous computing environments. To resolve this issue, the developer group has combined Linux-based software with Windows-based software to ensure that the platform could function well in multiple working environments.

Another distinct requirement for the platform is that it should have the ability to operate complex and multistep workflows automatically with different software. The platform can use quality control methods to evaluate the tool quality and integration efficiency.

The compatibility in high-performance computing and cloud environments makes the platform scalable to the established sequencing databases. Its large-memory allocation is integrated with multiple storage infrastructures.

One of the crucial characteristics for a bioinformatics platform to expand its lifespan is its community sharing. The publication and sharing of complete workflows not only promotes the dissemination and reproducibility of the workflows but also enhances the transparency of the data and its processing. The attention to data provenance also guarantees this.

The last essential requirement for building the desired platform is the wide adaption and sustainable user-friendly interface. Its web-based graphical user interface lowers the difficulty for the researchers with limited computational expertise to operate. The platform is sustained by the scientist's community rather than a single developer group, and each developer can publish their software and designs of workflows on their own.

18.2.4 Applications Based on Galaxy Platform

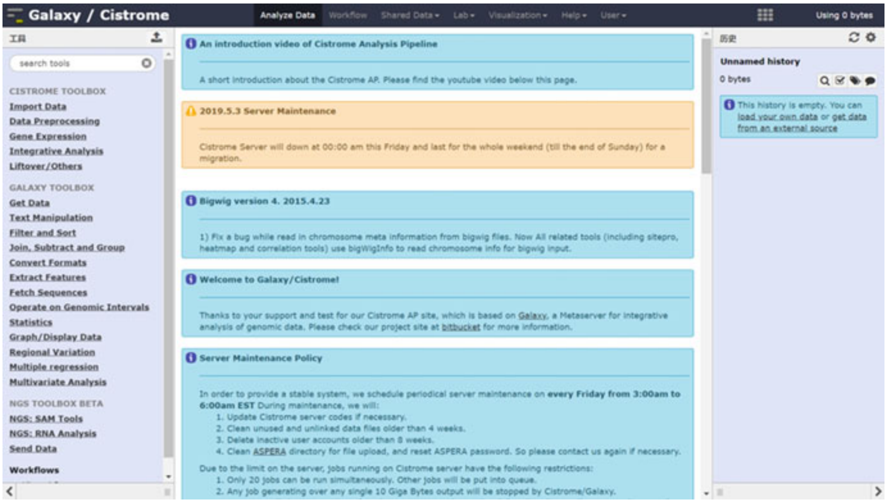
Some data analysis applications derived from Galaxy emerge to resolve typical questions. Here we introduce three applications of Galaxy: Cistrome, a new integrative platform based on Galaxy frameworks; RepeatExplorer, a computational pipeline or component aiming at repetitive DNA; and CloudMan, a cloud resource management system, as well as BioBlend, an automating pipeline analyses within Galaxy and CloudMan.

18.2.4.1 Cistrome: Galaxy-Based Integrative Platforms for Transcriptional Regulation

Chromatin immunoprecipitation (ChIP) combined with microarrays (ChIP-chip) and ChIP combined with NGS (ChIP-seq) are used for identifying cistromes, which refers to the set of cis-acting targets of a trans-acting factor on a genome-wide scale. However, the analysis of cistrome data requires both the hardware resources from the lab and the computational skills of the researchers to achieve the analyzing algorithms.

Under the conditions above, Cistrome (http://cistrome.org/ap/) has been built to provide a flexible bioinformatics workbench. Cistrome, an integrative platform for transcriptional regulation studies, is specifically designed for downstream data analysis accompanied by ChIP-chip or ChIP-seq technologies and includes fundamental analyses from peak calling to motif detection (Fig. 18.2).

To accomplish this, Galaxy framework provides a user-friendly, reproducible, and transparent workbench, on which the scientists can share, incorporate, and publish their data. Furthermore, its infrastructure makes each Cistrome tool to remember the run-time parameters in the server (Liu et al. 2011).



An introduction video of Cistrome Analysis Pipeline

A short introduction about the Cistrome AP. Please find the youtube video below this page.

2019.5.3 Server Maintenance

Cistrome Server will down at 00:00 am this Friday and last for the whole weekend (till the end of Sunday) for a migration.

Bigwig version 4. 2015.4.23

1) Fix a bug while read in chromosome meta information from bigwig files. Now All related tools (including sitepro, heatmap and correlation tools) use bigWigInfo to read chromosome info for bigwig input.

Welcome to Galaxy/Cistrome!

Thanks to your support and test for our Cistrome AP site, which is based on Galaxy, a Metaserver for integrative analysis of genomic data. Please check our project site at bitbucket for more information.

Server Maintenance Policy

In order to provide a stable system, we schedule periodical server maintenance on **every Friday from 3:00am to 6:00am EST** During maintenance, we will:

1. Update Cistrome server codes if necessary.
2. Clean unused and unlinked data files older than 4 weeks.
3. Delete inactive user accounts older than 8 weeks.
4. Clean ASPERA directory for file upload, and reset ASPERA password. So please contact us again if necessary.

Due to the limit on the server, jobs running on Cistrome server have the following restrictions:

1. Only 20 jobs can be run simultaneously. Other jobs will be put into queue.
2. Any job generating over any single 10 Giga Bytes output will be stopped by Cistrome/Galaxy.

This history is empty. You can load your own data or get data from an external source

Fig. 18.2 The web-based workspace of the Galaxy/Cistrome platform. In the left column, users can choose the available tools. The messages, tool options, and detailed analysis content will be shown in the middle. The analysis history is in the right column

18.2.4.2 RepeatExplorer: A Computational Pipeline for Characterization of Repetitive Elements

Repetitive DNA makes up a large part of eukaryotic nuclear genomes. The accurate quantification and sequence characterization of repetitive DNA is complex for most researchers due to the restricted computational resources and the lack of professional analyzing tools.

With the new approach for global repeat analysis developed by Novak et al. (Novák et al. 2010) and the availability of high-throughput sequencing data, Novak et al. developed a pipeline named RepeatExplorer (http://repeatexplorer.umbr.cas.cz/) for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads (Novak et al. 2013).

The main component of RepeatExplorer is the clustering pipeline, which performs all-to-all similarity comparisons of sequence reads followed by their graph-based clustering to identify groups of reads derived from repetitive elements.

Galaxy platform provides the adaption for the tools of the RepeatExplorer pipeline. These tools can be recombined to form specialized workflows. The Galaxy platform also facilitates easy execution, documentation, and sharing of analysis protocols and results.

18.2.4.3 CloudMan and BioBlend: A Cloud Resource Management System and an Automating Pipeline Analyses within Galaxy and CloudMan

With the availability of high-throughput sequencing data and the robust research future of analyzing sequence data, the computational infrastructure and support have gradually been a problem for researchers whose laboratory cannot reach the requirement in computing. Cloud computing, a computational model, is potential in the analysis of high-throughput sequencing data. However, the established projects are only targeted at specialized problems and are unsuitable for various computing circumstances.

CloudMan is an integrated solution that the researchers could create and control fully functional compute clusters with existing tools and packages provided on cloud resources. The intricacies of cloud computing resource acquisition, configuration, and scaling could be conducted on Amazon’s EC2 cloud infrastructure, and a personal computing cluster will be produced in minutes. (Afgan et al. 2010). The researchers have embedded Galaxy CloudMan on top of the Bio-Linux workstation machine image and integrated it with Galaxy.

BioBlend (http://bioblend.readthedocs.org/) is a unified API in a high-level language that wraps the functionality of Galaxy and CloudMan APIs (Sloggett et al. 2013). It is easier for researchers to automate end-to-end large data analysis using BioBlend, due to the convenient access for large datasets in the familiar Galaxy environment and the computing infrastructure provided.

References

Abouelhoda M, Issa SA, Ghanem M (2012) Tavaxy: integrating Taverna and Galaxy workflows with cloud computing support. BMC Bioinformatics 13:77

Afgan E, Baker D, Coraor N et al (2010) Galaxy CloudMan: delivering cloud compute clusters. BMC Bioinformatics 11:S4

Basenko E, Pulman J, Shanmugasundram A, Harb O, Crouch K, Starns D, Warrenfeltz S, Aurrecoechea C, Stoeckert C, Kissinger J, Roos D, Hertz-Fowler C (2018) FungiDB: an integrated bioinformatic resource for fungi and oomycetes. J Fungi 4(1):39

Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol 19:1–21

Blankenberg D, Coraor N, Von Kuster G, Taylor J, Nekrutenko A (2011) Integrating diverse databases into an unified analysis framework: a Galaxy approach. Database 2011:11

Boekel J, Chilton JM, Cooke IR, Horvatovich PL, Jagtap PD, Käll L, Lehtiö J, Lukasse P, Moerland PD, Griffin TJ (2015) Multi-omic data analysis using Galaxy. Nat Biotechnol 33(2):137–139

Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2(5):401–404

Chen M, Hofestädt R (2014) Approaches in integrative bioinformatics. Introduction. Springer, Berlin, pp 3–5

Deng M, Brägelmann J, Schultze JL, Perner S (2016) Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. BMC Bioinformatics 17:72

Desvillechabrol D, Legendre R, Rioualen C, Bouchier C, van Helden J, Kennedy S, Cokelaer T (2018) Sequanix: a dynamic graphical interface for Snakemake workflows. Bioinformatics 34(11):1934–1936

Feng Z, Chen M, Liang T, Shen M, Chen H, Xie X (2021) Virus-CKB: an integrated bioinformatics platform and analysis resource for COVID-19 research. Brief Bioinform 22(2):882–8955

Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6(269):l1

Goecks J, Nekrutenko A, Taylor J, Galaxy T (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):86

Jalili V, Afgan E, Gu Q, Clements D, Blankenberg D, Goecks J, Taylor J, Nekrutenko A (2020) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. Nucleic Acids Res 48(1):395–402

Köster J, Rahmann S (2018) Snakemake-a scalable bioinformatics workflow engine. Bioinformatics 34(20):3600

Kristensen VN, Lingjærde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A (2014) Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 14(5):299–313

Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, Zhang Y, Shin H, Wong SS, Ma J, Lei Y, Pape UJ, Poidinger M, Chen Y, Yeung K, Brown M, Turpaz Y, Liu XS (2011) Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol 12(8):83

Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29(6):792–793

Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, Carver T, Glover K, Pocock MR, Wipat A, Li P (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 20(17):3045–3054

Schatz MC (2010) The missing graphical user interface for genomics. Genome Biol 11(8):128

Sloggett C, Goonasekera N, Afgan E (2013) BioBlend: automating pipeline analyses within Galaxy and CloudMan. Bioinformatics 29(13):1685–1686

Zhang X, Jonassen I (2020) RASflow: an RNA-Seq analysis workflow with Snakemake. BMC Bioinformatics 21(1):110–119